Accepted Manuscript

To appear In Proceedings of the 15th International Conference on Automatic Face and Gesture Recognition (FG), November, 2020, Buenos Aires, Argentina.

CLIFER: Continual Learning with Imagination for Facial Expression Recognition

Nikhil Churamani and Hatice Gunes
Department of Computer Science and Technology, University of Cambridge, United Kingdom
Email: {*nikhil.churamani, hatice.gunes*}*@cl.cam.ac.uk*

***Abstract*— Current Facial Expression Recognition (FER) approaches tend to be insensitive to individual differences in expression and interaction contexts. They are unable to adapt to the dynamics of real-world environments where data is only available incrementally, acquired by the system during interactions. In this paper, we propose a novel continual learning framework with imagination for FER (CLIFER) that (i) implements *imagination* to simulate expression data for particular subjects and integrates it with (ii) a complementary learning-based dual-memory (episodic and semantic) model, to augment person-specific learning. The framework is evaluated on its ability to remember previously seen classes as well as on generalising to yet unseen classes, resulting in high F1-scores for multiple FER datasets: RAVDESS (episodic: F1$= 0.98 \pm 0.01$, semantic: F1$= 0.75 \pm 0.01$), MMI (episodic: F1$= 0.75 \pm 0.07$, semantic: F1$= 0.46 \pm 0.04$) and BAUM-1 (episodic: F1$= 0.87 \pm 0.05$, semantic: F1$= 0.51 \pm 0.04$).**

I. Introduction

Facial Expression Recognition (FER) approaches extract hierarchical feature representations using carefully hand-crafted features [1] or, more recently, data-driven methodologies [2], to analyse and understand human facial expressions. The recent success of deep learning has further enhanced their performance by reducing the dependency on the choice of features used, learning these directly from the data [3]. While this improvement is most observed in clean and noise-free environments, spontaneous expression recognition, in less-controlled real-world settings, is still challenging [4]. Thus, the Affective Computing (AC) [5] community is now focused on recognising expressions *in-the-wild* [6], robust to the movements of the observed individual, noise in the environment as well as occlusions [7]. Adapting FER models to contextual information as well as individual differences in expression, however, remains an open challenge [8]–[10].

When applied towards modelling of human behaviour such as recognising spontaneous expressions during human-robot interactions, analysing user experience while interacting with technology, affective game-playing, or diagnosing emotional or mental conditions in an individual, current FER models are not able to adapt to such dynamic interactions in real-time. They are not able to personalise towards the observed user, failing to adapt to individual characteristics such as distinctive facial features (for example, shape of the forehead, nose-width or thickness of the lips [11]) or characteristic attributes such as skin-colour or expressivity [9]. Some of the existing approaches that do focus on personalisation, depend on feature-level adaptations using a priori contextual attributions like gender and culture [8], apply unsupervised clustering [12] of data or selectively re-weight relevant samples for test subjects to improve performance [9]. Despite their sensitivity towards individualistic differences in expression, the applicability of these methods is limited as this contextual information may not always be available a priori in real-world situations. To mitigate these limitations, there is a need to develop models that (1) *continually learn* with each user, sensitive to their expression and (2) *dynamically adapt* to changing interaction conditions.

Unlike computational models, humans learn throughout their lifetime acquiring and integrating new information based on their experiences, without forgetting previous knowledge [13]. This ability to perpetually adapt acts as inspiration for most Machine Learning (ML) models aiming to achieve continual adaptation. Such Continual Learning (CL) approaches [14] address the problem of adaptability in models, enabling them to integrate new information without interfering with previously learnt knowledge. Although usually applied to object learning tasks [14], the basic principles of CL are transferable to FER systems, enabling them to personalise towards different subjects over repeated interactions. Such a learning approach can be beneficial for modelling user behaviour, understanding their current responses and predicting future behaviours.

The human ability to *imagine* allows them to simulate *imagined contact* [15], that is, imagined interactions with other individuals to augment future interactions. Such *imagination* can be particularly beneficial for FER models, enabling them to simulate additional (unseen) expression data for each subject, to personalise towards individual behaviour.

Inspired by CL principles, we propose a novel framework that can simulate *mental imagery* of different subjects and continually learn six facial expression classes, namely, *anger, happiness, fear, sadness, surprise* and *neutral*. The CLIFER framework consists of two components: (i) a generative auto-encoder model for *imagining* facial images for individuals to augment learning; and (ii) a dual-memory-based learning model for FER that adapts to novel data and balances long-term retention of knowledge. The framework is evaluated on its ability to remember previously seen expression classes as well on generalising to yet unseen classes for each subject, achieving high F1-scores across evaluations.

II. Background

In traditional deep learning models, incrementally learning new information may result in *catastrophic forgetting* [16], impacting previously accumulated knowledge. To guard

The work of Nikhil Churamani and Hatice Gunes is supported by EPSRC under grants ref: EP/R513180/1 and EP/R030782/1, respectively. The authors also thank German Parisi for his insights on the GDM model.

against such *interference*, models need to balance their ability to learn new information (*plasticity*) with preservation of knowledge (*stability*). To address this, most CL approaches take inspiration from cognitive processes in the human brain, focussed on memory-based learning [14].

