IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 November 2024, accepted 28 November 2024, date of publication 2 December 2024,
date of current version 10 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3509862

RESEARCH ARTICLE

Temporal-Polar Dynamics: Elevating Event-Based Micro-Expression Recognition

JUNWU LIN[1], CUNHAN GUO[2,3], AND XIAOFANG WU[4]
[1]New Engineering Industry College, Putian University, Putian 351100, China
[2]School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
[3]Southeast Academy of Information Technology, Beijing Institute of Technology, Beijing 100811, China
[4]College of Electromechanical and Information Engineering, Putian University, Putian 351100, China

Corresponding author: Cunhan Guo (guocunhan22@mails.ucas.ac.cn)

This work was supported in part by the Science Foundation of Fujian Province Project, China, under Grant 2022J011170; and in part by the Startup Fund for Advanced Talents of Putian University under Grant 2022058.

ABSTRACT Micro-expressions refer to brief, subtle facial movements often concealing genuine human emotions. However, the advancement of Micro-Expression Recognition (MER) is hindered by the low frame rates of frame-based cameras. Although the successor event camera has a high frame rate, there are currently difficulties in obtaining and unstable performance issues. Drawing inspiration from the operational principles of event camera, we introduces two event features. Beyond spatial information, these features encode temporal and polarity information of event. Following local normalization, we employ the temporal polar pixel-wise interaction module to extract local feature. Additionally, we construct a temporal polar dynamic network, merging local feature with dense global optical flow to map deeper features. Experimental results demonstrate the superiority of the proposed method across multiple datasets compared to state-of-the-art approaches. This work enriches the encoding of event features, enhancing their performance in micro-expression recognition tasks and contributing to the future proliferation of event camera technology.

INDEX TERMS Event feature, micro-expression, polarity, temporal information.

I. INTRODUCTION

In recent years, the realm of human face analysis, specifically within the domain of facial expression recognition with a focal point on micro-expressions, has garnered substantial attention. Micro-expressions denote fleeting, nuanced facial movements that authentically convey an individual's emotions, often concealed in high-risk situations. The potential applications of micro-expressions span criminal investigations, clinical diagnoses, and business negotiations, underscoring the pivotal importance of unveiling concealed emotions.

Nevertheless, the precise detection of micro-expressions faces challenges stemming from a deficiency in diverse datasets, potentially leading to overfitting. The construction of models that strike a balance between recognition accuracy and computational efficiency poses an additional obstacle in this field. Initially, researchers relied on manual feature engineering and traditional machine learning for micro-expression recognition tasks. The subsequent integration of deep learning marked a noteworthy advancement. However, the intricate task of constructing models capable of extracting distinctive features persists.

Historically, datasets were amassed using conventional frame-based cameras, often capturing restricted information due to their low frame rates. In contrast, event-based cameras emerge as a promising solution. These cameras generate pixel-level events when intensity changes surpass specified thresholds, resulting in ultra-high frame rates, minimal latency, and a wide dynamic range. This renders them well-suited for capturing micro-expression data. Nonetheless, the domain of event-based cameras is still in its early stages and grapples with diverse technical challenges that necessitate further exploration, including high cost, poor stability, and difficulty in obtaining.

The associate editor coordinating the review of this manuscript and approving it for publication was Bing Li.

Motivated by the underlying principles of event-based cameras, we have presented a **G**lobal-**L**ocal **E**vent **F**eature **F**usion **N**etwork (**GLEFFN**) for MER task [10]. This work marks the first successful attempt to solve the MER task from an event feature perspective. However, both temporal and polarity information were deprecated, which play important role in MER task [39], thus restricting the performance of the model. In addition, this work did not fully integrate global and local features, which can further optimize the strategy of feature fusion.

To address the aforementioned challenges, we introduced the **Te**mporal-**P**ol**a**r **D**ynam**i**c Network (**TePaDi**). The model adopted the global-local design philosophy of GLEFFN and continued to utilize Optical Flow (OF) and a global feature computation module for global feature computation. For local information features, we proposed two novel event features, encoding temporal and polarity information, and designed the Temporal Polar Pixel-wise Interaction module (TPPI) for local feature fusion. Finally, global-local feature fusion was performed to output recognition results. Our primary contributions are as follows:

i) We introduce two event features, which encode temporal and polarity information of event together with spatial information, contributing to the precision of the MER task.

ii) We propose a new designed global-local feature fusion network, named TePaDi, which enhances the feature interaction efficiency and effectiveness.

iii) Our approach showcases remarkable performance across a variety of datasets, with experiments emphasizing the capability of the event feature to effectively capture rapid movements.

II. RELATED WORKS

In the ever-evolving landscape of micro-expression recognition, this comprehensive examination delves into a myriad of methodologies, guided by a nuanced understanding of diverse approaches and the datasets they engage. Over the recent temporal interval, a multitude of empirical investigations has unfolded strategically, aiming to address potential biases within individual datasets. A fascinating journey that initially revolved around manually crafted features has gradually transitioned into the intricate realm of deep learning, unearthing latent insights. This evolutionary trajectory has culminated in a discernible shift towards the investigation of composite datasets, heralding a paradigmatic advancement in current research methodologies.

