Feature Representation Learning with Adaptive Displacement Generation and Transformer Fusion for Micro-Expression Recognition

Zhijun Zhai[1], Jianhui Zhao[1]*, Chengjiang Long[2], Wenju Xu[3], Shuangjiang He[4], Huijuan Zhao[4]
[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2]Meta Reality Labs, Burlingame, CA, USA
[3]OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA
[4]FiberHome Telecommunication Technologies Co., Ltd, Wuhan, Hubei, China
zhijunzhai@whu.edu.cn, jianhuizhao@whu.edu.cn, clong1@meta.com, wenjuxu123@gmail.com

Abstract

Micro-expressions are spontaneous, rapid and subtle facial movements that can neither be forged nor suppressed. They are very important nonverbal communication clues, but are transient and of low intensity thus difficult to recognize. Recently deep learning based methods have been developed for micro-expression (ME) recognition using feature extraction and fusion techniques, however, targeted feature learning and efficient feature fusion still lack further study according to the ME characteristics. To address these issues, we propose a novel framework Feature Representation Learning with adaptive Displacement Generation and Transformer fusion (FRL-DGT), in which a convolutional Displacement Generation Module (DGM) with self-supervised learning is used to extract dynamic features from onset/apex frames targeted to the subsequent ME recognition task, and a well-designed Transformer Fusion mechanism composed of three Transformer-based fusion modules (local, global fusions based on AU regions and full-face fusion) is applied to extract the multi-level informative features after DGM for the final ME prediction. The extensive experiments with solid leave-one-subject-out (LOSO) evaluation results have demonstrated the superiority of our proposed FRL-DGT to state-of-the-art methods.

1. Introduction

As a subtle and short-lasting change, micro-expression (ME) is produced by unconscious contractions of facial muscles and lasts only 1/25th to 1/5th of a second, as illustrated in Figure 1, revealing a person's true emotions underneath the disguise [8, 35]. The demands for ME recognition technology are becoming more and more extensive [2, 18],

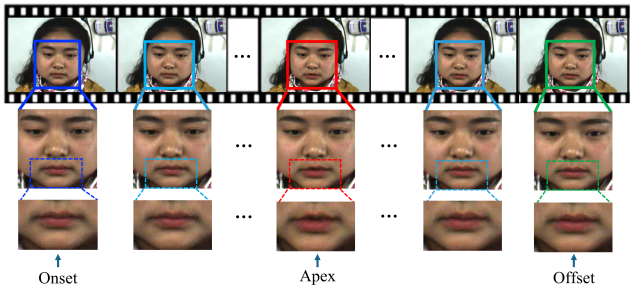


Figure 1. A video sequence depicting the order of which onset, apex and offset frames occur. Sample frames are from a "surprise" sequence in CASME II. Our goal is to design a novel feature representation learning method based on an onset-apex frame pair for facial ME recognition. (Images from CASME II ©Xiaolan Fu)

including multimedia entertainment, film-making, human-computer interaction, affective computing, business negotiation, teaching and learning, *etc*. Since MEs have involuntary muscle movements with short duration and low intensity in nature, the research of ME is attractive but difficult [1, 22]. Therefore, it is crucial and desired to extract robust feature representations to conduct ME analyses.

A lot of feature representation methods are already available including those relying heavily on hand-crafted features with expert experiences [4, 11, 25] and deep learning techniques [33, 36, 40]. However, the performance of deep learning networks is still restricted for ME classification, mainly due to the complexity of ME and insufficient training data [28, 47]. Deep learning methods can automatically extract optimal features and offer an end-to-end classification, but in the existing solutions, dynamic feature extraction is only taken as a data preprocessing strategy. It is not integrated with the subsequent neural network, thus failing to adapt the generated dynamic features to a specific training task, leading to redundancy or missing features. Such shortcoming motivates us to design a dynamic feature extractor to adapt the subsequent ME recognition task.

*Corresponding author.

In this paper, we propose a novel end-to-end feature representation learning framework named FRL-DGT, which is constructed with a well-designed adaptive Displacement Generation Module (DGM) and a subsequent Transformer Fusion mechanism composed of the Transformer-based local, global, and full-face fusion modules, as illustrated in Figure 2, for ME feature learning to model global information while focusing on local features. Our FRL-DGT only requires a pair of images, *i.e.*, the onset and the apex frames, which are extracted from the frame sequence.

Unlike the previous methods which extract optical flow or dynamic imaging features, our DGM and corresponding loss functions are designed to generate the displacement between expression frames, using a convolution module instead of the traditional techniques. The DGM is involved in training with the subsequent ME classification module, and therefore its parameters can be tuned based on the feedback from classification loss to generate more targeted dynamic features adaptively. We shall emphasize that the labeled training data for ME classification is very limited and therefore the supervised data for our DGM is insufficient. To handle this case, we resort to a self-supervised learning strategy and sample sufficient additional random pairs of image sequence as the extra training data for the DGM, so that it is able to fully extract the necessary dynamic features adaptively for the subsequent ME recognition task.

Regarding fusing the dynamic features extracted from DGM, we first adopt the AU (Action Unit) region partitioning method from FACS (Facial Action Coding System) [49] to get 9 AUs, and then crop the frames and their displacements into blocks based on the 9 AUs and the full-face region as input to the Transformer Fusion. We argue that the lower layers in the Transformer Fusion should encode and fuse different AU region features in a more targeted way, while the higher layers can classify MEs based on the information of all AUs. We propose a novel fusion layer with attentions as a linear fusion before attention mechanism [45], aiming at a more efficient and accurate integration of the embedding vectors. The fusion layers are interleaved with Transformer's basic blocks to form a new multi-level fusion module for classification to ensure it to better learn global information and long-term dependencies of ME.

