Contents lists available at ScienceDirect

Neurocomputing

HTNet for micro-expression recognition

Zhifeng Wang [a],*, Kaihao Zhang [a], Wenhan Luo [b], Ramesh Sankaranarayana [a]

[a] *College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia*
[b] *Sun Yat-sen University, Guangzhou, China*

A R T I C L E   I N F O

A B S T R A C T

Facial expression is related to facial muscle contractions and different muscle movements correspond to different emotional states. For micro-expression recognition, the muscle movements are usually subtle, which has a negative impact on the performance of current facial emotion recognition algorithms. Most existing methods use self-attention mechanisms to capture relationships between tokens in a sequence, but they do not take into account the inherent spatial relationships between facial landmarks. This can result in sub-optimal performance on micro-expression recognition tasks. Therefore, learning to recognize facial muscle movements is a key challenge in the area of micro-expression recognition. In this paper, we propose a Hierarchical Transformer Network (HTNet) to identify critical areas of facial muscle movement. HTNet includes two major components: a transformer layer that leverages the local temporal features and an aggregation layer that extracts local and global semantical facial features. Specifically, HTNet divides the face into four different facial areas: left lip area, left eye area, right eye area and right lip area. The transformer layer is used to focus on representing local minor muscle movement with local self-attention in each area. The aggregation layer is used to learn the interactions between eye areas and lip areas. The experiments on four publicly available micro-expression datasets show that the proposed approach outperforms previous methods by a large margin. The codes and models are available at: https://github.com/wangzhifengharrison/HTNet.

1. Introduction

Micro-expression refers to subtle muscle movements that last only for approximately 1/25–1/5 s. In recent years, extensive research has been conducted on utilizing computer vision-based methods to analyze micro-expressions [1,2]. However, the accuracy of recognizing micro-expressions still requires improvement. Even though the micro-expression datasets are collected in well-controlled laboratory environments, the current results are still unsatisfactory [3]. Utilizing computer vision for this task remains challenging due to the presence of subtle muscle movements that often accompany micro-expressions, making them difficult for both humans and computers to detect. On the other hand, the recognition of normal macro expressions has achieved high accuracy rates (over 95%) [4]. This stark difference in performance can be attributed to the fact that micro-expressions are incredibly hard to analyze due to their fleeting nature and subtle characteristics. As a result, researchers are still working to improve the accuracy and reliability of micro-expression recognition using computer vision techniques.

In the domain of micro-expression recognition, several researchers have proposed the adoption of the Local Binary Pattern (LBP) method

to extract facial features. The LBP technique has demonstrated commendable discrimination ability through its texture-based feature extraction approach, while maintaining a low computational complexity [5]. Moreover, other scholars have explored the use of optical flow features as inputs for estimating muscle motion [5]. Optical flow is derived from the differences in brightness between consecutive frames, enabling the estimation of subtle facial movements. Several optical flow-based methodologies have been investigated, including Bi-WOOF [6], MDMO [7], FHOFO [8], Optical Strain Weight, and Optical Strain Feature [9]. Furthermore, notable deep learning models such as VGG16 [10], GoogleNet [11], AlexNet [12], and OFF-Apex [3] have been utilized to process the TV-L1 optical flow. This optical flow is extracted from selected apex and onset frames of the image sequences. The apex frame, capturing the most salient information about the micro-expression, proves to be instrumental in accurately recognizing the expression.

In the field of facial expression recognition, analyzing micro-expressions is challenging due to the small and subtle nature of the facial movements. This makes it difficult to identify the specific facial muscles involved in the expression and monitor the pixel's motion over time. To
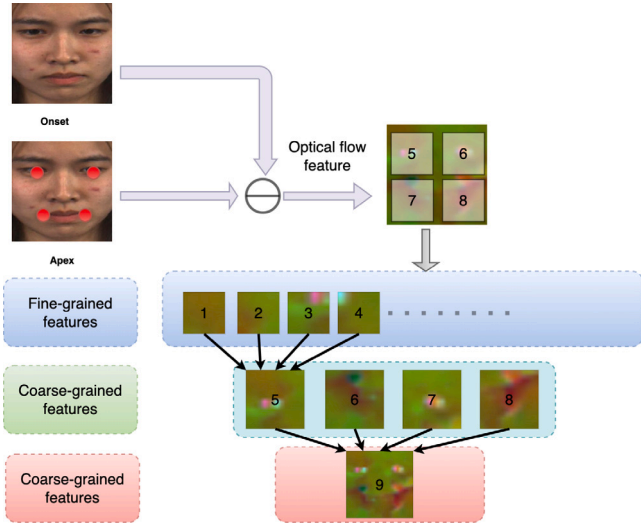

Fig. 1. The proposed HTNet paradigm. In the first step, we use onset frame and apex frame to get optical flow features. In the second step, we get middle level coarse-grained features by using facial landmarks. The middle level coarse-grained features will be evenly divided into four parts, which is fine-grained features. We use aggregation blocks to incorporate both local fine-grained and global coarse-grained interactions at different image scales. With this new mechanism, each pixel in same-level blocks is treated at a fine granularity, while the pixels in upper-level blocks are at a coarse granularity. This allows our model to effectively capture both short- and long-range visual dependencies.

