PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

ISRAEL ANDRÉ LAURENSI ROSA

ESQUECIMENTO CATASTRÓFICO EM MODELOS PROFUNDOS PARA RECONHECIMENTO DE EMOÇÕES EM FACES: ABORDAGENS PARA INTEGRIDADE DE MEMÓRIA

DOUTORADO EM INFORMÁTICA
PUCPR

CURITIBA
2026

Israel André Laurensi Rosa

Esquecimento Catastrófico em Modelos Profundos para Reconhecimento de Emoções em Faces: Abordagens para Integridade de Memória

Tese de doutorado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de doutor em Informática.

Pontifícia Universidade Católica do Paraná – PUCPR

Programa de Pós-Graduação em Informática – PPGIa

Orientador: Prof. Dr. Alceu de Souza Britto Junior

Coorientador: Prof. Dr. Alessandro Lameiras Koerich

Curitiba – PR, Brasil

2026

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

R788e
2026

Rosa, Israel André Laurensi
Esquecimento catastrófico em modelos profundos para reconhecimento de emoções em faces: abordagens para integridade de memória/ Israel André Laurensi Rosa ; orientador: Alceu de Souza Britto Junior ; coorientador: Alessandro Lameiras Koerich -- 2026.
155 f. : il. ; 30 cm
Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2026.
Bibliografia: f. 138-152

1. Informática. 2. Reconhecimento facial (Computação). 3. Aprendizado do computador. 4. Inteligência artificial. I. Britto Junior, Alceu de Souza. II. Koerich, Alessandro Lameiras. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título

CDD 22. ed. – 004

Bibliotecária: Roberta Moritz Moreira – CRB 9/2247

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 30 de abril de 2026.

36-2026

DECLARAÇÃO

Declaro para os devidos fins, que **ISRAEL ANDRÉ LAURENSI ROSA** defendeu a tese de Doutorado intitulada "**ESQUECIMENTO CATASTRÓFICO EM MODELOS PROFUNDOS PARA RECONHECIMENTO DE EMOÇÕES EM FACES: ABORDAGENS PARA INTEGRIDADE DE MEMÓRIA"**, na área de concentração Ciência da Computação no dia 04 de dezembro de 2025, no qual foi aprovado.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade firmo a presente declaração.

gov.br Documento assinado digitalmente
JEAN PAUL BARDDAL
Data: 01/05/2026 18:49:34-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Jean Paul Barddal
Coordenador do Programa de Pós-Graduação em Informática

Agradecimentos

Agradeço ao meu orientador, Prof. Dr. Alceu de Souza Britto Jr., pelo apoio, orientação e principalmente paciência durante todo o desenvolvimento deste trabalho. Suas contribuições foram essenciais para a conclusão desta tese. Sua parceria e incentivo, mesmo nos momentos mais difíceis, foram fundamentais para que eu pudesse superar os desafios encontrados ao longo do caminho.

Agradeço também ao meu coorientador, Prof. Dr. Alessandro Lameiras Koerich, pelas grandiosas sugestões e discussões que enriqueceram muito este trabalho. Há muito o que ser dito sobre suas contribuições, que colocaram diversas vezes este trabalho no rumo certo.

Agradeço ao Prof. Dr. Manoel Camillo Penna, que me apresentou ao mundo da pesquisa e me incentivou a seguir rumo ao doutorado. Sua orientação e apoio foram fundamentais.

Agradeço a minha esposa, Carolina, pelo companheirismo e apoio inabalável durante toda minha jornada no doutorado. Sua presença foi essencial para que eu pudesse manter o foco e a motivação necessários para concluir este trabalho. Seu amor e compreensão em todo esse período caótico e desafiador me mantiveram certo de que eu poderia ir em frente. Voamos para o outro lado do mundo e estamos realizando nossos sonhos juntos. Obrigado por estar sempre ao meu lado.

Agradeço especialmente a minha mãe Leda, pelo amor, apoio incondicional e incentivo constante em minha jornada acadêmica. Mesmo distante, seu suporte foi fundamental para que eu pudesse alcançar meus objetivos. Agradeço também ao meu irmão, Samuel, e à minha cunhada, Ivana, pelo carinho e apoio ao longo dessa caminhada. Agradeço também aos meus avós, Dorilde e Salvador Laurensi, por me ensinarem que empatia, bondade e paciência são as maiores virtudes que alguém pode ter.

Agradeço a todos da minha família Laurensi e Lourenci por todo apoio, que contribuíram de alguma forma para que essa jornada fosse mais leve.

Agradeço também a minha família Zilli, por todo o carinho, apoio e momentos felizes compartilhados. Vocês sempre me acolheram de braços abertos, e sou muito grato por fazer parte dessa família. Obrigado Isabela, Leodimeri, Valdir, Fábio, Ana e Vitor. Sem vocês, essa jornada teria sido muito mais difícil.

Agradeço aos meus amigos que moram no meu coração Rafael, Mariana, Luciana, Andrea, Maria Eduarda, Hudson, Janaína, Henrique e Vitor T., por todo o apoio, incentivo e amizade ao longo dessa jornada. A presença de vocês foi fundamental para que eu pudesse

superar os desafios do doutorado.

Agradeço também aos meus queridos amigos que fiz na Mongólia, por me receberem tão bem e me ajudarem a me adaptar a um país tão diferente do meu. Família é onde encontramos conexões verdadeiras, e vocês se tornaram minha família longe de casa.

Agradeço à Pontifícia Universidade Católica do Paraná (PUCPR) e ao Programa de Pós-Graduação em Informática (PPGIa) pelo suporte institucional e pelos recursos disponibilizados para a realização desta pesquisa, assim como a todos os professores e colegas com quem tive a oportunidade de aprender e colaborar ao longo desses anos.

"*Encumbered forever by desire and ambition*
There's a hunger still unsatisfied
Our weary eyes still stray to the horizon
Though down this road we've been so many times."

("High Hopes", Pink Floyd)

Resumo

O reconhecimento de expressões faciais é fundamental no aprendizado de máquina, abrindo portas para diversas aplicações. Contudo, redes neurais convolucionais frequentemente sofrem com o esquecimento catastrófico, o que compromete a retenção de conhecimento previamente aprendido. Esta tese investiga o esquecimento catastrófico em redes neurais convolucionais aplicadas ao reconhecimento de expressões faciais no aprendizado contínuo. Propõe-se o método *Emotion-Centered generative replay* (ECgr), baseado em *pseudo-rehearsal* com imagens sintéticas geradas por *Wassertein Generative Adversarial Network with Gradient Penalty* (WGAN-GP). Três estratégias de garantia de qualidade são avaliadas: uma abordagem supervisionada, uma técnica não supervisionada por clusterização e uma otimização do espaço latente guiada por amostras validadas. Os resultados demonstram que o ECgr mitiga significativamente o esquecimento em comparação ao *fine-tuning* tradicional, com ganhos adicionais quando combinado à filtragem supervisionada. A abordagem se aproxima do desempenho de treinamento conjunto, mesmo sem reutilizar dados originais das tarefas anteriores. Os experimentos confirmam as hipóteses de que o ECgr melhora o desempenho em aprendizado contínuo, que a filtragem supervisionada potencializa esse efeito e que a abordagem generaliza para outras bases de dados, como a Split-MNIST. No cenário incremental de domínio da Split-MNIST, o método ECgr, mesmo sem reutilizar dados reais, superou técnicas clássicas baseadas em regularização e aproximou-se de estratégias que utilizam parte dos dados reais das tarefas anteriores e métodos generativos mais complexos. Esses resultados reforçam a robustez e a generalização da abordagem proposta, destacando seu potencial como solução eficaz e versátil para mitigação do esquecimento catastrófico em aprendizado contínuo.

Palavras-chave: Reconhecimento de expressões faciais, Redes neurais convolucionais, Esquecimento catastrófico, *Pseudo-Rehearsal*, Regularização

Abstract

Facial expression recognition is a key task in machine learning, enabling a wide range of applications. However, convolutional neural networks often suffer from catastrophic forgetting, which compromises the retention of previously learned knowledge. This thesis investigates catastrophic forgetting in convolutional neural networks applied to facial expression recognition in the context of continual learning. It proposes the Emotion-Centered Generative Replay (ECgr) method, based on pseudo-rehearsal with synthetic images generated by Wassertein Generative Adversarial Network with Gradient Penalty (WGAN-GP). Three quality assurance strategies are evaluated: a supervised approach, an unsupervised clustering-based technique, and a latent space optimization guided by previously validated samples. Results show that ECgr significantly mitigates forgetting compared to traditional fine-tuning, with additional improvements when combined with supervised filtering. The method approaches the performance of joint training, even without reusing original data. Experimental results support the hypotheses that ECgr enhances continual learning performance, that supervised filtering further improves outcomes, and that the method generalizes to other datasets such as Split-MNIST. In the domain-incremental Split-MNIST scenario, ECgr, despite not reusing real data, outperformed classical regularization-based techniques and approached the results of strategies that use real data from previous tasks or more complex generative models. These findings highlight the robustness and generalizability of the proposed approach, reinforcing its potential as an effective and versatile solution for mitigating catastrophic forgetting in continual learning.

Keywords: Facial expression recognition, Convolutional Neural Networks, Catastrophic forgetting, Pseudo-rehearsal, Regularization

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADA	*Adaptive Distillation of Adapters*
ADAM	*Adaptive Model*
AU	*Action Unit*
AFEW	*Acted Facial Expressions in the Wild*
BiC	*Bias Correction*
BWT	*Backward Transfer*
CAAE	*Conditional Adversarial Auto-Encoder*
CFDC	*Cascaded Feature Drift Compensation*
CIFAR	*Canadian Institute for Advanced Research Dataset*
CK+	*Extended Cohn-Kanade Dataset*
CL	*Continual Learning*
CLAW	*Continual Learning with Adaptive Weights*
Class-IL	*Class-Incremental Learning*
CLEAR	*Continual Learning with Experience And Replay*
CLIFER	*Continual Learning with Imagination for Facial Expression Recognition*
CLS	*Complementary Learning System*
CGLO	*Coeficient Guided Latent Optmization*
CPG	*Compacting, Picking, and Growing*
CSIM	*Cosine Similarity*
CURL	*Continual Unsupervised Representation Learning*
CVT	*Contrastive Vision Transformer*
CNN	*Convolutional Neural Network*
DANA	*Deep Attention Network Architecture*

DEN	*Dynamically Expandable Networks*
DER	*Dynamically Expandable Representation*
Domain-IL	*Domain-Incremental Learning*
NR-DFERNet	*Noise-robust Dynamic Facial Expression Recognition Network*
DyTox	*Transformers for Continual Learning with Dynamic Token Expansion*
EASE	*Expandable Subspace Ensemble*
ECgr	*Emotion-Centered Generative Replay*
EMRNet	*Enhanced Micro-expression Recognition Network with Attention and Distance Correlation*
EWC	*Elastic Weight Consolidation*
FACS	*Facial Action Coding System*
FER	*Facial Expression Recognition*
FER2013	*Facial Expression Recognition 2013 Dataset*
FID	*Fréchet Inception Distance*
FM	*Forgetting Measure*
FWT	*Forward Transfer*
GAN	*Generative Adversarial Network*
GCAB	*Gated Class-Attention Block*
GEM	*Gradient Episodic Memory*
GR	*Generative Replay*
GWR	*Growing When Required*
HCI	*Human-Computer Interaction*
HiDe-Prompt	*Hierarchical Decomposed Prompting*
ICL	*Intra-dataset Continual Learning*
IM	*Intransience Measure*
IMM	*Incremental Moment Matching*

IS	*Inception Score*
iCaRL	*Incremental Classifier and Representation Learning*
JAFFE	*Japanese Female Facial Expression Database*
L2P	*Learning to Prompt*
LGR	*Latent Generative Replay*
LSTM	*Long Short-Term Memory*
LwF	*Learning without Forgetting*
MAS	*Memory Aware Synapses*
MER	*Meta-Experience Replay*
MMI	*MMI Facial Expression Database*
MNIST	*Modified National Institute of Standards and Technology Dataset*
MoCL-P	*Module Composition and Pruning for Continual Learning*
MSE	*Mean Squared Error*
MS-SSIM	*Multi-Scale Structural Similarity*
MUG	*Interactive Multimodal Grounding on User Interfaces Dataset*
PCA	*Principal Component Analysis*
PECL	*Parameter-Efficient Continual Learning*
PSNR	*Peak Signal-to-Noise Ratio*
PSP	*Parameter Superposition*
QA	*Quality Assessment*
RAF-DB	*Real-world Affective Faces Database*
RW	*Riemannian Walk*
SI	*Synaptic Intelligence*
SimpleCIL	*Simple Class Incremental Learning*
SLCA	*Slow Learner with Classifier Alignment*
SSA	*Spectral and Spatial Attention*

SSIM	*Structural Similarity Index Measure*
Task-IL	*Task-Incremental Learning*
TAM-CL	*Task Attentive Multimodal Continual Learning*
TCN	*Temporal Convolutional Network*
TFEID	*Taiwanese Facial Expression Image Database*
VA	*Valence-Arousal*
VaL	*Vision-and-Language*
VCL	*Variational Continual Learning*
VGG-Face	*Visual Geometry Group Face Descriptor*
ViT	*Vision Transformer*
WGAN	*Wassertein Generative Adversarial Network*
WGAN-GP	*Wassertein Generative Adversarial Network with Gradient Penalty*

Sumário

1 Introdução

As emoções são essenciais na interação e compreensão humanas. Elas são utilizadas no dia a dia como uma parte fundamental da comunicação e podem se manifestar de inúmeras formas: tristeza, felicidade, raiva, desgosto, alegria, entre muitas outras. Algumas expressões podem ser denominadas microexpressões faciais, as quais representam sutis alterações. Humanos, em geral, são capazes de perceber as alterações nas emoções de uma pessoa pela sua expressão facial, e esta habilidade é o que nos permite ter uma comunicação mais efetiva.

As emoções via expressões faciais também podem ser utilizadas em outros contextos, além da comunicação. Por exemplo, um sistema eletrônico pode reconhecer automaticamente as expressões faciais de uma pessoa para fornecer recomendações baseadas na sua emoção. O reconhecimento de expressões faciais (FER, do inglês *Facial Expression Recognition*) está presente em diversas aplicações, como carros autônomos, sistemas de recomendação personalizados e interação humano-robô (LI; DENG, 2022).

Uma maneira de reconhecer estas emoções humanas é por meio do uso de *Convolutional Neural Networks* (CNNs). Essas redes neurais são essenciais no reconhecimento de padrões em imagens e são amplamente utilizadas em visão computacional em contextos como classificação de imagens, detecção de objetos, segmentação de imagens, reconhecimento facial, dentre outros. No contexto de reconhecimento de emoções, estas redes neurais podem ser treinadas em grandes bases de dados contendo variadas emoções e, assim, serem capazes de distinguir entre estas diferentes emoções humanas ao analisar uma imagem facial.

1.1 Definição do problema

Uma limitação das CNNs e de redes neurais em geral é a sua fragilidade em esquecer tarefas já aprendidas quando submetidas a novos conhecimentos. Este problema é chamado de *catastrophic forgetting* ou esquecimento catastrófico (MCCLOSKEY; COHEN, 1989; GOODFELLOW et al., 2013b). Outros termos similares também utilizados são decadência ou esquecimento de memória e perda de aprendizado. Ao serem submetidas a um treinamento contínuo de tarefas, as CNNs frequentemente apresentam dificuldades em reter o conhecimento aprendido de tarefas anteriores, uma vez que o conhecimento de novas tarefas substitui o que já foi aprendido. Este comportamento é desfavorável para essas redes em cenários dinâmicos ou de aprendizado contínuo, pois diminui a sua capacidade de reter informações importantes.

O fenômeno de esquecimento de memória surge do processo de otimização das CNNs, no qual os parâmetros da rede neural são ajustados para se adequar às novas tarefas, muitas vezes sobrepondo recursos importantes do modelo neural relacionados ao conhecimento de tarefas anteriores. Isso compromete a capacidade de retenção de conhecimento em redes neurais quando expostas a novos dados de forma contínua. Diversos trabalhos no campo de aprendizado de máquina e aprendizado contínuo já foram propostos para mitigar esse problema de perda de memória, com diversas inspirações provindas de estudos biológicos do cérebro e comportamento humano, especialmente no campo da neuropsicologia (PARISI et al., 2019; KHETARPAL et al., 2022; ALEIXO et al., 2024).

Embora as CNNs continuem sendo amplamente utilizadas com sucesso em tarefas de visão computacional, incluindo reconhecimento facial, novas arquiteturas têm surgido como alternativas promissoras. Desde 2017, com a proposta dos *transformers* pelos autores Vaswani et al. (2017), modelos baseados em atenção vêm demonstrando forte desempenho em diversas áreas, inclusive na mitigação do esquecimento catastrófico (ERMIS et al., 2022; WANG et al., 2022a; WANG et al., 2022b; COTOGNI et al., 2025). Os *transformers* se destacam por capturar dependências globais nos dados por meio de mecanismos de atenção, o que pode contribuir para uma melhor retenção de conhecimento em cenários contínuos. Alguns estudos (DOUILLARD et al., 2022; WANG et al., 2022a) indicam que essas arquiteturas podem superar as CNNs em alguns contextos específicos de aprendizado contínuo, oferecendo oportunidades interessantes para futuras investigações. Entretanto, o foco deste trabalho não reside na obtenção do maior desempenho absoluto em tarefas de reconhecimento de emoções, mas na condução de um protocolo experimental sistemático e controlado para avaliar estratégias de mitigação do esquecimento catastrófico em arquiteturas convolucionais. Essa escolha está diretamente alinhada aos objetivos e questões de pesquisa propostos, que investigam como métodos generativos, técnicas de filtragem de qualidade e ponderações de custo podem preservar o conhecimento prévio ao longo do aprendizado incremental, por meio de experimentos repetidos, comparáveis e realizados sob condições controladas. As CNNs foram, portanto, adotadas principalmente por sua maturidade e ampla utilização na literatura de reconhecimento de expressões faciais, o que facilita a comparação com trabalhos anteriores e permite isolar de forma mais controlada os efeitos das estratégias propostas para mitigação do esquecimento catastrófico. Dessa forma, o foco permanece na avaliação do método proposto, e não na exploração de novas arquiteturas de rede. Assim, embora os *transformers* representem uma direção promissora, sua exploração extrapola o escopo deste trabalho e é indicada como uma perspectiva para estudos futuros.

O reconhecimento de expressões faciais foi escolhido como tarefa de estudo neste trabalho por representar um cenário realista e desafiador para investigação do esquecimento catastrófico. Em FER, o desempenho dos modelos depende da sensibilidade a pequenas variações faciais e da generalização sobre identidades e contextos variados, tornando-o

particularmente vulnerável à degradação de desempenho quando exposto a novos dados. Além disso, bases de dados de FER são frequentemente heterogêneas em termos de iluminação, pose, intensidade emocional e distribuição de classes, o que dificulta o aprendizado contínuo sem acesso ao conjunto completo de dados anteriores. Essas características tornam o FER uma escolha apropriada para avaliar a robustez de métodos de mitigação de esquecimento, pois simulam de forma natural os desafios enfrentados em sistemas de aprendizado incremental no mundo real, como em aplicações afetivas, educacionais e de saúde mental.

Além disso, quando se trata de reconhecimento de expressões faciais em contextos de aprendizado contínuo, é possível deparar-se com o problema da indisponibilidade de dados para retreinamento em tarefas previamente aprendidas. Esta é uma questão fundamental em áreas como reconhecimento facial e reconhecimento de emoções. Em cenários práticos, a coleta contínua de dados nem sempre é viável, seja por restrições de privacidade, limitações de armazenamento ou pelo custo de processamento. Essa escassez de dados cria o desafio de manter o aprendizado adquirido sem acesso aos dados originais das tarefas anteriores, essencial para evitar o fenômeno conhecido como esquecimento catastrófico. Para mitigar esse problema, técnicas como *pseudo-rehearsal* ou *replay* generativo, que geram dados sintéticos representativos das tarefas anteriores, surgem como soluções promissoras, permitindo que o modelo "reviva" experiências anteriores sem acesso direto aos dados originais (LANGE et al., 2022).

Neste trabalho, assume-se explicitamente que os conjuntos de dados utilizados em tarefas anteriores de reconhecimento de expressões faciais não estão mais disponíveis nas etapas subsequentes de aprendizado. Essa suposição é fundamental para o delineamento do método proposto, uma vez que, em cenários de aprendizado contínuo, a indisponibilidade dos dados originais impede o uso de estratégias que exijam estes dados para o retreinamento conjunto. Essa hipótese reflete condições plausíveis em ambientes reais, nos quais a manutenção ou reutilização de dados faciais é frequentemente inviável. Questões de privacidade e consentimento, como as regulamentadas por legislações de proteção de dados, podem restringir o armazenamento prolongado de imagens contendo faces humanas (WANG et al., 2024c).

Além disso, limitações relacionadas a *copyright* e propriedade intelectual também contribuem para a dificuldade de reter ou redistribuir bases de dados biométricas completas após o treinamento inicial. Muitas bases públicas de reconhecimento facial e de emoções impõem restrições explícitas de uso apenas para pesquisa acadêmica não comercial. Em contextos corporativos ou clínicos, os dados podem ainda estar sujeitos a acordos de confidencialidade e cláusulas contratuais que proíbem sua transferência entre instituições e projetos. Todos estes fatores tem levado diversos repositórios a limitar o acesso ou mesmo a descontinuar a disponibilização de seus conjuntos de dados (WANG et al., 2024c).

Essas barreiras legais e éticas tornam o reaproveitamento direto de dados originais uma prática cada vez mais inviável em ambientes reais de aprendizado contínuo, especialmente em aplicações sensíveis como o reconhecimento de emoções. Assim, a suposição adotada neste trabalho, de que os conjuntos de dados utilizados em tarefas anteriores não estão mais disponíveis nas tarefas subsequentes, não apenas torna o cenário experimental mais desafiador, mas também mais representativo das restrições práticas enfrentadas em sistemas reais. Nessa perspectiva, a adoção de estratégias baseadas em *replay* generativo e reconstrução do conhecimento por meio de amostras sintéticas alinha-se à tendência contemporânea de buscar métodos de aprendizado compatíveis com princípios de privacidade, ética e também conformidade regulatória.

A fim de ilustrar o problema abordado neste trabalho, é necessário contextualizar quantitativamente o esquecimento catastrófico em cenários reais. Para isto, foram selecionadas quatro bases de dados que compartilham o mesmo conjunto de sete expressões faciais básicas (alegria, tristeza, medo, surpresa, raiva, desgosto e neutro), permitindo uma comparação direta entre as tarefas. Utilizando as bases de dados de expressões faciais MUG (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010), JAFFE (LYONS; KAMACHI; GYOBA, 1998), TFEID (CHEN et al., 2009) e CK+ (LUCEY et al., 2010), foi conduzido um experimento ilustrativo com o objetivo de demonstrar qualitativamente o fenômeno de esquecimento catastrófico em um cenário de aprendizado sequencial. A Figura 1 ilustra alguns exemplos de imagens destas bases de dados. Os detalhes metodológicos sobre o balanceamento destas bases e a compatibilização das classes são detalhados na Seção 5.1.

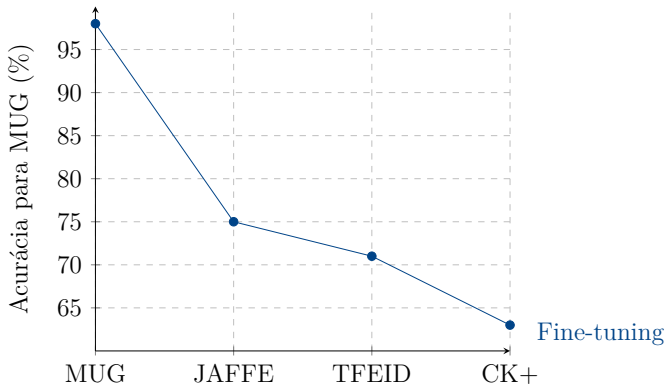
Figura 1 – Amostras das bases de dados MUG, JAFFE, TFEID e CK+.

Fonte: (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010; LYONS; KAMACHI; GYOBA, 1998; CHEN et al., 2009; LUCEY et al., 2010).

Assim, inicialmente, uma CNN foi treinada para a base de dados MUG. Em seguida,

de forma contínua, a CNN foi adaptada para as bases de dados JAFFE, TFEID e CK+, retreinando somente a última camada densamente conectada (do inglês *fully-connected*) – ou seja, utilizando o método de *fine-tuning*, em que as camadas convolucionais permanecem congeladas e os pesos da última camada *fully-connected* são retreinados para a nova tarefa. Ressalta-se que, embora as diferenças entre as bases, como iluminação, características étnicas ou condições de captura, introduzam desafios adicionais de *domain shift*, a ordem de apresentação foi mantida fixa neste experimento ilustrativo com o objetivo de evidenciar a degradação sequencial da memória do modelo. A Figura 2 ilustra os resultados obtidos. Para a base de dados MUG, a rede neural convolucional apresentou uma acurácia de 98%. Em seguida, ao ser adaptada para a base de dados JAFFE, a acurácia dos testes realizados com a base de dados MUG caiu para 75%. Para a última base de dados, CK+, a adaptação para um conjunto de dados diferente do inicial levou o modelo neural, inicialmente treinado na base de dados MUG, a atingir uma acurácia de 63%. É possível concluir que, ao expor uma CNN a diferentes bases de dados, de forma contínua, sem um método que busque mitigar o esquecimento catastrófico, a *performance* de tal rede neural nos aprendizados passados irá se deteriorar à medida que novas tarefas lhe sejam apresentadas. Cabe destacar que este experimento tem caráter apenas ilustrativo, com o objetivo de evidenciar o comportamento de degradação de desempenho quando o treinamento ocorre de forma sequencial. Aspectos metodológicos mais detalhados, como balanceamento entre bases, controle de variações visuais e análise da influência da ordem das tarefas, são tratados no Capítulo 5.

Figura 2 – Retreinamento para diferentes bases de dados de reconhecimento de expressões faciais, utilizando o método *fine-tuning*.



Fonte: autoria própria.

A partir da análise do resultado obtido neste experimento, é possível notar que o problema do esquecimento catastrófico em redes neurais, especialmente em tarefas de aprendizado incremental, torna-se ainda mais desafiador quando aplicado ao reconhecimento de emoções, devido à grande diversidade de expressões faciais e variações individuais entre sujeitos. Este trabalho propõe um método para mitigar esse problema por meio de

uma abordagem generativa combinada a estratégias de avaliação da qualidade das imagens sintéticas e a uma função de custo ponderada.

Diferentemente de abordagens tradicionais de *generative replay*, que geram dados sintéticos sem controle de qualidade ou ponderação durante o treinamento, o método proposto neste trabalho introduz três aspectos principais: (1) um algoritmo generativo chamado *Emotion-Centered Generative Replay* (ECgr), baseado em Generative Adversarial Networks (GANs) (GOODFELLOW et al., 2014), capaz de gerar imagens sintéticas representativas do conhecimento aprendido anteriormente; (2) três métodos independentes de avaliação da qualidade dessas imagens, com o objetivo de filtrar amostras que possam comprometer o desempenho do modelo ao serem reutilizadas no treinamento; e (3) uma função de custo que pondera a importância de cada imagem (real ou sintética) durante o aprendizado de novas tarefas.

Embora já existam diversas abordagens para lidar com o esquecimento catastrófico, muitas delas enfrentam limitações como a dependência de dados reais passados e a ausência de mecanismos que avaliem a confiabilidade de amostras sintéticas geradas. Apesar de existir na literatura algumas abordagens aplicadas ao reconhecimento de expressões faciais, ainda há uma lacuna quanto a estratégias que integrem esses mecanismos de controle em contextos realistas e com dados de alta variabilidade visual.

Até onde se tem conhecimento, esta é a primeira proposta a combinar geração sintética orientada por emoções, avaliação automática da qualidade das amostras e ponderação no processo de aprendizagem contínua em um único sistema para mitigação do esquecimento catastrófico em reconhecimento de expressões faciais. Os resultados experimentais obtidos, apresentados nos capítulos seguintes, indicam ganhos significativos de desempenho incremental, robustez e generalização.

1.2 Objetivos

Nesta seção apresentam-se os principais objetivos deste trabalho, expondo o objetivo principal e os detalhes específicos para atingir este objetivo.

1.2.1 Objetivo geral

Este trabalho tem como objetivo propor e validar um método generativo baseado em GANs, capaz de gerar imagens sintéticas que representem o conhecimento previamente adquirido por redes neurais convolucionais aplicadas ao reconhecimento de expressões faciais, acompanhado de três novas abordagens propostas de filtragem de qualidade das imagens geradas, assim como uma ponderação na função de custo, com o propósito de mitigar o esquecimento catastrófico.

1.2.2 Objetivos específicos

1. Investigar o desempenho de redes neurais convolucionais em tarefas de reconhecimento de expressões faciais sob o paradigma de aprendizado contínuo.

2. Projetar e implementar o método ECgr, fundamentado no uso de GANs específicas por classe para geração incremental de dados sintéticos representativos de emoções.

3. Desenvolver e integrar três estratégias distintas de garantia de qualidade (QA, do inglês *quality assessment*) das imagens geradas: (i) baseada em classificação supervisionada, (ii) baseada em agrupamento não supervisionado com validação interna, e (iii) baseada em otimização do espaço latente.

4. Propor uma função de custo ponderada para incorporar o nível de confiança na qualidade das imagens sintéticas durante o treinamento contínuo.

5. Combinar o método ECgr com os diferentes mecanismos de QA e a função de custo ponderada, avaliando o impacto individual e conjunto dessas estratégias no desempenho incremental.

6. Aplicar e analisar a generalização do método proposto em domínios distintos do reconhecimento de emoções faciais, como um estudo preliminar de portabilidade.

1.3 Questões de pesquisa

Nesta seção apresentam-se as principais questões de pesquisa relacionadas a este trabalho. Estas questões norteiam os experimentos conduzidos ao longo do trabalho, a fim de esclarecer todos os possíveis questionamentos em relação ao contexto em que o trabalho está introduzido e também em relação ao método proposto.

Q1 O método ECgr melhora a acurácia e retenção do conhecimento das CNNs no reconhecimento de expressões faciais em cenários de aprendizado contínuo?

Q2 A filtragem de imagens sintéticas com base nos métodos de QA propostos contribui para reduzir o *catastrophic forgetting* em tarefas sucessivas?

Q3 O método proposto melhora a acurácia e retenção do conhecimento das CNNs em cenários diferentes do de reconhecimento de expressões faciais?

1.4 Contribuições

Este trabalho apresenta quatro contribuições principais:

1. Um protocolo experimental utilizando conjuntos de dados em FER, com uma investigação do esquecimento catastrófico neste cenário em aspectos como a acurácia, retenção de conhecimento, etc., oferecendo uma discussão detalhada sobre os benefícios do método proposto em comparação com outras abordagens;

2. Proposição de um novo método de *pseudo-rehearsal* centrado em emoções para mitigar o esquecimento de memória em tarefas de reconhecimento de expressões faciais;

3. Três protocolos distintos para mitigar o esquecimento de memória a partir da filtragem de imagens sintéticas;

4. Proposta inédita da combinação de geração sintética com verificação automática de qualidade e ponderação de confiança, aplicadas conjuntamente em um cenário realista de aprendizado contínuo.

Outras contribuições deste trabalho incluem: a extensa revisão dos métodos de esquecimento catastrófico, sintetizados a partir das bases de dados mais utilizadas pelos estudos presentes na literatura especializada e a publicação de um artigo com o método proposto neste trabalho.

Além das contribuições supracitadas, este trabalho possui o seguinte trabalho aceito no 27° International Conference on Pattern Recognition (ICPR):

- LAURENSI, Israel A.; JR., Alceu de Souza Britto; BARDDAL, Jean Paul; KOERICH, Alessandro Lameiras. Alleviating Catastrophic Forgetting in Facial Expression Recognition with Emotion-Centered Models. 2024.

1.5 Organização do documento

Nesta seção apresenta-se a organização deste documento.

O Capítulo 2 apresenta a teoria e fundamentação dos tópicos mais relevantes com o método proposto neste trabalho.

O Capítulo 3 traz os principais trabalhos encontrados na literatura à luz dos temas abordados durante este trabalho, apresentando especificidades técnicas, resultados e implicações.

O Capítulo 4 expõe o método proposto, com uma visão geral e técnica do problema e dos recursos utilizados para compor o todo deste trabalho.

O Capítulo 5 apresenta os resultados obtidos com a aplicação do método proposto, suas implicações e discussões dos achados durante a condução de todos os testes realizados.

O Capítulo 6 traz as conclusões do trabalho desenvolvido, com uma análise da importância do projeto e do contexto a que se refere, bem como uma discussão sobre possíveis trabalhos futuros.

2 Fundamentação teórica

Neste capítulo, apresentam-se os principais fundamentos teóricos relacionados ao trabalho proposto, a saber: reconhecimento de emoções por meio de recursos computacionais, fundamentações biológicas a respeito do funcionamento da memória e teorias da neuropsicologia, aprendizado contínuo e o problema de esquecimento catastrófico. Optou-se por não detalhar novamente fundamentos amplamente consolidados como redes neurais convolucionais e aprendizado supervisionado, cujos princípios estão bem descritos nas obras clássicas dos autores Mitchell (1997), Bishop (2007), Alpaydin (2010), Krizhevsky, Sutskever e Hinton (2012) e Goodfellow, Bengio e Courville (2016). Considera-se que tais tópicos já estão suficientemente solidificados na literatura, de modo que sua não inclusão detalhada neste documento não compromete o pleno entendimento do trabalho. Ainda assim, reconhece-se que as CNNs continuam evoluindo, tanto em desempenho quanto em eficiência, com propostas modernas amplamente adotadas como as ResNet (HE et al., 2016), DenseNet (HUANG et al., 2017), EfficientNet (TAN; LE, 2019), RegNet (RADOSAVOVIC et al., 2020) e ConvNeXt (LIU et al., 2022), que representam marcos contemporâneos na arquitetura de redes convolucionais. Por outro lado, aplicações específicas, arquiteturas modernas e desafios atuais, especialmente no contexto do aprendizado contínuo, são discutidos ao longo do capítulo com base em literatura recente, de modo a refletir o estado da arte e os avanços mais relevantes para este estudo.

2.1 Reconhecimento de emoções

Nesta seção são apresentados os principais conceitos e fundamentos teóricos relacionados ao reconhecimento de emoções faciais, incluindo aspectos psicológicos, neurobiológicos e computacionais. O objetivo é fornecer uma base sólida para compreender como as emoções são reconhecidas e processadas por sistemas computacionais, além de discutir os desafios e avanços recentes na área.

2.1.1 Aspectos fundamentais

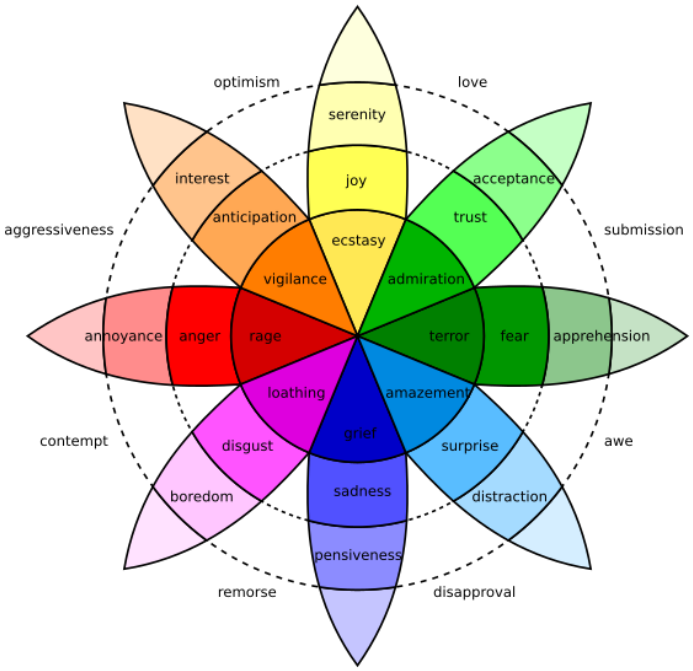
O reconhecimento de emoções faciais é um campo interdisciplinar que envolve aspectos da psicologia, neurociência, ciência da computação e inteligência artificial. A interpretação automática de emoções humanas a partir de sinais visuais exige uma compreensão básica do que são emoções, de como elas se manifestam fisicamente e de como podem ser operacionalizadas computacionalmente.

O conceito de emoção tem sido estudado por diversas disciplinas, cada uma ofere-

cendo definições distintas. De modo geral, emoções podem ser entendidas como respostas afetivas complexas que envolvem componentes fisiológicos, expressivos e subjetivos diante de eventos internos ou externos (EKMAN, 1992; IZARD, 1977). O autor Ekman (1992) propôs que certas emoções são biologicamente programadas e apresentam expressões faciais universais, isto é, independentemente do contexto cultural elas se manifestam de forma reconhecível por observadores humanos. Com base em experimentos interculturais, Ekman (1992) identificou seis emoções básicas: felicidade, tristeza, raiva, medo, surpresa e nojo. Cada uma dessas emoções possui um conjunto característico de contrações musculares faciais, formalizado por meio do sistema FACS (*Facial Action Coding System*).

O autor Izard (1977), por sua vez, também defendeu a existência de emoções discretas inatas, mas propôs um conjunto ligeiramente distinto, com 10 emoções básicas: interesse, alegria, surpresa, tristeza, raiva, nojo, desprezo, vergonha, culpa e medo (IZARD, 1977). O autor Plutchik (1982) propôs uma abordagem dimensional, representando as emoções em um modelo em forma de roda (a "roda das emoções"), onde emoções básicas se combinam em pares para gerar emoções compostas (PLUTCHIK, 1982). A Figura 3 ilustra essa proposta.

Figura 3 – Roda das Emoções proposta por Plutchik (1982), representando emoções básicas e suas combinações em pares.



Fonte: Adaptado da teoria de Plutchik (1982). Ilustração original por Machine Elf 1735 (2017) (domínio público).

No contexto computacional, essas emoções básicas são geralmente utilizadas como classes-alvo em tarefas de classificação de imagens faciais. No entanto, a própria definição

de emoção, sua quantificação e rotulação nos dados permanecem temas abertos e controversos, principalmente quando se considera a subjetividade da expressão emocional em contextos naturais. A distinção entre emoções básicas e compostas é muito importante para a modelagem de FER. Enquanto as emoções básicas possuem manifestações faciais mais padronizadas, consistentes e universais, as emoções compostas surgem da combinação de múltiplas emoções básicas e podem apresentar variações culturais e individuais mais acentuadas (DU; TAO; MARTINEZ, 2014). Por exemplo, a emoção de "culpa" pode ser considerada uma composição de tristeza e medo, enquanto "excitação" pode combinar alegria e surpresa. Do ponto de vista computacional, o reconhecimento de emoções compostas representa um desafio considerável, pois envolve maior ambiguidade expressiva e menor consistência nos dados de treinamento. Algumas bases de dados recentes, como o RAF-DB (LI; DENG; DU, 2017), incluem anotações de emoções compostas, ampliando o escopo do problema e exigindo classificadores mais robustos.

Além disso, há abordagens que adotam modelos dimensionais, como o modelo de Valência-Arousal (VA), que representa emoções em um espaço contínuo de duas dimensões: valência (agradável e desagradável) e excitação (alta e baixa energia) (RUSSELL, 1980). Isso permite representar nuances emocionais mais finas, sendo particularmente relevante para aplicações que exigem sensibilidade contextual ou resposta adaptativa.

O reconhecimento automático de emoções a partir de imagens faciais tem ampla aplicação em diversos domínios. Na interação humano-computador (HCI, do inglês *Human-Computer Interaction*), a detecção de emoções permite o desenvolvimento de interfaces mais empáticas e adaptativas, capazes de ajustar seu comportamento com base no estado emocional do indivíduo (PANTIC; ROTHKRANTZ, 2000). Exemplos incluem assistentes virtuais, robótica social e sistemas de ensino inteligente. Na área da saúde, o reconhecimento de expressões faciais pode ser utilizado para apoio diagnóstico e monitoramento de transtornos psicológicos, como depressão, nos quais a expressividade facial pode ser alterada. O uso de FER nesses contextos pode ajudar na avaliação objetiva de pacientes e no acompanhamento de terapias (DIBEKLIOGLU et al., 2015).

Apesar de seu potencial, o uso de FER em ambientes reais enfrenta limitações técnicas e éticas, incluindo variabilidade de iluminação, oclusões faciais, expressões sutis ou mascaradas, bem como preocupações com privacidade, viés e consentimento do uso dos dados pessoais. Tais aspectos tornam ainda mais relevante a pesquisa de abordagens robustas e generalizáveis para FER em contextos não controlados.

2.1.2 Bases de dados para FER

O desempenho de algoritmos para reconhecimento de emoções faciais está fortemente condicionado à qualidade, diversidade e natureza das bases de dados utilizadas durante o treinamento e avaliação dos modelos. As bases de dados disponíveis variam

amplamente quanto à quantidade de imagens, condições de captura, granularidade dos rótulos emocionais e se representam expressões artificiais ou espontâneas, estáticas ou dinâmicas.

Bases com expressões posadas (ou atuadas) são geralmente coletadas em ambientes controlados, com sujeitos instruídos a simular emoções específicas. Embora favoreçam a consistência e a rotulagem confiável, esse tipo de dado tende a ser artificial e pode não representar a variação emocional observada em ambientes naturais. Já as expressões espontâneas emergem em contextos mais naturais, frequentemente capturadas a partir de vídeos, redes sociais ou interações reais, e refletem nuances mais sutis e realistas, mas apresentam desafios adicionais de rotulagem e variabilidade interindividual. Essas bases de dados frequentemente são nomeadas com o termo *in-the-wild*, indicando que foram coletadas em condições não controladas, refletindo a complexidade do mundo real.

Outro aspecto relevante é a distinção entre bases estáticas, compostas por imagens individuais rotuladas, e bases dinâmicas, que contêm sequências temporais (vídeos ou quadros sucessivos), possibilitando a análise da evolução temporal das expressões, que é considerado um fator importante para modelos que exploram informações temporais, como redes recorrentes ou *transformers* com atenção temporal.

A Tabela 1 apresenta uma comparação entre as principais bases de dados utilizadas em pesquisas recentes de FER, destacando características como tamanho, tipo de expressão, natureza (estática/dinâmica), ambiente de captura e diversidade emocional. A FER2013 (GOODFELLOW et al., 2013a) é amplamente utilizada por seu tamanho e diversidade, embora contenha imagens de baixa qualidade e rótulos ruidosos. A RAF-DB (LI; DENG; DU, 2017) e a AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2019) contêm expressões espontâneas e foram construídas com anotações mais robustas, sendo adequadas para contextos *in-the-wild*. A AffectNet também inclui dimensões contínuas de emoção, como valência e *arousal*.

Tabela 1 – Principais bases de dados utilizadas para reconhecimento de emoções faciais.

Base de Dados	**Ano**	**Quantidade**	**Expressões**	**Tipo**	**Ambiente**	**Emoções**
FER2013	2013	35887 imagens	Posadas	Estática	*In-the-wild*	7 básicas
RAF-DB	2017	30000 imagens	Espontâneas	Estática	*In-the-wild*	7 compostas
AffectNet	2019	456349 imagens	Espontâneas	Estática	*In-the-wild*	8 compostas
CK+	2010	593 imagens	Posadas	Dinâmica	Controlado	7 básicas
JAFFE	1998	213 imagens	Posadas	Estática	Controlado	7 básicas
MMI	2005	1280 vídeos	Posadas	Dinâmica	Controlado	6 básicas
EmotioNet	2016	950000 imagens	Espontâneas	Estática	*In-the-wild*	16 AUs
AFEW	2012	54 vídeos	Espontâneas	Dinâmica	*In-the-wild*	7 básicas
TFEID	2009	268 imagens	Posadas	Estática	Controlado	7 básicas
MUG	2010	1462 vídeos	Posadas	Dinâmica	Controlado	6 básicas

Fonte: autoria própria.

Bases como *Extended Cohn-Kanade* (CK+) (LUCEY et al., 2010), *Japanese Female*

Facial Expression Database (JAFFE) (LYONS; KAMACHI; GYOBA, 1998), MMI (PANTIC et al., 2005), *Taiwanese Facial Expression Image Database* (TFEID) (CHEN et al., 2009) e MUG (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010) foram coletadas em ambientes controlados, com expressões posadas e, em alguns casos, sequências dinâmicas. São úteis para estudos mais estruturados e análises temporais. Por fim, a EmotioNet (BENITEZ-QUIROZ; SRINIVASAN; MARTINEZ, 2016) e a *Acted Facial Expressions in the Wild* (AFEW) (DHALL et al., 2012) oferecem grande variabilidade de expressões espontâneas, com foco em unidades de ação (AUs, do inglês *Action Units*) ou cenas de vídeos do cotidiano, sendo importantes para aplicações realistas.

É importante ressaltar que a escolha da base de dados afeta não apenas o desempenho dos modelos, mas também sua capacidade de generalização para contextos do mundo real. Com isso, métodos modernos têm buscado integrar aprendizado por transferência e aprendizado contínuo para mitigar limitações impostas por bases desbalanceadas ou restritas.

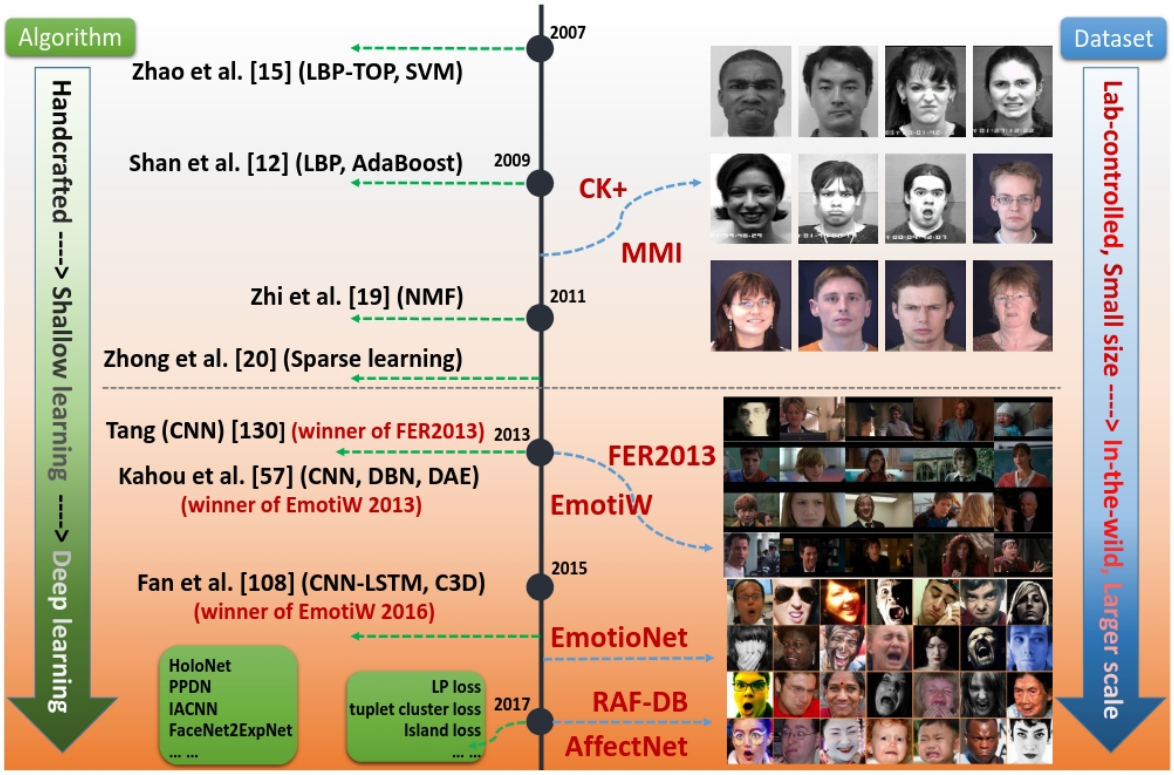
2.1.3 Reconhecimento de emoções faciais com *deep learning*

O avanço das técnicas de *deep learning* tem possibilitado o desenvolvimento de modelos cada vez mais robustos e precisos para identificar emoções humanas a partir de imagens faciais. Os autores Li e Deng (2022) propuseram uma revisão abrangente sobre o reconhecimento de emoções em *deep learning* e destacam que redes neurais profundas têm sido amplamente utilizadas para FER devido à sua capacidade de capturar características visuais complexas em imagens faciais e discriminar entre emoções com alta acurácia. CNNs, por exemplo, podem extrair padrões detalhados e sutis das expressões faciais, que são essenciais para distinguir emoções mesmo quando as diferenças são mínimas (LI; DENG, 2022).

Na Figura 4 é possível observar o levantamento feito pelos autores Li e Deng (2022) em relação aos avanços na área de reconhecimento de emoções, destacando dois grupos: algoritmos e bases de dados. Segundo os autores Li e Deng (2022), as abordagens tradicionais costumavam utilizar *handcrafted features*, ou seja, faziam o uso de algoritmos próprios para extração de características das imagens. O principal motivo disto é que não haviam muitas bases de dados disponíveis para treinar modelos profundos. No entanto, a partir de 2013, surgiram diversas competições no contexto de reconhecimento de emoções, como por exemplo FER2013 (GOODFELLOW et al., 2013a) e EmotiW (DHALL et al., 2015), que passaram a coletar e divulgar bases de dados suficientemente grandes para justificar o treinamento de redes neurais profundas e redes neurais convolucionais (LI; DENG, 2022).

Para auxiliar no aprendizado de características faciais, alguns autores, como os trabalhos de Liu et al. (2014), Zhang et al. (2016), Devries, Biswaranjan e Taylor (2014),

Figura 4 – Linha do tempo, proposta por Li e Deng (2022), sobre a evolução de algoritmos e bases de dados na área de reconhecimento de emoções.



Fonte: (LI; DENG, 2022).

utilizaram redes neurais pré-treinadas em outros contextos, a fim de adaptar essas redes para o reconhecimento de expressões faciais em bases de dados com um número reduzido de imagens. Diversas redes neurais, pré-treinadas em cenários diferentes do de FER, têm sido amplamente utilizadas para a transferência de aprendizado, como AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), VGG (SIMONYAN; ZISSERMAN, 2015), GoogleNet (SZEGEDY et al., 2015) e InceptionV3 (SZEGEDY et al., 2016). Outras redes foram criadas e adaptadas para o cenário de FER, como Facenet (SCHROFF; KALENICHENKO; PHILBIN, 2015) e VGG-Face (PARKHI; VEDALDI; ZISSERMAN, 2015). Também, outras redes como DANA (*Deep Attention Network Architecture*) (SADAK; KHALAF; SALAMA, 2024), *Enhanced Micro-expression Recognition Network with Attention and Distance Correlation* (EMRNet) (LIU et al., 2025) e *Noise-robust Dynamic Facial Expression Recognition Network* (NR-DFERNet) (LI et al., 2022) buscam aprimorar a discriminação entre expressões faciais por meio de mecanismos de atenção, extração de micro-expressões faciais e redução da interferência causada por imagens ruidosas.

No contexto de vídeos, redes sequenciais como *Long Short-Term Memory* (LSTMs) têm sido combinadas com CNNs para capturar a evolução temporal das expressões (MENG et al., 2019), enquanto abordagens com 3D-CNNs exploram convoluções espaço-temporais para modelar diretamente os conteúdos de vídeo (LI; DENG, 2022). Mais recentemente,

redes como *Temporal Convolutional Networks* (TCNs) também têm demonstrado bons resultados em tarefas temporais, oferecendo uma alternativa baseada em convoluções causais com memória de longo prazo (LEA et al., 2017).

Nos últimos anos, o reconhecimento de emoções faciais tem evoluído para arquiteturas híbridas e baseadas em *transformers*, superando as limitações das CNNs convencionais. Trabalhos recentes empregam mecanismos de autoatenção para capturar relações de dependência espacial e temporal mais amplas entre regiões faciais, como no *Vision Transformer* (ViT) e em variantes específicas para FER, por exemplo ViTFER (CHAUDHARI et al., 2022) e VTFF (MA; SUN; LI, 2023). Abordagens multimodais também têm recebido destaque, integrando informações de áudio e imagem (MENON et al., 2024), (ZHANG et al., 2024a). Trabalhos recentes exploram aprendizado auto-supervisionado para extração de representações de movimento facial para lidar com variações naturais de expressões (SUN et al., 2023). Em paralelo, diversos trabalhos lidam com discrepâncias entre domínios (por exemplo, entre bases de dados distintas ou ambientes de captura distintos), evidenciando que o reconhecimento de emoções precisa superar as diferenças de coleta para ser robusto (BIE et al., 2023), (GAO et al., 2024). Esses avanços indicam uma tendência de convergência entre aprendizado profundo, atenção, combinações multimodais e modelagem generativa, caracterizando uma nova fase na pesquisa no reconhecimento de emoções.

2.2 Neuropsicologia

O trabalho de Hebb (1949) marca um ponto crucial na psicologia e neurociência ao propor uma visão inovadora sobre o funcionamento do cérebro como uma máquina de aprendizado. Hebb (1949) propõem que o aprendizado e o comportamento são frutos não apenas de estímulos e respostas diretas, mas de redes complexas de neurônios interconectados. A ideia de que "neurônios que disparam juntos, conectam-se juntos" abre novos caminhos para entender como memórias e associações são criadas e solidificadas.

> When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased (HEBB, 1949).

O autor sugere que a estrutura do cérebro muda com cada nova experiência, com os neurônios formando "assembleias celulares" que armazenam fragmentos de experiências anteriores. Esse conceito é, de certa forma, uma metáfora para a resiliência do comportamento humano: assim como caminhos mais trilhados em uma floresta se tornam mais claros, nossas reações e pensamentos se reforçam quanto mais repetidos, moldando não

apenas nossa memória, mas também nosso comportamento e personalidade ao longo do tempo.

Ao expandir suas ideias, Hebb (1949) antecipa o impacto do ambiente e da experiência no desenvolvimento cerebral, o que se alinha com a noção de plasticidade neuronal. O autor sugere que o aprendizado é um processo dinâmico, em constante reconstrução, onde a memória e o comportamento não são rígidos, mas maleáveis. Dessa forma, ele lança as bases para uma compreensão mais rica de como o cérebro organiza o comportamento, tecendo uma rede intrincada de lembranças e habilidades que são tanto produto de nossa biologia quanto de nossas vivências. A Regra de Hebb estabelece que a estimulação repetida e persistente da célula pós-sináptica pela célula pré-sináptica leva a um aumento da eficácia sináptica e, com isso, segundo Parisi et al. (2019), os sistemas neurais se estabilizam para formar padrões funcionais de conectividade neural. Dessa forma, é possível representar a força dessa conexão sináptica por meio de:

$$\Delta w = x \cdot y \cdot n \tag{2.1}$$

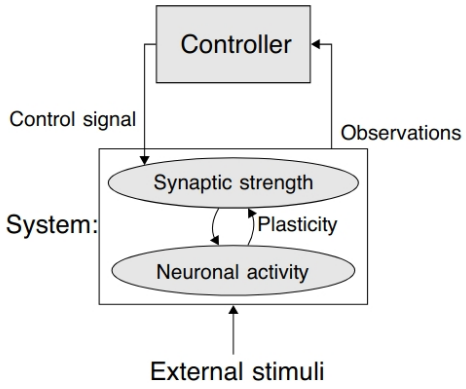
dado que x é a força pré-sináptica, y é a força pós-sináptica e n é a taxa de aprendizagem. Esta definição simplificada de como compreender a força sináptica entre neurônios pode impulsionar os estudos em aprendizado de máquina.

