Ingredients for Cooking a Large-scale Synthetic Database for Micro-expression recognition

Yuchi Liu Liang Zheng Tom Gedeon

Australian National University

Abstract. What are the computational properties of micro-expressions? How can we perform large-scale training given the small micro-expression recognition datasets due to the expensive data collection process? This paper, focusing on micro-expression analysis, undertakes early investigations into these two intriguing yet challenging questions. Our major contribution is the introduction of three types of synthetic micro-expression data based on different sources of face Action Units (AUs). First, we transfer the AUs of real-world micro-expressions to faces in the wild, yielding the "face-transferred micro-expressions". Second, we use the early frames from macro-expression videos as micro-expressions, we call this method "early-stage macro-expressions". Third, we employ the relationship between AUs and expression labels defined by human knowledge, thus creating "human-defined micro-expressions". These synthetic micro-expression images are naturally labeled and available in large amounts. We use them to train a standard convolutional neural network (CNN) and evaluate results on real-world micro-expression recognition datasets. Experimental results reveal some critical and complementary computational properties of micro-expressions: they generalize across faces, are close to early-stage macro-expressions, and can be manually defined. Moreover, the synthetic micro-expression images allow large-scale CNN training, resulting in very competitive and stable recognition performance on real-world data.

Keywords: Micro-expression learning, Micro-expression synthesising.

1 Introduction

This paper studies the problem of micro-expression (MiE) recognition. MiE recognition aims to classify a face image into predefined emotion categories, such as happiness, surprise, anger, sadness, and fear. In nature, MiEs are brief and sometimes involuntary facial expressions. This problem is highly challenging due to the short time duration and the low intensity of facial muscle movements.

In the area of MiE recognition, despite the progress in both hand-crafted and deep learning systems [2,3,4,5,6,7,8,9], there remain two important issues to be addressed. First, since it is very expensive to annotate MiE samples, existing MiE datasets only have a few hundred annotated subjects/videos [10,11,12,1]. With such limited scale data, deep networks might have a high risk of overfitting. Therefore, it would be desirable if large-scale datasets were available or created.

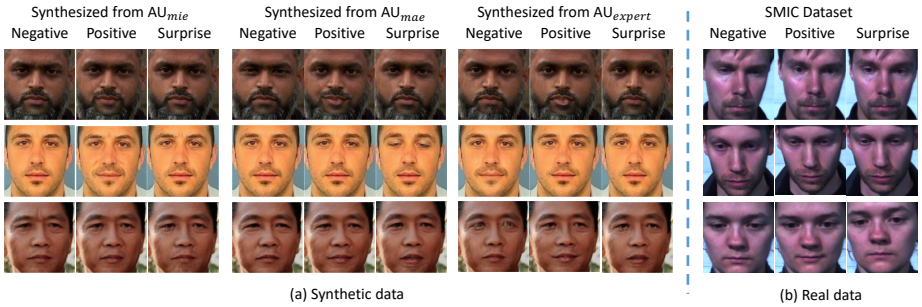


Fig. 1. Sample micro-expressions from (a) synthetic data and (b) real-world data (from the SMIC dataset [1]). In (a), we use GANimation to synthesize three types of MiEs corresponding to different Action Units (AUs), *i.e.,* AU$_{mie}$, AU$_{mae}$, AU$_{expert}$ to be described in Section 3.2. For real-world data and each type of synthetic data, the class labels of "negative", "positive" and "surprise" are shown. Note that for synthetic data, the labels (Negative, Positive, Surprise) are borrowed from the images/expert knowledge where the AUs are obtained.

Second, to our knowledge, mechanisms of MiEs are mainly studied in the psychology domain, such as the generation mechanism, psychological indications, and sociological roles. In comparison, efforts to reveal the computational properties of MiEs are limited. In the computer vision community, we might be interested in questions like *how does facial movements in MiEs relate to those in Macro-expressions (MaEs)? How are MiEs related to facial expression knowledge provided by human expert?*

Considering the above discussions, this paper studies two important issues regarding MiEs. First, we investigate how to synthesize large-scale MiE datasets that allow us to leverage the strength of deep networks. Using GANimation [13] as the default generative method for "cooking" databases, we use three types of Action Units (AUs) as effective "ingredients" to be fed into GANimation. These AUs come from MiE videos (AU$_{mie}$), MaE videos (AU$_{mae}$), or psychological studies (AU$_{expert}$). GANimation then allows us to mount these AUs on any face in the wild, and thus obtain a large number of synthetic MiEs. Examples of sythetic MiEs are shown in Fig. 1.

Second, we study the computational properties of MiEs through experimental analysis of the effectiveness of the synthetic training data on real-world test data. A rich range of insights arise. For example, we find that early-stage MaEs are computationally close to MiEs. Another finding is that expert knowledge of the AUs of MaEs can be transferred to MiEs. This suggests that MiEs can be defined by humans from a computational perspective. Furthermore, AUs computed from real-world MiEs provide MiE specific information and are irreplaceable.

In summary, the following contributions are made[1].

[1] Note that our contribution is neither the design of MiE generation networks nor the design of MiE recognition networks. In fact, we use existing generation methods and [14] as MiE classification network.

– We discover three types of AUs as effective inputs (ingradients) for MiE synthesis. Using the synthesis approach of GANimation, we demonstrate that large-scale training of MiE recognition is feasible. We report stable and competitive performance on real-world MiE recognition datasets.
– Through extensive experiment and ablation studies, interesting computational properties of MiEs are revealed. The insights can be valuable for future MiE studies.