A. HAND-CRAFTED FEATURES

The pivotal role of manually crafted features unfolds as a compelling narrative in recent endeavors, showcasing their resilience and efficacy in extracting nuanced information from datasets of modest proportions. Shreve et al. [29] innovatively employed the central difference method to dense optical flow fields, revealing strain magnitude for subsequent deployment in micro-expression recognition. A symphony of methods, including LBP-TOP [35], extended Local Binary Patterns to three orthogonal planes, achieving commendable success in deciphering spontaneous facial micro-expressions. Wang et al. [30] elevated LBP-TOP by introducing intersection points, sculpting a more compact and lightweight representation. This symphony continued with Huang et al. [13], fashioning Local Quantized Patterns into CLQP, a mosaic of sign-based, magnitude-based, and orientation-based distinctions. Chavali et al. [6] orchestrated a harmonious blend, combining Histogram of Oriented Gradient with Eulerian video magnification for micro-expression recognition, yielding promising outcomes. Temporally and spatially, Huang et al. [12] crafted an opus, extracting local binary pattern information, embracing integral projection technology. Despite their tenacity, these manually crafted features encounter challenges in plumbing the profound depths of information, constraining their evolutionary journey.

B. LEARNING-BASED STRATEGIES

The crescendo of deep learning, harmonizing with the availability of expansive datasets, has ushered in an era of learning-based feature extraction techniques in micro-expression recognition. Kim et al. [17] orchestrated a melodic convergence, employing Convolutional Neural Networks (CNNs) to weave spatial tales, entwined with the rhythmic beats of a Long Short-Term Memory (LSTM) recurrent neural network for temporal nuances. Khor et al. [16] conducted a dual-stream symphony, fusing convolutional features into a harmonious representation. Li et al. [19] unveiled a spatial-temporal composition, sculpting a 3D neural network to navigate the micro-expression recognition task. Zhi et al. [37] composed a sonnet, introducing supervised contrastive learning as the key to unlocking representation intricacies in micro-expressions, painting a canvas of effectiveness through competitive results.

While learning-based methods dance to the rhythm of data dependency, the existing repertoire of datasets often falls short for the grandeur of deep learning performances. In response, researchers have orchestrated a concerto, fusing manually crafted features with learning-based techniques. Chan et al. [5] curated a performance, leveraging the synergy of both realms—manually crafted features and deep networks—in the PCANet for micro-expression recognition, a crescendo that echoed promises.

C. COMPOSITED DATASET ANALYSIS

The inception of composite datasets emerges as an enthralling chapter, a coveted elixir sought by researchers to transcend the confines inherent in individual datasets. These datasets, akin to alchemists' concoctions, proffer a dual potion: expediting model training while fortifying defenses against the seductive whispers of overfitting. Woven from samples spanning diverse races and ages, their tapestry breathes vitality into dataset diversity, giving birth to the phenomenon of cross-dataset validation—an acid test for the wings of model transferability.

The epic of deep networks in micro-expression recognition draws inspiration from the realm of composite datasets. Liu et al. [23] painted an optical masterpiece, employing part-based average pooling to distill discriminative representation, clinching a podium finish in the 2nd Micro-Expression Grand Challenge. Xia et al. [31] composed a symphony, introducing a recurrent convolutional network (RCN) adorned with parameter-free modules—a ballet of wide expansion, shortcut connection, and attention units—garnering applause on the MEGC2019 stage. Li et al. [20] sculpted a magnum opus, the multi-scale joint feature network, intricately weaving optical flow images into the fabric of micro-expression recognition. The validation, a triumphant odyssey through the landscapes of three benchmark datasets (SMIC, CASME II, and SAMM) and a composite dataset (3DB), etched a testament to its efficacy. This multifaceted exploration vividly encapsulates the cutting edge in the captivating realm of micro-expression research.

D. EVENT BASED CAMERA APPLICATION

Event cameras, also known as neuromorphic vision sensors, represent a significant advancement in imaging technology. Unlike traditional cameras, which capture frames at regular intervals, event cameras detect changes in brightness asynchronously at a per-pixel level. This unique feature allows them to achieve high temporal resolution, even in challenging conditions such as low light and high-speed motion.

These cameras have found applications across various domains, including autonomous driving [28], [33], facial recognition [3], [24], and lip reading [4]. In autonomous vehicles, event cameras enable real-time perception of the environment, enhancing safety and efficiency [4]. Additionally, in facial recognition systems, they offer advantages in terms of speed and accuracy. Similarly, event cameras have shown promise in lip reading applications, where they excel in capturing rapid and subtle movements of the lips.

However, one of the most intriguing applications of event cameras lies in micro-expression recognition [1], [2]. Micro-expressions are fleeting facial expressions that reveal underlying emotions, often lasting just a fraction of a second. Traditional cameras struggle to capture these subtle cues effectively. Event cameras, with their high temporal resolution, excel in detecting and analyzing micro-expressions, providing valuable insights into human emotions and behavior.

Despite their effectiveness in micro-expression recognition, event cameras present challenges in terms of acquisition and stability. At present, there are only a few manufacturers capable of producing event cameras, but the cost is high, and the current selling price is dozens of times higher than ordinary cameras, making it difficult to obtain. The technology is still relatively nascent, and the spatial resolution of the camera is also relatively low. Furthermore, its stability is still poor, that the performance may be affected by environmental factors, such as lighting conditions and motion artifacts.

III. FEATURE EXTRACTION

In this section, we will provide a detailed explanation of our data pre-processing and the process of event. Two event features encoding temporal and polarity information will also be introduced.

A. TEMPORAL UP-SAMPLING

Standard frame-based cameras typically operate at a frame rate of around 60 frames per second (fps), with high-speed cameras reaching rates exceeding 200 fps. Despite this, micro-expressions persist for an extremely brief duration, spanning from 0.04 seconds to half a second, usually around 0.2 seconds. Consequently, even high-speed cameras capturing 200 fps can only provide 40 frames for a single micro-expression event. This limited frame count poses challenges for effective feature acquisition. To surmount this limitation, we employed a video temporal up-sampling method to enhance the frame rate.

