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

Zhijun Zhai¹, Jianhui Zhao¹, Chengjiang Long², Wenju Xu³, Shuangjiang He⁴, Huijuan Zhao⁴
¹School of Computer Science, Wuhan University, Wuhan, Hubei, China
²Meta Reality Labs, Burlingame, CA, USA
³OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA
⁴FiberHome 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], 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
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) \cite{13} and Swin Transformer \cite{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 $\alpha$. Three loss components are set for the output displacement: reconstruction loss $L_{rec}$, normalization loss $L_{nm}$, and smoothing loss $L_{sm}$.

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 $G_s(F_{onset}, D)$ 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 $(x+D_{x,y}, y+D_{y,x})$ of the approximate apex frame. Then
the displacement related loss $\mathcal{L}_{DGM}$ is calculated by:

$$\mathcal{L}_{rec} = |G_s(F_{onset}, D) - F_{apex}|,$$

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

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

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

where $(h, w)$ is the size of images, $\lambda_{rec}$, $\lambda_{nm}$, and $\lambda_{sm}$ 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.

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
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 $\lambda_{rev} = 10$, $\lambda_{nm} = 1$, $\lambda_{sm} = 0.2$, 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
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. 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 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Type</th>
<th>Full UF1</th>
<th>SMIC Part UF1</th>
<th>SAMM Part UF1</th>
<th>CASME II Part UF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP-TOP [44]</td>
<td>2014</td>
<td>Hand-Crafted</td>
<td>0.588</td>
<td>0.200</td>
<td>0.395</td>
<td>0.703</td>
</tr>
<tr>
<td>Bi-WOOF [25]</td>
<td>2018</td>
<td>Hand-Crafted</td>
<td>0.630</td>
<td>0.573</td>
<td>0.521</td>
<td>0.781</td>
</tr>
<tr>
<td>CapsuleNet [33]</td>
<td>2019</td>
<td>Deep-Learning</td>
<td>0.652</td>
<td>0.582</td>
<td>0.621</td>
<td>0.707</td>
</tr>
<tr>
<td>STSTNet [24]</td>
<td>2019</td>
<td>Deep-Learning</td>
<td>0.735</td>
<td>0.680</td>
<td>0.659</td>
<td>0.838</td>
</tr>
<tr>
<td>RCV-A [42]</td>
<td>2020</td>
<td>Deep-Learning</td>
<td>0.743</td>
<td>0.644</td>
<td>0.760</td>
<td>0.851</td>
</tr>
<tr>
<td>GEME [31]</td>
<td>2021</td>
<td>Deep-Learning</td>
<td>0.740</td>
<td>0.657</td>
<td>0.687</td>
<td>0.840</td>
</tr>
<tr>
<td>MERSiamC3D [51]</td>
<td>2021</td>
<td>Deep-Learning</td>
<td>0.807</td>
<td>0.736</td>
<td>0.748</td>
<td>0.882</td>
</tr>
<tr>
<td>FeatRef [52]</td>
<td>2022</td>
<td>Deep-Learning</td>
<td>0.784</td>
<td>0.701</td>
<td>0.737</td>
<td>0.892</td>
</tr>
<tr>
<td>FRL-DGT</td>
<td>2022</td>
<td>Deep-Learning</td>
<td>0.812</td>
<td>0.749</td>
<td>0.772</td>
<td>0.919</td>
</tr>
</tbody>
</table>

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.

---

1The 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.
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-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 3x3 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
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Classes</th>
<th>CASME II</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRL-DGT</td>
<td>4</td>
<td>0.750</td>
</tr>
<tr>
<td>FRL-DGT-S</td>
<td>5</td>
<td>0.748</td>
</tr>
<tr>
<td>STRCN-A [41]</td>
<td>4</td>
<td>0.542</td>
</tr>
<tr>
<td>FRL-DGT-S</td>
<td>4</td>
<td>0.562</td>
</tr>
<tr>
<td>ELRCN [16]</td>
<td>5</td>
<td>0.500</td>
</tr>
<tr>
<td>FRL-DGT-S</td>
<td>5</td>
<td>0.543</td>
</tr>
</tbody>
</table>

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

### 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 UFI and UAR tested with the same datasets, and the ablation experiments demonstrate the effectiveness of each proposed module.

---

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

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


[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