2 Related Work

Facial micro-expression recognition. Early MiE recognition systems use handcrafted features, such as the 3DHOG descriptor [2] and the LBP-TOP descriptor [3]. Both describe dynamic texture patterns. Variants and extensions of LBP-TOP have also been proposed [15,16,17].

Recently, deep learning based solutions have been proposed [4,5,6,7,8,9]. Petal *et al.* [4] use the VGG model pretrained on ImageNet [18] and the fine-tuning technique for MiE recognition. In ELRCN [8], the network input is enriched by the concatenation of the RGB image, optical flow and *optical strain*(the derivatives of the optical flow) [19]. To reduce computation cost and information redundancy, Liong *et al.* [9] select representative frames (onset frame, apex frame, offset frame) in each micro-expression video. Optical flow is also used as in [9].

Deep learning from synthetic data. Deep learning using synthetic data has drawn recent attention. Many use graphic engines to generate the visual environment and corresponding ground truths. Richter *et al.* [20] use GTA5 to simulate training images with pixel-level semantic label maps for semantic segmentation. In [21], human prior knowledge is used to constrain the distribution of target synthetic data. Tremblay *et al.* [22] randomize the parameters of the simulator to force the model to handle large variations in object detection. Learning-based approaches [23,24] try to find the best parameter distributions of simulators so that the gap between generated content and the real-world data is minimized.

Others use the generative adversarial networks (GANs) to generate images for subsequent learning. For example, in [25], the label smoothing regularization technique is adopted for the fake images. Camstyle [26] trains camera-to-camera person appearance translation to generate new training data. These newly generated images and the real images are also combined and regularized by the label smooth regularization. CYCADA [27] introduces semantic segmentation loss on synthetic images. It uses the CycleGAN structure to keep consistent semantics.

Face manipulation. Face manipulation can be considered as an image-to-image translation task. Generic image-to-image translation methods include Conditional GAN [28], CycleGAN [29] and DiscoGAN [30]. Some recent works combine conditional GAN and cycle consistency [31,13,32,33,34,35,36]. In this manner, key aspects like identity can be preserved while attributes-of-interest are modified. StarGAN [31] implements face translation between multiple domains with a single generator and is effective in facial emotion translation. In

[13], GANimation leverages the action unit (AU) vectors of continuous entries as the condition for manipulation. When there are changes to the AU vector, the generated faces will have corresponding expression changes. SMIT [32] can also make continuous manipulation on faces but the condition vector is coded with randomness. So this method does not guarantee the semantic meaning of the resulting images. Recently, AGGAN [33] is proposed to synthesize sharper expressions with clearer details by an attention-guided discriminator. In this paper, we use GANimation [13] for MiE generation.

3 Proposed Method

Our main contribution is the *discovery of effective ingredients* that allow us to synthesize a large-scale micro-expression database. The ingredients are three types of Action Units (AUs) (Section 3.2). These AUs are used by GANimation (Section 3.1) to synthesize micro-expressions according to carefully designed synthesis strategies (Section 3.3). The synthesized images are on a large scale. These synthetic images are used to train a MiE recognition model (Section 3.4).

3.1 GANimation Revisit

This paper employs the GANimation method [13] to synthesize MiEs. It is an image-to-image translation model for face expression manipulation. A facial expression is defined as $I_{y_r} \in \mathbb{R}^{H \times W \times 3}$, where the encoded Action Units (AUs) y_r [2] form a vector with each entry normalized between 0 and 1. Given a face image dataset $\{I_{y_r}^m\}_{m=1}^M$ and a target Action Unit set $\{{y_g}^m\}_{m=1}^M$, GANimation aims to learn a single mapping function $G : (I_{y_r}, y_g) \rightarrow I_{y_g}$ such that the generated face image not only belongs to the same identity as the original image but also satisfies the target AUs. Discriminator D_I is used to distinguish whether faces are real or fake. Discriminator D_y is designed to regress face AUs, *i.e.,* the original and generated images. The loss function of GANimation is:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_I(G, D_I, I_{y_r}, y_g) + \lambda_y \mathcal{L}_y(G, D_y, I_{y_r}, y_r, y_g) + \lambda_A(\mathcal{L}_A(G, I_{y_g}, y_r) \\ + \mathcal{L}_A(G, I_{y_r}, y_g)) + \lambda_{idt}\mathcal{L}_{idt}(G, I_{y_r}, y_r, y_g)\end{aligned}, \tag{1}$$

where $\mathcal{L}_I$ is the image adversarial loss based on WGAN-GP [37], and $\mathcal{L}_y$ is the conditional expression regression loss. $\mathcal{L}_A$ is called the attention loss with an l_2-weight penalty and a total variation regularization over the generated attention mask. Finally, the identity loss $\mathcal{L}_{idt}$ enforces G to learn cycle-consistency properties. λ_y, λ_A, and λ_{idt} determine the importance of each loss. In the training phase, GANimation optimizes the mini-max game [38].

During dataset synthesis, given the single face manipulator G, a face I_{y_r} with AUs y_r, and the target AUs y_g, we use G to generate a new facial expression I_{y_g} with the same personal identity but with the target AUs. In this manner,

[2] Note that y_r here only represents the target Action Units (AUs) in GANimation. It is different with the label y in emotion classification.