Video temporal up-sampling methods encompass four categories: frame blending, frame sampling, motion estimation and motion compensation, and optical flow-based approaches. In our investigation, we embraced an optical flow-based method known as Super SLoMo, introduced by NVIDIA [15]. Super SLoMo utilizes a U-Net to compute bidirectional optical flow between consecutive output images. To refine the approximate optical flow and predict soft visibility maps at stationary boundaries, another U-Net is employed. The final step involves transforming and linearly blending the two output images to generate the intermediate frame.

We applied Super SLoMo to up-sample the dataset; for instance, up-sampling the CASME II dataset from 200 fps to 1000 fps, resulting in approximately 200 frames per micro-expression, as depicted in Figure. 1. Leveraging the Super SLoMo technique effectively increased the temporal resolution of the original video, enabling a higher frame rate. This enhancement facilitated the capture and analysis of micro-expressions in finer detail, significantly improving the feature acquisition process. This, in turn, played a pivotal role in the success of our proposed global-local event feature fusion network in micro-expression recognition tasks.

B. EVENT FEATURE CALCULATION

We will first generate event from video, and then extract temporal image feature and polar image feature, respectively. Both of these two feature will be further normalized at a local manner to change into local features.

1) FROM VIDEO TO EVENT

The event camera, being a bio-inspired visual sensor, operates differently from conventional cameras. Instead of generating intensity image frames at a constant rate, it provides information about the changes in local pixel-level log photocurrent intensity ($\mathcal{L}$). When these changes in intensity surpass a predetermined threshold (C), the event camera records the

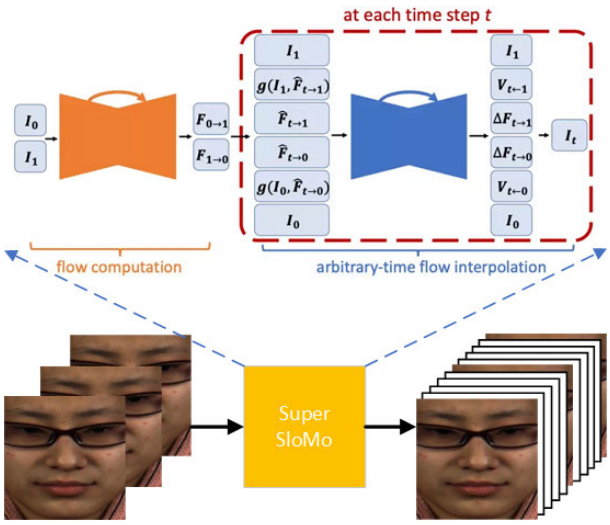


FIGURE 1. Temporal up-sampling with super SLoMo. These frame were carefully extracted and cut from the raw video based on the landmark of face to reduce the background disturbing and reduce the computational cost. All the videos were finally up-sampled to 1000 fps for fairness.

timestamp with microsecond resolution and generates an asynchronous event stream, referred to as an 'event'. The polarity (p) of an event is determined by the D-value, $\Delta\mathcal{L}(x, y, t)$, representing the difference in log photocurrent of a pixel between time $(t - \Delta t)$ and time t. This process can be mathematically described by Equation (1) and Equation (2).

$$\Delta\mathcal{L}(x, y, t) = \mathcal{L}(x, y, t) - \mathcal{L}(x, y, t - \Delta t) \tag{1}$$

$$p(x, y, t) = \begin{cases} 1 & , \Delta\mathcal{L}(x, y, t) > C \\ 0 & , -C < \Delta\mathcal{L}(x, y, t) < C \\ -1 & , \Delta\mathcal{L}(x, y, t) < -C \end{cases} \tag{2}$$

By employing this mechanism, a video can be transformed into an event stream data with an extremely long length in one dimension but containing sparse data in the other dimension. To facilitate further utilization, we recorded the pixels that generated the event and organized them into the same set, the $E_t^{(p)}$. Here, t means the time stamp and p is the polarity, which can be positive (+) or negtive (−).

2) TEMPORAL IMAGE

The concept behind temporal image encoding takes into consideration both the rate and spatial frequency of events. Suppose that at some truncated timestamp T, a pixel (x_1, y_1) generates event A over a duration of t_1 time units, while another pixel (x_2, y_2) generates event B over a duration of t_2 time units. If t_1 is less than t_2, indicating that the timestamp of event B is earlier than that of event A, and the pixel $(x_{1,2}, y_{1,2})$ does not generate new events after producing event A or event B until the specified timestamp, then the pixel (x_1, y_1) is considered to have a higher frequency to generate an event.

Follow this hypothesis, firstly, let F indicates the whole image sequence. Then, cut F into M pieces, the mth one can be represented as $F[\Delta T \times (m-1), \Delta T \times m], m \in [1, M)$. Here, the $\Delta T \times m$ is the truncated timestamp. In each segment, we calculated the timestamp of the last time when the event of each pixel was generated, denoted as $t_m^{(x,y)}$. Then we can calculate temporal image for each segment, as shown in Equation (3), where δ is used to prevent divisor from being 0, and is set 0.5 here.

$$TI_m^{(x,y)} = \frac{1}{\Delta T - t_m^{(x,y)} + \delta}, \tag{3}$$

Finally, we can get the temporal image by simply calculate the average of each segment, as shown in Equation (4)

$$TI = \frac{1}{M} \sum_{m=1}^{M} TI_m \tag{4}$$

3) POLAR IMAGE

For leveraging polarity information, we designed the polar image. Building upon the count image [10], we assign different weights γ to the statistics for each type of event. When finishing calculation of all N frames' polar image, calculate their average, as shown in Equation (5).

$$PI^{x,y} = \frac{1}{N} \sum_{i=1}^{N} \gamma_n^{(x,y)}. \tag{5}$$

Initially, we calculate events generated between two frames using frame differencing. We then determine the event types produced at that moment and, based on the Algorithm. 1, assess the weight assigned to each pixel during the statistics. Taking the most common scenario as an example, through frame differencing, we compute several events between two frames, including positive and negative events, as well as some pixels that do not generate events. For pixels generating positive events during calculation, $\gamma_n^{(x,y)} = 1$, while pixels not generating events correspond to $\gamma_n^{(x,y)} = -1$. Through this approach, we can effectively utilize the polarity of events. Although the features generated by this method may have lower interpretability, they contain more information and are suitable for in-depth exploration by computers.

