Micro-Expression Recognition via Fine-Grained Dynamic Perception

ZHIWEN SHAO, YIFAN CHENG, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, and Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, China, and also Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

FAN ZHANG, Inspur Zhuoshu Big Data Industry Development Co., Ltd., Jinan, China

XUEHUAI SHI, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

CANLIN LI, School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China

LIZHUANG MA, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

DIT-YAN YEUNG*, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

Facial micro-expression recognition (MER) is a challenging task, due to the transience, subtlety, and dynamics of micro-expressions (MEs). Most existing methods resort to hand-crafted features or deep networks, in which the former often additionally requires key frames, and the latter suffers from small-scale and low-diversity training data. In this paper, we develop a novel fine-grained dynamic perception (FDP) framework for MER. We propose to rank frame-level features of a sequence of raw frames in chronological order, in which the rank process encodes the dynamic information of both ME appearances and motions. Specifically, a novel local-global feature-aware transformer is proposed for frame representation learning. A rank scorer is further adopted to calculate rank scores of each frame-level feature. Afterwards, the rank features from rank scorer are pooled in temporal dimension to capture dynamic representation. Finally, the dynamic representation is shared by a MER module and a dynamic image construction module, in which the former predicts the ME category, and the latter uses an encoder-decoder structure to construct the dynamic image. The design of dynamic image construction task is beneficial for capturing facial subtle actions associated with MEs and alleviating the data scarcity issue. Extensive experiments show that our method (i) significantly outperforms the state-of-the-art MER methods, and (ii) works well for dynamic image construction. Particularly, our FDP improves by 4.05%, 2.50%, 7.71%, and 2.11% over the previous best results in terms of F1-score on the CASME II, SAMM, CAS(ME)^2, and CAS(ME)^3 datasets, respectively. The code is available at https://github.com/CYF-cuber/FDP.

CCS Concepts: • **Computing methodologies → Computer vision**.

Additional Key Words and Phrases: Micro-expression recognition, rank pooling, dynamic image construction, local-global feature-aware transformer

*Corresponding authors: Yifan Cheng, Xuehuai Shi, and Dit-Yan Yeung.

Authors' Contact Information: Zhiwen Shao, Yifan Cheng, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, and Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, China, and also Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, {zhiwen_shao;yifan_cheng}@cumt.edu.cn; Fan Zhang, Inspur Zhuoshu Big Data Industry Development Co., Ltd., Jinan, China, zhangfan_inspur@foxmail.com; Xuehuai Shi, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China, xuehuai@njupt.edu.cn; Canlin Li, School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China, li-cl@zzuli.edu.cn; Lizhuang Ma, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China, ma-lz@cs.sjtu.edu.cn; Dit-Yan Yeung, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, dyyeung@cse.ust.hk.

1 Introduction

Facial micro-expression recognition (MER) has recently gained increasing attention in the fields of computer vision and affective computing [27, 51, 53, 66]. It has applications in many areas, as micro-expressions (MEs) can reveal emotions those are attempted to conceal [11]. For instance, in mental health, MER can be used to spot signs of disorders like depression and monitors treatment progress. It can also be used to detect non-verbal pain in patients who cannot communicate well. Recently, Zhou *et al.* [77] proposed a multi-modal fine-grained depression detection method via fusing audio and text features. However, visual modality ME that can well reflect the degree of depression is ignored. In public security, MER can be used to aid criminal investigations in terms of lie detection and suspect identification. It can also be used to spot suspicious emotional states in the crowd so as to prevent threats. However, MEs are often neglected in recent lie detection works [18, 22]. Therefore, we explore a new MER solution to empower mental health and public security. MEs are facial subtle muscle actions, and are dynamic during a short duration with no more than 500 milliseconds [69]. Besides, most of the existing ME datasets are small-scale [7, 68], due to the large costs of manual labeling. With limited training data to capture challenging MEs, MER remains a difficult task.

One common solution is to adopt hand-crafted features associated with MEs. These features typically try to capture motion patterns [5, 72], encode spatio-temporal information [4, 74], or focus on local contrast information [8, 32]. However, hand-crafted features based on prior knowledge only process partial characteristics, and have limited capacity to model challenging MEs in diverse samples. Moreover, these features like optical flow [72] often additionally rely on key frames including onset, apex, and offset frames of MEs to improve the recognition performance, which limits the applicability.

Another alternative way is to use prevailing deep neural networks. Zhou *et al.* [79] computed the optical flow between onset and apex frames of the input video, and then fed horizontal and vertical components of the optical flow into a dual-inception network to predict the ME category. However, pre-extracted optical flow as well as key frames are required. Some other methods directly input raw frame images to deep networks so as to remove the limitations of hand-crafted features. For example, Reddy *et al.* [48] employed a 3D convolutional neural network (CNN) to capture spatial and temporal information, and Xia *et al.* [63] used macro-expression recognition to facilitate MER. However, the capture of fine-grained ME information is not explicitly handled, and these methods suffer from insufficient training data.

To tackle the above issues, we introduce a rank pooling technique [13] to perceive temporal evolution of appearance in ME videos, and improve the transformer [58] structure to capture both local and global characteristics. As mentioned in [67], dynamic facial expression recognition has wider practical applications, such as empowering smart cities. Our work is based on dynamic and continuous ME frame sequence rather than pre-annotated key frames. By the rank pooling, the spatial appearances and temporal motions of a video can be encoded as a dynamic image [4], which indicates the correlations between dynamic image and ME, as illustrated in Fig. 1. In particular, we propose an end-to-end Fine-grained Dynamic Perception framework called **FDP**, which jointly estimates ME and constructs dynamic image of the input video. First, a novel local-global feature-aware transformer is proposed to capture ME related local information while preserving the global relational modeling ability of vanilla transformer [58] for single frame representation learning. After extracting the local-global feature, a rank scorer is further employed to learning temporal rank information of each frame. Then, a 3D CNN based temporal pooling module is applied to capture temporal features from all the single rank features so as to learn video-wide dynamic representation. Finally, two modules of MER estimation and dynamic image construction are adopted to predict the ME category and the dynamic image, respectively.

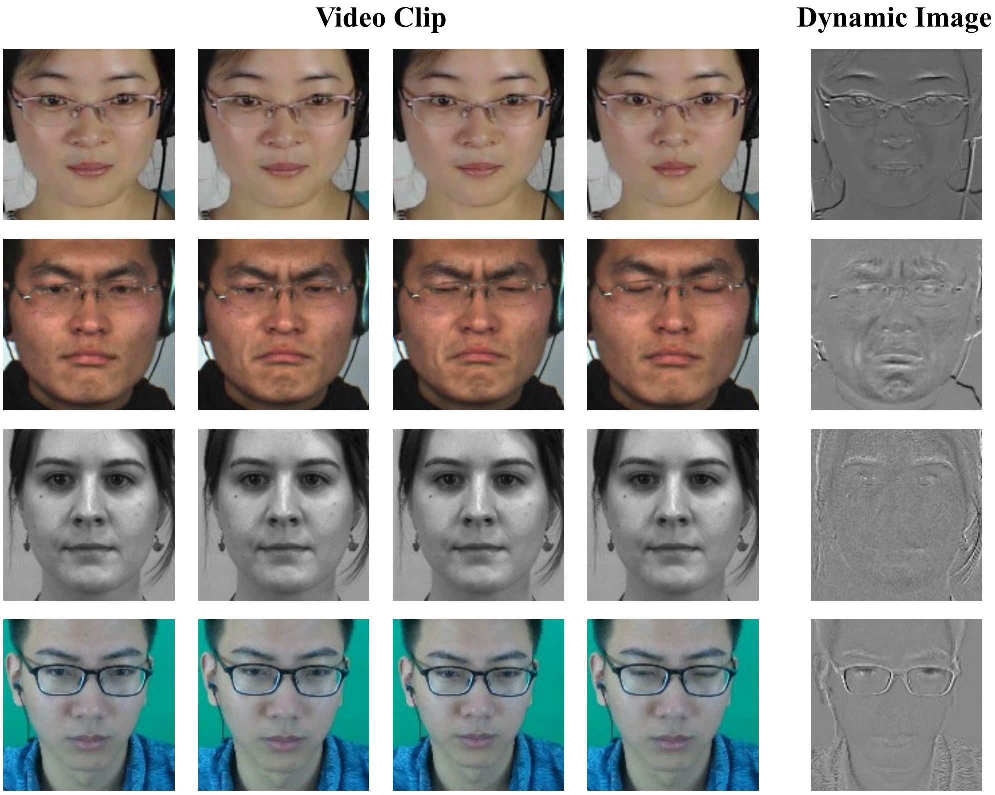


Fig. 1. Illustration of dynamic images [4] for several example video clips. Each row shows four sample frames of a ME video clip as well as the generated dynamic image. *The overall facial appearances and the highlighted motion areas can be observed from the dynamic images.*

The main contributions of this work are threefold:

• We propose a novel fine-grained dynamic perception framework with MER and dynamic image construction, which does not depend on pre-extracted hand-crafted features and key frames. The use of dynamic image construction task contributes to capturing facial subtle muscle actions related to MEs, which relaxes the dependence of our deep network on large-scale training samples.

