AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition

Hong-Xia Xie
National Chiao Tung Univresity
hongxiaxie.ee08g@nctu.edu.tw

Ling Lo
National Chiao Tung Univresity
lynn97.ee08g@nctu.edu.tw

Hong-Han Shuai
National Chiao Tung Univresity
hhshuai@nctu.edu.tw

Wen-Huang Cheng
National Chiao Tung Univresity
whcheng@nctu.edu.tw

ABSTRACT

Micro-expressions (MEs) are important clues for reflecting the real feelings of humans, and micro-expression recognition (MER) can thus be applied in various real-world applications. However, it is difficult to perceive and interpret MEs correctly. With the advance of deep learning technologies, the accuracy of micro-expression recognition is improved but still limited by the lack of large-scale datasets. In this paper, we propose a novel micro-expression recognition approach by combining Action Units (AUs) and emotion category labels. Specifically, based on facial muscle movements, we model different AUs based on relational information and integrate the AUs recognition task with MER. Besides, to overcome the shortcomings of limited and imbalanced training samples, we propose a data augmentation method that can generate nearly indistinguishable image sequences with AU intensity of real-world micro-expression images, which effectively improve the performance and are compatible with other micro-expression recognition methods. Experimental results on three mainstream micro-expression datasets, i.e., CASME II, SAMM, and SMIC, manifest that our approach outperforms other state-of-the-art methods on both single database and cross-database micro-expression recognition.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Computer vision**; **Image representations**.

KEYWORDS

micro-expression recognition; deep neural networks; data augmentation; AU graph relation learning

ACM Reference Format:
Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2020. AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394171.3414012

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414012

1 INTRODUCTION

"Most lies succeed because no one goes through the work to figure out how to catch them"

— Paul Ekman *(1934-)*

Different from facial expressions, Micro-expressions (MEs) are subtle and spontaneous facial muscle movements that usually last less than 200ms, which cannot be hidden even for professional actors [7]. Therefore, MEs are considered as the real reflections of human emotion [30] and are especially useful in high-risk situations, e.g., depression recovery test [54], negotiation with terrorists [39], criminal investigations [7]. Take the depression recovery test as an example. Patients with major depressive disorder require the interview with psychiatrists to make sure that they are recovered for leaving the hospitals; otherwise they may commit suicide after leaving. In this case, they may pretend to be fine by fake smiles, which could easily fool facial expression recognition systems. In contrast, MEs reveal their true mental states but are inclined to be ignored since they only show up for less than 200ms during the interview. Therefore, Micro-Expression Recognition (MER) plays an important role in interpreting people's genuine emotion, which can facilitate a variety of practical applications in our daily life [31, 39, 55].

In the last decade, early approaches of MER are based on handcrafted feature extraction according to the survey in [31], e.g. local binary pattern (LBP) and local quantized pattern (LQP) for spatial texture features, LBP on three orthogonal planes (LBP-TOP) for spatio-temporal features. However, LBP-TOP and its variants are unable to recognize the subtle motion of facial movements. With the advance of deep learning technologies, a recent line of research studies the MER problem by data-driven approaches [16, 17, 36, 43, 44]. For instance, 3DConvNet is used in [36] to discover the spatio-temporal relationship between ME sequences with high-level features. Monu et al. [44] first acquires minute variations of a ME sequence in an RGB image using dynamic imaging technique and then spots these variations with a simple convolution network. However, due to the limited amount of ME training samples, deep models tend to be restrained, and thus the deep relations between features remain tangled. One promising solution is to leverage the Action Units (AUs) of Facial Action Coding System (FACS) [10, 41] to identify the important face regions and to extract meaningful features from the interplay of different AUs for MER. However, current works only focus on detecting AUs without further modeling the relations in between for MER.

In fact, it is challenging to recognize MEs due to the following three issues: i) *Subtle changes in a short period.* The low intensity of spontaneous and brief facial movements are difficult to be directly extracted. Therefore, after being trained by professional micro-expression training tools, humans still detect and recognize MEs from videos with low accuracy [9]. ii) *Complicated interplay between facial regions.* Different MEs are based on different combinations of facial regions instead of one important region. For example, happiness requires flexing muscles around the mouth and contracting the muscles around eyes (called the orbicularis oculi). Without any of them, it can be a fake smile or a squinting. iii) *Insufficient and unbalanced training data.* The existing ME datasets [6, 24, 47, 48] contain a limited amount and are heavily unbalanced due to the collection difficulties, e.g., fear is hardly triggered. The limited amount of data makes the training of an end-to-end deep learning model challenging. On one hand, the recognition model suffers from overfitting issue due to the limited labeled data, resulting in low accuracy when recognizing unseen data. On the other hand, training an MER model on a dataset with the unbalanced classes favors the majority class, resulting in severely biased prediction results.

Therefore, to address these issues, we propose two modules: i) *AU-assisted Graph Attention Convolutional Network (AU-GACN)*, which extracts discriminative features for subtle MEs by fully exploiting the relation between AUs to address the first and second challenges, and ii) *AU Intensity Controllable Generative Adversarial Nets (AU-ICGAN)*, which effectively generates the synthetic data of MEs for addressing the third challenge and enabling the training of AU-GACN. Specifically, since AU labels can be vital hints for further emotion label classification, in addition to learning micro-expression classifiers by minimizing label errors between predicted emotion labels and the ground-truth emotion labels, AU-GACN adds AU classification as an auxiliary task. Firstly, we apply a lightweight 3D ConvNet backbone network to extract the spatio-temporal features for AUs. Afterward, the AU features are further used as the node features of AU relation graph. AU-GACN leverages self-attention graph pooling and graph convolutional networks to enhance the AU node features for ME classification. Since well-defined muscle information (i.e., AUs) can be vital hints to MER, instead of merely using discrete emotion classes [31], we propose a new loss combining the AU multi-label loss and ME classification loss to train the proposed AU-GACN.

Moreover, to address the third challenge, we propose AU-ICGAN for MER task to enrich the limited training samples, preventing our deep neural network from overfitting. Specifically, the proposed data augmentation approach is designed to synthesize facial images conditioned on AU intensity extracted from real-world micro expression images. Therefore, in addition to the GAN loss and AU intensity loss, we take the image structure similarity along with image sequence authenticity into consideration to make generated ME sequences more realistic. Therefore, the proposed AU-ICGAN can be more precise on simulating facial movements with microscale and resulting micro expression image sequences are convincing. The contributions of this paper are summarized as follows.

