Alleviating Catastrophic Forgetting in Facial Expression Recognition with Emotion-Centered Models

Israel A. Laurensi[1], Alceu de Souza Britto Jr.[1], Jean Paul Barddal[1], and Alessandro Lameiras Koerich[2]

[1] Graduate Program in Informatics (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR, Brazil
[2] École de Technologie Supérieure (ÉTS), Montréal, Canada

Abstract. Facial expression recognition is pivotal in machine learning, facilitating various applications. However, convolutional neural networks (CNNs) are often plagued by catastrophic forgetting, impeding their adaptability. The proposed method, emotion-centered generative replay (ECgr), tackles this challenge by integrating synthetic images from generative adversarial networks. Moreover, ECgr incorporates a quality assurance algorithm to ensure the fidelity of generated images. This dual approach enables CNNs to retain past knowledge while learning new tasks, enhancing their performance in emotion recognition. The experimental results on four diverse facial expression datasets demonstrate that incorporating images generated by our pseudo-rehearsal method enhances training on the targeted dataset and the source dataset while making the CNN retain previously learned knowledge.

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

1 Introduction

Emotions are essential in human interaction and comprehension. In such a context, facial expressions play an important role [15]. Thus, facial expression recognition (FER) is the functionality of numerous machine learning applications, including emotion-aware interfaces, personalized recommender systems, and human-robot interaction. One way to identify these emotions in complex systems is via convolutional neural networks (CNNs). These networks have achieved remarkable success in computer vision tasks such as image classification, object detection, and facial expression recognition. However, a significant limitation of CNNs is their susceptibility to catastrophic forgetting. When sequentially trained on different tasks or datasets, CNNs often struggle to retain previously learned information, which leads to degraded performance on previously mastered tasks. This phenomenon impairs the practical application of CNNs in dynamic environments where models must continuously adapt to new data while retaining accuracy in the previous scenarios.

Evaluating the catastrophic forgetting problem in FER - a complex learning scenario - allows us to observe the proposed method's ability to deal with datasets composed of diverse emotional expressions, unlike more straightforward tasks with more limited patterns. Moreover, such an evaluation sheds light on the model's adeptness at maintaining previously learned emotional recognition performance while assimilating the changes of a new domain, showing faces collected with other acquisition protocols, and representing people with different characteristics and cultures.

Catastrophic forgetting arises due to CNN's optimization process, which tends to adjust the model's parameters to fit the current task, often overshadowing previously acquired representations. Researchers have proposed numerous approaches to mitigate catastrophic forgetting, including regularization techniques, dynamic neural network architecture, and rehearsal-based methods [7, 8, 12, 13, 16]. Furthermore, several literature reviews have been published in this research field and in continual learning, offering comprehensive insights into the state-of-the-art methodologies, best practices, and emerging trends in mitigating catastrophic forgetting and advancing continual learning algorithms [6, 11, 14]. While these state-of-the-art methodologies have demonstrated promising results in specific scenarios, they have limitations such as increased computational complexity or limited capacity to effectively retain information from past tasks, especially in facial expression recognition scenarios.

In this paper, we propose a novel approach to overcome the limitations of existing methods and effectively address catastrophic forgetting in CNNs when applied to facial expression recognition. Our approach capitalizes on generative adversarial networks (GANs) capabilities to generate synthetic samples that resemble the original training data. Incorporating these synthetic samples during training enables the CNN to re-learn and retain knowledge from previous tasks, thereby mitigating catastrophic forgetting. To achieve this, we generated synthetic images of each emotion (class) present in the datasets, aiming to better capture the intrinsic characteristics of each facial expression associated with human emotion. We refer to this method as emotion-centered generative replay (ECgr). Moreover, we introduce a quality assurance (QA) algorithm as a crucial component of our approach. The QA algorithm assesses the generated synthetic samples based on the CNN's original classification accuracy. Only high-quality synthetic samples, which the original CNN can accurately classify, are retained for training. This filtering step ensures that only superior generated samples are utilized, thus augmenting the performance of the proposed method. In addition, we weigh the importance of the synthetic images, considering the CNN output score as an image quality assignment. Such a weight penalizes images that have been assigned a low confidence value by the CNN, which might positively influence the training convergence, as these images may be considered detrimental to the adaptation to the new dataset.

Our hypothesis centers around the effectiveness of employing a pseudo-rehearsal method: H1) the utilization of a pseudo-rehearsal method, particularly our emotion-centered generative replay, offers a potential solution for memory decay in CNNs; H2) the fusion of our emotion-centered generative replay and the proposed QA algorithm offers a promising strategy to counteract memory decay within neural networks; and H3) the combination of emotion-centered generative replay, QA, and a weighted loss function is hypothesized to further strengthen memory retention and performance in neural networks, potentially surpassing the benefits of either ECgr or QA alone. To assess the proposed method's efficiency and validate our hypothesis, we undertook facial expression recognition experiments across various emotion datasets and employed diverse training methodologies.