address these issues, Zhou et al. [13] suggest using a feature refinement network, specific expression feature learning network, and fusion techniques for expression recognition. By employing self-attention on global image features, their method can extract distinguishing characteristics for micro-expression recognition. However, global self-attention between image features will lack of fine-grained features in multiple scale. In addition, they do not take into account the inherent spatial relationships between facial landmarks. This can result in sub-optimal performance on micro-expression recognition tasks. To deal with these issues, our objective is to maintain self-attention in the local blocks at each hierarchy level to capture fine-grained features in the Fig. 1. We use aggregation blocks to incorporate both local fine-grained and global coarse-grained interactions at different image scales. With this new mechanism, each pixel in same-level blocks is treated at a fine granularity, while the pixels in upper-level blocks are at a coarse granularity. This allows our model to effectively capture both short- and long-range visual dependencies. In this study, we present a novel self-attention method utilizing Transformer layers to effectively capture both local and global interactions within a hierarchical structure. The low-level self-attention in the Transformer layers aims to capture fine-grained features within local regions. On the other hand, the high-level self-attention in these layers is designed to capture coarse-grained features spanning global regions. To facilitate interactions between different blocks at the same level, we propose an aggregation block. The overall architecture of our proposed method, referred to as HTNet, is illustrated in Fig. 2. Our contributions in this study can be summarized as follows:

(a) We introduce a novel self-attention mechanism of Transformer layers that effectively captures both local and global interactions within a hierarchical structure for recognizing micro-expressions in images. This is achieved by utilizing the proposed block aggregation function. The low-level self-attention focuses on capturing fine-grained features in local regions, while the high-level self-attention targets coarse-grained features in global regions.

(b) Our network specifically focuses on four facial areas – the left eye region, left lip region, right eye region, and right lip region – instead of considering the entire facial region. This approach helps

to mitigate the impact of background noise at edges of facial areas that may be picked up by the lab camera. The use of Transformer layers allows us to concentrate on modeling small-scale, subtle muscle motions within each area through local self-attention. Additionally, the aggregation layer learns the interactions between the eye areas and lip areas. We conduct experiments to investigate how varying block sizes affect the accuracy of micro-expression recognition.

(c) Through experiments conducted on four available datasets, we demonstrate that our proposed method outperforms previous approaches noticeably. This highlights the effectiveness and superiority of our model for micro-expression recognition tasks.

2. Related work

2.1. Handcrafted features

In the upcoming subsections, we will discuss two main types of handcrafted features: appearance-based features and geometric-based features.

2.1.1. Appearance-based features

In conventional techniques for expression recognition, appearance-based characteristics are commonly utilized [14–16]. One prevalent pattern is the local binary pattern from three orthogonal planes (LBP-TOP) [17]. In some LBP-TOP works, LBP-TOP is transferred to the tensor-independent RGB space, which enhances robustness [18]. To achieve lower computational complexity, LBP-SIP effectively minimizes redundancy in LBP-TOP patterns, providing a lightweight representation [19]. TICS [20] introduces a new color model called Tensor Independent Color Space (TICS) to improve how well they can recognize true emotions. Their method treats a color video clip of a micro-expression as a four-dimensional array, where they change the usual RGB color dimension into TICS for more independent color components, leading to more accurate recognition of these quick expressions.

2.1.2. Geometric-based features

Geometric-based features are typically divided into two categories: optical flow based and texture variations based features. These geometric-based features can identify motion deformations. Li et al. [21] explore a deep learning technique for localizing face landmarks and dividing the facial area into interest zones. Since facial micro-expressions are produced by facial muscular contractions, assessing the orientation of these contractions is crucial for identifying emotions. Using this technique, they categorize the face into different areas of concern that correlate to different muscle action patterns. Furthermore, the facial area's activity can be adequately reflected by the optical flow. Liu et al. [7] propose the use of a powerful MDMO feature extraction network for micro-expression identification. They calculate the optical flow for each image in the video sequence and divide the facial zones into various intriguing portions. Mean optical flow features are then calculated from each image, and an SVM classifier is employed on these mean optical flow features for emotion identification. Their method effectively considers both geographical position and regional statistical movement information, proving to be straightforward and efficient. Rather than using the LBP histogram, Liong et al. [6] suggest weighting the histogram of LBP and averaging the value of optical flow features for recognizing emotions on the face, namely Bi-WOOF. Wang et al. [20] proposed utilising dynamic textures on tensor independent color space for micro-expression recognition. They improves micro-expression recognition by transforming color video clips into a four-dimensional array to analyze spatial, temporal, and independent color components, yielding higher accuracy than conventional RGB color space. Lu et al. [22] presents a novel Delaunay-based temporal
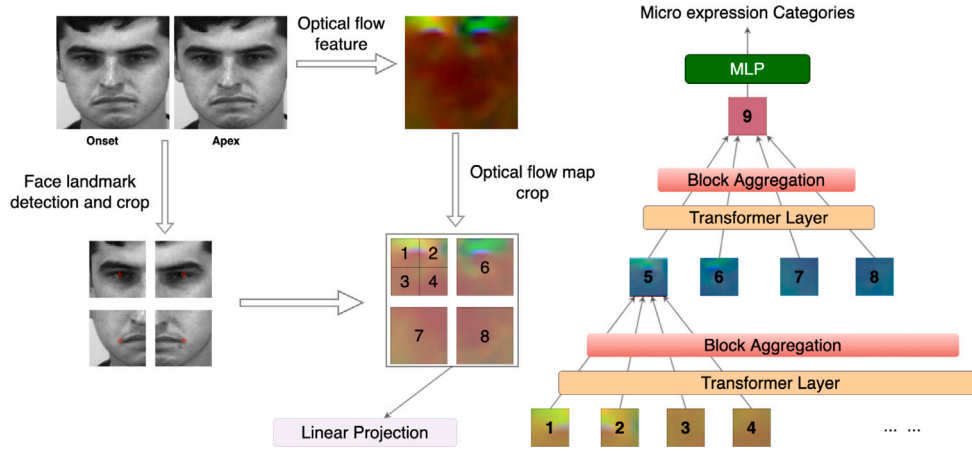
Fig. 2. HTNet: Overall architectures of hierarchical transformer network for micro-expression recognition. Low-level self-attention in transformer layers captures fine-grained features in local regions. High-level self-attention in transformer layers captures coarse-grained features in global regions. An aggregation block is proposed to create interactions between different blocks at the same level.

coding model (DTCM) that effectively encodes texture variations corresponding to muscle activities by normalizing facial image sequences and suppressing irrelevant features, leading to improved recognition rates, as validated by extensive experiments on major datasets. However, handcrafted feature extraction usually involves using extractors that are designed by experts, and this process often requires specialized knowledge.