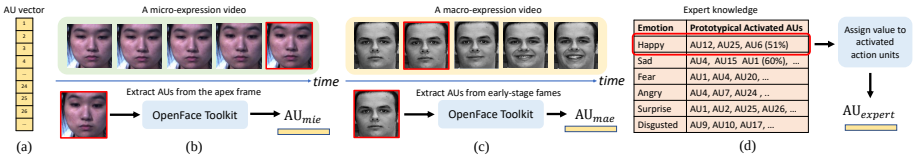


Fig. 2. Examples to compute AU_{mie}, AU_{mae} and AU_{expert}. (a) An AUs vector with fixed size. Each integer number inside it refers to one specific AU number. (Follow [13], we use 17 typical AU to make up a AU vector). (b) OpenFace Toolkit compute AU_{mie} from representative frames of MiE videos (*i.e.* the first frame or the last frame).(c) Early frames (between the first and medium frame) are inputted into OpenFace to obtain AU_{mae}. (d) We specify an emotion type from Expert mapping table and get the AU vector with activation probability. Then we assign activated AU number with intensity values and others with 0.

a large number of face images with diverse identities and designed expressions (AUs) can be generated.

3.2 Three Types of AUs as Input for GANimation

This paper discovers three types of AUs that can be used by GANimation for building an effective large-scale MiE database. The AUs are denoted as AU_{mie}, AU_{mae} and AU_{expert} to be detailed below.

First, **AU_{mie} denotes the AUs calculated from real-world MiEs.** Intuitively, the assumption is that the synthetic MiEs should share similar AU patterns with the real-world MiEs. Motivated by this, we use OpenFace [39] toolkit to extract AUs for representative frames (*i.e.* the apex frame) from real MiE video clips (See Fig. 2(b)) and obtain AU vectors for them. Then, we can use the pre-trained GANimaiton model to generate new MiEs based on the faces in EmotioNet [40] dataset and the detected AUs. Labels form the synthetic data are borrowed from the original micro-expressions.

Second, **AU_{mae} represents the AUs of early-stage frames in MaE video clips.** We assume these frames are similar to micro-expressions. Figure 2(c) shows an example of early-stage MaEs. Given a MaE video clip, the intensities of facial components movements at the beginning are subtle. However, these frames are still informative on emotion and their motion trends respect the emotion label. Under this circumstance, we also calculate AUs from mac-expression video clips by OpenFace. Then we pick AUs of earlystage frames and faces in EmotioNet to synthesize new micro-expressions. The emotion labels keep the same with the original macro-expressions.

Third, **we use AU_{expert} to denote the AUs derived from expert knowledge.** Because AUs can objectively describe the movements of facial parts, the correlation between AUs and emotion are widely investigated. For example, experts find that AU12 (Lip Corner Puller) is often activated when

emotion is "happy" and that AU4 (Brow Lowerer) is usually observed if face exhibits "angry". We consider micro-expressions as a subset of expressions with low muscle movement intensities and short time duration. Therefore, expert knowledge of the relationship between AUs and expressions should also be suitable for micro-expression. The research in [41] concludes a mapping table between emotion categories and prototypical AUs with probabilities. The given mapping $\mathcal{P}(\mathrm{AU} \mid y)$ describes the probabilities of different AU numbers to be activated for a given emotion. We choose six basic emotions (happy, sad, fear, angry, disgusted). $\{(\mathrm{AU}^m, y^m)\}_{m=1}^M$ are sampled according to $\mathcal{P}(\mathrm{AU} \mid y)$ by giving desired emotion labels set $\{y^m\}_{m=1}^M$ and assigning intensities to activated AUs (See Fig. 2(d)). yielding get AU_{expert}.

Note that, for any AU vector of the three types, we also have its emotion label. This emotion label is inherited from the source where the AUs are obtained, *i.e.,* the real-world micro-expressions, the real-world macro-expressions, and the expert knowledge (the relationship between AUs and emotion labels), respectively.

3.3 MiE Synthesis Strategies

A synthetic MiE sample $(\tilde{x}_{onset}, \tilde{x}_{apex}, y)$ is created by GANimation G given $(\mathrm{AU}_{onset}, \mathrm{AU}_{apex}, y)$, where y is the emotion label, and $\tilde{x} = G(x, \mathrm{AU})$. Below, we describe our detailed considerations from three aspects how to synthesize the database.

First, **a synthetic MiE sample is composed of an onset frame and an apex frame**. That is, we do not use the whole video clip as network input. Using only the two representative frames as classifier input is standard in the community when dealing with real-world data [9,14].

For AU_{mie}, we pick the AUs of the onset frame and the apex frame from real MiEs and denote them as $(\mathrm{AU}_{onset}, \mathrm{AU}_{apex})$. If not specified, we use the first frame and last frame of real MiEs.

For AU_{mae}, we choose the AUs from the early stage of a MaE video with n frames. Here, AU_{onset} is the AU of the first frame. To select AU_{apex}, we introduce a lower bound $\alpha \in [0, 1]$ and an upper bound $\beta \in [0, 1]$, such that,

$$i_{apex} = \lfloor n \times p \rfloor \quad p \in [\alpha, \beta]\,, \tag{2}$$

where p is random and i_{apex} indexes the frame of the selected AU_{apex}. We rand down the value $n \times p$ as $\lfloor n \times p \rfloor$.