Um ponto importante ao trabalhar com redes neurais (biológicas ou artificiais) é o dilema da plasticidade-estabilidade. Esse dilema representa um dos desafios centrais ao compreendermos como o cérebro, e sistemas artificiais, conseguem aprender continuamente sem comprometer o conhecimento pré-existente. A plasticidade, que é a capacidade de adaptação e aprendizado a partir de novas experiências, é vital para a sobrevivência e o desenvolvimento de habilidades em um ambiente de mudança constante. Porém, essa maleabilidade do sistema neural precisa ser balanceada pela estabilidade, que é a habilidade de preservar o conhecimento já adquirido. Assim, o cérebro implementa mecanismos neurofisiológicos, mantendo um certo grau de plasticidade para a adaptação e, ao mesmo tempo, protegendo contra o esquecimento catastrófico, onde novas informações poderiam sobrepor-se a aprendizados anteriores (ZENKE; GERSTNER; GANGULI, 2017).

Como pode ser observado na Figura 5, Zenke, Gerstner e Ganguli (2017) sugerem a presença de um controlador externo para monitorar e providenciar *feedback* para as atividades sinápticas que ocorrem nas conexões entre os neurônios. O motivo disto é devido ao fato de que a plasticidade Hebbiana e atividade neural são sistemas dinâmicos instáveis e, por isso, precisam de um mecanismo compensatório que observe as conexões sinápticas e a atividade neural e utilizem essas observações para providenciar um sistema de controle que possa estabilizar a dinâmica sináptica (ZENKE; GERSTNER; GANGULI, 2017).

Durante os primeiros anos de vida, no desenvolvimento inicial dos neurônios e conexões, a plasticidade assume um papel predominante, permitindo que redes neurais

Figura 5 – Aprendizado Hebbiano com um controlador externo.



Fonte: (ZENKE; GERSTNER; GANGULI, 2017).

cerebrais se moldam conforme recebem estímulos sensoriais externos. Esse período crítico proporciona uma janela na qual a estrutura neural se organiza de maneira profunda, estabelecendo bases sólidas para o desenvolvimento humano. Posteriormente, à medida que o organismo atinge uma estabilidade biológica, a plasticidade reduz-se gradualmente, adaptando-se a níveis menores e mais localizados. Essa transição é essencial para que o sistema conserve a estrutura funcional consolidada ao longo da vida, enquanto ainda permite ajustes pontuais de acordo com novas demandas ambientais (HENSCH et al., 1998).

Assim, o dilema da estabilidade-plasticidade não se resume a uma simples escolha entre manter ou esquecer, mas sim a um sofisticado jogo de forças que se ajusta dinamicamente. Esse equilíbrio é mantido por uma série de mecanismos internos que garantem que o aprendizado seja acumulativo e, ao mesmo tempo, adaptável. Em sistemas artificiais, a replicação desse processo representa um desafio, uma vez que a eficiência do aprendizado contínuo depende dessa harmonia entre absorver o novo e preservar o antigo.

Outro ponto crucial a ser analisado ao tratar de sistemas neurais é o *Complementary Learning System* (CLS). O conceito deste sistema oferece uma visão de como o cérebro processa e armazena informações de curto e longo prazo. Segundo a teoria, proposta por McClelland, McNaughton e O'Reilly (1995), o sistema do hipocampo desempenha um papel importante na aprendizagem de novas informações, permitindo que novas experiências sejam assimiladas com velocidade e precisão. No entanto, esse armazenamento inicial é apenas temporário, funcionando como uma memória de curto prazo que, posteriormente, será transferida e consolidada em sistemas neocorticais, garantindo assim a retenção duradoura das informações (MCCLELLAND; MCNAUGHTON; O'REILLY, 1995).

A distinção entre hipocampo e neocórtex no contexto do CLS é tanto funcional quanto estrutural. Segundo McClelland, McNaughton e O'Reilly (1995), enquanto o

hipocampo possui uma taxa de aprendizado rápida e utiliza representações esparsas para reduzir interferências, o neocórtex possui um ritmo de aprendizado mais lento e sobrepõe representações, construindo um arcabouço de conhecimento consolidado. Esse processo de transferência gradual é fundamental para o aprendizado contínuo e o desenvolvimento de representações estáveis, que permitem ao sistema neural lidar com a complexidade e as variações do ambiente.

O CLS também ajuda a explicar como ocorre, ao longo do tempo, a integração de novas informações sem que isso destrua as memórias antigas – como, por exemplo, o esquecimento catastrófico em redes neurais. A interação entre hipocampo e neocórtex cria um sistema adaptável e resiliente, no qual o aprendizado rápido não compromete a estabilidade do conhecimento previamente adquirido. Incorporar esse modelo em aprendizado de máquina e nas arquiteturas de redes neurais gera modelos que habilmente possam equilibrar plasticidade e estabilidade.

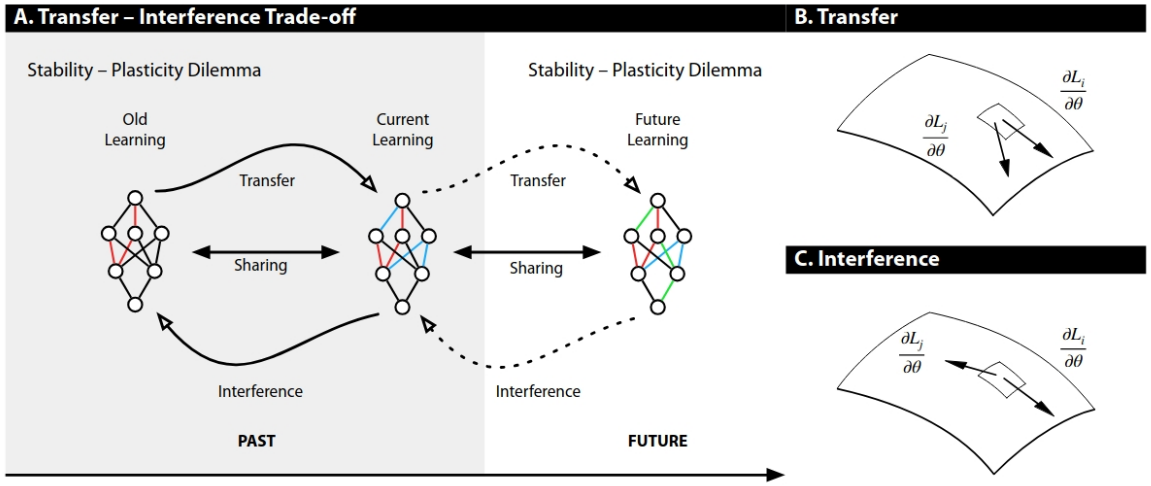
2.3 Aprendizado contínuo

Diversos estudos foram realizados no âmbito de *continual learning* (CL), como os trabalhos de Xu e Zhu (2018), Parisi et al. (2019), Lee e Lee (2020), Ven, Soures e Kudithipudi (2024), Wang et al. (2024a). Estes trabalhos normalmente focam nos problemas relacionados aos sistemas de aprendizado contínuo, por exemplo, o esquecimento catastrófico. Os autores Riemer et al. (2019) propuseram estender essa visão e analisar o contexto de aprendizado contínuo como um todo e, para isso, é preciso olhar para o problema de diferentes formas, considerando como as redes neurais se comportam com os pesos compartilhados ao longo do tempo e como elas se comportam diante do problema de estabilidade-plasticidade (RIEMER et al., 2019). A Figura 6 ilustra esse cenário.

Na Figura 7, é possível observar uma classificação proposta pelos autores Khetarpal et al. (2022) para o contexto geral de aprendizado contínuo. É importante notar que os autores evidenciam o *continual lifelong learning* como um cenário mais amplo e que requer diversos recursos e contextos interligados para que exista tal agente contínuo. De acordo com Khetarpal et al. (2022), cada categoria pode ser definida como:

- Adaptação de domínio: Adaptação de domínio é o processo de ajustar uma habilidade específica para um novo domínio, como transferir o aprendizado de um agente treinado em simulação para o mundo real;
- Transferência de aprendizado: Aprender cada tarefa do zero pode demandar muitos dados e é computacionalmente caro. A transferência de aprendizado permite que um agente use o conhecimento de tarefas anteriores para melhorar seu desempenho em novas tarefas. Isso normalmente envolve o treinamento em tarefas fonte para

Figura 6 – O compromisso entre transferência e interferência leva em conta o dilema estabilidade-plasticidade e sua dependência do compartilhamento de pesos.



Fonte: (RIEMER et al., 2019).

adaptar a uma tarefa alvo relacionada. O processo de aprendizado é dividido em pré-treinamento e *fine-tuning*;

- *Meta-training* e *meta-testing*: Esse protocolo, comum em meta-aprendizado, envolve um agente treinando em uma distribuição de tarefas para aprender a generalizar e depois testando em uma nova distribuição de tarefas. Esse processo facilita a otimização, pois o aprendizado ocorre em uma distribuição estacionária, ao contrário do aprendizado contínuo;

- Aprendizado multi-tarefa: Nesse cenário, o agente aprende uma série de tarefas sequenciais, adquirindo conhecimento em uma antes de passar para outra. É comum que as tarefas sejam intercaladas, mas o agente não controla a ordem. O objetivo é maximizar o desempenho em todas as tarefas, podendo incluir tanto tratativas universais quanto tratativas específicas para cada tarefa;

- Aprendizado contínuo (*lifelong*): Diferente de abordagens como *transfer learning* ou *multi-task learning*, o aprendizado contínuo exige que o agente opere em um cenário de evolução não estacionária de tarefas. Isso implica que o fluxo de dados e os objetivos mudam ao longo do tempo, exigindo aprendizado *online* e alta eficiência de recursos. Para ser bem-sucedido nesse contexto, o agente deve ser capaz de extrair e armazenar informações relevantes de um fluxo massivo de dados, atribuindo crédito a eventos cruciais enquanto mitiga o esquecimento catastrófico para reaproveitar conhecimentos prévios em novos domínios.

Estas definições são importantes para o entendimento do contexto como um todo ao tratar de aprendizado contínuo. Além destas definições, outros autores se propuseram

Figura 7 – Diferentes contextos de aprendizado contínuo.

Setting	Multiple Deployment Domains	Multiple Required Skills	Requires Online Learning	Requires Resource Efficiency & Sustainability	Non-stationary Task Evolution
Domain Adaptation	✓	X	X	X	X
Transfer Learning	✓	✓	X	X	✓
Meta-Training and Meta-Testing	✓	✓	X	X	X
Multi-task Learning	✓	✓	X	X	X
Continual (Lifelong) Learning	✓	✓	✓	✓	✓

Fonte: (KHETARPAL et al., 2022).

a categorizar os diferentes tipos de retreinamentos possíveis em cenários de aprendizado contínuo. Os autores Ven, Tuytelaars e Tolias (2022) propuseram uma categorização dividida em três grupos:

- *Task-incremental learning (Task-IL)*: O agente aprende múltiplas tarefas sequenciais e recebe o identificador da tarefa atual. Cada tarefa pode usar um classificador próprio, e não é necessário distinguir entre classes de tarefas diferentes durante a inferência;
- *Domain-incremental learning (Domain-IL)*: O agente aprende o mesmo conjunto de classes em domínios diferentes, sem receber o identificador da tarefa. O objetivo é generalizar para novos domínios sem esquecer os anteriores;
- *Class-incremental learning (Class-IL)*: Cada tarefa introduz novas classes e o agente deve distingui-las de todas as classes aprendidas previamente, sem acesso ao identificador da tarefa. É o cenário mais desafiador por exigir uma classificação global ao longo do tempo.

A fim de exemplificar a categorização apresentada, os autores Khetarpal et al. (2022) propuseram uma divisão da base de dados MNIST (DENG, 2012). Nesta divisão, a definição do problema para a categoria *task-incremental* torna-se a escolha entre dois dígitos do mesmo contexto, por exemplo, 0 ou 1; para a categoria *domain-incremental*, o problema torna-se definir se o dígito é par ou ímpar; e para o *class-incremental*, o problema torna-se a escolha entre todas as classes disponíveis (KHETARPAL et al., 2022).

A formalização desta categorização, segundo os autores Khetarpal et al. (2022), pode ser definida como: uma entrada $x \in \mathcal{X}$, um rótulo dentro do contexto $y \in \mathcal{Y}$ e um rótulo de contexto $c \in \mathcal{C}$. Os três cenários podem então ser definidos com base em como a função ou mapeamento que deve ser aprendido se relaciona com o espaço de contexto $\mathcal{C}$. Dessa forma, na categoria *task-incremental*, espera-se que um algoritmo aprenda um

mapeamento da forma $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathcal{Y}$; no *domain-incremental*, o mapeamento a ser aprendido é da forma $f : \mathcal{X} \rightarrow \mathcal{Y}$, enquanto no *class-incremental* a forma do mapeamento a ser aprendido é $f : \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{Y}$ (KHETARPAL et al., 2022).

2.4 Integridade de memória e *catastrophic forgetting*

O esquecimento catastrófico ou *catastrophic forgetting* é um problema em *machine learning* em que as redes neurais perdem conhecimentos previamente adquiridos ao serem submetidas a novas informações. Robins (1995) sugere a seguinte definição:

> If after its original training is finished a network is exposed to the learning of new information, then the originally learned information will typically be greatly disrupted or lost (ROBINS, 1995).

Carpenter e Grossberg (1988) definem o esquecimento em redes neurais como "*new learning washes away memories of prior learning if too many inputs perturb the system*" (CARPENTER; GROSSBERG, 1988). Os trabalhos de Carpenter e Grossberg (1988), Ratcliff (1990) e Robins (1995) foram cruciais para o entendimento de que as redes neurais, mesmo na sua forma mais simples, com poucas conexões e neurônios, já estavam suscetíveis ao problema da sobreposição de conhecimento. No trabalho de Ratcliff (1990), o autor apresenta uma análise aprofundada das limitações dos modelos de redes neurais utilizados na época, com foco no problema do esquecimento catastrófico. O autor demonstra que a capacidade de redes neurais de aprender de forma incremental e generalizar para novas situações é fundamental para a construção de sistemas de inteligência artificial mais autônomos e adaptáveis (RATCLIFF, 1990). No entanto, o estudo revela que os algoritmos de aprendizado tradicionais, como o *back-propagation*, tendem a sobrepor novas informações sobre as antigas, levando à perda gradual de conhecimento.

Através de análises teóricas e experimentais, Ratcliff (1990) identifica os mecanismos subjacentes ao esquecimento catastrófico e propõem diversas técnicas para mitigá-lo, como o uso de redes neurais recorrentes, o aprendizado por reforço e a regularização. Os resultados obtidos no trabalho de Ratcliff (1990) pavimentaram um caminho crucial para o desenvolvimento de sistemas de aprendizado de máquina capazes de lidar com a complexidade e a dinâmica dos ambientes reais.

2.4.1 Categorizações

Para compreender adequadamente os diversos estudos realizados no campo de *continual learning*, é necessário considerar algumas terminologias e categorizações recorrentes na literatura relacionadas ao problema de *catastrophic forgetting*.

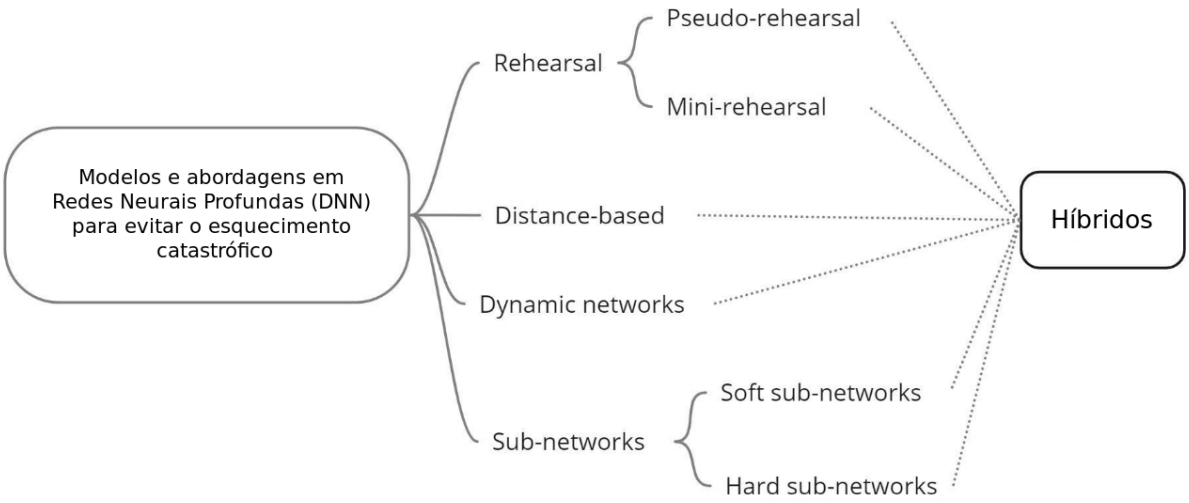
Os autores Parisi et al. (2019) propuseram uma categorização amplamente utilizada para os métodos desenvolvidos com o objetivo de mitigar o esquecimento catastrófico. Segundo esses autores, tais métodos podem ser organizados em três cenários principais (PARISI et al., 2019):

- Métodos de regularização: abordagens dessa categoria buscam reduzir o *catastrophic forgetting* ao impor restrições na atualização dos pesos das redes neurais. Essas estratégias são frequentemente inspiradas em modelos da neurociência teórica, os quais sugerem que o conhecimento consolidado pode ser protegido por meio de sinapses que apresentam diferentes níveis de plasticidade. Do ponto de vista computacional, essa ideia é geralmente modelada pela adição de termos de regularização à função de *loss*, penalizando alterações em parâmetros considerados relevantes para tarefas previamente aprendidas;
- Arquiteturas dinâmicas: compreendem abordagens que modificam a própria arquitetura da rede em resposta à chegada de novas informações, incorporando dinamicamente novos recursos computacionais, como a adição de neurônios, camadas ou módulos especializados;
- *Complementary Learning System* e *memory replay*: essa categoria é inspirada na teoria do *Complementary Learning Systems* (CLS), que modela a consolidação e a recuperação de memória por meio da interação entre estruturas análogas ao hipocampo e ao neocórtex em mamíferos. Nesse contexto, mecanismos de memorização e generalização são implementados por meio de estratégias de reapresentação de dados previamente aprendidos durante o treinamento de novas tarefas.

Ainda segundo Wang et al. (2024a), a categoria conhecida como *replay*, que envolve a reutilização de dados de tarefas anteriores durante o treinamento de novas tarefas, pode ser subdividida em três abordagens principais:

- *Experience replay*, em que uma parcela ou a totalidade dos dados de tarefas anteriores é armazenada e reutilizada no treinamento subsequente;
- *Generative replay*, também denominado *pseudo-rehearsal*, no qual um modelo generativo é utilizado para produzir novos dados sintéticos que aproximam a distribuição dos dados das tarefas anteriores, permitindo sua utilização no treinamento de tarefas futuras;
- *Feature replay*, abordagem semelhante ao *generative replay*, porém voltada à geração de representações intermediárias (*features*) em vez dos dados brutos.

Figura 8 – Categorização proposta por Aleixo et al. (2024) para o contexto de *catastrophic forgetting* em redes neurais profundas.



Fonte: Traduzido de (ALEIXO et al., 2024).

Apesar de eficazes em diversos cenários, métodos baseados em geração de dados podem introduzir novos desafios. Em particular, quando o modelo generativo também precisa ser atualizado ao longo do aprendizado contínuo, ele pode sofrer degradação de desempenho associada ao próprio *catastrophic forgetting*. Em abordagens de *feature replay*, pode ainda ocorrer o fenômeno conhecido como *representation shift*, no qual as representações geradas para dados de tarefas anteriores passam a se distorcer ao longo do treinamento, deixando de refletir adequadamente os dados originais que deveriam representar.

Os autores Wang et al. (2024a) também destacam a categoria denominada *optimization*. Nessa categoria encontram-se métodos que introduzem estratégias específicas de otimização durante o treinamento. A subcategoria *gradient projection* reúne abordagens que impõem restrições à atualização dos parâmetros de forma que os gradientes associados às novas tarefas permaneçam compatíveis com os espaços de gradiente definidos por tarefas anteriores. Esse princípio está relacionado, por exemplo, a métodos de *experience replay*, que procuram preservar regiões relevantes do espaço de gradientes associadas a dados previamente aprendidos.

A subcategoria *meta-learning* engloba abordagens que buscam aprender estratégias de atualização de parâmetros diretamente a partir dos dados, permitindo que o próprio processo de treinamento adapte a forma como os pesos da rede são atualizados. Já a subcategoria *loss landscape* refere-se a métodos que exploram explicitamente propriedades da paisagem de otimização, incluindo estratégias para atualização dinâmica de hiperparâmetros durante o treinamento, bem como abordagens baseadas em redes neurais não

supervisionadas (WANG et al., 2024a).

Outra categoria discutida por Wang et al. (2024a) é a de *representation-based approaches*, que reúne métodos focados na aprendizagem ou exploração de representações robustas dos dados. Nessa categoria, destacam-se três subcategorias principais:

- *Self-supervised learning*, em que modelos são treinados para aprender representações dos dados sem o uso explícito de rótulos;
- *Pre-training for downstream tasks*, que utiliza o conhecimento adquirido por redes pré-treinadas em grandes conjuntos de dados para facilitar o treinamento em tarefas subsequentes;
- *Continual pre-training*, que consiste na atualização contínua de representações por meio de técnicas de aprendizado auto-supervisionado ao longo do tempo.

A última categoria descrita por Wang et al. (2024a) corresponde às abordagens baseadas em arquiteturas de redes neurais. De maneira semelhante a outras categorizações presentes na literatura, esses métodos adaptam a estrutura do modelo ao longo do treinamento e podem ser divididos em três subcategorias principais:

- *Parameter allocation*, que compreende métodos nos quais diferentes subconjuntos de parâmetros são dedicados a tarefas específicas, podendo essa alocação ser fixa ou dinâmica;
- *Model decomposition*, que envolve a decomposição do modelo em múltiplos componentes especializados, permitindo a incorporação de novos recursos para tarefas adicionais, ao mesmo tempo em que mantém parâmetros compartilhados;
- *Modular networks*, que correspondem a abordagens em que novos módulos ou redes independentes são adicionados para acomodar novas tarefas, geralmente sem compartilhamento direto de parâmetros.

Os autores ressaltam que, embora essas categorias sejam apresentadas de forma separada, muitos métodos propostos na literatura combinam múltiplas estratégias para lidar com o aprendizado contínuo e o esquecimento catastrófico (ver Capítulo 3).

Com o avanço dos modelos de larga escala, como os *Vision Transformers* (DOSOVITSKIY et al., 2021) e modelos de linguagem pré-treinados, emergiu na literatura uma categoria denominada *Parameter-Efficient Continual Learning* (PECL) (SHI et al., 2024). Diferentemente das abordagens tradicionais, esses métodos partem do pressuposto de que um modelo já foi pré-treinado em larga escala e, portanto, possui representações genéricas e robustas que podem ser aproveitadas para o aprendizado contínuo. O objetivo central

dessas abordagens é adaptar o modelo pré-treinado a novas tarefas de forma eficiente, preservando o conhecimento já adquirido sem a necessidade de retreinar toda a rede. Nessa categoria destacam-se três abordagens principais (WANG et al., 2022a; SHI et al., 2024):

- *Prompt-based methods*: métodos que introduzem pequenos vetores treináveis, denominados *prompts*, na entrada ou em camadas intermediárias do modelo pré-treinado. Os parâmetros do *backbone* permanecem congelados, e apenas os *prompts* são atualizados durante o aprendizado de novas tarefas (WANG et al., 2022b; WANG et al., 2022a);
- *Adapter-based methods*: abordagens que inserem módulos adicionais (*adapters*) entre as camadas do modelo pré-treinado. Esses módulos são treinados para cada tarefa, enquanto os pesos originais da rede permanecem inalterados (ERMIS et al., 2022);
- *Representation-based methods*: métodos que exploram diretamente as representações geradas pelo modelo pré-treinado, utilizando classificadores lineares ou técnicas estatísticas para adaptar o modelo a novas tarefas sem modificar seus pesos (MCDONNELL et al., 2023).

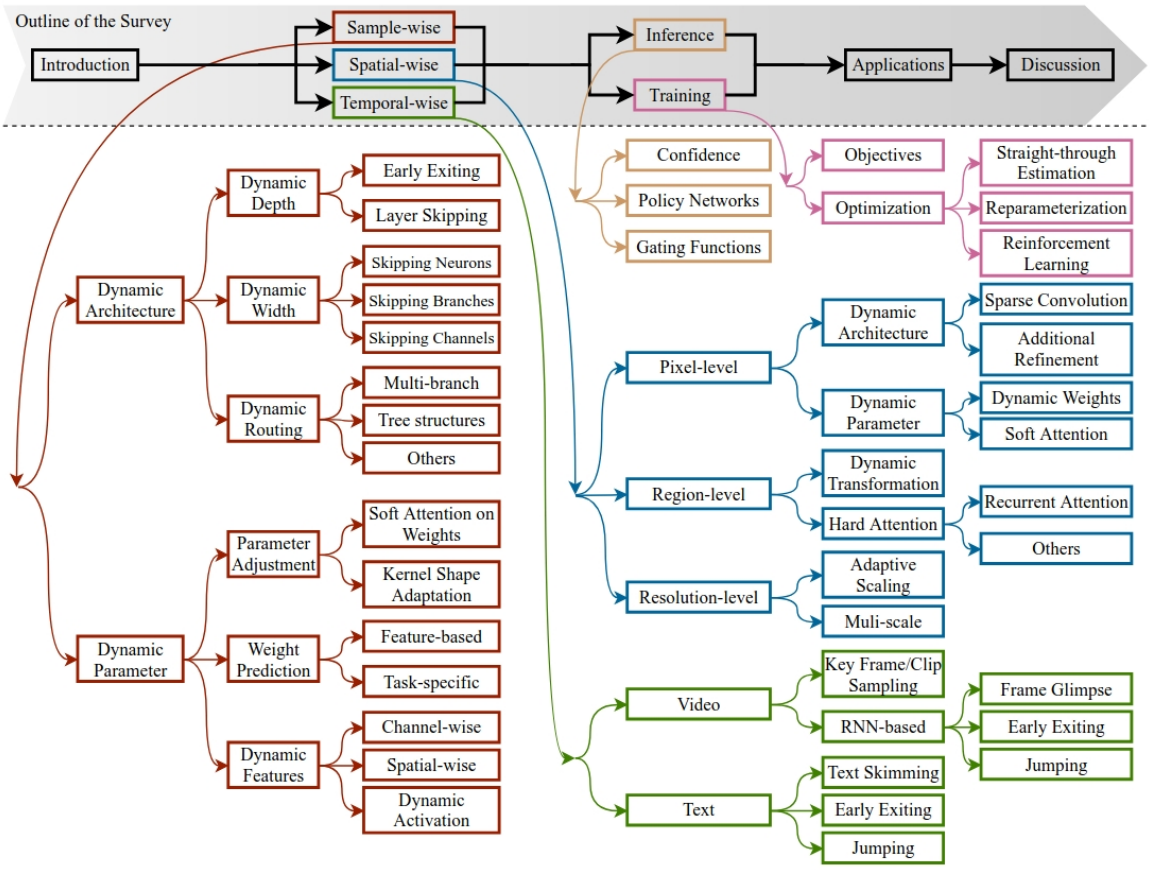
Cabe destacar que essa categoria é intrinsecamente dependente da qualidade e da generalização do modelo pré-treinado utilizado como *backbone*. Modelos como o ViT pré-treinado no ImageNet-21K (DOSOVITSKIY et al., 2021) têm se mostrado particularmente eficazes nesse contexto, uma vez que suas representações são suficientemente genéricas para cobrir uma ampla variedade de tarefas (SHI et al., 2024).

Outro tópico emergente na literatura de *deep learning* refere-se às chamadas *dynamic neural networks*. Essas arquiteturas são projetadas para adaptar sua estrutura ou seu fluxo computacional de acordo com diferentes tipos de entrada. Os autores Han et al. (2022) realizaram um levantamento abrangente dos principais métodos nessa área e propuseram diferentes categorizações para esse cenário de redes neurais adaptativas. A Figura 9 apresenta a categorização proposta por esses autores.

Cabe destacar que, embora essa linha de pesquisa compartilhe algumas terminologias com o campo de *continual learning*, as redes neurais dinâmicas não são necessariamente propostas com o objetivo direto de mitigar o *catastrophic forgetting*. Em muitos casos, sua motivação principal está relacionada à adaptabilidade do modelo a diferentes tipos de dados ou tamanhos de entrada, como sequências de vídeo ou textos. Ainda assim, tais arquiteturas podem contribuir indiretamente para a mitigação do esquecimento catastrófico ao permitir a incorporação de novos recursos computacionais ao longo do treinamento.

As categorizações discutidas nesta seção contribuem para organizar e consolidar o conhecimento na área, permitindo uma compreensão mais sistemática das diferentes abordagens propostas na literatura. É importante notar, contudo, que essas classificações

Figura 9 – Categorização proposta por Han et al. (2022) para redes neurais com arquiteturas dinâmicas.



Fonte: (HAN et al., 2022).

não são estáticas e podem evoluir à medida que novos métodos e paradigmas de aprendizado contínuo são desenvolvidos.

Por fim, ao longo deste trabalho, o termo *replay* será utilizado para se referir de forma geral à categoria de métodos que reutilizam informações de tarefas anteriores durante o treinamento, seja por meio do armazenamento de dados originais, seja pela geração de dados sintéticos com métodos generativos, como *autoencoders* e *GANs*. Nesse contexto, também serão empregados os termos *pseudo-rehearsal* e *generative replay*. O termo *rehearsal*, por sua vez, será utilizado especificamente para descrever abordagens que reutilizam diretamente parte ou a totalidade dos dados originais de tarefas anteriores durante o treinamento de novas tarefas.

2.4.2 Métricas para avaliar esquecimento

Além das categorizações, é importante definir as métricas comumente utilizadas para avaliar os métodos na área de esquecimento catastrófico. Os autores Wang et al. (2024a) levantaram as principais métricas utilizadas na literatura. A *performance* geral de

um sistema de aprendizado contínuo pode ser definida pela acurácia média e pela acurácia média incremental, definidas por

$$AM_k = \frac{1}{k}\sum_{j=1}^{k} a_{k,j}\,, \tag{2.2}$$

$$AMI_k = \frac{1}{k}\sum_{i=1}^{k} AM_i \tag{2.3}$$

onde k representa o número total de tarefas aprendidas até o momento, e $a_{k,j}$ corresponde à acurácia do modelo na tarefa j após o treinamento até a tarefa k.

Wang et al. (2024a) destacam outras duas importantes métricas relacionadas a perda de memória: medida de esquecimento (FM, do inglês *Forgetting Measure*) e a medida de estabilidade (IM, do inglês *Intransience Measure*). A primeira, segundo os autores Wang et al. (2024a) é dada por

$$FM_k = \frac{1}{k-1}\sum_{j=1}^{k-1} f_{j,k}\,, \tag{2.4}$$

em que $f_{j,k}$ pode ser definido como a diferença entre a sua acurácia máxima nas tarefas anteriores e a acurácia atual:

$$f_{j,k} = \max_{i\in 1,\ldots,k-1}(a_{i,j} - a_{k,j}), \forall j < k\,, \tag{2.5}$$

enquanto a medida de estabilidade pode ser definida formalmente como

$$IM_k = a_k^* - a_{k,k}\,, \tag{2.6}$$

onde a_k^* é a acurácia do modelo treinado com todos as bases de dados das tarefas anteriores unidas (*joint*), até a tarefa k. Dessa forma, a métrica IM mede a inabilidade de um modelo em aprender uma nova tarefa (WANG et al., 2024a). Os autores destacam ainda duas outras métricas, definidas, respectivamente, nas Equações 2.7 e 2.8, *backward transfer* (BWT) e *forward transfer* (FWT):

$$BWT_k = \frac{1}{k-1}\sum_{j=1}^{k-1}(a_{k,j} - a_{j,j})\,, \tag{2.7}$$

$$FWT_k = \frac{1}{k-1}\sum_{j=2}^{k}(a_{j,j} - a_j)\,. \tag{2.8}$$

BWT é compreendida como a influência média do aprendizado da tarefa k nas tarefas anteriores. Quando BWT for superior a 0, indica que não há esquecimento catastrófico. Quando BWT for menor que 0, indica esquecimento das tarefas anteriores. A FWT pode ser compreendida como o oposto de BWT, ou seja, a influência média das tarefas anteriores no aprendizado da tarefa k, dado que a_j é a acurácia de um modelo treinado na tarefa j. Se FWT for maior que 0, indica transferência positiva do aprendizado das tarefas anteriores para as novas tarefas, e, se for inferior a 0, indica que a transferência do aprendizado não foi benéfica (WANG et al., 2024a).

2.5 *Generative Adversarial Networks*

GANs têm sido extensivamente utilizadas no processo de geração de novas imagens a partir de diferentes contextos de aplicações. A modelagem de novas imagens pode ser aprendida a partir da distribuição de probabilidade de um conjunto de imagens qualquer (GOODFELLOW et al., 2014). Tal aprendizado pode ser aplicado à geração de novas imagens (KARRAS; LAINE; AILA, 2021), à transferência de estilos de um conjunto de imagens para outro (ZHU et al., 2017), à modelagem de novas imagens combinando espaços discriminativos de contextos diferentes (RADFORD; METZ; CHINTALA, 2016), dentre outras aplicações.

No trabalho precursor de Karras, Laine e Aila (2021), que introduziu o conceito de redes adversárias, existem duas redes neurais, o gerador e o discriminador. Segundo Karras, Laine e Aila (2021), para aprender uma distribuição p_g do gerador sobre os dados x, é definida uma distribuição das variáveis de ruído de entrada $p_z(z)$, e então representado um mapeamento para o espaço de dados como $G(z;\theta_g)$, em que G é uma função diferenciável representada por um *perceptron* multicamadas com parâmetros θ_g. Também é definido um segundo *perceptron* multicamadas $D(x;\theta_d)$ que gera um valor único. $D(x)$ representa a probabilidade de que x tenha sido originado dos dados reais, ao invés de p_g. Assim, o modelo D é treinado para maximizar a probabilidade de atribuir o rótulo correto tanto aos exemplos de treinamento quanto às amostras de G. Simultaneamente, G é treinado para minimizar $\log(1 - D(G(z)))$. Sendo assim, as redes D e G estão em uma constante competição definida como (KARRAS; LAINE; AILA, 2021)

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D(G(z)))\right] . \tag{2.9}$$

A *Wasserstein* GAN (WGAN) foi proposta por Arjovsky, Chintala e Bottou (2017) para mitigar problemas de instabilidade no treinamento de GANs convencionais. Em vez de utilizar a divergência de Kullback-Leibler (JOYCE, 2011) ou a divergência de Jensen-Shannon (MENéNDEZ et al., 1997), a WGAN baseia-se na distância de *Wasserstein*. A função de distância pode ser definida como

$$W(P_r, P_g) = \inf_{\gamma\in\Pi(P_r,P_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\|x - y\|\right] , \tag{2.10}$$

em que o $\gamma \in \Pi(P_r, P_g)$ diz respeito ao conjunto de todas as distribuições $\gamma(x, y)$, indicando o "esforço" necessário para transformar a distribuição P_r em P_g. Essa função denota então o custo do menor "esforço" para que essa transformação aconteça (ARJOVSKY; CHINTALA; BOTTOU, 2017).

Além disso, a WGAN implementa um *clipping* dos pesos em um intervalo predefinido. Esse ajuste permite que o treinamento seja mais estável e que a perda da WGAN tenha um significado mais direto em termos de qualidade da distribuição gerada.

A WGAN com *Gradient Penalty* (WGAN-GP) (GULRAJANI et al., 2017) aprimora ainda mais a WGAN adicionando uma penalidade ao gradiente no discriminador, removendo a necessidade de *clipping* e promovendo a propriedade de *Lipschitz*, motivo pelo qual Arjovsky, Chintala e Bottou (2017) propuseram o *weight clipping*, de maneira mais robusta. Então, a penalidade no gradiente é dada por

$$\lambda \mathbb{E}_{\hat{x}\sim\hat{P}_x}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] . \tag{2.11}$$

Assim, a função de custo de uma WGAN-GP é dada por

$$L = \mathbb{E}_{\tilde{x}\sim P_g}\left[D(\tilde{x})\right] - \mathbb{E}_{x\sim P_r}\left[D(x)\right] + \lambda \mathbb{E}_{\hat{x}\sim\hat{P}_x}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] , \tag{2.12}$$

no qual o primeiro termo é a função de distância de *Wassertein* somado a penalidade do gradiente (GULRAJANI et al., 2017).

A fim de avaliar a qualidade das imagens geradas pelas GANs, é importante comparar os resultados obtidos com o quão próximo se aproximam do objetivo inicial, ou seja, da distribuição da base de dados de entrada.

A métrica *Structural Similarity Index* (SSIM), proposta por Wang et al. (2004), é uma medida de similaridade perceptual que compara as características de baixo nível das imagens geradas com as imagens originais. A métrica é composta por três análises de informações: contraste, luminância e textura. Dessa forma, segundo os autores Wang et al. (2004), a medida SSIM tenta simular a percepção humana, ao invés de fazer uma comparação *pixel* a *pixel*, como é no erro quadrático médio (MSE, do inglês *Mean Squared Error*). A fórmula da SSIM é definida por, para duas imagens x e y em uma janela local

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} , \tag{2.13}$$

em que: μ_x e μ_y são as médias das imagens x e y, e dizem respeito a luminância; σ_x^2 e σ_y^2 são as variâncias locais, e dizem respeito ao contraste da imagem; σ_{xy} é a covariância local entre x e y; e C_1 e C_2 são constantes definidas como $C_1 = (k_1 L)^2$ e $C_2 = (k_2 L)^2$, onde L é o valor máximo possível de um *pixel* (geralmente 255) e k_1 e k_2 são constantes empiricamente definidas (WANG et al., 2004). O valor do SSIM varia entre -1 e 1, sendo que quanto mais próximo de 1, mais similares são as imagens.

Os autores Wang, Simoncelli e Bovik (2003) propuseram uma versão multi escala da SSIM, intitulada *Multi-Scale Structural Similarity* (MS-SSIM), com o propósito de calcular a similaridade com diferentes resoluções da imagem, de forma a não se manter fixa com uma única resolução, permitindo expandir e diminuir o nível de detalhes das imagens comparadas. A métrica é dada por

$$SSIM(x,y) = \left[l_M(x,y)\right]^{\alpha_M} \prod_{j=1}^{M} \left[c_j(x,y)\right]^{\beta_j} \left[s_j(x,y)\right]^{\gamma_j} , \tag{2.14}$$

dado que $l_M(x, y)$ é o cálculo da luminância, $c_j(x, y)$ é o contraste na escala j, $s_j(x, y)$ é a textura na escala j e M é o total de escalas. Os expoentes α_M, β_j e γ_j são utilizados para ajustar a importância de cada componente (WANG; SIMONCELLI; BOVIK, 2003). Os valores MS-SSIM variam entre 0 e 1, sendo 1 o máximo de similaridade entre as duas imagens.

Outra métrica de similaridade é a *Cosine Similarity* (CSIM) (HAN; KAMBER; PEI, 2012). Para comparar duas imagens, os vetores de características são necessários para calcular a similaridade, dado a fórmula

$$sim(x, y) = \frac{v_x \cdot v_y}{||v_x||||v_y||}\,, \tag{2.15}$$

em que v_x e v_y são os vetores de características das imagens x e y, respectivamente.

Peak Signal to Noise Ratio (PSNR) calcula a força de um sinal e a força do ruído que distorce a imagem (WINKLER; MOHANDAS, 2008). PSNR é derivado do MSE, definido como

$$PSNR(x, y) = 10 \cdot \log_{10} \frac{R^2}{MSE(x, y)}\,, \tag{2.16}$$

em que R representa o valor máximo de uma *pixel* da imagem.

A métrica IS (*Inception Score*), proposta pelos autores Salimans et al. (2016), é uma métrica que utiliza uma rede pré-treinada, como a InceptionV3 (SZEGEDY et al., 2016), para calcular a probabilidade de pertencimento de uma imagem gerada a diferentes classes. Intuitivamente, o cálculo deve sinalizar que imagens geradas com baixa qualidade terão uma ampla distribuição entre as classes, enquanto imagens com boa qualidade, estão mais concentradas em poucas classes. A métrica tem o objetivo de calcular a qualidade e a diversidade da imagem. A fórmula é dada por

$$IS = \exp\left(\mathbb{E}_x \left[D_{\mathrm{KL}}\left(p(y|x) \,\|\, p(y)\right)\right]\right)\,, \tag{2.17}$$

em que x é a imagem gerada, $p(x, y)$ é a distribuição das classes atribuída pela rede InceptionV3, D_{KL} é a divergência de Kullback-Leibler, medindo a diferença entre $p(y|x)$ e $p(y)$ (a distribuição marginal das classes para todas as imagens) sendo que $p(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|x_i)$, com N sendo o total de imagens.

Os autores Heusel et al. (2017) propuseram uma métrica mais robusta utilizando a rede InceptionV3 (SZEGEDY et al., 2016), chamada *Fréchet Inception Distance* (FID), que utiliza a distância de Fréchet (FRÉCHET, 1957) para comparar as distâncias multivariadas das características das imagens analisadas. Segundo os autores Heusel et al. (2017), a métrica utiliza as características extraídas das imagens reais e das imagens geradas, utilizando a InceptionV3 como uma rede pré-treinada para extração de características. As distribuições são comparadas com a equação

$$FID = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right), \tag{2.18}$$

dado que μ_r e μ_g são as médias dos vetores de características das imagens reais e geradas, respectivamente, Σ_r e Σ_g são as matrizes de covariância dos vetores e Tr é o traço da matriz (HEUSEL et al., 2017). Nesta métrica, quanto maior o valor de FID, menor é a semelhança entre as imagens comparadas, enquanto valores mais próximos de 0 indicam uma perfeita semelhança.

2.6 Considerações finais

Neste capítulo, revisam-se os conceitos fundamentais relacionados a este estudo, abrangendo os temas teóricos e práticos de reconhecimento de emoções, neuropsicologia, aprendizado contínuo, integridade de memória e *catastrophic forgetting* e GANs.

Na seção de reconhecimento de emoções, destacam-se os principais processos computacionais e psicológicos envolvidos na identificação de expressões faciais, além de um detalhamento da evolução dos métodos e bases de dados na área de FER – que, a partir de 2013, foi possível perceber um aumento no número de bases de dados e, consequentemente, abordagens para reconhecimento de emoções, devido, principalmente, às publicações dos autores Goodfellow et al. (2013a), Dhall et al. (2015). Na seção de neuropsicologia, abordam-se os principais mecanismos cognitivos do cérebro e do comportamento humano que servem de inspiração para os métodos propostos na literatura para mitigar o esquecimento catastrófico em redes neurais convolucionais.

Ao aprofundar os principais tópicos de aprendizado contínuo, destacam-se as limitações atuais das CNNs em contextos de aprendizado contínuo. Além disso, expuseram-se as principais categorizações propostas na literatura a respeito dos diferentes métodos e abordagens relacionados à integridade de memória. Também se destacam as métricas utilizadas em *continual learning*, oferecendo uma visão prática e teórica sobre como mensurar a degradação de desempenho em tarefas anteriores à medida que novos conhecimentos são incorporados a uma CNN.

Por fim, a seção sobre GANs embasa o desenvolvimento do método ECgr, proposto neste trabalho. A capacidade das GANs em gerar imagens sintéticas possibilita a criação de representações do conhecimento anterior da rede, mitigando os efeitos do esquecimento catastrófico. Também apresentam-se as principais métricas utilizadas para avaliar a qualidade das imagens geradas pelas GANs, com o propósito de embasar as métricas utilizadas neste trabalho para avaliar o método proposto.

Esses conceitos fornecem uma compreensão abrangente dos desafios e das possíveis soluções para o aprendizado contínuo em redes neurais. A fundamentação teórica apresentada neste capítulo representa a base conceitual para o desenvolvimento das metodologias e experimentos descritos nos capítulos subsequentes.

3 Estado da arte

O esquecimento catastrófico, um desafio inerente às redes neurais, tem instigado diversas pesquisas com o objetivo de mitigar o seu impacto. Neste capítulo, ocorre um aprofundamento em algoritmos importantes que lidam com todos os aspectos relacionados ao esquecimento catastrófico. A fim de melhor explorar os trabalhos relacionados a *continual learning* e *catastrophic forgetting*, adotou-se a categorização em 4 grupos: (i) regularização, (ii) *replay*, (iii) arquiteturas dinâmicas e (iv) baseados em modelos pré-treinados.

3.1 Regularização

Nesta seção são apresentados os trabalhos que utilizam métodos de regularização (ver Seção 2.4 para mais detalhes sobre as categorias) para mitigar o esquecimento catastrófico. Discutem-se as definições técnicas de cada trabalho, bem como seus resultados.

O Elastic Weight Consolidation (EWC) (KIRKPATRICK; AL., 2017) é uma técnica de regularização para mitigar o esquecimento catastrófico, o qual utiliza uma abordagem que penaliza mudanças nos parâmetros importantes para as tarefas anteriores. Inspirado na plasticidade seletiva do cérebro, os autores Kirkpatrick e al. (2017) introduziram um termo de regularização que mede a importância dos parâmetros para preservar o desempenho em tarefas anteriores enquanto permite a adaptação a novas informações. Isso significa que o EWC tenta manter uma taxa de erro baixa para a tarefa anterior e para a nova tarefa, como mostra a Figura 10. A função de custo regularizada do EWC é formulada como (KIRKPATRICK; AL., 2017):

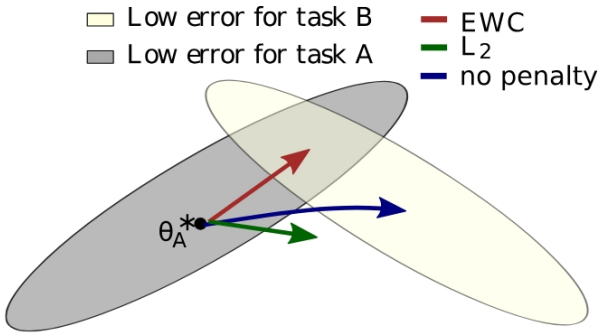
$$\mathcal{L}_{\text{EWC}}(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2 \tag{3.1}$$

em que:

- $\mathcal{L}(\theta)$ representa a função de custo original,
- λ é definido como hiper-parâmetro que controla a intensidade da regularização,
- F_i é o valor esperado da Informação de Fisher para o parâmetro i, indicando sua importância para tarefas anteriores,
- θ_i e θ_i^* são os valores atuais e iniciais dos parâmetros, respectivamente.

Essa formulação permite balancear a retenção do conhecimento adquirido e a adaptação do modelo, minimizando o impacto de novos aprendizados nos parâmetros considerados essenciais para as tarefas já aprendidas.

Figura 10 – Espaço de cálculo dos erros das redes neurais, representando os conjuntos de erros de duas tarefas, A e B. Os autores Kirkpatrick e al. (2017) demonstram que o cenário ideal é encontrar uma função que direciona a rede neural para a ponto ilustrado pela seta vermelha.



Fonte: (KIRKPATRICK; AL., 2017).

Learning without Forgetting (LwF) é uma técnica que explora a destilação de conhecimento para transferir informações de um modelo treinado em tarefas anteriores para um modelo novo. No método de Li e Hoiem (2018), um modelo retém o conhecimento anterior através de uma rede auxiliar, que age como um "guardião" das representações anteriores. Durante o treinamento em novas tarefas, o método ajusta o modelo sem comprometer os conhecimentos anteriores. Nos experimentos, LwF se destacou por manter acurácia em tarefas antigas ao mesmo tempo em que possibilita a adaptação a novas. A função de custo do LwF consiste em uma combinação da função de custo para as novas tarefas ($\mathcal{L}_{\text{new}}$) e da função de custo de destilação ($\mathcal{L}_{\text{distill}}$), formulada como (LI; HOIEM, 2018)

$$\mathcal{L}_{\text{LwF}} = \mathcal{L}_{\text{new}} + \alpha \cdot \mathcal{L}_{\text{distill}} \,, \tag{3.2}$$

em que $\mathcal{L}_{\text{new}}$ é a função de custo para as novas tarefas, $\mathcal{L}_{\text{distill}}$ representa a função de custo de destilação, que preserva a resposta da rede antiga para os dados passados e α é um hiper-parâmetro que pondera a importância da retenção de conhecimento anterior. Esse balanceamento possibilita que o modelo LwF retenha conhecimentos adquiridos previamente enquanto se adapta a novas informações, sem necessidade de acessar diretamente os dados das tarefas anteriores.

Os autores Zenke, Poole e Ganguli (2017) propuseram o *Synaptic Intelligence* (SI). Essa abordagem associa a importância às conexões sinápticas do modelo com base em suas contribuições para o desempenho em tarefas anteriores. Ao treinar uma nova tarefa, SI preserva os pesos de sinapses importantes, aplicando uma penalização diferencial para conexões relevantes, e protegendo essas sinapses de alterações excessivas. Inspirado no funcionamento das conexões neuronais, o SI proporciona uma preservação mais granular

do conhecimento passado. Em experimentos, os autores demonstraram que o SI oferece excelente retenção do conhecimento em cenários de aprendizado contínuo, destacando-se por seu desempenho em redes que enfrentam múltiplas tarefas sequenciais. A fórmula para a penalização diferencial do SI pode ser definida como

$$\mathcal{L}_{\text{SI}} = \mathcal{L}_{\text{new}} + \lambda \sum_i \omega_i (\theta_i - \theta_i^*)^2 \,. \tag{3.3}$$

Nesta fórmula, $\mathcal{L}_{\text{new}}$ representa a função de custo associada à tarefa atual, ω_i denota a importância acumulada da i-ésima sinapse, θ_i indica o peso atual do parâmetro, θ_i^* o valor desse parâmetro otimizado anteriormente e λ é um fator de regularização que controla o grau de proteção das sinapses relevantes.

Memory Aware Synapses (MAS), proposto por Aljundi et al. (2018), é uma técnica que identifica a importância dos parâmetros de rede baseando-se na sensibilidade da função de saída à mudanças nesses parâmetros. Esse método cria uma "memória sináptica" que preserva conexões importantes, garantindo que alterações excessivas não ocorram em sinapses cruciais para tarefas anteriores. Para o aprendizado de uma nova tarefa T_n, além da função de custo da nova tarefa $L_n(\theta)$, é adicionado um regularizador que penaliza as mudanças nos parâmetros considerados importantes:

$$L(\theta) = L_n(\theta) + \lambda \sum_{i,j} \Omega_{ij} (\theta_{ij} - \theta_{ij}^*)^2 \tag{3.4}$$

em que Ω_{ij} é o grau de importância de um parâmetro, medido pela magnitude de um gradiente g_{ij}, ou seja, o quanto uma alteração neste parâmetro irá influenciar no resultado final para um dado x_k:

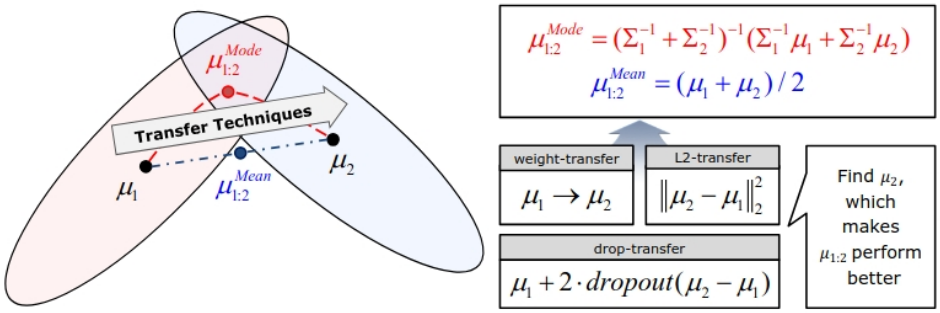
$$\Omega_{ij} = \frac{1}{N} \sum_{k=1}^{N} \| g_{ij}(x_k) \| \tag{3.5}$$

O algoritmo MAS, ao focar na contribuição de cada parâmetro para a *performance* global da rede, preserva o conhecimento sem necessidade de rótulos das tarefas anteriores.

Lee et al. (2018) criaram o método *Incremental Moment Matching* (IMM), que visa minimizar o esquecimento catastrófico combinando as distribuições de parâmetros das redes treinadas em tarefas anteriores com a rede da tarefa atual, chamado de *mean-IMM*. Por meio de uma combinação ponderada dos parâmetros, essa abordagem adapta-se à nova tarefa enquanto preserva informações das tarefas anteriores. Outra variação desse método apresentada pelos autores é o *mode-IMM*, que busca encontrar o máximo da combinação das distribuições posteriores Gaussianas. Ambos os métodos são úteis em redes que enfrentam um grande número de tarefas, pois reduzem a necessidade de recomeçar o treinamento ou calibrar novamente os parâmetros. A Figura 11 demonstra como os algoritmos funcionam.

Lee et al. (2018) mostraram que o IMM é eficaz em manter a estabilidade e a *performance* do modelo em uma ampla variedade de tarefas sequenciais, destacando-se pela simplicidade e eficácia.

Figura 11 – Ilustração do método IMM proposto por Lee et al. (2018). No lado esquerdo, a representação dos espaços de parâmetros, em que o *Mean-IMM* faz uma média dos parâmetros de duas redes neurais, enquanto o *mode-IMM* busca encontrar um máximo na mistura das distribuições Gaussianas. Para que o IMM seja aplicável, o espaço de busca da função de custo entre as médias μ_1 e μ_2 deve ser razoavelmente suave e semelhante a uma forma convexa.



Fonte: (LEE et al., 2018).

O método *Riemannian Walk* (RW), proposto por Chaudhry et al. (2018), aplica uma penalização aos parâmetros que sofrem alterações drásticas comparado ao estado de tarefas anteriores, limitando o alcance das atualizações conforme necessário para preservar conhecimento. Esse método incorpora um cálculo baseado em geometria diferencial para adaptar os parâmetros sem comprometer as informações de tarefas anteriores. A função final de custo pode ser definida como a combinação da matriz de Fisher, com o *KL divergence* e com o cálculo de importância:

$$L_k(\theta) = L_k(\theta) + \lambda \sum_{i=1}^{P} \left(F\theta_{k-1}^i + s_{t_{k-1}}^{t_0}(\theta_i) \right) \left(\theta_i - \theta_{k-1}^i \right)^2 \tag{3.6}$$

em que $s_{t_{k-1}}^{t_0}(\theta_i)$ é a pontuação acumulada entre a interação t_0 e t_{k-1} (importância) e λ é um hiper-parâmetro de regularização.

Os autores Adel, Zhao e Turner (2020) propuseram o *Continual Learning with Adaptive Weights* (CLAW), uma extensão do EWC na qual os coeficientes de regularização são ajustados dinamicamente conforme a importância dos parâmetros para cada tarefa. Essa adaptação busca reduzir a rigidez do EWC original, tornando o modelo mais flexível frente a sequências longas de tarefas. No entanto, assim como outros métodos baseados em regularização, o CLAW ainda depende da estimação precisa da importância dos parâmetros, o que pode ser sensível à ordem das tarefas e à qualidade dos dados disponíveis. De forma semelhante, os autores Zenke, Poole e Ganguli (2017) aprimoraram o método SI ao introduzir estratégias mais eficientes de acumulação da importância sináptica, permitindo preservar conexões relevantes com menor custo computacional. Apesar disso, esses métodos continuam limitados por não utilizarem dados explícitos de tarefas anteriores, o que pode restringir sua capacidade de retenção em cenários com alta variabilidade.

O *Meta-Experience Replay* (MER), proposto por Riemer et al. (2019), combina regularização com *replay* ao incorporar princípios de meta-aprendizagem. O método busca otimizar os parâmetros de forma que o aprendizado de novas tarefas cause o mínimo de interferência nas anteriores, ajustando dinamicamente a atualização dos pesos com base em sua utilidade ao longo do tempo. Embora apresente bons resultados em fluxos contínuos de dados, o MER depende da manutenção de um buffer de experiências passadas, o que pode ser inviável em cenários com restrições de armazenamento ou privacidade, diferentemente de abordagens baseadas exclusivamente em geração de dados sintéticos.