To summarize, our main contributions are as follows:

- We propose a novel end-to-end network FRL-DGT which fully explores AU regions from onset-apex pair and the displacement between them to extract comprehensive features via Transformer Fusion mechanism with the Transformer-based local, global, and full-face feature fusions for ME recognition.
- Our DGM is well-trained with self-supervised learning, and makes full use of the subsequent classification supervision information in the training phase, so that the trained DGM model is able to generate more targeted ME dynamic features adaptively.
- We present a novel fusion layer to exploit the linear fusion before attention mechanism in Transformer for fusing the embedding features at both local and global levels with simplified computation.
- We demonstrate the effectiveness of each module with ablation study and the outperformance of the proposed FRL-DGT to the SOTA methods with extensive experiments on three popular ME benchmark datasets.

2. Related Works

Micro-expression Recognition related technologies [1, 22, 23, 36, 43] mainly fall into two types. One category [25] (*e.g.*, EMRNet [26], STSTNet [24] and DSSN [15]) uses only onset and apex frames. Dynamic features (*e.g.*, optical flow [11]) between the two frames are extracted and fed into a 2D CNN, reducing the computation cost while retaining most of the features [14, 33, 48]. In the other category, a sequence of dynamic feature maps between every two adjacent frames are extracted and learned by a time series network or a 3D CNN (*e.g.*, ELRCN [16], STRCN-A [41] and 3DFlowNet [19]), taking the spatio-temporal features from the whole sequence as input. The entire image sequence can also be compressed into a single dynamic feature map (*e.g.*, dynamic imaging [4, 40]) and then processed (*e.g.*, LEARNet [40] and AffectiveNet [39]), maintaining both high-level and micro-level information. However, the LOSO validation accuracy of networks based on image sequences is generally lower than those using image pairs as input, probably because the information redundancy makes it difficult to focus on the most important features. Following the mainstream methods, our FRL-DGT model also takes an onset-apex frame pair as input, but we still extend it on the whole sequence to validate its superiority.

Dynamic Features are good at capturing subtle changes existing in the frame sequences. As a common type, optical flow and its variants (*e.g.*, Bi-WOOF [25] and MDMO [27]) can estimate the direction and magnitude of the displacement between two frames and extract the inter-frame motion information. Another useful type is dynamic image [4], which is a single RGB image obtained by compressing both the spatial information and temporal dynamic features from the image sequence. To some extent, both these two kinds of dynamic features have been successfully applied to ME recognition [25, 40]. Instead of using the common dynamic features, we design an adaptive DGM into the end-to-end pipeline so that we are able to extract the more targeted dynamic features for ME classification.

Visual Transformers [3, 6, 7, 38] have evolved rapidly with powerful variants developed for image and video classification. The input frames are split into evenly divided

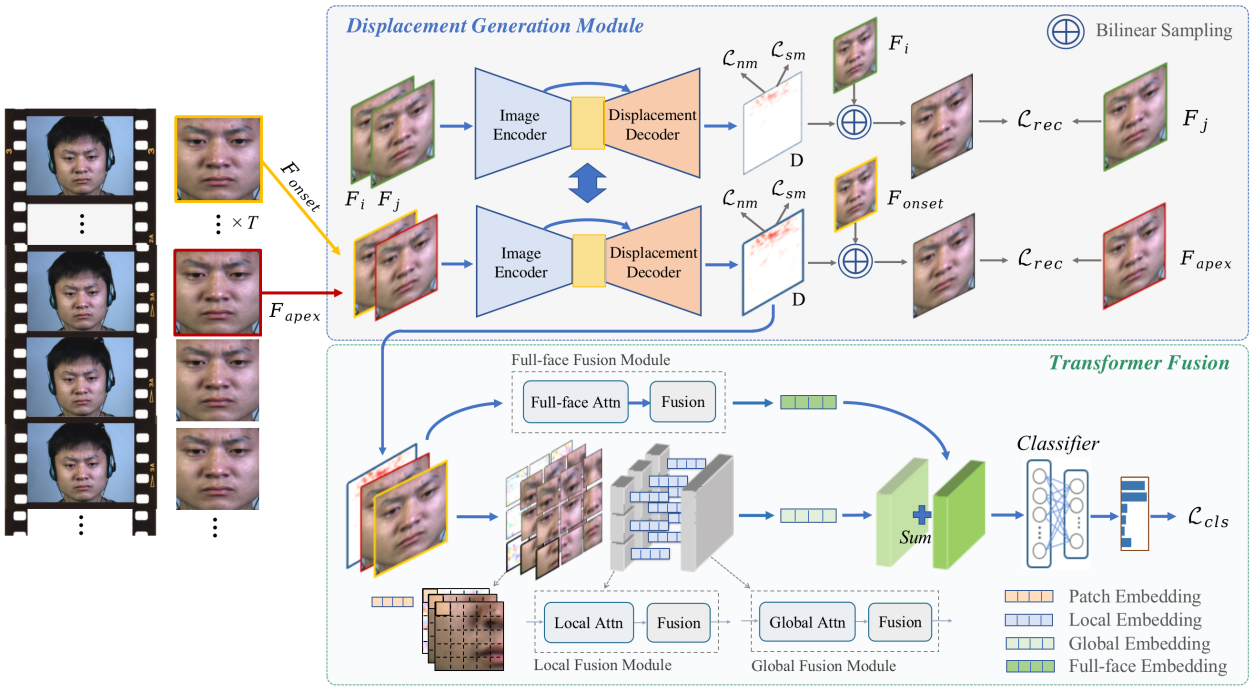


Figure 2. The overview of FRL-DGT which is an end-to-end structure. Given an input video clip, we firstly crop the front face regions and then select an onset-apex pair of frames as the input of Displacement Generation Module (DGM) to calculate the displacement adaptively. The Transformer Fusion composed of three modules, *i.e.*, local fusion module, global fusion module, and full-face fusion module, takes the onset-apex pair and the displacement between them as input to extract the final feature representation via both patch and integral feature fusion with Transformer for classification on ME categories. In addition, the random pairs of frames are used as auxiliary data to self-supervise the training of DGM. (Images from CASME II ©Xiaolan Fu)

fixed-size patches, which are linearly projected into tokens and fed into a Transformer encoder. Obviously, such division may divide the key parts into different patches, and is independent of image content. Inspired by Transformer iN Transformer (TNT) [13] and Swin Transformer [29], we take AU regions as input to our Transformer Fusion to extract local features and learn global information.