- In this paper, we propose *AU-assisted Graph Attention Convolutional Network (AU-GACN)*, which effectively integrates AU recognition task with MER and fully exploits AUs relational information. To the best of our knowledge, this is the first work that integrates AU detection with MER.
- To alleviate the limited and unbalanced problem of existing MER datasets, a new data augmentation method is proposed for MER, namely, *AU Intensity Controllable Generative Adversarial Nets (AU-ICGAN)*, which can generate diverse data for learning-based approaches on MER task.
- The experimental results verify that our method achieves better performance for both merging the datasets into 3 categories and the original dataset categories for MER.

2 RELATED WORK

2.1 Micro-Expression Recognition

In the last decade, many works [12, 47] used hand-crafted feature extractors, e.g., Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) [46], Histogram of Oriented Optical Flow (HOOF) [52]. However, the features are usually low-level representations, e.g., intensities, gradients, without considering explicit semantic information. Recently, it has been proved that deep learning-based approaches are powerful for extracting discriminative features in many applications [13, 14, 37]. Therefore, one promising solution is to exploit deep learning model to extract spatio-temporal features for MER. For example, Kim et al. [17] used convolutional neural networks(CNN) to encode micro-expressions spatial features, and then applied the long short-term memory (LSTM) recurrent neural networks to learn temporal features of micro-expressions. To further improve the performance, Enriched Long-term Recurrent Convolutional Network (ELRCN) [16] was proposed to i) enrich the spatial dimension by stacking the optical flow image, optical strain image and gray-scale raw image, and ii) enrich the temporal dimension by stacking deep features. However, the minor difference of micro-expressions in the spatial domain is inclined to be ignored without being jointly extracted in the temporal domain. In contrast, our proposed spatio-temporal learning backbone adopts 3D ConvNets with the advantage of learning features from both motion and appearance simultaneously.

2.2 Facial AU Detection

Instead of predicting micro-expression labels directly from the input image sequence, the facial muscle movements represented by AUs can serve as hints for MER task. There are many works about AU occurrence detection [8, 22, 23, 29], which is considered as a multi-label classification problem. For instance, several works considered the relationships on various AUs and modeled AU interrelations to improve recognition accuracy [4, 21, 26]. However, most works relied on probabilistic graphical models with manually extracted features [5, 53], which limits the extension for deep learning. Given that the graph has the natural ability of handling multi-relational data [4], Liu et al. [26] proposed the first work that employed GCN to model AU relationship. The cropped AU regions by EAC-Net [23] were fed into GCN as nodes, after that the propagation of the graph was determined by the relationship of AUs. Li et al. [21] used a structured knowledge-graph for AUs and integrated a Gated GNN (GGNN) to generate enhanced AU representation. Although existing researches effectively detect AUs, which characterize facial muscle movements, AU relational information has not been exploited for the MER task. To the best of our knowledge, this is the first work that integrates AU detection with MER.

2.3 Facial Expression Data Augmentation

Different from the data for facial expression recognition, existing databases for MER contain limited amounts of data since it is challenging to trigger MEs [6, 24, 47]. Data augmentation is a common technique to enlarge the database for improving the results, e.g., cropping [45], flipping [40], rotating [15, 44]. For facial expression recognition, Yu et al. [49] modified the saturation and brightness of each image to generate extra training data. Moreover, several noise models, e.g., Gaussian noise model, pepper and salt model, were employed on original images for data augmentation [28, 50]. However, the diversity of training samples is not improved since they cannot model unseen facial expressions. Recently, model-based technologies have been successfully applied to generate diverse training samples [1, 2, 19, 25, 51]. For example, [2, 51] generated images of one subject with different expressions using cGANs [32] by training the GAN model conditioned on different facial expressions. However, it is still difficult to generate realistic images with micro-expressions due to minor changes in a short period. In contrast, [34] trained their GAN model conditioned on continuous AU intensity, allowing to generate a large range of anatomically possible facial expressions. Compared with [34], our proposed network for data augmentation is trained not only conditioned on AU intensity extracted from real-world micro expression images but also taking image structure similarity along with image sequence authenticity into consideration to generate high-quality images with MEs.

3 PROPOSED METHOD

Given a micro-expression sequence, the goal is to classify the emotions. The basic assumption is that AU labels can be vital hints for further emotion label classification. Therefore, we add AU classification as an auxiliary task. Obtained AU nodes features can be further enhanced for micro-expression classification. However, three key challenges still arise in the design of the model for micro-expression classification: i) the low intensity of spontaneous and brief facial movements are difficult to be directly extracted for an MER network; ii) the interplay between different AUs are separately modeled and thus the relationship is missing; iii) there are few micro-expression datasets with heavily imbalanced categories for micro-expression classification, which is challenging for training an end-to-end deep learning model.

To address the first and second challenges, we propose an MER network that fully utilizes AU information and ME labels, namely, AU-GACN. The architecture of the proposed AU-GACN is shown in Figure 1. Firstly, we apply a lightweight 3D ConvNet backbone network, STPNet, to extract spatio-temporal features for MEs. Secondly, the extracted features for AUs are further used as the node features in AU relation graph. We propose to leverage GCNs with self-attention graph pooling in AU graph relation learning for enhancing the extracted features for ME classification. Since well-defined muscle information (i.e., AU) can be vital hints for micro-expression recognition compared to using discrete emotion classes directly [31], the AU multi-label loss and ME classification loss are combined for training AU-GACN. Moreover, to address the third challenge, we propose a data augmentation module, AU-ICGAN, for greatly enriching the limited training data.

3.1 Spatio-temporal ME Representation Learning

Facial dynamics can be explicitly analyzed by detecting their constituent temporal segments (i.e., onset, apex and offset), where MEs occur only in a short period. Therefore, the abilities to identify the period of micro-expression and extract discriminative features are vital for classifying the micro-expressions. Previous work shows that performing 3D spatio-temporal convolutions is effective for capturing more representative features [42]. Nevertheless, applying deep 3D CNN from scratch significantly increases the computational cost and memory demand.

Therefore, we adopt the lightweight 3D ConvNet backbone based on Pseudo-3D [35], STPNet, which factorizes the 3D convolutional filters (3×3×3) into 2D spatial convolutions (1×3×3) and 1D temporal convolutions (3×1×1), for spatio-temporal feature extraction. Since the low intensity of spontaneous and subtle facial movements are difficult to directly extract, we use AU labels to guide STPNet learning. Due to the space constraint, the layer-wise details of our backbone are presented in Appendix A of the supplementary materials. Notably, the last three layers are implemented for *Global Average Pooling (GAP)* with feature map size $[A, 1]$, where A is the number of AU in ME datasets (e.g., A=19 in CASME II dataset). The obtained feature maps are used to represent AU node features and fed into the AU graph building.