The human ability to perpetually acquire and integrate information without affecting past knowledge is governed by neurophysiological mechanisms in the brain that contribute towards early *plasticity* and later to experience-driven *stable* consolidation of knowledge [13]. Complementary Learning System (CLS) [17] in the *hippocampal* and *neocortical* regions in the brain regulate rapid learning of novel information along with long-term retention of knowledge. The hippocampus learns non-overlapping representations of novel experiences, forming an *episodic* memory. These representations are later replayed to the neocortex for the slow-learning of overlapping representations as a *semantic* understanding.

Furthermore, the Pre-Frontal Cortex (PFC) also contributes towards the consolidation of specific recent memories as well as selective memory recall [18]. Such recollection of specific visual memories enables the simulation of *mental imagery*, involving similar processing in the PFC as actual sensory experiences [19]. This allows an individual to recall past and *imagine* different future situations to evolve their understanding of their environment. A similar effect is witnessed in the case of *imagined contact* [15], where one mentally simulates interactions with out-group members in social settings, facilitating possible future interactions.

Applying this understanding of neurocognitive processes involved in learning can help develop long-term adaptation capabilities in models. Additionally, embedding *imagination* to simulate possible future interactions can provide for additional data needed for context-dependent adaptation. In particular for FER, such mechanisms enable continual learning in the model, personalising towards each user with *imagined* interactions compensating for lack of sensory experiences.

III. Related Work

Inspired by neural mechanisms described above, many existing approaches implement CLS-based dual-memory systems for continual learning [14]. Balancing episodic representations with a semantic understanding of data, these approaches incrementally integrate information for long-term retention and recollection. Furthermore, replaying these learnt experiences from memory, commonly referred to as *rehearsal*, in the absence of external stimulus [20], alleviates forgetting by interleaving past experiences with current perception. Additionally, instead of storing and reusing actual samples of previously seen data, a generative or a probabilistic model can be used to draw *pseudo-samples* from memory [21]. Such a *pseudo-rehearsal* can significantly reduce the memory footprint of these models.

Kemker and Kanan [22] propose the FearNet model that comprises of three underlying models: a hippocampal network capable of recalling recent novel experiences, a generative PFC network for long-term retention of information using *pseudo-rehearsal*, and a third network that determines which of the above model to use for any given sample. Parisi et al. [23] propose the Growing Dual Memory (GDM) model comprising of an episodic memory that adapts to novel data in an unsupervised manner, along with a semantic memory that learns compact, overlapping representations using task-relevant (labelled) annotations.

A common aspect of most CLS-based approaches is the *pseudo-rehearsal* of previously seen data [14]. This is important as in the CL paradigm, data is made available only sequentially, normally only one class at a time. For FER, however, a major challenge to overcome for CL models is the lack of person-specific data. Most benchmark datasets contain only limited interactions with an individual subject, usually over only 2-3 interaction sessions. This problem becomes even more pronounced in the real-world where the model only witnesses a few interactions with any user, under limited contexts. This lack of person-specific data makes it difficult for models to personalise towards each user, overlooking nuances in expression and interaction context.

To mitigate these challenges, adversarial training is emerging as a potent solution [24]. Generative models trained using adversarial learning can generate photo-realistic data samples. For FER, this is useful in generating additional data [25]–[27], either to learn a separate affective model for each individual [10] or augment the overall training-set for offline training [28]. For continual learning, however, generative models are usually only used for *pseudo-rehearsal* of previously seen samples [21], [22]. The ability of such models to generate facial images that showcase different expressions [25] can also be beneficial for simulating unseen data. Having seen only a few images of a subject, models can generate additional data (akin to *imagination* in humans) for the individual to preempt future interactions.

In this paper, we present a novel framework (CLIFER) that integrates CLS-based neuroinspired approaches for continual learning of facial expressions. The CLIFER framework uses the GDM architecture [23] as the basis for learning to classify different facial expressions, adapting to each subject and extends the GDM architecture with an auto-encoder-based generative model to facilitate *imagination* for augmenting learning.

IV. Proposed Framework

The CLIFER framework (see Fig. 1) employs *imagination* to conditionally generate facial images of a subject for different expression classes, namely, *anger, happy, fear, sad, surprise* and *neutral*. These generated images augment learning in the dual-memory GDM model that classifies these images, balancing learning of new classes with retention of previously acquired knowledge (see Algorithm 1).

A. Auto-Encoder based Imagination

Generating facial images for a subject requires the model to first encode their facial features, and then translate their images to showcase different expressions, preserving their identity. To achieve this, we use a Conditional Adversarial Auto-Encoder (CAAE)-based [25] *imagination* model (see