Ermis et al. (2022) introduziram o método *Adaptive Distillation of Adapters* (ADA), voltado ao aprendizado contínuo em arquiteturas *transformer*. A proposta utiliza módulos *Adapter*, que são camadas leves inseridas na rede principal, permitindo adaptar o modelo a novas tarefas sem modificar significativamente os parâmetros já aprendidos. Essa estratégia é particularmente relevante em aprendizado contínuo, pois reduz a interferência entre tarefas e limita o crescimento do número de parâmetros, mantendo a eficiência computacional. O método consiste no treinamento de novos *Adapters* e redes de classificação para cada tarefa, seguido por uma fase de consolidação via distilação, onde um *Adapter* é selecionado com base em medidas de transferibilidade e integrado ao modelo. A regularização ocorre por meio de uma perda de distilação entre os *logits* dos modelos antigos e do modelo consolidado, com dados selecionados via amostragem. Em experimentos nos conjuntos CIFAR-100 (KRIZHEVSKY, 2009) e ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), o ADA atingiu ganhos expressivos em eficiência computacional e acurácia.

Os autores Kara, Churamani e Gunes (2021) investigaram o uso de aprendizado contínuo como estratégia para mitigar vieses em sistemas de FER e de reconhecimento de AUs aplicados à robótica afetiva. Partindo do pressuposto de que robôs sociais estarão cada vez mais presentes em interações humanas, o estudo ressalta a importância de que tais sistemas sejam capazes de interpretar emoções de maneira justa e imparcial, independentemente de variações demográficas nos dados de entrada. O trabalho argumenta que o CL oferece uma solução promissora para lidar com o fluxo contínuo e não estacionário de dados durante essas interações, ao permitir a adaptação incremental sem sofrer de esquecimento catastrófico. A avaliação empírica realizada em cenários simulados de interação humano-robô demonstrou que abordagens baseadas em CL, especialmente aquelas fundamentadas em métodos de regularização, superam técnicas convencionais em termos de acurácia. Os resultados indicam que o CL pode desempenhar um papel central na construção de sistemas afetivos embarcados mais equitativos, adaptativos e eticamente responsáveis. Com base nas evidências obtidas em aplicações de robótica afetiva, os autores Churamani, Kara e Gunes (2023) expandiram a investigação sobre justiça algorítmica em FER, propondo uma abordagem mais geral baseada em aprendizado contínuo sob o paradigma de aprendizado incremental por domínio (*Domain-IL*). O objetivo foi analisar a eficácia do CL na mitigação de vieses causados por distribuições desbalanceadas de

dados, sem restringir o escopo ao contexto robótico. Para isso, os autores conduziram uma avaliação comparativa entre métodos tradicionais de mitigação de viés e abordagens baseadas em CL, considerando tanto o desempenho classificatório quanto métricas de equidade. Os experimentos foram realizados em duas tarefas distintas: reconhecimento de expressões faciais, utilizando a base RAF-DB (LI; DENG; DU, 2017), e detecção de Unidades de Ação, utilizando a base BP4D (ZHANG et al., 2024b). Os resultados confirmaram os achados anteriores, evidenciando que técnicas baseadas em CL são mais eficazes na promoção de justiça algorítmica, sem comprometer a acurácia. O estudo reforça o papel do CL como uma ferramenta versátil e robusta para o desenvolvimento de sistemas FER mais justos e generalizáveis.

Os autores Gao et al. (2023) propuseram uma nova abordagem denominada *SSA-ICL*, que integra mecanismos de atenção espectral e espacial com aprendizado contínuo. Os autores criaram o módulo *Spectral and Spatial Attention* (SSA), que combina informações espaciais e espectrais com o objetivo de enriquecer a capacidade representacional do modelo. Também propuseram outro módulo, denominado *Intra-dataset Continual Learning* (ICL), que subdivide um conjunto de dados em subconjuntos balanceados. Esses subconjuntos são utilizados em ciclos de treino incremental, promovendo a aprendizagem de representações mais robustas e equitativas. A avaliação empírica foi realizada nas bases de dados AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2019) e RAF-DB (LI; DENG; DU, 2017), demonstrando ganhos expressivos em relação a métodos com módulos de atenção convencionais, com melhorias de acurácia entre 3,8% e 6,7%. O método SSA-ICL também apresentou desempenho superior ou comparável a técnicas de estado da arte, alcançando 65,78% de acurácia na AffectNet e 89,44% na RAF-DB, consolidando-se como uma alternativa eficaz para lidar simultaneamente com desafios semânticos e estruturais em FER.

O método proposto por Cotogni et al. (2025) apresenta uma abordagem *exemplar-free* para aprendizado incremental com *Vision Transformers*, introduzindo dois mecanismos complementares: *Gated Class-Attention* (GCAB), que aplica máscaras aprendidas à atenção de classes para isolar neurônios relevantes por tarefa e prevenir interferência negativa, e *Cascaded Feature Drift Compensation* (CFDC), que utiliza projeções regulares para alinhar as características da tarefa atual com as anteriores. As máscaras de atenção são acumuladas entre tarefas. Avaliações em (KRIZHEVSKY, 2009) e ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) demonstraram que o método superou abordagens com até 500 exemplos armazenados, além de dobrar a *performance* de outros métodos *exemplar-free* com *transformers*, mantendo eficiência de memória e robustez a hiperparâmetros.

3.2 *Replay*

Nesta seção são apresentados os trabalhos que utilizam algum dos métodos reconhecidos na literatura de *replay* (ver Seção 2.4 para mais detalhes sobre as categorias). Discutem-se os principais métodos de cada trabalho, bem como seus resultados.

O trabalho pioneiro de Robins (1995) lançou as bases para o conceito de *pseudo-rehearsal*, que busca enfrentar o *catastrophic forgetting* em redes neurais ao introduzir uma estratégia de retenção através da geração de exemplos sintéticos que simulam memórias de tarefas anteriores. Ao invés de reter diretamente dados antigos, o método cria instâncias sintéticas para representar o conhecimento acumulado, o que permite que o modelo mantenha o aprendizado anterior enquanto se adapta a novas informações. Esse mecanismo imita uma "repetição indireta" do que foi aprendido, sendo um precursor para abordagens que mais tarde usariam redes generativas para preservar e integrar conhecimento. Com isso, Robins pavimentou o caminho para futuras técnicas de *rehearsal* e influenciou o desenvolvimento de estratégias mais complexas, como o *generative replay*.

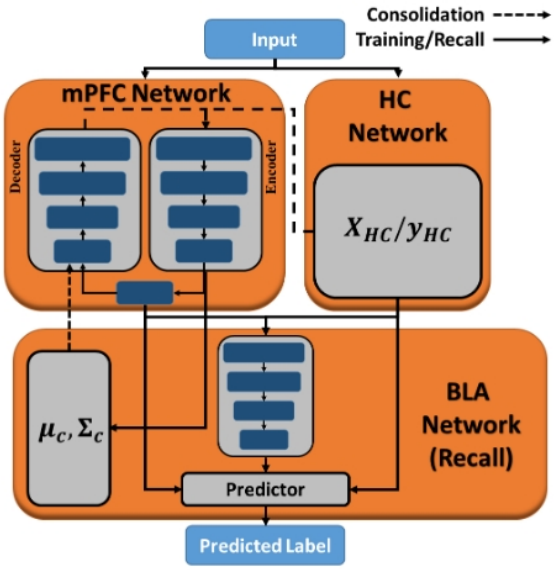
Shin et al. (SHIN et al., 2017) introduziram o *generative replay*, uma subcategoria do conceito de *replay*, que combina redes neurais profundas com modelos generativos para mitigar o esquecimento. Nesse método, uma rede generativa (como uma GAN) gera dados que simulam experiências passadas, que são então usadas para treinar a rede principal ao lado de novas informações. A inovação está no uso de exemplos sintetizados, o que não só reduz a necessidade de armazenamento massivo de dados antigos, como também oferece uma alternativa viável para redes neurais que lidam com dados sensíveis. Os resultados experimentais de Shin mostraram que a técnica consegue preservar a acurácia em tarefas anteriores enquanto absorve novos conhecimentos.

Inspirado em mecanismos cerebrais, Kemker e Kanan (2017) publicaram o FearNet, que traz uma abordagem híbrida ao aprendizado incremental, utilizando *rehearsal* e *pseudo-rehearsal* para lidar com o esquecimento. O modelo é dividido em múltiplos módulos que imitam o armazenamento e a integração de memórias, preservando tanto o conhecimento consolidado quanto a plasticidade para novos aprendizados. A Figura 12 demonstra a estrutura geral da rede proposta.

A inovação do algoritmo criado por Kemker e Kanan (2017) reside em sua capacidade de decidir automaticamente quando usar memórias passadas ou criar memórias temporárias para novas tarefas, adaptando-se a diferentes contextos. Kemker e Kanan (2017) demonstraram que essa arquitetura melhora significativamente a acurácia em tarefas incrementais, destacando-se entre modelos inspirados em estruturas neurais biológicas por seu equilíbrio entre retenção e adaptação.

A abordagem *Incremental Classifier and Representation Learning* (iCaRL), proposta por Rebuffi et al. (2017), introduz o conceito de *incremental rehearsal*, que seleciona

Figura 12 – Arquitetura da rede neural FearNet.



Fonte: (KEMKER; KANAN, 2017).

e armazena instâncias-chave de cada classe para preservar o desempenho em classificações já vistas. Em cada iteração de treinamento, o algoritmo retreina a rede usando tanto os exemplos novos quanto esses representantes armazenados, permitindo que o modelo mantenha a acurácia em tarefas anteriores. Este método de armazenamento eficiente se destaca pela forma econômica como armazena dados antigos.

Em resposta ao desafio de aprendizado incremental em larga escala, Wu et al. (2019) propuseram o método *Bias Correction* (BiC), que combina *replay* com um mecanismo explícito de correção de viés entre classes antigas e novas. Esse viés surge devido ao desbalanceamento entre os dados, já que as classes antigas são representadas por poucos exemplares armazenados, enquanto as novas possuem maior volume de dados. Seja n o número de classes já aprendidas nas tarefas anteriores e m o número de novas classes introduzidas na tarefa atual. O método utiliza uma função de custo composta por um termo de destilação de conhecimento e um termo de classificação:

$$L = \lambda L_d + (1 - \lambda) L_c \,, \tag{3.7}$$

em que L_d é definido por

$$\begin{aligned} L_d &= \sum_{x \in \hat{X}_n \cup X_m} \sum_{k=1}^{n} -\hat{\pi}_k(x) \log[\pi_k(x)], \\ \hat{\pi}_k(x) &= \frac{e^{\hat{o}_k(x)/T}}{\sum_{j=1}^{n} e^{\hat{o}_j(x)/T}}, \\ \pi_k(x) &= \frac{e^{o_k(x)/T}}{\sum_{j=1}^{n} e^{o_j(x)/T}}, \end{aligned} \tag{3.8}$$

e L_c é definido por

$$L_c = \sum_{(x,y)\in \hat{X}_n \cup X_m} \sum_{k=1}^{n+m} -\delta_{y=k} \log[p_k(x)] \,. \tag{3.9}$$

Segundo os autores Wu et al. (2019), na Equação 3.8, as amostras das novas classes são definidas como $X_m = \{(x_i, y_i), 1 \leq i \leq M,\, y_i \in [n+1, \ldots, n+m]\}$, onde M é o número de novas amostras, e x_i e y_i representam a imagem e o rótulo, respectivamente. Os exemplares selecionados das n classes antigas são denotados como $\hat{X}_n = \{(\hat{x}_j, \hat{y}_j), 1 \leq j \leq N_s,\, \hat{y}_j \in [1, \ldots, n]\}$, onde N_s é o número de imagens antigas selecionadas, com $\frac{N_s}{n} \approx \frac{M}{m}$. As saídas dos classificadores antigo e novo são denominadas como $\hat{o}_n(x) = [\hat{o}_1(x), \ldots, \hat{o}_n(x)]$ e $o_{n+m}(x) = [o_1(x), \ldots, o_n(x), o_{n+1}(x), \ldots, o_{n+m}(x)]$, respectivamente. Ao manter uma base limitada de exemplos antigos para relembrar o modelo, o método alcança um equilíbrio entre eficiência de memória e retenção de conhecimento. Os resultados demonstram que o método mantém a precisão mesmo em cenários com grandes quantidades de classes. Além disso, o método introduz uma camada de correção de viés aplicada às saídas do classificador, ajustando as probabilidades das novas classes para reduzir a tendência do modelo em favorecê-las. Essa estratégia permite melhorar o equilíbrio entre retenção e adaptação, embora ainda dependa da seleção de exemplares armazenados, o que pode limitar sua escalabilidade em cenários com muitas tarefas.

Gradient Episodic Memory (GEM), proposto por Lopez-Paz e Ranzato (2017), traz uma inovação ao introduzir a memória episódica de gradientes, que armazena amostras de tarefas anteriores e usa esses exemplos para restringir a atualização dos gradientes durante o treinamento de novas tarefas. Durante o treinamento de uma nova tarefa, o modelo não apenas busca minimizar a perda no novo conjunto de dados, mas também tenta preservar as representações que foram aprendidas anteriormente. Isso é alcançado através de um termo de regularização que penaliza as mudanças nas representações dos exemplos armazenados na memória. A principal contribuição do GEM é sua capacidade de restringir mudanças nos parâmetros do modelo, mantendo os gradientes alinhados com os de tarefas anteriores para evitar a interferência destrutiva.

Os autores Nguyen et al. (2018) propuseram uma abordagem que combina o *rehearsal* com métodos variacionais, denominado *Variational Continual Learning* (VCL). Neste algoritmo, as representações latentes de classes anteriores são armazenadas e utilizadas ao invés das próprias amostras, economizando espaço e melhorando a eficiência.

Os autores Tannugi, Britto e Koerich (2019) propuseram um estudo com o intuito de explorar o esquecimento catastrófico em cenários de FER. Os autores exploraram o uso da base de dados da tarefa anterior (*rehearsal*) no retreinamento para tarefas futuras, utilizando 50% ou 100% da base de dados da tarefa anterior. Os achados, segundo Tannugi, Britto e Koerich (2019), comprovam que o uso de todos os dados originais das tarefas anteriores auxilia para o não esquecimento destas tarefas e, em alguns casos, auxilia

também no retreinamento da tarefa atual, providenciando características importantes para ajudar a CNN no novo desafio.

Explorando o aprendizado contínuo não supervisionado, Rao et al. (2019) introduziram um método de *replay*, chamado *Continual Unsupervised Representation Learning* (CURL), que prioriza a preservação de representações latentes de dados não rotulados. Neste cenário, onde não há supervisão explícita para guiar a retenção de tarefas, o modelo depende de instâncias anteriores para conservar informações. Os testes nas bases de dados MNIST (DENG, 2012) e Omniglot (LAKE; SALAKHUTDINOV; TENENBAUM, 2015) indicam que o método é capaz de conservar a qualidade das representações enquanto aprende novas distribuições de dados, sendo especialmente útil em problemas com fluxo de dados contínuo.

Os autores Rolnick et al. (2019) propuseram variações no tamanho e seleção de amostras para um algoritmo de *experience replay*, similar ao *rehearsal*, mostrando que pequenas amostras selecionadas estrategicamente, com foco na diversidade dos dados, conseguem melhorar a retenção sem exigir armazenamento excessivo. A abordagem, chamada de *Continual Learning with Experience And Replay* (CLEAR), amplia a eficácia do *rehearsal* e diminui a necessidade de memória, tornando-se aplicável a redes de grande escala. Outros autores desenvolveram trabalhos similares, como o de Buzzega et al. (2020), que aplica uma técnica de regularização adicional no reuso de dados armazenados.

O método *Continual Learning with Imagination for Facial Expression Recognition* (CLIFER), proposto por Churamani e Gunes (2020), visa aprimorar a personalização e a adaptação contínua de sistemas FER. A abordagem combina: (i) um módulo de imaginação, baseado em um *Conditional Adversarial Auto-Encoder* (CAAE), que gera imagens sintéticas personalizadas para diferentes expressões preservando a identidade do usuário; e (ii) um sistema de memória dual inspirado na teoria do CLS, composto por redes *Growing When Required* (GWR) hierárquicas responsáveis por armazenar e generalizar representações afetivas. O uso de *pseudo-rehearsal* e das imagens geradas fortalece a retenção do conhecimento ao longo do tempo. O método foi avaliado em bases como RAVDESS, MMI e BAUM-1, e superou métodos tradicionais em *F1-score*, demonstrando potencial para aplicações em interações humano-robô.

Ao integrar memória episódica, o modelo proposto por Karam et al. (2023), tenta preservar o conhecimento anterior em uma memória auxiliar. Isso permite que a rede revisite o conhecimento passado, maximizando a retenção sem utilizar todos os dados originais das tarefas anteriores. O trabalho de Karam et al. (2023) está inserido no contexto de classificação de áudio, e os autores avaliaram o método nas bases de dados ESC-50 (PICZAK, 2015) e UrbanSound8K (SALAMON; JACOBY; BELLO, 2014). O método foi comparado com o EWC (KIRKPATRICK; AL., 2017) e iCarl (REBUFFI et al., 2017).

Os autores Mainsant et al. (2021) propuseram o algoritmo *Dream Net*, um método

de *pseudo-rehearsal* que utiliza uma rede auxiliar denominada *Memory Net*, a qual recebe como entrada um vetor de ruído aleatório. Por meio de um processo de auto-reinjeção de suas próprias saídas, essa rede é capaz de sintetizar amostras artificiais que incorporam tanto atributos visuais quanto rótulos correspondentes a experiências anteriores. Tais exemplos sintéticos são utilizados para reforçar o conhecimento previamente adquirido pela rede principal durante o aprendizado de novas tarefas, contribuindo assim para a mitigação do esquecimento catastrófico. O método foi avaliado na base de dados FER2013 (GOODFELLOW et al., 2013a), e os resultados demonstraram que o *Dream Net* é capaz de manter uma acurácia elevada em tarefas anteriores, mesmo quando exposto a novas classes, superando abordagens como o iCaRL (REBUFFI et al., 2017). Inspirados neste trabalho, os autores Antoni et al. (2023) propuseram uma adaptação do *Dream Net* como um sistema embarcado, que visa otimizar o desempenho em dispositivos com recursos limitados e com foco em privacidade de dados. O sistema embarcado é projetado para operar em ambientes com restrições de memória e processamento, mantendo a eficácia do aprendizado contínuo sem comprometer a privacidade dos dados (ANTONI et al., 2023).

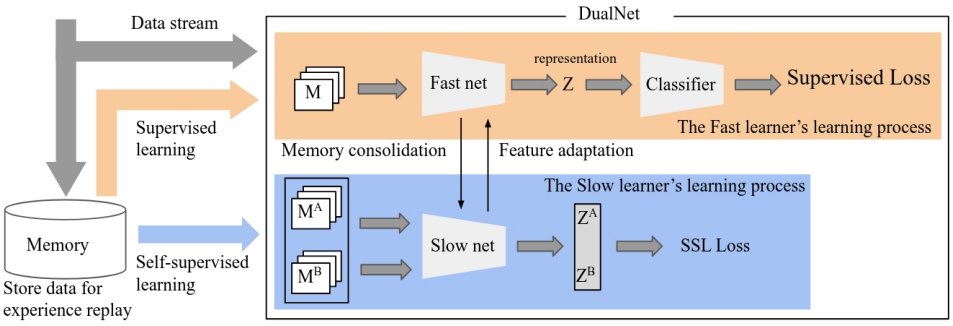
Os autores Wang et al. (2022b) propõem o *Contrastive Vision Transformer* (CVT), um modelo baseado em *transformers* projetado para aprendizado contínuo *online*. A arquitetura combina atenção externa com mecanismos contrastivos, utilizando focos aprendíveis por classe para reter representações anteriores e dois classificadores distintos: um *injection classifier* para dados novos e um *accumulation classifier* para integração com dados passados, ambos otimizados com funções de perda específicas. A perda total incorpora termos de classificação supervisionada e uma perda contrastiva, que reequilibra o aprendizado entre classes novas e antigas. Utilizando *buffer* de memória com amostragem, o modelo foi avaliado em CIFAR-100 (KRIZHEVSKY, 2009) e ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) em cenários *task-free* e *task-aware* e demonstrou resultados melhores em comparação com métodos concorrentes, mantendo uma arquitetura compacta de menos de 9.5M de parâmetros.

Os autores Stoychev, Churamani e Gunes (2023) propuseram o método *Latent Generative Replay* (LGR) com o objetivo de mitigar o esquecimento catastrófico em sistemas de reconhecimento de expressões faciais sob o paradigma de aprendizado contínuo. Diferentemente das abordagens tradicionais baseadas em *rehearsal* ou *pseudo-rehearsal*, que exigem o armazenamento ou a geração de amostras completas de tarefas anteriores, o LGR atua diretamente no espaço latente, sintetizando representações de baixa dimensionalidade das experiências passadas. O método foi incorporado a diversas estratégias consolidadas de aprendizado contínuo, substituindo a geração explícita de amostras pela reconstrução de vetores latentes. A abordagem foi avaliada nos conjuntos de dados CK+ (LUCEY et al., 2010), RAF-DB (LI; DENG; DU, 2017) e AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2019), segmentados em subconjuntos de duas classes sob o regime incremental de tarefas (*Task-IL*). Os resultados demonstraram que o LGR é capaz de preservar o

desempenho do modelo ao longo das tarefas, ao mesmo tempo em que reduz os requisitos de armazenamento e processamento, destacando-se como uma alternativa para retenção de conhecimento em ambientes com fluxo contínuo de dados.

Inspirando-se na teoria dos Sistemas de Aprendizagem Complementares (CLS) da neurociência, os autores Pham, Liu e Hoi (2024) propuseram o método *DualNets*, um sistema para aprendizado contínuo que combina dois sistemas de aprendizagem: um sistema rápido, voltado à aquisição supervisionada de representações específicas de tarefas, e um sistema lento, responsável pela aprendizagem de representações gerais e independentes de tarefa por meio de *Self-Supervised Learning*. A integração dessas representações complementares visa aprimorar a capacidade das redes neurais profundas de reter conhecimento ao longo de múltiplas tarefas. Avaliado em diferentes protocolos de aprendizado contínuo, incluindo cenários desafiadores como o ambiente *online* e sem informação de tarefa, o *DualNets* apresentou desempenho competitivo na base de dados miniImageNet, uma variação do ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A Figura 13 ilustra o *pipeline* do método proposto pelos autores.

Figura 13 – *Pipeline* do método *DualNets*, proposto por Pham, Liu e Hoi (2024).



Fonte: (PHAM; LIU; HOI, 2024).

3.3 Arquiteturas dinâmicas

Nesta seção são apresentados alguns trabalhos relacionados a categoria de arquiteturas dinâmicas (ver Seção 2.4 para mais detalhes sobre as categorias). Discutem-se os principais métodos de cada trabalho, bem como seus resultados e implicações.

Os autores Rusu et al. (2022) propuseram uma arquitetura que expande progressivamente a rede para cada nova tarefa, adicionando novas redes neurais (chamadas de colunas pelos autores), sem interferir nas redes já existentes. Segundo Rusu et al. (2022), a definição técnica do problema é dada por uma rede neural profunda com L camadas, cujas ativações ocultas são representadas por $h_i^{(1)} \in \mathbb{R}^{n_i}$, em que n_i é o número de unidades na camada $i \leq L$ e os parâmetros são dados por $\Theta^{(1)}$, treinados até a rede neural convergir. Ao receber uma segunda tarefa, os parâmetros $\Theta^{(1)}$ são congelados e uma nova coluna de

parâmetros $\Theta^{(2)}$ é instanciada (com inicialização aleatória), onde a camada $h_i^{(2)}$ recebe entradas tanto de $h_{i-1}^{(2)}$ quanto de $h_{i-1}^{(1)}$ por meio de conexões laterais. Isso se generaliza para K tarefas da seguinte forma

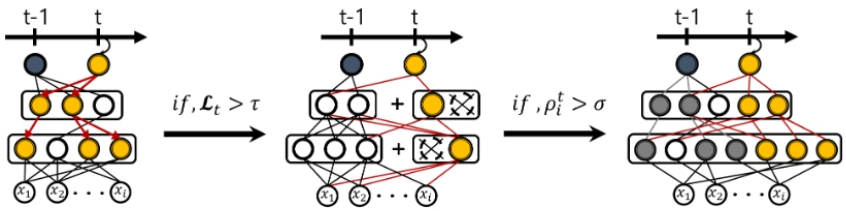
$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right) . \tag{3.10}$$

Essa estrutura permite que cada nova tarefa utilize tanto os parâmetros específicos de sua coluna quanto as ativações das tarefas anteriores, promovendo uma retenção de conhecimento por meio de conexões laterais entre as diferentes colunas de parâmetros (RUSU et al., 2022).

PathNet, proposto por Fernando et al. (2017), combina expansão dinâmica de redes com um mecanismo evolutivo para seleção de caminhos de ativação. A ideia central consiste em reutilizar subconjuntos de parâmetros (sub-redes) previamente treinados, evitando interferência entre tarefas ao congelar caminhos já otimizados. Para cada nova tarefa, diferentes caminhos são explorados e selecionados com base no desempenho, permitindo compartilhamento parcial de conhecimento e mitigação do esquecimento catastrófico. Os autores avaliaram o método em bases como MNIST (DENG, 2012), CIFAR (KRIZHEVSKY, 2009) e SVHN (NETZER et al., 2011).

No trabalho proposto por Yoon et al. (2018), a rede neural chamada *Dynamically Expandable Networks* (DEN), é expandida dinamicamente por meio da adição de neurônios ou camadas conforme necessário para acomodar novas tarefas. Um mecanismo de verificação identifica quais unidades são cruciais para o desempenho e previne a modificação das partes mais importantes da rede, enquanto, ao mesmo tempo, verifica a necessidade de expansão da rede neural, caso o treinamento não atinja um determinado *threshold* definido empiricamente (YOON et al., 2018).

Figura 14 – Fluxo de treinamento de uma rede que se expande e se adapta para novas tarefas.



Fonte: (YOON et al., 2018).

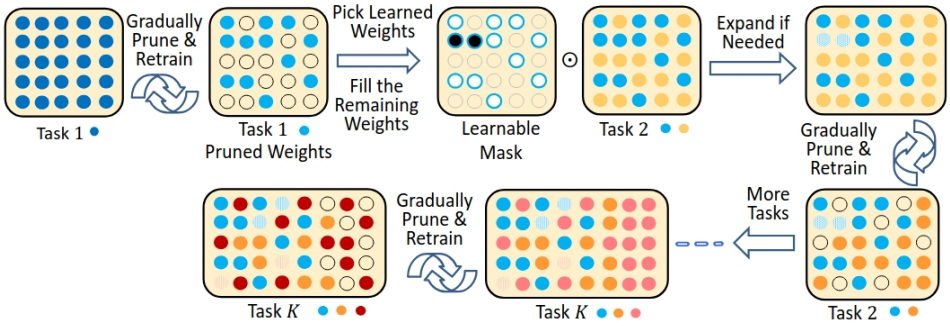
Na Figura 14 é possível observar o fluxo de treinamento da rede neural que se adapta a novas tarefas. O algoritmo proposto por Yoon et al. (2018) identifica os neurônios importantes para a nova tarefa e re-treina os parâmetros associados a eles. Após isso,

a expansão da rede é feita caso o retreinamento não consiga atingir um objetivo (*loss* > *threshold*, por exemplo). Na terceira etapa, a rede expandida é então reavaliada para identificar neurônios que mudaram muito dos seus valores originais. Esses neurônios são então duplicados (YOON et al., 2018). Os autores avaliaram o método DEN nas bases de dados MNIST (DENG, 2012), CIFAR-100 (KRIZHEVSKY, 2009) e AWA (LAMPERT; NICKISCH; HARMELING, 2009).

A arquitetura *Self-Net* (CAMP; MANDIVARAPU; ESTRADA, 2019) propõe uma abordagem baseada em compressão de modelos para aprendizado contínuo. Em vez de armazenar dados ou preservar explicitamente todos os parâmetros das tarefas anteriores, o método utiliza um *autoencoder* para aprender representações compactas dos pesos das redes treinadas para cada tarefa. Essas representações são então utilizadas para reconstruir os parâmetros quando necessário, permitindo reutilizar conhecimento prévio sem manter múltiplas redes completas em memória. Essa estratégia reduz significativamente o custo de armazenamento em comparação a abordagens como redes progressivas, porém introduz um erro de reconstrução que pode degradar o desempenho conforme o número de tarefas aumenta. O método foi avaliado em conjuntos como MNIST (DENG, 2012), CIFAR-10 (KRIZHEVSKY; NAIR; HINTON, 2009) e CIFAR-100 (KRIZHEVSKY, 2009), apresentando resultados competitivos em cenários controlados.

A abordagem CPG (*Compacting, Picking, and Growing*), proposta por Hung et al. (2019), usa uma técnica onde a rede compila parâmetros redundantes e ativa a expansão somente quando necessário. Os pesos importantes são mantidos para evitar o esquecimento de tarefas anteriores. Os pesos não redundantes ou que não são importantes são descartados, dando espaço para novos pesos. A Figura 15 ilustra o processo de adaptação de uma rede neural com o CPG. Os autores avaliaram o método nas bases de dados CIFAR-100 (KRIZHEVSKY, 2009), ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), CUBS (WAH et al., 2011), Stanford Cars (KRAUSE et al., 2013), Flowers (NILSBACK; ZISSERMAN, 2008), Wikiart (SALEH; ELGAMMAL, 2016) e Sketch (EITZ; HAYS; ALEXA, 2012).

Figura 15 – Processo de adaptação da rede neural no método CPG.

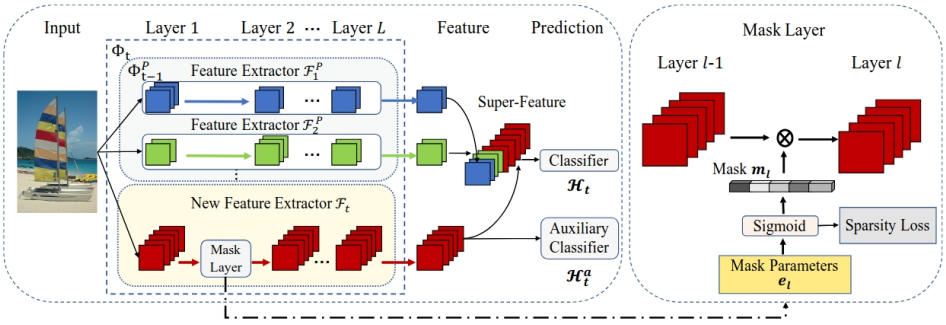


Fonte: (HUNG et al., 2019).

Parameter Superposition (PSP), trabalho proposto por Cheung et al. (2019), é uma técnica que permite armazenar vários modelos em um único conjunto de parâmetros. Isso é feito usando uma transformação linear que mapeia os parâmetros de cada modelo para um espaço de características comuns. Os parâmetros de todos os modelos são então somados para formar um único conjunto de parâmetros. Os modelos individuais podem ser recuperados do conjunto de parâmetros usando uma transformação linear inversa. Os autores Cheung et al. (2019) realizaram testes nas bases de dados MNIST (DENG, 2012), CIFAR-10 (KRIZHEVSKY; NAIR; HINTON, 2009), CIFAR-100 (KRIZHEVSKY, 2009) e fashionMNIST (XIAO; RASUL; VOLLGRAF, 2017), comparando com os métodos EWC (KIRKPATRICK; AL., 2017) e SI (ZENKE; POOLE; GANGULI, 2017). O método superou os métodos aos quais foi comparado (CHEUNG et al., 2019).

Os autores Yan, Xie e He (2021) criaram o DER (*Dynamically Expandable Representation*), que combina expansão dinâmica com *replay* seletivo, armazenando apenas as amostras mais representativas e expandindo a rede conforme a complexidade aumenta. A Figura 16 traz uma visão geral do método. Segundo Yan, Xie e He (2021), o DER funciona expandindo gradualmente a representação interna do modelo à medida que novas classes são introduzidas. Essa expansão ocorre de forma dinâmica, adicionando novas dimensões à representação apenas quando necessário. Além disso, o método utiliza uma estratégia de aprendizado em duas etapas: primeiro, as classes que já foram aprendidas são protegidas para evitar o esquecimento; em seguida, um novo extrator de características é treinado para as novas classes, garantindo que as novas informações sejam bem representadas. Os autores avaliaram o método em duas bases de dados: CIFAR-100 (KRIZHEVSKY, 2009) e ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Figura 16 – Visão geral do método *Dynamically Expandable Representation*.



Fonte: (YAN; XIE; HE, 2021).

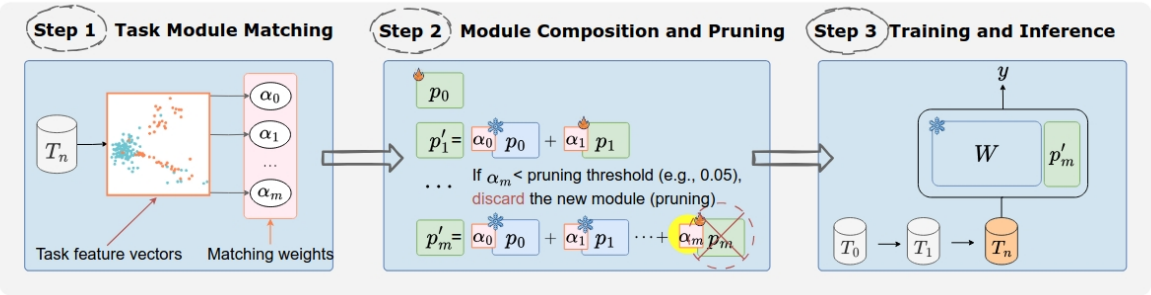
O trabalho de Douillard et al. (2022) propõe o *Transformers for Continual Learning with Dynamic Token Expansion* (DyTox), uma arquitetura baseada em *Vision Transformer* projetada para aprendizado contínuo com escalabilidade a múltiplas tarefas. A abordagem utiliza codificadores compartilhados entre tarefas (*self-attention blocks*) e um decodificador especializado (*task-attention block*), onde cada nova tarefa recebe um *token* aprendível

que direciona a saída a um classificador específico. A perda total combina classificação supervisionada, distilação de conhecimento e divergência inter-*token*. Além da expansão de *tokens*, o método adota técnicas de *rehearsal* com memória limitada e regularização. Avaliado nos conjuntos CIFAR-100 (KRIZHEVSKY, 2009) e ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), DyTox superou métodos como iCaRL (REBUFFI et al., 2017) e DER Yan, Xie e He (2021) na base de dados CIFAR-100 e desempenho de estado da arte com número de parâmetros significativamente menor no ImageNet.

Cai e Rostami (2024) propuseram uma implementação de uma arquitetura de atenção dinâmica que expande seletivamente camadas de atenção, inspirada no trabalho de Vaswani et al. (2017). A adaptação aos *transformers* fornece ao modelo a capacidade de preservar conhecimento com expansões mínimas. O trabalho dos autores, chamado de *Task Attentive Multimodal Continual Learning* (TAM-CL), utiliza tanto os modais de visão quanto de linguagem, denominados de *Vision-and-Language* (VaL) (CAI; ROSTAMI, 2024). Segundo Cai e Rostami (2024), essa implementação permite que o método seja utilizado em diversos cenários, pois se trata de uma abordagem multimodal.

O método de *continual learning* de Wang et al. (2024b), *Module Composition and Pruning for Continual Learning* (MoCL-P), utiliza a otimização de parâmetros durante treinamento para garantir que os parâmetros já treinados para uma tarefa anterior sejam mantidos intactos ao retreinar para novas tarefas. O fluxo geral do método pode ser observado na Figura 17.

Figura 17 – MoCL-P, método proposto por Wang et al. (2024b).



Fonte: (YAN; XIE; HE, 2021).

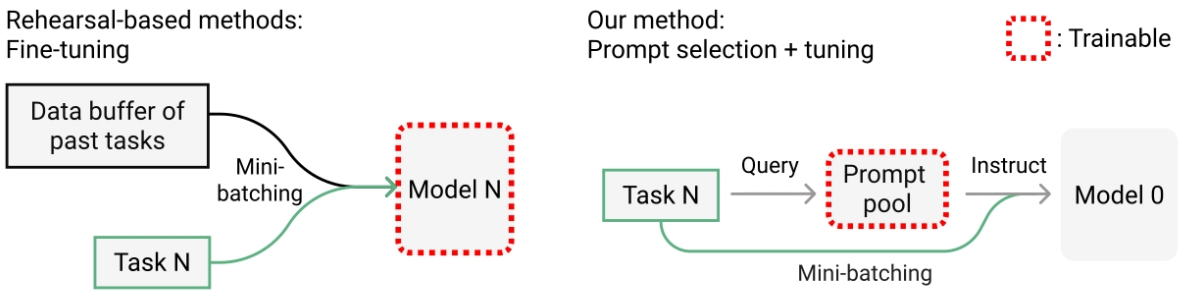
3.4 Modelos Pré-treinados

Nesta seção são apresentados trabalhos relacionados à categoria de métodos baseados em modelos pré-treinados (ver Seção 2.4 para mais detalhes sobre as categorias). São discutidos os principais métodos, resultados e implicações de cada trabalho.

Os autores Wang et al. (2022b) introduziram o método *Learning to Prompt* (L2P), sendo o primeiro a aplicar técnicas de *prompting* ao aprendizado contínuo utilizando

modelos pré-treinados (PTMs) como o ViT. A Figura 18 mostra a visão geral do método. Em vez de ajustar os pesos do modelo, o L2P mantém o PTM congelado e aprende um *pool* de *prompts* (parâmetros treináveis compactos) que atuam como instruções dinâmicas para o modelo. O método utiliza um mecanismo de consulta baseado em pares chave-valor: para cada entrada, uma função de consulta busca os *prompts* mais relevantes no *pool*, permitindo o compartilhamento de conhecimento entre tarefas similares e mantendo a plasticidade para novos conhecimentos. O L2P foi avaliado em cenários de *class-incremental* (*Class-IL*), *domain-incremental* e *task-agnostic*, sem uso de memória explícita, utilizando os *benchmarks* Split CIFAR-100, 5-datasets e CORe50. Os resultados mostraram que o L2P superou métodos baseados em regularização e alcançou resultados competitivos com métodos baseados em *rehearsal* sem a necessidade de um *buffer* de dados.

Figura 18 – Visão geral do método L2P, proposto por Wang et al. (2022b).



Fonte: (WANG et al., 2022b).

Os autores Wang et al. (2022a) propuseram o método *DualPrompt*, uma evolução do *Learning to Prompt* (L2P) que utiliza *prompts* complementares para gerenciar o conhecimento de forma mais eficiente. Inspirado na teoria de sistemas de aprendizagem complementares (*Complementary Learning Systems*, CLS), o método desacopla o espaço de *prompts* em dois tipos: *G-Prompts* (gerais), responsáveis por capturar conhecimentos invariantes às tarefas e inseridos nas camadas iniciais do modelo, e *E-Prompts* (especialistas), que aprendem instruções específicas de cada tarefa e são incorporados em camadas mais profundas. Essa estratégia permite a aprendizagem de características gerais compartilhadas enquanto isola conhecimentos específicos, mitigando o esquecimento catastrófico. O *DualPrompt* foi validado no cenário *Class-IL*, utilizando os *benchmarks* Split CIFAR-100, 5-datasets e o então introduzido Split ImageNet-R. Os autores reportaram que o método estabelece um novo estado da arte em cenários sem repetição de dados, superando o L2P e superando métodos baseados em *rehearsal* que utilizam *buffers* de dados.

Os autores Zhang et al. (2023) apresentaram o método *Slow Learner with Classifier Alignment* (SLCA), que propõe uma abordagem de ajuste fino seletivo para modelos pré-treinados. O trabalho identifica o sobreajuste progressivo como um dos principais desafios no uso de PTMs em aprendizado contínuo, no qual a representação do modelo se adapta

excessivamente às tarefas incrementais, comprometendo sua capacidade de generalização. Para mitigar esse problema, o SLCA introduz o componente *Slow Learner*, que aplica uma taxa de aprendizado significativamente reduzida na camada de representação e uma taxa mais elevada na camada de classificação. Adicionalmente, o método incorpora o *Classifier Alignment* (CA), uma estratégia *post-hoc* que modela as distribuições das classes por meio de estatísticas, como média e covariância, alinhando as previsões do classificador para equilibrar o desempenho entre tarefas antigas e novas. O SLCA foi avaliado no cenário *Class-IL* utilizando os *benchmarks* Split CIFAR-100, Split ImageNet-R e bases de granularidade fina como Split CUB-200 e Split Cars-196. Segundo os autores, o SLCA superou significativamente os métodos baseados em *prompts*, reduzindo a lacuna de desempenho em relação ao limite superior do treinamento conjunto (*joint training*) para menos de 2% em certos casos.

Zhou et al. (2023) conduziram um estudo crítico sobre o aprendizado incremental de classes com modelos pré-treinados, destacando a generalização e a adaptatividade como pilares fundamentais para o desempenho nesse contexto. O trabalho propõe o método *Adaptive Model* (ADAM), que combina a capacidade de generalização de um PTM congelado com a adaptatividade de um modelo ajustado especificamente na primeira tarefa incremental. Além disso, os autores introduzem o *SimpleCIL*, uma abordagem simplificada baseada no uso de um extrator de características congelado e um classificador prototípico para predição. Estas abordagens foram testadas sob o protocolo *Class-IL* em sete bases de dados: CIFAR-100, ImageNet-R, ImageNet-A, CUB-200, ObjectNet, OmniBenchmark e VTAB. Os achados indicaram que o uso de classificadores em modelos pré-treinados robustos pode superar abordagens de *prompting* muito mais complexas.

Os autores Wang et al. (2023) introduziram o método *HiDe-Prompt*, que aborda a subotimalidade de abordagens baseadas em *prompts* quando o pré-treinamento é realizado de forma autossupervisionada. O método propõe a decomposição hierárquica do objetivo de aprendizado contínuo em três componentes: predição dentro da tarefa (*Within-Task Prediction*, WTP), inferência da identidade da tarefa (*Task Identity Inference*, TII) e predição adaptativa à tarefa (*Task-Adaptive Prediction*, TAP). Esses componentes são otimizados explicitamente por meio de *prompts* específicos por tarefa e do uso de estatísticas de representações instruídas e não instruídas. Além disso, é empregada uma estratégia de regularização contrastiva para coordenar esses componentes, assegurando que as representações de diferentes tarefas sejam simultaneamente distinguíveis e compatíveis. O *HiDe-Prompt* foi avaliado em cenários *Class-IL* sob diversos paradigmas de pré-treino nos *benchmarks* Split CIFAR-100, Split ImageNet-R, 5-Datasets e Split CUB-200. Os autores reportaram um avanço nos resultados do estado da arte, com ganhos de até 15,01% sobre competidores como o CODA-Prompt em cenários autossupervisionados.

Mais recentemente, Zhou et al. (2024) propuseram o método *Expandable Subspace*

Ensemble (EASE), voltado à mitigação do dilema estabilidade-plasticidade sem o uso de exemplares. O EASE constrói subespaços específicos por tarefa por meio da adição de adaptadores leves (*adapters*) acoplados a um modelo pré-treinado congelado, garantindo que o aprendizado de novas classes não interfira no conhecimento previamente adquirido. Para lidar com a incompatibilidade entre classificadores de classes antigas e novos subespaços, o método introduz uma estratégia de complementação de protótipos guiada semanticamente. Essa abordagem utiliza similaridade entre classes em um espaço de coocorrência para sintetizar representações de classes antigas no novo subespaço, possibilitando decisões conjuntas sem a necessidade de armazenamento de dados anteriores. O método foi validado em sete *benchmarks* (incluindo CIFAR-100, ImageNet-R, ImageNet-A, ObjectNet e VTAB) sob o protocolo *Class-IL*. Os experimentos demonstraram que o EASE atinge o estado da arte, superando o ADAM e o CODA-Prompt, mantendo um custo de parâmetros similar aos métodos de *prompting* convencionais.

De forma geral, os métodos baseados em modelos pré-treinados apresentam desempenho superior aos métodos clássicos de aprendizado contínuo, especialmente em cenários *class-incremental*, devido à forte capacidade de generalização dos modelos de base. Observa-se que abordagens baseadas em *prompting* e *adapters* conseguem mitigar o esquecimento catastrófico sem a necessidade de armazenamento de dados, alcançando resultados próximos ao treinamento conjunto em diversos *benchmarks*. No entanto, esses métodos são predominantemente avaliados em bases de dados genéricas, como CIFAR-100 e ImageNet, sendo ainda pouco explorados em domínios específicos como o reconhecimento de expressões faciais. Além disso, a dependência de modelos pré-treinados de grande escala pode limitar sua aplicação em cenários com restrições computacionais ou com menor disponibilidade de dados. Esses aspectos evidenciam uma lacuna na literatura quanto ao uso de abordagens eficientes e adaptadas a domínios específicos, como o FER, especialmente quando se considera a integração com métodos baseados em *replay* generativo.

3.5 Considerações finais

Observa-se um crescimento significativo no uso de redes neurais profundas, especialmente CNNs, para mitigar o problema do esquecimento catastrófico. Em paralelo, modelos generativos, como GANs, têm sido amplamente explorados em abordagens de *pseudo-rehearsal*, devido à sua capacidade de aproximar a distribuição de dados de tarefas anteriores. No entanto, esses modelos ainda apresentam elevado custo computacional, o que limita sua aplicação em cenários mais restritos.

De forma geral, a literatura recente evidencia uma convergência para abordagens híbridas, que combinam estratégias de regularização, *replay* e arquiteturas dinâmicas. Essas combinações buscam equilibrar o dilema estabilidade-plasticidade, frequentemente

inspirando-se em mecanismos biológicos, como o uso de memórias auxiliares para retenção de conhecimento.

Em relação aos protocolos experimentais, observa-se que a maioria dos trabalhos adota cenários *task-incremental* ou *class-incremental*, com forte predominância de bases como MNIST e CIFAR-100. A base MNIST, em particular, destaca-se pela simplicidade estrutural e pela facilidade de adaptação a diferentes configurações experimentais, sendo amplamente utilizada como *benchmark* inicial para validação de métodos. Por outro lado, CIFAR-100 e ImageNet são frequentemente empregadas para avaliar a escalabilidade e robustez das abordagens propostas.

No que se refere aos resultados reportados, métodos clássicos baseados exclusivamente em regularização tendem a apresentar ganhos limitados quando comparados a abordagens baseadas em *replay*, especialmente aquelas que utilizam dados reais ou gerados. Trabalhos recentes baseados em modelos pré-treinados e *prompting* demonstram desempenho próximo ao treinamento conjunto (quando utilizam-se os dados reais de tarefas passadas para treinar em conjunto com as novas tarefas) em cenários padronizados, embora ainda sejam pouco explorados em domínios específicos.

Apesar dos avanços, observa-se uma limitação recorrente na literatura: a ausência de avaliações em domínios mais complexos e específicos, como o reconhecimento de expressões faciais. A maioria dos métodos é validada em bases genéricas, o que dificulta a análise de sua eficácia em cenários com maior variabilidade semântica e sensibilidade a ruídos. Além disso, diferenças nos protocolos experimentais e nas métricas utilizadas tornam a comparação direta entre trabalhos não trivial.

Nesse contexto, a Tabela 2 sintetiza os principais métodos analisados, destacando suas categorias, bases de dados e configurações experimentais. Essa análise evidencia tanto a diversidade de abordagens propostas quanto a necessidade de investigações mais direcionadas a cenários específicos, motivando o desenvolvimento do método apresentado neste trabalho. Na Tabela 3 são apresentados os resultados reportados por Ven, Tuytelaars e Tolias (2022) para a base de dados MNIST. Observa-se que métodos baseados em *replay* tendem a apresentar desempenho superior em cenários mais desafiadores, especialmente no cenário *Class-IL*.

Tabela 2 – Levantamento dos trabalhos na área de *continual learning* e que apresentam um método para mitigar o esquecimento catastrófico. São apresentados os nomes dos métodos, o ano em que foi publicado, a qual categoria pertence e quais foram as bases de dados utilizadas.

Método	**Nome Completo**	**Ano**	**Cenário**	**Categoria**	**Bases de dados**
EWC	*Elastic Weight Consolidation* (KIRKPATRICK; AL., 2017)	2017	*Task-IL*	Regularização	MNIST, Atari 2600
LWF	*Learning Without Forgetting* (LI; HOIEM, 2018)	2018	*Class-IL*	Regularização	ImageNet, VOC, CUB, MNIST, Places365, Scenes
SI	*Synaptic Intelligence* (ZENKE; POOLE; GANGULI, 2017)	2017	*Task-IL*	Regularização	MNIST, CIFAR-10, CIFAR-100
MAS	*Memory Aware Synapses* (ALJUNDI et al., 2018)	2018	*Task-IL*	Regularização	-
IMM	*Incremental Moment Matching* (LEE et al., 2018)	2018	*Task-IL*	Regularização	MNIST, CIFAR-10, Caltech-UCSD-Birds, Lifelog
RW	*Riemannian Walk* (CHAUDHRY et al., 2018)	2018	*Task-IL*	Regularização	MNIST, CIFAR-100
CLAW	*Continual Learning with Adaptive Weights* (ADEL; ZHAO; TURNER, 2020)	2020	*Task-IL*	Regularização	MNIST, fashionMNIST, Omniglot, CIFAR-100

MER	*Meta-Experience Replay* (RIEMER et al., 2019)	2019	*Task-IL*/*Class-IL*/*Online*	Regularização, *Rehearsal*	MNIST
GR	*Generative Replay* (SHIN et al., 2017)	2017	*Task-IL*	*Rehearsal*, *Pseudo-Rehearsal*	MNIST
FearNet	*FearNet* (KEMKER; KANAN, 2017)	2017	*Class-IL*	*Rehearsal*, *Pseudo-Rehearsal*	CIFAR-100, CUB, AudioSet
iCaRL	*Incremental Classifier and Representation Learning* (REBUFFI et al., 2017)	2017	*Class-IL*	*Pseudo-Rehearsal*	CIFAR-100, ImageNet
BiC	*Bias Correction* (WU et al., 2019)	2019	*Class-IL*	Regularização, *Rehearsal*	ImageNet, CIFAR-100, MS-Celeb-1M
CURL	*Continual Unsupervised Representation Learning* (RAO et al., 2019)	2019	*Class-IL*	*Pseudo-Rehearsal*, Arquitetura dinâmica	MNIST, Omniglot
GEM	*Gradient Episodic Memory* (LOPEZ-PAZ; RANZATO, 2017)	2017	*Task-IL*	*Pseudo-Rehearsal*	MNIST, CIFAR-100
CLEAR	*Continual Learning with Experience And Replay* (ROLNICK et al., 2019)	2019	*Online*	*Pseudo-Rehearsal*	-
VCL	*Variational Continual Learning* (NGUYEN et al., 2018)	2018	*Task-IL*/*Class-IL*	*Pseudo-Rehearsal*	MNIST

DualNet	*DualNet* (PHAM; LIU; HOI, 2024)	2024	*Class-IL/Online*	*Pseudo-Rehearsal*, Arquitetura dinâmica	CORE50, ImageNet
Prog. Networks	*Progressive Networks* (RUSU et al., 2022)	2022	*Task-IL*	Arquitetura dinâmica	Labyrinth, Atari, Pong
PathNet	*PathNet* (FERNANDO et al., 2017)	2017	*Task-IL*	Arquitetura dinâmica	MNIST, CIFAR, SVHN, Atari, Labyrinth
DEN	*Dynamically Expandable Network* (YOON et al., 2018)	2018	*Task-IL*	Regularização, Arquitetura dinâmica	MNIST, CIFAR-100, AWA
Self-Net	S*elf-Net: Lifelong Learning via Continual Self-Modeling* (CAMP; MANDIVARAPU; ESTRADA, 2019)	2019	*Task-IL*	Arquitetura dinâmica, *Pseudo-Rehearsal*	MNIST, CIFAR-10, CIFAR-100, Atari 2600
CPG	*Compacting, Picking, and Growing* (HUNG et al., 2019)	2019	*Task-IL*	Arquitetura dinâmica	CIFAR-100, ImageNet, CUBS, Stanford Cars, Flowers, Wikiart, Sketch
PSP	*Parameter Superposition* (CHEUNG et al., 2019)	2019	*Task-IL*	Regularização, Arquitetura dinâmica	MNIST, CIFAR-10, CIFAR-100
DER	*Dynamically Expandable Representation* (YAN; XIE; HE, 2021)	2021	*Class-IL/Online*	Regularização, Arquitetura dinâmica	ImageNet, CIFAR-100

EM	*Episodic Memory* (KARAM et al., 2023)	2023	*Class-IL*/*Online*	*Rehearsal*	ESC-50, UrbanSound8K
TAM-CL	*Task Attentive Multimodal Continual Learning* (CAI; ROSTAMI, 2024)	2024	Multimodal	Regularização, Arquitetura dinâmica	SNLI-VE, COCOQA, GQA, NLVR2, OKVQA
MoCL-P	*Module Composition and Pruning for Continual Learning* (WANG et al., 2024b)	2024	*Task-IL*	Regularização, Arquitetura dinâmica	MTL15, AfriSenti
DyTox	*Transformers for Continual Learning with Dynamic Token Expansion* (DOUILLARD et al., 2022)	2022	*Class-IL*	Arquiteturas Dinâmicas, *Replay*, Regularização	CIFAR-100, ImageNet
ADA	*Adaptive Distillation of Adapters* (ERMIS et al., 2022)	2022	*Class-IL*	Arquiteturas Dinâmicas, Regularização, *Replay*	CIFAR-100, ImageNet
CVT	*Contrastive Vision Transformer* (WANG et al., 2022b)	2022b	*Online*	*Replay*, Regularização	CIFAR-100, ImageNet
GCAB + CFDC	*Gated Class-Attention and Cascaded Feature Drift Compensation* (COTOGNI et al., 2025)	2025	*Class-IL*	Regularização, Arquiteturas Dinâmicas	CIFAR-100, ImageNet

L2P	*Learning to Prompt* (WANG et al., 2022b)	2022	*Class-IL*	Baseado em Modelos Pré-treinados	CIFAR-100, CORe50, 5-Datasets
DualPrompt	*DualPrompt* (WANG et al., 2022a)	2022	*Class-IL*	Baseado em Modelos Pré-treinados	CIFAR-100, ImageNet-R, 5-Datasets
SLCA	*Slow Learner with Classifier Alignment* (ZHANG et al., 2023)	2023	*Class-IL*	Regularização (*fine-tuning* seletivo), Baseado em Modelos Pré-treinados	CIFAR-100, ImageNet-R, CUB-200, Cars-196
ADAM + SimpleCIL	*Adaptive Model + Simple Class Incremental Learning* (ZHOU et al., 2023)	2023	*Class-IL*	Regularização, Baseado em Modelos Pré-treinados	CIFAR-100, ImageNet-R, ImageNet-A, CUB-200, VTAB
HiDe-Prompt	*Hierarchical Decomposed Prompting* (WANG et al., 2023)	2023	*Class-IL*	Baseados em Modelos Pré-treinados	CIFAR-100, ImageNet-R, CUB-200, 5-Datasets
EASE	*Expandable Subspace Ensemble* (ZHOU et al., 2024)	2024	*Class-IL*	Baseados em Modelos Pré-treinados	CIFAR-100, ImageNet-R, ImageNet-A, ObjectNet, VTAB

Fonte: autoria própria.

Tabela 3 – Resultados de métodos de aprendizado contínuo para os cenários *Task-IL*, *Domain-IL* e *Class-IL* na Split-MNIST, conforme reportado por Ven, Tuytelaars e Tolias (2022).