• We propose a novel local-global feature-aware transformer to capture local subtle information associated with MEs while preserving the global modeling capacity of transformer.

• Extensive experiments on CASME II, SAMM, CAS(ME)^2, and CAS(ME)^3 benchmarks demonstrate that our approach soundly outperforms the state-of-the-art MER methods, and achieves competitive performance for dynamic image construction.

2 Related Work

In this section, we review the previous methods those are closely associated with our approach, including non-deep learning based MER, deep learning based MER, rank pooling and dynamic image, and combination of CNN and vision transformer (ViT).

2.1 Non-Deep Learning Based MER

Since MEs are subtle and hardly distinguishable, earlier methods propose hand-crafted features based on prior knowledge about local characteristics and motion patterns. Zhao *et al.* [74] designed local binary patterns from three orthogonal planes (LBP-TOP) by considering co-occurrence statistics of motions in three directions. Wang *et al.* [60] further proposed local binary patterns with six intersection points (LBP-SIP) to avoid duplicated encoding in LBP-TOP. Ben *et al.* [2] proposed binary face descriptors including dual-cross patterns from three orthogonal planes (DCP-TOP) and hot wheel patterns from three orthogonal planes (HWP-TOP) to encode the discriminative features of ME videos. Another solution of hand-crafted features is based on histogram. Davison *et al.* [8] designed histogram of oriented gradients (HOG), and Li *et al.* [32] further proposed histogram of image gradient orientation (HIGO).

Besides, optical flow describes the motion pattern of each pixel across frames, which has been widely used in MER. Davison *et al.* [36] proposed bi-weighted oriented optical flow (Bi-WOOF) by using onset frame and apex frame to represent a ME. Happy *et al.* [16] developed histogram of oriented optical flow (HOOF) [5] to FHOOF by using fuzzy membership function to collect motion directions, and further developed FHOOF to fuzzy histogram of optical flow orientations (FHOFO) by ignoring subtle motion magnitudes. Dynamic image [4] is another newer way to encode facial motion information of a video, which has been introduced to MER [43, 59].

However, these hand-crafted features only focus on partial characteristics associated with MEs, and often additionally rely on key frames of MEs.

2.2 Deep Learning Based MER

Considering the power of deep neural networks [50, 52, 54], Reddy *et al.* [48] introduced a 3D CNN to capture spatial and temporal information from raw image sequences for MER. Wei *et al.* [62] proposed an attention-based magnification-adaptive network (AMAN) to magnify and focus on ME details. Since subtle MEs are hard to capture, some methods adopt correlated tasks to facilitate MER. Xie *et al.* [66] proposed an AU-assisted graph attention convolutional network (AU-GACN) to reason the relationships among AUs so as to assist the recognition of MEs. Xia *et al.* [63] introduced macro-expression recognition as an auxiliary task, and used adversarial learning to align the feature distributions between macro-expressions and MEs.

Since current deep networks suffer from small-scale and low-diversity ME datasets, other approaches combine hand-crafted features with deep learning. Hu *et al.* [19] incorporated local Gabor binary pattern from three orthogonal panels (LGBP-TOP) features and CNN features, and then trained MER by treating the classification of each ME category as a one-against-all classification problem. Liu *et al.* [37] extracted TV-L1 optical flow between key frames to input into a pre-trained ShuffleNet and then conducted classification via support vector machine (SVM). Verma *et al.* [59] first extracted the dynamic image of the input ME video, and then fed it into a lateral accretive hybrid network (LEARNet). Shao *et al.* [49] generated the optical flow between onset and apex frames of the input video, then input horizontal and vertical optical flow components to a dual-inception network, and finally jointly train MER and AU recognition based on a transformer [58].

All these methods suffer from insufficient training data, or dependence on hand-crafted features. In contrast, our method put MER and dynamic image construction into a joint learning framework, in which raw images are handled, and the auxiliary task alleviate the requirement of large-scale training data.

2.3 Rank Pooling and Dynamic Image

Rank pooling [13] is a video representation technique used in video analysis to aggregate information over time, typically in action recognition tasks [4, 6]. It aims to summarize a sequence of ranked frames into a single representative feature vector. The process involves ranking frames, temporal pooling and optimization. Compared to rank pooling, dynamic image [4] summarizes a whole video into a single image, which synthesizes a static representation that captures the motion dynamics. In the previous work [59], dynamic image is employed as a pre-extracted feature to directly input into deep networks.

Inspired by the above works, we design FDP using the same process as rank pooling, and further construct the dynamic image using rank features. In our approach, subtle ME actions are handled by learning video-wide temporal evolution, which includes ranking the frames in temporal dimension and constructing the dynamic image.

2.4 Combination of CNN and ViT

In the past few years, CNN and ViT have achieved great performance successively in many vision tasks [1, 3, 17, 34], in which the former works well in modeling local relationships and the latter is good at extracting global features. However, pure convolution struggles to capture long-range dependencies due to the limited reception field, and vanilla ViT that relies on self-attention mechanism is inefficient to encode low-level features.

Recently, hybrid structures of CNNs and ViTs are designed to improve the representation ability, in which local and global information are simultaneously focused while their respective weaknesses are avoided. Liu *et al.* [39] proposed a ConvTransformer with multi-head convolutional self-attention layers, to achieve video frame sequence learning and video frame synthesis. Yuan *et al.* [71] combined the advantages of CNN and transformer, in which the former works well in extracting low-level features and strengthening locality, and the latter can establish long-range dependencies by extracting patches from low-level features and can promote the correlations among neighboring tokens in the spatial dimension. Srinivas *et al.* [55] replaced the vanilla convolution with multi-head self-attention in the last several blocks of ResNet [17]. Guo *et al.* [15] proposed a CMT network by inheriting the merits of CNN and ViT, which is composed of depthwise convolutions with local perception units and a light-weight transformer module.

In our work, we integrate the merits of CNNs and ViTs by designing a local-global feature-aware transformer with local relational aggregator and global relational aggregator. Due to the capture of long-range dependencies and local information, our method is effective at modeling transient, subtle, and dynamic MEs.

3 Rank Pooling Inspired Micro-Expression Recognition and Dynamic Image Construction

3.1 Fine-Grained Dynamic Perception Framework

Given a video clip with t frames $\{\mathbf{I}_0, \mathbf{I}_1, \cdots, \mathbf{I}_{t-1}\}$, we first obtain single frame representation $\mathbf{F}_k$ of the k-th frame $\mathbf{I}_k$ in the input video, respectively. Then, a rank scorer $\mathbf{u}$ is employed to rate each single frame representation, in which the later frames are expected to receive higher scores according to the temporal order. Afterwards, the sequence of single frame representation $\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\}$ is fed into a temporal pooling module to learn video-wide dynamic representation $\mathbf{F}^{(d)}$. Finally, $\mathbf{F}^{(d)}$ is shared by MER module and dynamic image construction module for joint learning. Fig. 2 shows the overview of our framework, and Algorithm 1 shows the detailed processes.

Our FDP directly processes raw frame images without requiring key frames, and MER and dynamic image construction can contribute to each other in our joint learning framework.

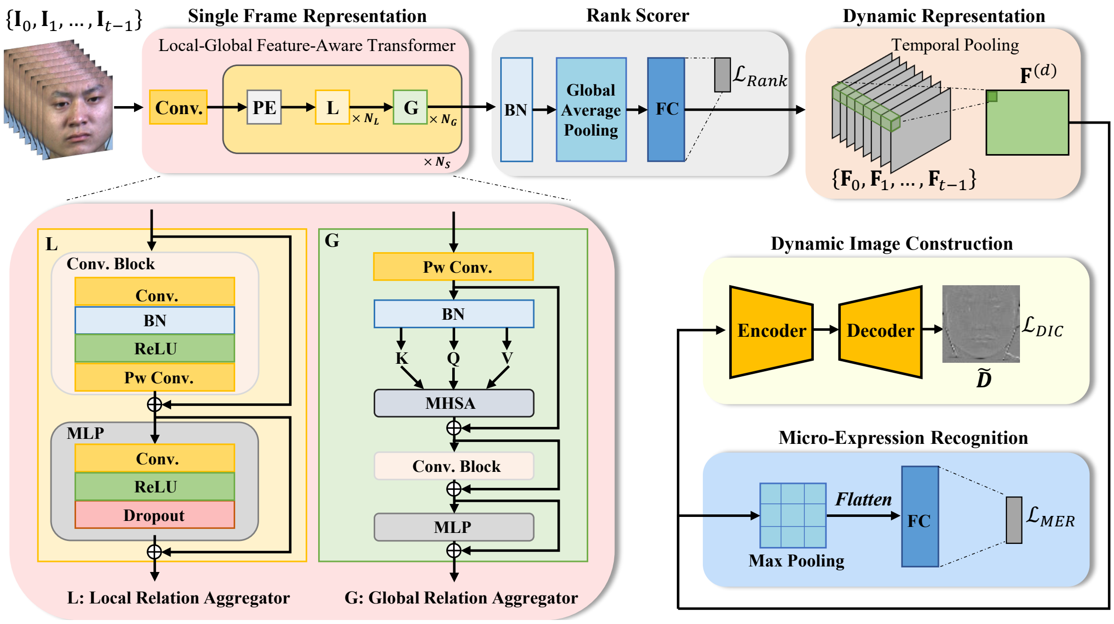


Fig. 2. The architecture of our FDP. Given a sequence of t frames $\{\mathbf{I}_0, \mathbf{I}_1, \cdots, \mathbf{I}_{t-1}\}$, FDP first extracts local-global feature $\mathbf{F}_k$ of each frame $\mathbf{I}_k$ by our proposed local-global feature-aware transformer. Then, the local-global features $\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\}$ are input to a fully connection layer based rank scorer to obtain the rank score of each frame, respectively. Afterwards, the sequence of local-global features $\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\}$ is fed into a 3D convolutional layer to extract video-wide dynamic representation $\mathbf{F}^{(d)}$. Finally, $\mathbf{F}^{(d)}$ is fed into MER module and dynamic image construction module to estimate the ME category and the dynamic image $\widehat{\mathbf{D}}$, respectively.

