MFDAN: Multi-level Flow-Driven Attention Network for Micro-Expression Recognition

Wenhao Cai, Junli Zhao, *Member, IEEE,* Ran Yi, Minjing Yu, Fuqing Duan, Zhenkuan Pan, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—Facial expressions are an essential part of human emotional communication, and micro-expressions (MEs), as transient and imperceptible non-verbal signals, can potentially reveal real human emotions. However, subtle motion variations, limited and unbalanced samples make micro-expression recognition (MER) challenging. In this paper, we design a novel dual-branch learning framework of multi-level flow-driven attention for micro-expression recognition (MFDAN), which innovatively integrates optical flow prior to guide the attention learning in the image encoding branch, enabling the model to focus on the most discriminative facial regions for subtle motion patterns. Firstly, we extract optical flow information by an optical flow encoding module. Then, in the image coding module, we construct a Transformer structure containing an optical flow-driven attention mechanism, which can effectively locate the interest region of micro-expressions in the image according to the position information of optical flow to capture more sensitive and fine-grained micro-expressions. By interoperating prior knowledge with data learning, and introducing the Dropkey operation and Focal Loss, our method can handle subtle micro-expression features on small imbalanced datasets. Through extensive experiments on three independent datasets and a composite database, including SMIC-HS, SAMM, and CASME II, robust leave-one-subject-out (LOSO) evaluation results show that our method outperforms state-of-the-art methods especially on the composite database.**

***Index Terms*—optical flow, micro-expression recognition, attention mechanism.**

I. Introduction

FACIAL expressions, including macro-expressions and micro-expressions, are an effective form of nonverbal communication, rich in emotional information, and crucial for understanding human emotional states [1]. Unlike macro-expressions, micro-expressions (MEs) generally occur when individuals try to hide their genuine emotions. MEs are shorter in duration and have more subtle facial muscle movements. They last between 0.04 and 0.2 seconds and are more difficult to perceive [2]. Since micro-expression (ME) is a spontaneous emotion that is difficult to control, it is more likely to reflect real human emotions, so automated ME analysis is widely used in national security, political psychology, lie detection, and depression treatment. This kind of analysis provides important clues to understanding human emotional intent and plays a vital role in different fields.

The MER task is to recognize sequences of micro-expression fragments into various emotional categories, which is challenging. Spontaneous ME is unconscious, subtle, fleeting, and has individual differences [3] due to the influence of emotional and cultural backgrounds. As a result, the collection and annotation of ME data are complex, leading to small and unbalanced ME datasets. Due to the short and uncontrollable nature of micro-expression, as well as the lack of training samples, it becomes crucial to extract robust and effective ME features accurately for performing ME analysis.

Early MER methods were mainly based on manual feature extraction [4], [5]. However, designing manual features is time-consuming, and these methods are inefficient and poorly adaptive for MER. With the rapid development of deep learning and its superiority in capturing features, the focus has gradually shifted to deep learning methods in recent years. While many studies have demonstrated the efficacy of deep learning in MER through the design of efficient shallow networks [6]–[8], there remains significant potential to enhance the current deep learning architectures to better extract the spatial features of facial micro-expressions. Some works have introduced the Transformer structure and attention mechanism to capture susceptible and discriminative features [9], [10]. However, MER still faces the challenges of locating discriminant expression regions due to slight motion and limited samples.

Compared with image, motion information in videos can provide an additional important clue, and many actions can be identified by motion information alone. As optical flow containing motion information, numerous MER techniques employ optical flow to characterize small surface movements and reduce the impact of identity features [11], [12]. To further

This work was supported by the National Natural Science Foundation of China under Grant (Nos.62172247, 62302297,61772294), Beijing Natural Science Foundation (L222008), Natural Science Foundation of Shandong Province(No.ZR2019LZH002), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), Shanghai Sailing Program (22YF1420300), Beijing Hospitals Authority Clinical Medicine Development of special funding support (ZLRK202330). (*Corresponding author: Junli Zhao.*)

Wenhao Cai, Junli Zhao and Zhenkuan Pan are with the College of Computer Science and Technology, Qingdao University, Qingdao,266071, China (e-mail:2018204652@qdu.edu.cn; zhaojl@yeah.net; zkpan@126.com).

Ran Yi is the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China(e-mail:ranyi@sjtu.edu.cn).

Minjing Yu is with the College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China(e-mail:minjingyu@tju.edu.cn).

Fuqing Duan is with the School of Artificial Intelligence, Beijing Normal University, Beijing, 100875, China(e-mail:fqduan@bnu.edu.cn).

Yong-Jin Liu is with the BNRist, MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China(e-mail:liuyongjin@tsinghua.edu.cn).

TABLE I
COMPARISON WITH PREVIOUS METHODS ON THE USE OF OPTICAL FLOW.

Method	Utilization of the optical flow	Optical flow information	Spatial information	Collaborative modeling
Bi-WOOF [5],Sparse MDMO [19]	Manual feature descriptor	YES	NO	NO
EMRNet [20]	RGB image	YES	NO	NO
STSTNet [15], SFAMNet [16], AM3F-FlowNet [17]	Three-channel image	YES	NO	NO
MERSiamC3D [21]	Optical flow sequence	YES	NO	NO
TSCNN [22],LFBVT [18]	Optical flow and image parallel modeling	YES	YES	NO
Ours(MFDAN)	Optical flow guides image spatial feature extraction	YES	YES	YES

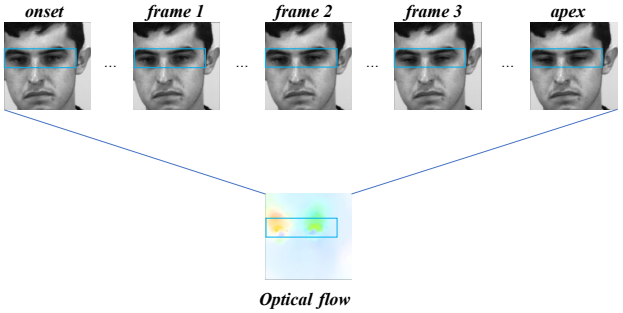

Fig. 1. ME motion is invisible to the naked eye, key feature areas are susceptible to redundant information, and the optical flow extracted from the onset frame and the apex frame has more distinctive features in regions of greater motion intensity. Our method adaptively enhances the focus on these image regions.

capture the subtle facial changes, references [13], [14] calculate the optical flow motion characteristics between the onset frame and the apex frame from the horizontal and vertical directions, respectively, and input them into the network in the form of images, where the onset frame represents the moment when the ME motion begins, and the apex frame represents the moment when the ME motion is the most intense. Other works [15]–[17] divide optical flow into three channels for feature extraction. However, these approaches only directly utilize the motion information from the optical flow, disregarding the spatial information and neglecting the correlation between dynamic and static components within the moving regions. Although some studies [10], [18] employ the spatio-temporal information of micro-expressions to process optical flow and image features independently through a two-branch network, their spatio-temporal features are spliced at a later stage in the fully connected layer of the network, without interaction and fusion in the feature extraction stage. In order to clearly compare the differences between the previous methods used of optical flow, we summarized their differences as shown in the TABLE I.

To solve the above problems, our method proposes the optical flow-driven attention mechanism to integrate optical flow and image features and comprehensively model the spatio-temporal information of micro-expressions, enabling a more effective capture of the subtle facial movements. This strategy can make the temporal and spatial features promote and improve each other in the extraction process. In Fig.1, the magnitude of the optical flow vector indicates the intensity of the motion. The longer the vector, the further away the pixels have moved. In other words, the optical flow contains more information in the region of large motion amplitude. Taking this as a starting point, we make full use of this prior information to enhance the adaptive representation of key facial areas of micro-expression. Specifically, based on the magnitude of the motion amplitude in the optical flow, we can locate that facial action occurs in a certain region of the image, that is, the region has a relatively high importance in spatial feature extraction. We propose a novel Multi-level Flow-Driven Attention Network (MFDAN) framework to thoroughly extract the key regional features to improve the performance of MER. Firstly, we employ two branches to extract temporal information and spatial information, respectively. Secondly, we learn the mutual relations between the two branches by innovatively integrating the window attention mechanism and utilizing cosine similarity to compute the similarity matrix between the optical flow and image, effectively mapping the positions with large motion amplitudes in the optical flow to the corresponding image regions. Finally, we fused temporal and spatial information in the fully connected layer to ensure learning of the global information of MEs. In essence, we introduce the prior information of feature pattern and change rules of the micro-expression into the MER model, so that the model can learn the feature representation related to the task more effectively. This will reduce the degree of dependence on data volume to a certain extent, providing an effective solution for small and unbalanced micro-expression datasets.