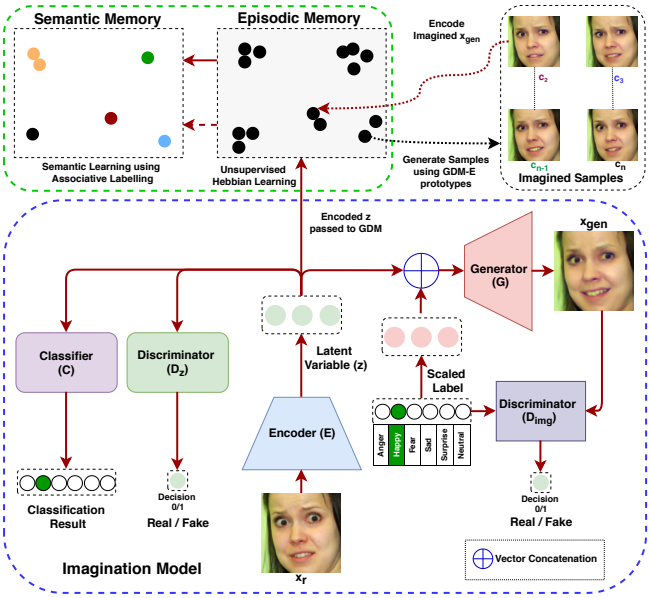


Fig. 1: The CLIFER Framework: x_r is encoded and input to different modules for further processing. For training the dual-memory and the classifier, encoded x_{gen} are also used.

Fig. 1) that takes an input image (x_r) and generates translated images (x_{gen}) for each of the 6 expressions. The model architecture is adapted from ExprGAN [25] given its ability to generate expression-translated photo-realistic images. Different components of the model, with the necessary adaptations, are described below:

1) Encoder (E): The Encoder takes a normalised facial image (each pixel in $x_r \in [-1, 1]$) and encodes it into a $50-$d latent vector z. It uses 4 stacked convolution layers (with $64, 128, 256, 512$ filters, of size (5×5) each) with *ReLU* activation. These are followed by a fully-connected layer with 50 units using *tanh* activation, resulting in the latent vector $z \in [-1, 1]$. Once trained, the Encoder model is used as a feature extractor for the overall framework.

2) Generator G: The generator takes the encoded z-vector, along with the one-hot coded target expression label y and generates photo-realistic images x_{gen} representing the label. y is transformed to range in $[-1, 1]$ (where -1 represents 0) for a fair concatenation to z. It is then scaled[1] and appended to z, enforcing the label on generated images. G implements 6 stacked convolution layers (using *ReLU*) performing transposed convolutions (with $1024, 512, 256, 128, 64, 32$ filters of size (5×5) each) to generate the resultant image $(G(E(x_r), y))$ with the same dimensions as x_r. An $\mathcal{L}_1$ *reconstruction loss* ($\mathcal{L}_{rec}$) is imposed on G, enabling reconstruction of images:

$$\min_{E,G} \mathcal{L}_{rec} = L_1(x_r, G(E(x_r), y)), \tag{1}$$

A pre-trained VGG-face model [29] is used for ID preservation(similar to [25]), enforcing similar facial features between input and generated images. L_{ID} is computed as:

$$\min_{E,G} \mathcal{L}_{ID} = \sum_{l} L_1(\phi_l(G(E(x_r), y), \phi_l(x_r)), \tag{2}$$

[1] $scale = |\frac{t}{l}|$, where t, l are dimensions of z, y, respectively

where ϕ_l represents the l-th layer VGG-face features. The activation from the first 5 convolution layers are compared for x_r and x_{gen}. Additionally, a total variation regularisation (L_{tv}) [30] is imposed (similar to [25]) that uses the sum of the absolute differences for neighbouring pixel-values in the input images to avoid 'ghosting' artefacts.

3) Discriminator (D_{Img}): D_{Img} evaluates the photo-realistic quality of x_{gen}, playing a $\min\max$ game with G. It implements 4 stacked convolution layers (using *LeakyReLU* with $\alpha = 0.2$) with the condition label also scaled[1] and concatenated to the input of the first 3 layers to further enforce the condition on generated images. The output of D_{Img} represents the probability of an image belonging to the real image distribution. The objective function for the $\min\max$ game between G and D_{img} is given as below:

$$\begin{aligned}\min_{G}\max_{D_{Img}} \mathcal{L}_{img} = \mathbb{E}_{x\sim p_{data}(x)}[log(D_{Img}(x))] \\ +\mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_{Img}(G(E(x), y)))],\end{aligned} \tag{3}$$

where $p_{data}(x)$ represents the distribution of training data.

4) Discriminator (D_z): D_z regularises the latent vector z, ensuring that it is uniformly distributed. This assures that the encoded features lie on the same 'face manifold' [26], to avoid generating distorted images. E and D_z play a $\min\max$ game with D_z predicting the probability of z originating from a uniform distribution $\mathcal{U}(-1, 1)$. The objective function for the $\min\max$ game between E and D_z is given below:

$$\begin{aligned}\min_{E}\max_{D_z} \mathcal{L}_z = \mathbb{E}_{z_{prior}\sim p_{prior}(z)}[logD_z(z_{prior})] \\ +\mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_z(E(x)))],\end{aligned} \tag{4}$$

where $p_{prior}(z)$ denotes the prior distribution imposed on z and $p_{data}(x)$ represents the true latent distribution.