Algorithm 1 The detailed processes of our FDP framework.

```
Input: A video clip $\{\mathbf{I}_0, \mathbf{I}_1, \cdots, \mathbf{I}_{t-1}\}$.
Output: The predicted ME category $\hat{c}$ and dynamic image $\widehat{\mathbf{D}}$.
Define single frame representation module as $\mathcal{F}$.
Define rank scorer as $\mathcal{R}$.
Define dynamic representation module as $\mathcal{D}$.
Define dynamic image construction module as $\mathcal{C}$.
Define micro-expression recognition module as $\mathcal{M}$.
for each $k \in \{0, 1, \cdots, t-1\}$ do
    Single frame representation $\mathbf{F}_k = \mathcal{F}(\mathbf{I}_k)$.
end for
Rank loss $\mathcal{L}_{Rank} = \mathcal{R}(\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\})$, only used for training.
Dynamic representation $\mathbf{F}^{(d)} = \mathcal{D}(\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\})$.
Predict ME category probabilities $\{\hat{p}_0, \cdots, \hat{p}_{m-1}\} = \mathcal{M}(\mathbf{F}^{(d)})$, and obtain $\hat{c} = \arg\max_{j \in \{0,1,\cdots,m-1\}} \hat{p}_j$.
$\widehat{\mathbf{D}} = \mathcal{C}(\mathbf{F}^{(d)})$.
Return $\hat{c}$ and $\widehat{\mathbf{D}}$.
```

3.2 Local-Global Feature-Aware Transformer

Inspired by [12] that conducting local and global features simultaneously for facial expression recognition, we design a local-global feature extractor. To learn high-quality frame representations while reducing the effect of noise and violent abrupt variations, we introduce local-global feature-aware transformer. The goal of our proposed local-global feature-aware transformer is to capture local correlated information while modeling global dependencies. It consists of a vanilla convolutional layer, a stack of N_S local-global relational aggregators, and a head block. Each local-global relational aggregator is a hybrid structure of CNN and ViT, which contains a patch embedding (PE) layer [10], a stack of N_L CNN based local relational aggregators, and a stack of N_G ViT based global relational aggregators. The head block is composed of a batch normalization (BN) layer [21]and a global average pooling layer [34] to extract the final local-global feature.

To enable the input of the first local-global relational aggregator, the patch embedding is obtained by extracting patches from the feature map of the vanilla convolutional layer. The subsequent patch embeddings are obtained from the output of the previous local-global relational aggregator. We will elaborate the local relational aggregator and the global relational aggregator in the following sections.

3.2.1 *Local Relational Aggregator.*

We design a local relational aggregator based on CNN while incorporating the paradigm of transformer [58]. Specifically, an input $\mathbf{X}^l = (\mathbf{X}^l_0, \mathbf{X}^l_1, \cdots, \mathbf{X}^l_{h-1})$ with h heads in channel dimension first goes through a multi-head convolution block. The i-th input $\mathbf{X}^l_i$ is fed into the i-th single-head convolutional block, which consists of a vanilla convolutional layer, a BN layer, and a rectified linear unit (ReLU) layer [42]. Then, the outputs of h heads are concatenated and are further interacted by a pointwise convolutional layer [20]. A residual structure with skip connection [17] is then utilized to suppress the vanishing gradient problem. Finally, a multilayer perceptron (MLP) layer with another residual structure is applied to obtain local feature.

Our proposed local relational aggregator inherits the advantage of convolution that can aggregate contexts in local regions with efficient computations, in which local token affinity is captured with a small amount of parameters.

3.2.2 *Global Relational Aggregator.*

Besides the capture of local details, it is also important to exploit global correlations in the broader token space. The architecture of our proposed global relational aggregator is illustrated in the bottom of Fig. 2. It is composed of a pointwise convolutional layer, a multi-head self-attention block [58], a multi-head convolution block, and a MLP layer, in which three residual structures are adopted.

Denote the input of the multi-head self-attention block be $\mathbf{X}^g = (\mathbf{X}^g_0, \mathbf{X}^g_1, \cdots, \mathbf{X}^g_{h-1})$. For the i-th head, we first calculate the queries $\mathbf{Q}_i$, keys $\mathbf{K}_i$, and values $\mathbf{V}_i$ as

$$\mathbf{Q}_i = \mathbf{W}_{\mathbf{Q}_i}\mathbf{X}^g_i, \tag{1a}$$

$$\mathbf{K}_i = \mathbf{W}_{\mathbf{K}_i}\mathbf{X}^g_i, \tag{1b}$$

$$\mathbf{V}_i = \mathbf{W}_{\mathbf{V}_i}\mathbf{X}^g_i, \tag{1c}$$

where $\mathbf{W}_{\mathbf{Q}_i}$, $\mathbf{W}_{\mathbf{K}_i}$, and $\mathbf{W}_{\mathbf{V}_i}$ are learnable weight matrices. To map $\mathbf{Q}_i$ and $\mathbf{K}_i$-$\mathbf{V}_i$ pair to a new output, the self-attention is defined as

$$\mathbf{A}_i = \sigma(\frac{\mathbf{Q}_i{\mathbf{K}_i}^T}{\sqrt{dim}})\mathbf{V}_i, \tag{2}$$

where $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ have the same channel dimension dim, $\frac{1}{\sqrt{dim}}$ is adopted to scale the dot product, and $\sigma(\cdot)$ is a Softmax function for weighted summing of the values $\mathbf{V}_i$. Then, a feed forward network is applied to each spatial position for further encoding:

$$\mathbf{Y}^g_i = FC(\mathbf{A}_i), \tag{3}$$

where $FC(\cdot)$ denotes a fully-connected layer, and one dropout layer [56] following the fully-connected layer is omitted.

The outputs of multiple heads are further fused to be the final output of the multi-head self-attention block. This block and the multi-head convolution block are cooperated to capture global dependencies, and the MLP layer is used to extract the final feature.

Our global relational aggregator can adaptively model long-range dependencies from distant regions by inheriting the self-attention [58] paradigm. By progressively stacking local and global relational aggregators, our local-global feature-aware transformer with merits of transformer and convolution can extract complete local-global feature.

3.3 Rank Scorer and Temporal Pooling

A video is ordered sequences of frames, where the frame order also dictates the evolution of the frame appearances [13]. To model the latent evolution information in frames, we introduce a linear function based rank scorer $\mathbf{u}$. Considering a pair of independent frame representations $\mathbf{F}_i$ and $\mathbf{F}_j$, we aim to learn $\mathbf{u}$ such that $i < j \Leftrightarrow S(\mathbf{F}_i) < S(\mathbf{F}_j)$. $S(\mathbf{x})$ denotes the rank score, which is defined as

$$S(\mathbf{x}) = \mathbf{u}^T \cdot \mathbf{x}. \tag{4}$$

Our optimization goal is to make the rank score increase in chronological order. Thus, the rank loss is defined as

$$\mathcal{L}_{Rank} = \sum_{k=0}^{t-1} |K(k) - S(\mathbf{F}_k)|, \tag{5}$$

where $K(\cdot)$ denotes direct proportionality function with a positive parameter. Simultaneously, the frame representations $\{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_{t-1}\}$ are reshaped into two-dimensional feature maps and concatenated in chronological order. Then, the temporal pooling achieved through 3D convolution is applied to obtain the dynamic representation $\mathbf{F}^{(d)}$.

3.4 Joint Learning of Tasks

3.4.1 Dynamic Image Construction. The dynamic image summarizes the appearances and dynamics of a whole video as one image. The introduction of the dynamic image construction task allows our framework to better extract dynamic features in a ME video, so as to facilitate the performance of MER.

The detailed structure of the dynamic image construction module is shown in Fig. 3. It contains an encoder network and a decoder network, which is a fully convolutional network without fully-connected layers. This design of full convolution is beneficial for element-wise prediction. Its end is a Sigmoid layer to produce the estimated single-channel dynamic image $\widehat{\mathbf{D}}$, in which each element in the output of the decoder network is mapped into $(0, 1)$ interval.