The main contributions of our work includes:

- We propose a novel optical flow-driven attention mechanism that leverages the prior knowledge of optical flow positions with large motion amplitudes to guide attention allocation, ensuring that the image feature extraction process precisely captures the crucial regions containing micro-expression variations.
- We propose a novel Multi-level Flow-Driven Attention Network (MFDAN) for MER, which introduces an innovative collaborative modeling scheme for optical flow and image features. The MFDAN architecture incorporates two flow-driven blocks combined with a window attention mechanism, effectively preserving the validity of spatial information while enabling comprehensive integration of spatio-temporal features.
- To address sample imbalance, we introduce Focal loss and employ the Dropkey operation to enhance our MER model's performance. Extensive experiments demonstrate that the proposed MFDAN outperforms state-of-the-art methods, and the significant improvement in cross-dataset evaluation showcases its exceptional generalization capability.

The rest of the paper is organized as follows: section II reviews essential MER-related work. Section III describes our methodology in detail. Section IV describes the experimental setup, preparation, and evaluation metrics, and section V presents the experimental results, which are analyzed and discussed. Finally, section VI summarizes the paper and the direction of future work.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3437481

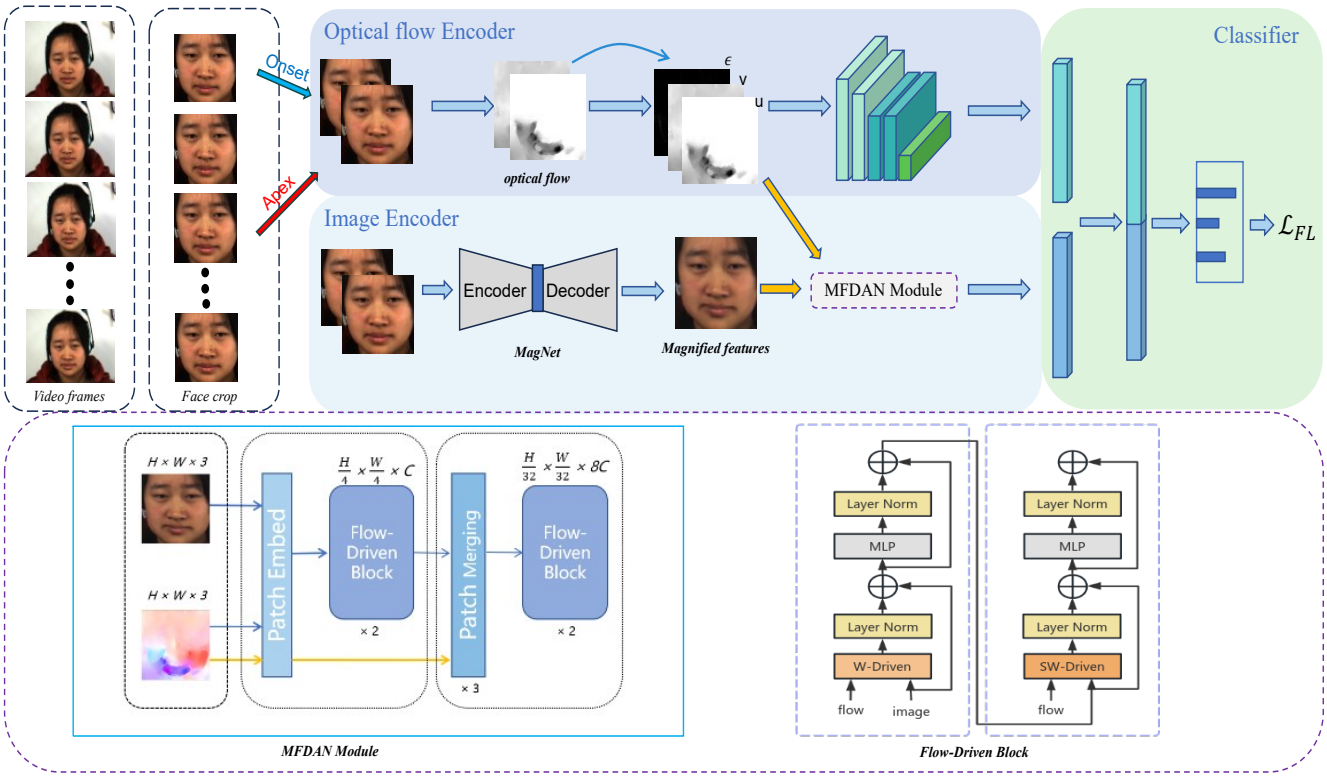

Fig. 2. The general framework of the proposed two-branch multilevel optical flow-driven attention network. We first crop the video frames, use a pair of onset and apex frames, and then extract the spatiotemporal information using optical flow and magnified images. The MFDAN Module consists of two layers of Flow-Driven Block.

II. Related Work

Currently, the MER methods are mainly divided into manual feature-based methods and deep learning-based methods.

A. Manual feature-based methods

The manual approach refers to designing visual descriptors artificially to capture the unique features of MEs. These features are then fed into an emotion recognition classifier for recognition. Zhao *et al.* [4] transform the vector code into the form of a histogram and design a Local Binary Pattern with three orthogonal planes(LBP-TOP). It uses the time and space properties of three orthogonal planes to enhance the ability to distinguish local texture feature information, then extracts and classifies features. Subsequently, in order to improve its low computational complexity, many improved algorithms based on LBP-TOP are proposed [23]–[25]. In addition, excellent progress has been made in extracting the motion features of MEs using optical flow. In order to reduce the feature dimension, Liu *et al.* [26] designs the main direction mean optical flow (MDMO), but this method makes the inherent underlying manifold structure lose the feature space. Later, Liu *et al.* [19] introduced a new distance measurement method and proposed sparse MDMO, which constructs all optical flow features in video into a dictionary to achieve sparse representation. Xu *et al.* [27] designed facial dynamics maps (FDM) to reduce the interference of noise and lighting changes in optical flow vector analysis. Liong *et al.* [5] proposed bi-weighted oriented optical flow (Bi-WOOF) and showed that the information in the onset and apex frame can already describe the characteristics of the entire micro-expression to a large extent. Inspired by this, our method uses the onset and apex frames to extract the optical flow information. However, the performance of the manual feature extraction method mainly depends on the manually designed feature extractor and requires a complex parameter adjustment process, which have poor robustness and generalization ability.

B. Deep learning-based methods

Deep learning methods refer to techniques that use deep neural networks to learn and extract facial information features in micro-expressions automatically, and these networks have made remarkable achievements in the field of computer vision. These achievements have stimulated the application of deep learning in the field of MER, and researchers have proposed various deep learning methods to improve the performance of MER. In this section, we introduce these approaches from the perspective of spatio-temporal information correlation, areas of interest, and Transformer network of MER.

Spatio-temporal information association: Patel *et al.* [28] adopted CNN network and transfer learning methods to realize MER, a milestone work to identify micro-expressions based on deep learning. Quang *et al.* [29] perform MER using CapsuleNet using only apex frames. Later, the optical flow is widely used as a pixel-level motion vector calculated between successive frames in deep learning methods for MER [22], [30]–[34]. Liu *et al.* [20] extract features from optical flow to

classify MEs. However, a single view cannot provide enough information due to the subtle motion and limited samples of MEs. Multiple inputs from different views can help us more fully understand the spatial distribution and temporal variation of micro-expressions. As a result, much of the recent work on MER uses a multi-branch network to synthesize information from different views, capturing subtle changes in micro-expressions from multiple inputs [35]–[37]. Liong *et al.* [13] and khor *et al.* [14] compute the optical flow characteristics from the onset and apex frames of each video and input the vertical and horizontal components of the optical flow into the two-stream CNN. Building upon OFF-ApexNet, Liong *et al.* [15] propose a Shallow Triple Stream Three-dimensional CNN (STSTNet) and obtained optical strain features to improve the performance further. Therefore, these methods do not directly interact and integrate spatio-temporal information. Following the mainstream approach, our MFDAN model also utilizes a two-branch structure with optical flow and image inputs. However, we develop unique collaborative modeling schemes to enhance feature learning, which significantly improves the extraction of subtle features, such as those in micro-expressions.