The contribution of this work is three-fold: i) a new pseudo-rehearsal method focused on the emotions to mitigate catastrophic forgetting when learning facial emotion recognition; ii) a loss function considering a penalization schema for low-quality synthetic images generated in the pseudo-rehearsal strategy; iii) a robust experimental protocol considering well-known FER datasets and a pipeline of experiments to discuss the contributions of the proposed emotion-centered generative replay in mitigating catastrophic forgetting when compared to a regular fine-tuning process or the possibility of joining datasets.

The remainder of this paper is structured as follows: Section 2 reviews related work on catastrophic forgetting and existing methods for its mitigation. Section 3 presents the proposed emotion-centered generative replay approach and outlines the architecture of the QA algorithm. Section 4 describes the experimental setup and presents the results of our comprehensive evaluations. Section 5 discusses the implications of our findings, and Section 6 concludes the paper, outlining potential directions for future research.

2 Related Works

Catastrophic forgetting has spurred numerous research works to minimize its effects. In this section, we explore prominent algorithms and insights inspired by neuropsychology, all aimed at addressing forgetting and improving memory retention within neural networks.

Learning without forgetting (LWF) stands out by employing knowledge distillation [8]. This technique transfers distilled knowledge from a model trained on prior tasks to a new model, thereby allowing the assimilation of new information while safeguarding the retention of past information. This intelligent utilization of previous knowledge effectively counteracts the plague of forgetting and amplifies the network's overall performance. Another regularization method, elastic weight consolidation (EWC) [7], introduces a nuanced regularization term in the scenario. This term identifies and assigns significance to pivotal network parameters linked to previous tasks, penalizing alterations to these parameters during subsequent training phases. By preserving these key parameters diligently, EWC balances between accommodating novel tasks and upholding the wisdom derived from past experiences.

Synaptic intelligence (SI) [16] offers an innovative perspective that stems from evaluating past task performance and assigns weight to synaptic connections based on their influence. The more a synapse contributes, the higher its importance; in contrast, less influential synapses are assigned lower importance. By preserving these critical connections, SI bridges the gap between old and new information, thus mitigating forgetting while embracing novelty.

Deep generative replay (DGR) [12] utilizes generative models to create simulated instances from prior tasks during training. This approach effectively enriches the current task's dataset. The augmented instances, fused with real-time data, offer the network a diverse and comprehensive pool of examples. With past knowledge seamlessly integrated, DGR effectively combats the erosion of previously gained insights, presenting itself as a powerful tool for memory retention.

Beyond these algorithms, insights obtained from neuropsychological research paint a broader picture. Investigations into context-dependent learning have illuminated the crucial role of training and testing contexts in determining network performance. Thus, using contextual cues, algorithms can be designed to exploit the training and testing context better, thereby enhancing memory retention while countering the forgetting phenomenon [11].

In light of these contributions, it is crucial to contextualize our work within the broader realm of the current state-of-the-art. The proposed methodology harmonizes the concepts of emotion-centered generative replay and QA. With CNNs as the focal point, our approach aims to prevent catastrophic forgetting in facial expression recognition, a domain where precise emotion identification heavily depends on image quality.

3 Proposed Method

In this section, we describe the methodology employed in our study to address the challenges of catastrophic forgetting in facial emotion recognition tasks. Our approach combines emotion-centered generative replay using a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) and a QA algorithm. Fig. 1 presents a general overview of the proposed method.

The use of WGAN-GPs is attributed to the stable learning power of these networks, a factor crucial when dealing with catastrophic forgetting. After all, attempting to address this issue through training and employing a generative method may lead to catastrophic forgetting in the generative networks. WGAN-GPs [5] implement a penalty on the gradient norm during training and optimization of the WGAN [2], thereby ensuring more stable training and yielding higher-quality generated images.

We have formalized our methodology using algorithmic representations to provide a more concrete understanding of the theoretical concepts presented. In Subsection 3.2, we provide detailed algorithms replicating our approach's offline preparation and training stages. These algorithms encapsulate the step-by-step processes of generating synthetic images and performing continuous retraining.