5) Classifier C: A classifier model is used to further enforce the label condition on the generated images. C is implemented as a Multilayer Perceptron (MLP) resulting in a *SoftMax* output. For the input image, the label comes from

Algorithm 1 CLIFER: GDM with Imagination

1: Pre-train the Imagination Model as described in Section IV-A.
2: Initialise **GDM-E** and **GDM-S** (based on [23]), each starting with two random neurons $A = \{w_1, w_2\}$.
3: Sample input images for a subject, one class at a time, for C classes.
4: **for** $c = 1$ to C **do**
5: ***Learning:***
6: Sample M mini-batches $X_r^m = \{x_{r^1}, x_{r^2}, .., x_{r^N}\}$ of N input (real) images for a subject, each representing an episode.
7: **for** $m = 1$ to M **do**
8: **for** $n = 1$ to N **do**
9: Encode x_{r^n} to latent vector z_n: $E(x_{r^n}) = z_n$. The encoded X_r^m acts as input to **GDM-E**.
10: Select 2 BMUs (b_n, s_n) closest to each z_n.
11: Update **GDM-E** weights, connectivity, label matrices [23].
12: **end for**
13: Input winner neurons $(B^m = \{b_1, b_2, ..b_N\})$ for each z_n, along with the corresponding labels as a mini-batch to **GDM-S** and execute **Steps 10-11** for **GDM-S**.
14: **end for**
15: ***Imagination:***
16: Input B^m to the Generator along with different labels vectors to conditionally generate X_{gen} with different expressions.
17: Repeat Steps $6 - 14$ using X_{gen}.
18: **end for**

the data sample, while for the generated image, the condition label is used. The classifier loss is defined as a *weighted sample cross-entropy loss*, adopted from the loss used in [31] for ID classification.

$$\min_{\theta_C} \mathcal{L}_C = \phi(x_r, y_r) + \lambda_c \phi(x_{gen}, y_{cond}), \tag{5}$$

where $\lambda_c = \exp^{-ep}$ ($ep \in [0, n)$ is the current epoch) is the weight for x_{gen}. θ_C represents model parameters and ϕ represents categorical cross-entropy loss.

6) Objective Function: The above-described models are trained together with the overall objective function computed as the weighted-sum of all individual losses.

$$\begin{aligned}\min_{E,G,D_z} \max_{D_{Img},C} \mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{ID} + \lambda_3 \mathcal{L}_z \\ + \lambda_4 \mathcal{L}_{img} + \lambda_5 \mathcal{L}_C + \lambda_6 \mathcal{L}_{tv}\end{aligned} \tag{6}$$

The weights for the different loss terms are set to λ_1=1, λ_2=0.33, λ_3=0.01, λ_4=0.01, λ_5=0.1, λ_6=8.5×10^{-5}, inspired from [25] and evaluated empirically by manual inspection of the generated facial images for photo-realism as well as identity preservation. The model is trained using *Adam* optimiser (lr=0.0002) with a batch-size of 49, similar to [25].

B. Complementary Learning-based Dual-Memory Model

The CLIFER framework uses the Growing Dual Memory (GDM) architecture [23] as the basis for incrementally acquiring and integrating knowledge. It consists of two hierarchically arranged recurrent self-organising neural networks representing the *episodic* (GDM-E or Growing Episodic Memory as in [23]) and *semantic* (GDM-S or Growing Semantic Memory as in [23]) memory, respectively. Each memory implements a Growing When Required (GWR) neural network with Gamma-filtering [23] for spatio-temporal processing of features. *Neurogenesis* in the models, that is, regulating when to add new neurons, is controlled by their ability to represent input data with new neurons and connections added when the activation of the model falls below a given *activation* threshold. Each GWR model implements Gamma-filtering with $K \in [0, n]$ context-descriptors that govern its temporal resolution. In this implementation, we set $K = 0$ to focus on individual frames.

1) Episodic Memory (GDM-E): Each input image is encoded using the pre-trained Encoder model (see section IV-A) into a $50-$d feature vector that acts as input to the GDM model. GDM-E sequentially receives these input vectors and rapidly learns (using a high learning-rate of $\epsilon = 0.2$) non-overlapping representations for each data point. This is achieved using a distance-based similarity measure, implementing unsupervised Hebbian-based learning (see [23] for details). As it receives data, one class at a time, it creates feature prototypes for each input sample. GDM-E is trained using lenient *activation* ($\alpha = 0.4$) and *habituation* thresholds ($\eta_{thresh} = 0.5$) that allow for new neurons to be added whenever an input differs even slightly from existing prototypes. The different learning hyper-parameters for the model are optimised using the Hyperopt[2] Python Library.

[2] `http://hyperopt.github.io/hyperopt/`

2) Semantic Memory (GDM-S): With GDM-E learning feature prototypes for individual data points, GDM-S learns compact overlapping representations that can generalise across a particular class. After each episode, here defined as a mini-batch of sequential input from each video sample, GDM-S receives the Best Matching Units (BMUs) or *winner* neurons from GDM-E, along with label annotations. A frequency-based associative labelling scheme [23] is used to associate feature prototypes with their respective labels (depicted by the mode of the histogram) and regulate *neurogenesis*. New neurons are added to GDM-S only if the existing neurons are not able to correctly classify the input. This reduces the negative impact of the new information from GDM-E on the acquired knowledge by GDM-S. It uses a slower learning rate ($\epsilon = 0.02$), and stricter *activation* ($\alpha = 0.2$) and *habituation* thresholds ($\eta_{thresh} = 0.2$), such that new neurons are added very slowly. These thresholds, along with associative labelling, make sure that existing neurons are *habituated* extensively before new neurons are added, resulting in the model learning overlapping representations.