Regions of interest (RoIs): The facial action unit comprises basic facial movement patterns, each corresponding to a specific facial muscle movement and region [38] [39]. Since the facial movement of micro-expression may be more pronounced in some areas and relatively weak in others, different facial regions do not contribute equally to the MER. To enhance the characterization of MEs local regions and better describe local variations, some works such as [40] [41] divide the entire face into several areas on average. Other methods, such as [10], [42], [43], select areas of interest from faces for feature extraction according to the scheme provided by the Facial Action Coding System (FACS), thus mitigating the effect of invalid information regions. Liong *et al.* [44] utilize the cropped eye and mouth regions for MER, and Ruan *et al.* [45] utilize different weights for the areas of interest. Nevertheless, the cropping and artificial processing of these regions of interest impose limitations on developing an end-to-end approach for deep learning networks. Our MFDAN can adaptively find these key feature regions and effectively extract the subtle motion features in the key areas.

Transformer's Attempt at MER: With Vision Transformer (ViT) [46] applying self-attention mechanisms to the field of image, the transformer architecture has demonstrated excellent performance in a variety of visual understanding tasks [47], [48]. Through the self-attention mechanism, the network can learn the relevance and importance of facial features, helping the network focus on the essential feature areas of the face and more effectively capture the information of tiny expressions when processing facial data. The researchers have also tried them in the field of ME. Lei *et al.* [49] utilize the encoder part of the Transformer for feature extraction. Zhao *et al.* [9] incorporates a local attention module into a 3D residual prototype network to emphasize key areas in the face while making the network more sensitive to the details of micro-expressions. Liu *et al.* [50] introduce the MobileViT module combining convolution and self-attention mechanism to improve the recognition efficiency, then propose a ShuffleNet model with pre-trained optical flow combined with a small self-attention module [51].

Fig. 3. The optical flow encoder network uses the optical flow information of three channels as input. It captures the motion features of the micro-expression through the convolutional and local response normalization layers.

The window attention mechanism of Swin Transformer [52] is unique in that it segments the image into small fixed-sized chunks (windows) and applies the attention mechanism to these windows. Qin *et al.* [53] proved the superiority of Swin in face feature extraction by designing a multi-task model. This innovative approach is well suited for micro-expression analysis because micro-expression features are usually present in small parts of the face, and these subtle changes need to be accurately captured and extracted. Inspired by this, we introduce the sliding window self-attention mechanism. Together with the optical flow-driven attention, our Transformer model can focus more finely on these micro-expression regions and capture subtle facial expression changes, resulting in more accurate and reliable micro-expression features.

III. Our Method

This section presents our proposed MFDAN (multi-level flow-driven attention network) model for recognizing micro-expressions. As shown in the Fig 2, the cropped onset and apex frames are extracted from the micro-expression clip. Our framework is divided into an optical flow encoder module, an image encoder module, and a classifier. The optical flow encoder uses optical flow as input to encode spatial displacement features to characterize the subtle motion changes of ME. The image encoder performs flow-driven attention in a designed MFDAN module to adaptively enhance feature representation of areas of interest on micro-expression images. By combining these two coding methods, the model obtains robust spatio-temporal features that effectively capture the subtle dynamics of micro-expressions. In addition, to address the sample imbalance problem, we adopt Focal loss as the loss function to better optimize the model and mitigate overfitting. We also incorporate the Adan optimizer and the Dropkey operation to improve our MER's performance and generalization ability.

A. Optical Flow Encoder Module

First, we extract optical flow from the video and apply identical pre-processing operations on the video frame to ensure that the optical flow and the corresponding image have the same feature representation or scale. Specifically, we extract the total variation regularization with L1-norm (TV-L1) optical flow using the onset and apex frames that were cropped and aligned to obtain the horizontal and vertical optical flow fields u and v, with a shape of $\mathbb{R}^{H\times W\times 2}$, expressed as:

$$OF = \{(u_{xy}, v_{xy}) \mid x = 1, \ldots, H; y = 1, \ldots, W\}, \quad (1)$$

where H and W represent the height and width of the image. In the third channel we use optical strain [14], which represents the local facial movement or deformation in the image sequence and can measure the degree of pixel-level deformation and direction of facial features in the image sequence. Optical strain is defined as:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} = \frac{\partial u}{\partial x} & \epsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ \epsilon_{yx} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \epsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix}, \quad (2)$$

where the diagonal component $(\epsilon_{xx}, \epsilon_{yy})$ represents the horizontal and vertical normal strain components, respectively, while $(\epsilon_{xy}, \epsilon_{yx})$ represents the shear strain components obtained by mixing partial derivatives. The optical strain size of each pixel can be calculated by the sum of squares as follows:

$$|\epsilon| = \left| \sqrt{\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 + \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2} \right|. \quad (3)$$

This was then normalized and adjusted to an image size of $\mathbb{R}^{224\times224\times3}$.

For optical flow encoder module, we use the shallow network structure of SSSN [14] for the extraction because the optical flow features are more prominent, as shown in Fig.3. The network has advantages in optical flow feature extraction. It can quickly extract effective optical flow characteristics in a short time and effectively resist some changes and noise in the input data. It consists of two convolution modules, each extracting the feature representation of the input data through a convolution layer and a pooling layer. It uses the ReLU activation function for nonlinear transformation. The local response normalization layer enhances the robustness and differentiation of features. Finally, the time information is extracted through the fully connected layer.

B. Image Encoder with MFDAN Module

In this work, we first input one onset frame and one apex frame into a learning-based video motion magnification network (MagNet) [54] to get the face image after motion amplification. This process aims to improve the visibility of micro-expressions, making the models more accessible to capture and identify. Then, we propose an innovative optical flow and image early fusion scheme and design the MFDAN Module, a novel optical flow-driven attention module to extract facial feature information, which performs region learning and feature extraction by two *Flow-Driven-Block*. This layered structure enables the model to perform optical flow-driven attention at different scales and efficiently handle global and local relationships. It is worth noting that the optical flow plays a driving and guiding role here, while the temporal information is extracted by the optical flow encoder. The structural design of the Flow-Driven Block is inspired by Swin [52] for efficient capture of critical information. Finally, we obtained information on the critical parts with significant ME features by MFDAN.

Motion amplification. To improve the network's sensitivity to subtle expression changes and enhance the intensity of facial micro-expressions, inspired by the work of Lei *et al.* [49], we input the onset frame and the corresponding apex frames into the MagNet [54] to obtain an image that has been zoomed in by motion, where the onset and apex frames undergo the same processing as when extracting the optical flow. During the training process, a randomized amplification factor operation is performed to increase the training data's richness and enhance the network's robustness and generalization ability. Similarly, we normalize and adjust it to the same size as the optical flow so that the optical flow and the image maintain the consistency of the information source at the exact corresponding region location.

MFDAN Module. In this module, different from the general video processing methods, we combine optical flow and image in the early stage by proposing optical flow-driven attention. MFDAN Module obtains critical information about the image by two designed Flow-Driven blocks from low-level and high-level features.

First, we split the input image and optical flow $(X_m, X_f) \in \mathbb{R}^{H\times W\times3}$ into non-overlapping patches. The patch size is 4×4, and then each patch is flattened in the channel direction to a token $X_t \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times48}$. Each token is then mapped into an embedding vector of length C through the Linear Embedding layer. This chunking and embedding helps to extract local features of the image and optical flow data, providing finer-grained inputs for subsequent processing and attention computation.