2.2. Deep learning methods

In recent years, an array of deep learning techniques has emerged for extracting facial characteristics in expression identification [23–26]. Xia et al. [27] proposed the application of recurrent convolutional networks to establish connections between facial position information and record facial muscle contractions across various regions. This model incorporates multiple recurrent convolutional layers and a classification layer to capture visual characteristics for facial emotion recognition. Gan et al. [3] proposed an automated approach to localize the apex frame and employ optical flow as input for their OFF-ApexNet. Leveraging CNN, their network extracts new feature descriptors from optical flow. On the other hand, some researchers have explored the adoption of shallow CNNs, which effectively derive high-level visual attributes from three components of movement estimation — horizontal, vertical optical flow fields, and optical strain for inferring emotional states [28]. To address the issue of composite-database domain shift, Xia et al. [29] developed an RCN model to examine how a smaller model affects the recognition of micro-expressions. Within the RCN, they designed three parameter-free modules, including an attention unit, shortcut connection, and wide expansion, to prevent an increase in the number of learnable parameters. However, convolution, a process used in image analysis, always works on fixed window sizes, meaning it may not pick up on relationships or patterns that are far apart. To tackle this, Kumar et al. [30] use a novel two-stream graph attention network that utilizes relationships between facial landmarks and optical flow patches to detect and classify subtle and brief facial micro-expressions. In addition, Zhang et al. [31] introduce a new type of architecture using transformers that does not rely on traditional convolutional networks for recognizing micro-expressions. This architecture includes a spatial encoder to learn about space patterns, a temporal aggregator to analyze timing, and a classification head. But, their approach overlooks the spatial relationships between facial landmarks, which potentially lead to less-than-ideal results in micro-expression recognition. Meanwhile, Lei et al. [23] introduced the use of Transformer as an encoder to model the connections between nodes and edges of the face, constructing a face graph based on the edges and nodes of facial features. However,

they do not consider extracting features in a hierarchical structures. Local fine-grained and global coarse-grained interactions at different image scales are missing.

3. Proposed method

As illustrated in Fig. 2, the proposed HTNet consists of transformer layers and block aggregation at each level of the hierarchical structure. The transformer layer performs self-attention on each image block independently. At the low-level network, the self-attention function in the transformer layer captures fine-grained features. Subsequently, the block aggregation process aggregates small image blocks into larger ones, allowing for the creation of interactions between different blocks at the same level. This aggregation leads to the capture of coarse-grained features after each block aggregation. It is important to note that all the blocks within the same level share the same set of parameters. Finally, the MLP block in our model is applied to the final feature maps for micro-expression classification. This modular and hierarchical design enables HTNet to effectively extract and integrate features at different scales and levels of granularity for improved micro-expression recognition performance.

3.1. Apex frame spotting and optical flow map extraction

Optical flows, generated from the onset and apex frames, serve as a valuable means to describe movement displacements in facial areas. These optical flows have demonstrated promising results in micro-expression recognition datasets [3,28].

In order to obtain an optical feature map, it is essential to use the onset and apex frame indices. However, in the case of micro-expression datasets, the onset frame index is already provided, necessitating the determination of only the apex frame index from video sequences. To achieve this, we adopt the D&C-RoIs method, which has been used in prior research on micro-expression [3] if the dataset do not provide the index of apex frame. The D&C-RoIs approach effectively establishes the relationship between the onset frame and subsequent frames, allowing for the accurate identification of the apex frame index. This ensures the reliable extraction of optical flow features for subsequent micro-expression recognition tasks.

$$d = \frac{\sum_{i=1}^{B} h_{1i} \times h_{h2i}}{\sqrt{\sum_{i=1}^{B} h_{1i}^2 \times \sum_{i=2}^{B} h_{2i}^2}}, \tag{1}$$

where B is the number of bins in gray-scale histograms, h_1 is gray-scale histogram of the first frame, and h_2 is the current frame. d is the rate of difference of the LBP features between two frames. The highest rate of

difference for the LBP features will be selected, which can identify the index of the frame in which the most facial movements occur [28,32].

Then, we obtain optical flow feature maps using the onset and apex frames. It is possible to formulate the optical flow feature map as follows:

$$V = (u(x, y), v(x, y))|x = 1, 2, \ldots, X, y = 1, 2, \ldots, Y, \tag{2}$$

where X and Y represent the frame's width W and height H respectively, $u(x, y)$ and $v(x, y)$ is the horizontal and vertical component of optical flow feature map V, $V = [V_x, V_y]$, where $V \in R^{W \times H \times 2}$.