3) Pseudo-rehearsal: As the GDM model encounters data sequentially, it is possible that samples from one class may overwrite feature representations learnt from previous classes. To guard against this forgetting, previously encoded feature prototypes, in the form of trajectories of neural activations from GDM-E are periodically (at the end of each episode) replayed to both memories. For all the neurons, an activation trace is maintained, recording the order in which they were fired, replayed in absence of external stimuli [23].

4) Imagination: Different from *pseudo-rehearsal*, implementing *imagination* in the GDM model allows simulation of additional data for all the classes to improve feature representation. After receiving data samples from a particular class, BMUs from the GDM-E are passed to the *imagination* model (see Section IV-A) which generates facial images for the subject encoding each expression class. These imagined images are encoded and played to both GDM-E and GDM-S, augmenting learning in the model.

V. Experimentation and Results

A. Experiments

To evaluate the proposed framework on its ability to continually learn to recognise facial expressions, we conduct two experiments:

1) Experiment 1 - Remembering Seen Facial Expressions: Remembering previously seen classes establishes the model's ability to guard against *catastrophic forgetting*, forming an essential quality of CL systems. Thus, after witnessing each new class, the model is evaluated on data from all the classes seen so far.

2) Experiment 2 - Adapting to New Facial Expressions: Learning with data from only one class at a time can also bias the model, negatively impacting its ability to generalise to unseen data. As FER models need to be applied to real-world scenarios, it is important to preserve the *generalisability* of the model. Thus, the framework is also evaluated on its ability to generalise to unseen facial expression classes. After

witnessing each new class, we evaluate the model on all seen and unseen classes for each subject.

B. Experimental Conditions

To provide an extensive and fair evaluation of the proposed framework, we compare different settings of the GDM model, that is, with and without the pseudo-replay mechanism, and the CLIFER framework using imagination, with a baseline MLP classifier. The four conditions are:

1) MLP Baseline: An MLP-based classifier (based on the one discussed in Section IV-A) is trained using traditional batch-learning, one class at a time. This forms the baseline to compare traditional ML with different variants of GDM.

2) GDM: This setting consists only of the dual-memory set-up without any pseudo-rehearsal mechanism to establish a baseline measurement for the GDM model for FER.

3) GDM + Replay: In this setting, the *pseudo-rehearsal* mechanism is added to the GDM model to guard against forgetting of previously seen classes.

4) CLIFER Framework: The GDM model is embedded with the auto-encoder to imagine additional images for each subject, without any explicit *pseudo-rehearsal* mechanism.

C. Datasets

The proposed framework needs to be evaluated for each user, with data from different expression classes made available only incrementally. As a result, we train and evaluate the framework on FER datasets that contain class labels for different expressions as well as ID labels to separate data for each subject. For evaluation, we select a subset of the 'big six' expression classes [32], namely, *anger, sadness, happiness, surprise* and *fearful* along with *neutral* samples. We omit *disgust* as the number of data samples for each subject for *disgust* vary a lot in the selected datasets (ranging from a handful in one to more than a hundred in another). The different datasets used in this work are explained below:

1) iCV-MEFED: The iCV-MEFED dataset [33] consists of facial images from 125 actors expressing 8 different emotional expressions, namely, the 'big six' along with *neutral* and *contempt*. For each image, annotations are provided for *dominant* (primary) and *complementary* (secondary) expressions. We split the data into six classes only on the basis of dominant class labels, combining the complementary labels to enhance data variability under each class.

2) RAVDESS: The RAVDESS dataset [34] consists of recordings from 24 actors performing monologues representing 8 different expressions namely, the 'big six' along with *neutral* and *calm*. For each class, except *neutral*, monologues are expressed in two intensities; *high* and *low*. We again split data into six classes and combine different intensity data for each class following the same variability assumption.

3) MMI: The MMI Expression database [35] consists of recordings from subjects reacting to different affective stimuli. We sample 10 different subjects providing data for the six classes used in this work. As annotations are provided at clip-level, we use the middle 3 (peak) frames from each clip (similar to other works [2]) to represent each class and the first frame to represent *neutral*.