For AU_{expert}, We first set AU_{onset} vector $\mathbf{0}$, because we assume that there is no emotion leaked in the first frame of MiEs. To determine AU_{apex}, we leverage the mapping between emotion labels and AUs made by expert knowledge [41], which provides the activation probability of each entry in an AU vector for a specific emotion class. Different from the case of AU_{mae}, we introduce $\mu \in [0, 1]$ and $\nu \in [0, 1]$ to denote the lower bound and upper bound intensity value of activated AUs in AU_{apex}. The activated AU number is assigned as:

$$j_{activated} = q \quad q \in [\mu, \nu]\,, \tag{3}$$

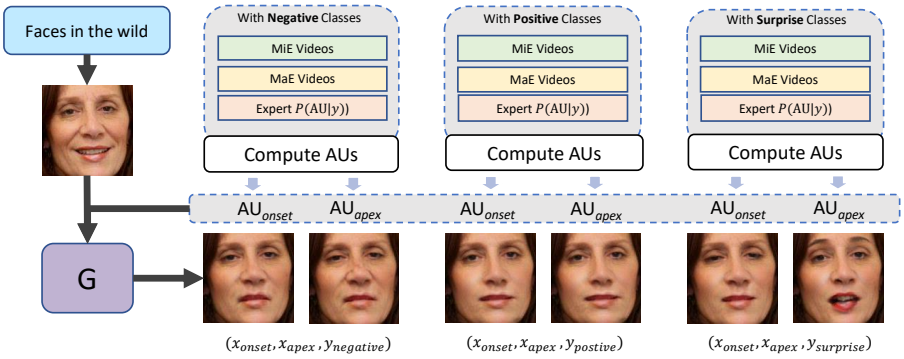


Fig. 3. MiE synthesis base on three types of AUs. AU_{onset} and AU_{apex} are computed from MiE videos, MaE videos or Expert Knowledge. They are conditions for the generator G. To synthesize MiEs (*i.e.* based on AU_{mie}), each face in the wild will generate 3 synthetic MiE samples (two frames, one label) with 3 different emotion classes (Negative, Positive, and Surprise).

where q is random and $j_{activated}$ means the intensity for the AU number j of the AU_{apex}.

Second, **the synthetic database should be balanced in terms of the number of training samples both per class and per identity.** For example, one face in the wild is used to generate 3 MiEs with indifferent emotion labels (Negative, Positive, Surprise). Each MiE consists of two frames representing AU_{onset} and AU_{spex}, respectively (see Fig. 3).

Third, **the emotion labels need to be the same among the three types of AUs.** AU_{mie} and AU_{mae} are computed from several MiE and MaE datasets where the emotion classes are different. Besides, the number of basic emotion types in AU_{expert} is six which is also different from that in AU_{mie} and AU_{mae}. To solve this. we choose a common label set with three emotion classes (positive, negative and surprise), as in the Second Micro-Expression Grand Challenge (MEGC2019 [42]).

Discussions. From a technical point of view, training with synthetic MiEs has two advantages. First, large-scale training sets are thus, for the first time, available in the MiE recognition area. The resulting large-scale training reduces the risk of over-fitting. Second, the synthetic MiEs are based on a large number of faces in the wild. This is critical for learning face-invariant expression patterns, *i.e.,* being robust against ages, races, illumination and head poses. This explains why the resulting model has good generalization ability on the real-world test set.

From a scientific point of view, the three types of synthetic micro-expressions provide us with a unique opportunity to understand micro-expressions from a computational perspective. As to be discussed in Section 4, micro-expressions can be viewed as having strong relationships with expert knowledge, magnitude-

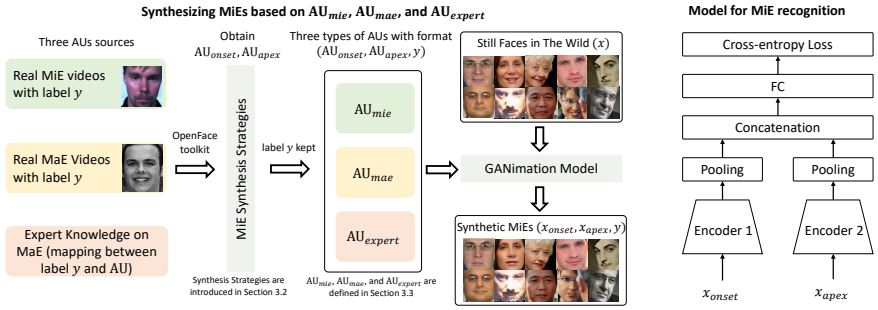


Fig. 4. System pipeline. We adopt the MiE synthesis strategies in Section 3.3 and obtain three types of AUs (AU_{mie}, AU_{mae}, and AU_{expert}) from real MiE videos, real MaE videos and expert knowledge, respectively (Section 3.2). Then, we generate MiE samples with a pretrained GANimation model, faces in the wild (various illumination, viewpoint, emotion, *etc*) and the three types of AUs (Section 3.1). Finally, we train a state-of-the-art CNN model using our synthetic data and test it on real-world MiEs (Section 3.4).

reduced macro-expressions, and natural micro-expressions generalized to other faces. We anticipate that more knowledge can be gained through data synthesis.

3.4 Deep MiE Classifier

We adopt the two-branches network [14] which won the first place in [42]. As shown in Fig. 4, the network has two branches which do not share weights. An MiE sample has an onset frame and an apex frame. The two frames are fed into the two branches of ResNet-18 [43] backbones, respectively, and their embeddings after global average pooling are concatenated. The classier has two fully connected (FC) layers with dimension 128 and 32, respectively.

4 Experiment

4.1 Dataset and AU calculation

During training, we use both MiE and MaE datasets. During testing, only the MiE dataset is used.