Next, our Flow-Driven-Block structure is similar to Swin [52], with the difference that we use our optical flow-driven attention when performing window attention(will be introduced in Sec.III-C). As shown in the Fig 2, each Block consists of two layers of structure. We design the Window-driven-attention (*W-Driven*) structure and Shift-window-driven-attention (*SW-Driven*) structure during the attention computation driven by optical flow. The *W-Driven* structure is used for in-window self-attention computation, focusing on critical facial regions by manipulating attention inside the window. The *SW-Driven* structure is used for window-to-window information transfer, which utilizes the sliding window mechanism to interoperate and integrate information between different windows. The module output is further processed through a multilayer perceptron MLP with hidden layers and a layer norm layer.

The successive Flow-Driven-Block is calculated by:

$$\hat{z}^l = LN(\text{W-Driven}(\hat{z}^{l-1})) + z^{l-1}, \quad (4)$$
$$z^l = LN(\text{MLP}(\hat{z}^l)) + \hat{z}^l, \quad (5)$$
$$\hat{z}^{l+1} = LN(\text{SW-Driven}(z^l)) + z^l, \quad (6)$$
$$z^{l+1} = LN(\text{MLP}(\hat{z}^{l+1})) + \hat{z}^{l+1}, \quad (7)$$

where $\hat{z}^l$ and z^l are respectively denoted as the output features of the (S)W-Driven module and the MLP module, and LN is the Layernorm layer.

Note that we use the same optical flow features as inputs when performing window attention calculations in both W-Driven and SW-Driven. This means that since the optical flow feature remains the same, it does not affect the region-guiding effect of the sliding window attention. After the low-

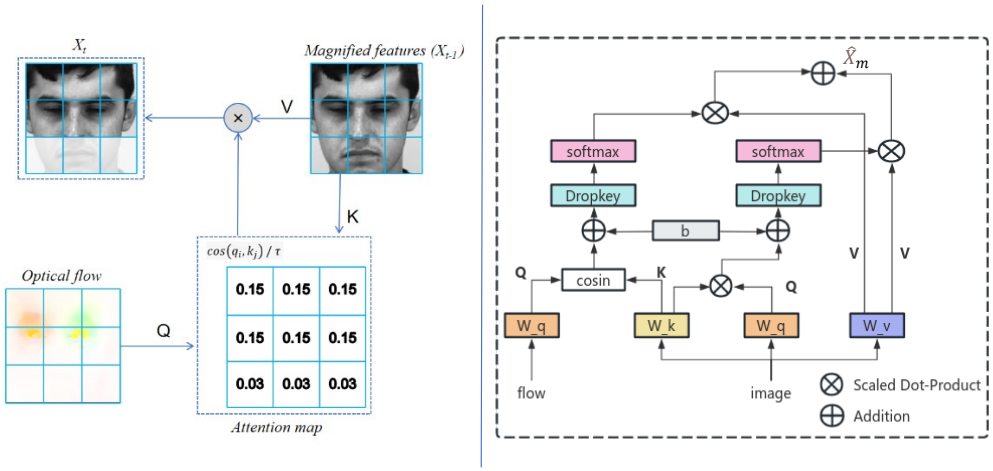

Fig. 4. The left figure shows the main idea of the optical flow-driven attention. The attention matrix is generated by calculating the cosine similarity between the optical flow and the image to increase the weight of the critical parts of the image. And the right figure shows the main flow of the (S)W-Driven structure.

level feature processing, we down-sample the feature map three times to obtain high-level features. Similarly, before entering the next block, we down-sampled the optical flow information that had not gone through the first Flow-Driven-Block three times to obtain optical flow features with the same receptive field and an equal number of windows as the image feature maps to fulfill the dimensionality requirement for the second time to perform the optical flow-driven attention. With this design, we can maintain the correspondence between the optical flow features and the image features to ensure their interact and correlate information in the same spatial extent so that no additional inconsistency or information loss will be introduced when performing the second optical flow-driven attention. And thus the critical features of the facial region can be better captured, and better discriminative high-level features can be extracted. We use two layers of transformers to introduce optical flow-driven attention from low and high layers. This structure can help the network focus on moving areas at the low-level feature level and enable the network to classify ME based on specific motion information at the high-level feature level.

C. Flow-Driven Attention

The image and optical flow are preprocessed to the same size so that the position of each pixel corresponds to the position of its optical flow. Along with this finding, we propose flow-driven attention, which uses the motion amplitude information in optical flow images to infer essential areas in the images, thereby helping the network understand image features and strengthen its attention to critical parts. It is carried out in the (S)W-Driven structure. As shown in Fig.4, according to the idea of the attention mechanism, we use cosine similarity to calculate the similarity matrix of optical flow and image so that the network can selectively focus on the moving region.

First, the image X_m and optical flow X_f are passed in through Equation 8:

$$(X_m{}', X_f{}') = F(X_m, X_f), \tag{8}$$

where X_m is the input after motion amplification and X_f is the input to the optical flow. F(·) is the Patch embed operation, which maps the image and optical flow to get their respective features $(X_m{}', X_f{}') \in \mathbb{R}^{n_h n_w \times n_c}$.

The queries (Q), keys (K), and values (V) are feature vectors extracted from the feature map at different locations. The query is used to calculate correlations with other locations, the key is used to provide reference features, and the value is used to generate the final feature representation. The corresponding Q, K, and V are obtained by a linear transformation of the image and the feature map of the optical flow. The formula is as follows:

$$Q_m^i, K_m^i, V_m^i, Q_f^i, K_f^i, V_f^i = W_{qkv}(X_m{}', X_f{}'), \tag{9}$$

where W_{qkv} is the parameter of the linear transformation, and it is worth noting that the linear layer of the image and optical stream input is weight-shared, and *i* represents the attention head.

Scaling dot-product attention is used in traditional self-attention modules, provided its queries, keys, and values come from the same feature graph. The self-attention mechanism can establish the correlation between pixels within the image to capture the local and global relationships in the image. The formula is as follows:

$$S_{self}(Q_m^i, K_m^i) = \frac{Q_m^i {K_m^i}^T}{\sqrt{d}} + B. \tag{10}$$

In contrast, the input to our optical flow driver module uses the optical flow feature as the query, the image feature as the key, and the attention matrix is S_{flow}. The formula is expressed as:

$$S_{flow}(Q_f^i, K_m^i) = \cos(Q_f^i, K_m^i)/\tau + B, \tag{11}$$

where B is the relative position offset and τ is a learnable scaling factor. As shown in the Fig.4, the optical flow describes

the displacement of a pixel point, while the image represents the pixel value of a pixel point, which belong to different domains. Based on this perspective, we argue that the attention weight map obtained using the dot product is not representative of the attention scores of the feature regions, so we use cosine similarity to compute the similarity between them. Cosine similarity does not vary depending on the scale, meaning it is not affected by the size of the vectors. In other words, it quantifies the degree of similarity between the optical flow and the image. It thus captures their similarity more efficiently and robustly, enabling the subsequent network to learn the features of the critical regions.

D. Dropkey Regularization

The micro-expression dataset has a small sample size and subtle micro-expression features, which are susceptible to the influence of noisy data interference, so we use a kind of operation called Dropkey [55], which is a novel dropout-before-softmax scheme to regularize the attention weights while maintaining their probability distribution, which intuitively helps to penalize the peaks of the weights, and enhances the model's ability of generalization by regularizing the attention weights.

As shown in the right picture of Fig.4. Unlike Dropout, which treats Keys as dropout units instead of weights, DropKey randomly masks a certain percentage of keys in the input key mapping in each training iteration. Specifically, a mask matrix $M \in \mathbb{R}^{n_h n_w \times n_h n_w}$ with the same dimension as the attention matrix is first randomly generated, and the hyperparameter dropout rate is set to d to achieve a mandatory make part of the attention score in the similarity matrix to be $-\infty$. The formula for the attention score M_j is as follows:

$$M_j = \begin{cases} 0 & \text{probability} = 1 - d \\ -\infty & \text{probability} = d \end{cases} . \tag{12}$$

Two outputs through the attention mechanism are calculated by:

$$Attn(Q_m^i, K_m^i, V_m^i) = Softmax(S_{self}(Q_m^i, K_m^i) + M_j)V_m^i, \tag{13}$$

$$Attn(Q_f^i, K_m^i, V_m^i) = Softmax(S_{flow}(Q_f^i, K_m^i) + M_j)V_m^i. \tag{14}$$

Ultimately, we sum the features after the optical flow-driven attention with the self-features, which is done so that the spatial features provided by the regions without motion changes will not be affected. In a one-shot (S)W-cross, the output $\hat{X}_m$ is obtained by:

$$\hat{X}_m = Attn(Q_m^i, K_m^i, V_m^i) + Attn(Q_f^i, K_m^i, V_m^i). \tag{15}$$

In this way, we also perform a self-attention operation on the image itself to prevent losing some vital information about the face. The contribution of original image spatial features to classification is preserved, and discriminative features are extracted through our optical flow-driven attention mechanism.