3. Proposed Method

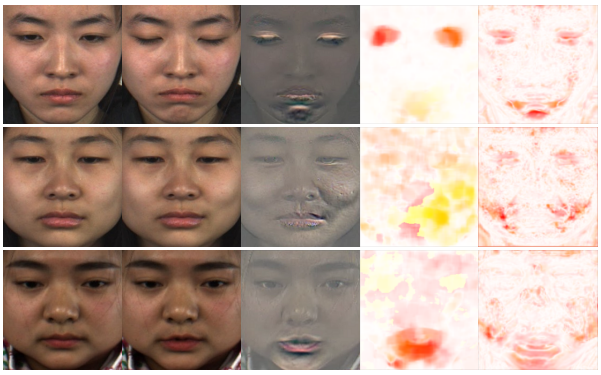
We propose a novel end-to-end ME recognition network, named FRL-DGT, as shown in Figure 2. It takes an onset-apex image pair from a frame sequence as input, and generates displacement features between them through the DGM trained with self-supervision (Section 3.1). The displacement is concatenated with the corresponding onset-apex pair, cropped according to the AU regions and full-face region, and then fed into the Transformer Fusion (Section 3.2) to obtain strong feature representation for ME classification.

3.1. Displacement Generation Module with Self-supervised Learning

The Displacement Generation Module (DGM) here is designed to extract the adaptive dynamic features for the specific ME task. It takes the onset-apex image pair as input and outputs a pixel displacement feature map D between the two frames. The structure of DGM follows an encoder-decoder style, first downsampling the high-resolution input images to obtain low-dimensional dynamic features, and then upsampling them to obtain displacements between image pairs, as shown in Figure 2. The basic block in DGM is a stack of convolutional layers, batch normalization layers, and nonlinear activation layers. Note that we normalize the displacements before output, increasing the intensity of minor expressions and decreasing the intensity of major ones, which plays a role of adaptive expression adjustment.

Similar to optical flow features, the displacement values represent the relative pixel position shifts in x and y directions from onset frame to apex frame. To limit the range of movable pixel positions and make the model easier to learn, we multiply the output displacement in the range [-1, 1] by a scaling factor α. Three loss components are set for the output displacement: reconstruction loss $\mathcal{L}_{rec}$, normalization loss $\mathcal{L}_{nm}$, and smoothing loss $\mathcal{L}_{sm}$.

Let F_{onset} be the onset frame, F_{apex} be the original apex frame, $D_{x,y}$ be the displacement value at (x, y), and $\mathbb{G}_s(F_{onset}, D)$ represents the approximate apex frame obtained by bilinear sampling the onset frame according to the generated displacement D. That is, bilinear sampling moves the pixel at location (x, y) in the onset frame to (x+$D^x_{x,y}$, y+$D^y_{x,y}$) of the approximate apex frame. Then



(a) (b) (c) (d) (e)

Figure 3. Visualization of the generated displacement on three ME categories: surprise, positive, and negative (from top to bottom). On each row from left to right are (a) Onset, (b) Apex, (c) Dynamic image, (d) Optical flow, and (e) Our displacement, respectively. (Images from CASME II ©Xiaolan Fu)

the displacement related loss $\mathcal{L}_{DGM}$ is calculated by:

$$\mathcal{L}_{rec} = |\mathbb{G}_s(F_{onset}, D) - F_{apex}|, \tag{1}$$

$$\mathcal{L}_{nm} = \frac{1}{w \times h} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} |D_{x,y}|, \tag{2}$$

$$\mathcal{L}_{sm} = \frac{\sum_{y=0}^{h-1} \sum_{x=1}^{w-1} |D_{x,y} - D_{x-1,y}|}{h \times (w-1)} + \frac{\sum_{x=0}^{w-1} \sum_{y=1}^{h-1} |D_{x,y} - D_{x,y-1}|}{w \times (h-1)}, \tag{3}$$

$$\mathcal{L}_{DGM} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{nm}\mathcal{L}_{nm} + \lambda_{sm}\mathcal{L}_{sm}, \tag{4}$$

where (h, w) is the size of images, λ_{rec}, λ_{nm} and λ_{sm} are different weights assigned to each loss component.

Note that only $\mathcal{L}_{DGM}$ is not enough to generate the dynamic features specific for ME recognition, we have to combine it to classification loss $\mathcal{L}_{cls}$ together into the end-to-end training. To avoid the underfitting issue due to the limited ME data samples, we resort to a self-supervised learning strategy by sampling sufficient number of image pairs with the $\mathcal{L}_{DGM}$ loss applied only. It is worthy mentioning that self-supervision here is very critical due to the number of training images in the field of ME recognition.

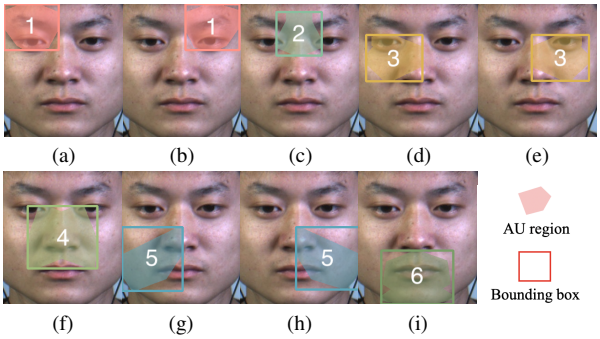
To make it easy to understand our DGM module, we visualize the generated displacements in Figure 3. Our generated displacement is more targeted to AU regions and more sensitive to capture the subtle changes of related MEs.