4) LOCAL NORMALIZATION

In order to capture finer details of the local regions, it is essential to normalize the temporal image feature and polar image feature from a local region perspective rather than a global view. As depicted in Figure. 2a, we have meticulously designed nine distinct local regions. These regions exhibit overlaps and are interconnected, facilitating a more concise representation for each region. This thoughtful design aids in reducing the distance between representations of individual regions, thereby simplifying the process of feature mapping and model fitting.

After determining the local regions, a local region normalization was applied to the previous feature, as depicted in

Algorithm 1 Polar Information Encoding

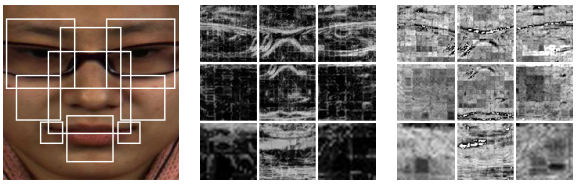
for (n, x, y)in range$((0, N), (0, width), (0, height))$ **do**
  if $\forall p_n \in E_n^{(+)}$ **then**
    $\gamma_n^{(x,y)} = 1$ if $(x, y) \in E_n^{(+)}$, else $\gamma_n^{(x,y)} = -1$
  else if $\forall p_n \in E_i^{(-)}$ **then**
    $\gamma_n^{(x,y)} = 0$ if $(x, y) \in E_n^{(-)}$, else $\gamma_n^{(x,y)} = -1$
  else
    if $then(x, y) \in E_n^{(+)}$
      $\gamma_n^{(x,y)} = 1$
    else if $then(x, y) \in E_n^{(-)}$
      $\gamma_n^{(x,y)} = 0$
    else
      $\gamma_n^{(x,y)} = -1$
    end if
  end if
end for

Equation (6) and Equation (7), where $[k]$ represents region k.

$$TI^{(x,y)} = \frac{TI^{(x,y)} - min(TI[k])}{max(TI[k]) - min(TI[k])}, (x, y) \in [k], \quad (6)$$

$$PI^{(x,y)} = \frac{PI^{(x,y)} - min(PI[k])}{max(PI[k]) - min(PI[k])}, (x, y) \in [k], \quad (7)$$

Finally, we resize and rearrange the nine part of feature according to the regions, as shown in Figure. 2b and Figure. 2c. The resulting feature is referred to as the Local Temporal Image (LTI) and Local Polar Image (LPI), representing the calculated local event feature.



(a) Local regions (b) Local temporal image (c) Local polar image

FIGURE 2. Local features.

IV. MODEL ARCHITECTURE

By comprehensively considering global perspective, local perspective, and the fusion of multi-modal features, we propose TePaDi to perform MER in this section. Next, the overall architecture and core modules of TePaDi are separately explained in detail.

A. OVERVIEW OF NETWORK ARCHITECTURE

Concerning the model design, we embrace the concept of global-local feature fusion, and the overall framework of the model is illustrated in the diagram. TePaDi consists of three main components: the global branch, the local branch, and the Global-Local Fusion Module (GLFM). The global branch primarily focuses on further extracting features from the global optical flow, providing the model with a comprehensive understanding from a global perspective. The local branch employs the Temporal Polar Pixel-wise Interaction module to fuse features from local polar image and local time image, achieving information fusion across spatial, temporal, and polarity dimensions in event features. The resulting global and local features are fed into a cross-attention module for feature fusion, ultimately yielding the MER results.

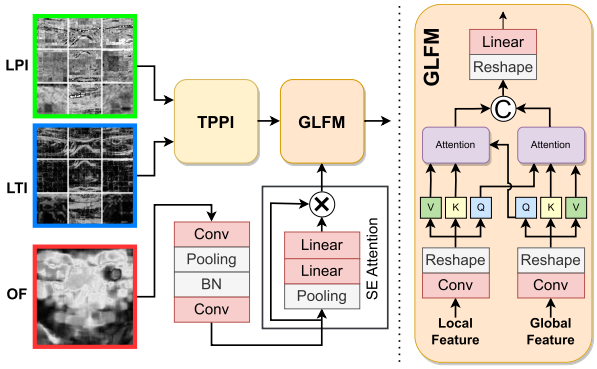


FIGURE 3. Overview of network architecture.

B. GLOBAL FEATURE CALCULATION BRANCH

The primary goal of the global feature branch is to provide a holistic representation of the subject's motion information from a global perspective. Given the swift pace and subtle motion amplitude of micro-expressions, attempting to discern a micro-expression solely from a static picture is nearly insurmountable. Moreover, processing the entire recorded video incurs substantial computational costs. To tackle these challenges, we opted for the dense Optical Flow method proposed by Farneb''ack in 2003 for the global feature branch.

The fundamental concept of the OF method involves employing polynomial expansion to approximate the neighborhood of each pixel, expressing the optical flow as $[u, v]$. To obtain more discriminative representations, we transformed these vectors into $[mag, ang]$ by converting Cartesian coordinates to polar coordinates.

The deep feature mapping involved in the computation of optical flow output comprises two convolution layers with a kernel size of 3, designed to capture intricate information while preserving computational efficiency. These convolution layers employ a progressively set stride of 2, with a strategically placed max-pooling layer between them to enhance speed and resource efficiency. To mitigate the risk of overfitting, a batch normalization layer follows the max-pooling operation. These sequential operations lead to a gradual reduction in the mapping feature size of each channel while compensating for the loss of feature capture ability by increasing the number of channels. To model channel dependencies and dynamically adjust response

values, we incorporate a squeeze-and-excitation attention mechanism.