Finally, the feature vectors obtained by the optical flow encoder and image encoder are spliced in the fully connected layer, and the resulting features are input into the classifier for ME classification.

E. Focal Loss

The micro-expression dataset is highly imbalanced in categories, and most of the work on categorization loss uses the cross-entropy loss, which causes the model to be more inclined to predict the majority of the categories, thus ignoring the minority of the categories, making the model perform poorly. For this reason, we introduce Focal Loss to solve the problem of category imbalance in the MER task, which mainly mitigates the problem by introducing a tunable parameter. The formula is as follows:

$$\mathcal{L}_{FL}(P_t) = -(1 - P_t)^{\gamma} \log(P_t), \tag{16}$$

where P_t is the probability of each category, and γ is a constant, and when it is 0, focal loss is consistent with the normal cross-entropy loss function. Focal Loss adopts a dynamic weighting approach, which makes the model pay more attention to the difficult-to-categorize samples, and this effectively avoids the problem of the model's overfitting for the difficult-to-categorize samples.

IV. Experimental Setup

In this section, we describe our experimental configuration and preparation, which includes the dataset, preprocessing, and evaluation metrics.

A. Datasets

Our experiments are performed on The Chinese Academy of Sciences Micro-Expression II(CASME II) [56], the Spontaneous Actions and Micro-Movement(SAMM) [57], and the Spontaneous Micro-Expression Corpus(SMIC) [58], which are the three most commonly used datasets for micro-expression recognition. In CASME II, the camera had a sampling rate of 200 fps, a resolution of 640 × 480, and a facial resolution of 280 × 340, providing 247 micro-expression samples from 26 subjects of the same ethnicity, categorized into five categories. In SAMM, there is a frame rate of up to 200fps and a facial resolution 400×400. The dataset contains 159 samples from 32 participants and 13 ethnicities, divided into seven categories. Each sample has emotion labels, apex labels, and action unit labels. SMIC-HS consists of a sample of 164 participants divided into three categories. They cover 16 participants from 3 ethnicities, recorded at a resolution of 640 × 480 and a frame rate of 100 fps. These samples lacked apex frames and action unit labels.

In addition, we use the Composite Database Evaluation (CDE) protocol [59] from the second Micro-Expression Grand Challenge Competition (MEGC2019) to harmonize different category settings across datasets. Specifically, the CDE reorganizes CASME II, SAMM, and SMIC into three categories: **Negative** {"Repression", "Anger", "Contempt", "Disgust", "Fear", "Sadness"}, **Positive** {"Happiness"}, and **Surprise** {"Surprise"}. Finally, a total of 442 samples are taken from 68 subjects. Detailed descriptions of these three datasets and the composite dataset are shown in TABLE II. Among them, the onset(start time of the ME), apex(time of the highest intensity of the ME), and offset(end time of the ME) frames have been

TABLE II
SUMMARY OF DATASET DISTRIBUTION

Expression Class	SMIC-HS	CASME II	SAMM	Combined
Negative	70	88	92	250
Positive	51	32	26	109
Surprise	43	25	15	83
Total	164	145	133	442

labeled and provided in the datasets SAMM and CASME II. Since the labeling of the apex frames is not performed in SMIC, we use the apex frame or the middle frame information provided by Quang *et al.* [29] instead of the apex frames.

B. Data Pre-processing

For the CASME II dataset, the onset and apex frames are first extracted from the video sequence. The face is aligned to ensure that it is always horizontal and facial marker points is extracted by OpenCV and Dlib toolkit. Specifically, we use the 68 facial marker detectors provided by the Dlib toolkit to obtain the coordinates of the left eye A(x_L,y_L) and the right eye B(x_R, y_R) and compute the center coordinates C(x_C,y_C). The face image is then rotated horizontally by affine transformation by calculating the tangent value of the horizontal and vertical distances from the left eye to the center point to the desired rotation angle θ. The image is then cropped to the face position. Using center cropping, we cut the image to 420 × 420 size for the SAMM dataset. For the SMIC-HS dataset, we use the provided dataset that has been center-cropped.

C. Evaluation Metrics

Regarding the evaluation metrics, we evaluate our method using leave-one-subject-out (LOSO) cross-validation, where ME samples from one subject are retained as the test set, and all other samples are used as the training set, repeated S times, with S being the total number of subjects. Performance is measured using the unweighted F1 score (UF1) and the unweighted average recall (UAR).

$$UF1 = \frac{1}{C}\sum_{i=1}^{C}\frac{2\times TP_C}{2\times TP_C + FP_C + FN_C}, \tag{17}$$

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{2\times TP_C}{N_C}, \tag{18}$$

where C is the number of categories, N_C is the number of samples in category C. True positives (TP), false positives (FP) and false negatives (FP) were obtained based on the confusion matrix.

D. Experimental Details

We use the PyTorch framework, and all experiments are performed on an NVIDIA GeForce RTX 4060 Laptop GPU and an Intel(R) Core(TM) i7-13700H CPU processor. The embedding vector size in the Transformer block is 96, the attention heads of the Flow-Driven-Block module are 3 and 6, respectively, and the window size is 7 × 7. The γ in the Focal loss is set to 2.0, the learning rate is 0.0001, and the batch size is 16. In particular, concerning the optimizer, we use the Adan optimizer developed by Xie *et al.* [60], which consumes only half of the computational resources to obtain results close to those of the SOTA optimizer, which facilitates the training of the Transformer model, accelerates the convergence of the model, and improves the performance of the model by combining a rewritten Nesterov impulse with an adaptive optimization algorithm and introducing decoupled weight decay.

V. RESULTS AND DISCUSSION

We compare our model with the SOTA methods and perform sufficient ablation experiments to verify the validity of each module in our model.

A. Comparison with MER methods

Comparison to State-of-the-art Methods. We have conducted comparison experiments with the SOTA representative works of MER in recent years. TABLE III demonstrates that our approach achieves an accuracy of over 91% for both UF1 and UAR metrics on the CASME II dataset. On the SAMM dataset, our method outperforms all other compared methods. Some recent works [9], [20], which either pre-train on large-scale macro-expression datasets or use action unit (AU) annotations, may achieve higher scores on SMIC and CASME II dataset. However, the SMIC dataset poses challenges due to its low frame rate (100 fps), significant background noise, and inaccurate apex frame labels. Future research will focus on addressing these issues by improving low frame rate data handling, apex frame labeling, and background noise processing techniques. It is particularly noteworthy that our model exhibits superior overall performance on a combined dataset comprising three individual datasets, indicating its strong generalization ability across diverse samples from different sources. This can be attributed to our collaborative modeling approach, which effectively captures standard features and expression patterns while mitigating the impact of individual differences. Consequently, our method enhances the model's tolerance to samples with varying expressions of the same emotion, resulting in high stability. By utilizing context information and better understanding micro-expression features and their change patterns, our approach is able to process micro-expression data and tackle the challenges posed by limited and imbalanced datasets.