3.2 AU Graph Relation Learning

AU is an effective description of facial muscle movements, which can be used for ME analysis. According to statistical and facial anatomy information, different AUs show strong relationships under different facial expressions [41], e.g., happiness can be the combination of AU12 (Lip Corner Puller) and AU13 (Cheek Puffer). Thus, by leveraging AU node representations obtained from STPNet, we construct AU relation graph to explore their structured relationships. Notably, the AU relation graph is composed of a node set V and an edge set E, where each node represents a corresponding AU associated with representations obtained from STPNet, and edge relationships are gathered from the training set based on common facial expressions and facial anatomy analysis [41].

After building AU relation graph, we aim to enhance the AU node features by considering the co-occurred AUs. Therefore, we propose to leverage Graph Convolutional Network (GCN) [18] to model the label correlation. Specifically, GCN takes the node feature description $X \in \mathbb{R}^{d\times N}$ and adjacency matrix $Adj \in \mathbb{R}^{N\times N}$ as input (where N represents the number of nodes and d stands for the dimension of feature description of each AU node). The outputs of L-layer GCN are embedded nodes represented by the last hidden layer X^l. We can represent each GCN layer as,

$$X^l = \sigma(Adj \times X^{l-1} \times W^{l-1}) \tag{1}$$

where σ is the non-linear activation function and $W^{l-1} \in \mathbb{R}^{d\times d'}$ is the $(l-1)$-th weighted matrix as d and d' stand for the input and output dimension of layer l, respectively.

Similar to the pooling operation in typical CNNs, graph pooling layers can reduce the number of parameters and retain a portion of nodes of input graphs, which avoids overfitting. Inspired by [20], after concatenating three sequential *GCN layers*, the AU node

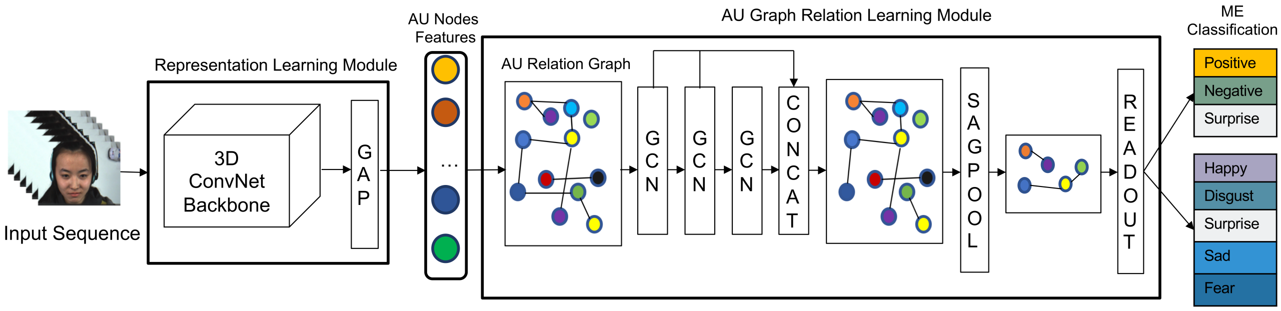

Figure 1: The overall architecture of our proposed AU-GACN for MER. The input sequence is fed into two main modules: *representation learning module* and *AU graph relation learning module*. It is fed to the representation learning module for spatio-temporal feature extraction first. *GAP* is used to output AU node features from the backbone features, which is used for the AU graph building in the second part. After the graph constructed by AU node passes through three *GCN* layers, only important nodes are left through self-attention graph pooling (*SAGPOOL* layer). The prediction of micro-expression categories is processed through *readout* layer. In this work, we consider both three emotion labels (i.e., Positive, Negative, Surprise) and the original emotion labels of the ME datasets.

features are aggregated to the *self-attention graph pooling layer* (SAGPOOL in Figure 1). To select useful nodes, we retain a portion $p \in (0,1]$ of nodes of input graph based on the *top-k* strategy [3].

Specifically, SAGPOOL first calculates self-attention scores $Z \in \mathbb{R}^{N\times 1}$ from the node embeddings of the last layer of GCN. Afterward, SAGPOOL selects the top-k nodes as follows.

$$idx = top - rank(Z, [kN]), Z_{mask} = Z_{idx}, \tag{2}$$

where idx is the index set of the selected nodes. Finally, SAGPOOL calculates the pooled feature map as follows.

$$X_{out} = X_{idx,:} \odot Z_{mask}, Adj_{out} = Adj_{idx,idx} \tag{3}$$

where $X_{idx,:}$ is the row-wise (i.e. node-wise) indexed feature matrix, $\odot$ is the broadcasted elementwise product, and $Adj_{idx,idx}$ is the row-wise and col-wise indexed adjacency matrix. X_{out} and Adj_{out} are the new feature matrix and the corresponding adjacency matrix, respectively. Finally, the *readout layer* (including global average pooling and global max pooling) aggregates node features to construct a fixed size representation, both AU graph features and topology contribute to ME classification.

3.3 AU-ME Supervised Loss

Since ME data is limited, it is challenging to differentiate emotion labels among input frames. AU features learning can be an intermediate step for ME classification. Therefore, we treat AU classification as an auxiliary task for MER and propose a new loss, namely AU-ME supervised loss, to combine the AU loss and ME loss. To the best of our knowledge, this is the first MER work that combines AU loss and ME loss to boost recognition accuracy.

Since multiple AUs can co-occur simultaneously, AU classification is a typical multi-label classification task. We compute the BCE Loss L_{AU} for AU nodes which are generated from the STPNet:

$$L_{AU} = -w[y \cdot log\sigma(x) + (1-y) \cdot log(1-\sigma(x))] \tag{4}$$

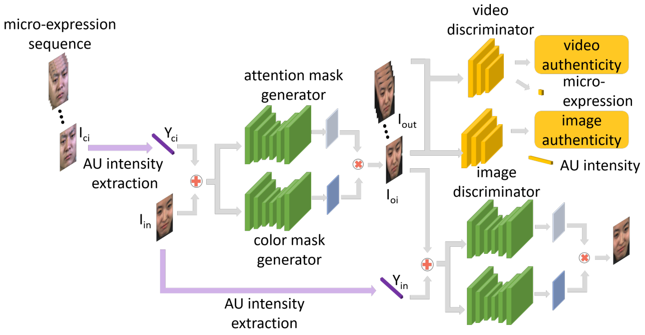

Figure 2: Overview of AU-ICGAN for data augmentation.

where w is a learnable parameter, x is the input and y is the target class. Then we classify N AUs (N is the number of AU) into one micro-expression category. ME loss L_{ME} is calculated by typical cross-entropy loss. The total loss combines AU loss and ME loss.

$$L_{total} = L_{AU} + \lambda \cdot L_{ME} \tag{5}$$

where λ is a balanced parameter between AU loss and ME loss.