3.1 Emotion-Centered Generative Replay

We initiate by training a set of WGAN-GPs, one for each of the seven emotion classes present in the 'source' dataset - fear, anger, happiness, sadness, disgust,

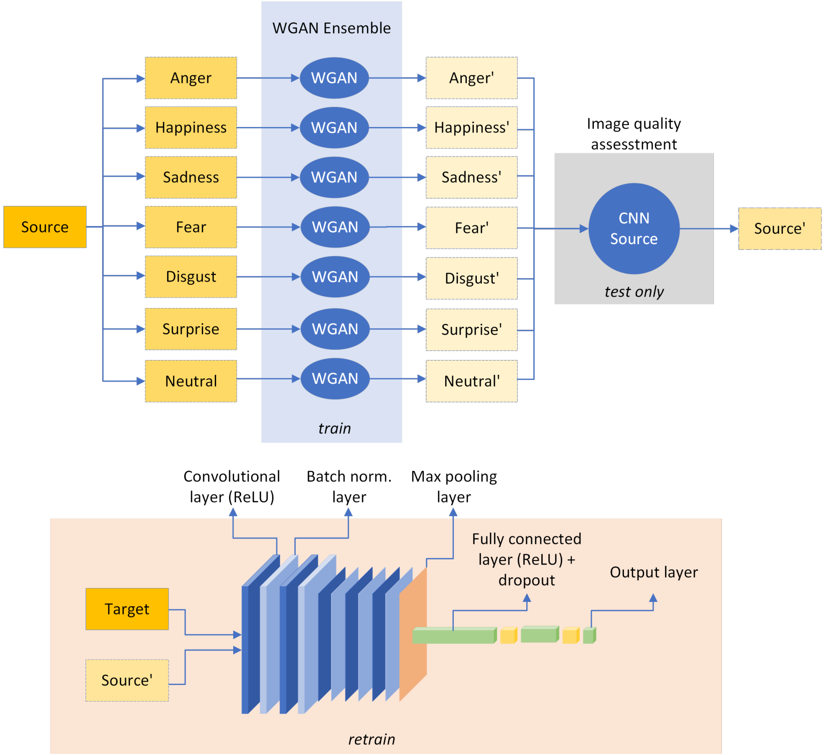

Fig. 1: An overview of the proposed method, separated into two key components. At the top, the emotion-centered WGAN-GP with CNN QA is depicted. This component involves training a WGAN-GP for each class in the source dataset to generate synthetic data resembling that class. At the bottom, the fine-tuning strategy is illustrated, where our synthetic dataset is replayed alongside the target dataset.

surprise, and neutral. Using these trained WGAN-GPs, we generate augmented datasets for each class. These generated images capture the intricate details of respective emotions, diversifying training data towards better generalization.

The WGAN-GP is built by two different networks: discriminator and generator. The discriminator network is crucial for distinguishing between real and synthetic images. It consists of several layers, including convolutional layers with leaky rectified linear united (ReLU) activation functions. These layers help the discriminator extract relevant features from input images. Additionally, dropout layers are applied to prevent overfitting. This network contains approximately 4.3 million trainable parameters.

The generator network, detailed in Table 1, creates synthetic images from random noise. It uses dense, batch normalization, and convolutional layers with leaky ReLU activations to upscale and refine feature maps. The output matches the desired image size. With around 1.5 million parameters, this architecture produces images to challenge the discriminator.

Table 1: WGAN-GP generator and discriminator architecture.

Generator	**Output Shape**	**Discriminator**	**Output Shape**
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)		
Total params: 1,586,500		**Total params: 4,303,360**	

We employ our QA algorithm to ensure the quality of the generated images. The QA algorithm filters out low-quality or incorrect images generated by the WGAN-GP, retaining high-quality images that the original classifier correctly classifies. The QA process is performed using the CNN trained on the source dataset. Given an empirically defined threshold, the images correctly classified by the network are used for future retraining, and the misclassified images are discarded. The QA process enhances the reliability of the emotion-centered generative replay, preventing the classifier from being influenced by poor-quality or misleading synthetic images. These images are then integrated into an improved dataset, which merges the synthetic images with the initial source data.

During retraining, the new dataset and the target dataset are employed. This unified dataset facilitates CNN training, where knowledge from the original emotion classes is combined with the new target emotions, minimizing forgetting.

3.2 General Pipeline

To address the challenge of catastrophic forgetting, our proposed approach involves a two-stage process: offline preparation and a training phase.

Offline preparation stage. Initialization occurs as depicted in Algorithm 1, where a set of datasets represented by T is defined, encompassing datasets A, B, C, and D. Each dataset d_t within T is traversed through an iterative

Algorithm 1 Offline stage

```
$T \leftarrow A, B, C, D$
$T' \leftarrow \emptyset$
for each dataset $d_t$ in $T$ do
    $G_{d_t} \leftarrow \emptyset$
    Train classifier $C_{d_t}$ on dataset $d_t$
    for each class $c$ in dataset $d_t$ do
        Train $\texttt{WGANGP}_c$ on class $c$
        Add $\texttt{WGANGP}_c$ to ensemble $G_{d_t}$
        Generate $\texttt{SI}_c$ using $\texttt{WGANGP}_c$
        Pass $\texttt{SI}_c$ through $C_{d_t}$ to generate dataset ${d_t}_c^{qa}$
        Add ${d_t}_c^{qa}$ to dataset $d'_t$
    end for
    Add $d'_t$ to $T'$
end for
return collection of synthetic datasets $T'$
```

process. For each specific dataset d_t, a classifier, denoted as C_{d_t}, is trained using that particular dataset.