In our approach, we employ the first-order derivatives of the optical flow field to calculate the variations in the optical flow fields, commonly known as optical strain. The optical strain provides an estimation of the degree of facial displacement, thereby offering valuable insights into the subtle movements that occur during micro-expressions. This computation of optical strain allows us to capture and analyze the intricate facial dynamics, contributing to the accurate recognition of micro-expressions:

$$V_z = \sqrt{\frac{\partial V_x}{\partial x}^2 + \frac{\partial V_y}{\partial y}^2 + \frac{1}{2}(\frac{\partial V_x}{\partial y}^2 + \frac{\partial V_y}{\partial x}^2)}, \tag{3}$$

where $\frac{\partial V_x}{\partial x}^2$, $\frac{\partial V_y}{\partial y}^2$, $\frac{\partial V_x}{\partial y}^2$ and $\frac{\partial V_y}{\partial x}^2$ are the partial first-order derivatives of V. Finally, three-dimension optical flow feature maps are formed and represented as $V_m = [V_x, V_y, V_z]$ and $V_m \in R^{W \times H \times 3}$.

In our network, we adopt a region-specific approach by focusing on four specific facial areas – the left eye region, left lip region, right eye region, and right lip region – instead of considering the entire facial region. To extract the four specific facial optical flow feature maps from the entire optical flow feature maps V_m, we employ the Multi-task Cascaded Convolutional Networks (MTCNN) [33] to obtain the face landmark coordinates from the apex images. Then, we use the coordinates of facial landmarks as center to get four facial optical flow maps. Specifically, the optical flow map for the left eye is centered on the left-eye facial landmark, and similarly, the right-eye optical flow map is centered on the right-eye facial landmark. The four facial optical flow feature maps are subsequently cropped from V_m, representing the left-eye optical flow feature map, left-lip visual feature map, right-eye optical flow feature map, and right-lip feature map. Each of these four feature maps has a size of $\frac{W}{2} \times \frac{H}{2} \times 3$, which is half the size of the whole optical flow image. Following the extraction of the four optical flow features, we combine them and feed the combined features into our HTNet for micro-expression recognition. The entire process is visually depicted in Fig. 2.

3.2. Transformer layer

In our approach, each image block has a size of $P \times P$ when processing an input optical flow image with dimensions $H \times W \times 3$. After linear projection and partitioning on the optical flow images, each patch has a feature dimension of $P \times P \times 3$. Subsequently, the patches are flattened, resulting in an input for our model denoted as $X \in \mathbb{R}^{b \times H_n \times n \times d}$, where H_n represents the number of blocks in each level, b is the batch size, n is the sequence length, and $H_n \times n = \frac{H \times W}{P^2}$. Multiple transformer layers are applied within each block, and the hierarchy determines the number of transformer layers utilized. Each transformer layer consists of Layer Normalization (LN), Multi-Head Self-Attention (MSA) layer, and Feed-Forward Fully Connected Network (FFN). To encode spatial information, a trainable positional embedding vector is incorporated into all sequence vectors in $\mathbb{R}^d$. This ensures that spatial relationships and positional information are effectively captured and encoded within the feature representations:

$$\begin{aligned} Y_l'^i &= Y_{l-1}^i + MSA(LN(Y_{l-1}^i)) \\ Y_l^{i+1} &= Y_l'^i + FFN(LN(Y_l'^i)) \\ Y_l'^{i+1} &= Y_l^{i+1} + MSA(LN(Y_l^{i+1})) \\ Y_l^{i+2} &= Y_l'^{i+1} + FFN(LN(Y_l'^{i+1})) \end{aligned}, \tag{4}$$
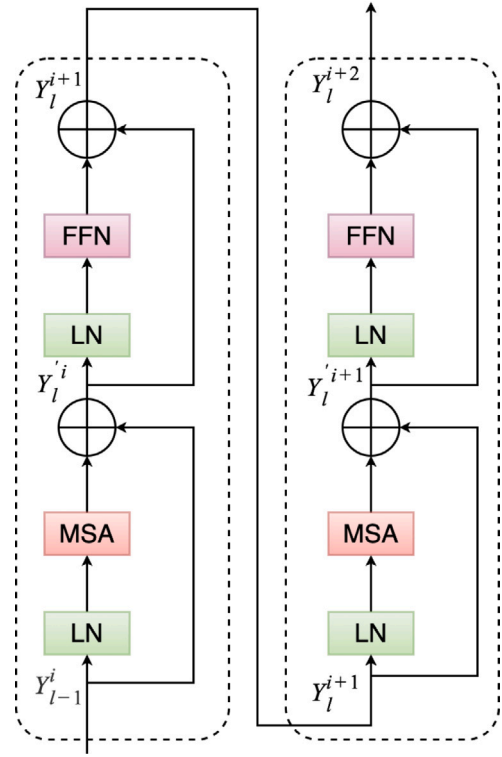

Fig. 3. Several transformer layers will be applied in each block in parallel. Hierarchy determines the amount of transformer layers. The transformer layers is composed of Layer normalization (LN), multi-head self-attention (MSA) layer and feed-forward fully connected network (FFN). The spatial information will be encoded by adding a trainable positional embedding vector to all sequence vectors in R^d.

where $l = 1, 2, \ldots, L$, l is the index of the lth block in each hierarchy layer i, and L is the overall number of blocks in each hierarchy layer(see Fig. 3). The FFN includes two layers: $max(0, xW_1 + b)W_2 + b$. At each block i within the same level, the multi-head self-attention mechanism is applied. In this self-attention component, the input $X \in \mathbb{R}^{n \times d}$ is transformed into three parts, namely queries Q, keys K, and values V, where n represents the sequence length and d is the dimension of the inputs. Subsequently, the scaled dot-product attention is employed on Q, K, and V:

$$MSA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V. \tag{5}$$

LN will be applied in each block as follows:

$$LN(x) = \frac{x - \mu}{\delta} o\lambda + \beta, \tag{6}$$

where μ is the mean of features and δ is standard deviation of the feature, o is the element-wise dot and λ and β are learnable parameters.