3.4 AU-ICGAN

Deep models for MER, including the proposed AU-GACN, requires a large and balanced training database to prevent the model from overfitting. Since existing training samples are limited and biased, we aim to generate synthetic data based on current manually-annotated ME datasets for data augmentation. As AUs represent the facial muscle movements that correspond to the displayed emotions, the level of expressive AU can be represented as a continuous value. Therefore, we propose a novel architecture, namely, AU Intensity

Controllable Generative Adversarial Network (AU-ICGAN), to synthesize ME training samples using the intensity of different AU combinations.

Specifically, given a face RGB image $I_{in} \in^{H\times W\times 3}$ and a sequence of condition sets Y_c as input, where $Y_c \in^{K\times N}$ denotes the level of expressive AUs for K frames[1], we aim to learn a generative model that generates a sequence of images $I_{out} \in^{K\times H\times W\times 3}= [I_{o_1}, I_{o_2}, \cdots, I_{o_N}]$, of which the face identity is the same as the original input face I_{in}, and maximizes the likelihood between I_{out} and the set of action-unit target Y_c. In other words, for generator, $G(I_{in}|Y_c)$ is trained to transform the original face from single image I_{in} into an image sequence I_{out} with spontaneous change of micro expression of desired condition Y_c.

Inspired by [34], our proposed network for data augmentation contains two generators for generating: i) the attention mask that ensures the generator focuses only on those regions responsible of synthesizing the desired expression and preserve the rest part of the image, and ii) the RGB color transformation of the entire image that can be integrated with the attention mask for generating the face image with the target MEs. Moreover, for the discriminator, WGAN-GP [11] based critic is adopted to evaluate the quality of generated image sequences.

Different from [34], the proposed discriminators contain two branches: i) image quality and ii) video quality. The branch of evaluating image quality uses a similar structure as that of PatchGAN network to classify whether the given image is real or fake patch-by-patch, and a regression head to regress the resulting AU intensity and target AU intensity. On the other hand, the branch of evaluating video quality utilizes a 3DConvNet architecture to distinguish the synthetic sequences from the real ones and also recognizes the micro expression to guarantee the quality of the synthetic dataset. The dual-discriminator design can thus supervise the generator to synthesize the reasonable ME sequence for the recognition task. Figure 2 illustrates the overall architecture of AU-ICGAN.

To train the proposed AU-ICGAN, seven losses are designed to guide the optimization, i.e., adversarial loss L_{adv}, consistency loss L_c, attention loss L_{att}, AU intensity loss L_{AI}, SSIM loss L_{SSIM}, ME loss L_{ME} and sequence authenticity loss L_{sa}. Specifically, the first four terms are similar to [34]. The adversarial loss term L_{adv} is originally proposed by WGAN-GP [11] to optimize the distribution of synthesized ME images toward the distribution of real-world ME images. The consistency loss term L_c ensures the generator learns the correct mapping and preserves the person's texture without using paired-data. Moreover, the attention loss term L_{att} helps the model smoothen the learned attention mask and prevent the mask from saturation. Finally, the AU intensity loss L_{AI} determines if the synthesized images are exactly with the desired intensity of each AU respectively. The combination of above four loss terms is for single image generation.

However, for MER, the facial muscle movements are usually subtle and the trained recognition network is thus sensitive to the details of local patterns. Since the original consistency loss tends to lose such information, we further introduce the loss of Structural SIMilarity (SSIM) index between the generated image and groundtruth image, denoted as L_{SSIM}. SSIM is an indicator that computes the similarity between two images and thus can evaluate the quality of the synthesized images in terms of local patterns. Similar to consistency loss, we encourage the generator to preserve as many details as possible to synthesize high-quality images.

$$L_{SSIM} = \mathbb{E}_{I_{in}\sim\mathbb{P}_{in}}\left[\sum_{i=1}^{L}(1 - SSIM(G(I_{in}|Y_{c_i}), I_{in}))\right]. \tag{6}$$

In addition to single image quality, to ensure the generated images are arranged in a reasonable ME sequence, we utilize several constraints for the legitimacy of the generated sequences. We first introduce the micro-expression loss, denoted as L_{ME}, to ensure that the synthesized image sequences satisfy the desired micro expression. Note that since we generate one image at a time, the generator with the AU loss L_{AU} is only conditioned on the AU intensity of each image while the resulting ME of the whole sequence remains unconditioned. Therefore, with ME loss, the generator can restrict the synthesized results to meet the desired AU and ME in both image and sequence level. L_{ME} can be defined as:

$$L_{ME} = \mathbb{E}_{I_{in}\sim\mathbb{P}_{in}}\left[\| D_{ME}(G(I_{in}|Y_c)), ME_c \|_2^2\right] \tag{7}$$

where ME_c is the ME label of the image sequence with AU intensity Y_c. Last but not least, the sequence authenticity loss L_{sa} is employed to examine if the synthesized sequence is reasonable. During the process of generating a sequence image by image, the combined fake sequence may be inconsistent since the image adversarial loss does not take the full sequence into consideration. To maintain the consistency of a whole sequence, we use L_{sa} to shorten the domain distance between real-world ME data samples and synthetic ones:

$$\begin{aligned} L_{sa} = &\mathbb{E}_{I_{in}\sim\mathbb{P}_{in}}\left[D_S(G(I_{in}|Y_c))\right] - \mathbb{E}_{I_{in}\sim\mathbb{P}_{in}}\left[D_S(I_c)\right] \\ &+\lambda_{gp}\mathbb{E}_{\tilde{I}\sim\mathbb{P}_{\tilde{I}}}\left[(\| \nabla_{\tilde{I}} D_I(\tilde{I}) \|_2 - 1)^2\right] \end{aligned} \tag{8}$$

The overall objective can be expressed as:

$$\begin{aligned} L_{total} = &\lambda_1 L_{adv} + \lambda_2 L_c + \lambda_3 L_{att} + \lambda_4 L_{AI} \\ &+ \lambda_5 L_{SSIM} + \lambda_6 L_{ME} + \lambda_7 L_{sa}, \end{aligned} \tag{9}$$

where all theλ are the hyper-parameter correspond to the importance of each term during training. The minimax problem of GAN training is then defined by:

$$G^* = \arg\min_G \max_D L_{total} \tag{10}$$

In summary, the proposed AU-ICGAN is able to generate a large amount of micro expression image sequences given the proper AU intensity of ME and images of human faces. In addition, the proposed AU-ICGAN also alleviates the unbalanced issue, i.e., some MEs are rarely to be collected. Therefore, the number of each emotion category is carefully tailored to arrange the proportion of different emotions in the proposed synthetic dataset to facilitate the training of AU-GACN.