Método	***Task-IL***	***Domain-IL***	***Class-IL***
None – limite inferior	84,32 (±0,99)	60,13 (±1,66)	19,89 (±0,02)
Joint – limite superior	99,67 (±0,03)	98,59 (±0,05)	98,17 (±0,04)
EWC	99,06 (±0,15)	63,03 (±1,58)	20,64 (±0,52)
SI	99,20 (±0,11)	66,94 (±1,13)	21,20 (±0,57)
LwF	99,60 (±0,03)	71,18 (±1,42)	21,89 (±0,32)
FROMP	99,12 (±0,13)	84,86 (±1,02)	77,38 (±0,64)
DGR	99,50 (±0,03)	95,57 (±0,30)	90,35 (±0,24)
BI-R	99,61 (±0,03)	97,26 (±0,15)	94,41 (±0,15)
ER	98,98 (±0,07)	93,75 (±0,24)	88,79 (±0,20)
A-GEM	98,54 (±0,10)	87,67 (±1,33)	65,10 (±3,64)
Generative Classifier	–	–	93,82 (±0,06)
iCaRL	–	–	92,49 (±0,12)

Fonte: adaptado de (VEN; TUYTELAARS; TOLIAS, 2022).

4 Método proposto

Neste capítulo são apresentados os métodos propostos para mitigar o problema de *catastrophic forgetting*. O estudo é conduzido no contexto do reconhecimento de expressões faciais, um domínio conhecido por sua complexidade no treinamento de redes neurais profundas. Bases de dados nesse cenário apresentam grande variabilidade entre indivíduos, incluindo diferenças de idade, sexo, etnia e outras características faciais, além de variações na forma como as emoções são expressas. Essa variabilidade torna o reconhecimento de expressões faciais um ambiente particularmente desafiador para investigação do esquecimento catastrófico, pois exige que os modelos preservem conhecimento previamente adquirido ao mesmo tempo em que aprendem novas informações. Nesse contexto, este trabalho propõe um método baseado na geração de imagens sintéticas para auxiliar na retenção de conhecimento ao longo das tarefas de aprendizado contínuo.

Para evitar ambiguidades terminológicas ao longo deste capítulo, adotam-se as seguintes definições padronizadas: denomina-se **tarefa anterior** aquela cuja base de dados já foi submetida à CNN que está sendo treinada de forma contínua, e da qual se deseja preservar o conhecimento; a **tarefa atual** refere-se à nova tarefa apresentada à CNN durante o aprendizado contínuo; e a **tarefa futura** corresponde àquela que sucederá a tarefa atual, sendo mencionada apenas de forma genérica em descrições do fluxo incremental. Além disso, utiliza-se o termo **imagens originais** (ou **dados originais**) para designar as imagens reais das bases de dados, e **imagens sintéticas** para aquelas geradas pelo modelo generativo (WGAN-GP), que representam as tarefas anteriores.

No cenário considerado neste trabalho, assume-se que os conjuntos de dados utilizados em tarefas anteriores não estão disponíveis durante o treinamento de tarefas subsequentes, conforme discutido na Seção 1.1. Essa hipótese reflete restrições frequentemente presentes em aplicações reais de reconhecimento de emoções, nas quais questões de privacidade, consentimento ou limitações de armazenamento impedem a reutilização direta de dados biométricos.

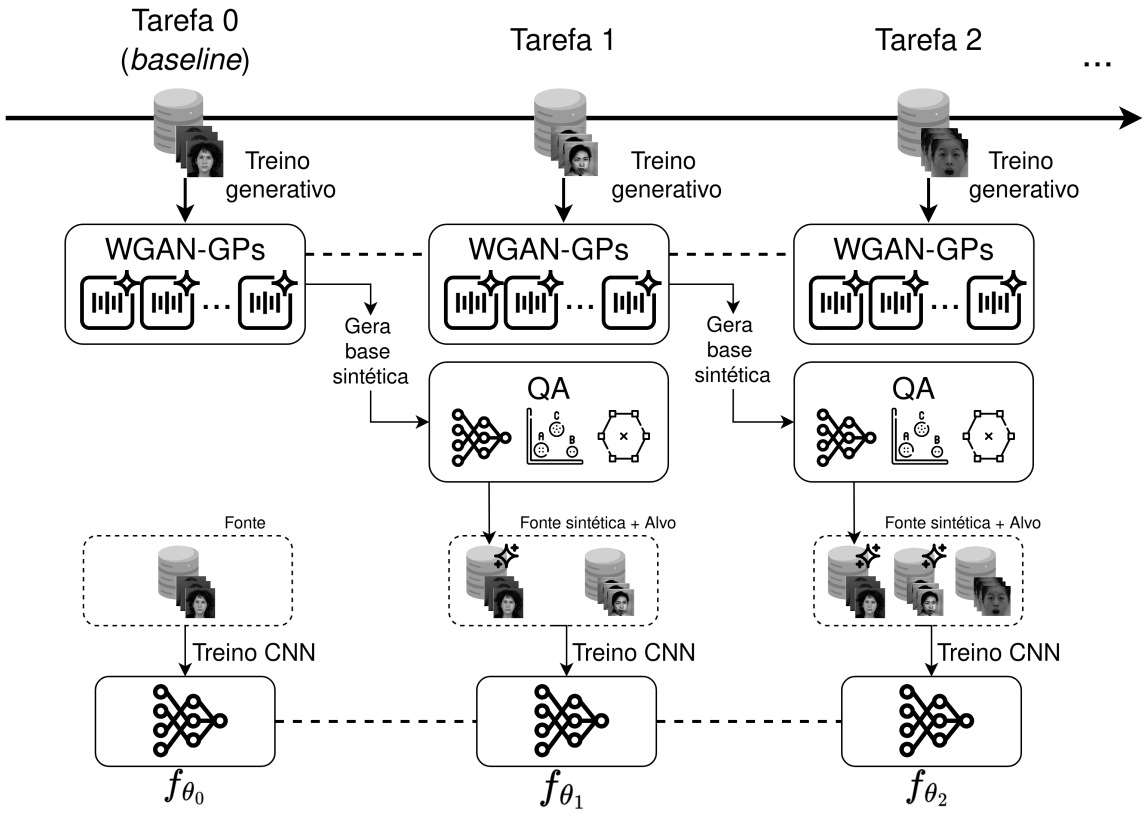
Dessa forma, o modelo não possui acesso às imagens originais de tarefas passadas e deve recorrer a amostras sintéticas para preservar o conhecimento previamente adquirido. Esse cenário motiva a adoção de uma estratégia baseada em *generative replay*, em que um modelo generativo (neste trabalho, uma WGAN-GP) é utilizado para aproximar distribuições representativas das tarefas anteriores a partir do conhecimento aprendido durante o treinamento.

A fim de classificar o método proposto perante à categorização vista na literatura, é possível defini-lo como a junção das categorias de regularização e *replay*. Mais detalha-

damente, o método de geração de imagens pode ser considerado como *pseudo-rehearsal*, afinal, não se utilizam as imagens originais para o retreinamento de tarefas anteriores.

Formaliza-se a metodologia usando representações algorítmicas para fornecer uma compreensão mais concreta dos conceitos teóricos apresentados. Na Seção 4.4, fornece-se um algoritmo detalhado sobre o método proposto, *Emotion-Centered generative replay* (ECgr).

Figura 19 – Visão geral do método proposto.



Fonte: autoria própria.

A Figura 19 apresenta a visão geral do método proposto. Inicialmente, um conjunto de WGAN-GPs é treinado separadamente para cada classe do conjunto de dados da tarefa anterior. Com esses geradores, são produzidas imagens sintéticas específicas para cada classe, resultando em uma base de dados sintética diversificada. Em seguida, aplica-se um processo de *Quality Assurance* (QA) das imagens sintéticas para selecionar apenas as imagens mais representativas e confiáveis. Três estratégias de QA são utilizadas de forma independente: (i) uma CNN supervisionada treinada na tarefa anterior, que retém imagens cuja predição é correta e com alta confiança; (ii) um método de agrupamento não supervisionado, que utiliza o coeficiente de silhueta para avaliar a coerência estrutural das amostras no espaço de *features*; e (iii) uma abordagem baseada em espaço latente, onde apenas imagens dentro do *convex hull* de cada classe no espaço da WGAN-GP são mantidas. As imagens selecionadas pelo QA compõem a base de dados sintética final, usada durante o

retreinamento. Nesta fase, as imagens reais da tarefa atual e as imagens sintéticas da tarefa anterior são combinadas, e uma função de custo ponderada é empregada para reduzir o impacto de amostras sintéticas potencialmente ruidosas, atribuindo a cada imagem um peso proporcional à sua qualidade. Esse fluxo permite preservar o conhecimento adquirido anteriormente sem acesso aos dados originais das tarefas anteriores, ao mesmo tempo em que adapta o modelo à nova tarefa, mitigando o esquecimento catastrófico de forma eficaz.

4.1 Método generativo e aprendizado contínuo

Para mitigar o problema de *catastrophic forgetting* em aprendizado contínuo, propomos um método baseado na geração controlada de dados sintéticos com WGAN-GPs (*Wasserstein* GANs com penalização de gradiente) por classe, seguido de um processo de retreinamento incremental de classificadores convolucionais. O método é estruturado em duas etapas principais: (i) o treinamento supervisionado de geradores por classe, e (ii) o aprendizado incremental com dados reais e sintéticos. Considera-se um conjunto supervisionado de tarefas $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$, cada uma composta pelas mesmas C classes. No contexto deste trabalho, considera-se o conjunto de sete expressões faciais básicas: alegria, tristeza, medo, surpresa, raiva, desgosto e neutro. Todas as bases de dados utilizadas nos experimentos foram selecionadas por compartilharem esse mesmo conjunto de classes, permitindo a condução dos experimentos de aprendizado contínuo de forma consistente, sendo assim, todos os conjuntos de dados incluem rótulos $y \in \{1, \ldots, C\}$. O Algoritmo 1 apresenta o método proposto que combina a geração de dados sintéticos com WGAN-GPs e o processo de QA para filtrar as imagens geradas, seguido de um retreinamento incremental de classificadores.

Na primeira etapa, para cada tarefa $\mathcal{D}_k$, um conjunto de geradores $\{G_k^{(1)}, G_k^{(2)}, \ldots, G_k^{(C)}\}$ é treinado, onde cada $G_k^{(c)}$ aprende a gerar amostras da classe c em $\mathcal{D}_k$. A função objetivo usada para o treinamento das WGAN-GPs é:

$$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}}[D(G(\mathbf{z}))] + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} \left[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2 \right] \tag{4.1}$$

onde D é o discriminador, G é o gerador, λ é o peso do termo de regularização, e $\mathbb{P}_{\hat{\mathbf{x}}}$ é a distribuição interpolada entre dados reais e gerados. A otimização ocorre via algoritmo Adam até a convergência da função crítica (ou até um número máximo de iterações).

Na etapa de aprendizado incremental, a cada nova tarefa $\mathcal{D}_k$, os geradores $\{G_j^{(c)}\}_{j=1}^{k-1}$ das tarefas anteriores são utilizados para gerar dados sintéticos $\mathcal{D}'_j$. Esses dados são então combinados com os dados reais da nova tarefa, formando o conjunto unificado $\mathcal{D}_k^{\text{aug}} = \mathcal{D}_k \cup \bigcup_{j=1}^{k-1} \mathcal{D}'_j$. A quantidade de imagens sintéticas geradas por classe é controlada por um parâmetro fixo do algoritmo, denotado por N_{sint}, garantindo uma proporção controlada entre amostras reais e sintéticas durante o treinamento. As imagens sintéticas geradas são rotuladas com a mesma classe utilizada durante o treinamento do respectivo

gerador e são combinadas diretamente com as imagens reais da tarefa atual. Assim, durante o treinamento do classificador, ambas são tratadas como exemplos supervisionados da mesma classe. A distinção entre amostras reais e sintéticas ocorre apenas no nível da função de custo, na qual imagens sintéticas podem receber pesos diferenciados de acordo com sua qualidade estimada, conforme descrito na Seção 4.3. Então, um novo classificador f_{θ_k} é treinado sobre esse conjunto. A função de perda utilizada é a entropia cruzada supervisionada:

$$\mathcal{L}_i(\boldsymbol{y}^{(i)}_{\text{true}}, \boldsymbol{y}^{(i)}_{\text{pred}}) = -\sum_{j=1}^{C} y^{(i)}_{\text{true},j} \cdot \log(y^{(i)}_{\text{pred},j}) \,, \tag{4.2}$$

em que C é o número total de classes, $y^{(i)}_{\text{true},j}$ é o valor verdadeiro para a classe j da amostra i, e $y^{(i)}_{\text{pred},j}$ é a probabilidade estimada pelo classificador para essa mesma classe.

Esse processo implica que, a cada nova tarefa, o conjunto de dados sintéticos utilizado durante o treinamento corresponde à união das bases sintéticas geradas para representar todas as tarefas anteriores. Formalmente, define-se

$$T'_{<k} = \bigcup_{j=1}^{k-1} \mathcal{D}'_j \tag{4.3}$$

em que cada conjunto $\mathcal{D}'_j$ contém as imagens sintéticas geradas pelas WGAN-GPs treinadas na tarefa j e aprovadas pelo processo de avaliação de qualidade. Esse conjunto acumulado atua como uma memória sintética das tarefas anteriores, permitindo aproximar suas distribuições durante o treinamento da tarefa atual sem a necessidade de armazenar os dados originais.

Esse mecanismo segue o princípio de *pseudo-rehearsal*, no qual amostras geradas artificialmente são utilizadas para preservar conhecimento previamente aprendido. Na prática, isso significa que, ao treinar o modelo na tarefa k, são utilizadas simultaneamente as imagens reais da tarefa atual e as imagens sintéticas representativas de todas as tarefas $1, \ldots, k-1$.

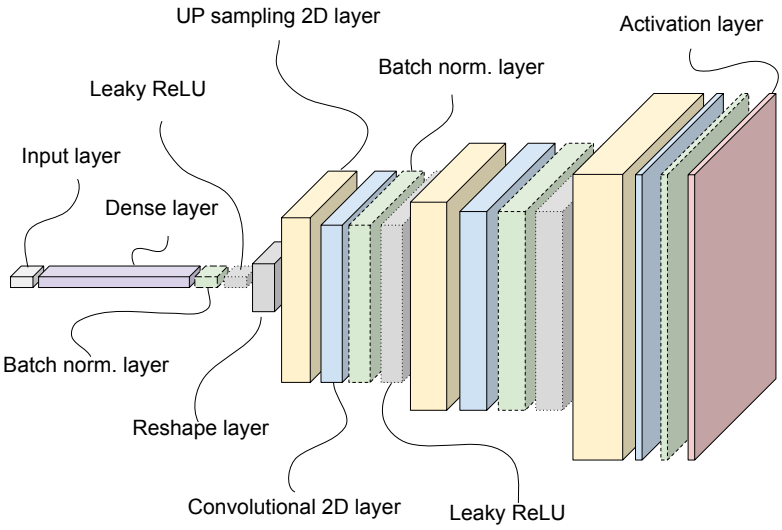
4.1.1 Geradores

Os geradores utilizados neste trabalho são WGAN-GPs (GULRAJANI et al., 2017). A escolha dessa arquitetura foi motivada por sua maior estabilidade de treinamento em comparação com GANs tradicionais, especialmente em cenários com conjuntos de dados relativamente pequenos (GULRAJANI et al., 2017; ANAYA-SáNCHEZ et al., 2025; YUDA et al., 2024). A penalização de gradiente substituiu o *weight clipping* do WGAN original, eliminando os problemas de gradientes explosivos e colapsantes que comprometem a convergência em GANs convencionais (GULRAJANI et al., 2017). Trabalhos anteriores demonstram que essa abordagem mantém atualizações estáveis e consistentes ao longo do treinamento, reduzindo a ocorrência de *mode collapse* (WU et al., 2024; SEHSAH;

MOUSA; FAROUK, 2025). Além disso, evidências empíricas indicam que a WGAN-GP mantém desempenho robusto mesmo com bases de dados menores (ANAYA-SáNCHEZ et al., 2025; SUN et al., 2022), condição relevante no presente trabalho.

Embora existam arquiteturas mais recentes, como a *Two-branch Disentangled Generative Adversarial Network* (TDGAN) (XIE; HU; CHEN, 2021), voltadas especificamente para a manipulação controlada de atributos faciais como identidade e expressão, este trabalho concentra-se na análise do impacto da utilização de imagens sintéticas no desempenho de modelos em cenários de aprendizado contínuo. O objetivo não é avaliar a fidelidade visual ou o realismo absoluto das imagens, mas sim investigar como diferentes proporções de imagens geradas afetam métricas como acurácia no conjunto de testes e mitigação do esquecimento catastrófico. Dessa forma, a escolha pela WGAN-GP é compatível com os objetivos definidos, centrados na análise quantitativa do uso de imagens sintéticas no treinamento incremental. É importante ressaltar que, no contexto de *generative replay*, o fator mais relevante é a capacidade do gerador de aproximar a distribuição das classes aprendidas anteriormente (LIU et al., 2020), de modo que mesmo imagens que não sejam visualmente perfeitas podem ser úteis para preservar representações discriminativas no classificador durante o aprendizado contínuo.

Figura 20 – Rede neural do gerador da WGAN-GP.



Fonte: autoria própria.

Assim, inicia-se o treinamento de um conjunto de WGAN-GPs, uma para cada uma das classes presentes no conjunto de dados fonte. Usando estas WGAN-GPs treinadas, geram-se as bases de dados sintéticas, por classe. Essas imagens geradas capturam os detalhes de cada classe das bases de dados.

Tabela 4 – Arquitetura das redes gerador e discriminador da WGAN-GP.

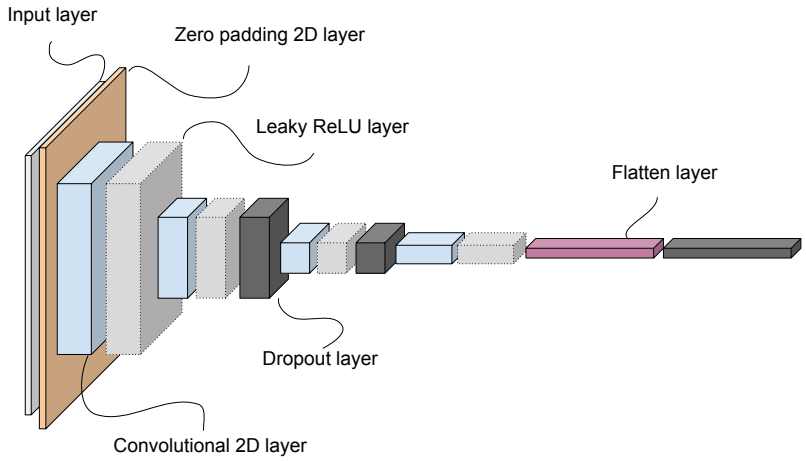
Gerador	**Saída**	**Discriminador**	**Saída**
Input Layer	(128)	*Input Layer*	(48, 48, 1)
Dense	(9216)	*Zero Padding* 2D	(52, 52, 1)
Batch Normalization	(9216)	*Convolutional* 2D	(26, 26, 64)
Leaky ReLU	(9216)	*Leaky ReLU*	(26, 26, 64)
Reshape	(6, 6, 256)	*Convolutional* 2D	(13, 13, 128)
Up Sampling 2D	(12, 12, 256)	*Leaky ReLU*	(13, 13, 128)
Convolutional 2D	(12, 12, 128)	*Dropout*	(13, 13, 128)
Batch Normalization	(12, 12, 128)	*Convolutional* 2D	(7, 7, 256)
Leaky ReLU	(12, 12, 128)	*Leaky ReLU*	(7, 7, 256)
Up Sampling 2D	(24, 24, 128)	*Dropout*	(7, 7, 256)
Convolutional 2D	(24, 24, 64)	*Convolutional* 2D	(4, 4, 512)
Batch Normalization	(24, 24, 64)	*Leaky ReLU*	(4, 4, 512)
Leaky ReLU	(24, 24, 64)	*Flatten*	(8192)
Up Sampling 2D	(48, 48, 64)	*Dropout*	(8192)
Convolutional 2D	(48, 48, 1)		
Batch Normalization	(48, 48, 1)		
Activation	(48, 48, 1)		
Parâmetros: 1586500		**Parâmetros: 4303360**	

Fonte: autoria própria.

A WGAN-GP é constituída por duas redes distintas, a rede do discriminador e a rede do gerador, descritas na Tabela 4. A rede do gerador, ilustrada na Figura 20, cria imagens sintéticas a partir de um ruído aleatório. A entrada consiste em um vetor de 128 números, gerados a partir de uma distribuição Gaussiana normal. Além da entrada, a rede utiliza camadas densas, *batch normalization* e camadas convolucionais com ativações *leaky ReLU*. A saída corresponde ao tamanho de imagem desejado, 48x48. Com cerca de 1,5 milhão de parâmetros, essa arquitetura produz imagens para desafiar o discriminador.

A rede do discriminador é crucial para distinguir entre imagens reais e sintéticas. Ela consiste em várias camadas, incluindo camadas convolucionais com funções de ativação *leaky ReLU*. Essas camadas ajudam o discriminador a extrair características relevantes das imagens de entrada. Além disso, são aplicadas camadas de *dropout* para evitar o *overfitting*. Cabe observar que as camadas de *dropout* são utilizadas apenas durante o treinamento como mecanismo de regularização. Durante a fase de inferência e avaliação dos modelos, essas camadas são desativadas, garantindo comportamento determinístico nas predições. Essa rede contém aproximadamente 4,3 milhões de parâmetros treináveis. A Figura 21 ilustra a rede do discriminador.

Figura 21 – Rede neural do discriminador da WGAN-GP.



Fonte: autoria própria.

4.1.2 Classificadores

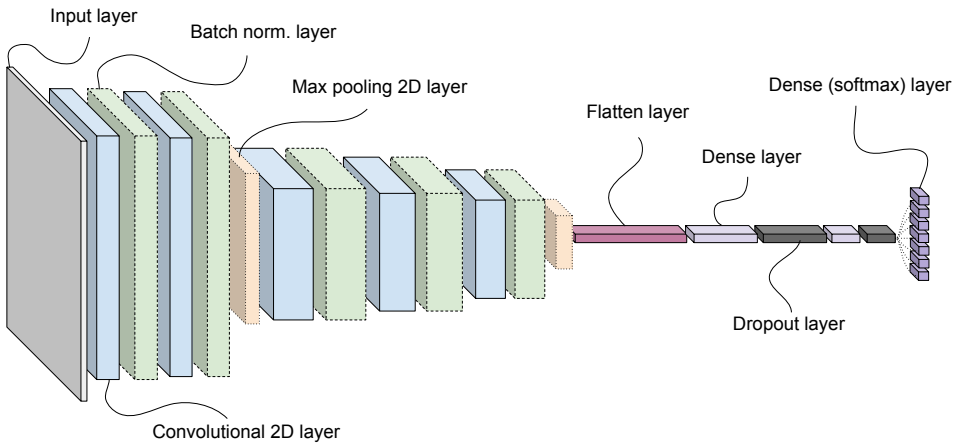
A CNN usada nos experimentos, adaptada do trabalho dos autores (TANNUGI; BRITTO; KOERICH, 2019), detalhada na Tabela 5, começa com camadas convolucionais 2D (64 filtros cada) e camadas de normalização. Também possui camadas convolucionais adicionais, *max-pooling* para redução de dimensionalidade e normalização adicional para extração de características de alto nível. Os mapas de características são achatados e passados por camadas totalmente conectadas com *dropout* para evitar *overfitting*, sendo utilizadas apenas durante o treinamento como mecanismo de regularização. Durante a fase de inferência e avaliação dos modelos, essas camadas são desativadas, garantindo comportamento determinístico nas predições. A camada final utiliza ativação *softmax* para gerar as probabilidades de cada classe. No geral, essa arquitetura de CNN compreende aproximadamente 19,3 milhões de parâmetros. Na Figura 22 é possível observar a arquitetura da CNN.

Tabela 5 – Arquitetura da CNN utilizada nos experimentos.

Camada	**Saída**	**Parâmetros**
Convolution 2D	(47, 47, 64)	320
Batch Normalization	(47, 47, 64)	256
Convolution 2D	(46, 46, 64)	16448
Batch Normalization	(46, 46, 64)	256
Max Pooling 2D	(23, 23, 64)	0
Convolution 2D	(21, 21, 128)	73856
Batch Normalization	(21, 21, 128)	512
Convolution 2D	(19, 19, 128)	147584
Batch Normalization	(19, 19, 128)	512
Convolution 2D	(17, 17, 128)	147584
Batch Normalization	(17, 17, 128)	512
Max Pooling 2D	(8, 8, 128)	0
Flatten	(8192)	0
Dense	(2048)	16779264
Dropout	(2048)	0
Dense	(1024)	2098176
Dropout	(1024)	0
Dense (Softmax)	(7)	7175
Parâmetros: 19272455		

Fonte: autoria própria.

Figura 22 – Arquitetura da CNN utilizada nos experimentos.



Fonte: autoria própria.

4.2 Abordagens para avaliação de qualidade de imagens sintéticas

Após a geração das imagens sintéticas pelas WGAN-GPs, torna-se necessário avaliar a qualidade dessas amostras antes de utilizá-las no retreinamento da rede neural. Embora os geradores sejam capazes de aproximar a distribuição das imagens reais, é comum que algumas amostras apresentem artefatos ou representações inconsistentes das expressões

faciais.

Nesta seção, propomos três abordagens para realizar a avaliação da qualidade das imagens sintéticas geradas pelas WGAN-GPs. O objetivo destes métodos é selecionar subconjuntos das imagens geradas que não apresentem inconsistências semânticas, uma vez que imagens geradas com inconsistências semânticas podem introduzir ruído no processo de treinamento e degradar o desempenho do classificador, e, portanto, sejam benéficas para compor o conjunto de dados de treinamento em um cenário de aprendizado contínuo. Dessa forma, a etapa de avaliação de qualidade busca reduzir o impacto dessas amostras potencialmente prejudiciais no aprendizado contínuo. As três abordagens propostas são:

- Uma abordagem supervisionada baseada em uma rede neural convolucional (CNN) treinada em tarefas anteriores;
- Uma abordagem não-supervisionada baseada em agrupamento com validação interna de qualidade;
- Uma abordagem de otimização do espaço latente das WGAN-GPs.

A seguir, descreve-se detalhadamente os três métodos propostos para avaliação de qualidade de imagens sintéticas. O primeiro método utiliza uma CNN treinada de forma incremental para estimar a qualidade das imagens com base na sua capacidade de classificação. O segundo método é baseado em um processo de agrupamento, combinando diferentes representações extraídas por modelos de *face embedding*, e avaliando a qualidade da alocação das instâncias a partir do coeficiente de silhueta dos *clusters*. O terceiro método propõe uma estratégia de otimização no espaço latente das WGAN-GPs com o objetivo de identificar vetores que geram imagens sintéticas de alta qualidade. Inicialmente, imagens são geradas a partir de amostras latentes aleatórias e, em seguida, filtradas por uma CNN previamente treinada nas tarefas anteriores. As imagens que são corretamente classificadas pela CNN são consideradas relevantes, e os respectivos vetores latentes utilizados para gerá-las são então preservados. A partir desse subconjunto de vetores latentes validados, constrói-se o involucro convexo (*convex hull*) que define a região do espaço latente associada a imagens de boa qualidade. Amostras adicionais são então geradas interpolando vetores dentro desse poliedro convexo. Estes métodos têm como objetivo principal identificar subconjuntos de imagens sintéticas confiáveis para serem utilizadas como memória no processo de *pseudo-rehearsal*.

4.2.1 Método supervisionado baseado em CNN

O método de QA proposto tem o objetivo de filtrar as imagens geradas pelas WGAN-GPs. Assim, dado um número k de tarefas, uma CNN treinada continuamente em

todas as tarefas até $k-1$ e a união de todas as bases sintéticas das tarefas anteriores, o objetivo da filtragem é

$$\mathcal{F}(T'_{<k}, \boldsymbol{\theta}) = \left\{ \mathbf{x} \in \bigcup_{d' \in T'_{<k}} d' \mid \arg\max(\mathrm{CNN}(\mathbf{x}; \boldsymbol{\theta})) = y \right\}, \tag{4.4}$$

em que $\mathbf{x}$ representa uma imagem de uma base de dados sintética d' pertencente ao conjunto $T'_{<k}$, gerada pelas WGAN-GPs; $\boldsymbol{\theta}$ são os pesos da CNN treinada continuamente até a tarefa $k-1$; $\arg\max(\mathrm{CNN}(\mathbf{x}; \boldsymbol{\theta}))$ fornece o rótulo predito $\hat{y}$ pela rede neural para a imagem $\mathbf{x}$; e y é o rótulo original associado à amostra. Assim, a função $\mathcal{F}(T'_{<k}, \boldsymbol{\theta})$ retorna o subconjunto de imagens sintéticas consideradas semanticamente consistentes segundo o critério aprendido pela CNN treinada nas tarefas anteriores.

4.2.2 Método não-supervisionado com *clusters*

Uma forma de identificar as inconsistências nas imagens sintéticas é analisar a organização estrutural das imagens no espaço de características. Espera-se que imagens pertencentes à mesma classe emocional apresentem proximidade nesse espaço quando representadas por extratores de características adequados. Assim, imagens sintéticas que não seguem esse padrão tendem a aparecer como pontos mal posicionados em relação aos agrupamentos formados pelas demais amostras. Com base nessa premissa, este trabalho utiliza um método de agrupamento não supervisionado para avaliar a coerência estrutural das imagens geradas no espaço de características.

O método não supervisionado proposto é aplicado após a geração das imagens sintéticas pelas WGAN-GPs. Inicialmente, as imagens geradas são representadas por diferentes extratores de características, produzindo múltiplos conjuntos de *embeddings*. Em seguida, para cada conjunto de representações, aplica-se um algoritmo de agrupamento para analisar a organização estrutural das amostras no espaço de características. Posteriormente, utiliza-se o conhecimento de rótulos verdadeiros para avaliar se a melhor atribuição de *cluster* corresponde à classe real.

Seja $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ o conjunto de extratores de características utilizados neste método. Foram considerados dez extratores previamente treinados disponíveis nas bibliotecas Keras e DeepFace: ConvNeXt, DeepID, Facenet512, SFace, ResNet50, EfficientNetV2, GhostFaceNet, OpenFace, Facenet e InceptionV3. Cada extrator E_e transforma uma imagem $\mathbf{x}_i$ em um vetor de características $\mathbf{z}_i^{(e)} \in \mathbb{R}^{d_e}$, onde d_e representa a dimensionalidade do espaço de características produzido por esse extrator.

Para cada conjunto de *embeddings* gerado por um extrator E_e, aplica-se um algoritmo de agrupamento (K-Means) para particionar os dados em r *clusters*. O resultado é um vetor de rótulos

$$\mathbf{c}^{(e)} = [c_1^{(e)}, c_2^{(e)}, \ldots, c_n^{(e)}] \tag{4.5}$$

sendo que $c_i^{(e)} \in \{1, 2, \ldots, r\}$ representa o rótulo do *cluster* atribuído à instância i sob o modelo E_e.

O algoritmo de agrupamento utilizado foi o K-Means. O número de *clusters* r foi definido como igual ao número de classes emocionais presentes no conjunto de dados, de forma consistente com a estrutura das tarefas de reconhecimento de expressões faciais. Assim, cada *cluster* tende a representar uma região do espaço de características associada a uma classe emocional específica.

Para cada instância i, calcula-se o coeficiente de silhueta $s_i^{(e)} \in [-1, 1]$, correspondente à atribuição feita por E_e. Esse valor quantifica a qualidade da separação de i em relação aos *clusters*. Valores próximos de 1 indicam que a instância está bem alocada, enquanto valores negativos sugerem alocação incorreta.

O coeficiente de silhueta é um índice interno de validação de agrupamentos que mede o grau de separação entre *clusters* com base nas distâncias entre amostras no espaço de características. Para cada instância i, define-se $a(i)$ como a distância média entre i e os demais elementos do mesmo *cluster*, e $b(i)$ como a menor distância média entre i e os elementos de outros *clusters*. O coeficiente de silhueta é então definido por

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4.6}$$

cujos valores pertencem ao intervalo $[-1, 1]$. Na implementação utilizada neste trabalho, os valores são posteriormente reescalados para o intervalo $[0, 1]$ por meio da transformação $s'(i) = (s(i) + 1)/2$, apenas para facilitar sua interpretação como medida de confiança.

Para cada instância i, seleciona-se o extrator E_{e^*} cuja atribuição de *cluster* $c_i^{(e^*)}$ maximize o coeficiente de silhueta:

$$e^* = \arg\max_e \left(s_i^{(e)} \right). \tag{4.7}$$

Esse procedimento define um *ensemble* baseado em máxima confiança por silhueta, permitindo identificar, para cada instância, a combinação mais apropriada de extrator e agrupamento.

Assim como no método supervisionado, o objetivo aqui é definir uma função de filtragem $\mathcal{F}\text{clust}$ que, dada a união das bases sintéticas geradas até a tarefa $k-1$, retorne um subconjunto de instâncias consideradas de qualidade suficiente segundo o critério não supervisionado baseado em agrupamento:

$$\mathcal{F}\text{clust}(T' < k, \mathcal{E}) = \mathbf{x}_i \in \bigcup d' \in T'_{<k} d' \mid \arg\max_e s_i^{(e)} > \tau \text{ e } \delta(c_i^{(e^*)}, \mathbf{y}_i) = 1 \,, \tag{4.8}$$

em que $\mathbf{x}_i$ é uma imagem sintética da base d' pertencente ao conjunto T', $s_i^{(e)}$ é o coeficiente de silhueta da instância $\mathbf{x}_i$ no extrator $E_e \in \mathcal{E}$, $\arg\max_e s_i^{(e)}$ é o extrator que produziu o maior valor de silhueta para a instância $\mathbf{x}_i$, $c_i^{(e^*)}$ é o rótulo do *cluster* atribuído à instância

pela combinação de extrator e agrupamento e $\mathbf{y}_i$ é o rótulo verdadeiro da instância. O termo $\delta(c_i^{(e^*)}, \mathbf{y}_i) = 1$ indica que há correspondência entre o rótulo do *cluster* e a classe real (via mapeamento majoritário), e τ é um limiar mínimo de silhueta para considerar a instância bem agrupada.

A saída final é o subconjunto filtrado $T'^{<}k = \mathcal{F}\text{clust}(T'^{<}k, \mathcal{E})$, que representa as instâncias sintéticas consideradas aptas para uso em etapas posteriores do aprendizado contínuo. Cabe destacar que o procedimento de filtragem é aplicado a todos os conjuntos de imagens sintéticas gerados para as tarefas anteriores.

4.2.3 Método baseado em otimização no espaço latente da WGAN-GP

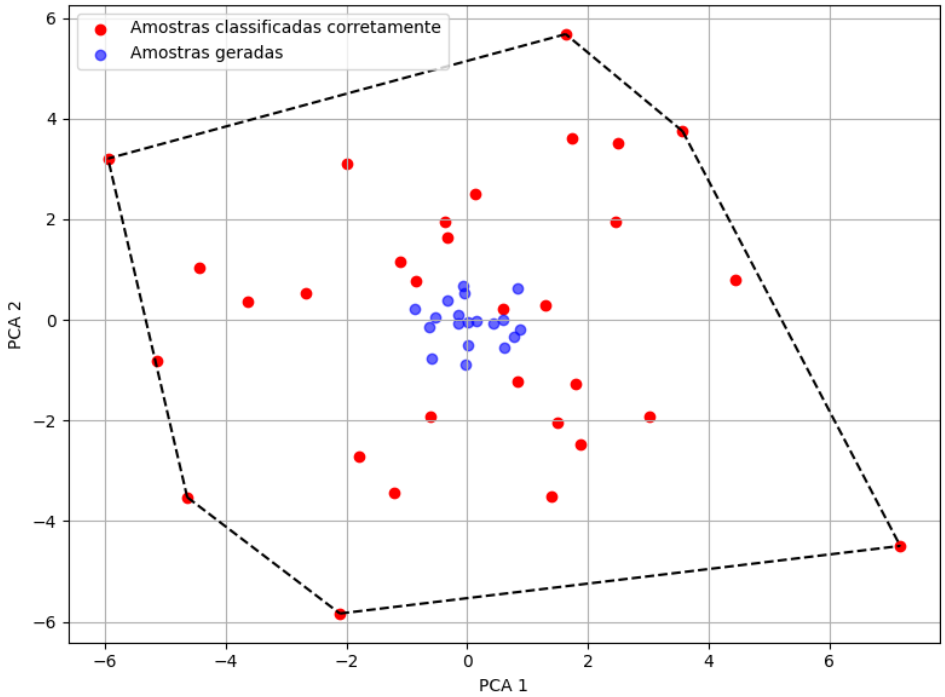
O método proposto para otimização no espaço latente das WGAN-GPs, denominado *Coefficient Guided Latent Optimization* (CGLO), utiliza uma CNN previamente treinada como referência semântica para identificar regiões promissoras do espaço latente do gerador. Diferentemente do método de filtragem baseado em CNN apresentado na Seção 4.2.1, cujo objetivo é apenas selecionar imagens sintéticas já geradas que são corretamente classificadas pela rede, o CGLO utiliza essas imagens aprovadas como pontos de partida para explorar novas regiões do espaço latente e gerar amostras adicionais potencialmente úteis para o treinamento contínuo.

Em outras palavras, enquanto o método da Seção 4.2.1 atua apenas como um filtro de qualidade sobre imagens já geradas, o CGLO utiliza essas amostras filtradas para guiar a geração de novas amostras no espaço latente do gerador, buscando regiões próximas a vetores latentes associados a imagens consideradas semanticamente consistentes pela CNN.

Inicialmente, um conjunto $\mathcal{Z}_{\text{raw}}$ de vetores latentes $\mathbf{z}_i \in \mathbb{R}^d$ é amostrado aleatoriamente do espaço latente da WGAN-GP, e as imagens correspondentes $\mathbf{x}_i = G(\mathbf{z}_i)$ são geradas. Essas imagens são então avaliadas por uma CNN supervisionada f_θ, treinada previamente nas tarefas anteriores, que serve como filtro de qualidade. Apenas as imagens classificadas corretamente por f_θ são mantidas (ver Seção 4.2.1), e os respectivos vetores latentes $\mathbf{z}_i$ compõem o conjunto filtrado $\mathcal{Z}_{\text{filt}}$. Cabe destacar que a CNN utilizada como filtro pode apresentar erros de classificação. Nesse contexto, o método assume que, ao longo das tarefas anteriores, a rede já tenha aprendido representações suficientemente consistentes das classes emocionais. Assim, eventuais erros individuais da CNN podem introduzir ruído no processo de seleção, porém o uso de múltiplas amostras e a exploração posterior do espaço latente tendem a reduzir o impacto de classificações incorretas isoladas.

A etapa seguinte consiste em explorar novas regiões promissoras do espaço latente, sob a hipótese de que vetores latentes associados a amostras de alta qualidade encontram-se próximos entre si nesse espaço. Para isso, calcula-se o invólucro convexo (*convex hull*) do subconjunto filtrado $\mathcal{Z}_{\text{filt}}$, que contém vetores latentes $\mathbf{z}_i$ previamente avaliados como

Figura 23 – Amostras geradas dentro do *convex hull* das imagens corretamente classificadas por uma CNN previamente treinada. Vetores reduzidos utilizando PCA.



Fonte: autoria própria.

promissores.

Para gerar novos vetores candidatos $\mathbf{z}_{\text{opt}}$, amostram-se combinações convexas de k vetores distintos $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\} \subset \mathcal{Z}_{\text{filt}}$:

$$\mathbf{z}_{\text{opt}} = \sum_{i=1}^{k} \alpha_i \mathbf{z}_i, \quad \text{com } \alpha_i \geq 0 \text{ e } \sum_{i=1}^{k} \alpha_i = 1. \tag{4.9}$$

Essa formulação garante que $\mathbf{z}_{\text{opt}}$ pertença ao invólucro convexo dos vetores selecionados, pois, por definição, uma combinação convexa de vetores está contida nesse conjunto. A distribuição de Dirichlet é utilizada para amostrar os coeficientes α_i, garantindo que os pesos sejam não negativos e que sua soma seja igual a 1. Na implementação adotada, os parâmetros da distribuição são definidos de forma uniforme, de modo que nenhuma direção específica do espaço latente seja priorizada. Essa escolha favorece a geração de combinações convexas diversificadas entre os vetores latentes selecionados, permitindo explorar regiões intermediárias do espaço latente sem impor um viés explícito para amostras mais concentradas ou mais dispersas. Cabe observar que essa formulação não impõe explicitamente restrições para evitar possíveis sobreposições entre regiões latentes associadas a diferentes classes. Entretanto, como os vetores latentes utilizados na construção do invólucro convexo são provenientes de imagens previamente validadas pela CNN para uma classe específica, assume-se que tais vetores estejam predominantemente concentrados em regiões do espaço latente associadas a essa classe.

A Figura 23 ilustra exemplos dessas amostras $\mathbf{z}_{\text{opt}}$ no invólucro convexo de $\mathcal{Z}_{\text{filt}}$, projetadas em duas dimensões via análise de componentes principais (PCA, do inglês *Principal Component Analysis*) para fins de visualização.

As novas amostras $\mathbf{z}_{\text{opt}}$ são então novamente passadas pelo gerador da WGAN-GP, gerando imagens $\mathbf{x}_{\text{opt}} = G(\mathbf{z}_{\text{opt}})$, que também são avaliadas pela CNN f_θ. Este processo iterativo permite a exploração eficiente de regiões do espaço latente que maximizam a probabilidade de gerar imagens sintéticas que podem ser consideradas benéficas para o retreinamento e que estejam alinhadas com o conhecimento aprendido nas tarefas anteriores.

Formalmente, seja $\mathcal{Z}_{\text{opt}}$ o conjunto de vetores latentes gerados via combinações convexas a partir de $\mathcal{Z}_{\text{filt}}$, e $G(\cdot)$ o gerador da WGAN-GP. As imagens sintéticas geradas são então avaliadas por uma CNN supervisionada $\text{CNN}(\cdot;\boldsymbol{\theta})$, treinada nas tarefas até $k{-}1$. Define-se o subconjunto de imagens $T'{<}\,k$ aceitas pelo filtro como

$$\mathcal{F}_{\text{cglo}}(T'_{<k}, G, \boldsymbol{\theta}) = \{\mathbf{x} = G(\mathbf{z}) \mid \mathbf{z} \in \mathcal{Z}_{\text{opt}} \wedge \arg\max(\text{CNN}(\mathbf{x};\boldsymbol{\theta})) = y(\mathbf{x})\}\,, \tag{4.10}$$

em que $\mathbf{z} \in \mathcal{Z}_{\text{opt}}$ representa o conjunto de vetores latentes; $G(\cdot)$ denota o gerador da WGAN-GP; $\text{CNN}(\cdot;\boldsymbol{\theta})$ é o modelo supervisionado treinado até a tarefa $k{-}1$; e $y(\mathbf{x})$ corresponde ao rótulo verdadeiro da imagem gerada $\mathbf{x}$.

Assim, a função $\mathcal{F}_{\text{cglo}}(T'_{<k}, G, \boldsymbol{\theta})$ retorna o subconjunto de imagens sintéticas consideradas coerentes com o conhecimento aprendido até a tarefa $k{-}1$, e que são então utilizadas como memória sintética no processo de aprendizado contínuo.

4.3 Função de custo ponderada

Nesta seção, descreve-se como a qualidade das imagens sintéticas é incorporada ao processo de aprendizado por meio de uma função de custo ponderada. A ideia central é atribuir pesos às imagens sintéticas de acordo com sua qualidade estimada durante a etapa de avaliação. Imagens de baixa qualidade terão menor influência no processo de aprendizado contínuo.

Dividem-se os métodos de avaliação de qualidade em três tipos:

- **(i) Avaliação baseada em CNN:** usa a confiança da predição da CNN.
- **(ii) Avaliação baseada em *cluster*:** usa o coeficiente de silhueta da instância nos agrupamentos.
- **(iii) Avaliação baseada em otimização do espaço latente (CGLO):** utiliza a confiança da CNN associada às amostras geradas em regiões selecionadas do espaço latente.

4.3.1 Perda ponderada

A função de custo usada durante o aprendizado contínuo é modificada para considerar um peso $w_{\text{img}}^{(i)}$ para cada imagem i:

$$\mathcal{L}_i(\mathbf{y}_{\text{true}}^{(i)}, \mathbf{y}_{\text{pred}}^{(i)}, w_{\text{img}}^{(i)}) = -\sum_{j=1}^{C} w_{\text{img}}^{(i)} \cdot y_{\text{true j}}^{(i)} \cdot \log(y_{\text{pred j}}^{(i)}) \,, \tag{4.11}$$

onde C é o número de classes, e $w_{\text{img}}^{(i)}$ controla o impacto da instância na otimização da rede.

4.3.2 QA com base na CNN

Para este método, o peso é a probabilidade atribuída pela CNN à classe correta da imagem sintética:

$$w_{\text{img}}^{(i)} = p_{\text{pred j}}^{(i)} \,, \tag{4.12}$$

em que $p_{\text{pred j}}^{(i)}$ é a confiança da predição da CNN na classe verdadeira. Intuitivamente, se a imagem sintética é ruim, a CNN irá classificá-la incorretamente com baixa confiança, resultando em um peso pequeno.

Cabe destacar que a estimativa de qualidade baseada na CNN depende da confiabilidade do modelo treinado nas tarefas anteriores. Caso a rede apresente incerteza ou erros de classificação, essa incerteza é naturalmente refletida no valor de w_{img}, reduzindo o impacto de amostras potencialmente inconsistentes durante o processo de aprendizado contínuo.

4.3.3 QA com base em *cluster*

Neste caso, a qualidade da imagem sintética é medida com base no coeficiente de silhueta da sua representação no espaço de *features*. Como os coeficientes de silhueta podem variar de -1 a 1, eles são normalizados para o intervalo $[0, 1]$:

$$w_{\text{img}}^{(i)} = \frac{s_i + 1}{2} \,, \tag{4.13}$$

em que s_i é o coeficiente de silhueta da instância i. Essa normalização permite que o valor seja interpretado como um peso.

4.3.4 QA com base em otimização do espaço latente (CGLO)

Para o método CGLO, o peso $w_{\text{img}}^{(i)}$ de cada imagem válida é definido como a confiança da CNN na classe correta, da mesma forma descrita na Seção 4.3.2:

$$w_{\text{img}}^{(i)} = p_{\text{pred j}}^{(i)} \,, \tag{4.14}$$

onde $p^{(i)}_{\text{pred j}}$ é a probabilidade atribuída pela CNN à classe verdadeira da imagem i.

Apesar de utilizar a mesma fórmula de ponderação do método supervisionado, o CGLO introduz um critério adicional de qualidade ao restringir a geração às regiões mais confiáveis do espaço latente, o que aumenta a chance de obter imagens sintéticas úteis e representativas para a tarefa.

4.3.5 Imagens reais

As imagens reais da base de dados alvo são sempre consideradas com peso máximo, ou seja

$$w^{(i)}_{\text{img}} = 1, \quad \text{para instâncias reais.} \tag{4.15}$$

Assim, a rede prioriza as imagens da nova tarefa no processo de aprendizado, independentemente do método de QA utilizado para as imagens sintéticas.

4.4 Algoritmo geral do método proposto

O Algoritmo 1 apresenta o fluxo geral do método proposto para aprendizado contínuo com *pseudo-rehearsal* utilizando WGAN-GPs e avaliação de qualidade (QA) das imagens sintéticas. Para cada tarefa, o algoritmo treina geradores WGAN-GP para cada classe, gera imagens sintéticas das tarefas anteriores, aplica o método de QA selecionado para filtrar essas imagens e, finalmente, treina um classificador CNN na combinação dos dados reais da nova tarefa e das imagens sintéticas filtradas.

4.5 Considerações finais

O método proposto neste trabalho busca mitigar o esquecimento catastrófico no aprendizado contínuo de redes neurais convolucionais aplicadas ao reconhecimento de expressões faciais. Para isso, é utilizada uma abordagem de *pseudo-rehearsal* baseada em geração de imagens sintéticas por meio de WGAN-GP, permitindo que a rede neural "reviva" tarefas anteriores sem acesso aos dados originais destas tarefas. Esse processo é enriquecido por uma etapa de garantia de qualidade (QA), composta por três estratégias complementares: uma CNN pré-treinada, validação por agrupamento não supervisionado, e um mecanismo baseado no espaço latente da própria WGAN-GP. As imagens sintéticas aprovadas pelo QA são ponderadas na função de custo conforme sua confiabilidade estimada, de modo que tenham influência na atualização dos pesos da rede. O fluxo completo divide-se em uma fase onde os modelos generativos são treinados e avaliados, e uma fase em que o retreinamento da CNN é realizado com a combinação ponderada de imagens sintéticas da tarefa anterior e imagens reais da tarefa atual. Com isso, o método

Algorithm 1 Aprendizado contínuo com *pseudo-rehearsal* e QA

Entrada: Tarefas $\mathcal{T} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, `QA_METHOD`, extratores $\mathcal{E}$, N_{sint}
Saída: Conjunto de classificadores $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_K}\}$
Inicialize $\mathcal{G} \leftarrow \emptyset$, $\mathcal{F} \leftarrow \emptyset$
for $k = 1$ **até** K **do** ▷ Para cada nova tarefa
    $C_k \leftarrow \{y \mid (\mathbf{x}, y) \in \mathcal{D}_k\}$ ▷ Identifica classes da tarefa k
    for cada $c \in C_k$ **do**
        $\mathcal{D}_k^{(c)} \leftarrow \{(\mathbf{x}, y) \in \mathcal{D}_k \mid y = c\}$
        Treinar gerador $G_k^{(c)}$ via WGAN-GP em $\mathcal{D}_k^{(c)}$; Adicione $G_k^{(c)}$ a $\mathcal{G}$
    end for
    $\mathcal{D}^{\text{sint}} \leftarrow \emptyset$
    if $k > 1$ **then**
        for $j = 1$ **até** $k - 1$ **do** ▷ Para cada tarefa passada
            $C_j \leftarrow \{c \mid G_j^{(c)} \in \mathcal{G}\}$ ▷ Recupera classes da tarefa j
            for cada $c \in C_j$ **do**
                for $n = 1$ **até** N_{sint} **do** ▷ Gera N_{sint} amostras
                    Amostre $\tilde{\mathbf{z}} \sim \mathcal{N}(0, \mathbf{I})$
                    $\tilde{\mathbf{x}} \leftarrow G_j^{(c)}(\tilde{\mathbf{z}})$
                    $amostra\_aceita \leftarrow false$
                    Aplicar QA:
                    if `QA_METHOD` = `CNN` **then**

                        if $\arg\max f_{\theta_j}(\tilde{\mathbf{x}}) = c$ **then** $amostra\_aceita \leftarrow true$
                        end if
                    else if `QA_METHOD` = `CLUSTER` **then**
                        Extraia *embeddings* $\{\mathbf{z}_i^{(e)}\}$ utilizando os extratores $\mathcal{E}$
                        for cada extrator $E_e \in \mathcal{E}$ **do**
                            Agrupe com r *clusters* (K-means)
                            Calcule coeficiente de silhueta $s_i^{(e)}$
                        end for
                        $e^* \leftarrow \arg\max_e s_i^{(e)}$
                        if $s_i^{(e^*)} > \tau$ **e** $c_i^{(e^*)} = c$ **then**
                        end if
                    else if `QA_METHOD` = `LATENT` **then**
                        Amostre $\tilde{\mathbf{z}} \in \text{Conv}(\mathcal{Z}_{\text{filt}}^{(j,c)})$
                        $\tilde{\mathbf{x}} \leftarrow G_j^{(c)}(\tilde{\mathbf{z}})$

                        if $\arg\max f_{\theta_j}(\tilde{\mathbf{x}}) = c$ **then** $amostra\_aceita \leftarrow true$
                        end if
                    end if

                    if $amostra\_aceita$ **then** $\mathcal{D}^{\text{sint}} \leftarrow \mathcal{D}^{\text{sint}} \cup \{(\tilde{\mathbf{x}}, c)\}$
                    end if
                end for
            end for
        end for
    end if
    $\mathcal{D}_k^{\text{treino}} \leftarrow \mathcal{D}_k \cup \mathcal{D}^{\text{sint}}$
    Treine classificador f_{θ_k} em $\mathcal{D}_k^{\text{treino}}$; Adicione f_{θ_k} a $\mathcal{F}$
end for
return $\mathcal{F}$

busca preservar o conhecimento adquirido anteriormente, mesmo na ausência dos dados originais das tarefas anteriores, promovendo um aprendizado contínuo mais robusto.

Os métodos apresentados nesta seção visam responder às questões de pesquisa expostas na Seção 1.3. Especificamente, a avaliação da combinação dos métodos apresentados e a avaliação destes mesmos métodos individualmente nos traz clareza quanto à viabilidade de utilizar os procedimentos supracitados e de suas capacidades de atuar em conjunto. Ao definir os métodos de maneira formal, possibilita-se a adaptabilidade para aplicação destes métodos para mitigar o esquecimento de memória em cenários e contextos diferentes do que este trabalho está inserido, que é o reconhecimento de emoções.

5 Resultados experimentais

Neste capítulo, apresentam-se os principais resultados experimentais obtidos a partir da aplicação do método proposto nas bases de dados. Avaliam-se os métodos propostos em diversos cenários e discutem-se os resultados à luz das questões de pesquisa que guiam este trabalho. Na Seção 5.1, descrevem-se todas as bases de dados utilizadas nos experimentos, com a descrição detalhada de cada conjunto de imagens. Na Seção 5.1.3 estão descritos todos os procedimentos realizados para preparação das bases de dados para as etapas de treinamento e teste das redes neurais com os métodos propostos. A Seção 5.2 discute os resultados qualitativos das imagens geradas a partir do uso de GANs. A Seção 5.3 traz os resultados dos protocolos experimentais aplicados às bases de dados. Por fim, a Seção 5.4 discute sobre todos os resultados obtidos, levando em consideração os objetivos e as questões de pesquisa que norteiam este trabalho.

5.1 Bases de dados

Nesta seção são apresentadas as diferentes bases de dados utilizadas durante os experimentos. No âmbito deste trabalho, cinco bases de dados foram utilizadas, sendo quatro delas no contexto de reconhecimento de emoções e uma no contexto de classificação de dígitos.

5.1.1 Bases de emoções

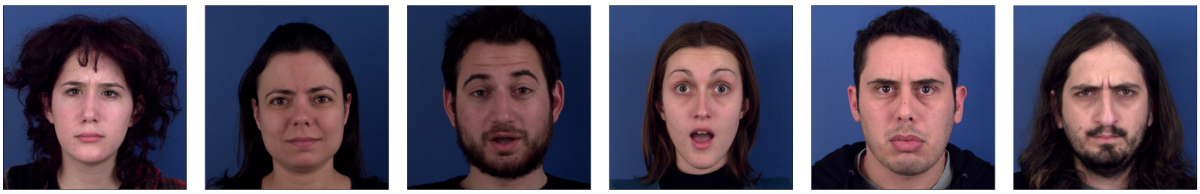
No domínio das bases que representam emoções, foram utilizadas quatro bases de dados, a saber: MUG (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010), JAFFE (LYONS; KAMACHI; GYOBA, 1998), TFEID (CHEN et al., 2009) e CK+ (LUCEY et al., 2010). Nesta seção apresentam-se os detalhes de cada uma delas.

5.1.1.1 MUG

O conjunto de imagens MUG (*Multimodal Understanding Group*) (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010) é um conjunto de dados abrangente projetado para o reconhecimento de expressões faciais e pesquisas multimodais. Ele contém uma coleção de imagens faciais e gravações de áudio, tornando-o adequado para tarefas de reconhecimento audiovisual de emoções. Na Figura 24 é possível observar alguns exemplos da base de dados.

O conjunto de dados inclui uma diversidade de 86 indivíduos realizando expressões faciais. Os participantes estavam sentados em uma cadeira em frente a uma câmera, com

Figura 24 – Amostras da base de dados MUG.