MiE dataset. To our knowledge, there exist three commonly used micro-expression recognition datasets, *i.e.,* CASME II [10], SAMM [11,12], and SMIC [1]. Following the protocol in MEGC2019, we merge them into a single dataset. Specifically, emotion labels in all the three datasets are mapped into a common label set including *negative*, *positive*, and *surprise*. The combined dataset has 442 samples (145 from CASME II, 133 from SAMM, and 164 form SMIC) from 68 subjects (24 from CASME II, 28 from SAMM, and 16 from SMIC). We call this *composite real-world dataset* in this paper.

Table 1. Training data notations. There are two broad categories: real data and synthetic data. Each category has different combinations of the data sources. We use 3-fold cross-subject for settings where MiEs (or their AUs) are used in training.

Data Type	Training data	Notation	3-Fold
Real	MiEs	R_1	✓
	MiEs + early-stage MaEs	R_2	✓
Synthetic	AU_{mie}	S_1	✓
	AU_{mae}	S_2	✗
	AU_{expert}	S_3	✗
	AU_{mie} + AU_{mae}	S_{12}	✓
	AU_{mie} + AU_{expert}	S_{13}	✓
	AU_{mae} + AU_{expert}	S_{23}	✗
	AU_{mie} + AU_{mae} + AU_{expert}	S_{123}	✓

MaE dataset. We use the CK+ dataset [44], a mainstream MaE dataset containing 327 high-quality video samples. We also map its labels to three classes, *i.e.,* negative, positive, and surprise.

Source of AUs. We defined three types of AUs in Section 3.2, *i.e.,* AU_{mie}, AU_{mae}, and AU_{expert}. AU_{mie} and AU_{mae} are calculated from the composite MiE dataset and the CK+ dataset, respectively. AU computation software OpenFace [39] is used. AU_{expert} comes from the research work in [41] where the correlation between emotions categories and AUs is studied.

4.2 Experiment Settings

Evaluation metric. Following MEGC2019, two metrics are used: Unweighted F1-score (UF1) and Unweighted Average Recall (UAR). They are used for balanced judgement considering label imbalance in the composite dataset. UF1 is also called macro-averaged F1-score. It is the unweighted average F1-score for all classes. Similarly, UAR is the unweighted mean of UAR for all classes.

Training GANimation. We follow [13] for GANimation training including procedure and hyper-parameters. A subset of 200,000 images from the EmotioNet dataset [40] are used for training. Their AUs are computed by OpenFace. We use the model after 25 epochs as MiE generator.

Training CNN MiE classifier. We adopt the parameter settings in [14]. Input images are resized to 256×256. Two data augmentation techniques including 224×224 random cropping and 10 degree random rotation are used. The backbone is ResNet-18 and the dimension of the concatenated feature is 2,048. Dropout rate in FC layers is set to 0.5. When the model is directly trained on the composite real-world dataset, learning rate is initialized to 1×10^{-4} and linearly decays to 0 in the last 40 epochs (80 epochs in total). When training on synthetic data, we have much more training samples. So, we set learning rate to 1×10^{-5} and linearly decay it to 0 in the last 10 epochs (30 epochs in total). Batch size is 32.

Table 2. MiE recognition performance comparison for different training data R_1 and R_2 by using the standard MiE recognition framework [14] with different CNN backbones [45,46,43].

Dataset	MobileFaceNet [45]		VGG-16 [46]		ResNet-18 [43]	
	UF1	UAR	UF1	UAR	UF1	UAR
R_1	41.22	43.17	42.31	43.84	47.96	47.87
R_2	44.44 ↑	47.82 ↑	46.42 ↑	47.89 ↑	50.34 ↑	51.17 ↑

Training and testing protocols. We train the MiE recognition model in [14] on the real data and/or our synthetic data. There are two types of real-world data: natural MiEs and early-stage MaEs. For synthetic data, three types of data exist (see Section 3.2). In total we have 9 combinations for training data, shown in Table 1.

Whenever real-world MiEs (and their AUs) are used in training, we use the 3-fold cross-subject protocol. All the subjects are evenly and randomly split into 3 sets. Existing MiE recognition works usually conduct leave-one-subject-out or 10-fold cross-validation. Considering the training cost in our large scale synthetic dataset, we use the 3-fold protocol as an alternative. In every fold, 2 sets are used for training and the other set is used for testing. Results are averaged from the 3 folds. In S_2, S_3, and S_{23}, because real-world MiEs (or their AUs) are not used in training, we directly train on the synthetic dataset and test on the composite real-world dataset. We repeat every experiment for 5 times and report the mean scores.

4.3 Preliminary Evaluation

Comparisons of training sets R_1 and R_2 composed of real-world data. The first training set is composed of real MiEs, and the second is a combination of real MiEs and early-stage MaEs (Table 2). To compare them, we select the early frames of MaE videos from the CK+ dataset by setting $\alpha = 0.3$ and $\beta = 0.5$ (Eq. 2). Results are shown by the blue and red lines in Fig. 6. We clearly observe that R_2 (blue line) yields higher F1 score (+2.38%) and UAR (+3.30%) than R_1 (red line). As show in Table 2, Such an improvement trend is consistent when we try different CNN backbones to extract features. Therefore, we validate that MaEs, when having a reduced magnitude and looking similar to micro-expression, benefit MiE recognition. This also suggests that data synthesis using AUs calculated from early-stage MaEs can be effective (to be verified later).

Impact of parameters for computing AU_{mae} and AU_{expert}. Equation 2 involves two parameters α and β, determining where to select AUs_{apex} from an MaE video. We choose two parameter settings: $\alpha = 0.1$, $\beta = 0.3$ and $\alpha = 0.3$, $\beta = 0.5$. We randomly select 5,000 faces from EmotioNet for training set synthesis (S_2). Each face produces 3 different MiEs, so we have 15,000 MiE samples.