Our proposal then iterates over each class c in dataset d_t. In this context, a WGAN-GP is trained per class, denoted $\texttt{WGANGP}_c$, and these are subsequently combined to form an ensemble, denoted as G_{d_t}. Through these $\texttt{WGANGP}_c$, synthetic images ($\texttt{SI}_c$) are generated to reflect the characteristics of each class. Continuing the process, these synthetic images are input to the classifier C_{d_t}, thus resulting in a new dataset, ${d_t}_c^{qa}$, consisting of the images that are correctly classified by C_{d_t}. These refined synthetic images are combined into a new dataset d'_t. This procedure is executed for each dataset d_t, and all the resulting datasets d'_t are unified into a collection labeled T', encapsulating the sets of synthetic datasets corresponding to each original dataset d_t in T.

The time complexity of Algorithm 1 is $O(n \cdot (f(p) + m \cdot g(p)))$, where n is the number of datasets, m is the number of classes per dataset, and p is the number of images per class. The term $f(p)$ represents the time complexity for training a classifier on p images, while $g(p)$ denotes the complexity for training a WGAN-GP on p images.

Continual learning stage. Our approach began with individual training for ECgr before merging ECgr and QA. For a comparative evaluation, we utilize joint training and fine-tuning methods. Joint training simultaneously incorporates the source and target data while fine-tuning adapts the CNN to new data, training only the fully connected layers.

As shown in Algorithm 2, we define a set of subsequent datasets, indicated by T, which comprises datasets B, C, and D. Then, we iterate over each combination of the original dataset and subsequent dataset, referred to as d_t and d'_t respectively, from set T and its counterpart T'. For each dataset combination, we create a unified dataset, d_t^u, by merging d_t and d'_t. Subsequently, we train a classifier, $C_{d_t^u}$ on the unified dataset d_t^u.

Algorithm 2 Continual learning stage

```
$T \leftarrow B, C, D$
$C_T \leftarrow \emptyset$
**for** each dataset $d_t$, $d'_t$ in $T$, $T'$ **do**
    $d^u_t \leftarrow d_t + d'_t$
    Train classifier $C_{d^u_t}$ on unified dataset $d^u_t$
    Add trained $C_{d^u_t}$ to $C_T$
**end for**
**return** ensemble $C_T$
```

For Algorithm 2, the time complexity is $O(n \cdot (r(m) + f(m)))$. Here, n denotes the number of dataset pairs processed from sets T and T', while m represents the size of individual datasets d_t and d'_t. The term $r(m)$ stands for the overhead for merging datasets d_t and d'_t, whereas $f(m)$ represents the computational cost of training a classifier on a dataset of size m.

The CNN used in our experiments, detailed in Table 2, begins with 2D convolutional layers (64 filters each) and batch normalization. It includes additional convolutional layers, max-pooling for downsampling, and further batch normalization for higher-level feature extraction. The feature maps are flattened and passed through fully connected layers with dropout to prevent overfitting. The final layer uses softmax activation to output class probabilities. Overall, this CNN architecture comprises approximately 19.3 million parameters.

Table 2: CNN network architecture.

Layer (type)	**Output Shape**	**Params**
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

Total params: 19,272,455

The training aims to optimize the Eq. (1), where a weight w is applied to each prediction. This weight is determined by the CNN's confidence percentage when predicting for all y_{pred}.

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

In summary, our general pipeline encompasses an offline preparation phase involving training WGAN-GPs and QA-based synthetic image generation. In the training stage, synthetic and original datasets are combined, and the continual retraining approach adapts the classifier to multiple datasets while incorporating different strategies.

4 Experiments

To evaluate the performance of our methodology, we utilize several datasets that contain human facial images displaying various emotions. The datasets considered in our study include TFEID, MUG, CK+, and JAFFE. These datasets provide diverse emotional contexts, allowing us to assess our approach's robustness and generalization capabilities. All datasets have the following classes: fear, anger, happiness, sadness, disgust, surprise, and neutral.

The Multimodal Understanding Group (MUG) [1] dataset consists of approximately 1462 facial images, each annotated with the corresponding facial expression labels. The Japanese Female Facial Expression (JAFFE) [10] dataset, despite its relatively small size, containing approximately 213 facial images, is valuable for evaluating and comparing facial expression recognition models. The Taiwanese Facial Expression Image Database (TFEID) [3] provides a suitable testbed for evaluating emotion recognition algorithms, with 1128 samples. Lastly, the extended Cohn-Kanade dataset (CK+) [9] is commonly used for facial expression recognition research. It includes a substantial number of facial images, compiled into 123 videos of different subjects, totaling approximately 593 videos, with 327 labeled videos covering various emotional expressions.

4.1 Results

This section offers an in-depth analysis of the outcomes achieved by employing different retraining strategies, each suited to minimize memory degradation and maximize knowledge retention.