Comparison to the methods with optical flow. To evaluate our network's effectiveness for optical flow, we compare it with recent state-of-the-art algorithms that use TV-L1 optical flow as input, as well as some dual streams. Bi-WOOF [5] is a traditional approach without any deep learning techniques, and its performance is relatively low compared to deep learning-based models. CapsuleNet [29] uses individual apex frames as inputs instead of motion information, resulting in significantly

TABLE III
FOR COMPARISON WITH STATE-OF-THE-ART METHODS, SMIC-HS, CASME II, AND SAMM SAMPLES ARE REGROUPED INTO NEGATIVE, POSITIVE, AND SURPRISE CATEGORIES. THE REGROUPED SAMPLES FROM THE THREE DATABASES ARE COMBINED INTO A SINGLE DATASET, AND PERFORMANCE IS EVALUATED USING THE LOSO VALIDATION METHOD. RED REPRESENTS THE FIRST BEST, AND BLUE REPRESENTS THE SECOND BEST.

Methods	SMIC-HS		CASME II		SAMM		Avg		3DB-composite	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [4]	0.5882	0.5785	0.7026	0.7429	0.3954	0.4102	0.5588	0.5772	0.5882	0.5785
EMRNet [20]	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152	0.7977	0.7728	0.7885	0.7824
UAI-CNN [61]	0.7451	0.7621	0.8280	0.8065	0.7056	0.6815	0.7596	0.7500	0.7603	0.7355
FGRL-AUF [49]	0.7192	0.7215	0.8798	0.8710	0.7751	0.7890	0.7914	0.7938	0.791	0.793
ME-PLAN [9]	0.7127	0.7256	0.8632	0.8778	0.7164	0.7418	0.7641	0.7817	0.772	0.786
FRL-DGT [10]	0.743	0.749	0.919	0.903	0.772	0.758	0.8113	0.8033	0.812	0.811
SelfME [62]	0.6972	0.7012	0.9078	0.9290	N\A	N\A	N\A	N\A	N\A	N\A
MFDAN(Ours)	0.6815	0.7043	0.9134	0.9326	0.7871	0.8196	0.7940	0.8188	0.8453	0.8688

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS USING TV-L1 OPTICAL FLOW AND USING DUAL-STREAM NETWORKS.

Methods	SMIC-HS		CASME II		SAMM		Avg		3DB-composite	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Bi-WOOF [5]	0.6296	0.6227	0.7805	0.8026	0.5211	0.5139	0.6437	0.6464	0.6296	0.6227
CapsuleNet [29]	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989	0.6366	0.6295	0.6520	0.6506
GoogLeNet [63]	0.5123	0.5511	0.5989	0.6414	0.5124	0.5992	0.5412	0.5972	0.5573	0.6049
VGG16 [64]	0.5800	0.5964	0.8166	0.8202	0.4870	0.4793	0.6279	0.6320	0.6425	0.6516
OFF-ApexNet [13]	0.6817	0.6695	0.8764	0.8680	0.5409	0.5392	0.6997	0.6922	0.7196	0.7096
Dual-Inception [33]	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663	0.7045	0.6983	0.7322	0.7278
STSTNet [15]	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810	0.7257	0.7503	0.7353	0.7605
FeatRef [65]	0.7011	0.7083	0.8915	0.8873	0.7372	0.7155	0.7766	0.7704	0.7838	0.7824
MFDAN(Ours)	0.6815	0.7043	0.9134	0.9326	0.7871	0.8196	0.7940	0.8188	0.8453	0.8688

worse performance compared to approaches using optical flow, while all other deep network models utilize the feature information of optical flow. Dual-Inception [33] and STSTNet [15] also design multi-stream networks to extract information from both spatial and temporal streams. Compared with these methods, our network fully exploits optical flow information through collaborative modeling, and realizes the information interoperation between optical flow and image. On the CASME II and SAMM datasets, our method achieves the highest results, validating the effectiveness of our approach. Our network effectively captures spatio-temporal patterns in micro-expressions and precisely focuses on key regions through the optical flow-driven attention mechanism. By integrating image features and optical flow features, our model achieves superior performance in micro-expression recognition. This collaborative modeling approach empowers our network to more accurately capture the dynamic and static features of micro-expressions, surpassing the performance of other methods.

Confusion matrix. Fig. 5 shows the confusion matrix of our model on the three datasets. The confusion matrix shows that our model, on the CASME II dataset, performs well for both negative, positive, and surprise, with a 96% recognition rate for surprise. In contrast, on the SMIC dataset, the recognition rate of surprise and positive is high, but the recognition rate of negative is poor. The SAMM dataset has a large number of negative emotion samples, resulting in poor performance in identifying positive emotions on this dataset. The dataset is unbalanced, in addition to the non-uniformity of the features exhibited by positive emotions due to the influence of individual differences. Our model can achieve a more stable recognition of the three types of emotions on the composite dataset. The experimental results show that for MER task, the total number of samples for each emotion, the number of differences in the emotions, and the movement amplitude of the micro-expressions are essential factors.

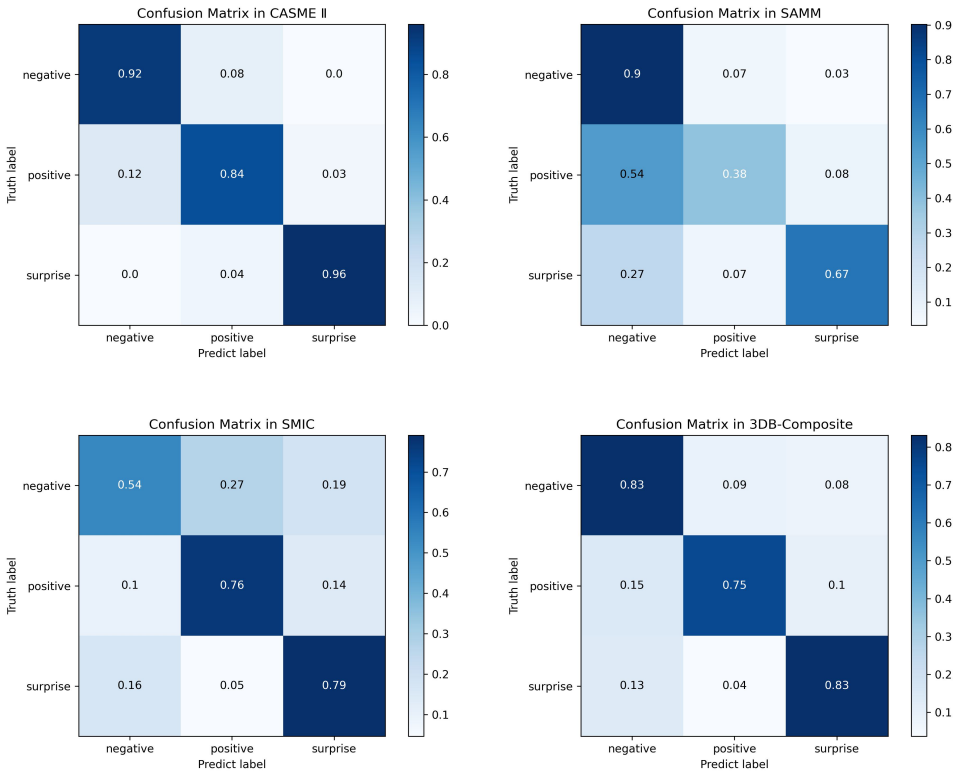

Fig. 5. Our proposed MFDAN has a confusion matrix with 3 ME classes on three datasets and the composite dataset.

B. Ablation Experiments

We conduct ablation experiments on the CASME II dataset to verify the effectiveness of each module, as shown in TABLE V.

Flow-Driven-Attention. By conducting ablation experiments, we evaluate the performance of optical flow-driven attention. In the M1 experiment, the UF1 and UAR of CASME II without flow-driven attention are 0.8493 and 0.8674, respectively, which are significantly lower. This indicates that in the micro-expression recognition task, due to the small amplitude of micro-expression movements, the network is easily interfered with by redundant regional information, which makes it difficult to focus on extracting the emotional features in the key areas. Therefore, it is essential to fully utilize the correlation information between optical flow and image features. However, the recognition ability is significantly improved when we introduce the optical flow-driven attention operation. This is because the optical flow-driven attention operation guides the extraction of image features through the optical flow information, which enables the model to focus on the motion region in a targeted way, which helps to extract the features of the critical areas and thus improves the performance of micro-expression recognition. The experimental results show that compared with the late fusion mode of the two-stream network in the fully connected layer, we use the optical flow-driven attention as an early space-time information fusion scheme of the two-stream network is effective. This provides valuable guidance for the further research and application of MER.