The encoder network consists of four consecutive encoder blocks, each of which is composed by a convolutional layer and a max-pooling layer. Particularly, the former is followed by a BN layer and a leaky ReLU layer [40]. The decoder network with four decoder blocks is the counterpart of the encoder network. Each decoder block contains a deconvolutional layer and a convolutional layer followed by BN and leaky ReLU. The deconvolutional layer is used to upsample feature maps, which is treated as the inverse process of the max-pooling layer. The skip connections [17] between encoder blocks and decoder blocks are beneficial for exploiting encoder information in the decoding process and accelerating the training procedure.

The dynamic image construction loss is defined as

$$\mathcal{L}_{DIC} = MSE(\widehat{\mathbf{D}}, \mathbf{D}), \tag{6}$$

where $\mathbf{D}$ denotes the ground-truth dynamic image of input video clip $\{\mathbf{I}_0, \mathbf{I}_1, \cdots, \mathbf{I}_{t-1}\}$, and $MSE(\cdot)$ denotes mean squared error (MSE) loss.

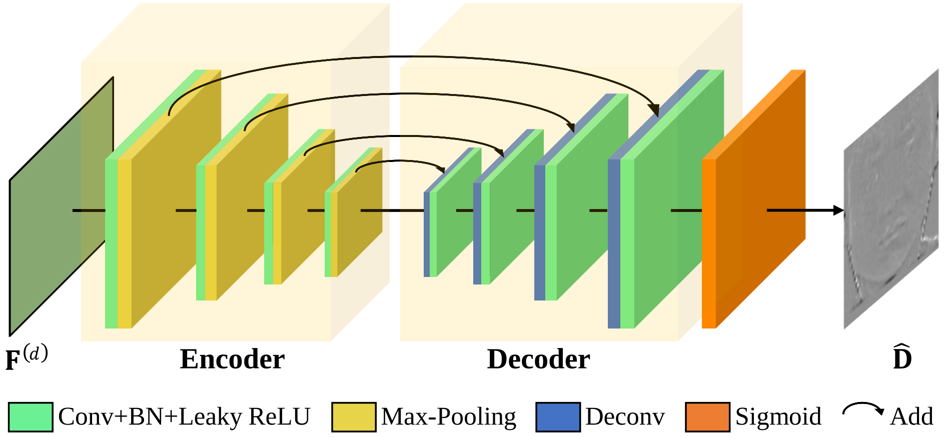


Fig. 3. The structure of dynamic image construction module. It is an encoder-decoder as a fully convolutional network without fully-connected layers. The curved arrow denotes skip connection, in which the encoder feature map is element-wise added to the decoder feature map.

3.4.2 *Micro-Expression Recognition.*

The overall architecture of the MER module is illustrated in the right side of Fig. 2. It consists of a max-pooling layer and two fully-connected layers, in which the former is utilized to reduce the dimensions of the dynamic representation $F^{(d)}$ while maintaining important information, and the latter is used for ME classification. The MER loss is defined as a cross-entropy loss:

$$\mathcal{L}_{MER} = -\sum_{j=0}^{m-1} p_j \log(\hat{p}_j), \tag{7}$$

where m denotes the number of ME categories, and $\hat{p}_j$ denotes the predicted probability that the video sample is in the j-th category. p_j denotes the ground-truth probability, which is 1 if the video sample is in the j-th category and is 0 otherwise.

3.4.3 *Full Loss.*

In our joint learning framework, the full loss is combined by $\mathcal{L}_{MER}$, $\mathcal{L}_{DIC}$ and $\mathcal{L}_{Rank}$:

$$\mathcal{L} = \mathcal{L}_{MER} + \lambda_d \mathcal{L}_{DIC} + \lambda_r \mathcal{L}_{Rank}, \tag{8}$$

where λ_d and λ_r are parameters to weigh the importance of MER, dynamic image construction and rank tasks.

4 Experiments

4.1 Datasets and Settings

4.1.1 *Datasets.*

We evaluate our method on four popular spontaneous ME datasets, including CASME II [68], SAMM [7], CAS(ME)^2 [47], and CAS(ME)^3 [31].

- **CASME II** consists of 255 videos captured from 26 subjects in steady and high-intensity illumination. To elicit the MEs, subjects are induced to experience a high arousal with motivations to disguise. Each video is recorded with the frame rate of 200 frames per second (FPS) and the frame size of 280 × 340. The average duration of MEs is 66.21 frames. Following the previous methods [30, 63], we use ME categories of happiness, disgust, repression, surprise, and others for five-classes evaluation, and use ME categories of positive, negative, and surprise for three-classes evaluation.

Table 1. The number of videos for each ME category in CASME II [68], SAMM [7], and CAS(ME)^3 [31]. The used five categories of CASME II and SAMM, as well as used seven categories of CAS(ME)^3 are highlighted with its number in bold. "-" denotes this category is not included.

Dataset Class	CASME II	SAMM	CAS(ME)^3
Happiness	**32**	**26**	**64**
Anger	-	**57**	**70**
Contempt	-	**12**	-
Disgust	**63**	9	**281**
Fear	2	8	**93**
Repression	**27**	-	-
Surprise	**28**	**15**	**201**
Sadness	4	6	**64**
Others	**99**	**26**	**170**

Table 2. The number of videos for each ME category in CASME II [68],and SAMM [7], in terms of three-classes evaluation, as well as CAS(ME)^2 [47] in terms of four-classes evaluation.

Dataset Class	CASME II	SAMM	CAS(ME)^2
Positive	32	26	8
Surprise	28	15	9
Negative	96	92	21
Others	-	-	19

• **SAMM** includes 159 videos at 200 FPS from 29 subjects, which are collected using gray-scale cameras in constrained lighting conditions without flickering. The MEs are elicited from stimuli tailored to each subject. The average duration of MEs is 74.31 frames. Similar to the previous works [30, 63], we select ME categories of happiness, anger, contempt, surprise, and others for five-classes evaluation, and select ME categories of positive, negative, and surprise for three-classes evaluation.

• **CAS(ME)^2** contains 87 long videos, each of which includes spontaneous macro-expressions or MEs. These videos are further cropped as 300 macro-expression video clips and 57 ME video clips. The average duration of MEs is 12.58 frames. We only evaluate on the ME video clips, in terms of four classes (positive, negative, surprise, and others).

• **CAS(ME)^3** provides $1,109$ MEs and $3,490$ macro-expressions from 100 subjects, in which each subject is asked to watch 13 emotionally stimuli and keep their faces expressionless. The recorded videos have the resolution of $1,280 \times 720$. The average duration of MEs is 28.61 frames. We conduct experiments on 943 ME videos from Part A, with seven categories (happiness, disgust, surprise, anger, fear, sadness and others).

All ME video clips in these datasets are labeled. The number of samples for each ME category are summarized in Table 1 and Table 2, and the attributes of each dataset are shown in Table 3. The dynamic image of each video in these datasets is generated using [4] as the ground-truth annotation.

4.1.2 Evaluation Metrics. Similar to most previous works [30, 63], leave-one-subject-out (LOSO) cross-validation is applied in the single dataset evaluation, in which each subject is taken as the test set in turn while the remaining subjects are taken as the training set. We report popular metrics, including accuracy (Acc) and F1-score (F1) for CASME

Table 3. The attributes of CASME II [68], SAMM [7], CAS(ME)^2 [47], and CAS(ME)^3 [31].

Dataset / Attribute	CASME II	SAMM	CAS(ME)^2	CAS(ME)^3
Number of Subject	26	29	22	100
Frames Per Second	200	200	30	30
Number of ME Samples	255	159	57	1109
Average Duration (frames)	66.21	74.31	12.58	28.61

II, SAMM, and CAS(ME)^2, as well as F1 and unweighted average recall (UAR) for CAS(ME)^3. UAR is defined as

$$UAR = \frac{1}{m} \sum_{j=0}^{m-1} \frac{TP_j}{TP_j + FN_j}, \tag{9}$$

where m is the total number of ME categories, and TP_j, FP_j, and FN_j denote the number of true positives, false positives, and false negatives for the j-th category, respectively.

To investigate the generalization ability of our method, we also perform a cross-dataset evaluation. We conduct a two-fold cross-validation on CASME II and SAMM datasets, in which one dataset is used for training while the other dataset is used for testing. Following the settings in previous approaches [24, 38, 74], we report two metrics of weighted average recall (WAR) and UAR. WAR is defined as

$$WAR = \sum_{j=0}^{m-1} \frac{TP_j}{N}, \tag{10}$$

where N denotes the total number of samples.

In the following sections, Acc, F1, WAR, and UAR results are all reported in percentages, in which % is omitted for simplicity.

4.1.3 Implementation Details. In our experiments, we extract a video clip with t frames as the input of our FDP by uniformly-space sampling from the raw video. Each frame image is cropped with few background regions and is aligned to $3\times72\times72$ via similarity transformation, in which facial shape is preserved without changing the ME. During training, each image is randomly cropped into $3 \times 64 \times 64$ and is further horizontally flipped to improve the diversity of training data. During testing, each image is centrally cropped into $3 \times 64 \times 64$ so as to be consistent with the training input size.