By training on S_2 and testing on the composite real-world dataset, we report results in Fig. 5 (A). When $\alpha = 0.3$, $\beta = 0.5$, both two metrics are higher than

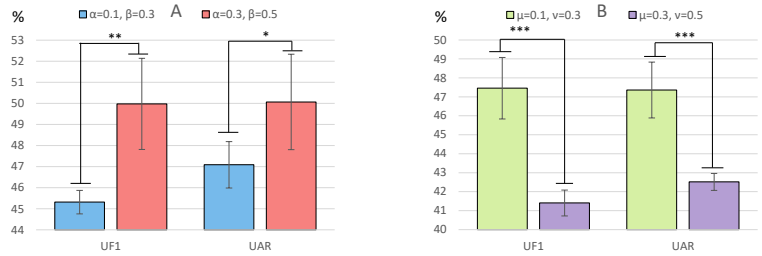


Fig. 5. Impact of parameters for computing AU_{mae} and AU_{expert}. UF1 and UAR are reported. **A:** we train on S_2. Two settings are tested: $\alpha = 0.1$, $\beta = 0.3$ and $\alpha = 0.3$, $\beta = 0.5$. **B:** we train on S_3. Also two groups of hyper-parameters are investigated: $\mu = 0.1$, $\nu = 0.3$ and $\mu = 0.3$, $\nu = 0.5$, *resp.* "n.s." means the difference is not statistically significant (*i.e.*, p-value > 0.05). $*$ denotes statistically significant (*i.e.*, $0.01 < p$-value < 0.05). $**$ and $***$ mean statistically very significant (*i.e.*, $0.001 < p$-value < 0.01) and statistically extremely significant (*i.e.*, p-value < 0.001), *resp.* In the following experiment, we adopt $\alpha = 0.3$, $\beta = 0.5$, $\mu = 0.1$, $\nu = 0.3$.

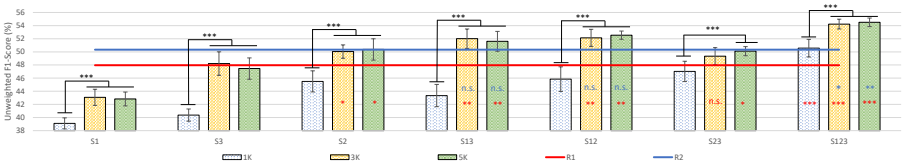
$\alpha = 0.1$, $\beta = 0.3$ with statistical significance. This indicates that frames within the time range of $30\% \sim 50\%$ for MaEs are more similar to MiEs than those within $10\% \sim 30\%$.

Similarly, we analyze the impact of μ and ν for AU computing from expert knowledge. We set $\mu = 0.1$, $\nu = 0.3$ or $\mu = 0.3$, $\nu = 0.5$. Similarly, we use the synthetic training set S_3 and the same test set. The average results of the two metrics are shown in Fig. 5 (B). We clearly observe the statistically significant drop on UF1 and UAR if we change the intensity range of AUs from $[0.1, 0.3]$ to $[0.3, 0.5]$. It suggests that the facial movement magnitude of MiEs is likely to count $10\% \sim 30\%$ of that for MaEs.

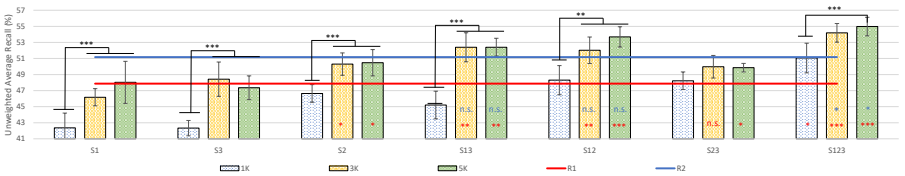
4.4 Effectiveness of the Synthetic Database

Comparison of databases synthesized by a single type of AU. In the first three columns of Fig. 6(a) and Fig. 6(b), we compare the F1 and UAR scores obtained by each of the three types of AUs, *i.e.,* AU_{mie}, AU_{expert} and AU_{mae}.

The most prominent observation is that AU_{mae} produces higher F1 and UAR scores than the other types of AUs. If we use 5,000 identities for synthesis, AU_{mae} can even yield competitive accuracy with the baseline approach: training with the combination of real MiEs and early-stage MaEs. The superiority of AU_{mae} comes from two major aspects. First, compared to AU_{mie}, AU_{mae} has a higher diversity. In the AUs sampling process of AU_{mae}, we can randomly select frames using parameters α and β. Nevertheless, in AU sampling with AU_{mie}, apex frames are fixed for given real MiEs. Second, compared with AU_{expert}, AU_{mae} is calculated from real expressions. In comparison, AU_{expert} is generated by rules and some AUs may not normally exist in the real world. In this case, AUs based on AU_{expert} are noisy, thus compromising the subsequent training process.



(a) UF1 score (%) comparison



(b) UAR score (%) comparison

Fig. 6. Performance comparison between training with real-world data (lines) and training with synthetic data (bars, various combinations). **Top:** UF1 score. **Bottom:** UAR score. Notations "n.s.", *, **, and *** have the same meaning as in Fig. 5. The blue and red lines are UF1 scores obtained by training with R2 (real-world MiE + real-world MaE) and R1 (real-world MiE), respectively. Note that $*$ and "n.s." in blue or red inside bars denote statistical significance of comparisons between the corresponding bars (training with synthetic data) and the red or blue line (training with real data). When three types of AUs are used to "cook" the synthetic database, both UF1 and UAR scores exceed those obtained by training with real data with a very high confidence.