[1] Y_c can be generated from any micro expression image sequences with annotations, where every subtle facial movement in K images is encoded by N AU intensity.

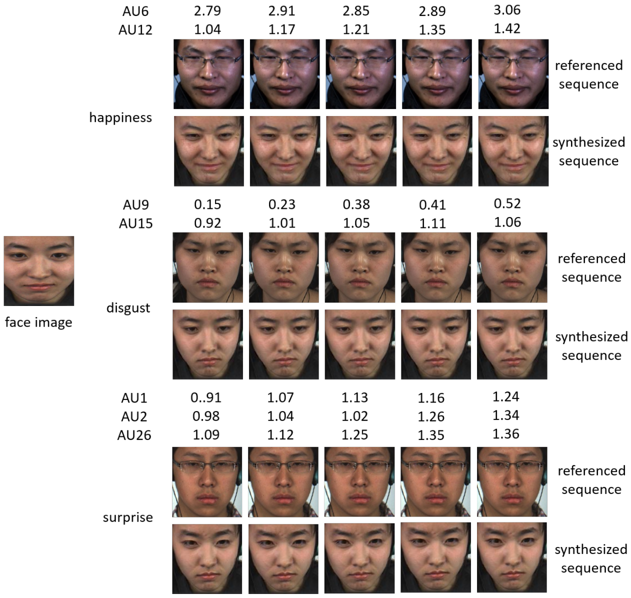

Figure 3: Synthetic dataset generated by AU-ICGAN from CASME II dataset with corresponding intensities of dominant AUs.

4 EXPERIMENTS

In this section, the experimental details will be introduced, including the datasets introduction, experimental setup, baselines, experimental results and ablation study [2].

4.1 Datasets

Three spontaneous ME datasets are utilized to evaluate our performances in the experiments. Chinese Academy of Sciences Micro-Expression II (CASME II) [47], contains 255 ME sequences from 26 subjects of 7 categories, which are happiness, disgust, surprise, fear, sadness, repression and others. The second dataset, Spontaneous Actions and Micro-Movements (SAMM) [6] has 159 ME sequences from 29 subjects. Different from CASME II, instead of repression, angry and contempt are obtained in SAMM, resulting in total of 8 categories. The Spontaneous Micro-expression Corpus (SMIC) [24], is composed of 164 ME sequences from 16 subjects. Sequences in SMIC are divided into 3 categories of positive, negative and surprise. Notably, SMIC is not labeled with AU, and the index of apex frames are not provided. Therefore, we only use SMIC as the testing dataset for cross-database evaluation in our experiments.

4.2 Experimental setup

We conduct experiments on CASME II and SAMM datasets for single database evaluation and then introduce SMIC for cross-database evaluation to test the robustness of our approach. Notably, aside from the original emotion classes of 7 and 8 in CASME II and SAMM, we further divide the data into three categories: positive, negative and surprise following the rules provided in MEGC2018 [38]. The experiments of single database are performed on CASME II and SAMM with their original emotion labels along with the 3 newly generated class labels respectively. For cross-database evaluation, since our proposed approach requires training with AU instance information, we take training samples from CASME II and SAMM respectively and acquire the samples from SMIC for testing. All the experiments are conducted on a workstation running Ubuntu 18.04 with 3.2GHz CPU, 64GB RAM, and NVIDIA GeForce GTX 2080 Ti GPU. We use PyTorch for the network implementation.

Data Augmentation. To train the proposed AU-ICGAN, we leverage an auxiliary macro emotion dataset with a large number of human face images to jointly train the GAN-based network with a micro expression dataset. Specifically, AU-ICGAN is trained on the affectnet database of facial expression [33], along with the *training* data of CASME II and SAMM. In existing datasets, as some types of MEs are difficult to collect, such as fear and sadness, existing training samples suffer from a serious problem of unequal emotion category distribution. Therefore, equipped with the well-trained AU-ICGAN, we carefully generate the synthesized data to provide 300 to 400 samples for each category depending on the distribution in the real-world datasets. In addition, the numbers of synthesized samples are specifically tailored for each subject by making different subjects to have all types of MEs. The resulting numbers of image sequences are similar.

As a result, AU-ICGAN generates 2432 and 2161 training samples from CASME II and SAMM datasets, respectively, which are nearly ten times greater than that of each original ME database. Moreover, our synthetic dataset contains more diverse training data than the real-world databases. Figure 3 shows a visual inspection of synthetic results along with the target AU intensity generated from CASME II dataset.[3] The results show that most of the generated images have no significant artifacts and are nearly indistinguishable from real-world images. Table 1 summarizes the expression levels and the number of image sequences presented in real-world databases and our synthetic dataset.

Evaluation Methods. In this paper, we adopt Accuracy (ACC) and F1-score as evaluation metrics. Two main-stream validation protocols, *Leave One Subject Out (LOSO)* and *Leave One Video Out (LOVO)* are used here. For LOSO validation, the model leaves out all samples of one single subject for model performance evaluation, and all other data are used as training data. The overall performance is then evaluated by calculating the average of all results of different subjects. Similar to LOSO, LOVO validation protocol requires the model to spare the frames from one video for validation purpose while all other data are sampled for training. Both LOSO and LOVO validation are leveraged to verify the performance as well as the robustness of the model under different situations. Note that the model is initialized in both LOSO and LOVO protocols when repeating the training process for different validation objects.

4.3 Baselines

We compare our performance with several state-of-the-art methods including MicroExpSTCNN [36], ELRCN [16], CapsuleNet [43] and MER-GCN [27]. To perform the classification of different classes,

[2] Due to the space constraint, more experimental visulizations are presented in: https://github.com/AU-GACN/AU-assisted-Graph-Attention-Convolutional-Network-forMicro-Expression-Recognition

[3] Due to the space constraint, the results of SAMM is presented in Appendix A.2 in the supplementary materials.

Dataset	Happiness	Disgust	Surprise	Fear	Sadness	Anger	repression	Contempt	Others	Total
CASME II	32	63	28	2	4	-	27	-	99	255
synthetic CASME II	384	353	388	294	307	-	389	-	317	2432
SAMM	26	9	15	8	6	57	-	12	26	159
synthetic SAMM	264	281	275	282	284	233	-	278	264	2161

Table 1: A summary of the amount of training samples in real-world and the proposed synthetic dataset.