3.2. Transformer Fusion on Onset-Apex Pair and Extracted Displacement

3.2.1 AU Regions

AU regions are divided with reference to the rules in [30], *i.e.*, 68 landmark points are obtained using the dlib package, and the face is divided into 43 basic region-of-interest (RoIs). As shown in Figure 4, 9 AUs are selected and each AU corresponds to multiple RoIs based on the relationship between MEs and facial muscle movements [9, 10].

In order to make the size of each bounding box appropriate, AU regions #1, #3, and #5 (marked in red, yellow, and blue, respectively in Figure 4) are all divided into left and right parts of each. AU regions #1 and #2 are mainly concerned with changes above the eyes, such as eyebrow lifting and frowning. AU regions #3 and #4 are responsible for changes in the middle of the face, around the nose and lower eyelids. While the changes at the mouth, chin and cheeks are focused on by AU regions #5 and #6.



(a) (b) (c) (d) (e)

(f) (g) (h) (i)

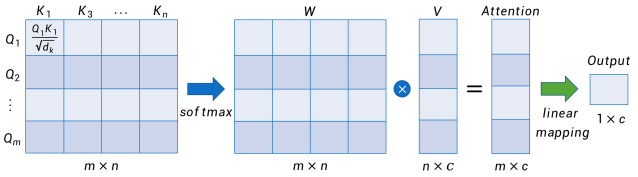
Figure 4. Visualization of 9 AUs as input of Transformer Fusion module. (a) AU region #1 left, (b) AU region #1 right, (c) AU region #2, (d) AU region #3 left, (e) AU region #3 right, (f) AU region #4, (g) AU region #5 left, (h) AU region #5 right, and (i) AU region #6. (Images from CASME II ©Xiaolan Fu)

3.2.2 Feature Fusion Modules with Transformer

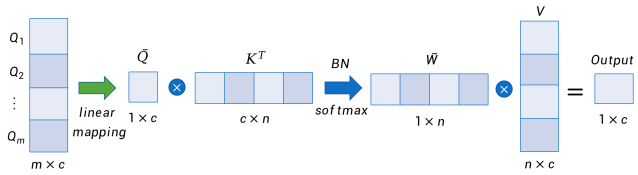
The three Transformer-based feature fusion modules (local, global and full-face fusions) perform learning and fusion of embedding vectors in a hierarchical manner based on K cropped AU regions and the full-face region. The target AU boxes can be set to different resolutions, and the i-th AU box corresponds to size (H_i, W_i). The number of channels M is the same for all AU regions (gray or colorful), which is equal to that of the input feature map.

Similar to the method in Vision Transformer (ViT) [7] with patch size $P \times P$, we divide each AU $x_r \in \mathbb{R}^{H_i \times W_i \times M}$ into a sequence of image patches $x_p \in \mathbb{R}^{N \times P \times P \times M}$, where $N = H_i W_i / P^2$ is the resulting number of patches. The linear projection maps each patch to a C-dimensional embedding vector (*i.e.*, vector length is C) without using the class token. Then both local and global feature fusions are applied to the selected AUs to obtain strong feature representation for ME classification, as shown in Figure 2.

Fusion with Attention. Each embedding vector is dot-multiplied with the weight matrices to obtain the query, key and value, which are all C-dimensional vectors. For merging, the vectors of $N \times C$ output from the attention layer can be mapped to C dimensions by a linear transformation, as shown in Figure 5 (a). In contrast, for the fusion layer



(a) Fusion after attention



(b) Fusion before attention

Figure 5. Comparison of attention and fusion mechanisms. Different from the original attentions with (a) linear fusion, our proposed fusion layer (b) obtains the linear mapped queries first, which has the advantages of reducing noise and simplifying computation.

in Figure 5 (b), we perform linear mapping to the queries of each key before dot product to remove the influence of noise, then pass it through a Batch Normalization layer and a Softmax function to obtain a probability vector that represents the relative importance of it in overall sequence, and the probabilities are used to weight the sum of all values.

The fusion layer adopts a multi-head mechanism that allows the model to learn different importance distributions of embedding vectors in multiple subspaces. For each head, we set $c = C/h$, where h is the number of attention heads. Packing the k-th head part (h parts in total) of the embedding vector of queries, keys and values together into matrices $Q \in \mathbb{R}^{m\times c}$, $K \in \mathbb{R}^{n\times c}$ and $V \in \mathbb{R}^{n\times c}$ respectively (Note that $m = n$ in our case), the fused embedding vector of the k-th head is obtained by:

$$\overline{W}_{raw} = Lin([Q_1, Q_2, ..., Q_m]) \cdot K^T, \tag{5}$$

$$head_k = Softmax(BN(\overline{W}_{raw})) \cdot V, \tag{6}$$

where Lin performs the linear transformation of Q, and BN stands for the batch normalization layers. The final fusion result is obtained by concatenating the fused embedding vector F from h heads:

$$F = Concat(head_1, \ldots, head_h), \tag{7}$$

Note that when all weights of linear mappings are equal, it corresponds to an averaging operation. The fusion before attention mechanism is used in the fusion layers of local, global and full-face modules in Transformer Fusion.

Local Fusion Module. The first level contains K local modules, and different AUs are processed by different local modules respectively. The i-th local module performs feature extraction on the N patches of the i-th AU and fuses them into a local embedding vector that contains the spatial (onset-apex pair) and dynamic (displacement) features.

Global Fusion Module. The second level contains a spatial module that learns and fuses the K local embedding vectors outputted from the previous level to obtain the expression information contained in each frame, represented by a global embedding vector.

Full-face Fusion Module. In addition to the hierarchical features of AUs, we perform attention learning and spatial feature fusion on entire face images to obtain full-face features as auxiliary classification information. Note that each module in local, global and full-face feature fusions consists of an Attn block for feature learning and a fusion layer for information synthesis. Attention Learning (Attn) block is a stack of multiple Transformer layers, including Layer Normalization, Multi-head Attention and Multi-layer Perceptron. The resulting embedding vector is fed into the subsequent fully connected layers for ME classification.