Our FDP is implemented based on PyTorch [44], with a solver of Adam [25], an initial learning rate of 1×10^{-4}, and a mini-batch size of 36. The number of frames in the input video clip is set as $t = 8$, in which each clip is uniformly-spaced sampled from the raw video with a random offset. The trade-off parameter λ_d and λ_r are set to 100 and 0.1, respectively. The structure parameters of local-global feature-aware transformer are set as: $N_S = 4$, $N_L = 2$, and $N_G = 1$. All the experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. FDP takes about 5.8 GB GPU memory for about 3.5 hours during training, which demonstrates light-weight transformer structure in our method.

4.2 Comparison with State-of-the-Art Methods

We compare our FDP with state-of-the-art MER methods under the same evaluation setting. These methods can be classified into non-deep learning (NDL) based methods and deep learning (DL) based methods. The latter can be further

Table 4. Comparison with state-of-the-art methods on CASME II [68] and SAMM [7] for five categories. DL, NDL, PF, RI, and KF denote deep learning based methods, non-deep learning based methods, pre-extracted hand-crafted features, raw images, and key frames, respectively. "-" denotes the result is not reported in its paper. The best results are highlighted in bold, and the second best results are highlighted by an underline.

Method	Paper	Type	CASME II		SAMM	
			Acc	F1	Acc	F1
SparseSampling [28]	TAFFC'17	NDL	49.00	51.00	-	-
Bi-WOOF [36]	SPIC'18	NDL+KF	58.85	61.00	-	-
HIGO+Mag [32]	TAFFC'18	NDL	67.21	-	-	-
FHOFO [16]	TAFFC'19	NDL	56.64	52.48	-	-
DSSN [23]	ICIP'19	DL+PF+KF	70.78	72.97	57.35	46.44
Graph-TCN [30]	MM'20	DL+RI+KF	73.98	72.46	75.00	69.85
MicroNet [64]	MM'20	DL+RI+KF	75.60	70.10	74.10	73.60
LGCcon [33]	TIP'21	DL+PF	62.14	60.00	35.29	23.00
AU-GCN [29]	CVPRW'21	DL+PF+KF	74.27	70.47	74.26	70.45
GACNN [27]	CVPRW'21	DL+PF	81.30	70.90	**88.24**	82.79
GEME [43]	NeuCom'21	DL+PF	75.20	73.54	55.88	45.38
MERSiamC3D [75]	NeuCom'21	DL+PF+KF	81.89	83.00	68.75	64.00
MiNet&MaNet [63]	IJCAI'21	DL+RI	79.90	75.90	76.70	76.40
MER-Supcon [76]	PRL'22	DL+PF+KF	73.58	72.86	67.65	62.51
AMAN [62]	ICASSP'22	DL+RI	75.40	71.25	68.85	66.82
SLSTT [73]	TAFFC'22	DL+PF	75.81	75.30	72.39	64.00
Dynamic [57]	TAFFC'22	DL+RI+KF	72.61	67.00	-	-
I^{2}Transformer [49]	APIN'23	DL+PF+KF	74.26	77.11	68.91	73.01
FDP	Ours	DL+RI	**88.42**	**87.05**	86.69	**85.29**

categorized into pre-extracted feature (PF) based methods and raw image (RI) based methods, according to the type of network input.

In particular, NDL based methods include LBP-TOP [74], 3DHOG [46], MDMO [38], SparseSampling [28], Bi-WOOF [36], HIGO+Mag [32], and FHOFO [16]. DL+PF based methods include AlexNet [26], Khor *et al.* [24], DSSN [23], STSTNet [35], RCN [65], LGCcon [33], AU-GCN [29], GACNN [27], GEME [43], MERSiamC3D [75], MER-Supcon [76], SLSTT [73], FR [78], HTNet [61], and I^{2}Transformer [49]. DL+RI based methods include Peng *et al.* [45], Graph-TCN [30], MicroNet [64], MiNet&MaNet [63], AMAN [62], and Dynamic [57]. Besides, some of these methods rely on key frames (KF) of MEs, or employ outside training data such as macro-expression datasets.

4.2.1 Single Dataset Evaluation. Table 4 and Table 5 show the comparison results on single datasets of CAMSE II and SAMM for five categories and three categories, respectively. It can be seen that DL based methods often outperform NDL based methods, which proves the power of deep networks. Note that some recent state-of-the-art methods like GACNN [27] achieve excellent results. This is mainly because these methods rely on auxiliary information such as hand-crafted features and key frames, which assist them to capture ME related information. In contrast, our FDP is significantly better on most evaluations by directly processing raw images. Besides, compared to the methods like MiNet&MaNet using additional macro-expression datasets, FDP performs better with only benchmark training samples.

Moreover, we evaluate our method on more challenging datasets CAS(ME)^2 and CAS(ME)^3 in Table 6 and Table 7, respectively. Note that CAS(ME)^2 and CAS(ME)^3 datasets exhibit more varieties than CASME II and SAMM datasets.

Table 5. Comparison with state-of-the-art methods on CASME II [68] and SAMM [7] for three categories. The best results are highlighted in bold.

Method	Paper	Type	CASME II		SAMM	
			Acc	F1	Acc	F1
OFF-ApexNet [14]	SPIC'19	DL+PF+KF	88.28	86.97	68.18	54.23
AU-GACN [66]	MM'20	DL+RI	71.20	35.50	70.20	43.30
GACNN [27]	CVPRW'21	DL+PF	89.66	86.95	88.72	81.18
MER-Supcon [76]	PRL'22	DL+PF+KF	89.65	88.06	81.20	71.25
FDP	Ours	DF+RI	**92.72**	**90.71**	**91.25**	**86.67**

Table 6. Comparison with state-of-the-art methods on CAS(ME)^2 [47]. The reported results of LBP-TOP are from [47], and other methods are implemented using its released code. The best results are highlighted in bold.

Method	Paper	Type	Acc	F1
LBP-TOP [74]	TPAMI'07	NDL	40.95	-
MicroExpSTCNN [48]	IJCNN'19	DL+RI	67.35	54.43
AU-GCN [29]	CVPRW'21	DL+PF+KF	69.38	65.21
SLSTT [73]	TAFFC'22	DL+PF	75.51	73.98
FDP	Ours	DL+RI	**83.67**	**81.69**

Table 7. Comparison with state-of-the-art methods on CAS(ME)^3 [31].The results of previous methods except for HTNet [61] are reported by [31]. The best results are highlighted in bold

Method	Paper	Type	F1	UAR
AlexNet [26]	NeurIPS'12	DL+KF	25.70	26.34
STSTNet [35]	FG'19	DL+PF+KF	37.95	37.92
RCN [65]	TIP'20	DL+PF+KF	39.28	38.93
FR [78]	PR'22	DL+PF+KF	34.93	34.13
HTNet [61]	arXiv'23	DL+PF+KF	57.67	54.15
FDP	Ours	DL+RI	**59.78**	**57.84**

The recent released CAS(ME)^3 has the largest number of samples. Compared with CAMSE II and SAMM, it has an abundant number of samples in all seven categories, all of which can be used for training. Meanwhile, CAS(ME)^3 is also the most challenging one because its data contains more noise compared to other datasets. In this challenging case, our FDP still significantly outperforms other methods.

It can be observed that our FDP achieves the overall best performance across datasets with varying categories, scales, and noise levels. Specifically, FDP processes raw video sequences, requiring no category-specific adjustments or auxiliary data. This confirms its applicability to any ME-containing video. In addition, by directly processing raw images without dependencies on key frames, pre-extracted features, or external datasets, FDP eliminates pre-processing dependencies. This enables deployment in real-world scenarios where prior information is inaccessible. Therefore, our FDP is a practical MER solution.

4.2.2 Cross-Dataset Evaluation. Table 8 presents the cross-dataset evaluation results. The common three ME categories of happiness, surprise, and others for the two datasets are used. It can be observed that our approach achieves the best

Table 8. WAR and UAR results for three ME categories (happiness, surprise, and others) of cross-dataset evaluations. Avg. denotes the average results of two cross-dataset evaluations. The results of methods except for I^2Transformer [49] are reported by [70]. CASME II→SAMM denotes training on CASME II and testing on SAMM. The best results are highlighted in bold, and the second best results are highlighted by an underline.

Method	Paper	Type	CASME II→SAMM		SAMM→CASME II		Avg.	
			WAR	UAR	WAR	UAR	WAR	UAR
LBP-TOP [74]	TPAMI'07	NDL	33.8	32.7	23.2	31.6	28.5	32.2
3DHOG [46]	ICDP'09	NDL	35.3	26.9	37.3	18.7	36.3	22.8
MDMO [38]	TAFFC'16	NDL	44.1	34.9	26.5	34.6	35.3	34.8
Peng *et al.* [45]	FG'18	DL+RI+KF	48.5	38.2	38.4	32.2	43.5	35.2
Khor *et al.* [24]	FG'18	DL+PF+KF	54.4	44.0	57.8	33.7	56.1	38.9
I^2Transformer [49]	APIN'23	DL+PF+KF	51.2	-	**66.2**	-	58.7	-
FDP	Ours	DL+RI	**58.2**	**51.8**	62.2	**56.0**	**60.2**	**53.9**

average performance especially for the UAR metric, which demonstrates the strong generalization ability of our FDP. This can be attributed to two merits of our method. First, our proposed local-global feature-aware transformer has strong capacities of relational reasoning and feature learning by simultaneously modeling local and global contexts. Second, the joint learning with dynamic image construction is beneficial for extracting ME related features, and thus improves the robustness on unseen samples.