On Quality of Synthetic Images. In this subsection, we present a comprehensive discussion of the qualitative aspects of the synthetic data. As shown in Fig. 2, the left side features an image from the original dataset as a reference for the dataset's inherent visual characteristics. On the right side, seven columns display synthetic images generated for each class within the dataset. These columns show the diversity and fidelity of the synthetic samples produced by our ECgr approach. Fig. 3 shows examples of images that were rejected during the QA

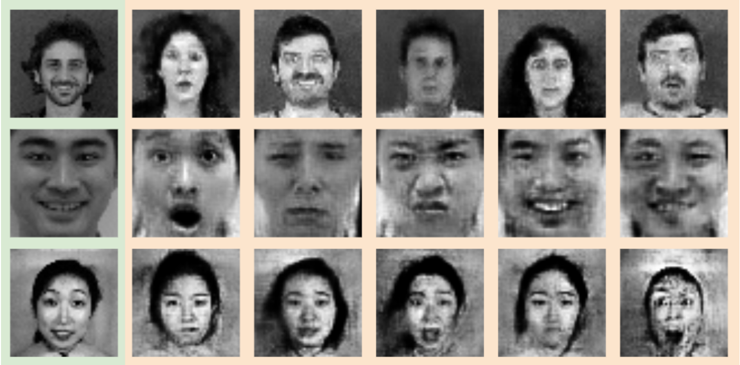

Fig. 2: Sample results for different classes from the MUG, JAFFE, and TFEID synthetic datasets generated by WGAN-GP. The first column (in green) displays the original samples from the MUG, TFEID, and JAFFE datasets (from top to bottom, respectively). In contrast, the second-to-last column (in orange) features the corresponding synthetic images for each dataset.

process. These rejected images are of low quality and do not convey emotion, resulting in incorrect classification by the CNN.

On Continual Learning. In this section, we discuss the main results observed from the tests conducted with facial expression datasets, utilizing the combination of different methods outlined in this study.

Initially, we trained a CNN on the MUG dataset. We then adapted this CNN for continuous learning across other datasets. Each training process was replicated 20 times. For methods involving image generation, the synthetic datasets differ across various replications of CNN adaptation.

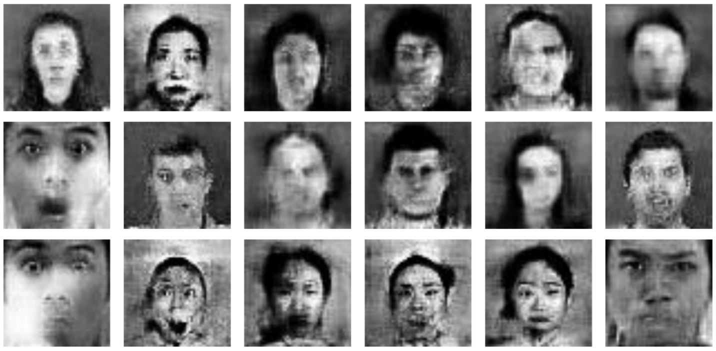

Fig. 3: Some rejected samples identified by the QA algorithm from the synthetic datasets of MUG, JAFFE, and TFEID.

In Tables 3, 4 and 5, the columns baseline, joint, and fine-tune represent, respectively: testing datasets with the CNN trained on the source dataset; adapting the CNN trained on the source plus target dataset; adapting the CNN trained on the source dataset using only the target dataset. Additionally, the ECgr and QA methods were evaluated separately (ECgr) and then combined (ECgr+QA) to determine the impact of using synthetic image filtering in continuous training. Furthermore, this scenario assessed whether using weights (ECgr+wQA) on synthetic images has any effect compared to training without this technique.

Table 3: Results on MUG's model fine-tuned to JAFFE dataset in terms of ECgr, QA, weighted QA, and the combination of ECgr with QA and wQA, alongside with fine-tune, joint, and current for a direct comparison.

	Current model	**Joining datasets**	**Fine Tuning**	**Proposed**		
				ECgr	**ECgr+QA**	**ECgr+wQA**
Source dataset						
MUG	0.98±0.00	1.00±0.00	0.75±0.03	0.88±0.04	0.93±0.02	0.94±0.03
Target dataset						
JAFFE	0.28±0.00	0.74±0.06	0.77±0.03	0.78±0.03	0.78±0.05	0.79±0.04
Mean	0.63	0.87	0.76	0.83	0.85	0.86

Table 3 shows the results when adapting the CNN trained on the MUG dataset (source) to the JAFFE dataset (target). Considering the baseline, joint, and fine-tune methods, we can assume that the upper limit is the joint method, which represents the ideal case where all datasets are available for training, and the lower limit is the fine-tune method, in which the source dataset is no longer available. The combined method ECgr+wQA yielded the best results in this initial adaptation involving only one dataset, with a result very close to joint training. Tables 4 and 5, show a change in this scenario, as more datasets are introduced in continuous training, the ECgr+QA method tends to outperform. Regarding the result obtained in the retraining for the JAFFE dataset, it is possible to justify this outcome, where the combined method (ECgr+QA) came close to the joint method, as the adaptation can still be considered trivial since only one dataset is being adapted. Thus, the complexity for the CNN to assimilate synthetic images needs to be higher.