Method	CASME II (3 categories)		CASME II		SAMM (3 categories)		SAMM	
	ACC	F1-score	ACC	F1-score	ACC	F1-score	ACC	F1-score
STCNN	0.610	0.253	0.368	0.132	0.676	0.271	0.289	0.094
ELRCN	0.623	0.342	0.443	0.325	0.691	0.272	0.358	0.066
CapsuleNet	0.568	0.347	0.331	0.194	0.575	0.392	0.259	0.111
MER-GCN	0.544	0.303	0.405	0.163	0.534	0.283	0.294	0.010
Ours	**0.712(+8.9%)**	**0.355(+0.8%)**	**0.561(+11.8%)**	**0.394(+6.9%)**	**0.702(+1.1%)**	**0.433(+4.1%)**	**0.523(+16.5%)**	**0.357(+24.5%)**

Table 2: The accuracy (ACC) and F1-Score of different methods under the LOSO protocol on CASME II and SAMM datasets.

Method	CASME II (3 categories)		CASME II		SAMM (3 categories)		SAMM	
	ACC	F1-score	ACC	F1-score	ACC	F1-score	ACC	F1-score
STCNN	0.607	0.398	0.407	0.184	0.694	0.424	0.384	0.149
ELRCN	0.609	0.429	0.396	0.197	0.682	0.227	0.359	0.066
CapsuleNet	0.582	0.356	0.324	0.155	0.592	0.384	0.266	0.158
MER-GCN	0.532	0.412	0.398	0.173	0.542	0.257	0.273	0.108
Ours	**0.634(+2.5%)**	**0.521(+9.2%)**	**0.519(+11.2%)**	**0.424(+22.7%)**	**0.721(+2.7%)**	**0.454(+3%)**	**0.426(+4.2%)**	**0.228(+7%)**

Table 3: The accuracy (ACC) and F1-Score of different methods under the LOVO protocol on CASME II and SAMM datasets.

all the baselines are only modified with the last layer to predict the same number of classes. All the hyperparameters of the baselines follow the setting of the original paper. For ELRCN, we use spatio-temporal architecture with only optical flow images as the input since the setting is reported to have the best performance in the paper. We re-implement all the 4 baseline network architectures for MER and apply both LOSO and LOVO validation strategy for evaluations.

4.4 Quantitative Result

Table 2 reports the results under LOSO protocol. The proposed AU-GACN outperforms other methods in all configurations and datasets. Since dividing all MEs into 3 categories reduces the recognition difficulty, the overall ACC and F1-score of CASME II and SAMM datasets in 3 categories are higher than that for the classification of 7 categories (CASME II) and 8 categories (SAMM). Moreover, the proposed AU-GACN also has more obvious outstanding performance in 7 categories and 8 categories. For CASME II, the proposed AU-GACN outperforms other baselines by 11.8% and 6.9% in terms of ACC and F1-score, respectively. For SAMM, the improvement is more obvious, i.e., 16.5% and 24.6% improvement in terms of ACC and F1-score, respectively. To complement the LOSO protocol, we also report the comparison results under the LOVO protocol in Table 3. Compared with the results of other baselines in Table 3, our proposed AU-GACN still performs better than the baselines in all configurations and datasets, especially for the 7 categories of CASME II dataset (22.7%) since our method can be good at capturing the subtle facial features and further improving the accuracy of MER.

The inferior performance of other methods may be caused by the intra-class variations of each subject as these deep models learn certain appearances from the samples of each subject. Specifically, ELRCN uses optical flow as representation features which may eliminate the intra-class information of each subject since only geometric information of subjects is reserved. Meanwhile, CapsuleNet relies on the apex frame as representative features, which affects its recognition accuracy on the MER dataset lacking reliable apex frame annotations. Moreover, in most cases, the performance of MER-GCN is worse than other methods since i) the AU node features are not learned from the input image and ii) the mapping of AUs to emotions is through a simple linear layer without considering the weights of nodes.

LOSO protocol allows us to evaluate the generalization of different methods in recognizing MEs on unseen subjects during testing. We separately take training samples from CASME II and SAMM, and acquire the samples from SMIC for testing. Table 4 shows the cross-database validation results. The proposed AU-GACN achieves better performance than other methods, which shows the ability of learning intra-class variations among different subjects. Note that our proposed method outperforms MER-GCN in terms of F1-score but performs slightly worse than MER-GCN in terms of ACC in the task of CASME II to SMIC. The main reason for this result is that the data number of merged 3 classes in CASME II(32 for positive class, 73 for negative and 25 for surprise) shows the class imbalance.

Method	CASME II -> SMIC		SAMM -> SMIC	
	ACC	F1-score	ACC	F1-score
STCNN	0.314	0.190	0.325	0.190
CapsuleNet	0.322	0.152	0.324	0.179
MER-GCN	0.367	0.272	0.361	0.178
AU-GACN	0.344	**0.319**	**0.451**	**0.309**

Table 4: The recognition accuracy (ACC) and F1-Score of different methods for cross-database evaluation.

Method	CASMM II		SAMM	
	ACC	F1-score	ACC	F1-score
STCNN	0.368	0.132	0.289	0.094
STCNN + AU-ICGAN	0.411	0.253	0.367	0.167
CapsuleNet	0.331	0.194	0.259	0.111
CapsuleNet + AU-ICGAN	0.368	0.267	0.275	0.193
MER-GCN	0.405	0.163	0.294	0.010
MER-GCN + AU-ICGAN	0.441	0.181	0.328	0.124
AU-GACN	0.492	0.273	0.489	0.310
AU-GACN + AU-ICGAN	0.561	0.394	0.523	0.357

Table 5: The recognition accuracy (ACC) and F1-Score of different methods with and without using our proposed data augmentation techniques under the LOSO protocol.

Method	CASMM II		SAMM	
	ACC	F1-score	ACC	F1-score
STCNN	0.407	0.184	0.384	0.149
STCNN + AU-ICGAN	0.403	0.273	0.383	0.159
CapsuleNet	0.324	0.155	0.266	0.158
CapsuleNet + AU-ICGAN	0.338	0.257	0.259	0.161
MER-GCN	0.398	0.173	0.273	0.108
MER-GCN + AU-ICGAN	0.439	0.186	0.344	0.118
AU-GACN	0.420	0.347	0.409	0.212
AU-GACN + AU-ICGAN	0.519	0.424	0.426	0.228

Table 6: The recognition accuracy (ACC) and F1-Score of different methods with and without using our proposed data augmentation techniques under the LOVO protocol.