Synthesis with multiple types of AUs. We compare various combinations of the three types of AUs and the corresponding training sets are denoted by S_{12}, S_{13}, S_{23}, S_{123}. MiE recognition of these synthetic training sets are summarized in the last four columns of Fig. 6(a) and Fig. 6(b). Three major observations are made.

First, when 3,000 or 5,000 face identities are used, synthetic databases with various combinations of AUs are competitive or superior to real ones like R_1 and R_2. For example, when measured by F1 score, S_{12} achieves 4.66% and 2.05% improvement over R_1 and R_2, respectively, with statistical significance.

Second, the three types of AUs are complementary to each other when synthesizing training sets. For example, S_{12}, synthesized from AU_{mie}, AU_{mae}, and 5,000 identities, is superior to both S_1 and S_2 by 9.31% and 5.09%. Further, among the four possible combinations, database S_{123} produces the highest recognition accuracy: 54.51% and 54.97% on UF1 and UAR. Its advantage over the real data datasets R_1 and R_2 is more significant than that over S_{12}, S_{13} and S_{23}.

Third, each type of AU is found indispensable in our system. When we remove one type of AU from S_{123}, both UF1 and UAR results decrease. For example, when AU_{mie} is absent, the resulting database S_{23} (5,000 face identities) produces an UF1 score that is lower than S_{123} by 4.4%. Similar observations can be made from the other two types of AUs.

Table 3. The **top** part of the table shows comparisons of different pre-training and fine-tuning strategies. ImageNet pre-training is always used. Results are reported using both two metrics. When S_{123} is used for pre-training, significant improvement is observed. In the **bottom** part, we re-implement two existing expression recognition methods using the same setting.

Pretraining		**Fine-tuning**		Apex [47]		Enrich [8]		Two-Branches [14]	
ImageNet	S_{123}	R_1	R_2	UF1	UAR	UF1	UAR	UF1	UAR
✓		✓		51.31	51.39	46.11	46.30	47.89	48.21
✓			✓	50.16	51.51	51.61	51.28	50.50	51.48
✓	✓			52.00	51.94	55.26	54.04	54.51	54.97
✓	✓	✓		56.80	**57.36**	54.94	55.46	58.17	**59.10**
✓	✓		✓	**58.12**	55.87	**57.73**	**57.71**	**58.81**	58.71

Data Augmentation Methods	**UF1**	**UAR**
FaceNet2ExpNet [48]	43.04	43.71
Inconsistent Annotation [49]	54.08	53.64

The above discussions indicate 1) the competitiveness of synthetic databases against real-world ones, 2) the complementary nature of different types of AUs, and 3) the necessity of including all the three types of AUs in our system.

The influence of the size of synthetic training sets. We synthesize training sets of different sizes by setting the number of face identities to 1,000, 3,000 and 5,000. From Fig. 6(a) and Fig. 6(b), if only 1,000 identities are used in data synthesis, the performance of most synthetic training sets are lower than R_1. increasing the number of identities from 1,000 to 3,000 can increase UF1 and UAR significantly. However, when we use more than 3,000 identities, model accuracy becomes saturated consistently across different synthetic training sets. Specifically, according to the two sample T-test, there is no statistically significant performance gap between 3,000 and 5,000.

Effectiveness when being used for network pre-training. We then discuss the effectiveness of the synthetic dataset S_{123} (generated from 5,000 identities) for network pretraining. In our experiment, ImageNet pretraining is always employed. After ImageNet pre-training and an optional S_{123} pre-training, the network is fine-tuned by R_1 or R_2. Besides the Two-Branches [14] framework we used in above study, we also validate the effectiveness of pre-training on other two MiE recognition frameworks which only uses the apex frame (Apex [47]) or enrich the input by concatenating frames before feature extraction (Enrich [8]). Results are shown in Table 3. We find that pre-training with S_{123} gives significant improvement over pretraining with ImageNet only for all MiE recognition frameworks. For example, when R_1 or R_2 is used for fine-tuning, the improvement of Two-Branches in UF1 score can be as large as +10.28% and +8.31%, respectively. We achieve the best result of **UF1=58.81% and UAR=59.10%**, respectively, when R_1 is used for fine-tuning. This validates the value of S_{123} in network pre-training.

Implementing and comparing with several state-of-the-art methods. We implement two recent expression recognition methods [48,49] under the 3-fold protocol. They focus on pretraining and data augmentation, thus enabling a fair comparison with ours. Ding *et al.* [48] use face recognition as supervision to pre-train the MiE recognition network. Zeng *et al.* [49] calculate pseudo labels from multiple models to enrich and smooth datasets. From Table 3, our result without synthetic data pretraining is inferior to [49]. After pretraining using S_{123}, we exceed [49] by +4.73% and +5.46% in UF1 and UAR, respectively.

4.5 Understandings of Micro-expressions

Our experiment and analysis reveal important computational properties of MiEs.

Early-stage MaEs resemble real MiEs. To our knowledge, we are the first to employ MaEs for MiE recognition. Although the two types of facial expressions differ significantly in the magnitude of facial movement, we find the initial phase of MaEs is an effective approximation to MiEs. Because there are abundant MaE samples, databases synthesized by AU_{mae} alone are effective.

Expert knowledge is transferable to MiEs. Related to the previous observation, although the AUs annotated by experts describe MaEs, we find the magnitude reduced expert AUs are also effective in synthesizing MiEs. We therefore can infer that MiEs are a subset of normal expressions with low intensity. Moreover, by examining the complementary nature of the three types of AUs, we infer that expert knowledge adds some computational ingredients for true MiEs, which do not appear in MaEs and real MiEs but can be humanely defined.