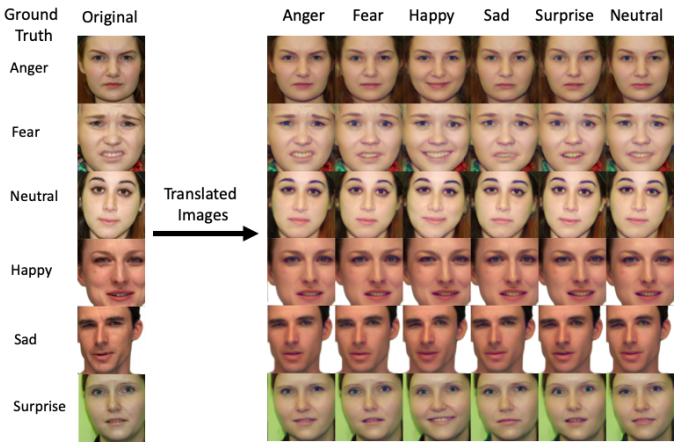


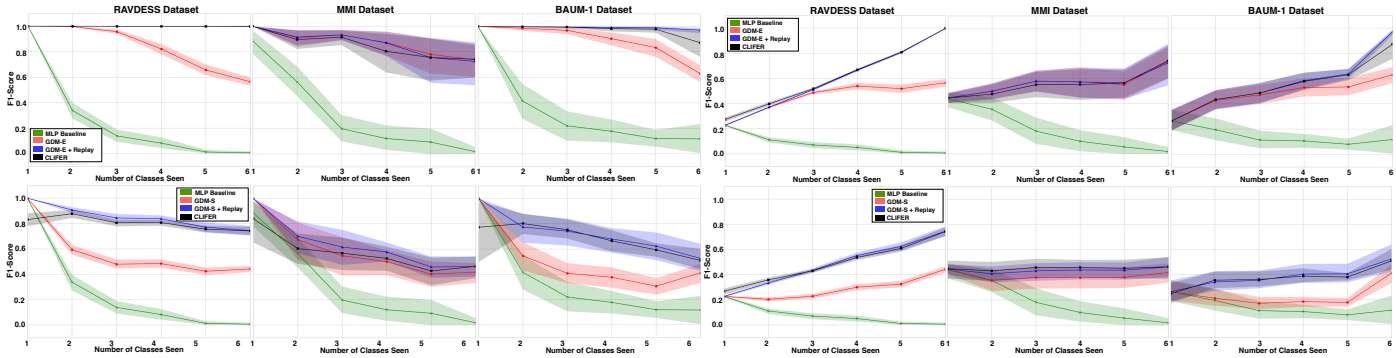
Fig. 2: Generated facial images for randomly selected subjects from iCV-MEFED and RAVDESS datasets.

4) BAUM-1: The spontaneous expressions collection from the BAUM-1 dataset [36] is selected to evaluate the framework for spontaneous FER. We sample 9 different subjects providing data for all six classes used in this work. Since annotations are provided at clip level, we split each clip into $500-$millisecond chunks (15 frames per-chunk) and select the 2 penultimate frames from each chunk, providing *apex* facial frames for the respective class. This results in $15-20$ frames extracted per clip for a particular class. A similar chunking mechanism is used for the RAVDESS dataset.

D. Results

1) Imagining Facial Images: The proposed framework relies on the auto-encoder model to *imagine* facial images of the subjects, conditioned on different expression class labels. For this, face-centred (96×96) RGB images are passed as input to the auto-encoder model. The model translates the input image to each of the 6 classes used in this work, generating images of the subject. The model is trained, a priori, by combining both the training and validation sets from iCV-MEFED and RAVDESS datasets. While iCV-MEFED provides stereotypical expression samples for each class, the monologue set-up of RAVDESS allows for variability in these expressions for robust feature encoding. The resultant class-conditioned generated images are used to augment learning in the dual-memory model. Examples of generated images, after training, can be seen in Fig. 2. The model successfully translates input images to represent different expressions, preserving the subject's identity.

2) Continual Learning for FER: The GDM architecture aims to learn discernible feature representations for each class to mitigate interference from new samples, making learning class-order agnostic. In practice, however, for FER we found the model's performance to be sensitive to the order of learning different classes for each subject. To quantify this, we selected 6 different class orders, starting with each of the 6 classes used in this work. The rest of the order was selected randomly as evaluating all possible permutations (720, in this case) was not computationally feasible. The GDM model (under all variants) was trained and tested separately for each subject from the RAVDESS,



(a) Experiment 1: GDM-E (top) and GDM-S (bottom) performance on remembering seen classes, after witnessing each new class.

(b) Experiment 2: GDM-E (top) and GDM-S (bottom) performance on generalising to new classes, after witnessing each new class.

Fig. 3: F1-Scores with 95% confidence intervals for the experiment conditions on RAVDESS, MMI and BAUM-1 datasets.

Dataset	GDM				GDM + Replay				CLIFER			
	GDM-E		GDM-S		GDM-E		GDM-S		GDM-E		GDM-S	
	First	Final	First	Final	First	Final	First	Final	First	Final	First	Final
RAVDESS	**(H=65.0, p<<0.05)**	(H=8.1, p=0.15)	**(H=23.6, p<0.05)**	(H=6.1, p=0.29)	**(H=64.8, p<<0.05)**	(H=0.2, p=0.99)	**(H=59.7, p<<0.05)**	(H=0.3, p=0.99)	**(H=36.1, p<<0.05)**	(H=0.2, p=0.99)	**(H=49.7, p<<0.05)**	(H=2.3, p=0.79)
MMI	**(H=31.9, p<<0.05)**	(H=0.6, p=0.98)	**(H=30.6, p<<0.05)**	**(H=16.6, p<0.05)**	**(H=32.9, p<<0.05)**	(H=0.1, p=0.99)	**(H=32.1, p<<0.05)**	**(H=17.0, p<0.05)**	**(H=33.8, p<<0.05)**	(H=0.4, p=0.99)	**(H=17.6, p<0.05)**	**(H=14.6, p<0.05)**
BAUM-1	**(H=25.6, p<0.05)**	(H=0.4, p=0.99)	**(H=32.2, p<<0.05)**	(H=7.7, p=0.17)	**(H=26.1, p<<0.05)**	(H=1.2, p=0.94)	**(H=21.9, p<<0.05)**	(H=0.9, p=0.97)	**(H=25.5, p<<0.05)**	(H=1.1, p=0.94)	**(H=22.3, p<0.05)**	(H=0.3, p=0.99)

TABLE I: Kruskal-Wallis H-test results for Experiment 2 comparing accuracy after the First and Final class, across orders.