Also, accuracy(ACC) is less sensitive to skewed data than F1-score, which provides a better measure of the performance of the MER classifier when dealing with imbalanced data here. This explains why our ACC is lower but F1-score is higher than MER-GCN.

4.5 Ablation Study

The impact of AU-ICGAN. To validate the effectiveness of the synthetic dataset generated by our proposed AU-ICGAN, we develop a scheme to remedy the scarcity of real-world datasets. First, we pre-train the model on the synthetic dataset. Then, the pre-trained model is further finetuned on target real-world datasets. The same scheme is applied to other baseline models for fair evaluation. Table 5 and Table 6 show the comparison results under both LOSO and LOVO protocol. We can observe that for most models, accuracy is slightly improved with the data augmentation module included, denoting that with more training samples, the recognition ability will be improved on most of the deep models. Moreover, with models pre-trained on the balanced synthetic dataset, the F1-score of all models are slightly increased, showing the recognition ability for rare emotions in real-world datasets are improved.

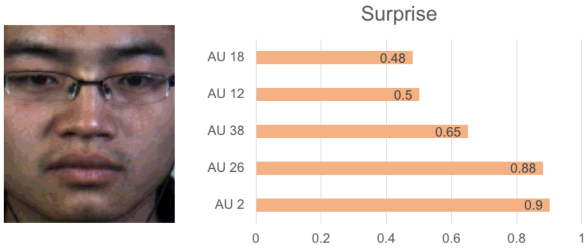

Figure 4: AU nodes representation

Explainability of AU Node Features The *representation learning module* is responsible for extracting the facial spatio-temporal features, and the *GAP* layer can map the 2D feature map extracted by the 3D ConvNet backbone into A single numbers (A is the number of AU in ME datasets). Intuitively, A numbers can represent the score of each AU category. To visually display the possibility of each AU node, Figure 4 shows an example of the top-5 dominant AU nodes proportion from CASME II dataset. The ME category of the apex frame from CASMEII is ’‘surprise”, and the decisive AU nodes are AU2 and AU26. It can be seen from the Figure 4 that AU2 and AU26 account for the highest score. The AU node features directly assign the meaning of each channel category, which increases the explainability of the network due to the nature of GAP.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose AU-assisted Graph Attention Convolutional Network (AU-GACN) for micro-expression recognition, which effectively integrates AU recognition task with MER and fully exploits AUs relational information. To alleviate the limited and unbalanced problem of existing MER datasets, a new data augmentation method, AU Intensity Controllable Generative Adversarial Nets (AU-ICGAN), is proposed, which can generate a large amount of diverse data for learning-based approaches on MER tasks. Experimental results manifest that the proposed micro-expression recognition approach outperforms the state-of-the-art methods. In the future, we plan to incorporate the AU relation graph with AU-ICGAN for improving the quality of generated ME sequences.

ACKNOWLEDGEMENTS

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2221-E-009-114-MY3, MOST-109-2634-F-009-018, MOST-109-2223-E-009-002-MY3, MOST-109-2218-E-009-025, MOST-109-2634-F-007-013, MOST-109-2218-E-002-015 and MOST-108-2218-E-002-055.

REFERENCES

[1] Iman Abbasnejad, Sridha Sridharan, Dung Nguyen, Simon Denman, Clinton Fookes, and Simon Lucey. 2017. Using Synthetic Data to Improve Facial Expression Analysis with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1609–1618.

[2] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. 2019. Identity-Free Facial Expression Recognition using conditional Generative Adversarial Network. *arXiv:1903.08051* (2019).

[3] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. 2018. Towards sparse hierarchical graph classifiers. In *Proceedings of the Workshop on Relational Representation Learning (R2L) at Conference on Neural Information Processing Systems*.

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5177–5186.

[5] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. 2017. Learning spatial and temporal cues for multi-label facial action unit detection. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 25–32.

[6] Adrian Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2016. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* PP (06 2016), 1–1.

[7] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106.

[8] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. 2019. Cross-domain au detection: Domains, learning approaches, and measures. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 1–8.

[9] Mark Frank, Malgorzata Herbasz, Kang Sinuk, A Keller, and Courtney Nolan. 2009. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *The Annual Meeting of the International Communication Association*.

[10] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3 (1978).

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.

[12] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. 2014. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *Proceedings of the IEEE International Joint Conference on Neural Networks*. 3473–3479.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. In *Proceedings of the ACM International Conference on Multimedia*. 275–283.

[15] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 2983–2991.

[16] Huai-Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao Lin. 2018. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 667–674.

[17] Dae Hoe Kim, Wissam J Baddar, and Yong Man Ro. 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *Proceedings of the ACM international conference on Multimedia*. 382–386.

[18] Thomas N Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations*.

[19] Ying-Hsiu Lai and Shang-Hong Lai. 2018. Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-View Facial Expression Recognition. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition*. 263–270.

[20] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-attention graph pooling. In *Proceedings of the International Conference on Machine Learning*.

[21] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. 2019. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8594–8601.

[22] Wei Li, Farnaz Abtahi, and Zhigang Zhu. 2017. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1841–1850.

[23] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2017. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 103–110.

[24] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–6.

[25] Feng Lin, Richang Hong, Wengang Zhou, and Houqiang Li. 2018. Facial Expression Recognition with Data Augmentation and Compact Feature Learning. In *Proceedings of the IEEE International Conference on Image Processing*. 1957–1961.

[26] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. 2020. Relation Modeling with Graph Convolutional Networks for Facial Action Unit Detection. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 489–501.

[27] Ling Lo, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. 2020. MER-GCN: Micro Expression Recognition Based on Relation Modeling with Graph Convolutional Network. *arXiv:2004.08915* (2020).

[28] Andre Lopes, Edilson Aguiar, Alberto De Souza, and Thiago Oliveira-Santos. 2016. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern Recognition* 61 (07 2016).

[29] Chen Ma, Li Chen, and Junhai Yong. 2019. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection. *Neurocomputing* 355 (2019), 35–47.

[30] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. 2019. Emotion Recognition using Multimodal Residual LSTM Network. In *Proceedings of the ACM International Conference on Multimedia*. 176–183.

[31] Walied Merghani, Adrian K Davison, and Moi Hoon Yap. 2018. A review on facial micro-expressions analysis: datasets, features and metrics. *arXiv:1805.02397* (2018).

[32] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv:1411.1784* (2014).

[33] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[34] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. In *Proceedings of the The European Conference on Computer Vision*.

[35] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[36] Sai Prasanna Teja Reddy, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. 2019. Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks*. 1–8.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in neural information processing systems*. 91–99.