The process of computing the global feature is articulated in Equation (8), where GF signifies the global feature derived from optical flow, $\mathcal{F}$ encompasses a series of convolution operations, and $\mathcal{O}$ denotes the dense optical flow operation.

$$GF = SE(\mathcal{F}(\mathcal{O}(x))) \tag{8}$$

C. LOCAL FEATURE CALCULATION BRANCH

In the preceding sections, we employed local normalization to transform the temporal image and polar image into local features, namely LTI and LPI. Now, we will use them as inputs for local feature calculation. To ensure comprehensive fusion of temporal information and polarity information, we have designed a temporal polar pixel-wise interaction module, as depicted in the Figure. 4.

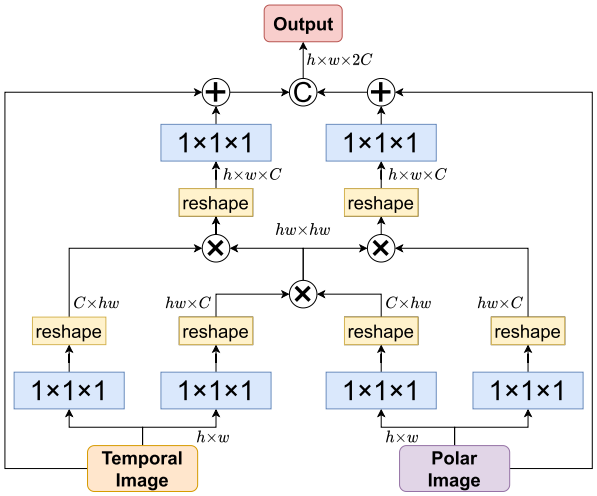


FIGURE 4. Temporal polar pixel-wise interaction module.

As our LTI and LPI features undergo the same local region normalization, we consider them to be aligned at the pixel-wise level. Therefore, in the TPPI module, our primary objective is to align and integrate temporal information and polar information. Initially, we utilize pixel-wise convolution for feature extraction, and the computed features undergo cross-multiplication to calculate their similarity matrix, as illustrated in Equation (9), where $\mathcal{S}$ stands for the similarity matrix, f_{pc} represents the pixel-wise convolution operation, and $\odot$ means the matrix multiply.

$$\mathcal{S} = f_{pc}(LTI) \odot f_{pc}(LPI), \tag{9}$$

Subsequently, based on the calculated similarity, we refine the two features separately. Finally, we adopt a residual connection method to avoid issues like gradient vanishing or exploding during the model training process. The specific calculation process is illustrated in Equation (10).

$$\begin{aligned} LTI &= LTI + f_{pc}(\mathcal{S} \odot f_{pc}(LTI)) \\ LPI &= LPI + f_{pc}(\mathcal{S} \odot f_{pc}(LPI)). \end{aligned} \tag{10}$$

Finally, we intergrate these two features together in channel dimension to form the local feature LF.

$$LF = [LTI, LPI] \tag{11}$$

D. GLOBAL-LOCAL FEATURE FUSION MODULE

Facial expression control often involves a series of muscle movements, which, when combined, shape the overall expression. To this end, we explore facial features from both local and global perspectives, aiming to comprehend both subtle local variations and broader overall changes, thereby facilitating better micro-expression recognition. Once global and local features are obtained, we designed the Global-Local Feature Fusion Module (GLFM) to mix them for deep information acquisition, as shown in Figure. 3. Firstly, we applied two convolutional operations to compress the features into 3 channels each, and then flattened them to form global and local queries (Q), keys (K), and values (V), denoted by subscripts g for global and l for local.

Subsequently, we exchanged Q_g and Q_l, and independently used attention mechanisms for feature extraction, as shown in the Equation (12). The exchange of features not only facilitates the fusion of global and local information but also guides the feature extraction process through the generation of fusion attention. Finally, we concatenated the features and used fully connected layers for information aggregation, resulting in the final recognition outcome.

$$\begin{aligned} GA &= (Q_l \odot K_g) \odot V_g \\ LA &= (Q_g \odot K_l) \odot V_l \\ Out &= fc([GA, LA]). \end{aligned} \tag{12}$$

V. EXPERIMENT DETAILS

n this section, we shall furnish comprehensive particulars of our experiments, encompassing dataset processing, experimental configurations, and the introduction of a random data augmentation strategy to enhance data diversity.

We opted for the evaluation of our method's performance using three datasets, namely the CASME II dataset [32], SMIC dataset [7], and SAMM dataset [8]. To ensure a robust evaluation, the leave one subject out cross-validation strategy was employed to partition the datasets into training and validation sets. It is noteworthy that the modest scale of the adopted datasets may give rise to model overfitting, posing a challenge, particularly in the training of large models. To mitigate this issue, we formulated five data augmentation methods and devised two distinct application strategies. The five data augmentation methods include random moving, noise addition, random flipping, random erasing, and grid mask, as depicted in Figure. 5.