Discussion: Compared to CNN networks, basic Transformer structures lack translation equivariance and locality, thus generalizing poorly when the amount of training data is insufficient. To overcome the weakness, we calibrate the face to a fixed position and size before performing recognition, so that we do not need to pay much attention to rotation and translation invariance. Then we build the Transformer architecture in the form of multi-level fusion, which allows better focus on local features. Moreover, linear fusion before attention can suppress noise and thus is beneficial to error elimination, meanwhile, our method reduces the amount of calculation by early using a vector instead of a matrix.

3.3. Implementation Details

Each of the selected $K = 9$ AUs is with a size of (90, 90), the patch size is (18, 18), and the number of patches is $N = 25$. The dimension C of embedding vectors is 256, while both the attention and fusion layers use $h = 8$ heads. In the formulas of DGM, we take $\lambda_{rec} = 10$, $\lambda_{nm} = 1$, $\lambda_{sm} = 0.2$, and displacement scaling factor $\alpha = 0.2$. The batch size is 32, and the optimizer is Adam with an initial learning rate of 0.002 for DGM, and the Adam with cosine annealing strategy for Transformer Fusion. The gradient in DGM back-propagated from the classification loss is scaled by 10^{-6} to prevent it from being dominant.

Note that to increase the amount of image pairs for training and improve the sensitivity of our network to the subtle expression changes, we use MagNet [32] to augment the datasets as in [17], and also perform a randomization operation: select a frame before or after the original apex frame randomly when loading the data for training.

4. Experiments

4.1. Datasets and Metrics

Since the Composite Database Evaluation (CDE) is commonly used for evaluation and comparison in the field

Method	Year	Type	Full		SMIC Part		SAMM Part		CASME II Part	
			UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [44]	2014	Hand-Crafted	0.588	0.579	0.200	0.528	0.395	0.410	0.703	0.743
Bi-WOOF [25]	2018	Hand-Crafted	0.630	0.623	0.573	0.583	0.521	0.514	0.781	0.803
CapsuleNet [33]	2019	Deep-Learning	0.652	0.651	0.582	0.588	0.621	0.599	0.707	0.702
STSTNet [24]	2019	Deep-Learning	0.735	0.761	0.680	0.701	0.659	0.681	0.838	0.869
RCN-A [42]	2020	Deep-Learning	0.743	0.719	0.633	0.644	0.760	0.672	0.851	0.812
GEME [31]	2021	Deep-Learning	0.740	0.750	0.629	0.657	0.687	0.654	0.840	0.851
MERSiamC3D [51]	2021	Deep-Learning	0.807	0.799	0.736	0.760	0.748	0.728	0.882	0.876
FeatRef [52]	2022	Deep-Learning	0.784	0.783	0.701	0.708	0.737	0.716	0.892	0.887
FRL-DGT	2022	Deep-Learning	0.812	0.811	0.743	0.749	0.772	0.758	0.919	0.903
EMRNet [26]*	2019	Deep-Learning	0.789	0.782	0.746	0.753	0.775	0.715	0.829	0.821
FGRL-AUF [17]*	2021	Deep-Learning	0.791	0.793	0.719	0.722	0.775	0.789	0.880	0.871
ME-PLAN [50]*	2022	Deep-Learning	0.772	0.786	0.713	0.726	0.716	0.742	0.863	0.878

Table 1. Performance comparison of the SOTA methods and our proposed FRL-DGT in terms of UF1 and UAR. The best and second best results are marked in red and blue colors, respectively. Methods with * use different datasets, and they have underlined higher scores.

of ME recognition because of the small amount of collected samples, we use 3 ME datasets CASME II [44], SAMM [5] and SMIC [21, 37] for composite training[1]. And we adopt the same way as in MEGC2019 Challenge [34] to unify different category settings across datasets, mapping them to 3 general classes: **Negative**{ "Repression", "Anger", "Contempt", "Disgust", "Fear", "Sadness" }, **Positive**{"Happiness"}, and **Surprise**{"Surprise"}. Unrelated or undefined emotion categories such as "Others" are omitted to reduce confusion for model training.

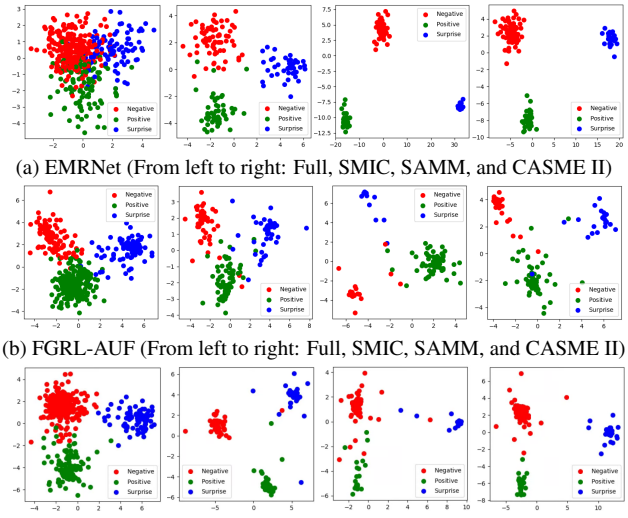
The 3 ME datasets have 442 image sequences from 68 subjects, 25,469 images in total with 58 averaged frames per sequence. To clarify, the onset (starting time of ME), apex (time with the highest intensity of ME), and offset (ending time of ME) frames are already labeled and provided in the benchmark datasets SAMM and CASME II. We follow [26] to obtain the apex frames of image sequences in SMIC, which are not officially labeled. For real scenarios, there are a lot of proven methods to locate the onset, apex, and offset frames from a video [1, 20]. After obtaining the key frames, we perform face calibration with dlib package to handle image-plane rotation and translation.

Regarding the evaluation metrics, we follow [26, 33], taking Unweighted F1-score (UF1) and Average Recall (UAR) with leave-one-subject-out (LOSO) cross validation to evaluate the ME recognition performance.