[38] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. 2018. Facial Micro-Expressions Grand Challenge 2018 Summary. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 675–678.

[39] Madhumita Takalkar, Min Xu, Qiang Wu, and Zenon Chaczko. 2018. A survey: facial micro-expression recognition. *Multimedia Tools and Applications* 77, 15 (2018), 19301–19325.

[40] Madhumita A Takalkar and Min Xu. 2017. Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*. 1–7.

[41] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[43] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. 2019. CapsuleNet for micro-expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–7.

[44] Monu Verma, Santosh Kumar Vipparthi, Girdhari Singh, and Subrahmanyam Murala. 2020. LEARNet: Dynamic Imaging Network for Micro Expression Recognition. *IEEE Transactions on Image Processing* 29 (2020), 1618–1627.

[45] Chongyang Wang, Min Peng, Tao Bi, and Tong Chen. 2018. Micro-Attention for Micro-Expression Recognition. *arXiv:1811.02360* (2018).

[46] S. Wang, W. Yan, X. Li, G. Zhao, C. Zhou, X. Fu, M. Yang, and J. Tao. 2015. Micro-Expression Recognition Using Color Spaces. *IEEE Transactions on Image Processing* 24, 12 (2015), 6034–6047.

[47] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. *PLOS ONE* 9, 1 (2014), e86041.

[48] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. 2013. CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of the IEEE international conference and workshops on automatic face and gesture recognition*. 1–7.

[49] Zhenbo Yu, Qinshan Liu, and Guangcan Liu. 2017. Deeper cascaded peak-piloted network for weak expression recognition. *The Visual Computer* (09 2017).

[50] Marcus Vinicius Zavarez, Rodrigo F Berriel, and Thiago Oliveira-Santos. 2017. Cross-Database Facial Expression Recognition Based on Fine-Tuned Deep Convolutional Network. In *Proceedings of the SIBGRAPI Conference on Graphics, Patterns and Images*. 405–412.

[51] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint Pose and Expression Modeling for Facial Expression Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3359–3368.

[52] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 915–928.

[53] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. 2016. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing* 25, 8 (2016), 3931–3946.

[54] Chuanlin Zhu, Xinyun Chen, Jianxin Zhang, Zhiying Liu, Zhen Tang, Yuting Xu, Didi Zhang, and Dianzhi Liu. 2017. Comparison of ecological micro-expression recognition in patients with depression and healthy individuals. *Frontiers in behavioral neuroscience* 11 (2017), 199.

[55] Yaochen Zhu, Zhenzhong Chen, and Feng Wu. 2019. Multimodal Deep Denoise Framework for Affective Video Content Analysis. In *Proceedings of the ACM International Conference on Multimedia*. 130–138.

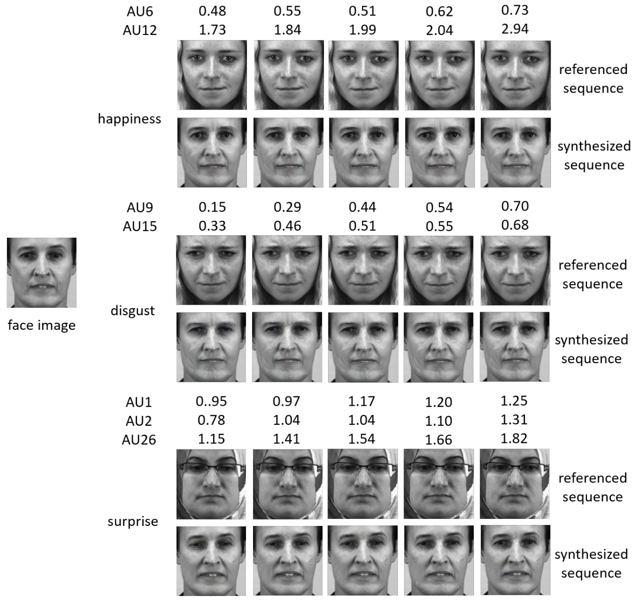
AU6 0.48 0.55 0.51 0.62 0.73
AU12 1.73 1.84 1.99 2.04 2.94
referenced sequence
happiness
synthesized sequence
AU9 0.15 0.29 0.44 0.54 0.70
AU15 0.33 0.46 0.51 0.55 0.68
referenced sequence
face image
disgust
synthesized sequence
AU1 0..95 0.97 1.17 1.20 1.25
AU2 0.78 1.04 1.04 1.10 1.31
AU26 1.15 1.41 1.54 1.66 1.82
referenced sequence
surprise
synthesized sequence

Figure 5: Synthetic dataset generated by our proposed data augmentation method from SAMM. The corresponding intensity of dominant AU for different emotions are shown as well.

A SUPPLEMENTARY MATERIALS

A.1 Implementation Details of Backbone Architecture

The main components are three bottleneck building blocks, i.e., *P3D-A*, *P3D-B* and *P3D-C*, which are residual-based variants considering performance and time efficiency. The layer-wise details of our backbone are presented in Table 7.

Table 7: Backbone Architecture, where Bottleneck-A, Bottleneck-B, Bottleneck-C indicates the three shortcut connections in P3D [35]. L and C denotes the length of sequence and the number of AU, respectively.

Layer	K:Kernel Size; S:Stride	Output Shape
`Input`	-	[3, L, 256, 256]
`Conv3d`	K=(1,7,7), S=2	[64, L, 128, 128]
`BatchNorm3d`	-	[64, L, 128, 128]
`Relu`	-	[64, L, 128, 128]
`MaxPool3d`	K=(2,3,3), S=2	[64, $L/2$, 64, 64]
`Bottleneck-A`	K=(1,3,3),(3,1,1)	[64, $L/2$, 64, 64]
`MaxPool3d`	K=(2,1,1), S=2	[64, $L/4$, 64, 64]
`Bottleneck-B`	K=(1,3,3),(3,1,1)	[64, $L/4$, 32, 32]
`MaxPool3d`	K=(2,1,1), S=2	[64, $L/8$, 32, 32]
`Bottleneck-C`	K=(1,3,3),(3,1,1)	[64, $L/8$, 16, 16]
`MaxPool3d`	K=(2,1,1), S=2	[64, $L/16$, 16, 16]
`AdaptiveAvgPool3d`	K=(1,1,1)	[64, 1, 1, 1]
`Conv2d`	K=(1,1)	[1, C, 1, 1]
`reshape`	-	[C, 1]

A.2 Synthetic Results from SAMM

Figure 5 shows the visual inspection of synthetic results from SAMM. Similar to synthetic CASME II, the generated ME images with corresponding different AU intensity have no significant artifacts and are nearly indistinguishable from real ME sequences.