MMI and BAUM-1 datasets. While RAVDESS and MMI datasets provide an evaluation on *posed* samples, BAUM-1 evaluates the model on *spontaneous* FER.

We compare model performance on the 3 datasets, evaluating accuracy scores after seeing the first class as well as after all classes. Kruskal-Wallis H-test results (see Table I) result in a significant difference ($p < 0.05$) in model performance for Experiment 2 between the 6 class orders. Starting with *neutral* results in the best performance, on average. Although not significant, a similar effect is seen for Experiment 1. To further substantiate these results, we select 5 more class orders, starting with *neutral* and followed by randomly selected classes. This also results in a similar effect yet the difference within these 5 orders is not significant. As a result, for further evaluations, we set the order of learning classes to start with *neutral*, followed by (randomly selected) *happy, surprise, anger, fear* and *sadness*. The results for the RAVDESS, MMI and BAUM-1 datasets for the two experiments can be seen in Fig. 3a and Fig. 3b, respectively. The GDM model outperforms the MLP baseline on all the 3 datasets for both the experiments. The GDM + Replay and the proposed CLIFER framework (GDM with *imagination*) perform better than the standard GDM model, with CLIFER, on average, performing the best across all settings for RAVDESS (episodic: F1=0.98 ± 0.01, semantic: F1=0.75 ± 0.01), MMI (episodic: F1=0.75 ± 0.07, semantic: F1=0.46 ± 0.04) and BAUM-1 (episodic: F1=0.87 ± 0.05, semantic: F1=0.51 ± 0.04) datasets.

VI. Discussion

Our results show that the proposed CLIFER framework, embedded with *imagination* capabilities, can remember previously seen classes as well as generalise to yet unseen classes. Comparing the different settings of the GDM model, we see that the proposed framework outperforms the MLP baseline (see Section V-B) trained using traditional batch-learning. The MLP experiences *catastrophic forgetting* when training data is presented to it, one class at a time. The GDM model, however, can guard against such forgetting by learning distinct feature prototypes, in both episodic and semantic memories. CLIFER performs the best, on average, across all evaluations achieving similar, if not better in most cases, results to the GDM + Replay condition which relies on an explicit *pseudo-rehearsal* mechanism. This reduces the need for maintaining and repeatedly replaying neural trajectories of past BMUs to both the memories at the end of each episode. Instead, CLIFER uses the auto-encoder-based *imagination* model to generate additional data for the subject for seen as well as yet unseen classes, augmenting learning. Thus, learning with *imagination* performs better than other conditions in Experiment 2. As the model *imagines* data for all the different classes, it *implicitly* replays data from previously seen classes as well, enabling its high performance on Experiment 1. The GDM + Replay condition also achieves high performance in both evaluations as a result of the explicit *rehearsal* mechanism guarding it against forgetting.

The proposed framework is found to be sensitive to the *order* in which the different classes are learnt. Even though it outperforms the MLP baseline for all the orders, the order significantly impacts the model's learning behaviour. This is counter-intuitive compared to the standard application of CL for object learning [14] which is found to be order-agnostic. This can be because most objects learnt by these models have very distinctive physical features. Hence, learnt feature representations do not overlap significantly. For FER, however, as the models learn how a specific individual expresses different emotions, the learnt feature representations may overlap significantly resulting in the order impacting model performance. Other approaches in curriculum-based learning [37] have also witnessed a specific order of learn-

ing (starting with high-intensity samples) enhancing model performance although they do not evaluate the models for continual learning. In our experiments starting with *neutral* results in the best performance, particularly in Experiment 2. This can be due to two reasons. Firstly, as *neutral* represents a normative baseline for an individual's facial expressions, learning this norm *first* allows the model to form distinct feature prototypes for subsequent classes, sensitive to the smallest deviation from this norm. Secondly, *imagination* also impacts model performance as the imagined images are used to augment learning. We see (in Fig. 2) that the generated images carry forward some of the features from the original image, for example, slightly raised eyebrows are seen in the images generated from a *surprise* sample. The images generated from *neutral* thus, have the least influence of the original image, resulting in images encoding distinctive expressions.

Such overlap in learnt feature representations for different classes is also seen to affect learning in GDM-S, resulting in lower performance scores than GDM-E. The task for GDM-S is to consolidate knowledge over time by forming compact overlapping feature representations. This should intuitively result in better performance than GDM-E as the model witnesses more classes. In our experiments, however, this results in a lower performance, despite the *neurogenesis* mechanism in GDM-S guarding against this. As GDM-S consolidates knowledge over different episodes, the overlap in feature prototypes from different classes may cause interference and degrade the performance of GDM-S. Improving feature representation in the auto-encoder can guard against such interference, not only resulting in better feature prototypes but also improving the images generated to augment learning.