Fonte: (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010).

um fundo azul. Foram utilizadas duas fontes de luz de 300W cada, montadas em suportes a uma altura de aproximadamente 130 cm. Em cada suporte, um guarda-chuva foi fixado para difundir a luz e evitar sombras. A câmera foi capaz de capturar imagens a uma taxa de 19 quadros por segundo. Cada imagem foi salva no formato JPG, com resolução de 896x896 *pixels* e tamanho variando entre 240 e 340 KB. Participaram do banco de dados 35 mulheres e 51 homens, todos de origem caucasiana e com idades entre 20 e 35 anos. Os homens estão com ou sem barba. Os indivíduos não usam óculos, exceto por 7 deles na segunda parte do banco de dados. Não há oclusões, exceto por alguns fios de cabelo caindo sobre o rosto (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010). Para atingir o objetivo, antes das gravações, segundo os autores Aifanti, Papachristou e Delopoulos (2010), foi dado aos participantes um breve tutorial sobre as emoções básicas. Os participantes foram informados sobre como realizar as seis expressões faciais (raiva, desgosto, medo, felicidade, tristeza, surpresa e neutro) de acordo com os "protótipos de emoção" definidos no Guia do Investigador do manual FACS (EKMAN; FRIESEN; HAGER, 2002). O objetivo era evitar expressões incorretas, ou seja, expressões que não correspondem realmente ao seu rótulo. Após os participantes aprenderem as diferentes formas de realizar as seis expressões, eles puderam escolher livremente imitar uma delas. As sequências de imagens começam e terminam em um estado neutro e seguem o padrão temporal de início, ápice (emoção imitada) e final.

Segundo os autores Aifanti, Papachristou e Delopoulos (2010), em termos de escala, a base de dados MUG consiste em aproximadamente 1462 imagens faciais, cada uma anotada com os rótulos correspondentes a uma das 6 expressões faciais: raiva, desgosto, medo, felicidade, tristeza e surpresa. A expressão facial neutra também está presente nas sequências de imagens e pode ser utilizada com uma expressão facial.

5.1.1.2 JAFFE

O conjunto de dados JAFFE (*Japanese Female Facial Expression*) (LYONS; KAMACHI; GYOBA, 1998) é um *benchmark* amplamente utilizado no reconhecimento de expressões faciais, projetado para estudar a similaridade semântica e a codificação de imagens faciais. Diferente de outras bases, esta foca exclusivamente em um perfil demo-

gráfico específico para controlar variáveis culturais. A base é composta por 213 imagens provenientes de 10 modelos. Em relação à composição e coleta:

- Gênero e Etnia: O grupo é composto exclusivamente por 10 mulheres japonesas.
- Distribuição das Classes: Cada modelo posou para 3 ou 4 exemplos de cada uma das seis expressões básicas (felicidade, tristeza, surpresa, raiva, desgosto e medo) e uma face neutra.
- Ambiente de Coleta: As imagens foram capturadas em ambiente controlado, com iluminação uniforme (luzes de tungstênio) e as modelos utilizaram um espelho semi-refletivo para visualizar suas expressões durante a captura. O cabelo foi mantido preso para expor todas as zonas expressivas do rosto.
- Dados Técnicos: As imagens originais foram digitalizadas em tons de cinza com resolução de 256x256 *pixels*.

Figura 25 – Amostras da base de dados JAFFE.

Fonte: (LYONS; KAMACHI; GYOBA, 1998).

Um aspecto relevante desta base é o seu processo de validação semântica. Um total de 92 estudantes universitárias japonesas avaliaram cada imagem em uma escala de cinco pontos para cada uma das seis expressões básicas. Essas avaliações permitiram a criação de vetores semânticos médios para cada imagem, facilitando a comparação entre o desempenho de algoritmos e a percepção humana. Vale notar que os autores Lyons, Kamachi e Gyoba (1998) indicam que a expressão de medo é considerada problemática para sujeitos japoneses, o que pode influenciar os resultados de classificação nesta classe específica.

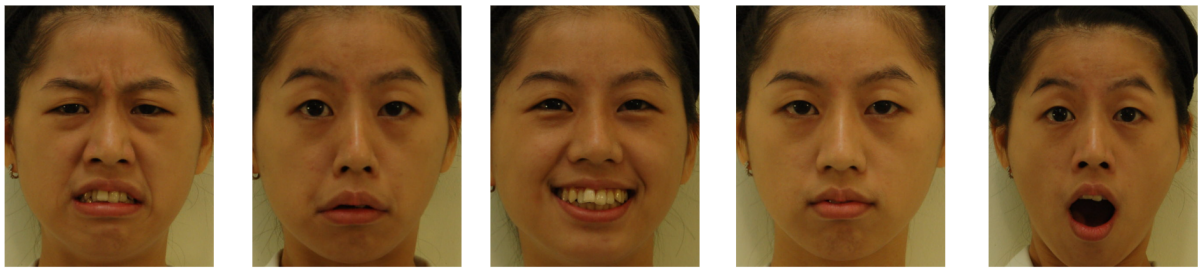
5.1.1.3 TFEID

A base de dados TFEID (*Taiwanese Facial Expression Image Database*) (CHEN et al., 2009) foi criada para fornecer um conjunto de dados de expressões faciais com modelos de origem asiática, visando mitigar vieses culturais em algoritmos de reconhecimento. O banco de dados completo contém um total de 6604 imagens estáticas provenientes de 30 modelos. Essas imagens cobrem várias expressões faciais, como raiva, desgosto,

medo, felicidade, tristeza, surpresa e neutro, fornecendo uma base adequada para avaliar algoritmos de reconhecimento de emoções. A Figura 26 demonstra algumas imagens desta base de dados. Em relação à composição demográfica dos modelos:

- Gênero: O grupo é composto por 14 homens e 16 mulheres.
- Idade: A faixa etária dos modelos varia dos 20 aos 60 anos.
- Perfil: Os participantes foram recrutados em grupos de teatro profissional e escolas de artes cênicas em Taiwan.

Figura 26 – Amostras da base de dados TFEID.



Fonte: (CHEN et al., 2009).

Embora o banco possua milhares de imagens, os autores Chen et al. (2009) escolheram um subconjunto de 406 imagens mais representativas (selecionadas de 12 modelos específicos) que foram submetidas a um rigoroso processo de validação por mais de 200 avaliadores humanos. Além disso, a base disponibiliza as coordenadas de 43 pontos de referência faciais (*landmarks*) para 291 dessas imagens, facilitando pesquisas baseadas em extração de características geométricas.

5.1.1.4 CK+

A base de dados Extended Cohn-Kanade (CK+) (LUCEY et al., 2010) é uma expansão da base CK original, desenvolvida para mitigar limitações na validação de rótulos de emoção e estabelecer métricas comuns de desempenho. Este conjunto expandido é composto por 593 sequências de vídeo provenientes de 123 sujeitos diferentes. Cada sequência registra a transição facial desde um estado neutro até o ápice da expressão solicitada, podendo ser uma das sete expressões: raiva, desgosto, medo, felicidade, tristeza, surpresa e desprezo, além da expressão neutra. Em relação à composição demográfica e técnica dos modelos:

- Gênero: O grupo de participantes é majoritariamente feminino, compreendendo 69% de mulheres.

- Idade: A faixa etária dos modelos varia entre 18 e 50 anos.
- Diversidade Étnica: A herança cultural dos participantes é composta por 81% de Euro-Americanos, 13% de Afro-Americanos e 6% de outros grupos.
- Dados Técnicos: As imagens possuem resoluções de 640x480 ou 640x490 *pixels*, capturadas em tons de cinza de 8 bits ou cores de 24 bits.

Figura 27 – Amostras da base de dados CK+.

Fonte: (LUCEY et al., 2010).

Um diferencial crítico da CK+ é o rigoroso processo de validação dos rótulos de emoção, que utiliza o Sistema de Codificação de Ação Facial (FACS) para garantir que a expressão realizada corresponda à categoria emocional, indo além do que foi meramente solicitado ao modelo. Após esse processo de triagem, 327 das 593 sequências foram confirmadas como representações de uma das sete emoções discretas. Além disso, a base disponibiliza as coordenadas de 68 pontos de referência faciais (*landmarks*) rastreados para cada quadro, facilitando a extração de características geométricas e o alinhamento de modelos de forma.

5.1.2 Base de dígitos

Embora o foco desta tese seja o reconhecimento de expressões faciais, a base de dados MNIST (DENG, 2012) foi incluída em um experimento complementar com o objetivo de avaliar o comportamento do método proposto em um cenário distinto de classificação de imagens. A utilização da MNIST permite analisar o método em um ambiente mais controlado e amplamente utilizado na literatura de aprendizado contínuo, facilitando a comparação conceitual com outros trabalhos.

Nesse contexto, o uso da MNIST não tem como objetivo substituir o domínio principal desta pesquisa, mas verificar se o método proposto mantém seu comportamento quando aplicado a um problema com características diferentes do reconhecimento de emoções. Dessa forma, esse experimento serve como uma avaliação adicional da generalização do método ECgr para cenários de aprendizado incremental em outros domínios de classificação.

Essa base de dados contém 70000 imagens de dígitos escritos à mão. As imagens são divididas em 10 classes, que representam os dígitos de 0 a 9. Todas as imagens são em

preto e branco, com o dígito em preto e o fundo branco. Os dígitos estão centralizados em imagens com tamanho de 28x28 *pixels*. A Figura 28 ilustra alguns exemplos das imagens na base de dados MNIST.

Figura 28 – Amostras da base de dados MNIST.

Fonte: (DENG, 2012).

5.1.3 Pré-processamento

Fez-se necessário realizar o pré-processamento das bases de dados de emoções, para extração dos *frames* individuais dos vídeos, dado que as bases MUG, TFEID e CK+ obtidas para este trabalho estavam no formato de vídeo. As bases de dados JAFFE e MNIST disponibilizam as imagens referentes a cada classe. É importante observar que as quatro bases de dados utilizadas compartilham sete classes de emoções: raiva, desgosto, medo, felicidade, tristeza, surpresa e expressão neutra. Algumas bases de dados, especificamente TFEID e CK+, também incluem a classe desprezo. Contudo, essa classe foi excluída dos experimentos, de modo que nenhuma sequência de vídeo ou imagem correspondente a essa expressão foi considerada no processo de extração ou no conjunto final de dados.

As imagens faciais foram extraídas utilizando o detector de faces da biblioteca Dlib (KING, 2009), juntamente com OpenCV (BRADSKI, 2000). Todas as imagens faciais extraídas foram redimensionadas e normalizadas para o tamanho de 48x48 *pixels*, garantindo que todas as bases de dados utilizadas compartilhassem a mesma resolução espacial. A utilização de um detector de faces reduz significativamente a influência do fundo

das imagens, uma vez que o recorte realizado pelo detector concentra-se principalmente na região facial. Dessa forma, variações de fundo presentes nas diferentes bases de dados tendem a ter impacto limitado no processo de treinamento dos modelos. Também, todas as imagens foram normalizadas para a escala de cinza. O mesmo protocolo de extração de faces foi aplicado para todas as bases de dados. Nem todos os *frames* dos vídeos disponibilizados nas bases de dados foram utilizados para extração das faces, a fim de evitar o desbalanceamento de classes, visto que nos vídeos há uma predominância da expressão facial neutra. A Tabela 6 demonstra os detalhes de cada base de dados utilizada.

A opção por utilizar imagens em escala de cinza e resolução 48x48 está relacionada à eficiência computacional do método proposto. Trabalhar com imagens de menor resolução reduz substancialmente o custo de armazenamento, treinamento dos modelos e geração de imagens pelas WGAN-GPs, o que é particularmente relevante no contexto de aprendizado incremental com múltiplas classes e geradores específicos por classe. Essa padronização também facilita o treinamento da CNN principal e dos mecanismos de QA, garantindo consistência nas representações e minimizando o custo de pré-processamento.

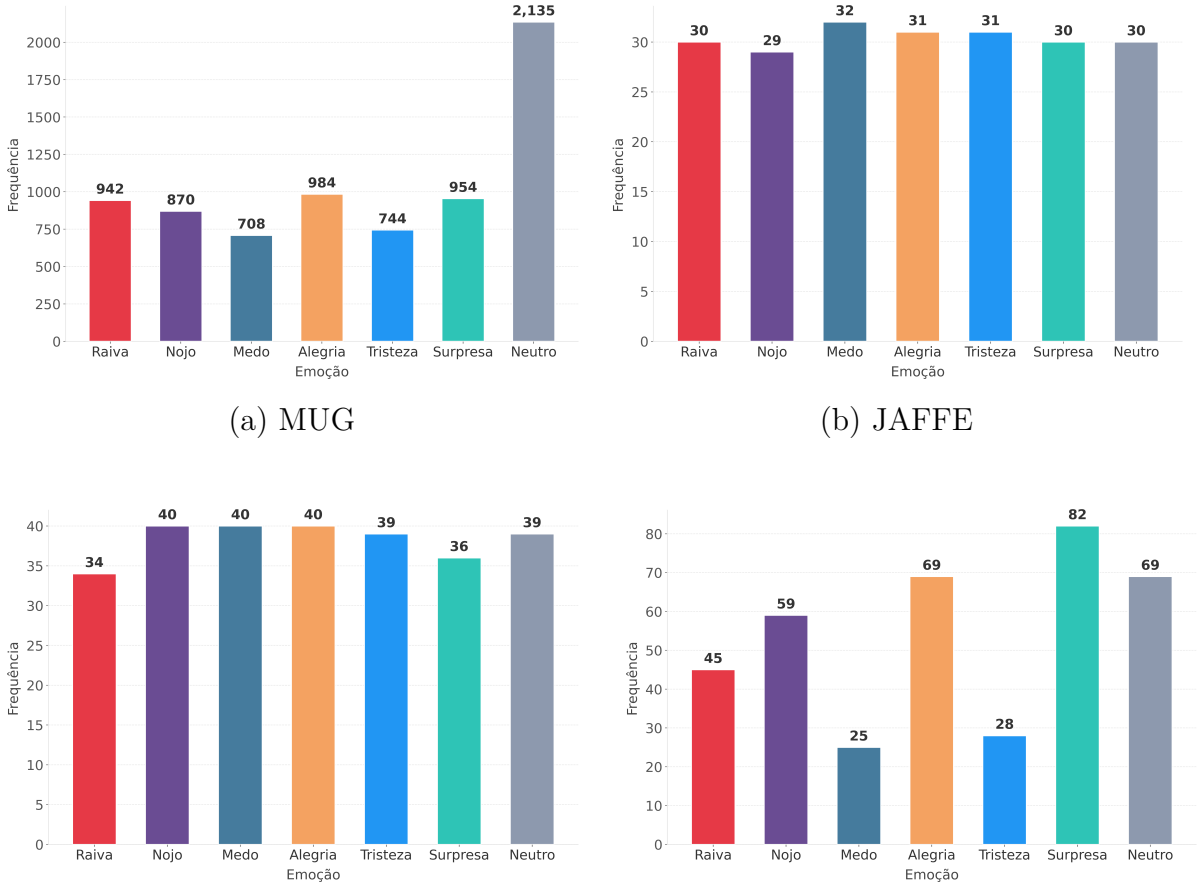
Tabela 6 – Detalhes das bases de dados de expressão facial utilizadas para a avaliação do método proposto.

Base de dados	**Qtd. de imagens**	**Classes**
MUG	7337	Raiva, desgosto, medo, felicidade, tristeza, surpresa e neutra
JAFFE	213	Raiva, desgosto, medo, felicidade, tristeza, surpresa e neutra
TFEID	268	Raiva, desgosto, medo, felicidade, tristeza, surpresa e neutra
CK+	377	Raiva, desgosto, medo, felicidade, tristeza, surpresa e neutra
MNIST	70000	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Fonte: autoria própria.

Após a etapa de extração e normalização das imagens, a distribuição final das classes em cada base de dados pode ser observada na Figura 29. Essa distribuição corresponde às amostras efetivamente utilizadas nos experimentos após o processo de pré-processamento.

Figura 29 – Distribuição das amostras por classe emocional nos conjuntos de dados de emoções utilizados. Cada gráfico apresenta a frequência de cada emoção na respectiva base de dados.



(a) MUG (b) JAFFE

(c) TFEID (d) CK+

Fonte: autoria própria.

5.2 Sobre a qualidade das imagens sintéticas

Nesta seção, apresentam-se uma discussão aprofundada dos aspectos qualitativos dos dados sintéticos gerados.

5.2.1 Geração de imagens sintéticas

A geração de imagens sintéticas desempenha um papel central no método proposto, pois visa mitigar o esquecimento catastrófico ao relembrar amostras de tarefas anteriores por meio da técnica de *pseudo-rehearsal*. Para isso, utilizam-se geradores WGAN-GP treinados separadamente por classe e por tarefa. Esta subseção apresenta os detalhes práticos do treinamento desses geradores, suas configurações e a forma como as imagens foram geradas e avaliadas antes de sua utilização no processo de aprendizado incremental.

Para cada tarefa $\mathcal{D}_k$ e para cada classe $c \in \{1, \ldots, C\}$, um gerador $G_k^{(c)}$ foi treinado de forma independente. Essa escolha permitiu que cada gerador aprendesse

a distribuição específica da respectiva classe, o que contribuiu para maior estabilidade durante o treinamento e maior fidelidade das imagens geradas.

O vetor de entrada $\mathbf{z} \in \mathbb{R}^{128}$ é amostrado de uma distribuição normal padrão $\mathcal{N}(0, \mathbf{I})$. A dimensão foi escolhida empiricamente após experimentos de ablação com diferentes tamanhos de vetor latente, incluindo 64, 128 e 256 dimensões. O valor $d_z = 128$ apresentou o melhor equilíbrio entre qualidade e estabilidade, sendo utilizado nas demais bases.

A arquitetura do gerador projeta esse vetor latente para uma imagem de saída com dimensão $(48 \times 48 \times 1)$. Essa escolha visou balancear fidelidade visual e custo computacional. Conforme já apresentado, a Tabela 4 resume as arquiteturas utilizadas, com cerca de 1,5 milhão de parâmetros no gerador e 4,3 milhões no discriminador. Conforme a recomendação dos autores Radford, Metz e Chintala (2016), o algoritmo de otimização utilizado foi o Adam com taxa de aprendizado $\alpha = 0{,}0002$, $\beta_1 = 0{,}5$, $\beta_2 = 0{,}9$ e penalidade de gradiente $\lambda = 10$. O treinamento foi conduzido por 20000 iterações para cada gerador. Monitora-se a função crítica (Equação 4.1) e inspeções qualitativas de amostras geradas a cada 500 iterações.

Para a geração de imagens sintéticas, adota-se uma proporção de 50% em relação ao tamanho do conjunto de dados da tarefa alvo, definida empiricamente e discutida em maior detalhe na Seção 5.3.2.4, gerando amostras sintéticas em número igual por classe. Salientam-se duas observações importantes: (i) a distribuição das amostras originais (isto é, a proporção entre classes considerando apenas as amostras reais) permanece inalterada, pois os dados reais não são modificados; (ii) ao agregar as amostras sintéticas ao conjunto de treino, a distribuição *a priori* do conjunto combinado (reais + sintéticas) é alterada, no sentido de que classes originalmente minoritárias recebem um incremento absoluto igual ao das majoritárias, resultando em um conjunto combinado mais balanceado. Uma discussão mais aprofundada sobre os impactos de diferentes proporções de dados sintéticos encontra-se na Seção 5.3.2.4.

5.2.2 Análise das imagens sintéticas

Nas Figuras 30, 31 e 32, é possível observar algumas amostras sintéticas geradas para as diferentes bases de dados. No lado esquerdo está a imagem do conjunto de dados original, funcionando como uma referência para as características visuais inerentes à base de dados. Já no lado direito, sete colunas exibem imagens sintéticas geradas para cada classe dentro do conjunto.

Para avaliar a qualidade das imagens geradas pelas GANs, utilizam-se três métricas (ver Seção 2.5 para mais detalhes): (i) CSIM (HAN; KAMBER; PEI, 2012), que avalia similaridade de cosseno entre representações de características; (ii) SSIM (WANG et al., 2004), que mede similaridade estrutural entre pares de imagens; e (iii) FID (HEUSEL et

Figura 30 – Amostras sintéticas da base de dados MUG. À esquerda, um exemplo real da base de dados e à direita, nas 7 colunas, exemplos de imagens sintéticas para cada classe.

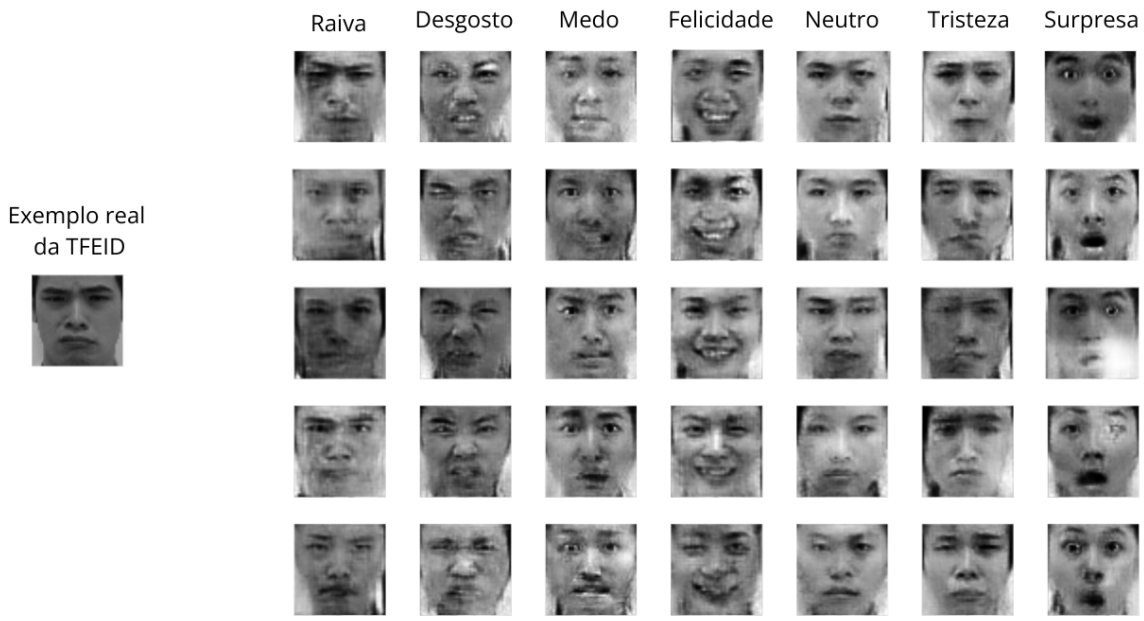
Fonte: autoria própria.

Figura 31 – Amostras sintéticas da base de dados JAFFE. À esquerda, um exemplo real da base de dados e à direita, nas 7 colunas, exemplos de imagens sintéticas para cada classe.

Fonte: autoria própria.

al., 2017), que mede a distância entre distribuições de características de imagens reais e sintéticas. Posteriormente, a qualidade semântica das imagens também é avaliada por

Figura 32 – Amostras sintéticas da base de dados TFEID. À esquerda, um exemplo real da base de dados e à direita, nas 7 colunas, exemplos de imagens sintéticas para cada classe.



Fonte: autoria própria.

meio de uma CNN no módulo de QA supervisionado e CGLO. As métricas CSIM e FID necessitam de um vetor de características para realizar a comparação de similaridade entre as características das imagens reais e sintéticas. Para isso, fez-se o uso da rede InceptionV3 (SZEGEDY et al., 2016), pré-treinada na base de dados ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), cujas entradas esperadas são imagens RGB com 299x299 *pixels*. Assim, para compatibilizar os domínios, cada imagem foi redimensionada para 299x299 e replicada nos três canais (R=G=B), de modo a atender o formato de entrada da InceptionV3. Embora essa adaptação possa introduzir alguma perda de fidelidade no mapeamento original, o objetivo neste trabalho é a comparação relativa entre diferentes variantes do método, todas avaliadas sob o mesmo procedimento. Como tanto as imagens reais quanto as sintéticas passam pela mesma adaptação, o impacto absoluto sobre as métricas é mitigado, permitindo uma análise comparativa válida entre abordagens. Para o cálculo da métrica SSIM, que depende da comparação entre pares de imagens, comparam-se todas as imagens sintéticas com todas as imagens reais da mesma classe, dentro de cada base de dados. A partir dessas comparações cruzadas, calculam-se a média e o desvio padrão dos valores de SSIM por classe e por base de dados. Em seguida, agrupam-se esses valores para obter estatísticas globais para cada um dos conjuntos de dados que tiveram imagens sintéticas geradas (MUG, JAFFE e TFEID).

A Tabela 7 apresenta os valores das métricas FID, CSIM e SSIM para cada classe da base de dados MUG.

Tabela 7 – CSIM, SSIM e FID por classe da base de dados MUG. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Classe	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
Raiva	97,45±8,59	0,81±0,01	0,22±0,01
Desgosto	229,04±15,51	0,70±0,02	0,22±0,00
Medo	107,94±10,89	0,81±0,01	0,21±0,01
Felicidade	120,21±13,32	0,80±0,01	0,18±0,01
Tristeza	91,27±9,16	0,82±0,01	0,22±0,01
Surpresa	101,04±8,67	0,82±0,01	0,22±0,01
Neutra	276,10±18,50	0,66±0,02	0,22±0,01

Fonte: autoria própria.

As Tabelas 7, 8 e 9 demonstram os resultados obtidos a partir do cálculo das métricas por classe para cada uma das bases de dados em que foi feito o uso de imagens sintéticas. A Tabela 10 contém os resultados com as médias de cada base de dados, dado o cálculo dos seus valores por classe. Para melhor compreensão dos valores das métricas, os símbolos ↑ e ↓ indicam, respectivamente, quando uma métrica é melhor se o seu valor é mais alto ou quando é melhor se o seu valor é mais baixo. Quanto mais próximo de 0 for o valor da métrica FID, mais similares são as imagens comparadas. Quanto às métricas CSIM e SSIM, com valores de 0 a 1, quanto mais próximo de 1 mais similares são as imagens. Para o cálculo das métricas, foram utilizadas as imagens sintéticas geradas na última etapa do treinamento contínuo, em que a CNN previamente treinada em cada uma das bases (MUG, JAFFE e TFEID) foi adaptada para o novo conjunto-alvo, CK+. Em cada experimento, foram geradas 27 imagens sintéticas por classe, totalizando 189 imagens por base de dados (o que corresponde a 50% do total do conjunto CK+). Os valores apresentados correspondem à média e ao desvio padrão obtidos a partir de 20 repetições do experimento.

Para a base de dados MUG (Tabela 7), observa-se que as métricas FID e CSIM indicam melhor qualidade para as classes raiva, tristeza e surpresa, enquanto as classes desgosto e neutra apresentam os piores resultados.

Em relação à métrica SSIM, os valores obtidos indicam baixa similaridade estrutural entre imagens sintéticas e reais. Isso sugere que, embora as imagens sintéticas preservem características semânticas das expressões faciais (refletidas pelos valores de FID e CSIM) elas não mantêm com precisão a estrutura local das imagens reais, como alinhamento de traços faciais, textura e contraste. A métrica SSIM é particularmente sensível a pequenas variações de posição e iluminação, o que pode penalizar diferenças mesmo quando o conteúdo perceptivo geral permanece adequado. Isso sugere que a variação intra-classe e o possível desalinhamento entre imagens reais e sintéticas afetam negativamente o SSIM.

A Tabela 8 apresenta os valores das métricas FID, CSIM e SSIM para cada classe da base de dados JAFFE.

Tabela 8 – CSIM, SSIM e FID por classe da base de dados JAFFE. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Classe	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
Raiva	179,06±9,66	0,79±0,01	0,38±0,01
Desgosto	220,33±10,38	0,75±0,01	0,34±0,01
Medo	191,79±10,32	0,78±0,01	0,40±0,01
Felicidade	133,28±5,64	0,82±0,01	0,41±0,01
Tristeza	158,37±8,92	0,80±0,01	0,41±0,01
Surpresa	238,03±14,31	0,74±0,01	0,41±0,01
Neutra	165,74±5,82	0,80±0,00	0,41±0,01

Fonte: autoria própria.

Na base de dados JAFFE (Tabela 8), as métricas FID e CSIM indicam melhor desempenho para as classes felicidade e tristeza, enquanto surpresa e desgosto apresentam os piores resultados de similaridade.

Entre as três bases analisadas, JAFFE apresentou os maiores valores de SSIM, indicando maior similaridade estrutural entre imagens sintéticas e reais. Em particular, as classes felicidade e neutra obtiveram os maiores valores dessa métrica, sugerindo que essas expressões preservam melhor características locais como contornos faciais e padrões de textura. Esse comportamento pode estar relacionado à menor variabilidade intra-classe e à maior padronização das expressões presentes nessa base de dados.

A Tabela 9 apresenta os valores das métricas FID, CSIM e SSIM para cada classe da base de dados TFEID.

Tabela 9 – CSIM, SSIM e FID por classe da base de dados TFEID. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Classe	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
Raiva	208,39±8,24	0,73±0,01	0,05±0,01
Desgosto	212,76±8,84	0,73±0,01	0,02±0,00
Medo	194,08±7,88	0,74±0,01	0,10±0,01
Felicidade	204,68±9,58	0,73±0,01	0,09±0,01
Tristeza	172,95±5,71	0,76±0,01	0,12±0,00
Surpresa	209,40±9,45	0,71±0,01	0,16±0,01
Neutra	164,50±4,46	0,77±0,00	0,11±0,01

Fonte: autoria própria.

Para a base de dados TFEID (Tabela 9), as métricas FID e CSIM indicam melhor qualidade para as classes neutra e tristeza, enquanto desgosto e raiva apresentam os piores resultados.

Os valores de SSIM reforçam essa observação, pois as classes com menor desempenho, especialmente desgosto e raiva, também apresentam baixa similaridade estrutural em relação às imagens reais. Esse resultado indica dificuldades na preservação de padrões locais e texturas faciais, sugerindo maior complexidade na síntese dessas expressões nessa base de dados.

Tabela 10 – CSIM, SSIM e FID por base de dados, a partir da média das classes. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Base de dados	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
MUG	146,15±3,82	0,77±0,00	0,21±0,01
JAFFE	183,80±3,83	0,78±0,00	0,40±0,01
TFEID	195,25±2,29	0,74±0,00	0,10±0,01

Fonte: autoria própria.

Ao analisar conjuntamente os resultados das três bases de dados (MUG, JAFFE e TFEID), observa-se que o desempenho das imagens sintéticas varia entre as bases. A Tabela 10 mostra que, em termos médios, a base MUG apresentou o melhor resultado em FID. Por outro lado, a base JAFFE apresentou o melhor desempenho em SSIM e em CSIM, ainda que com um FID mais alto que o da MUG. Esses resultados sugerem que, embora as imagens sintéticas dessa base possam divergir mais em termos de distribuição estatística global, segundo a análise da métrica FID, elas preservam melhor a estrutura visual e os detalhes locais. A base TFEID, por sua vez, obteve os piores desempenhos médios nas três métricas.

De modo geral, observa-se um padrão em que expressões como tristeza, felicidade e neutra tendem a apresentar os melhores resultados nas métricas, enquanto desgosto e raiva estão entre as mais desafiadoras, independentemente da base. Esse comportamento sugere que certas emoções, por apresentarem expressões mais consistentes e definidas visualmente, são mais facilmente modeladas pelas redes generativas WGAN-GPs. Além disso, o SSIM se mostra sistematicamente inferior nas bases com maior variabilidade intra-classe, como a MUG e, especialmente, a TFEID, o que reforça sua sensibilidade a desalinhamentos espaciais, variações de textura e diferenças de iluminação. Também, observaram-se valores elevados de FID, principalmente para as bases de dados JAFFE e TFEID, que são conjuntos com um número reduzido de imagens, o que dificulta a geração de imagens semanticamente relacionadas as da base de dados original. Ainda assim, o objetivo principal do método não é produzir imagens fotorrealistas, mas gerar

amostras semanticamente coerentes capazes de auxiliar o processo de *pseudo-rehearsal* no aprendizado contínuo.

Esses resultados indicam que, mesmo quando as imagens sintéticas apresentam limitações estruturais segundo algumas métricas, elas preservam características semânticas suficientes para atuar como memória sintética no processo de aprendizado contínuo.

5.2.3 Análise da filtragem supervisionada

Nesta seção, analisam-se a eficácia da etapa de filtragem supervisionada aplicada às imagens geradas pelas WGAN-GPs. Como descrito na Seção 4.2.1, essa filtragem consiste na aplicação de uma CNN previamente treinada de forma incremental até a tarefa T_{k-1}, com o objetivo de selecionar imagens sintéticas condizentes com o conhecimento aprendido até o momento.

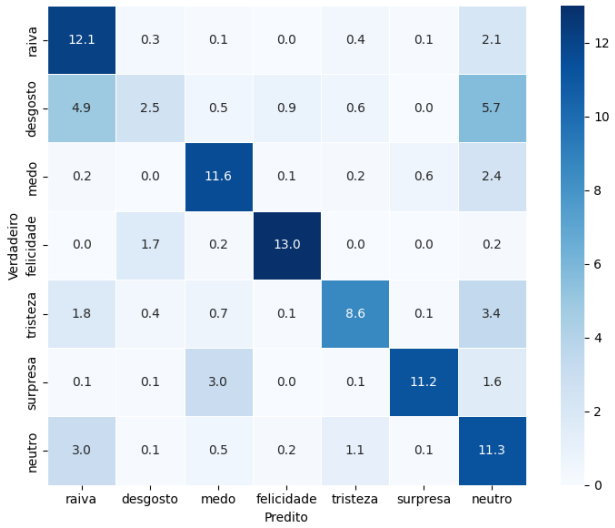
A CNN utilizada na filtragem é a mesma arquitetura empregada no treinamento contínuo ao longo das tarefas, conforme descrito na Seção 4.1.2 e ilustrado na Figura 22. Essa arquitetura, adaptada do trabalho dos autores Tannugi, Britto e Koerich (2019), compreende aproximadamente 19,3 milhões de parâmetros e apresenta uma sequência de camadas convolucionais, normalização, *max-pooling* e camadas densas com *dropout*, culminando em uma camada *softmax* para predição das classes.

A função de filtragem supervisionada $\mathcal{F}(T'_{<k}, \boldsymbol{\theta})$, definida formalmente na Equação 4.4, avalia cada imagem sintética $\mathbf{x}$ gerada pelas WGAN-GPs e a mantém no conjunto final somente se a predição da CNN, $\hat{y} = f_{\boldsymbol{\theta}}(\mathbf{x})$, for igual ao seu rótulo original y. Essa etapa é essencial para garantir a qualidade semântica das amostras utilizadas na construção da memória sintética para a tarefa T_k.

Na prática, essa filtragem supervisionada atua como um mecanismo de controle de qualidade, eliminando amostras inconsistentes (ruidosas ou fora do domínio aprendido). Para ilustrar o comportamento da filtragem supervisionada, realizam-se uma análise específica da transição entre as tarefas MUG para JAFFE, utilizando os 20 *folds* disponíveis. Para cada *fold*, registra-se a matriz de confusão da CNN ao classificar as imagens sintéticas geradas pelas WGAN-GPs, após aplicação da função de filtragem $\mathcal{F}(T'_{<k}, \boldsymbol{\theta})$.

A Figura 33 mostra o *heatmap* médio da matriz de confusão, obtido a partir da média dos 20 *folds*. Os valores representam a frequência média de classificações por classe. A acurácia média final da CNN testada na base sintética da MUG (denominada MUG^A) foi de $65,56\% \pm 0.04$. Observa-se que classes associadas a expressões faciais mais distintivas, como felicidade, medo, surpresa e raiva, apresentaram maior taxa de aceitação correta, refletida pelos valores elevados nas diagonais correspondentes (por exemplo, 13,0 para felicidade, 12,1 para raiva e 11,6 para medo). Essa superior reconhecibilidade está de acordo com estudos dos autores Ekman (1992) e Pantic e Rothkrantz (2000), que apontam essas

Figura 33 – *Heatmap* médio da matriz de confusão resultante do processo de filtragem supervisionada no conjunto sintético MUG^A, gerado por WGAN-GPs treinadas nas classes da base de dados original MUG, obtido ao longo de 20 *folds* durante a adaptação incremental de uma CNN previamente treinada na base MUG e adaptada para a base JAFFE. Os valores representam a frequência média de classificações por classe ao longo dos *folds*, evidenciando a distribuição das predições da CNN.



Fonte: autoria própria.

emoções como dotadas de padrões musculares faciais universalmente reconhecíveis, alta intensidade expressiva e elevada consistência entre culturas. Em contraste, emoções com manifestações mais sutis ou frequentemente envolvidas em padrões emocionais compostos, como tristeza e desgosto, apresentaram maior índice de confusão entre classes, sugerindo menor seletividade da CNN para essas categorias.

A filtragem supervisionada via QA supervisionado teve um impacto na qualidade das amostras sintéticas quando analisadas sob as métricas FID e CSIM. Em todas as bases de dados, observa-se um aumento nos valores de FID, indicando que, segundo a métrica FID (na qual valores menores indicam maior similaridade), houve aumento da distância estatística entre os domínios sintético e real após o filtro. A Tabela 11 demonstra os resultados médios por base de dados. Ao comparar com as métricas obtidas sem a etapa de QA (Tabela 10), percebe-se que o FID nas bases de dados MUG e TFEID aumentaram, sugerindo uma perda de representatividade no conjunto filtrado. A métrica CSIM apresentou queda nas bases de dados JAFFE e TFEID.

A métrica SSIM apresentou um aumento em todas as bases de dados, indicando uma preservação estrutural mais robusta após a filtragem. A base MUG teve um aumento de 0,21 para 0,29, a JAFFE de 0,39 para 0,43 e a TFEID de 0,09 para 0,18. Esses resultados sugerem que o processo de QA supervisionado pode ter contribuído para melhorar a preservação da estrutura local das imagens sintéticas.

Tabela 11 – CSIM, SSIM e FID por base de dados, a partir da média das classes após o filtro de QA supervisionado. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Base de dados	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
MUG	172,39±36,08	0,74±0,03	0,29±0,02
JAFFE	194,56±25,50	0,75±0,02	0,43±0,03
TFEID	234,06±23,97	0,69±0,02	0,18±0,05

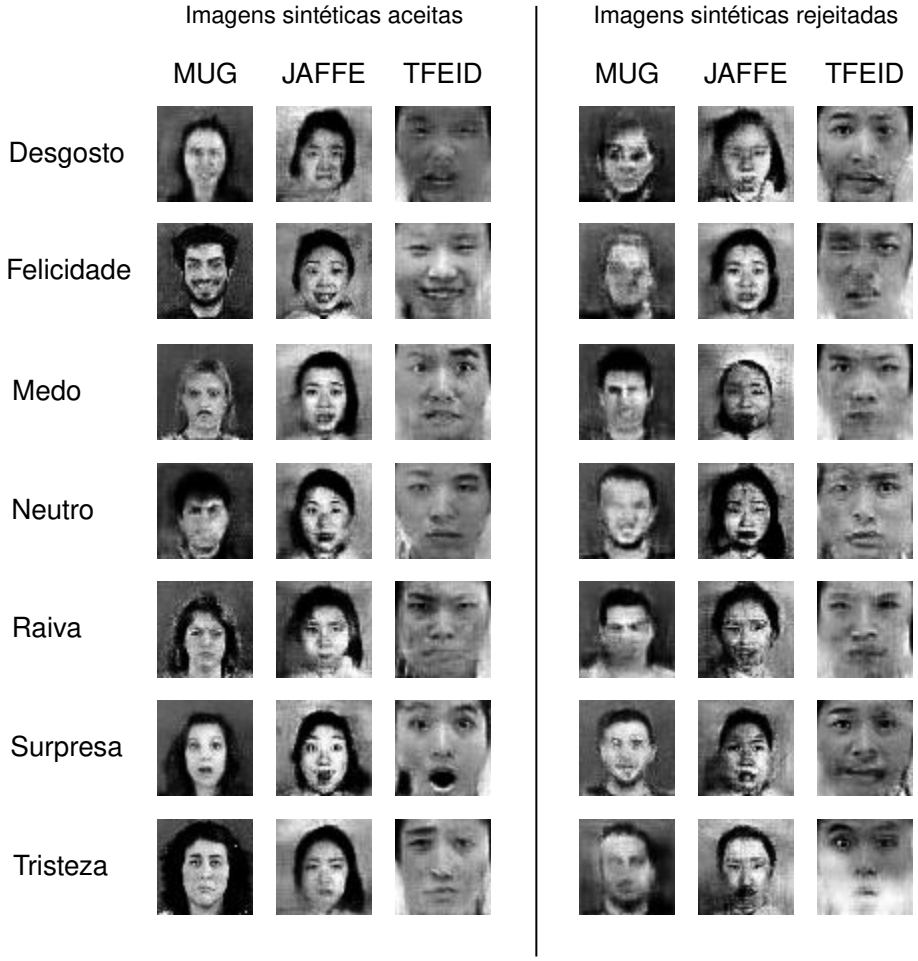
Fonte: autoria própria.

A Figura 34 ilustra amostras sintéticas das bases MUG, JAFFE e TFEID que foram aceitas ou rejeitadas pela CNN do módulo de QA supervisionado. Esse módulo foi treinado incrementalmente utilizando os conjuntos sintéticos MUG^A, JAFFE^A, TFEID^A e a base original CK+. As imagens rejeitadas, mostradas à direita de cada grupo, correspondem a amostras que a CNN não conseguiu classificar corretamente, indicando que a emoção representada não foi suficientemente distinguível para o modelo. Sob a premissa de que imagens com baixa confiabilidade não contribuem positivamente para a adaptação do modelo, essas amostras são descartadas e não utilizadas no processo de retreinamento.

5.2.4 Análise da filtragem não supervisionada

Na implementação do método não supervisionado baseado em agrupamento, foram utilizados seis extratores de características amplamente aplicados em tarefas de reconhecimento facial (DeepID, Facenet512, SFace, GhostFaceNet, OpenFace e Facenet) e quatro extratores genéricos pré-treinados em tarefas de classificação geral de imagens (ConvNeXt, ResNet50, EfficientNetV2 e InceptionV3). Os modelos foram empregados conforme suas versões disponíveis nas bibliotecas DeepFace ou Keras. Todos os extratores foram aplicados diretamente sobre as imagens sintéticas previamente normalizadas, produzindo vetores de características com dimensionalidades específicas conforme apresentado na Tabela 12.

Figura 34 – Amostras aceitas (à esquerda) e rejeitadas (à direita) das bases MUG, JAFFE e TFEID, conforme identificadas pela CNN do algoritmo de QA supervisionado, treinada incrementalmente nos conjuntos sintéticos MUG^A, JAFFE^A, TFEID^A e na base original CK+. As bases com sufixo A correspondem a imagens geradas por WGAN-GPs treinadas em suas respectivas bases originais.



Fonte: autoria própria.

Tabela 12 – Extratores de características utilizados na etapa de agrupamento da filtragem não supervisionada, com suas respectivas dimensionalidades e especializações (facial ou genérica)

Extrator	**Dimensão do *embedding***	**Especialização**
ConvNeXt	1536	Genérico
DeepID	160	Face
Facenet512	512	Face
SFace	128	Face
ResNet50	2048	Genérico
EfficientNetV2	1280	Genérico
GhostFaceNet	512	Face
OpenFace	128	Face
Facenet	128	Face
InceptionV3	2048	Genérico

Fonte: autoria própria.

Para cada conjunto de *embeddings* gerado por um extrator E_e, aplicou-se o algoritmo K-Means com $r = C$ *clusters*, onde C é o número de classes reais da tarefa corrente. Essa escolha não pressupõe que as classes formem estruturas perfeitamente separáveis no espaço de características, mas foi adotada como uma aproximação prática para permitir a comparação entre os agrupamentos obtidos e os rótulos das classes. O critério de validação interna adotado para avaliar a qualidade do agrupamento foi o coeficiente de silhueta normalizado, calculado individualmente para cada instância e para cada extrator. Para que uma instância fosse considerada de qualidade suficiente, era necessário que ao menos um dos extratores atribuísse um rótulo de *cluster* que coincidisse com o rótulo verdadeiro da instância, independentemente do valor da silhueta. O valor de silhueta correspondente foi então utilizado como medida de confiança para ponderar a importância da instância. Todas as operações foram realizadas com suporte vetorizado utilizando a biblioteca *scikit-learn* para o K-Means e para o cálculo dos coeficientes de silhueta.

Com base na comparação entre os resultados apresentados na Tabela 10 (sem filtragem) e na Tabela 13 (após o filtro de QA com clusterização não supervisionada), é possível observar impactos distintos nas métricas de qualidade das imagens sintéticas geradas.

Na base MUG, observa-se um aumento na métrica FID, indicando uma piora na qualidade estatística global das imagens em relação à distribuição das imagens reais. Contudo, o valor de SSIM subiu, sugerindo uma melhor preservação da estrutura local das imagens sintéticas. A base JAFFE apresentou um padrão semelhante: o FID aumentou, assim como o SSIM, que apresentou o maior valor entre todas as bases após a filtragem, sugerindo que a filtragem via agrupamento favoreceu a preservação de detalhes estruturais em imagens dessa base de dados. Na base TFEID, observa-se o mesmo comportamento, em que o FID e SSIM aumentaram, indicando que o método de clusterização pode ter removido amostras degradadas, preservando apenas imagens sintéticas estruturalmente mais próximas das reais.

Tabela 13 – CSIM, SSIM e FID por base de dados, a partir da média das classes após o filtro de QA não supervisionado com clusterização. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Base de dados	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
MUG	154,64±62,39	0,77±0,06	0,29±0,02
JAFFE	191,42±31,46	0,78±0,03	0,42±0,03
TFEID	202,31±16,16	0,74±0,02	0,17±0,04

Fonte: autoria própria.

5.2.5 Análise da filtragem por otimização do espaço latente

A implementação do método *Coefficient-guided Latent Optimization* (CGLO) baseia-se em três etapas principais: (1) geração inicial de vetores latentes, (2) filtragem supervisionada por uma CNN previamente treinada, e (3) amostragem de novas regiões do espaço latente via combinações convexas.

Inicialmente, amostram-se vetores latentes $\mathbf{z} \in \mathbb{R}^d$ a partir de uma distribuição normal padrão, isto é, $\mathbf{z} \sim \mathcal{N}(0, I)$, onde $d = 128$ corresponde à dimensionalidade do espaço latente da WGAN-GP. A quantidade de vetores amostrados nessa etapa inicial é dinâmica, variando proporcionalmente ao tamanho do conjunto de dados da nova tarefa T_k.

As imagens correspondentes são geradas por meio do gerador G da WGAN-GP, ou seja, $\mathbf{x} = G(\mathbf{z})$, e avaliadas por uma CNN supervisionada $f_{\theta_{k-1}}$, treinada previamente nas tarefas $T_{<k}$. Apenas as imagens classificadas corretamente por $f_{\theta_{k-1}}$, isto é, para as quais $\arg\max(f_{\theta_{k-1}}(\mathbf{x})) = y(\mathbf{x})$ são mantidas. Os vetores latentes dessas imagens compõem o conjunto filtrado $\mathcal{Z}_{\text{filt}}$, utilizado na próxima etapa.

Na etapa de otimização latente, o método busca explorar regiões promissoras do espaço latente assumindo que vetores latentes associados a amostras consideradas válidas pelo processo de QA encontram-se relativamente próximos entre si. Para cada classe c presente nas tarefas T_{k-1}, calcula-se um invólucro convexo (*convex hull*) sobre os vetores latentes $\mathcal{Z}^c_{\text{filt}} \subset \mathcal{Z}_{\text{filt}}$, que correspondem às imagens dessa classe corretamente classificadas pela CNN. Este processo resulta em um polígono convexo no espaço latente associado àquela classe, definido de forma determinística a partir dos vetores filtrados.

Novos vetores latentes $\mathbf{z}^c_{\text{opt}}$ são então gerados aleatoriamente dentro desse invólucro convexo, utilizando combinações convexas dos vértices que definem o polígono. A amostragem dos coeficientes convexos $\boldsymbol{\alpha}$ é realizada a partir de uma distribuição de Dirichlet simétrica, de forma que $\mathbf{z}^c_{\text{opt}}$ respeite a Equação 4.9. Por definição, qualquer ponto interno ao invólucro pode ser representado como uma combinação convexa de até $d + 1$ vértices, com $d = 128$ neste caso.

As imagens correspondentes, $\mathbf{x}^c_{\text{opt}} = G(\mathbf{z}^c_{\text{opt}})$, são então avaliadas novamente pela CNN $f_{\theta_{k-1}}$, utilizando o mesmo critério de filtragem definido anteriormente. Apenas as imagens corretamente classificadas são incorporadas ao conjunto final sintético da classe c. Esse processo é repetido para todas as classes de todas as tarefas T_{k-1}, resultando em um conjunto $\mathcal{Z}_{\text{opt}}$ organizado por classe e tarefa, que compõe a memória sintética utilizada para o aprendizado contínuo.

Com base na comparação entre os resultados obtidos antes e depois da aplicação do filtro CGLO (Tabela 10 e Tabela 14, respectivamente), nota-se que a aplicação do filtro CGLO não resultou em melhorias nas métricas FID e CSIM. Pelo contrário, observou-se

um aumento nos valores de FID e uma redução nas médias de CSIM em todas as bases de dados, indicando uma degradação na qualidade global das imagens sintéticas do ponto de vista da distribuição estatística (FID) e da similaridade perceptual de alto nível (CSIM).

Tabela 14 – CSIM, SSIM e FID por base de dados, a partir da média das classes após o filtro de QA CGLO. FID-INC e CSIM-INC correspondem às métricas FID e CSIM utilizando a rede InceptionV3 para extração de características.

Base de dados	**FID-INC ↓**	**CSIM-INC ↑**	**SSIM ↑**
MUG	166,44±27,72	0,74±0,03	0,14±0,02
JAFFE	183,63±22,65	0,75±0,02	0,25±0,02
TFEID	212,70±14,55	0,71±0,01	0,02±0,02

Fonte: autoria própria.

A base de dados MUG, que anteriormente apresentava o melhor resultado em FID, sofreu uma deterioração com o uso do filtro CGLO. Na base de dados TFEID, o FID aumentou e o CSIM diminuiu. A base JAFFE apresentou queda na similaridade perceptual (CSIM). Esses resultados sugerem que, embora o CGLO seja uma técnica orientada à refinamento qualitativo das imagens, seu impacto global nas distribuições estatísticas pode ser adverso quando avaliado sob o ponto de vista das métricas FID e CSIM. Além disso, observa-se que o desvio padrão das métricas aumentou após a aplicação do CGLO, especialmente na base de dados MUG (FID: de ±3,82 para ±27,72), o que indica maior variabilidade entre as repetições dos experimentos.

A análise dos novos valores de SSIM após a aplicação do filtro CGLO revela uma queda na preservação da estrutura local das imagens sintéticas. Observa-se que todas as bases apresentaram reduções. Esses resultados indicam que, embora o CGLO possa melhorar aspectos perceptuais mais globais, ele tende a comprometer os detalhes estruturais das imagens, possivelmente por introduzir distorções e suavizações durante a otimização no espaço latente que não estão presentes nas bases de dados originais.

A Figura 35 apresenta amostras sintéticas da base de dados JAFFE utilizadas em uma das etapas do treinamento com o algoritmo CGLO. À esquerda, estão as imagens sintéticas que serviram como referência, ou seja, imagens geradas pelas WGAN-GPs, treinadas em cada classe da base de dados original JAFFE, que foram corretamente classificadas pela CNN e constituíram parte do *convex hull* do espaço latente da WGAN-GP; à direita, as correspondentes imagens sintéticas geradas a partir dessas referências que pertencem ao invólucro convexo do espaço latente.

5.2.6 Discussão

Para facilitar a comparação entre todas as abordagens avaliadas, a Tabela 15 resume os valores médios das métricas para cada base de dados. Assim, a partir da análise

Figura 35 – Amostras sintéticas corretamente classificadas pela CNN usadas como referências (à esquerda) para gerar as novas imagens sintéticas (à direita) conforme o algoritmo CGLO, do conjunto de dados JAFFE.

Fonte: autoria própria.

dos resultados obtidos nas métricas FID, CSIM e SSIM, com os resultados agrupados de todas as abordagens, é possível observar que os melhores resultados para FID e CSIM foram alcançados a partir do método ECgr (sem filtro de qualidade). Ao analisar a métrica SSIM, a aplicação da filtragem apresenta ganhos na similaridade estrutural das imagens sintéticas, em que a abordagem ECgr+QA (com filtro supervisionado) obteve os melhores resultados para as bases JAFFE e TFEID, enquanto a abordagem ECgr+*cluster* (com filtro não supervisionado) teve o melhor desempenho na base MUG.

Na Figura 36, são apresentados os valores de CSIM, utilizando a rede InceptionV3 para extração de características e as suas respectivas imagens sintéticas, no intuito de ilustrar imagens sintéticas com valores "altos" e "baixos" de similaridade de cosseno em relação ao vetor médio das *features* reais de cada classe. A primeira coluna exibe uma imagem real de referência de cada classe, enquanto as colunas Alto e Baixo mostram, respectivamente, a imagem sintética com maior e menor similaridade de cosseno. Os valores numéricos correspondem ao valor de CSIM obtido para cada imagem sintética em relação à média das *features* reais da classe. É possível observar que, para a maioria das classes,

Tabela 15 – Resultados das métricas CSIM, SSIM e FID por base de dados, calculadas a partir da média das classes, considerando todas as abordagens: sem filtragem (ECgr), com filtro supervisionado (ECgr+QA), com filtro não supervisionado (ECgr+*cluster*) e com filtro por otimização do espaço latente (ECgr+CGLO). As variantes FID-INC e CSIM-INC utilizam a rede InceptionV3 para extração de características. As setas ↑ e ↓ indicam se a métrica é melhor quanto maior ou menor, respectivamente.

Métrica	**Base de dados**	**ECgr**	**ECgr+QA**	**ECgr+*cluster***	**ECgr+CGLO**
FID-INC ↓	MUG	**146,15±3,82**	172,39±36,08	154,64±62,39	166,44±27,72
	JAFFE	183,80±3,83	194,56±25,50	191,42±31,46	**183,63±22,65**
	TFEID	**195,25±2,29**	234,06±23,97	202,31±16,16	212,70±14,55
CSIM-INC ↑	MUG	**0,77±0,00**	0,74±0,03	0,77±0,06	0,74±0,03
	JAFFE	**0,78±0,00**	0,75±0,02	0,78±0,03	0,75±0,02
	TFEID	**0,74±0,00**	0,69±0,02	0,74±0,02	0,71±0,01
SSIM ↑	MUG	0,21±0,01	0,29±0,02	**0,29±0,02**	0,14±0,02
	JAFFE	0,40±0,01	**0,43±0,03**	0,42±0,03	0,25±0,02
	TFEID	0,10±0,01	**0,18±0,05**	0,17±0,04	0,02±0,02

Fonte: autoria própria.

as imagens sintéticas com maior similaridade de cosseno tendem a preservar melhor as características visuais e emocionais das expressões faciais, enquanto aquelas com menor similaridade apresentam distorções ou características menos reconhecíveis.

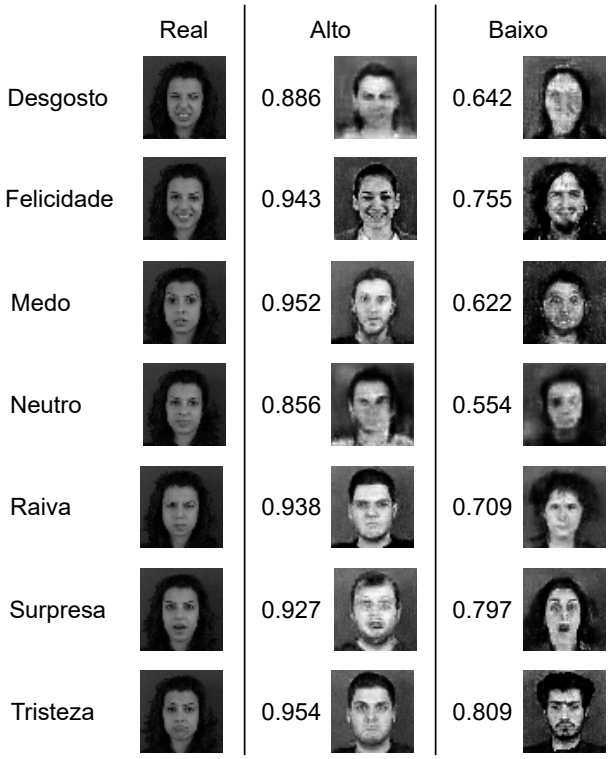
5.3 Sobre o aprendizado contínuo

Nesta seção, discutem-se os principais resultados observados a partir dos testes conduzidos com conjuntos de dados de expressões faciais, utilizando a combinação dos diferentes métodos discutidos neste estudo.