Following the transformer layer, we employ block aggregation to merge the output of the transformer layer. Specifically, we group every four small blocks into one larger block through the block aggregation process.

3.3. Block aggregation

The block aggregation function utilized in our HTNet shares similarities with several Pyramid designs. However, a notable distinction is that our model employs local attention on each image block, rather than global attention on the entire image. This approach proves to be beneficial for enhancing the model's performance since micro-expression recognition heavily relies on localized facial muscle motion
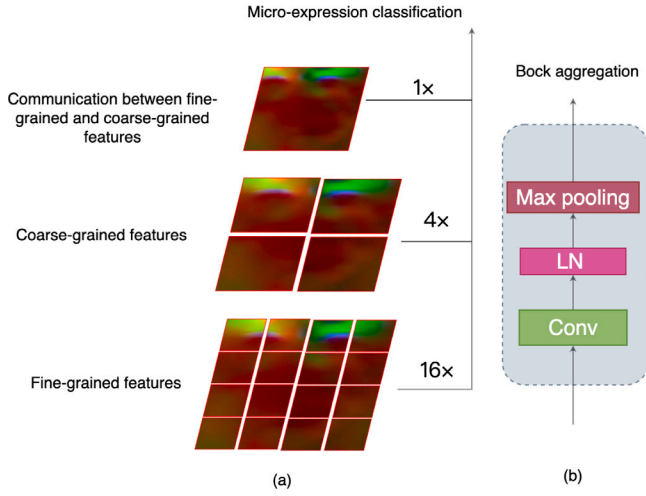
Fig. 4. The block aggregation include 3×3 convolutional layer and followed by LN and 3×3 max pooling. At the bottom of our model, the facial optical flow map includes 16 facial blocks, which are 4×4 feature maps. By using 3×3 convolutional layer on this feature map, partial feature maps within four facial areas will be merged and the block size will become 2×2 blocks which corresponding to four facial areas-left eye region, left lip region, right eye region and right lip region.

Table 1

The experiments are implemented on SAMM [34], SMIC [35], CASME II [36] and CAS(ME)³ [37] databases. SAMM, SMIC, and CASME II are merged into one composite dataset, and the same labels in these three datasets are adopted for micro-expression tasks.

Database	SAMM	CASME II	SMIC	CAS(ME)³
Subjects	28	24	16	100
Samples	133	145	164	943
Frame rate	200	200	100	30
Negative	92	88	70	508
Positive	26	32	51	64
Surprise	15	25	43	201
Onset index	✓	✓	✓	✓
Offset index	✓	✓	✓	✓
Apex index	✓	✓	×	✓

areas. By focusing on specific face regions and leveraging local attention, our model effectively captures the essential features for inferring micro-expression states while disregarding irrelevant facial areas.

In our HTNet model, each block independently processes the optical flow map, and block communication occurs during block aggregation. Block aggregation operates on four neighboring blocks, facilitating the exchange of information between them and enabling the fusion of local and global features. Specifically, low-level block aggregation focuses on exchanging information within the four facial areas — left eye region, right eye region, left lip region, and right lip region. This process extracts fine-grained features that capture detailed facial dynamics. On the other hand, high-level block aggregation facilitates global information exchange between the four facial areas, resulting in the extraction of coarse-grained features that capture broader facial expressions. The overall process is depicted in Fig. 4. At hierarchy l, the optical flow image size is represented as $X_l \in \mathbb{R}^{b \times h \times w \times d^l}$. After block aggregation, the optical flow image size becomes $X_{l+1} \in \mathbb{R}^{b \times 2h \times 2w \times d^{l+1}}$. Subsequently, all four facial areas are merged into one facial feature map, denoted as $X \in \mathbb{R}^{b \times H \times W \times D}$, with $d^{l+1} > d^l$ to preserve and enhance the features effectively.

The block aggregation process in HTNet consists of a 3×3 convolutional layer, followed by Layer Normalization (LN) and 3×3 max pooling. At the bottom of our model, the facial optical flow map consists of 16 facial blocks, each represented by a 4×4 feature map. The first 3×3 convolutional layer merges partial feature maps within the four facial areas, reducing the block size to 2×2, corresponding to the left eye area, left lip region, right eye area, and right lip region. Subsequently, another 3×3 convolutional layer facilitates information exchange between the four facial areas, further reducing the block size to 1×1, and extracting full optical feature maps. Finally, the extracted full feature maps are fed into the MLP layer for micro-expression classification. This hierarchical process effectively captures and integrates features at different levels of granularity, contributing to improved micro-expression recognition.

3.4. Loss function

In this paper, we employ the cross-entropy loss function to train our model. The computation of the cross-entropy loss L can be expressed by the following formula:

$$L = \sum_i (-w_i log(p_t^i)),$$
$$p_t = p_c^y (1 - p_c)^{1-y}, \tag{7}$$

where w_i is the weight of samples in the dataset, y is the label and $y_i \in 0, 1$.

4. Experiments

4.1. Datasets

The experiments are conducted on four databases: SAMM [34], SMIC [35], CASME II [36] and CAS(ME)³ [37] databases. To ensure consistency and comparability, SAMM, SMIC, and CASME II are merged into a composite dataset, where the same emotion labels from these three datasets are adopted for the micro-expression recognition tasks. In these datasets, the emotion categories are divided as follows: the "positive" emotion category includes the "happiness" emotion class, and the "negative" emotion category includes "sadness","disgust", "contempt", "fear" and "anger" emotion classes while "surprise" emotion category only includes "surprise" class.