4.2.3 Systematic Discussion and Structured Gap Analysis. The above results demonstrate that our FDP outperforms state-of-the-art MER methods in terms of both single dataset evaluation and cross-dataset evaluation. There are two main limitations in previous MER methods:

- **Dependency on Auxiliary Inputs and Pre-Processing**

Pre-extracted Feature (PF) Reliance: Top-performing methods like GACNN [27], STSTNet [35], RCN [65], and FR [78] rely on pre-extracted optical flow or other hand-crafted features. This introduces complexity and sensitivity to the quality of feature extraction techniques. The need for separate feature computation hinders end-to-end learning and real-time applicability.

Key Frame (KF) Reliance: Previous methods such as DSSN [23], AU-GCN [29], MERSiamC3D [75], MER-Supcon [76], and RCN [65] require accurate detection of key frames. This dependency is problematic when key frames are not provided or not detected correctly in real scenarios, leading to cumulative errors and limited robustness.

External Data Dependency: Existing approaches like MiNet&MaNet [63] utilize additional macro-expression datasets for training. This reduces practicality, as such data may be not readily available or directly relevant.

- **Limited Generalization and Robustness**

Dataset Sensitivity: The performances of some methods vary across datasets and category amounts. For instance, GACNN [27] excels on SAMM with five categories but drops significantly on CASME II with five categories. This indicates overfitting to specific dataset characteristics or evaluation benchmarks.

Category Scalability: Existing methods often struggle when the number of ME categories increases. Their performances generally degrade in five-classes evaluation (see Table 4) compared to three-classes evaluation (see Table 5), showing the limitations in feature discriminability and model capacity for recognizing fine-grained categories.

Table 9. Acc and F1 results for different variants of FDP on SAMM [7] in terms of five categories. L: local relational aggregator; G: global relational aggregator. The best results are highlighted in bold.

Method	Acc	F1
FDP	**86.69**	**85.29**
FDP w/o $\mathcal{L}_{DIC}$	83.98	80.40
FDP w/o $\mathcal{L}_{Rank}$	83.54	80.02
FDP w/o L	81.79	79.78
FDP w/o G	78.54	75.32
FDP w/o L&G	60.70	59.05

Noise Vulnerability: The performances often degrade on challenging datasets like CAS(ME)^3, where noises are prevalent. Some methods relying on key frames or hand-crafted features are particularly susceptible, demonstrating poor noise robustness.

Existing methods exhibit critical gaps in practicality, robustness, and generalization. Their performances usually rely on pre-processing, external data, or specific dataset conditions. In contrast, our FDP overcomes these limitations, and the results from Tables 4 to 8 demonstrate the superiority.

4.3 Ablation Study

In this section, we conduct ablation experiments to investigate the effectiveness of dynamic image construction module, rank scorer, local-global feature-aware transformer, and backbone structure on MER. The results of different variants of FDP are shown in Table 9. These experiments are all evaluated on SAMM dataset in terms of five categories.

4.3.1 Dynamic Image Construction. Compared with the FDP, the performance of FDP w/o $\mathcal{L}_{DIC}$ is degraded after removing the dynamic image construction module. This demonstrates that dynamic image construction task in our joint learning framework contributes to MER. The estimation of dynamic image can guide the dynamic representation $\mathbf{F}^{(d)}$ shared by the MER module to capture spatial appearances and temporal patterns.

4.3.2 Rank Scorer. When removing $\mathcal{L}_{Rank}$ of the FDP, the Acc and F1 results of FDP w/o $\mathcal{L}_{Rank}$ are decreased to 83.54 and 80.02, respectively, which shows the effectiveness of rank scorer. This is mainly because the supervision of rank scorer can enhance the model's understanding on the evolution of micro-expression actions.

4.3.3 Local-Global Feature-Aware Transformer. Here we evaluate the main components of local-global feature-aware transformer, including local relational aggregator and global relational aggregator. When removing both two relational aggregators, the Acc and F1 results of FDP w/o L&G are significantly decreased to 60.70 and 60.15, respectively. If we remain either relational aggregator, the results improve a lot. However, the performance is still worse than FDP. This demonstrates the effectiveness of local-global feature-aware transformer with local-global relational reasoning and feature learning, which largely determines the performance of FDP as the backbone.

4.3.4 Backbone Structure. Table 10 shows the number of parameters of different backbone structures, in which the results are obtained by replacing our proposed local-global feature-aware transformer with new backbone. When directly using the classical vision transformer ViT-Base as the backbone, we obtain low Acc and F1 results with a large

Table 10. SAMM [7] results (five categories) and the number of parameters (#Params.) for different backbone structures of FDP. The best results are highlighted in bold.

Backbone	Type	Acc	F1	#Params.
ViT-Base [9]	T	70.58	67.24	88.39M
Ours	C+T	79.82	74.46	28.71M
Ours	C′+T	83.93	79.21	27.14M
Ours	C*+T	**86.69**	**85.29**	14.82M
ResNet18 [17]	C	76.47	73.25	**13.27M**

T: Transformer
C: Conv.
C′: Conv. + BN + ReLU + Pointwise Conv.
C*: Multi-head Conv.

number of parameters. If using a single vanilla convolutional layer to replace our proposed multi-head convolution block, the performance is significantly improved over ViT-Base. This is because convolution works better than ViT-Base in extracting local features, which demonstrates the importance of local features for MER. When further adding pointwise convolution, the results are better while the number of parameters is slightly decreased. This is attributed to the enhanced feature learning ability and the reduced number of output channels by pointwise convolution.

When changing the single head to multiple heads, our final version of FDP achieves the best performance using the least parameters. This is because our proposed multi-head convolution captures more diverse ME information from the input data by allowing different groups of channels to learn independent features. We also compare with a classical convolutional network ResNet18. Although it requires less parameters, its performance is significantly worse. Compared to typical vision transformers and convolutional networks, our FDP performs better by integrating their both advantages.

We also notice that FDP shows markedly superior training efficiency compared to ViT-Base [9] due to its lightweight hybrid architecture. The multi-head convolution reduces optimization complexity by autonomously capturing diverse spatial representations without manual feature engineering. Besides, the self-stabilizing properties of the pointwise convolution layers reduce the requirement of learning rate scheduling, which substantially diminishes tuning effort. Moreover, FDP exhibits robustness to hyperparameter changing comparing to other variants, demonstrating better hyperparameter insensitivity.

4.3.5 *Weights of Losses.*

As shown in Eq. (8), the full loss is composed of MER loss $\mathcal{L}_{MER}$, dynamic image construction loss $\mathcal{L}_{DIC}$, and rank loss $\mathcal{L}_{Rank}$. Table 11 shows the results of our FDP using different weights of loss terms. The first three rows show that when keeping λ_r unchanged and λ_d increasing, both Acc and F1 results increase. This is because the numerical scale of $\lambda_d \mathcal{L}_{DIC}$ gradually approaches to that of $\mathcal{L}_{MER}$, enabling each loss term to contribute to the optimization process. However, when λ_d increases to 1000, both Acc and F1 results drop significantly. Due to the overly large dynamic image construction loss term, $\mathcal{L}_{MER}$ becomes insignificant in the full loss. When fixing λ_d as 100 and changing λ_r, the similar phenomenon can be found in the last three rows. Therefore, our method performs the best when the magnitude values of loss terms are balanced.

Table 11. Acc and F1 results for our FDP with different loss term weights on SAMM [7] in terms of five categories. The best results are highlighted in bold.

λ_d	λ_r	Acc	F1
1	0.1	83.98	80.40
10	0.1	84.50	81.19
100	**0.1**	**86.69**	**85.29**
1000	0.1	50.73	23.67
100	0.01	83.20	79.73
100	1	77.43	73.89
100	10	51.47	35.17

Table 12. Statistics of Wilcoxon rank-sum test [41] and P-values on benchmark F1 results from Tables 4 to 7.

Benchmark	Statistics	P-value
CASME II (Five Categories)	4.977	3.227e-7
SAMM (Five Categories)	4.243	1.106e-5
CASME II (Three Categories)	2.309	1.046e-2
SAMM (Three Categories)	2.309	1.046e-2
CAS(ME)^2	1.964	2.477e-2
CAS(ME)^3	2.611	4.512e-3

4.4 Significance Test

4.4.1 Statistical Significance between State-of-the-Art Methods and Our Method. To prove the significant superiority of our method to previous methods, we conduct a significance test based on the results from Tables 4 to 7. Considering the results of different methods do not follow a normal distribution, we conduct the Wilcoxon rank-sum test [41]. Specifically, we make the following hypotheses:

- H_0: There is no significant difference between our method and state-of-the-art methods.
- H_1: Our method is significantly superior to state-of-the-art methods.

The test statistics and P-values of the F1 results are shown in Table 12. It can be seen that all P-values are less than 0.05. Therefore, we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. Our FDP significantly outperforms state-of-the-art methods in terms of statistics.