5.3.1 Parâmetros gerais

Em todos os experimentos, os modelos foram treinados por até 200 épocas, com avaliação a cada época no conjunto de validação. A cada repetição, o modelo com melhor desempenho validado foi salvo (*early stopping*), evitando o sobreajuste e garantindo que apenas o ponto ótimo de generalização fosse considerado para teste. O processo foi repetido 20 vezes, utilizando diferentes sementes aleatórias em cada execução, com o objetivo de capturar a variabilidade estatística dos resultados. Cada semente controla os principais aspectos estocásticos do experimento, incluindo a inicialização dos pesos da rede (via inicializador Glorot Uniform), a divisão estratificada entre treino, validação e teste, o comportamento das camadas de *dropout* e o embaralhamento dos dados dentro de cada partição. Ainda assim, pode haver variabilidade residual associada a fatores não

Figura 36 – Exemplos de imagens sintéticas da base MUG avaliadas pela métrica CSIM, utilizando *features* extraídas pela rede InceptionV3. Para cada classe, a primeira coluna mostra uma imagem real de referência, enquanto as colunas Alto e Baixo exibem, respectivamente, a imagem sintética com maior e menor similaridade de cosseno em relação ao vetor médio das *features* reais da classe. Os valores numéricos correspondem ao valor de CSIM obtido.



Fonte: autoria própria.

determinísticos da execução. O otimizador utilizado em todas as etapas foi o Adam, com taxa de aprendizado fixa de 0,001. Essa configuração garante comparações justas entre os métodos avaliados, isolando seus efeitos em cenários de aprendizado contínuo.

É importante ressaltar que todas as partições das bases de dados em treino, validação e teste foram feitas utilizando uma semente fixa, para todas as rodadas de treinamento. Dessa forma, foi possível garantir, por exemplo, que as imagens utilizadas uma vez para treinamento, sempre serão as mesmas para treinamento daquela base de dados em específico. Após a divisão, as bases de dados sofrem um embaralhamento, mas somente dentro das suas partições, com o objetivo único de aleatorizar a ordem das imagens no treinamento, validação e teste. No conjunto de treinamento, essa aleatorização tem como objetivo evitar viés na ordem de apresentação das amostras ao modelo. Nos conjuntos de validação e teste, o embaralhamento não impacta os resultados, sendo utilizado apenas para manter consistência no carregamento dos dados.

Mantém-se, nesta etapa, a mesma estratégia de geração de dados sintéticos citada anteriormente (Seção 5.2.1), estabelecendo uma proporção de 50% em relação ao conjunto de dados da tarefa alvo. As amostras reais permanecem inalteradas, garantindo que a proporção entre classes do conjunto original seja preservada. Por outro lado, a inclusão de amostras sintéticas, em quantidade igual por classe, modifica a distribuição *a priori* do conjunto combinado, resultando em um cenário mais equilibrado entre classes originalmente desbalanceadas.

Essa escolha implica uma modificação controlada na distribuição a priori das classes, o que, em cenários desbalanceados, atua como uma forma implícita de balanceamento de dados. Embora essa alteração possa deslocar ligeiramente a distribuição estatística do conjunto combinado, ela foi intencional, pois o objetivo principal é preservar representações de classes antigas e reduzir o viés do modelo em favor das classes mais recentes ou mais numerosas. Em outras palavras, o acréscimo uniforme de amostras sintéticas não visa reproduzir fielmente a distribuição original, mas compensar o desequilíbrio entre tarefas e reforçar a diversidade de padrões relembrados pelo modelo.

Do ponto de vista do aprendizado incremental, esse balanceamento artificial tende a favorecer a estabilidade do modelo, pois reduz a tendência de degradação do desempenho em classes com menor representatividade (KIM; JEONG; KIM, 2020; HE, 2024). Em cenários desbalanceados, a geração uniforme de amostras sintéticas contribui para que o modelo mantenha sensibilidade a classes menos frequentes, evitando que o processo de atualização sucessiva das tarefas acentue o esquecimento seletivo. Por outro lado, reconhece-se que essa estratégia pode introduzir um leve viés em direção às classes com maior número de amostras geradas, especialmente quando a qualidade das sínteses varia entre classes. No presente trabalho, os resultados empíricos indicam que o ganho em retenção de conhecimento supera esse efeito, resultando em maior consistência das métricas de desempenho ao longo das etapas de treinamento. Na Seção 5.3.2, é realizada uma discussão mais detalhada a respeito dos resultados do aprendizado contínuo e o esquecimento catastrófico.

5.3.2 Reconhecimento de emoções

Inicialmente, treina-se uma CNN no conjunto de dados MUG. Em seguida, adapta-se essa CNN para o processo de aprendizado contínuo em outras bases de dados sequencialmente, nesta ordem: JAFFE, TFEID e CK+. Esta sequência foi definida com o propósito de expor o modelo a um conjunto de dados mais desafiador e variado (MUG), forçando-o a aprender representações mais gerais desde o início. Também, a base de dados MUG, dentre as avaliadas neste trabalho, é a que possui mais amostras. Em seguida, os conjuntos mais controlados e menores (JAFFE, TFEID e CK+) serviram para consolidar o conhecimento aprendido.

Cada conjunto de dados foi dividido em 80% para treino, 10% para validação e

10% para teste, respeitando a distribuição estratificada por classe. A Tabela 16 apresenta os números absolutos de amostras utilizados em cada etapa para fins de reprodutibilidade. Inicialmente, uma fração dos dados reais foi reservada para o teste. Em seguida, os dados restantes foram utilizados para compor o treinamento e validação. Quando uma base de dados sintética adicional (gerada pelos modelos generativos) estava disponível, ela foi incorporada ao conjunto de treinamento real antes da divisão entre treino e validação. As bases de dados sintéticas não foram utilizadas durante o teste, somente durante o treinamento e validação.

Tabela 16 – Número de amostras em cada subconjunto (treinamento, validação e teste) para os conjuntos de dados MUG, JAFFE, TFEID e CK+. A divisão segue a proporção de 80% para treino, 10% para validação e 10% para teste.

Base de dados	**Treinamento**	**Validação**	**Teste**
MUG	5868	735	734
JAFFE	169	22	22
TFEID	214	27	27
CK+	301	38	38

Fonte: autoria própria.

Para cada processo de treinamento para uma nova base de dados, replica-se o treinamento 20 vezes para obtermos uma média entre os resultados. Para métodos que envolvem a geração de imagens, os conjuntos sintéticos são diferentes para cada replicação de retreinamento da CNN, ou seja, a geração das imagens é feita a cada rodada, de forma que as imagens sintéticas sejam sempre diferentes.

A Subseção 5.3.2.2 traz um detalhamento sobre cada conjunto de teste utilizando os diferentes métodos de QA: QA supervisionado baseado em CNN (Seção 4.2.1); QA não-supervisionado com *clusters* (Seção 4.2.2); e QA baseado em otimização no espaço latente da WGAN-GP (Seção 4.2.3).

5.3.2.1 Protocolos de treinamento utilizados

Para contextualizar os resultados obtidos, são considerados três protocolos de treinamento utilizados como referência em cenários de aprendizado contínuo:

- *Baseline*: corresponde à acurácia obtida por uma CNN treinada exclusivamente no conjunto de dados fonte e avaliada diretamente no conjunto de testes da tarefa alvo, sem qualquer adaptação adicional.
- *Joint*: corresponde à acurácia obtida por uma CNN treinada utilizando simultaneamente os dados do conjunto fonte e do conjunto alvo. Esse cenário representa um limite superior idealizado, pois assume acesso simultâneo a todos os dados.

- *Fine-tuning*: corresponde à acurácia obtida por uma CNN inicialmente treinada no conjunto fonte e posteriormente adaptada ao conjunto alvo, restringindo o re-treinamento às camadas finais da rede (especificamente, a última camada *fully connected*).

Esses protocolos são utilizados como referência para comparação com os métodos baseados em *generative replay* avaliados neste trabalho. Além dos protocolos de referência, também foram avaliadas diferentes variantes do método proposto. O método ECgr foi analisado isoladamente e em combinação com os mecanismos de avaliação de qualidade (QA) das imagens sintéticas: QA supervisionado baseado em CNN (ECgr+QA), QA não supervisionado baseado em agrupamento (ECgr+*cluster*) e QA baseado em otimização no espaço latente (ECgr+CGLO).

Adicionalmente, investigou-se o impacto do uso de uma função de custo ponderada para as imagens sintéticas, resultando nas variantes ECgr+wQA, ECgr+w*cluster* e ECgr+wCGLO. Essas variantes permitem avaliar se a ponderação das imagens sintéticas de acordo com sua qualidade estimada influencia o desempenho do modelo em comparação com o treinamento sem essa ponderação.

5.3.2.2 Avaliação do método ECgr

Os resultados apresentados nesta seção são relacionados ao uso do método proposto ECgr (Seção 4.1) e a combinação deste método com os métodos de QA (Seções 4.2.1, 4.2.2 e 4.2.3) e de otimização da função de perda com peso ponderado (Seção 4.3.1).

Nas Tabelas 20, 25 e 31, as abordagens *baseline*, *joint* e *fine-tuning* correspondem aos protocolos experimentais descritos na Subseção 5.3.2.1. Além disso, os métodos ECgr e os métodos de QA foram avaliados: separadamente (ECgr) e combinados (ECgr+QA, ECgr+*cluster*, ECgr+CGLO), com o objetivo de investigar o impacto da filtragem de imagens sintéticas no aprendizado contínuo. Também se avalia o uso de pesos na função de custo (ECgr+wQA, ECgr+w*cluster*, ECgr+wCGLO), a fim de verificar se a ponderação das imagens sintéticas influencia o desempenho em comparação com o treinamento sem essa técnica.

A fim de facilitar o entendimento dos resultados apresentados, utilizam-se a partir deste momento A_F para denotar a acurácia na base de dados fonte, A_V para denotar a acurácia na base de dados alvo e $\bar{A}$ para denotar a média aritmética $\frac{A_F+A_V}{2}$ entre as acurácias fonte e alvo, como métrica global do desempenho incremental.

5.3.2.2.1 MUG para JAFFE

Nesta etapa, a base de dados JAFFE é considerada como a base-alvo atual. Os métodos são avaliados a partir de diferentes estratégias de adaptação. Para os métodos

baseados em ECgr, utilizam-se versões sintéticas da base MUG, geradas pelas WGAN-GPs treinadas especificamente para esse domínio. Essa versão sintética é indicada por MUG^A. O tamanho da base MUG^A é fixado em 50% da base-alvo JAFFE, sendo combinada com esta durante o processo de adaptação. Dessa forma, nessa etapa foram geradas 107 imagens sintéticas da base MUG, o que corresponde a 50% das 213 imagens reais da base JAFFE (arredondamento para cima), com distribuição igual entre as classes, resultando em cinco classes com 15 imagens e duas classes com 16, conforme mostra a Tabela 17. Para o método de *fine-tuning*, apenas JAFFE é usada como base-alvo, e para o *joint*, é feita a combinação direta de MUG original e JAFFE.

Tabela 17 – Número de imagens sintéticas geradas por classe na base de dados MUG^A, utilizando 50% do tamanho da base-alvo JAFFE, para a adaptação da CNN treinada na base MUG para a JAFFE.

Base de dados	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**	**Total**
MUG^A	15	15	15	15	15	16	16	107

Fonte: autoria própria.

A Tabela 18 mostra o número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada na base MUG para a base-alvo JAFFE. A base de testes é composta somente por imagens da base-alvo JAFFE. Nos métodos com ECgr que utilizam imagens sintéticas, a base de dados sintética MUG^A é combinada com a base JAFFE durante o treinamento e validação. Nos métodos com QA supervisionado (ECgr+QA e ECgr+wQA) e não-supervisionado (ECgr+*cluster*, ECgr+w*cluster*), os números de imagens para treino e validação são apresentados como médias e desvios padrão, considerando as 20 repetições do experimento. Essa variação pode acontecer devido ao processo de geração e filtragem das imagens sintéticas. Nos métodos ECgr+CGLO e ECgr+wCGLO, apesar de haver o processo de filtragem, são geradas novas imagens até atingir o limite de 50% da base-alvo, conforme definido na Seção 4.2.3, resultando em números fixos de imagens para treino e validação.

Antes de analisar os resultados do aprendizado contínuo, é importante considerar os números de imagens aceitas e recusadas no processo de filtragem dos métodos generativos com QA (ECgr+QA, ECgr+wQA, ECgr+*cluster*, ECgr+w*cluster*, ECgr+CGLO e ECgr+wCGLO). A Tabela 19 apresenta o número médio de imagens aceitas e recusadas para cada classe no processo de filtragem das imagens sintéticas da base de dados MUG, geradas com as WGAN-GPs, na adaptação da CNN treinada na base MUG para a JAFFE, durante 20 repetições.

Percebe-se que os métodos com filtragem via CNN (ECgr+QA e ECgr+CGLO e suas respectivas versões ponderadas) tendem a recusar as imagens da classe Desgosto e Tristeza. Isso pode evidenciar que a geração dessas classes específicas é mais desafiadora para as WGAN-GPs, resultando em imagens sintéticas menos representativas. Já os

Tabela 18 – Número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada na base MUG para a base-alvo JAFFE. Nos métodos baseados em ECgr, a base de dados sintética MUG^A é combinada com a base JAFFE durante o treinamento e validação. No *fine-tuning*, apenas JAFFE é usada como base-alvo, e no *joint*, é feita a combinação direta da MUG original com a JAFFE.

Método	**Treino**	**Validação**	**Teste**	**Total**
Fine-tuning	169	22	22	213
Joint	6038	756	22	6816
ECgr	264	34	22	320
ECgr+QA	231,95±3,43	29,60±0,49	22	283,55
ECgr+wQA	213,00±4,76	29,55±0,59	22	264,55
ECgr+*cluster*	242,55±2,91	31,00±0,45	22	295,55
ECgr+w*cluster*	240,75±5,26	30,65±0,65	22	293,40
ECgr+CGLO	264	34	22	320
ECgr+wCGLO	264	34	22	320

Fonte: autoria própria.

métodos de clusterização tendem a recusar a emoção Medo, sugerindo que as imagens geradas para essa classe podem não se agrupar de maneira coesa, dificultando a identificação de padrões claros.

Tabela 19 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados MUG^A, geradas com as WGAN-GPs, na adaptação da CNN treinada na base MUG para a JAFFE, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+ QA	Aceitas	12,40±1,64	3,40±2,54	11,50±1,76	13,20±1,36	8,45±2,09	10,20±1,58	11,40±1,79
	Recusadas	2,60±1,64	11,65±2,60	3,50±1,76	1,80±1,36	6,55±2,09	5,80±1,58	4,60±1,79
ECgr+ wQA	Aceitas	12,20±1,15	2,70±1,34	11,30±1,87	13,55±1,10	7,95±1,90	10,30±2,11	11,55±2,11
	Recusadas	2,80±1,15	12,30±1,34	3,70±1,87	1,45±1,10	7,05±1,90	5,70±2,11	4,45±2,11
ECgr+ *cluster*	Aceitas	10,40±1,43	12,90±1,48	8,90±1,94	12,85±1,53	11,10±1,62	10,55±1,93	15,85±0,49
	Recusadas	4,60±1,43	2,10±1,48	6,10±1,94	2,15±1,53	3,90±1,62	5,45±1,93	0,15±0,49
ECgr+ w*cluster*	Aceitas	9,65±2,06	12,60±1,54	8,90±2,05	12,30±1,56	10,35±2,16	10,60±1,85	16,00±0,00
	Recusadas	5,35±2,06	2,40±1,54	6,10±2,05	2,70±1,56	4,65±2,16	5,40±1,85	0,00±0,00
ECgr+ CGLO	Aceitas	11,80±1,54	2,80±1,64	11,55±1,39	13,45±1,15	9,00±1,56	10,20±1,47	11,10±2,07
	Recusadas	3,20±1,54	12,25±1,71	3,45±1,39	1,55±1,15	6,00±1,56	5,80±1,47	4,90±2,07
ECgr+ wCGLO	Aceitas	12,35±1,60	2,55±1,32	11,80±1,70	12,85±1,09	8,60±2,04	10,30±2,13	11,25±1,41
	Recusadas	2,65±1,60	12,55±1,47	3,20±1,70	2,15±1,09	6,40±2,04	5,70±2,13	4,75±1,41

Fonte: autoria própria.

A Tabela 20 mostra os resultados ao adaptar a CNN treinada na base de dados MUG (fonte) para a base JAFFE (alvo). Considerando os métodos *baseline*, *joint* e *fine-tuning*, assume-se que o limite superior é o método *joint*, representando o caso ideal onde todos as bases de dados estão disponíveis para treinamento e o limite inferior é o método *fine-tuning*, onde a base de dados fonte não está mais disponível e o retreinamento é feito utilizando somente a base de dados alvo. O método *joint* alcançou o valor mais alto de acurácia média, com $\bar{A} = 0,8998$. Já o limite inferior, *fine-tuning*, obteve $\bar{A} = 7722$, com

uma perda de desempenho na base de dados fonte de 0,2515 $(0,9800 - 0,7285)$.

A aplicação do método ECgr melhorou o equilíbrio entre as tarefas, com $A_F = 0,8702$, $A_V = 0,8558$ e $\bar{A} = 0,8635$. Além disso, a filtragem por meio do QA supervisionado, ECgr+QA, melhorou a retenção do conhecimento anterior ($A_F = 0,9256$) e apresentou desempenho competitivo na base alvo ($A_V = 0,8227$), com $\bar{A} = 0,8742$. O método combinado ECgr+wQA apresentou os melhores resultados nessa adaptação inicial envolvendo apenas uma base de dados. Essa abordagem obteve $A_F = 0,9452$, $A_V = 0,8386$ e $\bar{A} = 0,8919$, ficando apenas 0,0079 pontos do *joint*.

Tabela 20 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada na base de dados MUG e adaptada para a base JAFFE, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual, para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da JAFFE com a versão sintética da MUG (MUG^A). No *fine-tuning*, a base-alvo é composta apenas pela JAFFE; no *joint*, é formada pela combinação das bases MUG original e JAFFE.

Método	**Base fonte**	**Base alvo**	**Média**
	MUG	**JAFFE**	
Baseline	0,9800±0,00	0,2800±0,00	0,6297±0,00
Joint	0,9836±0,00	0,8159±0,06	0,8998±0,09
Fine-tuning	0,7285±0,04	0,8159±0,05	0,7722±0,06
ECgr	0,8702±0,04	0,8568±0,06	0,8635±0,05
ECgr+QA	0,9256±0,03	0,8227±0,05	0,8742±0,07
ECgr+wQA	0,9452±0,01	0,8386±0,06	**0,8919±0,07**
ECgr+*cluster*	0,8540±0,04	0,8386±0,08	0,8463±0,06
ECgr+w*cluster*	0,8710±0,05	0,8568±0,06	0,8639±0,06
ECgr+CGLO	0,9079±0,03	0,8318±0,06	0,8699±0,06
ECgr+wCGLO	0,9067±0,03	0,8295±0,06	0,8681±0,06

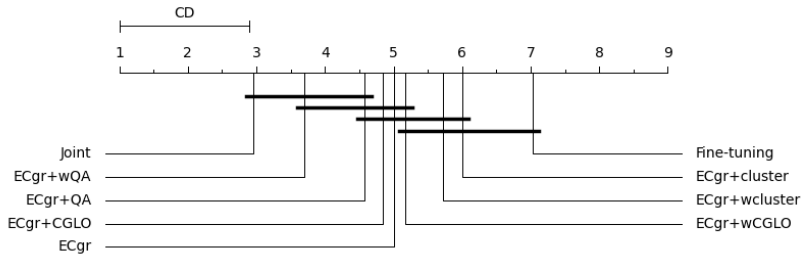
Fonte: autoria própria.

Outras variantes como ECgr+*cluster* ($\bar{A} = 0,8463$) e ECgr+w*cluster* ($\bar{A} = 0,8639$) também apresentaram melhorias sobre o *fine-tuning*, porém com desempenho inferior às combinações com QA supervisionado. Métodos baseados em CGLO, como ECgr+CGLO ($\bar{A} = 0,8699$) e ECgr+wCGLO ($\bar{A} = 0,8681$), mostraram desempenhos estáveis, mas não superaram o QA supervisionado com a aplicação de pesos na função de perda (ECgr+wQA).

Em síntese, os resultados evidenciam que o método ECgr+wQA atinge o melhor compromisso entre retenção e plasticidade, aproximando-se do desempenho ideal obtido utilizando o método *joint*, e superando o *fine-tuning* convencional.

O teste de Friedman (FRIEDMAN, 1937), com um nível de significância de $\alpha = 0,5$, revelou que há uma diferença significativa entre os métodos comparados para a primeira etapa do treinamento contínuo. A fim de averiguar quais métodos possuem uma diferença

Figura 37 – Teste *post-hoc* de Nemenyi na adaptação da CNN treinada na base de dados MUG para a JAFFE.



(a) Teste *post-hoc* de Nemenyi.

Método	*Rank*
Joint	2,950
ECgr+wQA	3,700
ECgr+QA	4,575
ECgr+CGLO	4,850
ECgr	5,000
ECgr+wCGLO	5,175
ECgr+w*cluster*	5,725
ECgr+*cluster*	6,000
Fine-tuning	7,025

(b) *Rank* por método.

Fonte: autoria própria.

estatisticamente significativa, foi aplicado o teste *post-hoc* de Nemenyi (NEMENYI, 1963), introduzido no contexto de aprendizado de máquina por Demšar (2006). A Figura 37 demonstra os resultados do teste *post-hoc*, em que, na Figura 37 (a) é possível observar o gráfico de diferença crítica, com as barras horizontais conectando os métodos que não obtiveram uma diferença estatisticamente significativa entre si, considerando a distância crítica (em inglês, *critical distance* – CD) e, na Figura 37 (b), estão listados os *ranks* de cada método, a partir do teste de Nemenyi. Os resultados corroboram com a síntese obtida a partir da análise das médias das acurácias, em que o método *joint*, no qual o treinamento é realizado de forma conjunta com dados dos dois conjuntos (MUG e JAFFE), obteve o melhor desempenho médio (*rank* = 2,950). Isso era esperado, uma vez que o aprendizado conjunto evita os efeitos do esquecimento catastrófico ao expor o modelo continuamente a dados de ambos os domínios. Em contrapartida, o método *fine-tuning*, apresentou o pior desempenho (*rank* = 7,025), evidenciando os impactos negativos do esquecimento catastrófico quando não há controle sobre a perda de plasticidade da rede. Ainda de acordo com o teste, nota-se que os métodos ECgr+QA (*rank* = 4,575) e ECgr+wQA (*rank* = 3,700) se destacam e não apresentam diferença significativa em relação ao limite superior, *joint*. Isto evidencia que, além da capacidade de retenção do conhecimento e de adaptabilidade para a nova tarefa destes métodos, a etapa de filtragem das imagens sintéticas também foi benéfica nesta etapa do treinamento contínuo. Já os métodos que incluem técnicas de clusterização (ECgr+*cluster* e ECgr+w*cluster*) obtiveram *ranks* mais elevados (6,000 e 5,725, respectivamente) e não indicaram diferença significativa em relação ao método de *fine-tuning*, considerado o limite inferior, indicando que a simples adição de agrupamento na filtragem das imagens não contribuiu para o desempenho do modelo final.

5.3.2.2.2 MUG e JAFFE para TFEID

Nesta etapa, a base de dados TFEID é considerada como a base-alvo atual. Para o método de *fine-tuning*, apenas JAFFE é usada como base-alvo, e para o *joint*, é feita a combinação direta de MUG original e JAFFE. Para os métodos baseados em ECgr, utilizam-se versões sintéticas das bases MUG e JAFFE, geradas pelas WGAN-GPs treinadas especificamente para esses domínios. Essas versões sintéticas são indicadas por MUG^A e JAFFE^A. O tamanho total de cada base sintética corresponde a 50% da base-alvo TFEID, sendo todas combinadas a esta durante o processo de adaptação. Dessa forma, nessa etapa foram geradas 134 imagens sintéticas da base MUG e 134 imagens sintéticas da base JAFFE, totalizando 268 imagens sintéticas, com distribuição igual entre as classes, resultando em seis classes com 19 imagens e uma classe com 20, conforme mostra a Tabela 21.

Tabela 21 – Número de imagens sintéticas geradas por classe nas bases de dados MUG^A e JAFFE^A, utilizando 50% do tamanho da base-alvo TFEID, para a adaptação da CNN treinada nas bases MUG e JAFFE para a TFEID.

Base de dados	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**	**Total**
MUG^A	19	19	19	19	19	19	20	134
JAFFE^A	19	19	19	19	19	19	20	134

Fonte: autoria própria.

A Tabela 22 mostra o número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada nas bases MUG e JAFFE para a base-alvo TFEID. Nos métodos com ECgr que utilizam imagens sintéticas, as bases de dados sintética MUG^A e JAFFE^A são combinadas com a base TFEID durante o treinamento e validação. Nos métodos com QA supervisionado (ECgr+QA e ECgr+wQA) e não-supervisionado (ECgr+*cluster*, ECgr+w*cluster*), os números de imagens para treino e validação são apresentados como médias e desvios padrão, considerando as 20 repetições do experimento. Nos métodos ECgr+CGLO e ECgr+wCGLO, apesar de haver o processo de filtragem, são geradas novas imagens até atingir o limite de 50% da base-alvo, conforme definido na Seção 4.2.3, resultando em números fixos de imagens para treino e validação. A base de testes é composta somente por imagens da base-alvo TFEID.

As Tabelas 23 e 24 apresentam, respectivamente, o número médio de imagens aceitas e recusadas das bases de dados MUG^A e JAFFE^A para cada classe no processo de filtragem dos métodos generativos com QA (ECgr+QA, ECgr+wQA, ECgr+*cluster*, ECgr+w*cluster*, ECgr+CGLO e ECgr+wCGLO). Ao analisar a Tabela 23, é possível perceber que a classe Desgosto é a que mais sofre com a rejeição das imagens sintéticas geradas pelas WGAN-GPs nos métodos em que a filtragem é feita de forma supervisionada (ECgr+QA e ECgr+CGLO e suas respectivas versões ponderadas), da mesma forma que já acontecia na primeita etapa do aprendizado contínuo, porém, agora, com valores menores

Tabela 22 – Número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada nas bases MUG e JAFFE para a base-alvo TFEID. Nos métodos baseados em ECgr, as bases de dados sintéticas MUG^A e JAFFE^A são combinadas com a base TFEID durante o treinamento e validação. No *fine-tuning*, apenas TFEID é usada como base-alvo, e no *joint*, é feita a combinação direta da MUG e JAFFE original com a TFEID.

Método	**Treino**	**Validação**	**Teste**	**Total**
Fine-tuning	214	27	27	268
Joint	6252	783	27	7062
ECgr	452	57	27	536
ECgr+QA	350,15±7,07	44,45±0,80	27	421,60
ECgr+wQA	343,20±7,93	43,35±1,01	27	413,55
ECgr+*cluster*	392,00±3,45	49,50±0,50	27	468,00
ECgr+w*cluster*	391,45±6,70	49,55±0,80	27	468,00
ECgr+CGLO	452	57	27	536
ECgr+wCGLO	452	57	27	536

Fonte: autoria própria.

de imagens recusadas. Na Tabela 24, observa-se que as imagens sintéticas da base JAFFE^A são recusadas em maior quantidade pelos métodos generativos com QA supervisionado, com exceção da classe Felicidade.

Tabela 23 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados MUG^A, geradas com as WGAN-GPs, na adaptação da CNN treinada nas bases MUG e JAFFE para a TFEID, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+ QA	Aceitas	15,85±1,57	9,95±2,39	16,30±1,17	17,70±1,03	11,35±1,79	12,45±1,76	12,00±1,86
	Recusadas	3,15±1,57	9,05±2,39	2,70±1,17	1,30±1,03	7,65±1,79	6,55±1,76	7,00±1,86
ECgr+ wQA	Aceitas	16,05±1,32	3,55±1,57	14,85±2,23	18,30±0,92	12,30±1,69	11,80±2,33	11,90±2,57
	Recusadas	2,95±1,32	15,45±1,57	4,15±2,23	0,70±0,92	6,70±1,69	7,20±2,33	7,10±2,57
ECgr+ *cluster*	Aceitas	12,85±1,87	16,10±1,83	11,85±2,11	15,85±0,99	12,60±1,82	13,95±1,88	18,80±0,41
	Recusadas	6,15±1,87	2,90±1,83	7,15±2,11	0,00±0,00	6,40±1,82	5,05±1,88	1,20±0,41
ECgr+ w*cluster*	Aceitas	13,00±2,32	16,25±1,48	11,70±2,49	16,20±1,58	12,95±2,95	12,15±2,21	18,95±0,22
	Recusadas	6,00±2,32	2,75±1,48	7,30±2,49	0,00±0,00	6,05±2,95	6,85±2,21	1,05±0,22
ECgr+ CGLO	Aceitas	13,90±2,22	7,40±1,50	16,90±1,12	17,60±0,99	14,20±1,79	12,30±2,18	11,55±1,90
	Recusadas	5,10±2,22	11,60±1,50	2,10±1,12	1,40±0,99	4,80±1,79	6,70±2,18	7,45±1,90
ECgr+ wCGLO	Aceitas	14,10±1,65	5,30±1,69	14,20±1,40	17,75±1,16	14,25±2,17	13,65±2,70	12,15±2,39
	Recusadas	4,90±1,65	13,70±1,69	4,80±1,40	1,25±1,16	4,75±2,17	5,35±2,70	6,85±2,39

Fonte: autoria própria.

A adaptação incremental do modelo previamente treinado com os conjuntos MUG e JAFFE para o domínio TFEID revela nuances importantes sobre a estabilidade e a plasticidade dos métodos avaliados. Conforme apresentado na Tabela 25, o desempenho da abordagem *baseline*, ou seja, o modelo previamente treinado nas bases de dados MUG e JAFFE sob a técnica de *fine-tuning*, sem qualquer estratégia de adaptação para a nova tarefa, é significativamente inferior na base alvo ($\bar{A} = 0,2963$). O modelo *joint*, que representa o limite superior ao treinar simultaneamente com todas as bases, alcança

Tabela 24 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados JAFFE^A, geradas com as WGAN-GPs, na adaptação da CNN treinada nas bases MUG e JAFFE para a TFEID, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+	Aceitas	4,70±2,00	6,15±2,01	6,75±1,94	16,10±1,52	7,40±2,37	9,85±2,62	7,05±2,50
QA	Recusadas	14,30±2,00	12,85±2,01	12,25±1,94	2,90±1,52	12,60±2,37	9,15±2,62	11,95±2,50
ECgr+	Aceitas	5,70±1,45	6,35±2,03	5,90±2,61	15,55±1,15	6,80±2,04	7,40±1,90	9,10±2,17
wQA	Recusadas	13,30±1,45	12,65±2,03	13,10±2,61	3,45±1,15	13,20±2,04	11,60±1,90	9,90±2,17
ECgr+	Aceitas	17,35±1,39	17,10±1,37	7,80±2,04	14,40±1,96	6,90±2,53	18,25±0,85	16,70±1,26
cluster	Recusadas	1,65±1,39	1,90±1,37	11,20±2,04	0,00±0,00	12,10±2,53	0,75±0,85	3,30±1,26
ECgr+	Aceitas	18,15±0,88	17,70±1,03	7,30±2,39	14,45±1,88	6,75±2,27	17,90±0,91	16,55±1,61
w*cluster*	Recusadas	0,85±0,88	1,30±1,03	11,70±2,39	0,00±0,00	12,25±2,27	1,10±0,91	3,45±1,61
ECgr+	Aceitas	3,95±1,67	6,35±1,57	5,15±1,90	16,95±1,19	8,95±2,80	9,15±2,43	8,00±1,49
CGLO	Recusadas	15,05±1,67	12,65±1,57	13,85±1,90	2,05±1,19	11,05±2,80	9,85±2,43	11,00±1,49
ECgr+	Aceitas	4,10±1,74	6,30±1,98	5,90±1,55	16,10±1,97	7,60±2,35	8,30±2,11	8,75±3,02
wCGLO	Recusadas	14,90±1,74	12,70±1,98	13,10±1,55	2,90±1,97	12,40±2,35	10,70±2,11	10,25±3,02

Fonte: autoria própria.

$\bar{A} = 0,8520$. Já o *fine-tuning*, mesmo preservando parcialmente os conhecimentos anteriores ($A_{MUG} = 0,7685$ e $A_{JAFFE} = 0,7432$), obteve $\bar{A} = 0,7650$, valor inferior ao *joint* e também abaixo dos métodos baseados em ponderação e *pseudo-rehearsal*.

A estratégia ECgr por si só apresentou uma média de $\bar{A} = 0,8021$, indicando uma melhoria tanto na retenção quanto na adaptação ao novo domínio. Quando combinada com QA supervisionado, no entanto, nota-se uma queda no desempenho sobre o conjunto JAFFE ($A_{JAFFE} = 0,6773$), ainda que o desempenho no TFEID ($A_{TFEID} = 0,8185$) se mantenha competitivo. O resultado médio de $\bar{A} = 0,7972$ revela um prejuízo na estabilidade do modelo.

O uso da ponderação por qualidade com o método supervisionado (ECgr+wQA) melhora significativamente o equilíbrio entre as bases, obtendo o maior desempenho médio entre os métodos propostos: $\bar{A} = 0,8276$. Essa estratégia apresenta retenção do conhecimento na base de dados MUG ($A_{MUG} = 0,9276$), perda na JAFFE ($A_{JAFFE} = 0,7182$) e adaptação positiva ao novo conjunto TFEID ($A_{TFEID} = 0,8370$), superando inclusive o *joint* na base alvo. Esse comportamento sugere que a ponderação contribui não apenas para preservar o conhecimento anterior, mas também para reforçar o aprendizado do novo domínio. Outros métodos como ECgr+*cluster* ($\bar{A} = 0,7925$) e ECgr+CGLO ($\bar{A} = 0,7933$) apresentaram desempenho competitivo em relação aos demais, com destaque para ECgr+wCGLO, que atinge $\bar{A} = 0,8145$, demonstrando que estratégias com ponderação adicional tendem a superar suas versões não ponderadas.

A partir do teste não-paramétrico de Friedman (FRIEDMAN, 1937) ($\alpha = 0,5$), foi constatado que há uma diferença significativa entre as médias das acurácias dos métodos avaliados no cenário de adaptação da CNN treinada previamente nas bases de dados MUG e JAFFE para a TFEID. O teste *post-hoc* de Nemenyi foi aplicado, conforme demonstra a Figura 38 e percebe-se que, os melhores métodos são *joint* (*rank*=3,067) e ECgr+wQA

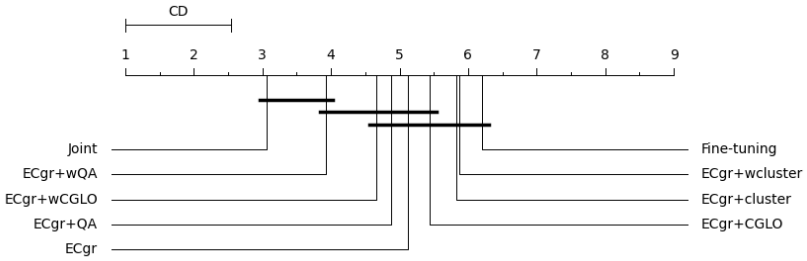
Tabela 25 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada nas bases de dados MUG e JAFFE e adaptada para a base TFEID, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da TFEID com as versões sintéticas das bases MUG e JAFFE (MUG$^+$ e JAFFE$^+$). No *fine-tuning*, a base-alvo é composta apenas pela TFEID; no *joint*, é formada pela combinação das bases MUG original, JAFFE original e TFEID.

Método	**Base fonte**		**Base alvo**	**Média**
	MUG	**JAFFE**	**TFEID**	
Baseline	0,7575±0,00	0,7273±0,00	0,2963±0,00	0,5937±0,00
Joint	0,9850±0,00	0,7841±0,06	0,7870±0,04	0,8520±0,10
Fine-tuning	0,7685±0,01	0,7432±0,04	0,7833±0,04	0,7650±0,03
ECgr	0,8287±0,05	0,7682±0,09	0,8093±0,04	0,8021±0,08
ECgr+QA	0,8958±0,02	0,6773±0,09	0,8185±0,06	0,7972±0,11
ECgr+wQA	0,9276±0,03	0,7182±0,09	0,8370±0,04	**0,8276±0,11**
ECgr+*cluster*	0,8411±0,03	0,7068±0,10	0,8296±0,04	0,7925±0,09
ECgr+w*cluster*	0,8587±0,04	0,6727±0,10	0,8185±0,05	0,7833±0,11
ECgr+CGLO	0,8732±0,03	0,6864±0,10	0,8204±0,06	0,7933±0,11
ECgr+wCGLO	0,9031±0,03	0,7273±0,09	0,8130±0,06	0,8145±0,10

Fonte: autoria própria.

(*rank*=3,933), os quais não possuem diferença significativa entre as acurácias observadas nos três conjuntos de dados.

Figura 38 – Teste *post-hoc* de Nemenyi na adaptação da CNN treinada na base de dados MUG e JAFFE para a TFEID.



(a) Teste *post-hoc* de Nemenyi.

Método	*Rank*
Joint	3,067
ECgr+wQA	3,933
ECgr+wCGLO	4,667
ECgr+QA	4,883
ECgr	5,117
ECgr+CGLO	5,433
ECgr+*cluster*	5,833
ECgr+w*cluster*	5,867
Fine-tuning	6,200

(b) *Rank* por método.

Fonte: autoria própria.

Os métodos ECgr+wCGLO, ECgr+QA, ECgr e ECgr+CGLO não apresentaram diferença significativa entre si. O método *fine-tuning*, com o *rank* mais alto (6,200), não apresentou diferença significativa com os métodos ECgr+w*cluster*, ECgr+*cluster*, ECgr+CGLO, ECgr, ECgr+QA e ECgr+wCGLO, o que demonstra que estes métodos

não demonstraram uma alta capacidade de retenção do conhecimento nesta etapa do experimento.

5.3.2.2.3 MUG, JAFFE e TFEID para CK+

Nesta etapa, a base de dados CK+ é considerada como a base-alvo atual. Para os métodos baseados em ECgr, utilizam-se versões sintéticas das bases MUG, JAFFE e TFEID, geradas pelas WGAN-GPs treinadas especificamente para esses domínios. Essas versões sintéticas são indicadas por MUG^A, JAFFE^A e TFEID^A. O tamanho total de cada base sintética corresponde a 50% da base-alvo CK+, sendo todas combinadas a esta durante o processo de adaptação. Para o método de *fine-tuning*, apenas CK+ é usada como base-alvo, e para o *joint*, é feita a combinação direta das bases originais MUG, JAFFE, TFEID e CK+.

A Tabela 26 mostra o número de imagens geradas por classe de cada base de dados sintética. Para cada base de dados (MUG, JAFFE e TFEID, que correspondem às tarefas anteriores), foram geradas 189 imagens sintéticas, totalizando 567 imagens sintéticas, com distribuição igual entre as classes, resultando em 27 imagens sintéticas para cada uma das sete classes.

Tabela 26 – Número de imagens sintéticas geradas por classe nas bases de dados MUG^A, JAFFE^A e TFEID^A, utilizando 50% do tamanho da base-alvo CK+, para a adaptação da CNN treinada nas bases MUG, JAFFE e TFEID para a CK+.

Base de dados	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**	**Total**
MUG^A	27	27	27	27	27	27	27	189
JAFFE^A	27	27	27	27	27	27	27	189
TFEID^A	27	27	27	27	27	27	27	189

Fonte: autoria própria.

O número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada nas bases MUG, JAFFE e TFEID para a base-alvo CK+ pode ser observado na Tabela 27. As bases de dados sintética MUG^A, JAFFE^A e TFEID^A são combinadas com a base CK+ durante o treinamento e validação nos métodos generativos. Nos métodos com QA supervisionado (ECgr+QA e ECgr+wQA) e não-supervisionado (ECgr+*cluster*, ECgr+w*cluster*), os números de imagens para treino e validação são apresentados como médias e desvios padrão, considerando as 20 repetições do experimento. A base de testes é composta somente por imagens da base-alvo CK+.

As Tabelas 28, 29 e 30 apresentam, respectivamente, o número médio de imagens aceitas e recusadas das bases de dados MUG^A, JAFFE^A e TFEID^A para cada classe no processo de filtragem dos métodos generativos com QA (ECgr+QA, ECgr+wQA, ECgr+*cluster*, ECgr+w*cluster*, ECgr+CGLO e ECgr+wCGLO). Para a base MUG^A, percebe-se que as classes Desgosto e Tristeza são as mais rejeitadas, o que indica que os

Tabela 27 – Número de imagens utilizadas nas etapas de treino, validação e teste para cada método na etapa de adaptação da CNN treinada nas bases MUG, JAFFE e TFEID para a base-alvo CK+. Nos métodos baseados em ECgr, as bases de dados sintéticas MUG^A, JAFFE^A e TFEID^A são combinadas com a base CK+ durante o treinamento e validação. No *fine-tuning*, apenas CK+ é usada como base-alvo, e no *joint*, é feita a combinação direta da MUG, JAFFE e TFEID original com a CK+.

Método	**Treino**	**Validação**	**Teste**	**Total**
Fine-tuning	301	38	38	377
Joint	6553	821	38	7412
ECgr	805	101	38	944
ECgr+QA	544,65±8,14	68,70±1,10	38	651,35
ECgr+wQA	547,20±7,49	68,95±0,86	38	654,15
ECgr+*cluster*	683,90±7,40	86,10±0,89	38	808,00
ECgr+w*cluster*	675,33±4,03	85,33±0,47	38	798,66
ECgr+CGLO	805	101	38	944
ECgr+wCGLO	805	101	38	944

Fonte: autoria própria.

modelos generativos podem não estar gerando imagens representativas o suficiente para que essas classes sejam distinguidas por redes neurais convolucionais. Ao analisar a base de dados JAFFE^A, nota-se um padrão similar ao já observado na etapa anterior, em que as imagens sintéticas desta base são rejeitadas com frequência, assim como para a base sintética TFEID^A. No entanto, da mesma forma que se observa na base MUG^A, a classe Felicidade apresenta um número menor de imagens recusadas, sugerindo que as imagens geradas para essa classe são mais facilmente reconhecíveis.

Tabela 28 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados MUG^A, geradas com as WGAN-GPs, na adaptação da CNN treinada nas bases MUG, JAFFE e TFEID para a CK+, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+ QA	Aceitas	21,05±1,82	11,90±2,95	21,25±2,10	25,70±1,30	16,35±1,73	19,20±3,21	17,55±2,54
	Recusadas	5,95±1,82	15,10±2,95	5,75±2,10	1,30±1,30	10,65±1,73	6,80±3,21	9,45±2,54
ECgr+ wQA	Aceitas	22,95±1,70	6,15±2,08	21,20±2,50	26,00±0,92	17,90±2,34	17,50±1,91	16,55±2,65
	Recusadas	4,05±1,70	20,85±2,08	5,80±2,50	1,00±0,92	9,10±2,34	8,50±1,91	10,45±2,65
ECgr+ *cluster*	Aceitas	19,08±2,47	22,58±1,78	18,17±1,95	23,42±2,07	19,33±1,37	18,17±2,48	27,00±0,00
	Recusadas	7,92±2,47	4,42±1,78	8,83±1,95	0,00±0,00	7,67±1,37	8,83±2,48	0,00±0,00
ECgr+ w*cluster*	Aceitas	18,12±2,20	21,65±2,00	16,94±2,59	22,82±1,42	18,88±1,87	17,76±2,95	26,88±0,33
	Recusadas	8,88±2,20	5,35±2,00	10,06±2,59	0,00±0,00	8,12±1,87	9,24±2,95	0,12±0,33
ECgr+ CGLO	Aceitas	20,80±2,07	17,25±2,73	18,85±2,28	25,40±1,47	16,15±3,03	20,35±2,41	16,95±2,44
	Recusadas	6,20±2,07	9,75±2,73	8,15±2,28	1,60±1,47	10,85±3,03	5,65±2,41	10,05±2,44
ECgr+ wCGLO	Aceitas	21,75±2,77	13,75±2,31	21,00±2,20	25,90±0,91	18,90±2,67	17,70±2,85	16,85±1,53
	Recusadas	5,25±2,77	13,25±2,31	6,00±2,20	1,10±0,91	8,10±2,67	8,30±2,85	10,15±1,53

Fonte: autoria própria.

Na Tabela 31, ao adaptar a CNN previamente treinada nos conjuntos MUG, JAFFE e TFEID para o conjunto CK+, observa-se que os métodos baseados em imagens sintéticas continuam apresentando desempenho competitivo. O método ECgr isoladamente

Tabela 29 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados JAFFE, geradas com as WGAN-GPs, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+	Aceitas	3,90±2,17	4,85±1,95	8,20±2,82	24,25±1,41	12,95±2,19	19,75±2,43	8,70±2,75
QA	Recusadas	23,10±2,17	22,15±1,95	18,80±2,82	2,75±1,41	14,05±2,19	7,25±2,43	18,30±2,75
ECgr+	Aceitas	11,00±2,70	6,35±1,84	6,55±2,31	21,15±1,63	10,80±2,95	16,75±2,43	14,45±3,28
wQA	Recusadas	16,00±2,70	20,65±1,84	20,45±2,31	5,85±1,63	16,20±2,95	10,25±2,43	12,55±3,28
ECgr+	Aceitas	25,50±1,45	24,83±0,94	11,33±2,67	20,42±2,31	8,83±2,37	25,50±1,17	24,00±1,76
cluster	Recusadas	1,50±1,45	2,17±0,94	15,67±2,67	0,00±0,00	18,17±2,37	1,50±1,17	3,00±1,76
ECgr+	Aceitas	25,06±1,43	24,65±1,54	10,88±3,02	20,24±1,82	9,12±1,76	25,29±1,45	24,24±1,30
w*cluster*	Recusadas	1,94±1,43	2,35±1,54	16,12±3,02	0,00±0,00	17,88±1,76	1,71±1,45	2,76±1,30
ECgr+	Aceitas	5,75±1,59	11,40±3,05	5,50±2,78	25,10±1,41	8,60±2,28	16,50±2,16	19,60±2,41
CGLO	Recusadas	21,25±1,59	15,60±3,05	21,50±2,78	1,90±1,41	18,40±2,28	10,50±2,16	7,40±2,41
ECgr+	Aceitas	7,35±2,78	10,25±2,38	7,45±2,21	25,05±1,15	7,60±2,39	15,55±2,09	19,15±2,28
wCGLO	Recusadas	19,65±2,78	16,75±2,38	19,55±2,21	1,95±1,15	19,40±2,39	11,45±2,09	7,85±2,28

Fonte: autoria própria.

Tabela 30 – Número médio de imagens aceitas e recusadas (com desvio padrão) para cada classe no processo de filtragem das imagens sintéticas da base de dados TFEID, geradas com as WGAN-GPs, durante 20 repetições.

Método	**Tipo**	**Raiva**	**Desgosto**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
ECgr+	Aceitas	0,40±0,60	5,55±2,09	11,10±1,94	17,65±2,25	9,30±2,20	14,65±2,21	0,10±0,31
QA	Recusadas	26,60±0,60	21,45±2,09	15,90±1,94	9,35±2,25	17,70±2,20	12,35±2,21	26,90±0,31
ECgr+	Aceitas	0,80±1,06	4,90±2,34	13,55±2,26	21,60±2,48	7,00±2,66	13,55±2,52	0,45±0,69
wQA	Recusadas	26,20±1,06	22,10±2,34	13,45±2,26	5,40±2,48	20,00±2,66	13,45±2,52	26,55±0,69
ECgr+	Aceitas	19,17±2,17	21,67±2,53	23,42±1,88	22,25±1,36	26,75±0,45	27,00±0,00	5,50±2,24
cluster	Recusadas	7,83±2,17	5,33±2,53	3,58±1,88	0,00±0,00	0,25±0,45	0,00±0,00	21,50±2,24
ECgr+	Aceitas	19,71±2,52	21,41±2,85	22,47±1,18	22,59±1,84	26,59±0,51	26,65±0,61	6,06±2,33
w*cluster*	Recusadas	7,29±2,52	5,59±2,85	4,53±1,18	0,00±0,00	0,41±0,51	0,35±0,61	20,94±2,33
ECgr+	Aceitas	0,40±0,50	6,70±1,72	16,15±2,50	19,90±2,85	17,45±2,48	14,45±2,37	0,70±0,92
CGLO	Recusadas	26,60±0,50	20,30±1,72	10,85±2,50	7,10±2,85	9,55±2,48	12,55±2,37	26,30±0,92
ECgr+	Aceitas	0,85±0,99	5,15±2,30	6,50±2,09	26,20±0,83	10,10±1,48	10,40±2,95	15,10±2,15
wCGLO	Recusadas	26,15±0,99	21,85±2,30	20,50±2,09	0,80±0,83	16,90±1,48	16,60±2,95	11,90±2,15

Fonte: autoria própria.

obteve o melhor desempenho médio ($\bar{A} = 0,7617$), com acurácias de $A_{MUG} = 0,8345$, $A_{JAFFE} = 0,7159$, $A_{TFEID} = 0,6556$ e $A_{CK+} = 0,8408$. Já os métodos ECgr+QA e ECgr+wQA, obtiveram na base alvo $A_{CK+} = 0,8526$ e $A_{CK+} = 0,8566$, respectivamente, e apresentaram maior variabilidade e queda nas demais bases, especialmente em JAFFE e TFEID, refletindo-se em médias inferiores.

Esse comportamento sugere que, embora a introdução de imagens sintéticas com QA (e sua versão com função de perda ponderada) favoreça a adaptação à nova base, ela pode amplificar desequilíbrios herdados das fases anteriores, particularmente quando a filtragem depende de redes com viés acumulado. Isso se manifesta, por exemplo, na acurácia reduzida em $A_{JAFFE} = 0,5702$ para ECgr+QA e $A_{TFEID} = 0,5519$ para ECgr+wQA. Esses resultados indicam que o uso dos pesos na função de perda, embora benéfico localmente, exige cautela em estágios mais avançados de aprendizado contínuo, onde erros propagados podem comprometer a generalização global do modelo.

Contudo, neste ponto, torna-se mais evidente que o uso de um peso para as imagens sintéticas traz um problema intrínseco ao treinamento da CNN usada para o método de

filtragem supervisionado. Esta CNN pode carregar comportamentos para as etapas de treinamento subsequentes, onde erros de certas classes podem comprometer o treinamento ao se utilizar o percentual de confiança.

Tabela 31 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada nas bases de dados MUG, JAFFE e TFEID e adaptada para a base CK+, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da CK+ com as versões sintéticas das bases MUG, JAFFE e TFEID (MUG$^+$, JAFFE$^+$ e TFEID$^+$). No *fine-tuning*, a base-alvo é composta apenas pela CK+; no *joint*, é formada pela combinação das bases MUG original, JAFFE original, TFEID original e CK+.

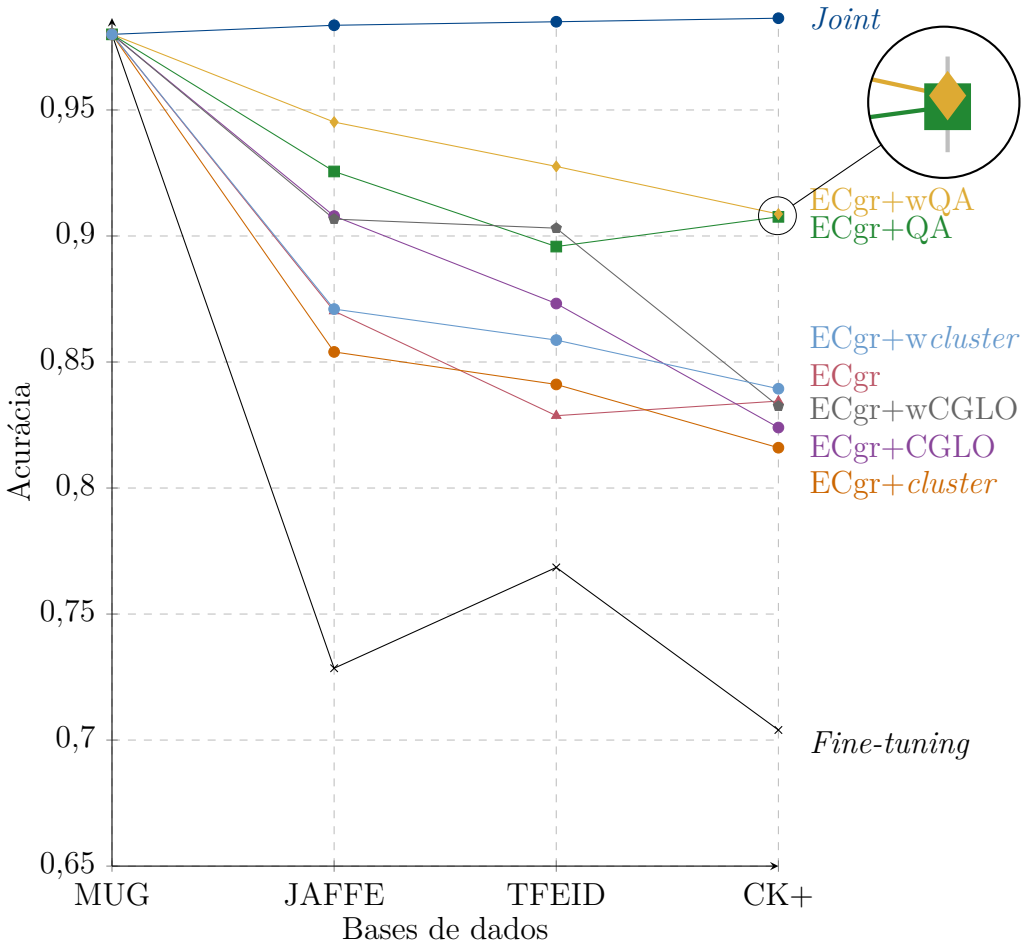
Método	**Base fonte**			**Base alvo**	**Média**
	MUG	**JAFFE**	**TFEID**	**CK+**	
Baseline	0,7602±0,00	0,7273±0,00	0,8519±0,00	0,3947±0,00	0,6835±0,00
Joint	0,9864±0,00	0,8455±0,07	0,8370±0,06	0,8645±0,03	0,8834±0,08
Fine-tuning	0,7048±0,06	0,5614±0,06	0,4519±0,04	0,8513±0,04	0,6424±0,16
ECgr	0,8345±0,03	0,7159±0,09	0,6556±0,10	0,8408±0,04	**0,7617±0,11**
ECgr+QA	0,9076±0,02	0,5702±0,05	0,6037±0,07	0,8526±0,03	0,7335±0,16
ECgr+wQA	0,9087±0,04	0,6841±0,09	0,5519±0,07	0,8566±0,03	0,7503±0,16
ECgr+*cluster*	0,8160±0,04	0,6432±0,09	0,5963±0,10	0,8447±0,03	0,7251±0,13
ECgr+w*cluster*	0,8394±0,04	0,6432±0,08	0,6833±0,07	0,8303±0,06	0,7452±0,15
ECgr+CGLO	0,8240±0,05	0,6023±0,08	0,6204±0,11	0,8513±0,05	0,7245±0,14
ECgr+wCGLO	0,8325±0,03	0,6455±0,09	0,6111±0,08	0,8474±0,04	0,7341±0,13

Fonte: autoria própria.