SAMM [34]: The SAMM dataset comprises 28 participants, 133 micro-expressions, and 147 long videos featuring 343 macro-expressions. The dataset is rich in Action Units coding, providing comprehensive facial expression information. The onset, offset, and apex indices for micro-expressions are also provided in SAMM. The original samples in the dataset have a resolution of 2040 by 1088 pixels, and the frame rate is set at 200 frames per second. Emotion categories in images in SAMM are divided into categories including "disgust", "fear", "contempt", "angry", "repression", "surprise", "happiness", and "others". After classifying into three emotion categories, the number of "negative", "positive", and "surprise" is 92, 26,15.

CASME II [36]: The CASME II dataset consists of data from 24 subjects, totaling 145 samples corresponding to 145 emotions. All samples are captured using lab cameras, with a frame rate set at 200 frames per second. The original size of the samples is 640×480 pixels. Samples in CASME II are divided into categories such as "happiness", "surprise", "disgust", "sadness", "fear", "repression", and "others". After merging into three emotion categories, the number of "negative", "positive", and "surprise" is 88, 32,25. The onset, offset, and apex index are annotated in CASME II.

CAS(ME)³ [37]: CAS(ME)³, the third generation of the Facial Spontaneous Micro-Expression database, is distinguished by its inclusion of depth information and high ecological validity, making it a valuable resource for micro-expression recognition. CAS(ME)³ Part A comprises data from 100 subjects, totaling 943 samples corresponding to 943 emotions. The samples are captured using a lab camera, with a frame rate of 30 frames per second, and have an original resolution of

1280×720 pixels. Samples in CAS(ME)^3 part A are categorized into "happiness", "anger", "fear", "disgust", "surprise", "others" and "sadness". The total number of "negative", "positive", and "surprise" is 508, 64, and 201.

SMIC[35]: The SMIC-HS dataset comprises data from 16 subjects, totaling 164 samples corresponding to 164 emotions. All samples are captured using a lab camera with a frame rate of 100 frames per second. The original image size of the samples is 640×480 pixels. Samples in SMIC are categorized into "negative", "surprise" and "positive". The number of "negative", "positive", and "surprise" is 70, 51, 43. The onset and offset are given in SMIC, but the apex index is not given in SMIC. The detailed information of these three datasets can be summarized in Table 1

4.2. Implementation details

Initially, in the case of the SMIC dataset, where the apex frame index is missing, we employed the D&C-RoIs technique [38] to determine the index of the apex frame. For the SAMM, CASME II, and CAS(ME)^3 datasets, the ground truth for the apex frame is available, which simplifies the process of obtaining the crucial micro-expression moment. After obtaining the onset and apex images from the dataset, we utilized Gunnar Farneback's algorithm [39] to extract optical flow from these images at both the onset and apex time points. Next, the three elements of optical flow images – horizontal, vertical, and optical strain elements – are with $28 \times 28 \times 3$ pixels. Afterward, we employ the Multi-task Cascaded Convolutional Networks (MTCNN) [33] to extract face landmark coordinates from the apex images. These face landmark coordinates are vital for localizing specific facial regions with high accuracy. Based on the face landmark coordinates, we extract four important face areas, namely the left-eye, left-lip, right-eye, and right-lip optical flow feature maps. These four optical flow feature maps are half the size of the entire optical flow image, measuring $14 \times 14 \times 3$ pixels. By focusing on these specific facial regions, we can effectively capture the relevant facial muscle movements associated with micro-expressions. Following the extraction of the four optical flow feature maps, our HTNet network receives and combines them. This comprehensive approach ensures that our HTNet is well-equipped to recognize and classify micro-expressions accurately and efficiently.

The experiments are conducted using PyTorch and Python 3.9 on Ubuntu 22.04 operating system. We set the learning rate for the training parameters to 5×10^{-5} and use a maximum of 800 epochs for training.

Setup: The standard evaluation method for micro-expression tasks is the leave-one-subject-out (LOSO) cross-validation. LOSO cross-validation is preferred as it allows for a fair comparison between different models and ensures that the model's performance is not biased by specific subject characteristics. This approach closely simulates real-world scenarios, where individuals with diverse backgrounds and expressions are encountered in various settings and locations.

Performance metrics The class distribution in the composite micro-expression dataset is unbalanced, with different emotions having varying frequencies. Specifically, the rate of surprise, positive, and negative emotions is approximately 1:1.3:3, respectively. To address this issue, we employ the Unweighted Average Recall (UAR) and Unweighted F1 score ($UF1$) to report our results.

(1) The Unweighted F1-score ($UF1$), also known as the macro-average F1-score, is a metric that is commonly used to evaluate performance in multi-class classification tasks with imbalanced class distributions. To calculate the $UF1$, we need to compute the False Positives (FP), True Positives (TP), and False Negatives (FN) for each class c in all folds of the leave-one-subject-out

(LOSO) cross-validation. Then, the F1-score for each class $F1_c$ can be computed using the formula:

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c}$$
$$UF1 = \frac{F1_c}{C} \tag{8}$$

In the formula (8), the C is the number of classes.