Table 4 shows, when adapting the CNN trained on MUG and JAFFE to the new dataset TFEID, that the best result lies between ECgr+QA and ECgr+wQA. Interestingly, in all results, the generative method - combined with QA or not - performed equally or better on the target dataset when compared to joint training. This reveals that synthetic images not only aid the CNN in recalling something it has already seen but also assist in training for new data, reinforcing knowledge when adapting to the same context, in this case, emotion recognition.

In Table 5, when adapting the CNN trained on MUG, JAFFE, and TFEID to CK+, we can observe a behavior similar to that observed when adapting to TFEID, where the best result lies between the ECgr+QA and ECgr+wQA methods. However, at this point, it becomes more apparent that using a weight for synthetic images brings an intrinsic problem to the training of the CNN being used for the filtering method. This CNN can carry certain behaviors into subsequent training steps, where errors from certain classes may compromise the entire training when using the confidence percentage.

We can better understand the results in the MUG dataset from the continuous training of all datasets with Fig. 4. It is noticeable that the best method for the MUG dataset is ECgr+QA. We can also observe the poor performance of the fine-tuning method in the context of continuous training, where the knowledge was significantly forgotten compared to methods that attempt to mitigate this behavior. While fine-tuning initially shows promise in adapting the model to new tasks or domains, its performance deteriorates over time as knowledge retention becomes increasingly challenging. Additionally, memory forgetting becomes trivial when all datasets are always available, as datasets can be combined for retraining. However, one must consider the high computational cost and storage requirements of joint training.

Given that synthetic images for each class are generated independently in our method, it is essential to examine class-specific memory loss. Fig. 5 compares the fine-tune and ECgr+QA methods, revealing subpar performance (F1 $<$ 0.6) for the anger and disgust classes. During the final retraining step, the ECgr+QA method also experiences performance deterioration for the fear class. This un-

Table 4: Results on MUG plus JAFFE's model fine-tuned to TFEID dataset in terms of ECgr, QA, weighted QA and the combination of ECgr with QA and wQA, alongside with fine-tune, joint and current for a direct comparison.

	Current model	**Joining datasets**	**Fine Tuning**	**Proposed**		
				ECgr	**ECgr+QA**	**ECgr+wQA**
Source datasets						
MUG	0.75±0.03	1.00±0.00	0.71±0.01	0.84±0.06	0.87±0.04	0.78±0.04
JAFFE	0.77±0.03	0.94±0.03	0.76±0.02	0.64±0.07	0.62±0.07	0.69±0.04
Mean	0.75	0.97	0.73	0.74	0.74	0.73
Target dataset						
TFEID	0.22±0.00	0.79±0.05	0.78±0.03	0.83±0.04	0.84±0.04	0.87±0.04
Updated mean	0.58	0.91	0.75	0.77	0.78	0.78

Table 5: Results on MUG plus JAFFE plus TFEID's model fine-tuned to CK+ dataset in terms of ECgr, QA, weighted QA and the combination of ECgr with QA and wQA, alongside with fine-tune, joint and current for a direct comparison.

	Current model	**Joining datasets**	**Fine Tuning**	**Proposed**		
				ECgr	**ECgr+QA**	**ECgr+wQA**
Source datasets						
MUG	0.71±0.01	1.00±0.00	0.63±0.05	0.85±0.04	0.87±0.03	0.73±0.04
JAFFE	0.76±0.02	0.99±0.01	0.57±0.08	0.61±0.05	0.55±0.04	0.59±0.05
TFEID	0.78±0.03	1.00±0.00	0.49±0.06	0.62±0.09	0.76±0.09	0.70±0.07
Mean	0.73	0.95	0.56	0.69	0.72	0.67
Target dataset						
CK+	0.53±0.00	0.81±0.03	0.79±0.03	0.83±0.03	0.82±0.03	0.81±0.02
Updated mean	0.68	0.99	0.62	0.72	0.75	0.71

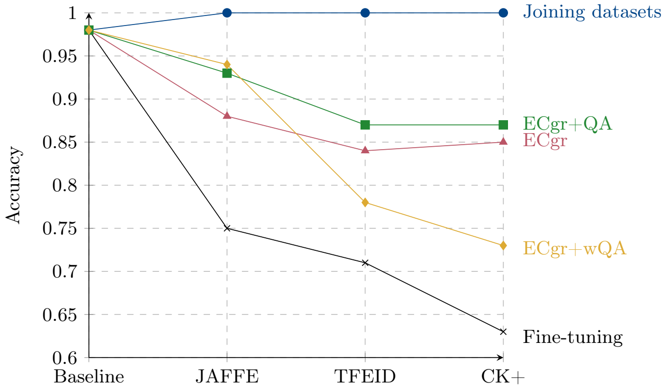

Fig. 4: Accuracy results on the MUG dataset, showcasing the continuous adaptation of a trained CNN across JAFFE, TFEID and CK+ datasets relative to the baseline accuracy.

derscores the difficulty of training these classes, as even minor facial changes can be misinterpreted as another emotional state.