Real MiEs provide domain specific supervision. Real MiEs from MiE datasets can provide close and domain specific information on what true MiEs should be like, and is thus indispensible. However, because real MiEs are very few, the AU_{mie} does not have large diversity, thus limiting its accuracy when being used alone.

5 Conclusion

The study of Micro-expression (MiE) recognition suffers from the lack of large-scale training data. In this paper, we discover three effective AUs, *i.e.,* AU_{mie}, AU_{mae}, and AU_{expert}, to synthesize large-scale micro-expressions datasets. They are sourced from real MiEs, early-stage MaEs, and expert knowledge, respectively, and are used together with faces in the wild to generate three types of MiEs with emotion labels. Experimental results on real-world MiE datasets show that the three ingredients are indispensable and complementary. We demonstrate that models trained on the synthetic dataset can outperform those trained on real MiEs. Moreover, when the synthetic data and real data are used for pre-training and fine-tuning, respectively, we observe further improvement. Importantly, this paper reveals some interesting computational properties of MiEs which might be beneficial for further research in this area.

References

1. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: Inducement, collection and baseline. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE (2013) 1–6 1, 2, 8
2. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. (2009) 1, 3
3. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis & Machine Intelligence (6) (2007) 915–928 1, 3
4. Patel, D., Hong, X., Zhao, G.: Selective deep features for micro-expression recognition. In: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE (2016) 2258–2263 1, 3
5. Kim, D.H., Baddar, W.J., Ro, Y.M.: Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In: Proceedings of the 24th ACM international conference on Multimedia, ACM (2016) 382–386 1, 3
6. Hao, X.l., Tian, M.: Deep belief network based on double weber local descriptor in micro-expression recognition. In: Advanced Multimedia and Ubiquitous Engineering. Springer (2017) 419–425 1, 3
7. Peng, M., Wang, C., Chen, T., Liu, G., Fu, X.: Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology **8** (2017) 1745 1, 3
8. Khor, H.Q., See, J., Phan, R.C.W., Lin, W.: Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE (2018) 667–674 1, 3, 13
9. Liong, S.T., Gan, Y., Yau, W.C., Huang, Y.C., Ken, T.L.: Off-apexnet on micro-expression recognition system. arXiv preprint arXiv:1805.08699 (2018) 1, 3, 6
10. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. PloS one **9**(1) (2014) e86041 1, 8
11. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: Samm: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing **9**(1) (2016) 116–129 1, 8
12. Davison, A., Merghani, W., Yap, M.: Objective classes for micro-facial expression recognition. Journal of Imaging **4**(10) (2018) 119 1, 8
13. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 818–833 2, 3, 4, 5, 9
14. Liu, Y., Du, H., Liang, Z., Gedeon, T.: A neural micro-expression recognizer. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE (2019) 2, 6, 8, 9, 10, 13
15. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In: Asian conference on computer vision, Springer (2014) 525–537 3
16. Huang, X., Wang, S.J., Zhao, G., Piteikainen, M.: Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: Proceedings of the IEEE international conference on computer vision workshops. (2015) 1–9 3

17. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing **175** (2016) 564–578 3
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255 3
19. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro-and micro-expression spotting in long videos using spatio-temporal strain. In: Face and Gesture 2011, IEEE (2011) 51–56 3
20. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision, Springer (2016) 102–118 3
21. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision 1–20 3
22. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018) 969–977 3
23. Kar, A., Prakash, A., Liu, M.Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., Fidler, S.: Meta-sim: Learning to generate synthetic datasets. arXiv preprint arXiv:1904.11621 (2019) 3
24. Ruiz, N., Schulter, S., Chandraker, M.: Learning to simulate. arXiv preprint arXiv:1810.02513 (2018) 3
25. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3754–3762 3
26. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5157–5166 3
27. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017) 3
28. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1125–1134 3
29. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232 3
30. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 1857–1865 3
31. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8789–8797 3
32. Romero, A., Arbeláez, P., Van Gool, L., Timofte, R.: Smit: Stochastic multi-label image-to-image translation. arXiv preprint arXiv:1812.03704 (2018) 3, 4

33. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. arXiv preprint arXiv:1903.12296 (2019) 3, 4
34. Lu, Y., Tai, Y.W., Tang, C.K.: Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 282–297 3
35. Song, J., Zhang, J., Gao, L., Liu, X., Shen, H.T.: Dual conditional gans for face aging and rejuvenation. In: IJCAI. (2018) 899–905 3
36. Zhang, G., Kan, M., Shan, S., Chen, X.: Generative adversarial network with spatial attention for face attribute editing. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 417–432 3
37. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. (2017) 5767–5777 4
38. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680 4
39. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE (2018) 59–66 5, 9
40. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5562–5570 5, 9
41. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15) (2014) E1454–E1462 6, 9
42. See, J., Yap, M.H., Li, J., Hong, X., Wang, S.J.: Megc 2019–the second facial micro-expressions grand challenge. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE (2019) 1–5 7, 8
43. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778 8, 10
44. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE (2010) 94–101 9
45. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition, Springer (2018) 428–438 10
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014) 10
47. Peng, M., Wu, Z., Zhang, Z., Chen, T.: From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE (2018) 657–661 13
48. Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE (2017) 118–126 13, 14

49. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European conference on computer vision (ECCV). (2018) 222–237 13, 14