É possível compreender melhor os resultados no conjunto de dados MUG a partir do treinamento contínuo de todos os conjuntos com a Figura 39. Nota-se que o melhor método para o conjunto MUG é o ECgr+wQA. A introdução do peso no QA favorece a retenção de conhecimento previamente adquirido, o que se reflete em uma acurácia consistente na base de dados MUG, mesmo após múltiplas etapas de adaptação. O método ECgr+QA, sem ponderação, também mantém resultados robustos, indicando que o uso de métricas de qualidade como critério de seleção já contribui significativamente para mitigar o esquecimento catastrófico. Por outro lado, métodos baseados em clusterização e no CGLO apresentam desempenho inferior, com quedas progressivas ao longo do tempo, o que sugere menor resiliência à adaptação contínua e indica uma maior sensibilidade à transferência de distribuições de dados entre diferentes tarefas.

Destaca-se o fraco desempenho do *fine-tuning*, que sofre uma deterioração substancial da acurácia após as primeiras etapas. Embora o método consiga adaptar-se rapidamente às novas bases, ele compromete severamente a retenção do conhecimento anterior, o que reforça seu comportamento de sobreajuste às tarefas mais recentes. Essa tendência fica evidente na queda abrupta de desempenho no conjunto MUG ao final do processo, o que o

Figura 39 – Acurácia média no conjunto de testes MUG para a CNN treinada de forma incremental, considerando as etapas sucessivas de adaptação com as combinações MUG^A+JAFFE, MUG^A+JAFFE^A+TFEID e MUG^A+JAFFE^A+TFEID^A+CK+, avaliadas pelos métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas. As bases com sufixo A indicam versões sintéticas geradas por WGAN-GP a partir das bases anteriores. No método *joint*, as combinações utilizadas são MUG, MUG+JAFFE, MUG+JAFFE+TFEID e MUG+JAFFE+TFEID+CK+, sem o uso de imagens sintéticas. No *fine-tuning*, a adaptação é feita utilizando somente a base-alvo de cada etapa.



Fonte: autoria própria.

torna inadequado para cenários de aprendizagem contínua sem mecanismos de preservação de memória.

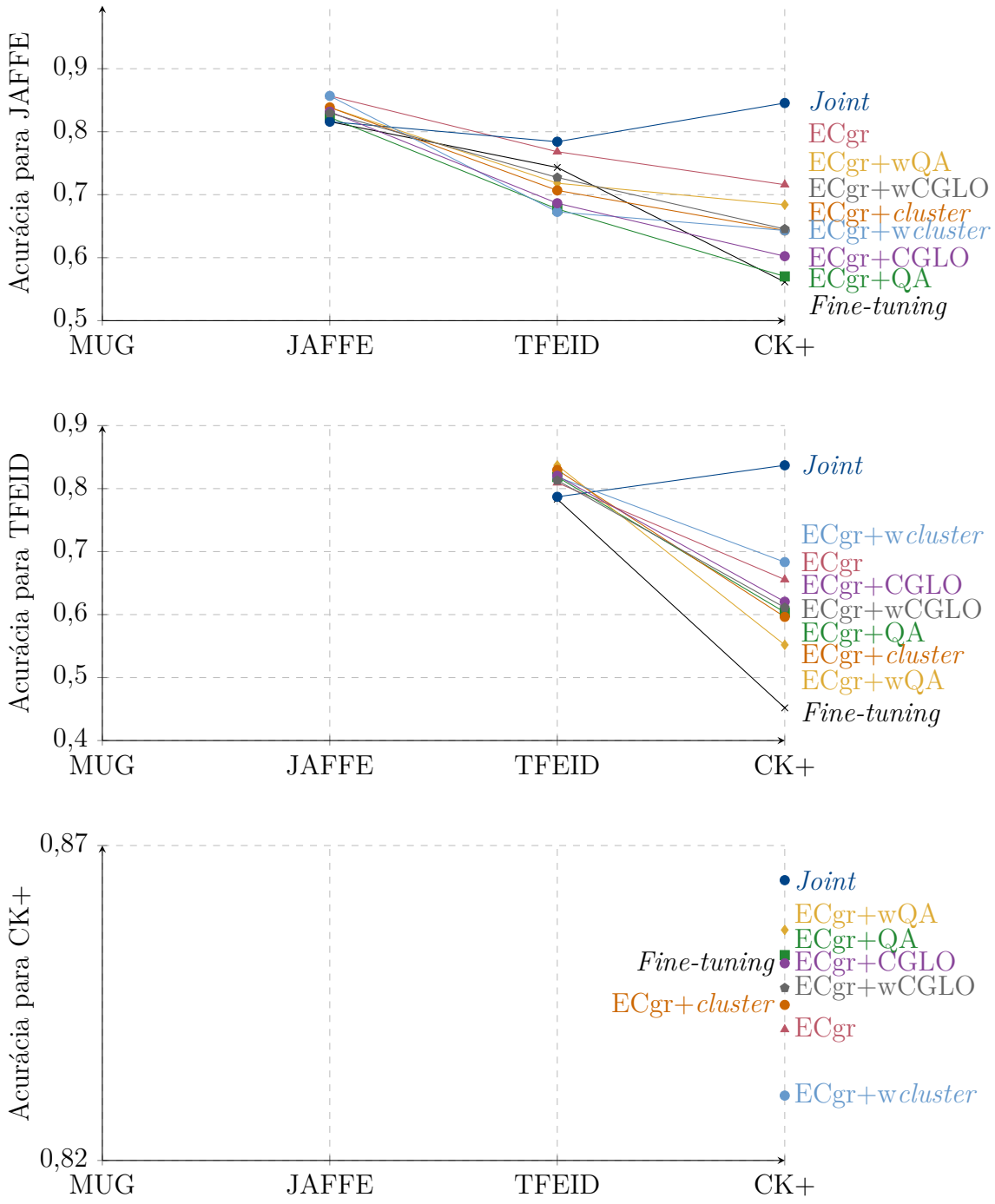
Já o método *joint* oferece o melhor desempenho geral, com acurácia quase estável em todos os conjuntos, dado que todos os dados estão disponíveis simultaneamente durante o treinamento. Esse cenário representa o limite superior de desempenho para os demais métodos. No entanto, sua aplicabilidade é restrita por fatores práticos, como consumo de memória, custos computacionais e restrições de privacidade, que frequentemente dificultam o acesso contínuo a todos os dados.

A fim de comparar os resultados nas demais bases de dados a partir da adaptação contínua das tarefas, a Figura 40 ilustra os resultados das acurácias para cada base de dados. O primeiro gráfico mostra os resultados da base de dados JAFFE. Para esta base de dados, a acurácia inicial do modelo treinado em MUG é baixa (0,2800), e todos os métodos apresentam aumento acentuado ao adaptar-se para JAFFE. O *fine-tuning* atinge 0,8159 ao ser adaptado para a JAFFE, mas nos demais passos de aprendizado contínuo, o conhecimento não foi retido, caindo para 0,5614 ao final. Os métodos baseados em ECgr apresentam comportamento mais estável, com ECgr atingindo o melhor resultado ao adaptar-se para a JAFFE (0,8568) e terminando também com o melhor resultado entre os métodos comparados (0,7159), o que evidencia que os métodos de filtragem das imagens prejudicaram os resultados na retenção do conhecimento na base de dados JAFFE.

No conjunto TFEID, o método *fine-tuning* alcança acurácia de 0,7833 após adaptação inicial, mas sofre degradação no passo seguinte, encerrando com 0,4519. Esse comportamento reforça o padrão de esquecimento catastrófico já observado nos outros conjuntos. Em contraste, os métodos baseados em ECgr e suas variações demonstram melhor estabilidade. Todos os métodos analisados superam o limite inferior *fine-tuning* mas ficam abaixo do limite superior *joint*, sendo o melhor deles ao final do processo o ECgr+w*cluster* (0,6833).

Para o conjunto CK+, por ser a última base de dados da sequência, não foi possível averiguar-se a retenção do seu conhecimento com os métodos propostos. No entanto, ao observar a acurácia nesta base de dados, é possível concluir que o melhor método é ECgr+wQA (0,8566), com uma acurácia próxima da observada no método *joint* (0,8645).

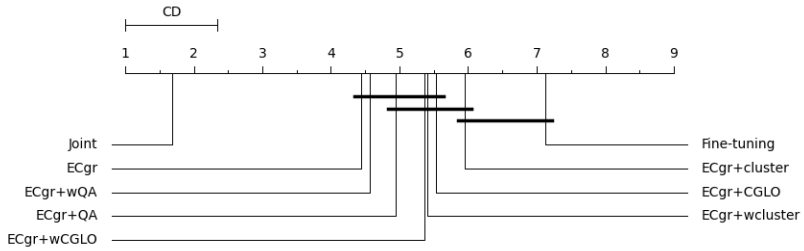
Figura 40 – Acurácia média em cada base-alvo (JAFFE, TFEID e CK+), ao longo do processo de adaptação incremental de uma CNN inicialmente treinada na base MUG, considerando as etapas sucessivas de adaptação com as combinações MUG^A+JAFFE, MUG^A+JAFFE^A+TFEID e MUG^A+JAFFE^A+TFEID^A+CK+, avaliadas pelos métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas. As bases com sufixo A indicam versões sintéticas geradas por WGAN-GP a partir das bases anteriores. No método *joint*, as combinações utilizadas são MUG, MUG+JAFFE, MUG+JAFFE+TFEID e MUG+JAFFE+TFEID+CK+, sem o uso de imagens sintéticas. No *fine-tuning*, a adaptação é feita utilizando somente a base-alvo de cada etapa.



Fonte: autoria própria.

Os resultados dos testes de Friedman revelam diferenças estatisticamente significativas entre os métodos avaliados, permitindo uma análise comparativa mais detalhada do desempenho relativo de cada abordagem com o teste *post-hoc* de Nemenyi, demonstrado na Figura 41. O método *joint* destacou-se com o menor valor de *rank* (1,688), confirmando sua superioridade em relação aos demais. Esse resultado está de acordo com a expectativa teórica, já que o treinamento conjunto aproveita diretamente os dados de todas as fontes. Em contrapartida, o método baseado em *fine-tuning* apresentou o pior desempenho (*rank* = 7,125), reforçando a hipótese de que a adaptação direta, sem mecanismos de mitigação do esquecimento, acarreta perdas expressivas de generalização.

Figura 41 – Teste *post-hoc* de Nemenyi na adaptação da CNN treinada na base de dados MUG, JAFFE e TFEID para a CK+.



(a) Teste *post-hoc* de Nemenyi.

Método	*Rank*
Joint	1,688
ECgr	4,438
ECgr+wQA	4,563
ECgr+QA	4,938
ECgr+wCGLO	5,363
ECgr+w*cluster*	5,400
ECgr+CGLO	5,538
ECgr+*cluster*	5,950
Fine-tuning	7,125

(b) *Rank* por método.

Fonte: autoria própria.

Entre as variações do método generativo ECgr, observa-se que a introdução de estratégias de controle de qualidade e ponderação influenciou o desempenho de maneira relevante. O ECgr puro apresentou um *rank* de 4,438, apresentando o menor valor de *rank* entre as variações do método nesta última etapa do treinamento contínuo. Já o ECgr+wQA (*rank* = 4,563) e o ECgr+QA (*rank* = 4,938) mostraram resultados semelhantes, porém inferiores ao ECgr padrão, sugerindo que a filtragem supervisionada e a ponderação da função de perda, apesar de promissoras em teoria, não garantem ganhos consistentes em todos os cenários. Estratégias como ECgr+wCGLO (*rank* = 5,363) e ECgr+w*cluster* (*rank* = 5,400) também não apresentaram melhorias significativas, posicionando-se entre os piores resultados. Os desempenhos mais baixos foram observados para ECgr+CGLO (*rank* = 5,538) e, sobretudo, ECgr+*cluster* (*rank* = 5,950).

Em síntese, os resultados do teste de Nemenyi reforçam dois aspectos centrais: (i) o método *joint* permanece como o cenário ideal, ainda que impraticável em contextos reais onde os dados anteriores não estão disponíveis; e (ii) as variações do ECgr, embora apresentem desempenho moderado, ainda carecem de ajustes mais sofisticados para superar limitações para mitigar o esquecimento e à consistência das filtragens das imagens sintéticas

utilizadas no retreinamento.

5.3.2.3 Influência da ordem das tarefas

Com o objetivo de analisar a influência da ordem de apresentação das tarefas no desempenho do método proposto, realizou-se um experimento com uma nova sequência de bases de dados: JAFFE, TFEID, MUG e CK+. Diferentemente da configuração original (MUG, JAFFE, TFEID e CK+), essa ordem foi definida iniciando pelo conjunto com menor número de amostras, permitindo avaliar o comportamento do modelo quando as primeiras representações aprendidas são mais limitadas em diversidade. Esse novo protocolo experimental permite investigar a sensibilidade do método à ordem das tarefas, uma vez que, em cenários de aprendizado contínuo, o desempenho pode variar significativamente dependendo da sequência de exposição aos dados.

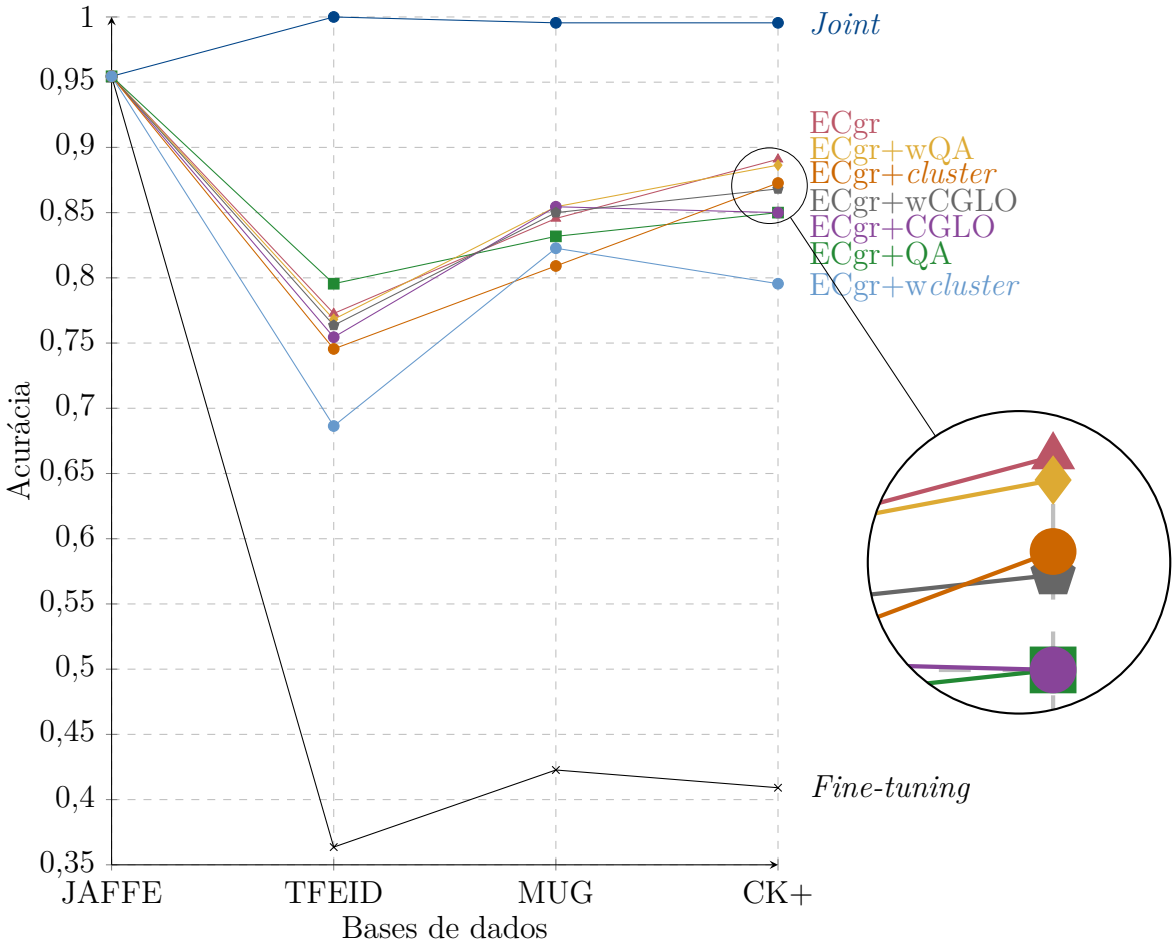
Os resultados, demonstrados na Figura 42, indicam que os métodos baseados em ECgr superam consistentemente o *fine-tuning* em ambas as ordens de apresentação das tarefas. Isso sugere que o ganho obtido não depende de uma sequência específica de bases de dados, evidenciando a robustez do método proposto em cenários distintos de aprendizado contínuo. Os valores numéricos completos correspondentes aos experimentos desta seção são apresentados no Apêndice A.

Por outro lado, observa-se que o desempenho ao longo das etapas varia conforme a ordem das tarefas, indicando que o problema de aprendizado contínuo permanece sensível à sequência de dados. Em particular, a adaptação para a base TFEID apresenta queda de desempenho em ambos os cenários, sugerindo maior dificuldade de generalização para esse domínio específico. Além disso, os resultados sugerem que a escolha da primeira tarefa influencia diretamente a qualidade das representações aprendidas. Quando o treinamento se inicia com uma base mais limitada (JAFFE), observa-se maior instabilidade nas etapas subsequentes, especialmente na adaptação para TFEID. Em contraste, iniciar com uma base mais diversa (MUG) tende a produzir representações mais robustas, favorecendo o desempenho nas tarefas seguintes.

A recorrente queda de desempenho na etapa envolvendo a base TFEID, observada em ambas as ordens experimentais, indica que determinadas bases podem atuar como pontos críticos no processo incremental. Esse comportamento sugere que características específicas do conjunto TFEID, como menor variabilidade e diferenças de distribuição, dificultam a adaptação do modelo e intensificam os efeitos do esquecimento.

Outro aspecto relevante é a maior estabilidade dos métodos propostos ao longo das etapas. Enquanto o *fine-tuning* apresenta quedas abruptas de desempenho, especialmente na nova ordem experimental, os métodos baseados em ECgr mantêm uma degradação mais suave, indicando melhor equilíbrio entre retenção e adaptação.

Figura 42 – Acurácia média no conjunto de testes JAFFE para a CNN treinada de forma incremental, considerando as etapas sucessivas de adaptação com as combinações JAFFE^{A}+TFEID, JAFFE^{A}+TFEID^{A}+MUG e JAFFE^{A}+TFEID^{A}+MUG^{A}+CK+, avaliadas pelos métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas. As bases com sufixo A indicam versões sintéticas geradas por WGAN-GP a partir das bases anteriores. No método *joint*, as combinações utilizadas são JAFFE, JAFFE+TFEID, JAFFE+TFEID+MUG e JAFFE+TFEID+MUG+CK+, sem o uso de imagens sintéticas. No *fine-tuning*, a adaptação é feita utilizando somente a base-alvo de cada etapa.



Fonte: autoria própria.

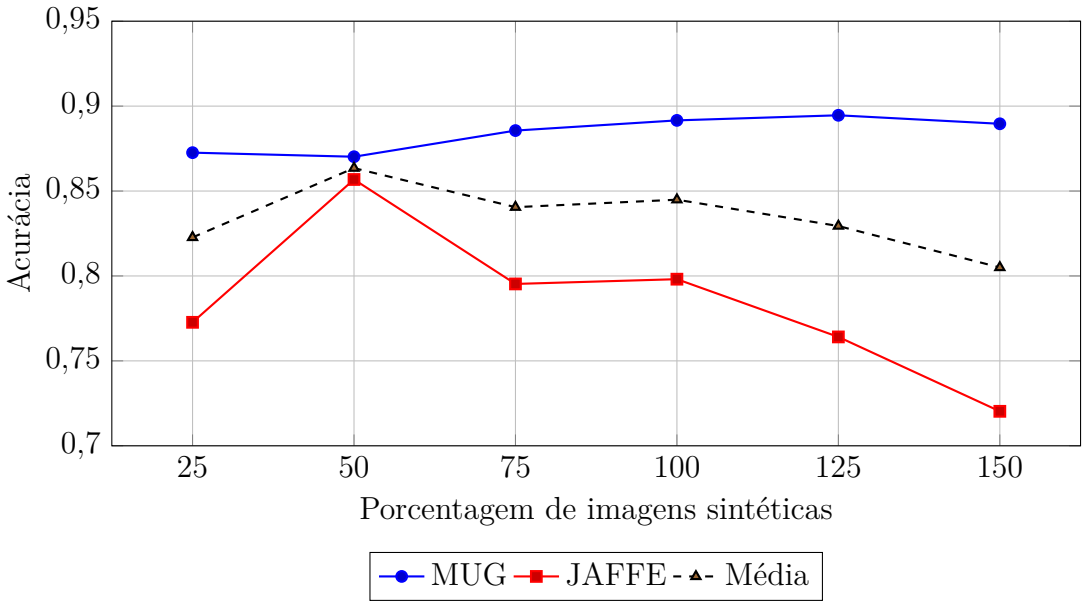
Em síntese, os experimentos com diferentes ordens de apresentação das tarefas demonstram que, embora o desempenho absoluto seja influenciado pela sequência dos dados, os métodos propostos mantêm superioridade consistente em relação ao *fine-tuning*. Além disso, os resultados evidenciam que a escolha da tarefa inicial e as características específicas de cada base de dados desempenham papel importante na dinâmica do aprendizado contínuo.

5.3.2.4 Avaliação em relação a proporção de imagens sintéticas

Nesta seção, analisa-se o impacto da proporção de imagens sintéticas geradas ao longo do processo de aprendizado contínuo. As porcentagens variam de 25% a 150% em relação ao tamanho do conjunto de dados da nova tarefa. O objetivo é entender como diferentes quantidades de dados sintéticos influenciam a retenção do conhecimento anterior e a adaptação à nova tarefa. Todos os resultados reportados nesta seção foram obtidos com o método ECgr (sem QA) e dizem respeito a partição de teste das bases de dados, respeitando a divisão citada previamente na Seção 5.3.2.

No primeiro cenário, o modelo é inicialmente treinado em MUG e, posteriormente, adaptado à base JAFFE. Os resultados indicam, como ilustra a Figura 43, que o desempenho em MUG permanece relativamente estável à medida que a porcentagem de imagens sintéticas aumenta, com uma melhora até 125%. Por outro lado, o desempenho em JAFFE atinge seu pico com 50% de imagens sintéticas, apresentando uma queda gradual nas porcentagens superiores. Isso sugere que, embora a geração de imagens sintéticas seja benéfica para manter o desempenho na tarefa anterior, o excesso de dados pode interferir na aprendizagem da nova tarefa. A média das acurácias entre os dois domínios confirma essa tendência, atingindo o melhor resultado em 50%, com posterior decréscimo. Portanto, observa-se que 50% de dados sintéticos é a proporção que melhor equilibra retenção e plasticidade nessa etapa do experimento.

Figura 43 – Variação da acurácia no conjunto de teste em função da porcentagem de imagens sintéticas utilizadas no treinamento, relativa ao tamanho da base de dados alvo (MUG → JAFFE).

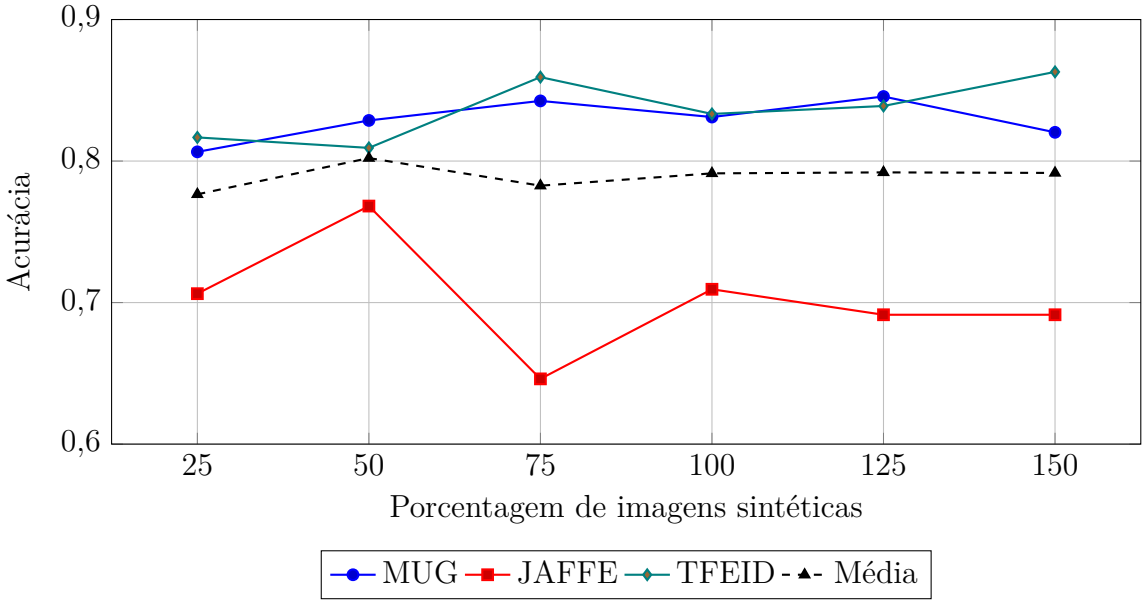


Fonte: autoria própria.

No segundo experimento, ao introduzir a terceira tarefa, a base de dados TFEID,

percebe-se que o desempenho na base de dados MUG continua estável. Já a base de dados JAFFE apresenta seu melhor resultado com 50% de imagens sintéticas. O desempenho do conjunto TFEID melhora progressivamente, com o primeiro pico em 75% e atingindo seu valor mais alto em 150%, conforme ilustra a Figura 44. Esse comportamento indica que, para tarefas com maior diversidade e complexidade, a introdução de dados sintéticos adicionais auxilia na generalização. No entanto, a média das acurácias entre as três tarefas ainda apresenta seu melhor valor em 50%, embora os resultados entre 100% e 150% também se mantenham competitivos.

Figura 44 – Variação da acurácia no conjunto de teste em função da porcentagem de imagens sintéticas utilizadas no treinamento, relativa ao tamanho da base de dados alvo (MUG → JAFFE → TFEID).

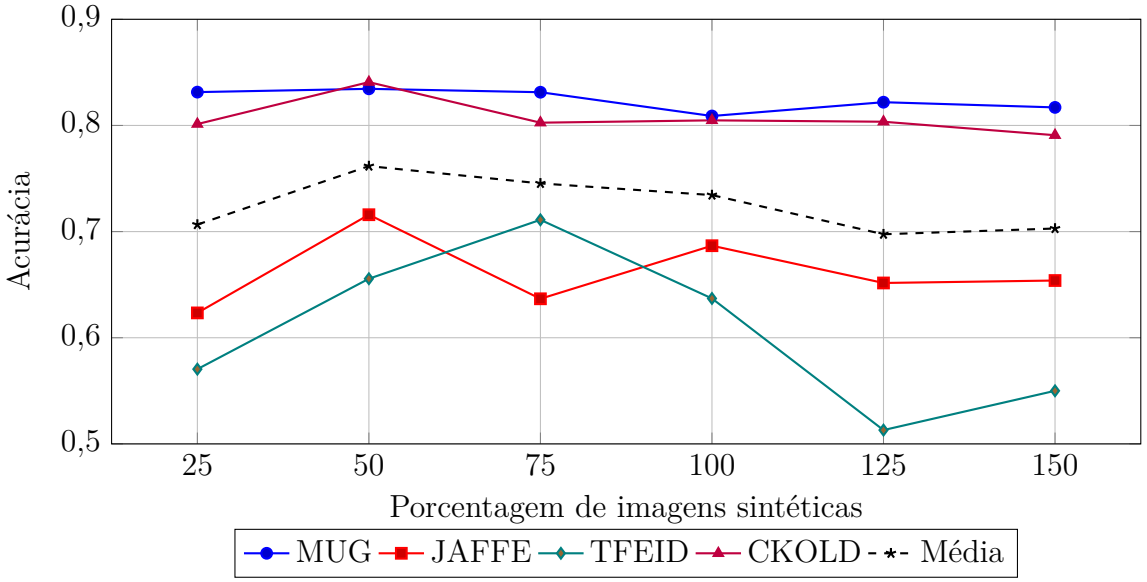


Fonte: autoria própria.

No último cenário, ilustrado na Figura 45, com quatro tarefas sequenciais, observa-se que MUG (primeira tarefa) apresenta uma queda na acurácia a partir de 100%. O conjunto JAFFE apresenta seu melhor desempenho com 50%, enquanto TFEID atinge o ápice em 75%, mas sofre queda nas porcentagens superiores. O desempenho na base de dados CK+ permanece estável ao longo de todas as porcentagens, atingindo o melhor valor em 50%. A média das acurácias entre as quatro tarefas atinge seu valor máximo em 50%, caindo gradualmente nas porcentagens subsequentes. Esse comportamento indica que, à medida que mais tarefas são incorporadas ao processo, a sensibilidade do sistema ao excesso de dados sintéticos aumenta. Quantidades elevadas de dados anteriores podem saturar a capacidade do modelo, prejudicando a assimilação de novos conhecimentos.

Nas Tabelas 32, 33 e 34 é possível encontrar as acurácias para as diferentes bases de dados utilizando cada uma das porcentagens analisadas. A análise dos três experimentos revela uma tendência que a proporção de 50% de imagens sintéticas em relação ao tamanho

Figura 45 – Variação da acurácia no conjunto de teste em função da porcentagem de imagens sintéticas utilizadas no treinamento, relativa ao tamanho da base de dados alvo (MUG → JAFFE → TFEID → CK+).



Fonte: autoria própria.

da nova tarefa é, de maneira consistente, a mais eficaz para preservar o conhecimento anterior sem comprometer a aprendizagem de novas tarefas. A introdução de maiores quantidades de dados sintéticos, embora beneficie casos pontuais, frequentemente resulta em interferência negativa, principalmente em cenários com maior número de tarefas. Assim, o uso de 50% de imagens sintéticas, em relação ao número de imagens na base de dados alvo, se mostrou uma escolha robusta para manter o equilíbrio entre retenção e plasticidade em cenários de aprendizado contínuo.

Tabela 32 – Acurácia média e desvio padrão para as bases MUG e JAFFE ao transferir um modelo inicialmente treinado na base MUG para a combinação das bases MUG^A e JAFFE original. As adaptações são realizadas utilizando diferentes proporções de imagens sintéticas da base MUG^A, geradas por WGAN-GPs e combinadas à base-alvo JAFFE. Cada linha da tabela corresponde a uma porcentagem distinta de MUG^A em relação ao tamanho total da base-alvo.

Porcentagem	**Base fonte**	**Base alvo**	**Média**
	MUG	**JAFFE**	
25%	0,8726±0,04	0,7727±0,04	0,8227±0,06
50%	0,8702±0,04	0,8568±0,06	**0,8635±0,05**
75%	0,8856±0,03	0,7953±0,03	0,8405±0,05
100%	0,8916±0,04	0,7981±0,06	0,8449±0,07
125%	0,8946±0,04	0,7641±0,07	0,8294±0,09
150%	0,8896±0,04	0,7203±0,05	0,8050±0,10

Fonte: autoria própria.

Tabela 33 – Acurácia média e desvio padrão para as bases MUG, JAFFE e TFEID ao transferir um modelo inicialmente treinado nas bases MUG^A+JAFFE para a combinação das bases MUG^A, JAFFE^A e TFEID original. As adaptações são realizadas utilizando diferentes proporções de imagens sintéticas das bases MUG^A e JAFFE^A, geradas por WGAN-GPs e combinadas à base-alvo TFEID. Cada linha da tabela corresponde a uma porcentagem distinta de MUG^A e JAFFE^A em relação ao tamanho total da base-alvo.

Método	**Base fonte**		**Base alvo**	**Média**
	MUG	**JAFFE**	**TFEID**	
25%	0,8065±0,05	0,7063±0,07	0,8167±0,05	0,7765±0,09
50%	0,8287±0,05	0,7682±0,09	0,8093±0,04	**0,8021±0,08**
75%	0,8425±0,04	0,6461±0,07	0,8593±0,03	0,7826±0,13
100%	0,8311±0,03	0,7094±0,05	0,8333±0,05	0,7913±0,09
125%	0,8456±0,03	0,6914±0,05	0,8389±0,05	0,7920±0,10
150%	0,8203±0,03	0,6914±0,05	0,8630±0,05	0,7916±0,10

Fonte: autoria própria.

Tabela 34 – Acurácia média e desvio padrão para as bases MUG, JAFFE, TFEID e CK+ ao transferir um modelo inicialmente treinado nas bases MUG^A + JAFFE^A + TFEID para a combinação das bases MUG^A, JAFFE^A, TFEID^A e CK+ original. As adaptações são realizadas utilizando diferentes proporções de imagens sintéticas das bases MUG^A, JAFFE^A e TFEID^A, geradas por WGAN-GPs e combinadas à base-alvo CK+. Cada linha da tabela corresponde a uma porcentagem distinta dessas bases sintéticas em relação ao tamanho total da base-alvo.

Método	**Base fonte**			**Base alvo**	**Média**
	MUG	**JAFFE**	**TFEID**	**CK+**	
25%	0,8314±0,04	0,6234±0,06	0,5704±0,12	0,8013±0,03	0,7066±0,19
50%	0,8345±0,03	0,7159±0,09	0,6556±0,10	0,8408±0,04	**0,7617±0,15**
75%	0,8313±0,05	0,6367±0,03	0,7111±0,13	0,8026±0,03	0,7454±0,15
100%	0,8089±0,03	0,6867±0,04	0,6370±0,07	0,8048±0,04	0,7344±0,13
125%	0,8219±0,03	0,6516±0,05	0,5130±0,08	0,8035±0,04	0,6975±0,19
150%	0,8170±0,04	0,6539±0,05	0,5500±0,09	0,7908±0,03	0,7029±0,17

Fonte: autoria própria.

5.3.2.5 Discussão

A fim de avaliar o impacto do treinamento de novas tarefas na retenção do conhecimento já obtido, avalia-se os modelos previamente treinados utilizando as métricas BWT (Equação 2.7) e FWT (Equação 2.8). Valores positivos de BWT sugerem que o desempenho em tarefas anteriores foi preservado ou até melhorado ao longo do treinamento, enquanto valores negativos indicam perda de conhecimento aprendido anteriormente. Já no caso do FWT, resultados acima de zero apontam para um efeito positivo do conhecimento prévio sobre tarefas futuras, ao passo que valores negativos revelam uma interferência prejudicial

na aprendizagem de novas tarefas.

Tabela 35 – Métricas BWT e FWT em relação ao treinamento contínuo nas bases de dados MUG, JAFFE, TFEID e CK+.

Método	**BWT ↑**	**FWT ↑**
Joint	0,0287	0,5191
Fine Tuning	-0,2870	0,5135
ECgr	**-0,1467**	0,5323
ECgr+QA	-0,1799	0,5279
ECgr+wQA	-0,1703	**0,5407**
ECgr+*cluster*	-0,1979	0,5343
ECgr+w*cluster*	-0,1631	0,5319
ECgr+CGLO	-0,1952	0,5312
ECgr+wCGLO	-0,1778	0,5266

Fonte: autoria própria.

O método *joint*, que representa o treinamento com todas as bases disponíveis simultaneamente, apresentou o melhor resultado para a métrica BWT (0,0287), confirmando a expectativa de que a ausência de restrições no acesso aos dados das tarefas anteriores resulta em um desempenho superior na mitigação do esquecimento. O resultado de FWT (0,5191) não foi o melhor dentre os métodos analisados, o que nos permite refletir sobre a eficácia da transferência positiva para as novas tarefas em relação à união de diversos conjuntos de dados para a adaptação a novas tarefas. Por outro lado, o *fine-tuning* exibe o menor valor de BWT (-0,2870), evidenciando um severo esquecimento das tarefas anteriores ao se treinar novas tarefas sem qualquer estratégia de retenção de conhecimento. Apesar disso, o valor de FWT (0,5135) permanece positivo, sugerindo que o conhecimento prévio ainda influencia beneficamente o aprendizado de tarefas subsequentes.

Dentre os métodos baseados em ECgr, observa-se uma melhora considerável no BWT quando comparados ao *fine-tuning*, embora ainda apresentem valores negativos, indicando que algum grau de esquecimento persiste. O método ECgr puro obteve um BWT de -0,1467, o melhor entre os métodos propostos, sinalizando que mesmo sem mecanismos adicionais de filtragem ou ponderação da função de perda, a utilização de imagens sintéticas geradas por classe contribui para a mitigação do esquecimento. A adição do método de qualidade (QA) ao ECgr melhora a transferência para novas tarefas (FWT = 0,5279), embora o BWT tenha sido menor (-0,1799). A versão com ponderação (wQA) alcança o maior valor de FWT entre todos os métodos (0,5407), sugerindo que o refinamento na escolha das imagens sintéticas tem um impacto direto na facilitação do aprendizado futuro, mesmo que o BWT (-0,1703) ainda denote uma perda de conhecimento prévio.

No caso do método ECgr+*cluster*, a abordagem de agrupamento na filtragem obteve um FWT competitivo (0,5343), embora apresente um dos menores BWTs entre os métodos propostos (-0,1979), indicando que, apesar da maior diversidade das imagens sintéticas, o modelo ainda sofre perda de desempenho em tarefas anteriores. A versão

ponderada (w*cluster*) melhora o BWT para -0,1631, mantendo o FWT elevado (0,5319), apontando para um equilíbrio mais favorável entre retenção e aprendizagem.

Por fim, o método ECgr+CGLO, também se mostrou eficaz na transferência de aprendizado, com um FWT de 0,5312. No entanto, seu BWT (-0,1952) permanece entre os mais baixos dos métodos propostos, sugerindo que o refinamento no espaço latente não é suficiente, por si só, para preservar o conhecimento anterior. A versão ponderada (wCGLO) obteve uma melhora no BWT (-0,1778), mantendo um FWT satisfatório (0,5266).

Em síntese, os resultados evidenciam que todos os métodos propostos são capazes de promover transferência positiva de aprendizado (FWT > 0), demonstrando a efetividade das estratégias de *pseudo-rehearsal* baseadas em geração de dados sintéticos. Contudo, o desafio do esquecimento ainda persiste (BWT < 0 para todos, exceto *joint*), sendo mais bem mitigado pelo método ECgr puro, enquanto a transferência para novas tarefas é melhor promovida por ECgr+wQA, destacando o potencial de abordagens híbridas que combinam filtragem supervisionada e ponderação.

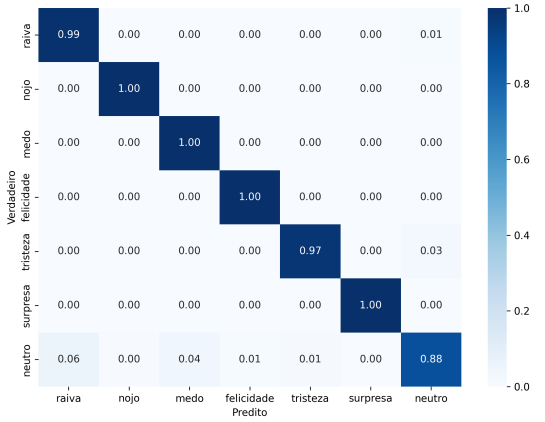
Um problema comum em FER é a variabilidade de desempenho por classe, e a ocorrência de confusões mais pronunciadas entre pares de expressões específicas. Para caracterizar isso, apresenta-se a matriz de confusão da base de dados MUG, a partir dos resultados obtidos com os métodos ECgr e ECgr+wQA (melhor BWT e melhor FWT, respectivamente) na primeira etapa (MUG para JAFFE) e na última etapa (MUG+JAFFE+TFEID para CK+). A Figura 46 apresenta, em cada linha, dois mapas de calor das matrizes de confusão, contendo a média (normalizada por classe) sobre as 20 execuções independentes de cada etapa, dispostos lado a lado, referentes aos métodos ECgr, na Figura 46 (a) e (b), e ECgr+wQA, na Figura 46 (c) e (d).

O método ECgr apresentou variações no desempenho por classe ao comparar o conjunto MUG após a primeira etapa de aprendizado contínuo (MUG→JAFFE) e após a última etapa (MUG+JAFFE+TFEID→CK+). A partir da comparação direta das matrizes de confusão, ilustradas na primeira linha da Figura 46 (a) e (b), é possível observar que as principais confusões estão concentradas nas classes Raiva e Neutro. Tanto na etapa inicial quanto final, a classe Raiva apresenta uma confusão com a classe Nojo, porém com uma redução na etapa final de 38% para 24%, indicando que o modelo se tornou mais robusto em distinguir essas duas emoções com as imagens sintéticas ao longo do aprendizado contínuo. A classe Neutro, por sua vez, apresenta, na etapa inicial, uma confusão de 29% com a classe Raiva, além de 7% com Nojo e também Tristeza. Na etapa final, no entanto, a confusão com a classe Raiva reduz para 3%, mas aumenta para 12% com a classe Medo e para 15% com a classe Tristeza, o que indica sustenta a afirmação de que o modelo se tornou mais robusto em reconhecer a emoção Raiva, mas que o modelo passou a confundir mais a classe Neutro com Medo e Tristeza ao longo do aprendizado contínuo. O mesmo padrão de confusão entre as classes Neutro e Raiva pode ser observado ao comparar as

Figura 46 – Comparação dos *heatmaps* das matrizes de confusão no conjunto de testes da base de dados MUG para os métodos ECgr e ECgr+wQA nas etapas inicial (MUG→JAFFE) e final (MUG+JAFFE+TFEID→CK+) do aprendizado contínuo.

(a) ECgr, etapa inicial.

(b) ECgr, etapa final.



(c) ECgr+wQA, etapa inicial.

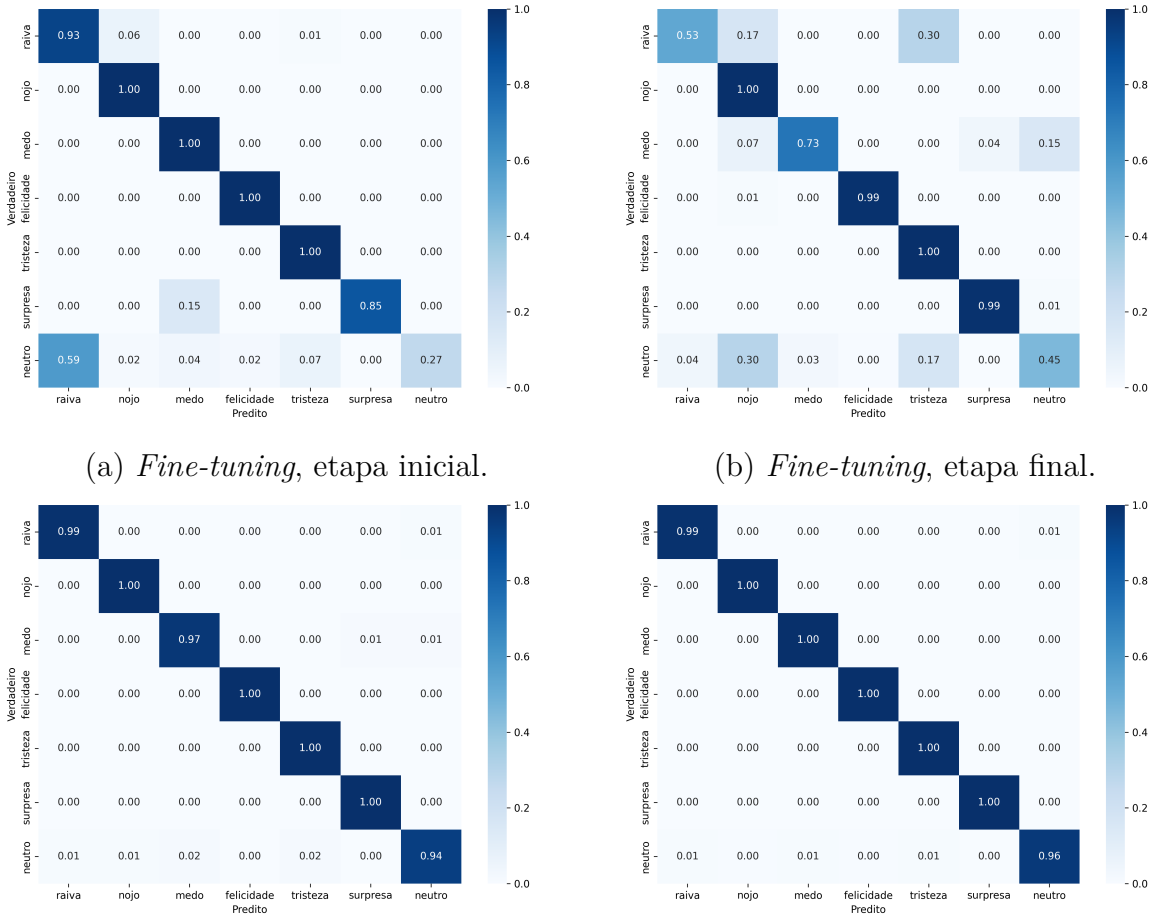
(d) ECgr+wQA, etapa final.

Fonte: autoria própria.

matrizes de confusão do método ECgr+wQA, apresentadas na segunda linha da Figura 46 (c) e (d). A classe Neutro apresenta uma confusão de 6% com Raiva na etapa inicial, que reduz para 2% na etapa final, mas aumenta a confusão com Medo (de 4% para 6%) e Tristeza (de 1% para 6%). Esses resultados evidenciam que classes com fronteiras mais sutis, como Medo, Nojo e Neutro, permanecem suscetíveis a alterações nas distribuições internas de confusão quando submetidas ao método de *pseudo-rehearsal* ECgr.

Na Figura 47 é possível observar a comparação dos métodos *fine-tuning* e *joint*. O método *fine-tuning* apresenta confusão na etapa inicial nas classes Raiva, Surpresa e Neutro, e, na etapa final, apresenta confusão, principalmente, nas classes Raiva, Medo e Neutro, com uma distribuição dispersa das confusões entre as emoções. O método *joint*, considerado como o limite superior de desempenho, por sua vez, apresenta uma matriz de confusão com estabilidade entre as etapas inicial e final, evidenciando que o treinamento conjunto

Figura 47 – Comparação dos *heatmaps* das matrizes de confusão no conjunto de testes da base de dados MUG para os métodos *fine-tuning* e *joint* nas etapas inicial (MUG→JAFFE) e final (MUG+JAFFE+TFEID→CK+) do aprendizado contínuo.



(a) *Fine-tuning*, etapa inicial.

(b) *Fine-tuning*, etapa final.

(c) *Joint*, etapa inicial.

(d) *Joint*, etapa final.

Fonte: autoria própria.

com todos as imagens das bases de dados disponíveis para o treinamento representa o cenário ideal para o aprendizado contínuo em FER.

O método EWC (KIRKPATRICK; AL., 2017) foi implementado e aplicado ao protocolo experimental de reconhecimento de emoção definido neste trabalho. Os resultados podem ser observados na Tabela 36. O método EWC apresentou o maior valor para BWT (-0,1030) dentre os métodos comparados, apesar de ainda indicar esquecimento (BWT < 0). No entanto, cabe destacar que a transferência do conhecimento antigo foi prejudicada no método EWC (0,5034), ressaltando a necessidade de equilíbrio entre técnicas de mitigação do esquecimento e de transferência de aprendizado. O melhor resultado para FWT, conforme discutido anteriormente, foi por meio do método ECgr+wQA (0,5407).

Tabela 36 – Métricas BWT e FWT em relação ao treinamento contínuo nas bases de dados MUG, JAFFE, TFEID e CK+, em comparação com o método EWC (KIRKPATRICK; AL., 2017).

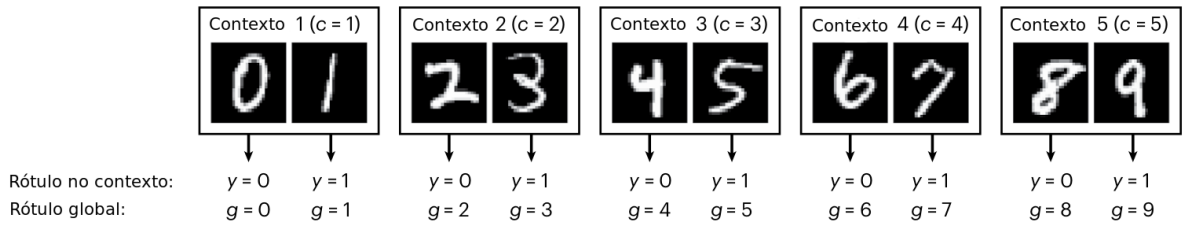
Método	**BWT ↑**	**FWT ↑**
Joint	0,0287	0,5191
Fine Tuning	-0,2870	0,5135
ECgr	-0,1467	0,5323
ECgr+QA	-0,1799	0,5279
ECgr+wQA	-0,1703	**0,5407**
ECgr+*cluster*	-0,1979	0,5343
ECgr+w*cluster*	-0,1631	0,5319
ECgr+CGLO	-0,1952	0,5312
ECgr+wCGLO	-0,1778	0,5266
EWC	**-0,1030**	0,5034

Fonte: autoria própria.

5.3.3 Reconhecimento de dígitos (MNIST)

A fim de averiguar a aplicabilidade do método proposto em um cenário que não o de reconhecimento de expressões faciais, avalia-se o método na base de dados MNIST (DENG, 2012), mais especificamente a Split-MNIST. A Figura 48 ilustra o cenário incremental de tarefas da base de dados Split-MNIST. Essa base é uma variante da MNIST, com as classes subdivididas em pares (0 e 1, 2 e 3, 4 e 5, 6 e 7, 8 e 9). Desta forma, o processo de treinamento incremental neste conjunto de dados consiste em introduzir um novo par a cada etapa, simulando a introdução de uma nova tarefa. É importante notar que os rótulos a serem aprendidos pela CNN são fixos em 0 e 1, enquanto os rótulos globais representam os dígitos de cada imagem, conforme mostra a Figura 48. Assim, é possível avaliar o método proposto diante de um cenário incremental de tarefas e não de classes.

Figura 48 – *Domain incremental* da base de dados Split-MNIST, de acordo com Ven, Tuytelaars e Tolias (2022) (traduzido).



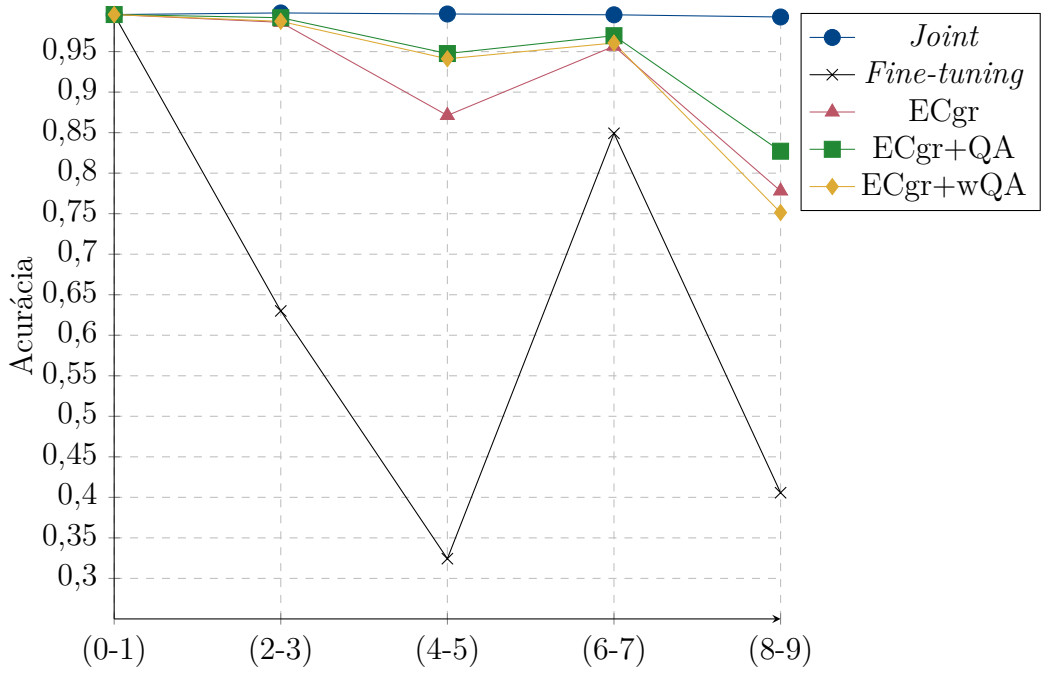
Fonte: adaptado de (VEN; TUYTELAARS; TOLIAS, 2022).

Aplica-se o método ECgr com WGAN-GPs, seguindo os passos descritos no Algoritmo 1. O treinamento iniciou com o par de classes 0 e 1 como base fonte, com pares subsequentes sendo utilizados no processo de aprendizado contínuo. As WGAN-GPs foram treinadas para cada dígito e o mesmo processo de combinação das bases de dados alvo com as bases de dados sintéticas, geradas pelas redes generativas, foi seguido durante o

retreinamento com os métodos ECgr, ECgr+QA e ECgr+wQA. Diante dos resultados obtidos nas bases de dados de FER, optou-se por não aplicar os demais métodos de filtragem nesta avaliação da base MNIST.

Como ilustra a Figura 49, o comportamento previamente observado nas bases de dados de emoção, também se manteve neste domínio. Os métodos ECgr+QA e ECgr+wQA superaram consistentemente o *fine-tuning* em todas as etapas de retreinamento. Em relação à avaliação qualitativa das imagens sintéticas, os dígitos 4 e 5 foram os mais desafiadores de se gerar, e o algoritmo QA apresentou maior dificuldade em lidar com esses dígitos.

Figura 49 – Resultados de acurácia para o subconjunto de dados da classe (0-1) do MNIST, demonstrando a adaptação contínua ao longo dos subconjuntos (2-3), (4-5), (6-7) e (8-9).



Fonte: autoria própria.

Os resultados apresentados na Tabela 37 evidenciam a evolução do desempenho dos métodos de aprendizado contínuo no cenário de domínio incremental aplicado à base Split-MNIST, conforme reportado pelos experimentos de Ven, Tuytelaars e Tolias (2022). A comparação com os métodos propostos neste trabalho é parcialmente válida, uma vez que, apesar da similaridade do protocolo experimental para o cenário de domínio incremental, podem existir variações não reportadas ou detalhes específicos de implementação que dificultam uma comparação sob os mesmos parâmetros. Os valores de acurácia são reportados a partir da média após o último subconjunto ter sido retreinado diante de 20 repetições do experimento.

A partir dos resultados reportados pelos autores Ven, Tuytelaars e Tolias (2022), o treinamento conjunto (*joint*) estabelece o limite superior de desempenho (98,59%). Por

outro lado, o *fine-tuning* sequencial, sem qualquer mecanismo de mitigação de esquecimento, representa o limite inferior (60,13%), refletindo a severidade do *catastrophic forgetting*. Entre os métodos clássicos de regularização (EWC, SI e LwF), observa-se um ganho sobre o *fine-tuning*, indicando que essas abordagens são pouco eficazes nesse cenário. Já os métodos baseados em *replay*, especialmente aqueles que utilizam dados reais ou gerados (DGR, BI-R, ER), alcançam desempenhos próximos ao limite superior, com destaque para BI-R (97,26%) e DGR (95,57%).

Os métodos propostos neste trabalho (ECgr, ECgr+QA e ECgr+wQA) apresentaram resultados competitivos. O ECgr isolado superou os métodos baseados em regularização, com 80,53%. Quando aliado à filtragem supervisionada (ECgr+QA), o desempenho aumentou para 82,00%, evidenciando a contribuição do mecanismo de qualidade no processo de *pseudo-rehearsal*. O resultado inferior do ECgr+wQA (77,83%) sugere que a atribuição de pesos às imagens sintéticas pode prejudicar o treinamento neste contexto.

Tabela 37 – Resultados de métodos de aprendizado contínuo para o cenário de domínio incremental na Split-MNIST, conforme reportado por Ven, Tuytelaars e Tolias (2022).