4.2. Comparison to State-of-the-art Methods

The comparative SOTA methods include representative works of two main ME recognition categories and mainstream approaches based on deep learning in recent years. As shown in Table 1, both UF1 and UAR of our framework

[1]The three datasets were received and exclusively accessed by the author Zhijun Zhai and Jianhui Zhao for purely academic research only. The author Zhijun Zhai produced the experimental results in this paper. Meta did not have access to the datasets as part of this research.



(a) EMRNet (From left to right: Full, SMIC, SAMM, and CASME II)

(b) FGRL-AUF (From left to right: Full, SMIC, SAMM, and CASME II)

(c) FRL-DGT (From left to right: Full, SMIC, SAMM, and CASME II)

Figure 6. The feature distributions of EMRNet, FGRL-AUF and our proposed FRL-DGT on the evaluation datasets.

FRL-DGT are higher than 0.810. FRL-DGT improves by 0.62% and 1.50% on UF1 and UAR over MERSiamC3D, the second best deep learning based ME classifier.

There are other existing efficient methods [12, 46], *e.g.*, EMRNet [26] introduces CK+ dataset to implement domain adaptation, FGRL-AUF [17] uses only CASME II and SAMM datasets for the annotated AUs, and ME-PLAN [50] constructs a pre-training by combining macro-expression samples in CK+, Oulu-CASIA and DFEW datasets. They may have higher scores on certain dataset, but their full scores are still less than our FRL-DGT. As shown in Figure 6, the distribution of extracted features from our FRL-DGT is more separated than those of EMRNet and FGRL-AUF, resulting in better classification results. Taken together, our network has excellent results in the same type of methods and is highly competitive among them.

Method	DGM	AU Regions	Full-face Fusion	Global Fusion	Local Fusion	Fu-B-Attn	Full		SMIC Part		SAMM Part		CASME II Part	
							UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
M0	→OpticalFlow	✓	✓	✓	✓	✓	0.741	0.718	0.671	0.662	0.695	0.662	0.846	0.834
M1	→OF+NORM	✓	✓	✓	✓	✓	0.758	0.739	0.671	0.667	0.778	0.730	0.869	0.831
M2	→DynamicImage	✓	✓	✓	✓	✓	0.739	0.720	0.684	0.679	0.762	0.745	0.759	0.716
M3	w/o self-supervise	✓	✓	✓	✓	✓	0.778	0.777	0.707	0.718	0.697	0.677	0.914	0.889
M4	✓	✓	✓	✓	✓	→Fu-A-Attn	0.797	0.792	0.746	0.746	0.734	0.719	0.898	0.885
M5	✓	→3x3 image patches	✓	✓	✓	✓	0.765	0.765	0.665	0.673	0.754	0.734	0.894	0.876
M6	✓	✓	×	✓	✓	✓	0.773	0.774	0.689	0.698	0.758	0.704	0.876	0.881
M7	✓	✓	✓	✓	×	✓	0.781	0.765	0.741	0.745	0.725	0.672	0.848	0.838
M8	✓	✓	✓	×	✓	✓	0.782	0.773	0.701	0.706	0.711	0.671	0.904	0.886
M9	✓	✓	✓	✓	✓	✓	0.812	0.811	0.743	0.749	0.772	0.758	0.919	0.903

Table 2. Ablation study of our proposed network. "→X" indicates replacing the corresponding component with X. ✓ and × represent yes or no, respectively. The best and second best results are marked in red and blue colors, respectively.

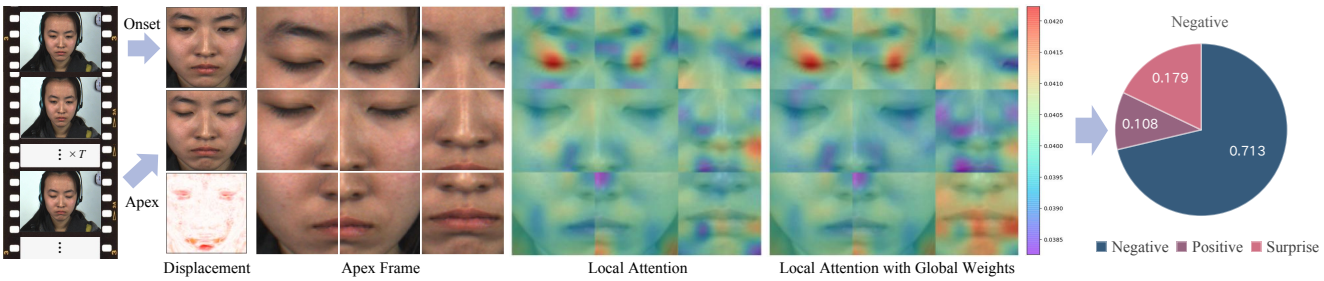


Figure 7. Visualization of the weights in fusion layers. (Images from CASME II ©Xiaolan Fu)

4.3. Ablation Study

Displacement Generating Module. The superiority of DGM over the two conventional methods is demonstrated in Table 2, where DynamicImage stands for dynamic imaging method and OpticalFlow stands for optical flow method. The ME performance by replacing DGM with optical flow or dynamic imaging is lower than that of FRL-DGT, which is an end-to-end network combining DGM and Transformer Fusion. We also replace DGM with the normalized OpticalFlow (OF+NORM), except for the boost of accuracy on SAMM by NORM, the overall performance is still inferior to that of DGM. The dynamic features obtained by the three approaches are visualized in Figure 3. It can be seen that all the three dynamic features can highlight the changing regions, but the displacement generated by DGM can be automatically adjusted according to the classification loss, obtaining additional hidden information. Thus, the following Transformer Fusion module can carry on more objective learning and get more reliable classification results. From M3 and M9 in Table 2, we can also find that self-supervised learning is very useful for DGM, helps improve the full UF1 score significantly from 0.778 to 0.812 by 4.37%, and the UAR score significantly from 0.777 to 0.811 by 4.38%.