To optimize the utilization of the suggested five data augmentation methods, we incorporated two distinct strategies: the fixed augmentation strategy and the random augmentation strategy. The fixed strategy employed all five methods collectively in a predefined sequence, whereas the random strategy involved the random selection of a number

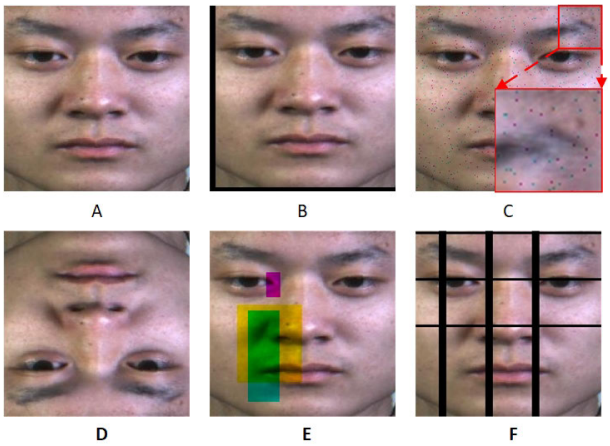


FIGURE 5. Random data augmentation method. A is the original image, and from B to E is image with random moving, noise adding, random flipping, random erasing and grid mask, respectively.

(ranging from 1 to 5) to specify the count of augmentation methods to be applied. With each selection, a single augmentation method was randomly chosen for implementation. The random augmentation strategy provides enhanced data augmentation capabilities, introducing a broader array of feature variations in the images, and has proven effective in mitigating and overcoming challenges associated with model overfitting.

A. EXPERIMENT CONFIGURATION AND EVALUATION METRICS

Adhering to the guidelines of the 2nd Micro-Expression Grand Challenge [27], our experiments consisted of two types: Holdout-Database Evaluation (HDE) and Composite Database Evaluation (CDE). In the CDE experiment, all three datasets were amalgamated into a single dataset for 3-class classification. In contrast, HDE involved 3-class classification experiments conducted separately for each of the three datasets. Notably, the CASME II and SAMM dataset were also utilized for a 5-class classification experiment.

To ensure fairness and avoid ambiguity stemming from facial features of different individuals, we adhered to the leave one subject out cross-validation principle. Each round of the experiment comprised 300 epochs, with an early stop strategy employed to prevent overfitting and conserve computational resources. The batch size, set to 250 for each epoch, utilized the maximum capacity of our GPU. Stochastic gradient descent was our optimizer of choice, with an initial learning rate of 0.01 and weight decay of 0.01. The learning rate underwent a 5% reduction every three epochs. Additionally, we fine-tuned the event threshold through experiments and set it to 1.0.

Importantly, the random data augmentation strategy was applied to the extracted features rather than the raw images. This precautionary measure aimed to prevent augmentation interference with the event and optical flow feature extraction processes. For model evaluation, we adopted metrics such as Unweighted Average Recall (UAR) and Unweighted F1-score (F1), following the MEGC 2019 guidelines [27] and Accuracy (Acc), as shown in the Equation (13), where C means the number of categories, N means the number of samples and the subscript c stands for each category.

$$\begin{cases} \text{UAR} = (\sum_{c=1}^{C} \frac{TP_c}{N_c})/C \\ \text{F1} = (\sum_{c=1}^{C} \frac{2TP_c}{2TP_c + FP_c + FN_c})/C \\ \text{Acc} = (\sum_{c=1}^{C} TP_c)/(\sum_{c=1}^{C} N_c) \end{cases} \tag{13}$$

B. EXPERIMENT RESULTS

resented in Table 1 are the results of the 3-class classification experiments, encapsulating both the HDE and CDE experiments. Our proposed method demonstrated superior performance in the composite database evaluation experiment, surpassing previous works with elevated unweighted average recall scores and F1 score. A comparative analysis with GLEFFN unmistakably highlights TePaDi's substantial performance improvement across multiple datasets, thereby affirming the efficacy of the two features we introduced.

From Table 2, in the multi-class classification recognition of independent datasets, our model also achieved good results. Among the total of 6 indicators, 3 indicators reached the state-of-the-art level, and the other 3 indicators were also at a relatively advanced level.

C. ABLATION STUDY

We chose to conduct ablation experiments on CASME II because this task is more challenging relative to others, providing a better opportunity to distinguish the strengths and weaknesses of different methods.

1) ABLATION STUDIES FOR EVENT FEATURE.

The ablation studies for LTI and LPI feature were carried out on 5-class classification task on CASME II dataset. The experiment result was shown as Table 3. When neither local feature was utilized, relying solely on optical flow as a global feature for micro-expression recognition yielded poor results. This highlights the strong dependence of MER tasks on local information. After separately incorporating LTI and LPI features, there was a noticeable improvement in the model's performance, suggesting that both temporal information and polarity information contribute positively to the model's effectiveness. Finally, when both LTI and LPI were simultaneously applied, the model achieved optimal results.

2) ABLATION STUDIES FOR DATA AUGMENTATION AND TEMPORAL UP-SAMPLING

We conducted ablation experiments focusing on data augmentation and temporal up-sampling, specifically in the five-class classification task of CASME II. The experimental results are summarized in the Table 4.

TABLE 1. Result of 3-class classification experiments.

	SMIC		CASME II		SAMM		3DB-combined	
Model	F1	UAR	F1	UAR	F1	UAR	F1	UAR
LBP-TOP [35]	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102	0.5882	0.5785
Bi-WOOF [22]	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139	0.6296	0.6227
OFF-ApexNet [9]	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392	0.7196	0.7096
Dual-Inception [38]	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663	0.7322	0.7278
STSTNet [21]	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810	0.7353	0.7605
SLSTT [11]	0.740	0.720	[0.901]	0.885	0.715	0.643	0.816	0.790
AU-GCN [18]	0.7292	0.7215	0.8798	0.8710	[0.7751]	0.7890	0.7914	0.7933
C3Dbed [26]	0.7760	0.7703	0.8978	0.8882	**0.8126**	**0.8067**	0.8075	0.8013
ME-PLAN [36]	N/A	N/A	0.8941	0.8962	0.7358	0.7687	0.7979	0.8041
FRL-DGT [34]	0.743	0.749	**0.919**	0.903	0.772	0.758	0.812	0.811
GLEFFN [10]	[0.7714]	[0.7856]	0.8825	**0.9110**	0.7458	0.7843	[0.8121]	[0.8208]
TePaDi[Ours]	**0.7830**	**0.7904**	0.8889	[0.9076]	0.7623	[0.7911]	**0.8266**	**0.8254**

TABLE 2. Result of multi-class classification experiments.