(2) Unweighted Average Recall (UAR): UAR is a metric that is particularly useful when evaluating the effectiveness of a model in the presence of imbalanced class ratios. n_c is the total number of samples in each class. The UAR can be defined as follows:

$$UAR = \frac{1}{C} \sum_c \frac{TP_c}{n_c} \tag{9}$$

4.3. Comparison to state of the arts

In Tables 2 and 3, we present the performance of our proposed method in comparison with previous handcrafted and deep learning methods on the micro-expression datasets, namely CASME II, CAS(ME)^3, SMIC, and SAMM. The evaluation metrics used are Unweighted F1-score ($UF1$) and Unweighted Average Recall (UAR). The results are reported. Bold text is used to highlight the best result achieved among the methods for each dataset and metric. LBP-TOP [19] introduce a novel approach for micro-expression recognition by utilizing local binary patterns with six intersection points (LBP-SIP) to extract facial features. Bi-WOOF [6] extracts essential facial features from the apex image, they introduce the Bi-Weighted Optical Flow method. This technique effectively captures the relevant facial movements and highlights the important information needed for micro-expression recognition. OFF-ApexNet [3] suggest utilizing optical flow fields from the onset and apex frames. In the optical flow fields, horizontal and vertical features are obtained and fed into CNN-based network for further feature enhancement. After that, the extracted features from OFF-ApexNet will be used for micro-expression classification. STSTNet [28] proposes employing a three shallow CNN-based model to obtain high-level discriminative representations for classifying micro-expression emotions. To address the challenges of cross-database micro-expression recognition, Dual-Inception [40] propose a novel approach using two inception networks to extract horizontal and vertical features from the optical flow maps. By feeding the horizontal component of the optical flow into one inception network and the vertical component into another inception network, they can independently capture relevant patterns and information from both directions. Given the finite number of training samples in facial emotion datasets, EMR [5] incorporates two domain adaptation strategies. The first strategy involves adversarial training methods. The second strategy is the expression magnification method. RCN [29] proposes approach that suggests using a lower-size image as input and adopting a smaller-architecture model, which has shown to be beneficial for improving the model's performance on composite dataset tasks. FeatRef. [13] include two stages, in the first stage, horizontal inception network and vertical inception network will extract horizontal and vertical muscle motion features. After that, the horizontal and vertical features will be merged and fed into three attention based network for classifying these extracted features into different micro-expression categories. Finally, the classification branch is used to fuse salient and discriminative features obtained from the inception module for inferring micro-expression. SLSTT-LSTM [31] create a new kind of system to recognize quick, small facial expressions. It is the first one ever made that only uses transformers, which are a type of model that does not need the usual image processing networks. Their system has three parts: one that learns about space and shapes, another that looks at changes over time, and a final part that decides what the facial expression is.

Table 2

The Unweighted F1-score ($UF1$) and Unweighted Average Recall (UAR) performance of handcraft methods, deep learning methods and our HTNet method under LOSO protocol on the composite (Full), SMIC, CASME II and SAMM. Bold text indicates the best result.

Approaches	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [19]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Bi-WOOF [6]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
AlexNet [12]	0.6933	0.7154	0.6201	0.6373	0.7994	0.8312	0.6104	0.6642
GoogLeNet [11]	0.5573	0.6049	0.5123	0.5511	0.5989	0.6414	0.5124	0.5992
VGG16 [10]	0.6425	0.6516	0.5800	0.5964	0.8166	0.8202	0.4870	0.4793
OFF-ApexNet [3]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
STSTNet [28]	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
CapsuleNet [41]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
Dual-Inception [40]	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
EMR [5]	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
RCN [29]	0.7432	0.7190	0.6326	0.6441	0.8512	0.8123	0.7601	0.6715
FeatRef [13]	0.7838	0.7832	0.7011	0.7083	0.8915	0.8873	0.7372	0.7155
SLSTT-LSTM [31]	0.816	0.790	0.740	0.720	0.901	0.885	0.715	0.643
HTNet(Ours)	**0.8603**	**0.8475**	**0.8049**	**0.7905**	**0.9532**	**0.9516**	**0.8131**	**0.8124**

4.3.1. Compared to handcrafted methods

Table 2 presents a comparative analysis of different methods, including LBP-TOP and Bi-WOOF, which extract facial features using appearance-based and geometric-based techniques, respectively. Both methods employ SVM as the classifier. In contrast, our proposed method, HTNet, demonstrates substantial improvements in UF1 and UAR for composite datasets. Specifically, the UF1 increases from 0.6296 to 0.8603, and the UAR improves from 0.6227 to 0.8475, showcasing an improvement of more than 20%. Furthermore, HTNet consistently outperforms handcrafted methods (LBP-TOP and Bi-WOOF) on CASME II, SMIC, and SAMM datasets. These results underscore the efficacy of HTNet in addressing domain shift and achieving superior performance in micro-expression recognition tasks compared to both handcrafted and deep learning approaches.

4.3.2. Compared to deep learning methods

In Tables 2 and 3, our HTNet outperforms most deep learning methods by a considerable margin. As shown in Table 2, HTNet achieved an UF1 and UAR of 0.8603 and 0.8475, respectively, on the full composite dataset, which represents an improvement of approximately 5% compared to previous state-of-the-art methods. Analyzing Table 2, we observe that ALexNet, GoogLeNet, and VGG16 achieved UF1 and UAR of 0.6933 and 0.7154, 0.5573 and 0.6049, and 0.6425 and 0.6516, respectively, on the full composite dataset. These deeper learning methods yielded inferior results compared to other shallower methods (STSTNet, Dual-Inception, RCN) in the full composite datasets. The reason for this outcome lies in the fact that these three methods use deeper networks, which may lead to learning noise information due to the limited number of training samples. Although shallower networks have achieved superiority performance compared with VGG16, one critical research issue is how to extract salient and discriminative features from micro-expression. FeatRef can extract salient and discriminative features from micro-expression, so their method achieved UF1 and UAR of 0.7838 and 0.7832. However, FeatRef ingores the local and global spatio-temporal patterns for micro-expression recognition. One of the existing methods which remains competitive against ours is SLSTT-LSTM [31]. They use transformer to extract image features in different time sequence and use LSTM to merge this features along different time. Although, SLSTT-LSTM is sensitive to the changes between different images in a processed sequence, it may ignore the spatial information in single optical flow images and the relationship between different areas in optical flow images is missing. In contrast, HTNet focuses on four facial areas instead of the entire facial region. With local self-attention in each face region, the transformer layer can concentrate on localizing subtle muscle contractions. Furthermore, the aggregation layer facilitates learning interactions between different resolutions of optical flow

Table 3

The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance of deep learning methods and our HTNet method under LOSO protocol on CAS(ME)^3 Part A. Bold text indicates the best result.