Transformer Fusion. The fusion layer with linear fusion before attention (Fu-B-Attn) of M9 is replaced with a normal linear fusion after attention (Fu-A-Attn) of M4. Comparison between M4 and M9 in Table 2 demonstrates that our Fu-B-Attn merges embedding vectors in a more effective way. For runtime, Fu-A-Attn takes 50.8ms while Fu-B-Attn only needs 47.6ms to run all fusions, which proves that our Fu-B-Attn can simplify computation.

Based on the results of M6, M7, and M8, there are contributions from three fusion modules in Transformer Fusion, i.e., local fusion module, global fusion module, and full-face fusion module, respectively. Figure 7 plots the original weight distribution in local fusion layer with each AU region and the weight distribution after global fusion, which shows that the global fusion further adjusts the attention to make it more focused on the eyes and mouth regions. Note that we compress the global fusion weights while maintaining the comparison order to facilitate visualization.

AU Regions. We explore the difference between using 9 different AUs and using 3×3 image patches divided evenly as input. From M5 and M9 shown in Table 2, we can clearly see that using specific AUs allows the Transformer Fusion to perform more targeted learning, where the lower layers can focus on extracting features from individual AUs, while the higher layers can classify based on information of all AU regions, resulting in better performance.

4.4. Discussions

Extension of FRL-DGT on the Whole Sequence. For the classification of an expression sequence, we extend our FRL-DGT to FRL-DGT-S, in which the input data has an additional time dimension and the sequence is converted to T frames by the Temporal-Interpolation-Model (TIM) [53]. The onset frame is concatenated with all other frames separately to obtain a sequence containing $T-1$ onset-other

(a) 4 classes (b) 5 classes

Figure 8. The confusion matrices of our proposed FRL-DGT and FRL-DGT-S on CASME II dataset with 4 and 5 ME classes.

image pairs as input to DGM-S. Then the output displacement concatenated with its corresponding frame is input to sequenced Transformer Fusion (TransFu-S) for classification, and TransFu-S requires a further fusion step to model the temporal dependencies between T frames in addition to the two-level fusion of TransFu. In our experiments, image sequences are interpolated to 10 frames using TIM, with the batch size 5 and same optimizer setting as FRL-DGT.

For the comparison of FRL-DGT-S and the SOTA methods which take image sequences as input, we conduct experiments on CASME II dataset. However, the classes used by ELRCN and STRCN-A are different, ELRCN uses 5 classes (*i.e.*, Happiness, Disgust, Repression, Surprise and Others) while STRCN-A uses 4 classes (*i.e.*, Positive, Negative, Surprise and Others), so we conduct experiments separately and compare with the corresponding methods.

The comparison results are listed in Table 3, where we cite the results from papers of ELRCN and STRCN-A, and our FRL-DGT-S outperforms them under the corresponding settings. We can also find from Table 3 that FRL-DGT has higher precision than FRL-DGT-S for both 4 classes and 5 classes, which is previously proved by related works and illustrated with confusion matrices in Figure 8. To investigate the reason why FRL-DGT-S is not as effective as FRL-DGT, we compare the interpolated frames after TIM and the original apex frame, which indicates that TIM may miss apex information, resulting in less good performance.

	Method	#(Classes)	CASME II		
			UF1	UAR	Acc
Pair	FRL-DGT	4	0.750	0.732	0.780
	FRL-DGT	5	0.748	0.735	0.757
Sequence	STRCN-A [41]	4	0.542	-	0.560
	FRL-DGT-S	4	0.562	0.549	0.643
	ELRCN [16]	5	0.500	0.440	0.524
	FRL-DGT-S	5	0.543	0.539	0.594

Table 3. Results of FRL-DGT on more ME classes and FRL-DGT-S taking image sequences as input.

Sensitivity To Onset and Apex. Perturbations on onset/apex frames of CASME II include 10/20/30% deviation between onset and apex, e.g., onset+10% is about 3 frames after onset. The averaged UF1/UAR from onset deviations are 0.848/0.838, 0.817/0.808, 0.786/0.769, while the results are 0.866/0.858, 0.830/0.834, 0.814/0.795 for apex.

Figure 9. Failure cases of our FRL-DGT. GT stands for Ground Truth. (Images from CASME II ©Xiaolan Fu)

Automatic Detection on Apex. Without using the labeled frames, we automatically detect apex frames with the algorithm of [33], causing the averaged UF1/UAR decrease from 0.812/0.811 to 0.658/0.648 for all the three datasets.

Running Time. All experiments are performed on 3080 Ti GPU with 12GB memory, and i9 CPU with 2.8GHz. The average time of reading one picture from dataset, detecting the face, and outputting classification result is 0.416s, while the average time is 0.384s only for ME recognition.

Failure Cases. There are some failure cases that are difficult to be accurately classified by the existing methods and our FRL-DGT. As shown in Figure 9, the glasses with reflection can confuse the extraction and classification of displacement features, leading to incorrect predictions.

5. Conclusion

For micro-expression recognition, we propose an end-to-end FRL-DGT, which takes onset-apex image pairs as input. The convolutional module DGM with self-supervised learning is used instead of traditional dynamic feature extractions, and the classification loss can back-propagate the gradient to DGM, modifying the information contained in the generated displacement. Our designed classification module Transformer Fusion consists of Transformer's basic layers and the novel fusion layer, utilizing cropped AU regions and the full-face region as input for multi-level learning and fusion, and using the linear fusion before attention mechanism with efficient and accurate integration of embedding vectors. The LOSO evaluation result of our FRL-DGT has higher precision than the SOTA methods on UF1 and UAR tested with the same datasets, and the ablation experiments demonstrate the effectiveness of each proposed module.