4.4.2 Statistical Significance between Backbones and Our Proposed Modules. To investigate the effectiveness of the proposed modules in our framework from the perspective of statistics, we conduct a significance test based on the results from Tables 9 to 10. We make the following hypotheses:

- H_0: Our framework has no significant effect.
- H_1: Our framework has a significant optimization effect.

The test statistics and P-values of the F1 results are presented in Table 13. As can be seen, all P-values are less than 0.05. Therefore, we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. It is demonstrated that our proposed modules are beneficial for MER.

Table 13. Statistics of Wilcoxon rank-sum test [41] and P-values on ablation F1 results from Tables 9 to 10.

Ablation	Statistics	P-value
Modules	2.611	4.512e-3
Backbones	2.309	1.046e-2

Table 14. Dynamic image construction results (lower is better) for different variants of FDP on SAMM [7]. The best results are highlighted in bold.

Method	Average MSE ($\times 10^{-3}$)
FDP	**1.44**
FDP w/o $\mathcal{L}_{MER}$	1.78
FDP w/o L	6.77
FDP w/o G	6.90
FDP w/o L&G	7.43

4.5 FDP for Dynamic Image Construction

We have validated the contribution of dynamic image construction task to MER in Sec. 4.3. To also investigate the effectiveness of MER task for dynamic image construction, we implement a new baseline FDP w/o $\mathcal{L}_{MER}$. It only achieves dynamic image construction by removing the MER module. Besides, FDP w/o L, FDP w/o G, and FDP w/o L&G are still evaluated to explore the influence of local-global feature-aware transformer on dynamic image construction. We report Average MSE as the evaluation metric, which is computed as the average of MSE between $\mathbf{D}$ and $\widehat{\mathbf{D}}$ over all samples.

Table 14 shows the average MSE on the SAMM benchmark. We can observe that FDP outperforms FDP w/o $\mathcal{L}_{MER}$ with the help of MER. This is attributed to the guidance of the MER task to capture facial subtle muscle actions, which is closely related to the dynamic image. Combining with the observations in Sec. 4.3, it can be concluded that MER and dynamic image construction facilitate each other in our joint learning framework.

Besides, compared with FDP w/o L, FDP w/o G, and FDP w/o L&G, FDP exhibits a large margin. This demonstrates that our local-global feature-aware transformer is a strong backbone network for capturing spatio-temporal clues.

4.6 Visual Results

Fig. 4 visualizes the dynamic image construction results of these methods on several video clip samples from CASME II, CAS(ME)^2, CAS(ME)^3, and SAMM. We can see that our FDP extracts dynamic information from ME videos with the best effects. Its estimations are close to the ground-truth annotations, which demonstrates that the fully convolutional encoder-decoder structure of dynamic image construction module is effective for element-wise prediction. Besides, compared to FDP w/o $\mathcal{L}_{MER}$, FDP captures more appearance and motion details. For example, FDP accurately captures the dynamics around eyes for the second video sample, while FDP w/o $\mathcal{L}_{MER}$ fails.

Moreover, the results of FDP and FDP w/o $\mathcal{L}_{MER}$ look more reasonable than FDP variants with incomplete local-global feature-aware transformer. This again proves the effectiveness of local-global feature-aware transformer for the dynamic image construction task. Due to the appearance and motion details captured by the dynamic image construction task, FDP can focus on facial subtle muscle actions associated with MEs.

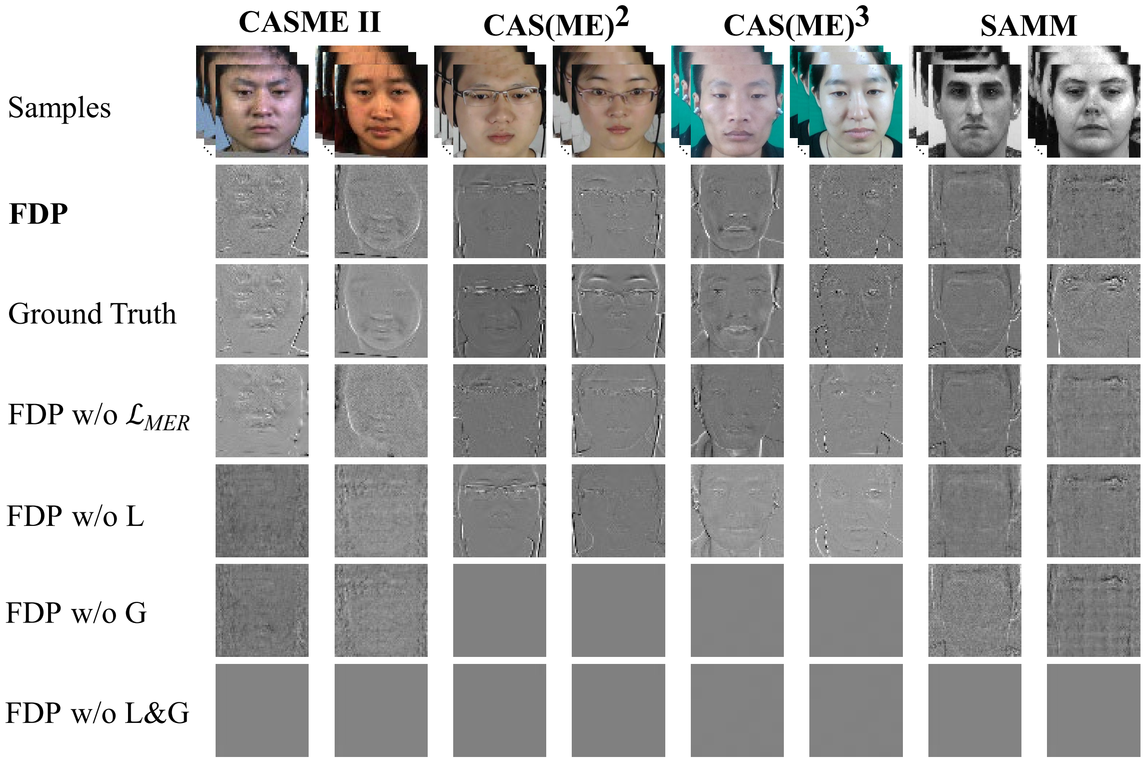


Fig. 4. Visualization of dynamic image construction results for example video clips from CASME II [68], CAS(ME)^2 [47], CAS(ME)^3 [31], SAMM [7]. The third row shows the ground-truth dynamic images, and other rows show the estimated dynamic images of different methods.

4.7 Limitations

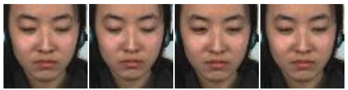
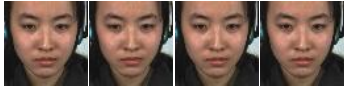
According to the above experiments, our method significantly outperforms the previous works. However, there are a few failure cases, as illustrated in Table 15. We notice that mistakenly recognized videos are very challenging, and even their ground-truth dynamic images fail to reflect clear facial subtle motions. For example, the correctly recognized video "006_2_4" of the subject "006" from SAMM has highlighted motions around eyebrows in its dynamic image, while the mistakenly predicted video "006_5_11" has no significant motions in its dynamic image. We will try to solve this challenging motion capture issue in the future work.

Besides, our method has limitations when dealing with real-world data. Since most of the existing ME datasets consist of frontal and well-lit face images collected in constrained environments, our method ignores the case of in-the-wild faces. We will also explore more robust techniques to the unconstrained scenarios.

5 Conclusion

In this paper, we have proposed a novel end-to-end fine-grained dynamic perception framework for joint MER and dynamic image construction, in which the rank technique benefits MER and two correlated tasks contribute to each other. Besides, we have developed a local-global feature-aware transformer to extract local-global features. Our framework

Table 15. Failure cases of our FDP on CASME II [68] and SAMM [7]. The incorrect predictions are highlighted in bold. "DI" denotes dynamic image.

Subject	Video		Ground Truth		Prediction
	Name	Illustration	DI	ME Cate.	
SAMM 006	006_2_4			Anger	Anger
	006_5_11			Anger	**Contempt**
CASME II 17	17_EP03_09			Happiness	Happiness
	17_EP13_09			Happiness	**Surprise**

does not rely on pre-extracted hand-crafted features and key frames, which is a promising solution to MER with good applicability.

We have compared our method with state-of-the-art works on the challenging CASME II, SAMM, CAS(ME)^2, and CAS(ME)^3 benchmarks. It is shown that our method outperforms previous works for both single dataset evaluation and cross-dataset evaluation. Besides, we have conducted an ablation study which indicates that main components in our framework are all beneficial for MER. Moreover, the experiments on dynamic image construction show excellent performance of our method, and the visual results demonstrate that our method can capture facial subtle muscle actions related to MEs.

In the future work, there are two aspects worthy of further exploring. First, regarding to the process of dynamic image construction, we hope to guide the network to pay more attention to the areas where MEs occur and ignore irrelevant information such as facial shape. Therefore, it is promising to develop the technique of disentangle irrelevant information like facial identity information. Second, the input of our network is a sequence of frames, in which some frames except for the key frames also play an important role in MER. It is promising to design a technique to locate important frames besides the key frames in a video clip, thereby facilitating the MER.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472424), the China Postdoctoral Science Foundation (No. 2023M732223), the Hong Kong Scholars Program (No. XJ2023037/HKSP23EG01), and the Research Impact Fund of the Hong Kong Government (No. R6003-21). It was also partially supported by the National Natural Science Foundation of China (Nos. 62402231, 72192821, and 62472282), the Opening Fund of the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2024C03), and the Science and Technology Planning Project of Henan Province (No. 242102211003).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *IEEE international conference on computer vision*. IEEE, 6836–6846.