	SMIC		CASME II (5 class)		SAMM (5 class)	
Model	F1	Acc	F1	Acc	F1	Acc
LBP-TOP [35]	0.5384	0.5366	0.4241	0.4646	-	-
LBP-SIP [30]	0.4451	0.4492	0.4656	0.448	-	-
STCLQP [14]	0.6402	0.6381	0.5839	0.5836	-	-
SSSN [16]	0.6329	0.6341	0.7115	0.7119	0.4513	0.5662
DSSN [16]	0.6462	0.6341	0.7297	0.7078	0.4644	0.5735
GEME [25]	0.6158	0.6463	**0.812**	**0.8178**	0.711	0.7132
SLSTT [11]	-	-	0.753	0.7581	0.640	0.7239
MER-Supcon [37]	-	-	0.7286	0.7358	0.6251	0.6765
C3Dbed [26]	[0.7784]	[0.7804]	0.7520	[0.7764]	[0.7216]	**0.7573**
GLEFFN [10]	-	-	0.7564	0.7607	-	-
TePaDi[Ours]	**0.7833**	**0.7917**	[0.7677]	0.7718	**0.7642**	[0.7491]

TABLE 3. Ablation studies for event feature.

No.	LTI	LPI	F1	UAR	Acc
1	N	N	0.7166	0.7495	0.7183
2	N	Y	0.7592	0.7688	0.7657
3	Y	N	0.7601	0.7691	0.7698
4	Y	Y	**0.7677**	**0.7748**	**0.7718**

By comparing the first six experimental outcomes, it becomes evident that employing data augmentation with a random strategy leads to better model fitting results. This is because the random strategy introduces a more diverse range of augmentations to the features, effectively enriching the dataset and thereby enhancing model performance. Conversely, using a fixed strategy for data augmentation not only limits the forms of augmentation but also predisposes the model to data bias, ultimately hindering performance improvements.

Furthermore, through the comparison of temporal up-sampling at different scales, we observed that increasing the up-sampling frame rate effectively enhances experimental outcomes. However, after reaching a frame rate of 1000fps, the rate of improvement slows down significantly and even turns negative. Concurrently, the events required for event feature extraction also increase substantially. Balancing between performance and efficiency, we ultimately chose 1000 fps as the target for temporal up-sampling. The primary reason for this phenomenon is that at lower frame rates, event information is sparse and lacks sufficient information content. As the frame rate increases, the augmented information significantly improves model performance. However, once the frame rate reaches a certain threshold, further increases become less effective in providing additional meaningful information and may even introduce noise, resulting in model performance saturation and limited further improvements.

TABLE 4. Ablation studies for data augmentation and temporal up-sampling.

No.	Data Augmentation	Temporal Up-sampling	F1	UAR	Acc	Time Cost
1	Fixed	-	0.7125	0.7177	0.7099	-
2	Random	-	0.7268	0.7315	0.7239	-
3	Fixed	500 fps	0.7324	0.7378	0.7319	0.5×
4	Random	500 fps	0.7375	0.7464	0.7401	0.5×
5	Fixed	1000 fps	0.7586	0.7615	0.7664	1×
6	Random	1000 fps	[0.7677]	[0.7748]	[0.7718]	1×
7	Random	1500 fps	**0.7684**	0.7702	**0.7745**	1.5×
8	Random	2000 fps	0.7649	**0.7812**	0.7723	2×

VI. DISCUSSION

As trailblazers in the application of event features to micro-expression recognition tasks, we have achieved commendable results with remarkably low computing costs, paving the way for new possibilities in subsequent event camera applications. Nonetheless, several critical issues demand attention and merit further investigation:

i. Temporal and Polarity Information Encoding: While we successfully encoded temporal and polarity information separately, achieving a certain level of effectiveness, there is still considerable room for improvement in terms of computational efficiency. Future endeavors may explore the potential benefits of simultaneously encoding both types of information into the same feature.

ii. Artificial Intelligence Generated Content (AIGC): The rise in popularity of AIGC presents a valuable opportunity. Leveraging AIGC to expand the dataset can prove highly

advantageous. By utilizing AIGC to generate synthetic data, pre-training the model becomes feasible, allowing it to learn rich facial expressions and muscle features. Subsequently, manually collected data can be employed for fine-tuning, effectively addressing the challenge of data scarcity.

iii. Micro-Expression Recognition in Streaming Video: A notable gap currently exists between the proposed approach and real-world applications. Developing a Micro-Expression recognition method based on streaming video is of paramount importance to achieve real-time and accurate detection, bridging the divide between research and practical applications.

VII. CONCLUSION

In conclusion, our study addresses the challenges in micro-expression recognition posed by the limitations of frame-based cameras with low frame rates. We have introduced two innovative event features inspired by the operational principles of event cameras, encoding not only spatial information but also temporal and polarity information of events. The incorporation of these features has significantly contributed to the precision of MER tasks. Furthermore, our proposed global-local feature fusion network enhances the efficiency and effectiveness of feature interaction, providing a robust framework for micro-expression analysis. Experimental results across multiple datasets have demonstrated the remarkable performance of our approach, surpassing state-of-the-art methods. This work enriches the encoding of event features, showcasing their potential in enhancing micro-expression recognition and contributing to the future advancement of event camera technology.

VIII. ABBREVIATIONS

TABLE 5. Units for magnetic properties.