We have also conducted experiments on a different domain using the MNIST dataset, and the results are presented in Appendix A.

5 Discussion

Firstly, the results support our hypothesis regarding using pseudo-rehearsal methods, specifically emotion-centered generative replay, to minimize memory decay. Our strategy demonstrated remarkable efficacy in alleviating catastrophic forgetting, consistently outperforming the fine-tuning methods across various tasks. The generation of synthetic data resembling past task patterns through WGAN-GPs proved positive in enabling the network to retain knowledge without using original data. This substantiates our anticipation that pseudo-rehearsal techniques, particularly our emotion-centered generative replay, are essential in counteracting memory decay.

Furthermore, synthesizing our WGAN-GP class-driven generative and QA methods substantiates our second hypothesis. Introducing a QA mechanism during replay significantly improved the quality of synthetic data, further augmenting the approach's effectiveness. The third hypothesis, in which applying a weight to synthetic images would benefit continuous training, can only be observed as positive in the retraining for the first dataset - from MUG to JAFFE. We observed that this technique was ineffective for more datasets after JAFFE. This may be directly related to the errors of the network that assigns these weights to the synthetic images—meaning the network may be making errors with high confidence, negatively affecting the synthetic images, which in turn are not fully considered in the retraining, leading the CNN to not remember these data.

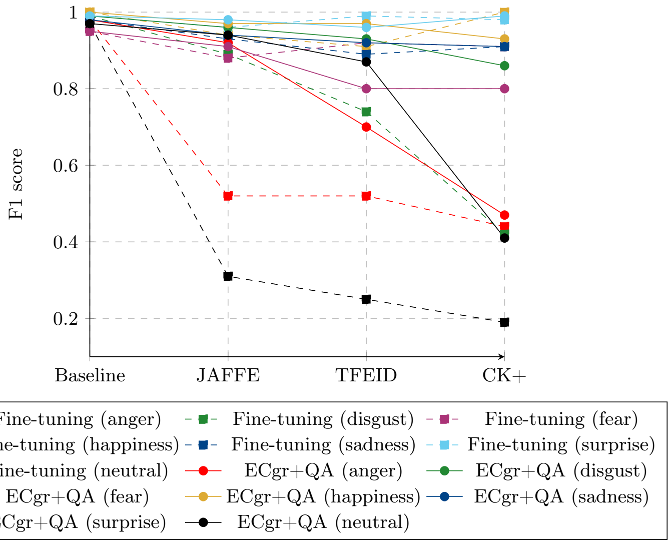

Fig. 5: Comparison of F1 scores by class on the MUG dataset between fine-tune and ECgr+QA, showcasing the continuous adaptation across JAFFE, TFEID, and CK+ datasets.

6 Conclusion

In this study, we presented a comprehensive investigation into the challenge of catastrophic forgetting in CNNs within the context of facial expression recognition, proposing a novel approach to mitigate its effects. We employed a pseudo-rehearsal method, specifically our emotion-centered generative replay (ECgr) with WGAN-GPs, to generate synthetic images for each dataset class and combined this with a filtering method to exclude images that could hinder retraining.

Across various tasks, ECgr consistently demonstrated superior performance compared to baseline and fine-tuned methods. Utilizing WGAN-GPs to synthesize task-specific data and our QA algorithm resulted in substantial knowledge retention. This confirms the potential of pseudo-rehearsal methods to effectively retrain CNNs without revisiting original datasets, offering a promising strategy for addressing memory decay, particularly in challenging scenarios like facial expression recognition.

Despite promising results with pseudo-rehearsal, its effectiveness may vary across network architectures, datasets, and tasks. Additionally, WGAN-GP-based data generation can be computationally expensive, limiting real-time use. These aspects highlight opportunities for future research, such as improved weight assignment algorithms and exploration of regularization techniques' synergy with pseudo-rehearsal approaches. Also, enhancements can be made to the quality of images generated by the WGAN-GP and in the architecture of classifiers, for example, using transformer networks. While the primary concern remains mitigating catastrophic forgetting, there is significant potential to improve results by optimizing synthetic data usage.

In an ideal scenario, combining classes from datasets such as MUG, JAFFE, and TFEID is recommended, as it enhances diversity and representation, leading to improved model performance. However, our method presents a viable alternative when such a combination is not feasible. This approach allows for flexibility in data augmentation and model training, providing a potential solution for scenarios with limited data availability or when data integration is challenging.