References

[1] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5826–5846, 2021. 1, 2, 6

[2] Allaert Benjamin, Bilasco Ioan Marius, and Djeraba Chaabane. Micro and macro facial expression recognition using advanced local motion patterns. *IEEE Transactions on Affective Computing*, 13(1):147–158, 2022. 1

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the IEEE Conference on International Conference on Machine Learning (ICML)*, 2021. 2

[4] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[5] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 9(1):116–129, 2016. 6

[6] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 2, 4

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

[9] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. 4

[10] Paul Ekman and Wallace V Friesen. Facial action coding system (facs): A technique for the measurement of facial action. *Rivista Di Psichiatria*, 47(2):126–138, 1978. 4

[11] David Fleet and Yair Weiss. Optical flow estimation. In *Handbook of Mathematical Models in Computer Vision*, pages 237–257. Springer, 2006. 1, 2

[12] Puneet Gupta. Merastc: Micro-expression recognition using effective feature encodings and 2d convolutional neural network. *IEEE Transactions on Affective Computing*, 2021. 6

[13] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2021. 3

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[15] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin. Dual-stream shallow networks for facial micro-expression recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019. 2

[16] Huai-Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao Lin. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018. 2, 8

[17] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 5, 6

[18] Peter Lewinski, Marieke L Fransen, and Ed SH Tan. Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, 7(1):1–14, 2014. 1

[19] Jing Li, Yandan Wang, John See, and Wenbin Liu. Micro-expression recognition based on 3d flow convolutional neural network. *Pattern Analysis and Applications*, 22(4):1331–1339, 2019. 2

[20] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing*, 9(4):563–577, 2017. 6

[21] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Proceedings of the 10th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2013. 6

[22] Yante Li, Xiaohua Huang, and Guoying Zhao. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Transactions on Image Processing*, 30:249–263, 2020. 1, 2

[23] Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(4):2028–2046, 2022. 2

[24] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019. 2, 6

[25] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018. 1, 2, 6

[26] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019. 2, 6

[27] Yongjin Liu, Bingjun Li, and Yukun Lai. Sparse mdmo: Learning a discriminative feature for micro-expression recognition. *IEEE Transactions on Affective Computing*, 12(1):254–261, 2018. 2

[28] Yang Liu, Xingming Zhang, Yante Li, Jinzhao Zhou, Xin Li, and Guoying Zhao. Graph-based facial affect analysis: A review. *IEEE Transactions on Affective Computing*, pages 1–20, 2022. 1

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021. 3

[30] Chen Ma, Li Chen, and Junhai Yong. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *Neurocomputing*, 355:35–47, 2019. 4

[31] Xuan Nie, Madhumita A. Takalkar, Mengyang Duan, Haimin Zhang, and Min Xu. Geme: Dual-stream multi-task gender-based micro-expression recognition. *Neurocomputing*, 427:13–28, 2021. 6

[32] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T. Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[33] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019. 1, 2, 6, 8

[34] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Sujing Wang. Megc 2019–the second facial micro-expressions grand challenge. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2019. 6

[35] Xunbing Shen, Qi Wu, and Xiaolan Fu. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University Science B*, 13(3):221–230, 2012. 1

[36] Madhumita Takalkar, Min Xu, Qiang Wu, and Zenon Chaczko. A survey: facial micro-expression recognition. *Multimedia Tools and Applications*, 77(15):19301–19325, 2018. 1, 2

[37] Pfister Tomas, Xiaobai Li, Guoying Zhao, and Pietikäinen Matti. Recognising spontaneous facial micro-expressions. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 6

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017. 2

[39] Monu Verma, Santosh Kumar Vipparthi, and Girdhari Singh. Affectivenet: Affective-motion feature learning for microexpression recognition. *IEEE MultiMedia*, 28(1):17–27, 2020. 2

[40] Monu Verma, Santosh Kumar Vipparthi, Girdhari Singh, and Subrahmanyam Murala. Learnet: Dynamic imaging network for micro expression recognition. *IEEE Transactions on Image Processing*, 29:1618–1627, 2019. 1, 2

[41] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia*, 22(3):626–640, 2019. 2, 8

[42] Zhaoqiang Xia, Wei Peng, Huaiqian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29:8590–8605, 2020. 6

[43] Hongxia Xie, Ling Lo, Honghan Shuai, and Wenhuang Cheng. An overview of facial micro-expression analysis: Data, methodology and challenge. *IEEE Transactions on Affective Computing*, 2022. 2

[44] Wenjing Yan, Xiaobai Li, Sujing Wang, Guoying Zhao, Yongjin Liu, Yuhsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS One*, 9(1):e86041, 2014. 6

[45] Jiaqi Yu, Yongwei Nie, Chengjiang Long, Wenju Xu, Qing Zhang, and Guiqing Li. Monte carlo denoising via auxiliary feature guided self-attention. *ACM Transactions on Graphics*, 40(6), 2021. 2

[46] Jianhui Yu, Chaoyi Zhang, Yang Song, and Weidong Cai. Ice-gan: Identity-aware and capsule-enhanced gan for micro-expression recognition and synthesis. *arXiv:2005.04370*, 2020. 6

[47] Liangfei Zhang and Ognjen Arandjelović. Review of automatic microexpression recognition in the past decade. *Machine Learning and Knowledge Extraction*, 3(2):414–434, 2021. 1

[48] Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelovic, and Guoying Zhao. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1973–1985, 2022. 2

[49] Tong Zhang, Yuan Zong, Wenming Zheng, C. L. Philip Chen, Xiaopeng Hong, Chuangao Tang, Zhen Cui, and Guoying Zhao. Cross-database micro-expression recognition: A benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):544–559, 2022. 2

[50] Sirui Zhao, Huaying Tang, Shifeng Liu, Yangsong Zhang, Hao Wang, Tong Xu, Enhong Chen, and Cuntai Guan. Meplan: A deep prototypical learning with local attention network for dynamic micro-expression recognition. *Neural Networks: the official journal of the International Neural Network Society*, 153:427–443, 2022. 6

[51] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. A two-stage 3d cnn based learning method for spontaneous micro-expression recognition. *Neurocomputing*, 448:276–289, 2021. 6

[52] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. 6

[53] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Towards a practical lipreading system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 7