Focal Loss. To demonstrate the superiority of using Focal Loss, we also evaluate the performance of the proposed MFDAN framework under the conditions of negative log-likelihood(NLL) loss, cross-entropy(CE) loss, and KL divergence, respectively, and conduct comparative experiments using data resampling. The experimental results are shown in the TABLE VI. The traditional classification loss effect is significantly lower than our overall design result, indicating that the micro-expression data set is unbalanced and the sample is small, considerably impacting micro-expression recognition. Focal loss solves the problem that easily classified samples contribute too much to the loss function by reducing the weight of easily classified samples and makes the model pay more attention to the samples that are difficult to classify to deal with the training problem in the case of severe class imbalance more effectively.

TABLE V
THE ABLATION STUDY OF OUR PROPOSED NETWORK.✓INDICATES THE MODULE IS USED, AND × INDICATES IT IS NOT USED.

Method	Adan	Flow-Driven	Dropkey	Focal Loss	CASME II	
					UF1	UAR
M1	✓	×	✓	✓	0.8493	0.8674
M2	✓	✓	✓	×	0.8053	0.8351
M3	✓	✓	×	✓	0.8623	0.8834
M4	×	✓	✓	✓	0.8993	0.9201
M5	✓	✓	✓	✓	0.9134	0.9326

TABLE VI
THE EFFECTIVENESS OF FOCAL LOSS WAS VERIFIED BY COMPARISON WITH MAINSTREAM CLASSIFIED LOSS AND RESAMPLING EXPERIMENTS.

Method	CASME II	
	UF1	UAR
MFDAN + NLL loss	0.8207	0.8285
MFDAN + CE loss	0.8053	0.8351
MFDAN + KL Divergence	0.7907	0.7958
MFDAN with resampling	0.8368	0.8570
MFDAN + Focal loss (Ours)	0.9134	0.9326

TABLE VII
DROPKEY DISCARD RATE HYPERPARAMETERS (D) WERE ANALYZED ON THREE DATASETS.

DropKey(d)	CASME II		SAMM		SMIC	
	UF1	UAR	UF1	UAR	UF1	UAR
×0.3	0.8886	0.9062	**0.7871**	**0.8196**	**0.6815**	**0.7043**
×0.4	**0.9134**	**0.9326**	0.7713	0.8154	0.6672	0.6965
×0.5	0.9100	0.9300	0.7501	0.7922	0.6733	0.7048
×0.6	0.8047	0.8221	0.7293	0.7557	0.6623	0.6999

Dropkey. When performing the dropout-before-softmax scheme, the complexity of the model is reduced, and overfitting is prevented by randomly dropping keys during the computation of attention. In the M3 experiment in TABLE V, the modeling performance without Dropkey operation is lower. Due to the insufficient sample size of the ME dataset, Dropout can be more effective in mitigating the occurrence of overfitting and achieving regularization to a certain extent. Moreover, we will set a dropkey rate *d*. The effect of Dropkey is shown in the TABLE VII. Too high a dropkey rate may lead to too much information loss and a decrease in the training effect of the model, while a low dropkey rate will affect the model's generalization ability. Therefore, a proper balance is needed when choosing the drop rate. CASME II performs best when we set the discard rate to 0.4, while SAMM and SMIC are 0.3.

Adan optimizer. We evaluate the impact of the Adan optimizer in our experiments and make a comparison with the now famous AdamW optimizer in Visual Transformer, which reached 89.93% for UF1 and 92.01% for UAR on CASME II in the M4 experiments using the AdamW optimizer. It is lower than our overall design results, and experimentally, the Adan optimizer is found to be faster in convergence, which confirms that the Adan optimizer can be used as a performance-enhancing optimizer for subsequent visual Transformer frameworks.

C. Discussions

We use the Grad-CAM [66] visualization technique to visualize the heat map of the lower norm layer in the MFDAN framework to visually evaluate the effectiveness of our proposed framework on the micro-expression recognition task. As shown in the Fig. 6 .It can be seen that our model enables to help the network to focus on the facial region with optical flow movement to extract the feature information of the part and contribute to the final classification. Moreover, when using optical flow-driven attention, we find through observation that it does not affect the spatial features provided by regions without motion information. In other words, through the mechanism of optical flow-driven attention, the model can pay targeted attention to the moving areas while still utilizing the features extracted from the spatial information for the regions with no apparent movement. Therefore, the introduction of optical flow-driven attention does not disrupt or affect the extraction of spatial information by the model. On the contrary, it further improves the performance of micro-expression recognition by focusing on the motion region and making the model more concentrated on extracting critical features related to micro-expressions. This collaborative modeling scheme of optical flow and image is of great significance in the micro-expression recognition task, which retains the validity of spatial information and fully uses the a priori knowledge of optical flow information. It can be used for reference in the field of video action recognition.

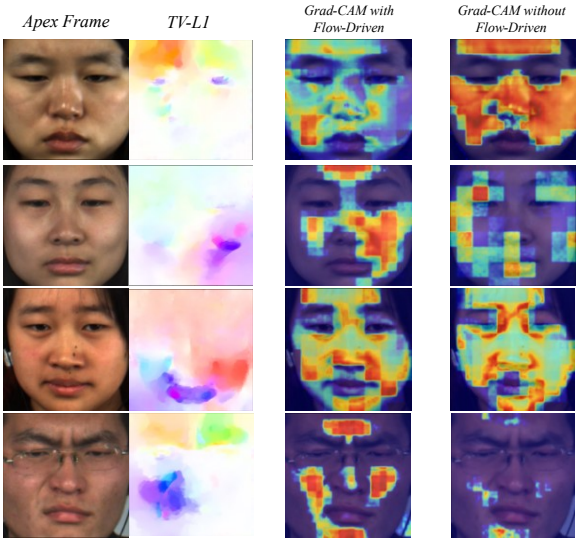

Fig. 6. With Grad-CAM, we are able to plot heat maps with and without Flow-Driven separately, and by comparing them with the optical flow visualization maps, it is intuitively clear that Flow-Driven is effective in helping the network find critical areas.

MFDAN currently uses TV-L1 optical flow as input, and nowadays, more and more advanced methods using self-supervised motion representation are being developed in MER. Although TV-L1 optical flow can represent the motion between the apex frame and the onset frame very well, the current ME datasets are performed on laboratory-controlled scenes. Suppose MER needs to be completed in natural environments. In that case, the motion information of the optical flow is exceptionally vulnerable to noise, such as ambient light, so future work will focus on adaptively extracting targeted micro-expression facial motion information instead of optical flow. MFDAN needs to determine the onset and apex frames in advance, which may seriously affect the performance if they are inaccurate. Therefore, in the future, a whole set of frame

systems from micro-expression apex localization to optical flow adaptive generation and recognition can be designed, and our optical flow-driven attention will be a crucial part of it, which will play an essential role in improving the performance.

VI. Conclusion

For micro-expression recognition task, we propose a new two-branch multilevel optical flow-driven attention network framework and design a new attention mechanism: optical flow-driven attention, in which the optical flow information is used as a query to guide the computational process of attention to optimize the extraction of image features. This mechanism makes full use of the motion information of the optical flow, which enables the attention to be more focused on the image regions related to motion. We also improve the accuracy and performance of image features with Focal Loss and Dropkey. Comparison experimental results on CASME II, SAMM, SMIC-HS, and composite datasets of them, and ablation experiments prove the effectiveness of each proposed module. And our MFDAN effectively solves the problem of insufficient and unbalanced micro-expression dataset, which is a new attempt. Introducing optical flow-driven attention brings new ideas and methods for solving image processing and computer vision tasks. In the future, we will address these limitations and extend the MFDAN framework to create a unified pipeline for ME discovery and MER tasks.

References

[1] P. Ekman, "Darwin, deception, and facial expression," *Annals of the new York Academy of sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[3] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: Datasets, features and metrics," *ArXiv*, 2018.

[4] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[5] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[6] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5826–5846, 2021.

[7] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, pp. 2028–2046, 2021.