A Appendix

A.1 Evaluation of the MNIST Dataset

We evaluate our methodology across different domains using the MNIST dataset [4]. We applied the ECgr method with WGAN-GPs, dividing the dataset into class pairs (0 and 1, 2 and 3, 4 and 5, 6 and 7, 8 and 9), following the steps in Algorithm 1 and 2. Training began with the 0 and 1 class pair as the source dataset, with subsequent pairs used in the continual learning process. For continual learning, WGAN-GPs were trained for each digit, and the same process of combining target datasets with synthetic datasets generated by the generative networks was followed during retraining using the ECgr, ECgr+QA, and ECgr+wQA methods. As shown in Fig. 6, the behavior previously observed in FER datasets also held in this domain. The ECgr+QA and ECgr+wQA methods consistently outperformed fine-tuning in all retraining steps. Regarding the qualitative assessment of synthetic images, digits 4 and 5 were the most challenging to generate, and the QA algorithm struggled the most with these digits.

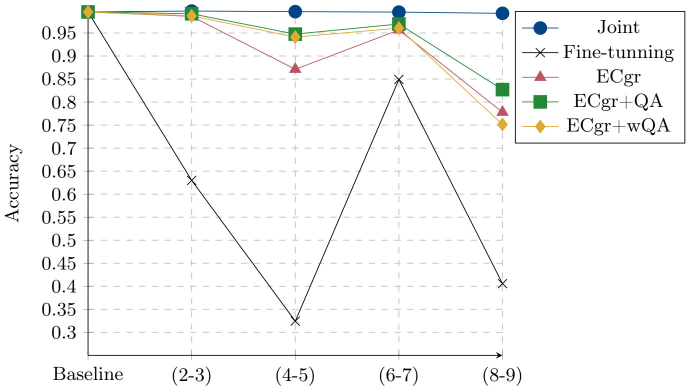

Fig. 6: Accuracy results for the MNIST (0-1) class pair subdataset, demonstrating continuous adaptation across subdatasets (2-3), (4-5), (6-7) and (8-9) relative to the baseline accuracy.

Time complexity. Time and computational complexity were evaluated on an Intel Core i7-8700 CPU and an NVIDIA GeForce GTX 1060 GPU. The algorithm took approximately 5200 seconds to complete 20 replications of a single CNN retraining on the MNIST dataset. Each batch, with 1024 images, took 3 to 5 seconds to process. Predictions for 1000 images took approximately 3 seconds.

Acknowledgements Thanks to CAPES SticAmSud (023-STIC-13), Univision Informática LTDA, Pontifícia Universidade Católica do Paraná (PUCPR) (grant 10844/2021) and CNPq (grant 306878/2022-4).

References

1. Aifanti, N., Papachristou, C., Delopoulos, A.: The mug facial expression database. In: 11th Intl Works Image Anal Multim Interac Services. pp. 1–4 (2010)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) 34th International Conference on Machine Learning. Proc. of Machine Learning Research, vol. 70, pp. 214–223. PMLR (2017)
3. Chen, C.C., ling Cho, S., Horszowska, K., Chen, M.Y., Wu, C.C., Chen, H.C., Yeh, Y.Y., Cheng, C.M.: A facial expression image database and norm for Asian population: a preliminary report. In: Farnand, S.P., Gaykema, F. (eds.) Image Quality and System Performance VI. vol. 7242, p. 72421D. SPIE (2009)
4. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012). https://doi.org/10.1109/MSP.2012.2211477
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
6. Khetarpal, K., Riemer, M., Rish, I., Precup, D.: Towards continual reinforcement learning: A review and perspectives. J Artif Intell Research **73**, 295–333 (2022)
7. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. of the National Academy of Sciences **114**(13), 3521–3526 (2017)
8. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)
9. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE CVPR Workshops. pp. 94–101 (2010)
10. Lyons, M., Kamachi, M., Gyoba, J.: The japanese female facial expression (jaffe) dataset. Zenodo (Apr 1998). https://doi.org/10.5281/zenodo.3451524
11. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
12. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) NIPS. vol. 30. Curran Associates, Inc. (2017)
13. Tannugi, D.C., Britto, A.S., Koerich, A.L.: Memory integrity of cnns for cross-dataset facial expression recognition. In: IEEE Intl Conf on Systems, Man and Cybernetics. pp. 3826–3831 (2019)
14. van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nature Machine Intelligence **4**(12), 1185–1197 (2022)
15. Zavaschi, T.H., Britto Jr, A.S., Oliveira, L.E., Koerich, A.L.: Fusion of feature sets and classifiers for facial expression recognition. Exp Syst App **40**(2), 646–655 (2013)
16. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. Proc Machine Learning Research **70**, 3987–3995 (2017)