VII. Conclusion and Future Work

In this paper, we present a novel framework for continual learning of facial expressions that combines neurocognitive principles of complementary learning with *imagination* to augment learning. At each step, the auto-encoder-based *imagination* model generates additional data that improves its ability to recall previously seen classes as well as generalise its learning to yet unseen expressions. The experiments highlight the ability of the proposed framework to incrementally learn and adapt to an individual's expressions, enhancing the real-world applicability of FER models in settings that require adaptation to individual subjects.

References

[1] L. Zhong *et al.*, "Learning active facial patches for expression analysis," in *Proc. CVPR*. IEEE, 2012, pp. 2562–2569.

[2] S. Li *et al.*, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018.

[3] D. Kollias *et al.*, "Training Deep Neural Networks with Different Datasets In-the-wild: The Emotion Recognition Paradigm," in *Proc. IJCNN*, 2018, pp. 1–8.

[4] E. Sariyanidi *et al.*, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE PAMI*, vol. 37 (6), pp. 1113–1133, 2015.

[5] R. Picard, *Affective Computing*. MIT Press, 1997.

[6] J. Kossaifi *et al.*, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[7] G. Zen *et al.*, "Learning Personalized Models for Facial Expression Analysis and Gesture Recognition," *IEEE Multimedia*, vol. 18 (4), pp. 775–788, 2016.

[8] O. Rudovic *et al.*, "Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy," *Science Robotics*, vol. 3 (19), 2018.

[9] W. Chu *et al.*, "Selective Transfer Machine for Personalized Facial Expression Analysis," *IEEE PAMI*, vol. 39 (3), pp. 529–545, 2017.

[10] P. Barros *et al.*, "A Personalized Affective Memory Model for Improving Emotion Recognition," in *Proc. ICML*. PMLR, 2019, pp. 485–494.

[11] R. Brunelli *et al.*, "Face recognition: features versus templates," *IEEE PAMI*, vol. 15 (10), pp. 1042–1052, 1993.

[12] G. Zen *et al.*, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proc. ICMI*. ACM, 2014, pp. 128–135.

[13] J. D. Power *et al.*, "Neural plasticity across the lifespan," *WIREs Dev Biol*, vol. 6, p. e216, 2017.

[14] G. Parisi *et al.*, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[15] R. J. Crisp *et al.*, "Can imagined interactions produce positive perceptions?: Reducing prejudice through simulated social contact." *American Psychologist*, vol. 64 (4), pp. 231–240, 2009.

[16] R. Kemker *et al.*, "Measuring Catastrophic Forgetting in Neural Networks," in *AAAI*, 2018.

[17] R. C. O'Reilly *et al.*, "Complementary learning systems," *Cognitive Science*, vol. 38 (6), pp. 1229–1248.

[18] T. Kitamura *et al.*, "Engrams and circuits crucial for systems consolidation of a memory," *Science*, vol. 356 (6333), pp. 73–78, 2017.

[19] S. D. Slotnick *et al.*, "Visual memory and visual mental imagery recruit common control and sensory regions of the brain," *Cognitive Neuroscience*, vol. 3, no. 1, pp. 14–20, 2012.

[20] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[21] H. Shin *et al.*, "Continual Learning with Deep Generative Replay," in *NIPS*, 2017, pp. 2990–2999.

[22] R. Kemker *et al.*, "Fearnet: Brain-inspired model for incremental learning," *CoRR*, vol. abs/1711.10563, 2018.

[23] G. I. Parisi *et al.*, "Lifelong Learning of Spatiotemporal Representations With Dual-Memory Recurrent Self-Organization," *Frontiers in Neurorobotics*, vol. 12, p. 78, 2018.

[24] J. Han *et al.*, "Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives," *IEEE CI Magazine*, vol. 14 (2), pp. 68–81, 2019.

[25] H. Ding *et al.*, "ExprGAN: Facial Expression Editing with Controllable Expression Intensity," *AAAI*, pp. 6781–6788, 2018.

[26] Z. Zhang *et al.*, "Age Progression/Regression by Conditional Adversarial Autoencoder," in *IEEE CVPR*, 2017, pp. 4352–4360.

[27] I. Abbasnejad *et al.*, "Using Synthetic Data to Improve Facial Expression Analysis With 3D Convolutional Networks," in *ICCV Workshops*. IEEE, 2017.

[28] D. Saez-Trigueros *et al.*, "Generating photo-realistic training data to improve face recognition accuracy," *CoRR*, vol. abs/1811.00112, 2018.

[29] O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. BMVC*, 2015.

[30] A. Mahendran *et al.*, "Understanding deep image representations by inverting them," in *IEEE CVPR*, 2015.

[31] Y. Shen *et al.*, "FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis," in *IEEE CVPR*, 2018.

[32] P. Ekman *et al.*, "Constants across cultures in the face and emotion," *JPSP*, vol. 17 (2), p. 124, 1971.

[33] J. Guo *et al.*, "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26 391–26 403, 2018.

[34] S. R. Livingstone *et al.*, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, 2018.

[35] M. F. Valstar *et al.*, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," in *Proc. LREC Workshop on EMOTION*, 2010, pp. 65–70.

[36] S. Zhalehpour *et al.*, "BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States," *IEEE Affective Computing*, vol. 8 (3), pp. 300–313, 2017.

[37] L. Gui *et al.*, "Curriculum learning for facial expression recognition," in *IEEE FG*, 2017, pp. 505–511.