[2] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. 2018. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters* 107 (2018), 50–58.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *International Conference on Machine Learning*. PMLR, 813–824.

[4] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. 2017. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2017), 2799–2813.

[5] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. 2009. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1932–1939.

[6] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. 2017. Generalized rank pooling for activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3222–3231.

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

[8] Adrian K Davison, Moi Hoon Yap, and Cliff Lansley. 2015. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1864–1869.

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

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[11] Paul Ekman. 2009. Lie catching and microexpressions. *The philosophy of deception* 1, 2 (2009), 5.

[12] Zixiang Fei, Bo Zhang, Wenju Zhou, Xia Li, Yukun Zhang, and Minrui Fei. 2025. Global multi-scale extraction and local mixed multi-head attention for facial expression recognition in the wild. *Neurocomputing* 622 (2025), 129323.

[13] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2016. Rank pooling for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2016), 773–787.

[14] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. 2019. OFF-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication* 74 (2019), 129–139.

[15] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. 2022. Cmt: Convolutional neural networks meet vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 12175–12185.

[16] SL Happy and Aurobinda Routray. 2019. Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Transactions on Affective Computing* 10, 3 (2019), 394–406.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[18] Md Kowsar Hossain Sakib, Md Rafiqul Islam, Shanjita Akter Prome, Thanh Thao Lam Nguyen, David Asirvatham, Neethiahnanthan Ari Ragavan, Xianzhi Wang, and Cesar Sanin. 2024. MVis4LD: Multimodal visual interactive system for lie detection. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 28–43.

[19] Chunlong Hu, Dengbiao Jiang, Haitao Zou, Xin Zuo, and Yucheng Shu. 2018. Multi-task micro-expression recognition combining deep and handcrafted features. In *International Conference on Pattern Recognition*. IEEE, 946–951.

[20] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. 2018. Pointwise convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 984–993.

[21] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*. PMLR, 448–456.

[22] Hamza Javaid, Aniqa Dilawari, Usman Ghani Khan, and Bilal Wajid. 2022. EEG guided multimodal lie detection with audio-visual cues. In *International Conference on Artificial Intelligence*. 71–78.

[23] Huai-Qian Khor, John See, Sze-Teng Liong, Raphael CW Phan, and Weiyao Lin. 2019. Dual-stream shallow networks for facial micro-expression recognition. In *IEEE International Conference on Image Processing*. IEEE, 36–40.

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

[25] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 1097–1105.

[27] Ankith Jain Rakesh Kumar and Bir Bhanu. 2021. Micro-expression classification based on landmark relations with graph attention convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1511–1520.

[28] Anh Cat Le Ngo, John See, and Raphael C-W Phan. 2017. Sparsity in dynamics of spontaneous subtle emotions: analysis and application. *IEEE Transactions on Affective Computing* 8, 3 (2017), 396–411.

[29] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. 2021. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1571–1580.

[30] Ling Lei, Jianfeng Li, Tong Chen, and Shigang Li. 2020. A novel graph-tcn with a graph structured representation for micro-expression recognition. In *ACM International Conference on Multimedia*. 2237–2245.

[31] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. 2022. CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 2782–2800.

[32] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. 2018. Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. *IEEE Transactions on Affective Computing* 9, 4 (2018), 563–577.

[33] Yante Li, Xiaohua Huang, and Guoying Zhao. 2021. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Transactions on Image Processing* 30 (2021), 249–263.

[34] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in network. In *International Conference on Learning Representations*.

[35] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. 2019. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 1–5.

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

[37] Yanju Liu, Yange Li, Xinhan Yi, Zuojin Hu, Huiyu Zhang, and Yanzhong Liu. 2022. Micro-expression recognition model based on TV-L1 optical flow method and improved ShuffleNet. *Scientific Reports* 12, 1 (2022), 17522.

[38] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. 2016. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. *IEEE Transactions on Affective Computing* 7, 4 (2016), 299–310.

[39] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Shilei Sun, Chunguo Li, and Luxi Yang. 2020. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185* (2020).

[40] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning Workshops*. PMLR.

[41] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[42] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*. PMLR, 807–814.

[43] Xuan Nie, Madhumita A Takalkar, Mengyang Duan, Haimin Zhang, and Min Xu. 2021. GEME: Dual-stream multi-task GEnder-based micro-expression recognition. *Neurocomputing* 427 (2021), 13–28.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 8024–8035.

[45] Min Peng, Zhan Wu, Zhihao Zhang, and Tong Chen. 2018. From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 657–661.

[46] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. 2009. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In *International Conference on Imaging for Crime Detection and Prevention*. 1–6.

[47] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. 2017. CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing* 9, 4 (2017), 424–436.

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

[49] Zhiwen Shao, Feiran Li, Yong Zhou, Hao Chen, Hancheng Zhu, and Rui Yao. 2023. Identity-Invariant Representation and Transformer-Style Relation for Micro-Expression Recognition. *Applied Intelligence* 53 (2023), 19860–19871.

[50] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2021. JÂA-Net: Joint Facial Action Unit Detection and Face Alignment via Adaptive Attention. *International Journal of Computer Vision* 129, 2 (2021), 321–340.

[51] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2022. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1274–1289.

[52] Zhiwen Shao, Yong Zhou, Jianfei Cai, Hancheng Zhu, and Rui Yao. 2023. Facial Action Unit Detection via Adaptive Attention and Relation. *IEEE Transactions on Image Processing* 32 (2023), 3354–3366.

[53] Zhiwen Shao, Hengliang Zhu, Junshu Tang, Xuequan Lu, and Lizhuang Ma. 2021. Explicit facial expression transfer via fine-grained representations. *IEEE Transactions on Image Processing* 30 (2021), 4610–4621.

[54] Zhiwen Shao, Hancheng Zhu, Yong Zhou, Xiang Xiang, Bing Liu, Rui Yao, and Lizhuang Ma. 2025. Facial Action Unit Detection by Adaptively Constraining Self-Attention and Causally Deconfounding Sample. *International Journal of Computer Vision* 133, 4 (2025), 1711–1726.

[55] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. 2021. Bottleneck transformers for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 16519–16529.

[56] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[57] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. 2022. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing* 13, 2 (2022), 1037–1043.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 5998–6008.

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

[60] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. 2015. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLOS ONE* 10, 5 (2015), e0124674.

[61] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. 2023. HTNet for micro-expression recognition. *arXiv preprint arXiv:2307.14637* (2023).

[62] Mengting Wei, Wenming Zheng, Yuan Zong, Xingxun Jiang, Cheng Lu, and Jiateng Liu. 2022. A novel micro-expression recognition approach using attention-based magnification-adaptive networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2420–2424.

[63] Bin Xia and Shangfei Wang. 2021. Micro-Expression Recognition Enhanced by Macro-Expression from Spatial-Temporal Domain. In *International Joint Conference on Artificial Intelligence*. 1186–1193.

[64] Bin Xia, Weikang Wang, Shangfei Wang, and Enhong Chen. 2020. Learning from macro-expression: a micro-expression recognition framework. In *ACM International Conference on Multimedia*. 2936–2944.

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

[66] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2020. Au-assisted graph attention convolutional network for micro-expression recognition. In *ACM International Conference on Multimedia*. ACM, 2871–2880.

[67] Shaoqi Yan, Yan Wang, Xinji Mai, Qing Zhao, Wei Song, Jun Huang, Zeng Tao, Haoran Wang, Shuyong Gao, and Wenqiang Zhang. 2024. Empower smart cities with sampling-wise dynamic facial expression recognition via frame-sequence contrastive learning. *Computer Communications* 216 (2024), 130–139.

[68] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLOS ONE* 9, 1 (2014), e86041.

[69] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* 37, 4 (2013), 217–230.

[70] Moi Hoon Yap, John See, Xiaopeng Hong, and Su-Jing Wang. 2018. Facial micro-expressions grand challenge 2018 summary. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 675–678.

[71] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. 2021. Incorporating convolution designs into visual transformers. In *IEEE International Conference on Computer Vision*. IEEE, 579–588.

[72] Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*. Springer, 214–223.

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

[74] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 915–928.

[75] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. 2021. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. *Neurocomputing* 448 (2021), 276–289.

[76] Ruicong Zhi, Jing Hu, and Fei Wan. 2022. Micro-expression recognition with supervised contrastive learning. *Pattern Recognition Letters* 163 (2022), 25–31.

[77] Li Zhou, Zhenyu Liu, Yutong Li, Yuchi Duan, Huimin Yu, and Bin Hu. 2024. Multi Fine-Grained Fusion Network for Depression Detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 8 (2024), 1–23.

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

[79] Ling Zhou, Qirong Mao, and Luoyang Xue. 2019. Dual-inception network for cross-database micro-expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 1–5.