[8] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: facial micro-expression recognition," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19 301–19 325, 2018.

[9] S. Zhao, H. Tang, S. Liu, Y. Zhang, H. Wang, T. Xu, E. Chen, and C. Guan, "Me-plan: A deep prototypical learning with local attention network for dynamic micro-expression recognition," *Neural Networks*, vol. 153, pp. 427–443, 2022.

[10] Z. Zhai, J. Zhao, C. Long, W. Xu, S. He, and H. Zhao, "Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 086–22 095.

[11] X. Zeng, X. Zhao, X. Zhong, and G. Liu, "A survey of micro-expression recognition methods based on lbp, optical flow and deep learning," *Neural Processing Letters*, pp. 1–32, 2023.

[12] L. Zhang and O. Arandjelović, "Review of automatic microexpression recognition in the past decade," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 414–434, 2021.

[13] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129–139, 2019.

[14] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *2019 IEEE international conference on image processing (ICIP)*, 2019, pp. 36–40.

[15] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, 2019, pp. 1–5.

[16] G.-B. Liong, S.-T. Liong, C. S. Chan, and J. See, "Sfamnet: A scene flow attention-based micro-expression network," *Neurocomputing*, vol. 566, p. 126998, 2024.

[17] C. Fu, W. Yang, D. Chen, and F. Wei, "Am3f-flownet: Attention-based multi-scale multi-branch flow network," *Entropy*, vol. 25, no. 7, p. 1064, 2023.

[18] J. Hong, C. Lee, and H. Jung, "Late fusion-based video transformer for facial micro-expression recognition," *Applied Sciences*, vol. 12, no. 3, p. 1169, 2022.

[19] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, "Sparse mdmo: Learning a discriminative feature for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 254–261, 2018.

[20] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, 2019, pp. 1–4.

[21] S. Zhao, H. Tao, Y. Zhang, T. Xu, K. Zhang, Z. Hao, and E. Chen, "A two-stage 3d cnn based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, 2021.

[22] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184 537–184 551, 2019.

[23] S.-T. Liong, J. See, R. C.-W. Phan, K. Wong, and S.-W. Tan, "Hybrid facial regions extraction for micro-expression recognition system," *Journal of Signal Processing Systems*, vol. 90, pp. 601–617, 2018.

[24] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016.

[25] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 32–47, 2017.

[26] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[27] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[28] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 2258–2263.

[29] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–7.

[30] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.

[31] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and Bennamoun, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *ArXiv*, vol. abs/1904.11592, 2019.

[32] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in psychology*, vol. 8, p. 1745, 2017.

[33] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, 2019, pp. 1–5.

[34] L. Zhou, Q. rong Mao, and M. Dong, "Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation," *ArXiv*, vol. abs/2012.13148, 2020.

[35] N. Liu, X. Liu, Z. Zhang, X. Xu, and T. Chen, "Offset or onset frame: A multi-stream convolutional neural network with capsulenet module for micro-expression recognition," in *2020 5th international conference on intelligent informatics and biomedical sciences (ICIIBMS)*, 2020, pp. 236–240.

[36] B. Sun, S. Cao, J. He, and L. Yu, "Two-stream attention-aware network for spontaneous micro-expression movement spotting," in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, 2019, pp. 702–705.

[37] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1511–1520.

[38] W. Zhang, L. Li, Y. Ding, W. Chen, Z. Deng, and X. Yu, "Detecting facial action units from global-local fine-grained expressions," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[39] P. Ekman and E. L. Rosenberg, "What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (facs)," 2005.

[40] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE transactions on affective computing*, vol. 9, no. 4, pp. 563–577, 2017.

[41] B. Chen, K.-H. Liu, Y. Xu, Q.-Q. Wu, and J.-F. Yao, "Block division convolutional network with implicit deep features augmentation for micro-expression recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1345–1358, 2022.

[42] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, and M. H. Yap, "Objective micro-facial movement detection using facs-based regions and baseline evaluation," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018, pp. 642–649.

[43] W. Merghani and M. H. Yap, "Adaptive mask for region-based facial micro-expression recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 765–770.

[44] S.-T. Liong, J. See, R. C.-W. Phan, K. Wong, and S.-W. Tan, "Hybrid facial regions extraction for micro-expression recognition system," *Journal of Signal Processing Systems*, vol. 90, pp. 601–617, 2018.

[45] B.-K. Ruan, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Mimicking the annotation process for recognizing the micro expressions," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 228–236.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[47] R. Ji, J. Li, L. Zhang, J. Liu, and Y. Wu, "Dual transformer with multi-grained assembly for fine-grained visual classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[48] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[49] L. Lei, T. Chen, S. Li, and J. 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*, 2021, pp. 1571–1580.

[50] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang, and Y. Liu, "Lightweight vit model for micro-expression recognition enhanced by transfer learning," *Frontiers in Neurorobotics*, vol. 16, p. 922761, 2022.

[51] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang, and Y. Liu, "Micro-expression recognition model based on tv-l1 optical flow method and improved shufflenet," *Scientific Reports*, vol. 12, no. 1, p. 17522, 2022.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[53] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng, "Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[54] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik, "Learning-based video motion magnification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 633–648.

[55] B. Li, Y. Hu, X. Nie, C. Han, X. Jiang, T. Guo, and L. Liu, "Dropkey," *arXiv preprint arXiv:2208.02646*, 2022.

[56] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[57] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE transactions on affective computing*, vol. 9, no. 1, pp. 116–129, 2016.

[58] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, 2013.

[59] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "Megc 2019–the second facial micro-expressions grand challenge," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–5.

[60] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan, "Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models," *arXiv preprint arXiv:2208.06677*, 2022.

[61] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images." in *CVPR Workshops*, 2019, pp. 12–20.

[62] X. Fan, X. Chen, M. Jiang, A. R. Shahid, and H. Yan, "Selfme: Self-supervised motion learning for micro-expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 834–13 843.

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[65] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition*, vol. 122, p. 108275, 2022.

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

Wenhao Cai is currently studying for a master's degree in the School of Computer Science and Technology at Qingdao University. His current research interests are computer vision, microexpression recognition, and 3D action recognition.

Junli Zhao is a professor and doctoral supervisor in College of Computer Science and Technology, Qingdao University, Qingdao. She received her PhD degree in Computer applied technology major in 2015 from Beijing Normal University, Beijing. She was a visiting scholar of State University of New York At Stony Brook. She is currently engaged in craniofacial informatics, computer graphics, computer vision, and virtual reality research. She has presided over and participated in more than ten projects, such as National Natural Science Foundation of China, Key Research and Development Projects of Shandong Province, Natural Science Foundation of Shandong Province and etc., published more than fifty papers in conference and journals on related topics.

Ran Yi is an assistant professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. She received the BEng degree and the PhD degree from Tsinghua University, China, in 2016 and 2021. Her research interests include computer vision, computer graphics and computational geometry.

Minjing Yu is currently an associate professor with the College of Intelligence and Computing, Tianjin University, China. She received the B.Eng. degree from Wuhan University, Wuhan, China, in 2014, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2019. Her research interests include cognitive computation, human-computer interaction and computer graphics.

Fuqing Duan received his Ph.D. degree in 2006 from the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation of the Chinese Academy of Sciences (CAS), China. Currently, he is a professor at the School of Artificial Intelligence, Beijing Normal University. His research interests are computer vision and computer graphics.

ZhenKuan Pan is a professor with the College of Computer Science and Technology, Qingdao University, Qingdao, China from 1996. He received PhD from the Department of Engineering Mechanics, Shanghai Jiao Tong University, Shanghai, China in 1992 and bachelor's degree from the Department of Engineering Mechanics, Northwestern Polytechnical University, Xi'an, China in 1987. He was a visiting scholar of UCLA, UCI in 2005 and 2015 respectively. His main research interests focus on computer vision, image processing, multibody system dynamics and control, with over 300 coauthored papers.

Yong-Jin Liu is a tenured full professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include computer vision, computer graphics and computer-aided design. For more information, visit http://cg.cs.tsinghua.edu.cn /people/ Yongjin/Yongjin.htm.