Método	***Domain-IL***
Fine-tuning – limite inferior	60,13 (±1,66)
Joint – limite superior	98,59 (±0,05)
EWC	63,03 (±1,58)
SI	66,94 (±1,13)
LwF	71,18 (±1,42)
FROMP	84,86 (±1,02)
DGR	95,57 (±0,30)
BI-R	97,26 (±0,15)
ER	93,75 (±0,24)
A-GEM	87,67 (±1,33)
Generative Classifier	93,82 (±0,06)
ECgr	80,53 (±0,00)
ECgr+QA	82,00 (±0,00)
ECgr+wQA	77,83 (±0,00)

Fonte: adaptado de (VEN; TUYTELAARS; TOLIAS, 2022).

Esses resultados reforçam a hipótese alternativa da questão de pesquisa Q1, de que o método ECgr melhora significativamente a acurácia e a retenção do conhecimento em cenários de aprendizado contínuo, e também corroboram a hipótese alternativa da questão Q3, de que o método proposto melhora o desempenho em contextos distintos do reconhecimento de emoções.

5.4 Discussões gerais

Esta seção apresenta uma avaliação geral do método proposto e dos experimentos realizados, à luz das questões de pesquisa que guiam este trabalho, expostas na Seção 1.3.

Primeiramente, os resultados fornecem evidências de que o uso de *pseudo-rehearsal*, em especial o método proposto neste trabalho, contribui para reduzir a perda de memória no cenário de aprendizado contínuo (Q1). Os dados mostram que o método ECgr, isoladamente, melhora a acurácia em comparação ao *fine-tuning*, e seu impacto é ampliado quando combinado com mecanismos de QA. A estratégia proposta demonstrou eficácia ao reduzir os efeitos do esquecimento catastrófico, superando consistentemente os resultados obtidos com *fine-tuning*. A geração de dados sintéticos que reproduzem padrões das tarefas anteriores por meio de WGAN-GPs mostrou-se útil ao permitir que a rede retivesse conhecimento sem recorrer aos dados originais das tarefas passadas.

Ao avaliar os resultados ao longo das diferentes etapas do protocolo experimental no reconhecimento de expressões faciais, observa-se diferença significativa entre o método ECgr e o método *fine-tuning* em todas as etapas (Figuras 37, 38 e 41). Esses resultados indicam que o uso de *generative replay* contribui para preservar o desempenho do modelo ao longo do aprendizado incremental.

É importante observar que a etapa de geração sintética exerce papel central na mitigação do esquecimento catastrófico. Ao produzir amostras que reproduzem a distribuição das tarefas anteriores, o sistema mantém o espaço de representação ativo, evitando o esquecimento de determinadas representações essenciais para a distinção entre as classes. Quando essa etapa é suprimida, observa-se que o modelo tende a superajustar-se para as classes mais recentes, com perda gradual de acurácia nas anteriores. Assim, mesmo sem acesso direto aos dados originais, a geração sintética atua como uma forma de consolidação da memória, essencial para preservar o desempenho incremental.

Há de se discutir, no entanto, que a abordagem proposta para gerar imagens sintéticas adiciona ao sistema a complexidade de treinar as WGAN-GPs para cada classe de cada base de dados previamente aprendida. Isto é, dado k tarefas com n classes cada, o total de WGAN-GPs necessárias será de $k \cdot n$, o que pode não ser viável em cenários dinâmicos em que o número de bases imponha uma limitação na capacidade de treinamento destes geradores.

Outro aspecto relevante investigado neste trabalho diz respeito ao impacto da filtragem de imagens sintéticas na redução do esquecimento catastrófico (Q2). Os resultados mostram que a introdução de um mecanismo de QA durante o *replay* generativo pode melhorar a qualidade das amostras utilizadas no treinamento incremental.

A abordagem supervisionada (QA e wQA) apresentou diferenças estatisticamente significativas em relação ao método *fine-tuning* em todas as etapas do treinamento. Por

outro lado, o método baseado em clusterização (*cluster* e w*cluster*) não apresentou diferenças estatisticamente significativas em relação ao *fine-tuning* nas etapas avaliadas. Já o método baseado em otimização do espaço latente da WGAN-GP (CGLO e wCGLO) apresentou diferença significativa apenas na última etapa do treinamento.

Esses resultados sugerem que a eficácia da filtragem depende fortemente do método utilizado para estimar a qualidade das imagens sintéticas, sendo o método supervisionado baseado em CNN o que apresentou maior consistência nos cenários avaliados.

Cabe destacar que a eficácia dos métodos de filtragem estão diretamente atreladas à qualidade das imagens geradas pelas WGAN-GPs, sendo que, a partir do momento que as WGAN-GPs forem aprimoradas à ponto que a métrica FID-FNET ≈ 0, com o intuito de gerar imagens as mais similares possíveis às da base de dados original na qual a rede generativa foi treinada, a contribuição do método de QA pode diminuir.

Por fim, em relação à adaptação do método proposto para cenários distintos do reconhecimento de expressões faciais (Q3), o método foi avaliado na base de dados MNIST e apresentou resultados satisfatórios, corroborando os achados obtidos no cenário de reconhecimento de emoções, nos quais o método ECgr+QA se mostrou superior nos experimentos realizados. A base de dados MNIST foi dividida em subgrupos de classes, com os dígitos em pares, 0 e 1, 2 e 3 e assim por diante. Dessa forma, este cenário apresentou cinco bases de dados para o aprendizado contínuo. Ainda assim, pôde-se perceber que o método proposto obteve uma *performance* acima do método *fine-tuning*.

6 Conclusão

Neste trabalho, investiga-se o problema do esquecimento catastrófico em redes neurais convolucionais (CNNs) aplicadas ao reconhecimento de expressões faciais no contexto de aprendizado contínuo. Como principal contribuição, propõe-se o método *Emotion-Centered generative replay* (ECgr), uma abordagem original baseada em *pseudo-rehearsal* com dados sintéticos gerados por WGAN-GP, capaz de operar sem a reutilização dos dados reais das tarefas anteriores. Diferentemente de trabalhos anteriores, o ECgr incorpora um mecanismo de seleção de amostras sintéticas com foco na preservação de conhecimento sem violar restrições de privacidade ou armazenamento. Para isso, foram propostas e avaliadas três estratégias inéditas de garantia de qualidade (QA) das amostras geradas, que exploram diferentes níveis de supervisão e guiam o processo de *replay* com foco na retenção de conhecimento relevante: (i) um método supervisionado, que utiliza a própria CNN previamente treinada para validar a consistência das imagens geradas; (ii) um método não supervisionado, baseado em clusterização para inferência de rótulos; e (iii) uma técnica de otimização do espaço latente das WGAN-GPs, guiada por imagens sintéticas previamente validadas. Essa combinação entre geração centrada em emoção e filtragem de qualidade representa um avanço na literatura de aprendizado contínuo com redes generativas, especialmente em domínios com sensibilidade ética, como o reconhecimento de emoções humanas.

Os resultados experimentais mostraram que o método ECgr, de forma isolada, já proporciona ganhos expressivos em relação ao *fine-tuning* tradicional. Quando combinado com o filtro de QA supervisionado (ECgr+QA), o método apresentou desempenho ainda mais robusto, mitigando significativamente os efeitos do esquecimento catastrófico. A estratégia proposta se aproximou dos resultados obtidos pelo treinamento conjunto (*joint training*), mesmo sem utilizar os dados originais das tarefas anteriores, o que reforça sua aplicabilidade em cenários onde o armazenamento contínuo de dados não é viável.

A partir das questões de pesquisa levantadas, os experimentos realizados permitiram avaliar o comportamento do método proposto em diferentes cenários de aprendizado contínuo.

Em relação à Q1, que investigava se o método ECgr seria capaz de melhorar a acurácia e retenção do conhecimento das CNNs em aprendizado contínuo, os resultados mostram que o ECgr apresentou desempenho superior ao *fine-tuning* em todas as etapas avaliadas, com diferenças estatisticamente significativas em duas das três etapas experimentais.

Quanto à Q2, que avaliava se a filtragem das imagens sintéticas por qualidade

contribuiria para reduzir ainda mais o esquecimento catastrófico, observou-se que apenas os métodos de QA supervisionado (QA e wQA) apresentaram melhorias estatisticamente significativas em relação ao *fine-tuning*. As demais estratégias de filtragem, baseadas em clusterização (*cluster* e w*cluster*) e otimização do espaço latente (CGLO e wCGLO), apresentaram resultados menos consistentes.

No que se refere à Q3, que investigou a generalização do método para além do reconhecimento de emoções, o método proposto foi aplicado à base MNIST em um cenário de aprendizado incremental. Os resultados obtidos nessa base indicam que a abordagem proposta também pode ser aplicada a outros domínios de classificação, apresentando desempenho superior ao método *fine-tuning*.

Os experimentos também indicaram que a eficácia dos filtros de QA está diretamente relacionada à qualidade das imagens geradas pelas WGAN-GPs. Assim, melhorias na capacidade gerativa desses modelos tendem a reduzir a dependência de mecanismos de filtragem. Também verificou-se que, apesar dos bons resultados no reconhecimento de expressões faciais, o método ECgr+QA demonstrou generalização ao ser aplicado com sucesso à base MNIST, evidenciando sua adaptabilidade a outras tarefas de classificação.

De modo geral, os resultados confirmam que a interação entre geração sintética e filtragem supervisionada é o principal fator responsável pela robustez observada. Enquanto a geração permite reter representações antigas de forma controlada, a filtragem atua como mecanismo de proteção contra deriva semântica, assegurando que apenas exemplos plausíveis contribuam para o processo de *pseudo-rehearsal*. Essa integração explica o desempenho superior do ECgr+QA em relação às variantes sem filtro e ao método de *fine-tuning*, consolidando a proposta como uma estratégia consistente para aprendizado contínuo sem acesso a dados originais.

A abordagem proposta também apresenta limitações. A necessidade de treinar uma WGAN-GP para cada classe e para cada tarefa aumenta o custo computacional do método proposto, o que pode dificultar sua aplicação em cenários com grande número de classes ou tarefas sucessivas. Essa escolha foi adotada para permitir maior controle experimental sobre a qualidade das amostras sintéticas geradas por classe. Ainda assim, essa limitação sugere caminhos promissores para investigações futuras, como o uso de modelos gerativos mais eficientes, mecanismos de compartilhamento de geradores entre classes e integração com métodos de regularização, como *weight consolidation*. Também, pretende-se investigar o possível impacto da interferência entre amostras sintéticas geradas em diferentes etapas do treinamento incremental, analisando se a introdução progressiva desses dados pode influenciar o processo de esquecimento gradual ao longo das tarefas.

Uma direção promissora para investigações futuras consiste na exploração de modelos generativos mais recentes, como os *Vision-Language Models* (VLMs) ou modelos de difusão condicionados por texto, para a geração de amostras sintéticas. Diferentemente

da abordagem adotada neste trabalho, baseada em múltiplas WGAN-GPs treinadas por classe, esses modelos poderiam permitir a geração controlada de imagens a partir de descrições semânticas das expressões faciais, potencialmente reduzindo a necessidade de treinamento de diversos geradores independentes. Considerando que o processo de geração de dados sintéticos é realizado de forma *offline* no método proposto, a adoção desses modelos pode representar uma alternativa interessante a ser investigada em estudos futuros.

Outra possibilidade para trabalhos futuros consiste em investigar estratégias alternativas para o método CGLO, de amostragem do espaço latente utilizado pelos geradores. No método proposto CGLO, os vetores latentes são amostrados a partir de uma distribuição gaussiana padrão, conforme prática comum em modelos GAN. Entretanto, abordagens alternativas poderiam explorar distribuições estimadas a partir dos dados reais ou de representações latentes extraídas por modelos previamente treinados, utilizando técnicas como *Kernel Density Estimation* ou modelagem da distribuição das *features* produzidas por uma CNN treinada na base de dados original. Essa estratégia poderia favorecer uma melhor correspondência entre o espaço latente e a distribuição dos dados reais, além de permitir cenários em que apenas representações compactas dos dados são armazenadas, contribuindo potencialmente para redução de armazenamento e maior preservação da privacidade.

Em suma, este trabalho contribui para o avanço de abordagens baseadas em *pseudo-rehearsal* com dados sintéticos no aprendizado contínuo, destacando o potencial do *replay* generativo. Os resultados obtidos reafirmam a possibilidade de atenuar significativamente o esquecimento catastrófico sem depender dos dados originais das tarefas anteriores, abrindo novas possibilidades para sistemas de aprendizado incremental em contextos sensíveis à privacidade, memória ou tempo de execução.

Referências

ADEL, Tameem; ZHAO, Han; TURNER, Richard E. *Continual Learning with Adaptive Weights (CLAW)*. 2020. Disponível em: <https://arxiv.org/abs/1911.09514>.

AIFANTI, Niki; PAPACHRISTOU, Christos; DELOPOULOS, Anastasios. The mug facial expression database. In: *11th Intl Works Image Anal Multim Interac Services*. [S.l.: s.n.], 2010. p. 1–4.

ALEIXO, Everton Lima; COLONNA, Juan G.; CRISTO, Marco; FERNANDES, Everlandio. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *Journal of the Brazilian Computer Society*, v. 30, n. 1, p. 175–211, Aug. 2024. Disponível em: <https://journals-sol.sbc.org.br/index.php/jbcs/article/view/3966>.

ALJUNDI, Rahaf; BABILONI, Francesca; ELHOSEINY, Mohamed; ROHRBACH, Marcus; TUYTELAARS, Tinne. Memory aware synapses: Learning what (not) to forget. In: FERRARI, Vittorio; HEBERT, Martial; SMINCHISESCU, Cristian; WEISS, Yair (Ed.). *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018. p. 144–161. ISBN 978-3-030-01219-9.

ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.]: MIT Press, 2010. (Adaptive computation and machine learning). ISBN 9780262012430.

ANAYA-SáNCHEZ, Héctor; ALTAMIRANO-ROBLES, Leopoldo; DíAZ-HERNáNDEZ, Raquel; ZAPOTECAS-MARTíNEZ, Saúl. Wgan-gp for synthetic retinal image generation: Enhancing sensor-based medical imaging for classification models. *Sensors*, v. 25, n. 1, 2025. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/25/1/167>.

ANTONI, Olivier; MAINSANT, Marion; GODIN, Christelle; MERMILLOD, Martial; REYBOZ, Marina. An embedded continual learning system for facial emotion recognition. In: AMINI, Massih-Reza; CANU, Stéphane; FISCHER, Asja; GUNS, Tias; NOVAK, Petra Kralj; TSOUMAKAS, Grigorios (Ed.). *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer Nature Switzerland, 2023. p. 631–635. ISBN 978-3-031-26422-1.

ARJOVSKY, Martin; CHINTALA, Soumith; BOTTOU, Léon. Wasserstein generative adversarial networks. In: PRECUP, Doina; TEH, Yee Whye (Ed.). *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 214–223. Disponível em: <https://proceedings.mlr.press/v70/arjovsky17a.html>.

BENITEZ-QUIROZ, C. Fabian; SRINIVASAN, Ramprakash; MARTINEZ, Aleix M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 5562–5570.

BIE, Mei; XU, Huan; LIU, Quanle; GAO, Yan; SONG, Kai; CHE, Xiangjiu. Da-fer: Domain adaptive facial expression recognition. *Applied Sciences*, v. 13, n. 10, 2023. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/13/10/6314>.

BISHOP, Christopher M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1. ed. Springer, 2007. ISBN 0387310738. Disponível em: <http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738>.

BRADSKI, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

BUZZEGA, Pietro; BOSCHINI, Matteo; PORRELLO, Angelo; ABATI, Davide; CALDERARA, SIMONE. Dark experience for general continual learning: a strong, simple baseline. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.F.; LIN, H. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 15920–15930. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf>.

CAI, Yuliang; ROSTAMI, Mohammad. *Dynamic Transformer Architecture for Continual Learning of Multimodal Tasks*. 2024. Disponível em: <https://arxiv.org/abs/2401.15275>.

CAMP, Blake; MANDIVARAPU, Jaya Krishna; ESTRADA, Rolando. *Self-Net: Lifelong Learning via Continual Self-Modeling*. 2019. Disponível em: <https://arxiv.org/abs/1805.10354>.

CARPENTER, G.A.; GROSSBERG, S. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, v. 21, n. 3, p. 77–88, 1988.

CHAUDHARI, Aayushi; BHATT, Chintan; KRISHNA, Achyut; MAZZEO, Pier Luigi. Vitfer: Facial emotion recognition with vision transformers. *Applied System Innovation*, v. 5, n. 4, 2022. ISSN 2571-5577. Disponível em: <https://www.mdpi.com/2571-5577/5/4/80>.

CHAUDHRY, Arslan; DOKANIA, Puneet K.; AJANTHAN, Thalaiyasingam; TORR, Philip H. S. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: _____. *Computer Vision – ECCV 2018*. Springer International Publishing, 2018. p. 556–572. ISBN 9783030012526. Disponível em: <http://dx.doi.org/10.1007/978-3-030-01252-6_33>.

CHEN, Chien-Chung; CHO, Shu ling; HORSZOWSKA, Katarzyna; CHEN, Mei-Yen; WU, Chia-Ching; CHEN, Hsueh-Chih; YEH, Yi-Yu; CHENG, Chao-Min. A facial expression image database and norm for Asian population: a preliminary report. In: FARNAND, Susan P.; GAYKEMA, Frans (Ed.). *Image Quality and System Performance VI*. [S.l.]: SPIE, 2009. v. 7242, p. 72421D.

CHEUNG, Brian; TEREKHOV, Alexander; CHEN, Yubei; AGRAWAL, Pulkit; OLSHAUSEN, Bruno. Superposition of many models into one. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf>.

CHURAMANI, Nikhil; GUNES, Hatice. Clifer: Continual learning with imagination for facial expression recognition. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. [S.l.: s.n.], 2020. p. 322–328.

CHURAMANI, Nikhil; KARA, Ozgur; GUNES, Hatice. Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition . *IEEE Transactions on Affective Computing*, IEEE Computer Society, Los Alamitos, CA, USA, v. 14, n. 04, p. 3191–3206, out. 2023. ISSN 1949-3045. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TAFFC.2022.3181033>.

COTOGNI, Marco; YANG, Fei; CUSANO, Claudio; BAGDANOV, Andrew D.; WEIJER, Joost van de. Exemplar-free continual learning of vision transformers via gated class-attention and cascaded feature drift compensation. *International Journal of Computer Vision*, v. 133, n. 7, p. 4571–4589, 2025. ISSN 1573-1405. Disponível em: <https://doi.org/10.1007/s11263-025-02374-x>.

DEMŠAR, Janez. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006.

DENG, Li. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, v. 29, n. 6, p. 141–142, 2012.

DEVRIES, Terrance; BISWARANJAN, Kumar; TAYLOR, Graham W. Multi-task learning of facial landmarks and expression. In: *2014 Canadian Conference on Computer and Robot Vision*. [S.l.: s.n.], 2014. p. 98–103.

DHALL, Abhinav; GOECKE, Roland; JOSHI, Jyoti; WAGNER, Michael; GEDEON, Tom. Collecting large, richly annotated facial-expression databases from movies. In: *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. [S.l.: s.n.], 2012. p. 1–6.

DHALL, Abhinav; MURTHY, O.V. Ramana; GOECKE, Roland; JOSHI, Jyoti; GEDEON, Tom. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2015. (ICMI '15), p. 423–426. ISBN 9781450339124. Disponível em: <https://doi.org/10.1145/2818346.2829994>.

DIBEKLIOGLU, Hamdi; HAMMAL, Zakia; YANG, Ying; COHN, Jeffrey F. Multimodal detection of depression in clinical interviews. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2015. (ICMI '15), p. 307–310. ISBN 9781450339124. Disponível em: <https://doi.org/10.1145/2818346.2820776>.

DOSOVITSKIY, Alexey; BEYER, Lucas; KOLESNIKOV, Alexander; WEISSENBORN, Dirk; ZHAI, Xiaohua; UNTERTHINER, Thomas; DEHGHANI, Mostafa; MINDERER, Matthias; HEIGOLD, Georg; GELLY, Sylvain; USZKOREIT, Jakob; HOULSBY, Neil. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

DOUILLARD, Arthur; RAMé, Alexandre; COUAIRON, Guillaume; CORD, Matthieu. Dytox: Transformers for continual learning with dynamic token expansion. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 9275–9285.

DU, Shichuan; TAO, Yong; MARTINEZ, Aleix M. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, v. 111, n. 15, p. E1454–E1462, 2014. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.1322355111>.

EITZ, Mathias; HAYS, James; ALEXA, Marc. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, v. 31, n. 4, p. 44:1–44:10, 2012.

EKMAN, Paul. An argument for basic emotions. *Cognition and Emotion*, Routledge, v. 6, n. 3-4, p. 169–200, 1992. Disponível em: <https://doi.org/10.1080/02699939208411068>.

EKMAN, Paul; FRIESEN, Wallace V.; HAGER, Joseph C. *Facial Action Coding System. Manual and Investigator's Guide*. Salt Lake City, UT: Research Nexus, 2002.

ERMIS, Beyza; ZAPPELLA, Giovanni; WISTUBA, Martin; RAWAL, Aditya; ARCHAMBEAU, Cédric. Memory efficient continual learning with transformers. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2022. (NIPS '22). ISBN 9781713871088.

FERNANDO, Chrisantha; BANARSE, Dylan; BLUNDELL, Charles; ZWOLS, Yori; HA, David; RUSU, Andrei A.; PRITZEL, Alexander; WIERSTRA, Daan. *PathNet: Evolution Channels Gradient Descent in Super Neural Networks*. 2017. Disponível em: <https://arxiv.org/abs/1701.08734>.

FRÉCHET, Maurice. Sur la distance de deux lois de probabilité. *Annales de l'ISUP*, Publications de l'Institut de Statistique de l'Université de Paris, VI, n. 3, p. 183–198, 1957. Disponível em: <https://hal.science/hal-04093677>.

FRIEDMAN, Milton. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, ASA Website, v. 32, n. 200, p. 675–701, 1937. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522>.

GAO, Hongxiang; WU, Min; CHEN, Zhenghua; LI, Yuwen; WANG, Xingyao; AN, Shan; LI, Jianqing; LIU, Chengyu. Ssa-icl: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition. *Neural Networks*, v. 158, p. 228–238, 2023. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S089360802200466X>.

GAO, Yuefang; XIE, Yuhao; HU, Zeke Zexi; CHEN, Tianshui; LIN, Liang. Adaptive global-local representation learning and selection for cross-domain facial expression recognition. *IEEE Transactions on Multimedia*, v. 26, p. 6676–6688, 2024.

GOODFELLOW, Ian; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative adversarial nets. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE, N.; WEINBERGER, K.Q. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. v. 27. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.

GOODFELLOW, Ian J.; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOODFELLOW, Ian J.; ERHAN, Dumitru; CARRIER, Pierre Luc; COURVILLE, Aaron; MIRZA, Mehdi; HAMNER, Ben; CUKIERSKI, Will; TANG, Yichuan; THALER, David; LEE, Dong-Hyun; ZHOU, Yingbo; RAMAIAH, Chetan; FENG, Fangxiang; LI, Ruifan; WANG, Xiaojie; ATHANASAKIS, Dimitris; SHAWE-TAYLOR, John; MILAKOV,

Maxim; PARK, John; IONESCU, Radu; POPESCU, Marius; GROZEA, Cristian; BERGSTRA, James; XIE, Jingjing; ROMASZKO, Lukasz; XU, Bing; CHUANG, Zhang; BENGIO, Yoshua. Challenges in representation learning: A report on three machine learning contests. In: LEE, Minho; HIROSE, Akira; HOU, Zeng-Guang; KIL, Rhee Man (Ed.). *Neural Information Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 117–124. ISBN 978-3-642-42051-1.

GOODFELLOW, Ian J.; MIRZA, Mehdi; XIAO, Da; COURVILLE, Aaron; BENGIO, Yoshua. *An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks*. 2013. Disponível em: <https://arxiv.org/abs/1312.6211>.

GULRAJANI, Ishaan; AHMED, Faruk; ARJOVSKY, Martin; DUMOULIN, Vincent; COURVILLE, Aaron. Improved training of wasserstein gans. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 5769–5779. ISBN 9781510860964.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining concepts and techniques, third edition. Morgan Kaufmann Publishers, Waltham, Mass., 2012. Disponível em: <http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>.

HAN, Yizeng; HUANG, Gao; SONG, Shiji; YANG, Le; WANG, Honghui; WANG, Yulin. Dynamic Neural Networks: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 44, n. 11, p. 7436–7456, nov. 2022. ISSN 1939-3539. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3117837>.

HE, Jiangpeng. Gradient reweighting: Towards imbalanced class-incremental learning. In: . [S.l.: s.n.], 2024. p. 16668–16677.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.

HEBB, Donald O. *The organization of behavior: A neuropsychological theory*. New York: Wiley, 1949. Hardcover. ISBN 0-8058-4300-0.

HENSCH, Takao K.; FAGIOLINI, Michela; MATAGA, Nobuko; STRYKER, Michael P.; BAEKKESKOV, Steinunn; KASH, Shera. Local gaba circuit control of experience-dependent plasticity in developing visual cortex. *Science*, v. 282 5393, p. 1504–8, 1998. Disponível em: <https://api.semanticscholar.org/CorpusID:18223265>.

HEUSEL, Martin; RAMSAUER, Hubert; UNTERTHINER, Thomas; NESSLER, Bernhard; HOCHREITER, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: GUYON, Isabelle; LUXBURG, Ulrike von; BENGIO, Samy; WALLACH, Hanna M.; FERGUS, Rob; VISHWANATHAN, S. V. N.; GARNETT, Roman (Ed.). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. [s.n.], 2017. p. 6626–6637. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html>.

HUANG, Gao; LIU, Zhuang; MAATEN, Laurens Van Der; WEINBERGER, Kilian Q. Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 2261–2269.

HUNG, Steven C. Y.; TU, Cheng-Hao; WU, Cheng-En; CHEN, Chien-Hung; CHAN, Yi-Ming; CHEN, Chu-Song. Compacting, picking and growing for unforgetting continual learning. In: _____. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

IZARD, Carroll. *Human Emotions*. Springer US, 1977. (Emotions, Personality, and Psychotherapy). ISBN 9781489922090. Disponível em: <https://doi.org/10.1007/978-1-4899-2209-0>.

JOYCE, James M. Kullback-leibler divergence. In: _____. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 720–722. ISBN 978-3-642-04898-2. Disponível em: <https://doi.org/10.1007/978-3-642-04898-2_327>.

KARA, Ozgur; CHURAMANI, Nikhil; GUNES, Hatice. *Towards Fair Affective Robotics: Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition*. 2021. Disponível em: <https://arxiv.org/abs/2103.09233>.

KARAM, Said; RUAN, Shanq-Jang; HAQ, Qazi Mazhar ul; LI, Lieber Po-Hung. Episodic memory based continual learning without catastrophic forgetting for environmental sound classification. *Journal of Ambient Intelligence and Humanized Computing*, v. 14, n. 4, p. 4439–4449, 2023. ISSN 1868-5145. Disponível em: <https://doi.org/10.1007/s12652-023-04561-5>.

KARRAS, Tero; LAINE, Samuli; AILA, Timo. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 43, n. 12, p. 4217–4228, dez. 2021. ISSN 1939-3539. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.2970919>.

KEMKER, Ronald; KANAN, Christopher. Fearnet: Brain-inspired model for incremental learning. *CoRR*, abs/1711.10563, 2017. Disponível em: <http://arxiv.org/abs/1711.10563>.

KHETARPAL, Khimya; RIEMER, Matthew; RISH, Irina; PRECUP, Doina. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, v. 75, p. 1401–1476, 2022.

KIM, Chris; JEONG, Jinseo; KIM, Gunhee. Imbalanced continual learning with partitioning reservoir sampling. In: _____. [S.l.: s.n.], 2020. p. 411–428. ISBN 978-3-030-58600-3.

KING, Davis E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, v. 10, p. 1755–1758, 2009.

KIRKPATRICK, James; AL. et. Overcoming catastrophic forgetting in neural networks. *Proc. of the National Academy of Sciences*, v. 114, n. 13, p. 3521–3526, 2017.

KRAUSE, Jonathan; STARK, Michael; DENG, Jia; FEI-FEI, Li. 3d object representations for fine-grained categorization. In: *2013 IEEE International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2013. p. 554–561.

KRIZHEVSKY, Alex. Learning multiple layers of features from tiny images. p. 32–33, 2009. Disponível em: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.

KRIZHEVSKY, Alex; NAIR, Vinod; HINTON, Geoffrey. Cifar-10 (canadian institute for advanced research). 2009. Disponível em: <http://www.cs.toronto.edu/~kriz/cifar.html>.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F.; BURGES, C.J.; BOTTOU, L.; WEINBERGER, K.Q. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. v. 25. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.

LAKE, Brenden M.; SALAKHUTDINOV, Ruslan; TENENBAUM, Joshua B. Human-level concept learning through probabilistic program induction. *Science*, v. 350, n. 6266, p. 1332–1338, 2015. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aab3050>.

LAMPERT, Christoph H.; NICKISCH, Hannes; HARMELING, Stefan. Learning to detect unseen object classes by between-class attribute transfer. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 951–958.

LANGE, Matthias De; ALJUNDI, Rahaf; MASANA, Marc; PARISOT, Sarah; JIA, Xu; LEONARDIS, Aleš; SLABAUGH, Gregory; TUYTELAARS, Tinne. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 44, n. 7, p. 3366–3385, 2022.

LEA, Colin; FLYNN, Michael D.; VIDAL, René; REITER, Austin; HAGER, Gregory D. Temporal convolutional networks for action segmentation and detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 1003–1012.

LEE, Cecilia S.; LEE, Aaron Y. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, Elsevier, v. 2, n. 6, p. e279–e281, 2020. ISSN 2589-7500. Disponível em: <https://doi.org/10.1016/S2589-7500(20)30102-3>.

LEE, Sang-Woo; KIM, Jin-Hwa; JUN, Jaehyun; HA, Jung-Woo; ZHANG, Byoung-Tak. *Overcoming Catastrophic Forgetting by Incremental Moment Matching*. 2018. Disponível em: <https://arxiv.org/abs/1703.08475>.

LI, Hanting; SUI, Mingzhe; ZHU, Zhaoqing; ZHAO, Feng. *NR-DFERNet: Noise-Robust Network for Dynamic Facial Expression Recognition*. 2022. Disponível em: <https://arxiv.org/abs/2206.04975>.

LI, Shan; DENG, Weihong. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, v. 13, n. 3, p. 1195–1215, 2022.

LI, Shan; DENG, Weihong; DU, JunPing. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: IEEE. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2017. p. 2584–2593.

LI, Zhizhong; HOIEM, Derek. Learning without forgetting. *IEEE Trans on Pattern Analysis and Machine Intelligence*, v. 40, n. 12, p. 2935–2947, 2018.

LIU, Gaqiong; HUANG, Shucheng; WANG, Gang; LI, Mingxing. Emrnet: Enhanced micro-expression recognition network with attention and distance correlation. *Artificial Intelligence Review*, v. 58, n. 6, p. 176, 2025. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-025-11159-0>.

LIU, Ping; HAN, Shizhong; MENG, Zibo; TONG, Yan. Facial expression recognition via a boosted deep belief network. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1805–1812.

LIU, Xialei; WU, Chenshen; MENTA, Mikel; HERRANZ, Luis; RADUCANU, Bogdan; BAGDANOV, Andrew D.; JUI, Shangling; WEIJER, Joost van de. Generative feature replay for class-incremental learning. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [S.l.: s.n.], 2020. p. 915–924.

LIU, Zhuang; MAO, Hanzi; WU, Chao-Yuan; FEICHTENHOFER, Christoph; DARRELL, Trevor; XIE, Saining. A convnet for the 2020s. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 11966–11976.

LOPEZ-PAZ, David; RANZATO, Marc' Aurelio. Gradient episodic memory for continual learning. In: GUYON, I.; LUXBURG, U. Von; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf>.

LUCEY, Patrick; COHN, Jeffrey F.; KANADE, Takeo; SARAGIH, Jason; AMBADAR, Zara; MATTHEWS, Iain. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: *IEEE CVPR Workshops*. [S.l.: s.n.], 2010. p. 94–101.

LYONS, M.; KAMACHI, M.; GYOBA, J. The japanese female facial expression (jaffe) dataset. *Zenodo*, abr. 1998.

MA, Fuyan; SUN, Bin; LI, Shutao. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, v. 14, n. 2, p. 1236–1248, 2023.

Machine Elf 1735. *Plutchik's Wheel of Emotions*. 2017. Wikimedia Commons. Domínio público. Disponível em: <https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg>.

MAINSANT, Marion; SOLINAS, Miguel; REYBOZ, Marina; GODIN, Christelle; MERMILLOD, Martial. Dream net: a privacy preserving continual leaming model for face emotion recognition. In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. [S.l.: s.n.], 2021. p. 01–08.

MCCLELLAND, James L; MCNAUGHTON, Bruce L; O'REILLY, Randall C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, American Psychological Association (APA), v. 102, p. 419–457, 1995.

MCCLOSKEY, Michael; COHEN, Neal J. Catastrophic interference in connectionist networks: The sequential learning problem. In: BOWER, Gordon H. (Ed.). Academic Press, 1989, (Psychology of Learning and Motivation, v. 24). p. 109–165. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0079742108605368>.

MCDONNELL, Mark; GONG, Dong; PARVANEH, Amin; ABBASNEJAD, Ehsan; HENGEL, Anton van den. RanPAC: Random projections and pre-trained models for continual learning. In: *Thirty-seventh Conference on Neural Information Processing Systems*. [s.n.], 2023. Disponível em: <https://openreview.net/forum?id=aec58UfBzA>.

MENG, Debin; PENG, Xiaojiang; WANG, Kai; QIAO, Yu. Frame attention networks for facial expression recognition in videos. In: *2019 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2019. p. 3866–3870.

MENON, Luciana Trinkaus; NEDUZIAK, Luiz Carlos Ribeiro; BARDDAL, Jean Paul; KOERICH, Alessandro Lameiras; JUNIOR, Alceu de Souza Britto. Dynamic modality and view selection for emotion recognition: An experimental study on missing modality evaluation. In: PALAIAHNAKOTE, Shivakumara; SCHUCKERS, Stephanie; OGIER, Jean-Marc; BHATTACHARYA, Prabir; PAL, Umapada; BHATTACHARYA, Saumik (Ed.). *Pattern Recognition. ICPR 2024 International Workshops and Challenges - Kolkata, India, December 1, 2024, Proceedings, Part IV*. Springer, 2024. (Lecture Notes in Computer Science, v. 15617), p. 151–166. Disponível em: <https://doi.org/10.1007/978-3-031-88217-3_11>.

MENéNDEZ, M.L.; PARDO, J.A.; PARDO, L.; PARDO, M.C. The jensen-shannon divergence. *Journal of the Franklin Institute*, v. 334, n. 2, p. 307–318, 1997. ISSN 0016-0032. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016003296000634>.

MITCHELL, Tom M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. v. 1.

MOLLAHOSSEINI, Ali; HASANI, Behzad; MAHOOR, Mohammad H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, v. 10, n. 1, p. 18–31, 2019.

NEMENYI, Peter. *Distribution-free multiple comparisons*. Tese (Ph.D. dissertation) — Princeton University, Princeton, NJ, USA, 1963.

NETZER, Yuval; WANG, Tao; COATES, Adam; BISSACCO, Alessandro; WU, Bo; NG, Andrew Y. Reading digits in natural images with unsupervised feature learning. In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. [s.n.], 2011. Disponível em: <http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf>.

NGUYEN, Cuong V.; LI, Yingzhen; BUI, Thang D.; TURNER, Richard E. *Variational Continual Learning*. 2018. Disponível em: <https://arxiv.org/abs/1710.10628>.

NILSBACK, Maria-Elena; ZISSERMAN, Andrew. Automated flower classification over a large number of classes. In: *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*. [S.l.: s.n.], 2008. p. 722–729.

PANTIC, M.; ROTHKRANTZ, L.J.M. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 12, p. 1424–1445, 2000.

PANTIC, M.; VALSTAR, M.; RADEMAKER, R.; MAAT, L. Web-based database for facial expression analysis. In: *2005 IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2005. p. 5 pp.–.

PARISI, German I.; KEMKER, Ronald; PART, Jose L.; KANAN, Christopher; WERMTER, Stefan. Continual lifelong learning with neural networks: A review. *Neural Networks*, v. 113, p. 54–71, 2019. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608019300231>.

PARKHI, Omkar; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep face recognition. In: . [S.l.: s.n.], 2015. v. 1, p. 41.1–41.12.

PHAM, Quang; LIU, Chenghao; HOI, Steven C. H. Continual Learning, Fast and Slow . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 46, n. 01, p. 134–149, jan. 2024. ISSN 1939-3539. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3324203>.

PICZAK, Karol J. Esc: Dataset for environmental sound classification. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2015. (MM '15), p. 1015–1018. ISBN 9781450334594. Disponível em: <https://doi.org/10.1145/2733373.2806390>.

PLUTCHIK, Robert. A psychoevolutionary theory of emotions. *Social Science Information*, v. 21, n. 4-5, p. 529–553, 1982. Disponível em: <https://doi.org/10.1177/053901882021004003>.

RADFORD, Alec; METZ, Luke; CHINTALA, Soumith. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. Disponível em: <https://arxiv.org/abs/1511.06434>.

RADOSAVOVIC, Ilija; KOSARAJU, Raj Prateek; GIRSHICK, Ross; HE, Kaiming; DOLLáR, Piotr. Designing network design spaces. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2020. p. 10425–10433.

RAO, Dushyant; VISIN, Francesco; RUSU, Andrei A.; TEH, Yee Whye; PASCANU, Razvan; HADSELL, Raia. *Continual Unsupervised Representation Learning*. 2019. Disponível em: <https://arxiv.org/abs/1910.14481>.

RATCLIFF, Roger. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, v. 97, n. 2, p. 285—308, April 1990. ISSN 0033-295X. Disponível em: <https://doi.org/10.1037/0033-295x.97.2.285>.

REBUFFI, Sylvestre-Alvise; KOLESNIKOV, Alexander; SPERL, Georg; LAMPERT, Christoph H. icarl: Incremental classifier and representation learning. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 5533–5542.

RIEMER, Matthew; CASES, Ignacio; AJEMIAN, Robert; LIU, Miao; RISH, Irina; TU, Yuhai; TESAURO, Gerald. *Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference*. 2019. Disponível em: <https://arxiv.org/abs/1810.11910>.

ROBINS, Anthony. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, Taylor & Francis, v. 7, n. 2, p. 123–146, 1995. Disponível em: <https://doi.org/10.1080/09540099550039318>.

ROLNICK, David; AHUJA, Arun; SCHWARZ, Jonathan; LILLICRAP, Timothy P.; WAYNE, Greg. *Experience Replay for Continual Learning*. 2019. Disponível em: <https://arxiv.org/abs/1811.11682>.

RUSSELL, J.A. A circumplex model of affect. *Journal of personality and social psychology*, v. 39, n. 6, p. 1161–1178, 1980. ISSN 0022-3514.

RUSU, Andrei A.; RABINOWITZ, Neil C.; DESJARDINS, Guillaume; SOYER, Hubert; KIRKPATRICK, James; KAVUKCUOGLU, Koray; PASCANU, Razvan; HADSELL, Raia. *Progressive Neural Networks*. 2022. Disponível em: <https://arxiv.org/abs/1606.04671>.

SADAK, Hany M.; KHALAF, Ashraf A. M.; SALAMA, Gerges M. Dana: Deep attention network architecture for facial emotions recognition using limited resources. In: *2024 International Telecommunications Conference (ITC-Egypt)*. [S.l.: s.n.], 2024. p. 44–49.

SALAMON, Justin; JACOBY, Christopher; BELLO, Juan Pablo. A dataset and taxonomy for urban sound research. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2014. (MM '14), p. 1041–1044. ISBN 9781450330633. Disponível em: <https://doi.org/10.1145/2647868.2655045>.

SALEH, Babak; ELGAMMAL, Ahmed. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, n. 2, Oct. 2016. Disponível em: <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/23376>.

SALIMANS, Tim; GOODFELLOW, Ian J.; ZAREMBA, Wojciech; CHEUNG, Vicki; RADFORD, Alec; CHEN, Xi. Improved techniques for training gans. In: LEE, Daniel D.; SUGIYAMA, Masashi; LUXBURG, Ulrike von; GUYON, Isabelle; GARNETT, Roman (Ed.). *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. [s.n.], 2016. p. 2226–2234. Disponível em: <https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html>.

SCHROFF, Florian; KALENICHENKO, Dmitry; PHILBIN, James. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015.

SEHSAH, Aalaa I.; MOUSA, Afaf; FAROUK, Gamal. A hybrid variational autoencoder and wgan with gradient penalty for tertiary protein structure generation. *Scientific Reports*, v. 15, n. 1, p. 14191, apr 2025. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-025-94747-y>.

SHI, Haizhou; XU, Zihao; WANG, Hengyi; QIN, Weiyi; WANG, Wenyuan; WANG, Yibin; WANG, Zifeng; EBRAHIMI, Sayna; WANG, Hao. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

SHIN, Hanul; LEE, Jung Kwon; KIM, Jaehong; KIM, Jiwon. Continual learning with deep generative replay. In: GUYON, I.; LUXBURG, U. Von; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *NIPS*. [S.l.]: Curran Associates, Inc., 2017. v. 30.

SIMONYAN, K; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: . [S.l.]: Computational and Biological Learning Society, 2015. p. 1–14.

STOYCHEV, Samuil; CHURAMANI, Nikhil; GUNES, Hatice. Latent generative replay for resource-efficient continual learning of facial expressions. In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. [S.l.: s.n.], 2023. p. 1–8.

SUN, Caihao; ZHANG, Xiaohua; MENG, Hongyun; CAO, Xianghai; ZHANG, Jinhua. Ac-wgan-gp: Generating labeled samples for improving hyperspectral image classification with small-samples. *Remote Sensing*, v. 14, n. 19, 2022. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/14/19/4910>.

SUN, Zheng; TORRIE, Shad A.; SUMSION, Andrew W.; LEE, Dah-Jye. Self-supervised facial motion representation learning via contrastive subclips. *Electronics*, v. 12, n. 6, 2023. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/12/6/1369>.

SZEGEDY, Christian; LIU, Wei; JIA, Yangqing; SERMANET, Pierre; REED, Scott; ANGUELOV, Dragomir; ERHAN, Dumitru; VANHOUCKE, Vincent; RABINOVICH, Andrew. Going deeper with convolutions . In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2015. p. 1–9. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594>.

SZEGEDY, Christian; VANHOUCKE, Vincent; IOFFE, Sergey; SHLENS, Jon; WOJNA, Zbigniew. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 2818–2826.

TAN, Mingxing; LE, Quoc. EfficientNet: Rethinking model scaling for convolutional neural networks. In: CHAUDHURI, Kamalika; SALAKHUTDINOV, Ruslan (Ed.). *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 6105–6114. Disponível em: <https://proceedings.mlr.press/v97/tan19a.html>.

TANNUGI, Dylan C.; BRITTO, Alceu S.; KOERICH, Alessandro L. Memory integrity of cnns for cross-dataset facial expression recognition. In: *IEEE Intl Conf on Systems, Man and Cybernetics*. [S.l.: s.n.], 2019. p. 3826–3831.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Ł ukasz; POLOSUKHIN, Illia. Attention is all you need. In: GUYON, I.; LUXBURG, U. Von; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

VEN, Gido M. van de; SOURES, Nicholas; KUDITHIPUDI, Dhireesha. *Continual Learning and Catastrophic Forgetting*. 2024. Disponível em: <https://arxiv.org/abs/2403.05175>.

VEN, Gido M. van de; TUYTELAARS, Tinne; TOLIAS, Andreas S. Three types of incremental learning. *Nature Machine Intelligence*, v. 4, n. 12, p. 1185–1197, 2022.

WAH, Catherine; BRANSON, Steve; WELINDER, Peter; PERONA, Pietro; BELONGIE, Serge. *The Caltech-UCSD Birds-200-2011 Dataset*. [S.l.: s.n.], 2011.

WANG, Liyuan; XIE, Jingyi; ZHANG, Xingxing; HUANG, Mingyi; SU, Hang; ZHU, Jun. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 2023.

WANG, Liyuan; ZHANG, Xingxing; SU, Hang; ZHU, Jun. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 8, p. 5362–5383, 2024.

WANG, Mingyang; ADEL, Heike; LANGE, Lukas; STRÖTGEN, Jannik; SCHUETZE, Hinrich. Learn it or leave it: Module composition and pruning for continual learning. In: ZHAO, Chen; MOSBACH, Marius; ATANASOVA, Pepa; GOLDFARB-TARRENT, Seraphina; HASE, Peter; HOSSEINI, Arian; ELBAYAD, Maha; PEZZELLE, Sandro; MOZES, Maximilian (Ed.). *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 163–176. Disponível em: <https://aclanthology.org/2024.repl4nlp-1.12>.

WANG, Xukang; WU, Ying Cheng; ZHOU, Mengjie; FU, Hongpeng. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in Big Data*, Frontiers Media SA, Switzerland, v. 7, p. 1337465, 2024. ISSN 2624-909X. ECollection 2024. Copyright © 2024 Wang, Wu, Zhou and Fu.

WANG, Zhou; BOVIK, A.C.; SHEIKH, H.R.; SIMONCELLI, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600–612, 2004.

WANG, Zhen; LIU, Liu; DUAN, Yiqun; KONG, Yajing; TAO, Dacheng. Continual learning with lifelong vision transformer. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 171–181.

WANG, Zhen; LIU, Liu; KONG, Yajing; GUO, Jiaxian; TAO, Dacheng. Online continual learning with contrastive vision transformer. In: AVIDAN, Shai; BROSTOW, Gabriel; CISSÉ, Moustapha; FARINELLA, Giovanni Maria; HASSNER, Tal (Ed.). *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022. p. 631–650. ISBN 978-3-031-20044-1.

WANG, Z.; SIMONCELLI, E.P.; BOVIK, A.C. Multiscale structural similarity for image quality assessment. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*. [S.l.: s.n.], 2003. v. 2, p. 1398–1402 Vol.2.

WANG, Zifeng; ZHANG, Zizhao; EBRAHIMI, Sayna; SUN, Ruoxi; ZHANG, Han; LEE, Chen-Yu; REN, Xiaoqi; SU, Guolong; PEROT, Vincent; DY, Jennifer et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. *European Conference on Computer Vision*, 2022.

WANG, Zifeng; ZHANG, Zizhao; LEE, Chen-Yu; ZHANG, Han; SUN, Ruoxi; REN, Xiaoqi; SU, Guolong; PEROT, Vincent; DY, Jennifer; PFISTER, Tomas. Learning to prompt for continual learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 139–149.

WINKLER, Stefan; MOHANDAS, Praveen. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE Transactions on Broadcasting*, v. 54, n. 3, p. 660–668, 2008.

WU, Jialing; LI, Wanyi; WU, Yilin; QIU, Shuang. Wasserstein generative adversarial network with gradient penalty for handwritten digit generation. In: *2024 International Conference on Intelligent Robotics and Automatic Control (IRAC)*. [S.l.: s.n.], 2024. p. 375–379.

WU, Yue; CHEN, Yinpeng; WANG, Lijuan; YE, Yuancheng; LIU, Zicheng; GUO, Yandong; FU, Yun. *Large Scale Incremental Learning*. 2019. Disponível em: <https://arxiv.org/abs/1905.13260>.

XIAO, Han; RASUL, Kashif; VOLLGRAF, Roland. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. Cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/. Disponível em: <http://arxiv.org/abs/1708.07747>.

XIE, Siyue; HU, Haifeng; CHEN, Yizhen. Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 31, n. 6, p. 2359–2371, 2021.

XU, Ju; ZHU, Zhanxing. Reinforced continual learning. In: BENGIO, S.; WALLACH, H.; LAROCHELLE, H.; GRAUMAN, K.; CESA-BIANCHI, N.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. v. 31. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2018/file/cee631121c2ec9232f3a2f028ad5c89b-Paper.pdf>.

YAN, Shipeng; XIE, Jiangwei; HE, Xuming. DER: Dynamically Expandable Representation for Class Incremental Learning . In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2021. p. 3013–3022. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00303>.

YOON, Jaehong; YANG, Eunho; LEE, Jeongtae; HWANG, Sung Ju. *Lifelong Learning with Dynamically Expandable Networks*. 2018. Disponível em: <https://arxiv.org/abs/1708.01547>.

YUDA, Emi; ANDO, Tomoki; KANEKO, Itaru; YOSHIDA, Yutaka; HIRAHARA, Daisuke. Comprehensive data augmentation approach using wgan-gp and umap for enhancing alzheimer's disease diagnosis. *Electronics*, v. 13, n. 18, 2024. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/13/18/3671>.

ZENKE, Friedemann; GERSTNER, Wulfram; GANGULI, Surya. The temporal paradox of hebbian learning and homeostatic plasticity. *Current Opinion in Neurobiology*, v. 43, p. 166–176, 2017. ISSN 0959-4388. Neurobiology of Learning and Plasticity. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959438817300910>.

ZENKE, Friedemann; POOLE, Ben; GANGULI, Surya. Continual learning through synaptic intelligence. *Proc Machine Learning Research*, v. 70, p. 3987–3995, 2017. ISSN 2640-3498 (Electronic).

ZHANG, Gengwei; WANG, Liyuan; KANG, Guoliang; CHEN, Ling; WEI, Yunchao. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2023.

ZHANG, Xiaoqin; LI, Min; LIN, Sheng; XU, Hang; XIAO, Guobao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 34, n. 5, p. 3192–3203, 2024.

ZHANG, X.; YIN, L.; COHN, J. F.; CANAVAN, S.; REALE, M.; HOROWITZ, A.; LIU, P. Bp4d. In: . TIB, 2024. Disponível em: <https://service.tib.eu/ldmservice/dataset/bp4d>.

ZHANG, Zhanpeng; LUO, Ping; LOY, Chen Change; TANG, Xiaoou. From facial expression recognition to interpersonal relation prediction. *CoRR*, abs/1609.06426, 2016. Disponível em: <http://arxiv.org/abs/1609.06426>.

ZHOU, Da-Wei; SUN, Hai-Long; YE, Han-Jia; ZHAN, De-Chuan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In: *CVPR*. [S.l.: s.n.], 2024. p. 23554–23564.

ZHOU, Da-Wei; YE, Han-Jia; ZHAN, De-Chuan; LIU, Ziwei. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *arXiv preprint arXiv:2303.07338*, 2023.

ZHU, Jun-Yan; PARK, Taesung; ISOLA, Phillip; EFROS, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 2242–2251.

Apêndices

APÊNDICE A – Resultados complementares: influência da ordem das tarefas

Este apêndice apresenta os valores numéricos completos dos experimentos conduzidos com a ordem de tarefas iniciando pela base de dados JAFFE e incluindo, uma a uma, as bases de dados TFEID, MUG e CK+. Os resultados correspondem às curvas de acurácia apresentadas na Seção 5.3.2.3, incluindo média e desvio padrão para cada método avaliado ao longo das etapas de aprendizado incremental. As Tabelas 38, 39 e 40 demonstram os resultados de cada etapa de adaptação do modelo, considerando a inclusão progressiva de novas bases de dados.

Tabela 38 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada na base de dados JAFFE e adaptada para a base TFEID, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual, para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da TFEID com a versão sintética da JAFFE (JAFFE^A). No *fine-tuning*, a base-alvo é composta apenas pela TFEID; no *joint*, é formada pela combinação das bases JAFFE original e TFEID.

Método	**Base fonte**	**Base alvo**	**Média**
	JAFFE	**TFEID**	
Baseline	0,9545±0,00	0,1852±0,00	0,5699±0,00
Joint	1,0000±0,00	0,8370±0,05	0,9185±0,05
Fine-tuning	0,3636±0,18	0,8111±0,04	0,5874±0,11
ECgr	0,7727±0,05	0,8111±0,06	0,7919±0,06
ECgr+QA	0,7955±0,09	0,7889±0,07	0,7922±0,08
ECgr+wQA	0,7682±0,08	0,8333±0,04	**0,8008±0,06**
ECgr+*cluster*	0,7455±0.05	0,8259±0,04	0,7857 ± 0,04
ECgr+w*cluster*	0,6864±0.09	0,8296±0,04	0,7580 ± 0,06
ECgr+CGLO	0,7545±0,08	0,8185±0,04	0,7865±0,06
ECgr+wCGLO	0,7636±0,07	0,8000±0,07	0,7818±0,07

Fonte: autoria própria.

Tabela 39 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada nas bases de dados JAFFE e TFEID e adaptada para a base MUG, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da MUG com as versões sintéticas das bases JAFFE e TFEID (JAFFE$^+$ e TFEID$^+$). No *fine-tuning*, a base-alvo é composta apenas pela MUG; no *joint*, é formada pela combinação das bases JAFFE original, TFEID original e MUG.

Método	**Base fonte**		**Base alvo**	**Média**
	JAFFE	**TFEID**	**MUG**	
Baseline	0,2727±0,00	0,7407±0,00	0,1485±0,00	0,3873±0,00
Joint	0,9955±0,01	0,9963±0,01	0,9812±0,00	0,9910±0,01
Fine-tuning	0,4227±0,03	0,1852±0,02	0,9800±0,00	0,5293±0,02
ECgr	0,8455±0,04	0,3926±0,08	0,9837±0,00	0,7406±0,04
ECgr+QA	0,8318±0,04	0,5556±0,06	0,9823±0,00	0,7899±0,03
ECgr+wQA	0,8545±0,03	0,4741±0,05	0,9809±0,00	0,7698±0,03
ECgr+*cluster*	0,8091±0,05	0,5222±0,06	0,9782±0,01	0,7698±0,06
ECgr+w*cluster*	0,8227±0,05	0,4704±0,04	0,9823±0,00	0,7585±0,04
ECgr+CGLO	0,8545±0,04	0,5333±0,05	0,9827±0,01	**0,7902±0,03**
ECgr+wCGLO	0,8500±0,04	0,5259±0,04	0,9839±0,00	0,7866±0,03

Fonte: autoria própria.

Tabela 40 – Acurácia média e desvio padrão no conjunto de testes para a CNN treinada nas bases de dados JAFFE, TFEID e MUG e adaptada para a base CK+, considerando os métodos ECgr, ECgr+QA, ECgr+*cluster*, ECgr+CGLO e suas respectivas versões ponderadas, juntamente com *fine-tuning*, *joint* e o modelo atual para uma comparação direta. Nos métodos ECgr, a base-alvo corresponde à combinação da CK+ com as versões sintéticas das bases JAFFE, TFEID e MUG (JAFFE$^+$, TFEID$^+$ e MUG$^+$). No *fine-tuning*, a base-alvo é composta apenas pela CK+; no *joint*, é formada pela combinação das bases JAFFE original, TFEID original, MUG original e CK+.

Método	**Base fonte**			**Base alvo**	**Média**
	JAFFE	**TFEID**	**MUG**	**CK+**	
Baseline	0,4091±0,00	0,2222±0,00	0,9782±0,00	0,3421±0,00	0,4879±0,00
Joint	0,9955±0,01	1,0000±0,00	0,9959±0,00	0,5026±0,05	0,8735±0,02
Fine-tuning	0,4091±0,05	0,3037±0,02	0,8128±0,07	0,6737±0,06	0,5498±0,05
ECgr	0,8909±0,03	0,4481±0,05	0,8351±0,03	0,6842±0,04	0,7146±0,04
ECgr+QA	0,8500±0,04	0,4630±0,06	0,8651±0,02	0,5842±0,02	0,6906±0,04
ECgr+wQA	0,8864±0,02	0,5593±0,05	0,9110±0,04	0,5289±0,06	0,7214±0,04
ECgr+*cluster*	0,8727±0,04	0,5556±0,06	0,8102±0,03	0,5474±0,04	0,6965±0,10
ECgr+w*cluster*	0,7955±0,03	0,5704±0,05	0,8606±0,05	0,5289±0,07	0,6888±0,08
ECgr+CGLO	0,8500±0,03	0,5222±0,07	0,8474±0,05	0,4868±0,05	0,6766±0,05
ECgr+wCGLO	0,8682±0,02	0,6259±0,07	0,9390±0,04	0,4579±0,06	**0,7227±0,05**

Fonte: autoria própria.