Abbreviation	Full Form
AIGC	Artificial Intelligence Generated Content
CDE	Composite Database Evaluation
GF	Global Feature
GLEFFN	Global-Local Event Feature Fusion Network
GLFM	Global-Local Fusion Module
HDE	Holdout-Database Evaluation
LF	Local Feature
LTI	Local Temporal Image
LPI	Local Polar Image
MER	Micro-Expression Recognition
OF	Optical Flow
TePaDi	Temporal-Polar Dynamics Network
TPPI	Temporal Polar Pixel-wise Interaction
UAR	Unweighted Average Recall

IX. DECLARATIONS

A. AVAILABILITY OF DATA AND MATERIALS

Not applicable.

B. COMPETING INTERESTS

The authors have no competing interests to declare that are relevant to the content of this article.

C. AUTHORS' CONTRIBUTIONS

Junwu Lin: methodology, software, validation, writing–reviewing, and editing; Cunhan Guo: conceptualization, methodology, software, writing–original draft preparation, visualization, and investigation; and Xiaofang Wu: data curation, validation, and editing.

ACKNOWLEDGMENT

(Junwu Lin and Cunhan Guo contributed equally to this work.)

REFERENCES

[1] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022.

[2] L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, F. Becattini, and A. Del Bimbo, "Neuromorphic event-based facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4109–4119.

[3] U. Bissarinova, T. Rakhimzhanova, D. Kenzhebalin, and H. A. Varol, "Faces in event streams (FES): An annotated face dataset for event cameras," *Sensors*, vol. 24, no. 5, p. 1409, 2024.

[4] H. Bulzomi, M. Schweiker, A. Gruel, and J. Martinet, "End-to-end neuromorphic lip-reading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 4100–4107.

[5] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[6] G. Chavali, S. K. N. V. Bhavaraju, T. Adusumilli, and V. G. Puripanda, "Micro-expression extraction for lie detection using Eulerian video (motion and color) magnication," Blekinge Inst. Technol., 2014.

[7] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[8] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.

[9] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019.

[10] C. Guo and H. Huang, "GLEFFN: A global-local event feature fusion network for micro-expression recognition," in *Proc. 3rd Workshop Facial Micro-Expression, Adv. Techn. Multi-Modal Facial Expression Anal.*, New York, NY, USA, Nov. 2023, p. 1724.

[11] F. Guowen and L. Xi, "Micro-expression recognition based on dual branch neural network," in *Proc. 2nd Int. Conf. Artif. Intell. Comput. Inf. Technol. (AICIT)*, Sep. 2023, pp. 1–4.

[12] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikäinen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1–9.

[13] X. Huang et al., "Texture description with completed local quantized patterns," in *Proc. 18th Scandinavian Conf. Image Anal. (SCIA)*, vol. 1. Espoo, Finland: Springer, Jun. 2013, p. 7944.

[14] 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, Jan. 2016.

[15] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9000–9008.

[16] H.-Q. Khor, J. See, S.-T. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 36–40.

[17] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 382–386.

[18] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1571–1580.
[19] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, Nov. 2018.
[20] X. Li, G. Wei, J. Wang, and Y. Zhou, "Multi-scale joint feature network for micro-expression recognition," *Comput. Vis. Media*, vol. 7, no. 3, pp. 407–417, Apr. 2021.
[21] 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 *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[22] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process., Image Commun.*, vol. 62, pp. 82–92, Mar. 2018.
[23] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–4.
[24] G. Moreira, A. Graça, B. Silva, P. Martins, and J. Batista, "Neuromorphic event-based face identity recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 922–929.
[25] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "GEME: dual-stream multi-task GEnder-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, Feb. 2021.
[26] H. Pan, L. Xie, and Z. Wang, "C3DBed: Facial micro-expression recognition with three-dimensional convolutional neural network embedding in transformer model," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106258.
[27] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019—The second facial micro-expressions grand challenge," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[28] W. Shariff, M. S. Dilmaghani, P. Kielty, J. Lemley, M. A. Farooq, F. Khan, and P. Corcoran, "Neuromorphic driver monitoring systems: A computationally efficient proof-of-concept for driver distraction detection," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 836–848, 2023.
[29] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro- and micro-expression spotting in video using strain patterns," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–6.
[30] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 525–537.
[31] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 8590–8605, 2020.
[32] 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, Jan. 2014, Art. no. e86041.
[33] C. Yang, P. Liu, G. Chen, Z. Liu, Y. Wu, and A. Knoll, "Event-based driver distraction detection and action recognition," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2022, pp. 1–7.
[34] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22086–22095.
[35] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[36] 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 Netw.*, vol. 153, pp. 427–443, Sep. 2022.
[37] R. Zhi, J. Hu, and F. Wan, "Micro-expression recognition with supervised contrastive learning," *Pattern Recognit. Lett.*, vol. 163, pp. 25–31, Nov. 2022.
[38] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[39] D. Cai, K. Chen, Z. Lin, D. Li, T. Zhou, Y. Ling, and M.-F. Leung, "JointSTNet: Joint pre-training for spatial-temporal traffic forecasting," *IEEE Trans. Consum. Electron.*, early access, Oct. 2024, doi: 10.1109/TCE.2024.3476129.

JUNWU LIN received the Ph.D. degree from Xiamen University, China. He is an Associate Professor of intelligent science and technology with Putian University. He hosts or participates in multiple scientific research projects and has published several articles in domestic and foreign journals. His primary research interests include artificial intelligence and intelligent control.

CUNHAN GUO is currently pursuing the Ph.D. degree in computer applied technology with the University of Chinese Academy of Sciences. He holds multiple invention patents and has published several papers in international journals and conferences. His primary research interests include multimodal cognition and computer vision.

XIAOFANG WU is currently pursuing the Postgraduate degree in mechanical engineering with Putian University. Her research focus is mainly on image processing based on deep learning.