Approaches	CASME^3 Part A	
	UF1	UAR
AlexNet [12]	0.257	0.2634
STSTNet [28]	0.3795	0.3792
RCN [29]	0.3928	0.3893
FeatRef [13]	0.3493	0.3413
HTNet(Ours)	**0.5767**	**0.5415**

feature maps. As a result, HTNet achieves better performance compared to previous methods on full composite datasets.

In Table 3, we conducted experiments on CAS(ME)^3 Part A and reported the Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) of deep learning methods, including our HTNet, under the leave-one-subject-out (LOSO) protocol. HTNet outperforms previous methods by a significant margin, demonstrating its effectiveness in micro-expression recognition.

4.4. Ablation study

This section presents an in-depth analysis of the impact of various parameters in our HTNet model. We investigate the influence of block size, hidden dimension, number of heads in the transformer, and the number of transformer layers at each hierarchy level. The initial experimental settings include a block size of 7×7 at the bottom-level, a hidden dimension of 256, three heads in the transformer, and $(2, 2, 8)$ transformer layers in each hierarchy level. For each ablation experiment, we modify a specific parameter while keeping other settings constant. The evaluation of each approach is conducted using the UAR metric and UF1 metric.

4.4.1. Impacts of block size

We conducted experiments to study the impact of different block sizes on the overall accuracy in composite micro-expression datasets, including SMIC, SAMM, and CASME II. The block size refers to the size of the facial regions considered in the HTNet model. We varied the bottom-level block size from 5 to 10, with the middle-level block size being twice the bottom-level size, and the top-level block size being four times the bottom-level size. The results, reported in Table 4, indicate that the choice of block size significantly affects the model's performance. Smaller block sizes may lead to subpar performance as

Table 4
Study the impacts of different block sizes on composite datasets-SMIC, SAMM and CASME II. The composite datasets Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance are reported.

| #Block size (Bottom level)		5 × 5	6 × 6	7 × 7	8 × 8	9 × 9	10 × 10
#Block size (Middle level)		10 × 10	12 × 12	14 × 14	16 × 16	18 × 18	20 × 20
#Block size (Top level)		20 × 20	24 × 24	28 × 28	32 × 32	36 × 36	40 × 40
Full	UF1	0.837	0.8271	**0.8603**	0.85	0.8511	0.8546
Full	UAR	0.8213	0.8037	0.8475	**0.846**	0.8445	0.8383
SMIC	UF1	0.7556	0.7394	**0.8049**	0.7833	0.7753	0.8008
SMIC	UAR	0.7469	0.7214	**0.7905**	0.7792	0.7708	0.7892
SAMM	UF1	**0.8237**	0.7903	0.8131	0.7909	0.8168	0.8068
SAMM	UAR	0.8033	0.775	**0.8124**	0.79	0.8088	0.7812
CASME II	UF1	0.9422	0.9482	0.9532	**0.964**	0.9722	0.957
CASME II	UAR	0.9308	0.9317	0.9516	0.962	**0.9658**	0.945

(a)

(b)
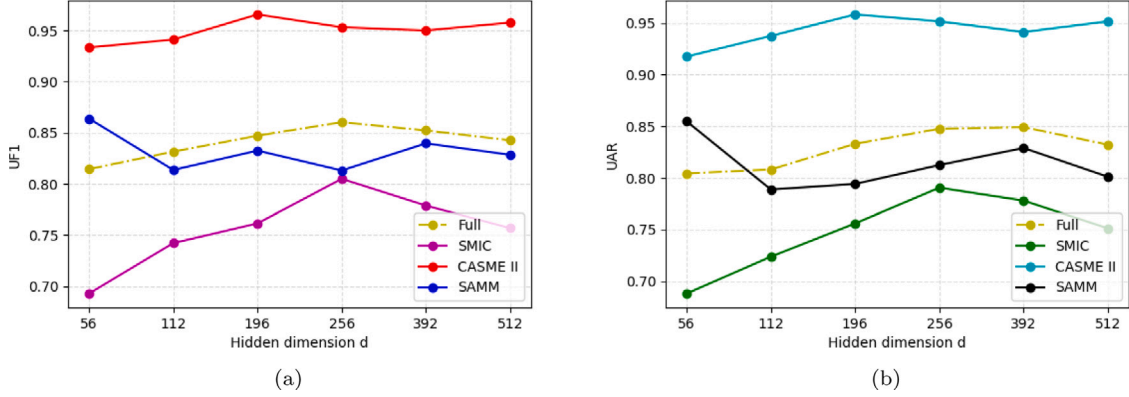
Fig. 5. We research how the number of dimensions affects the accuracy of composite datasets-SMIC, SAMM and CASME II. The composite datasets' Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance are reported.

(a)

(b)

Fig. 6. We investigate the effects of the transformer layer's head count on accuracy in composite datasets-SMIC, SAMM and CASME II. The composite datasets' Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance are reported.
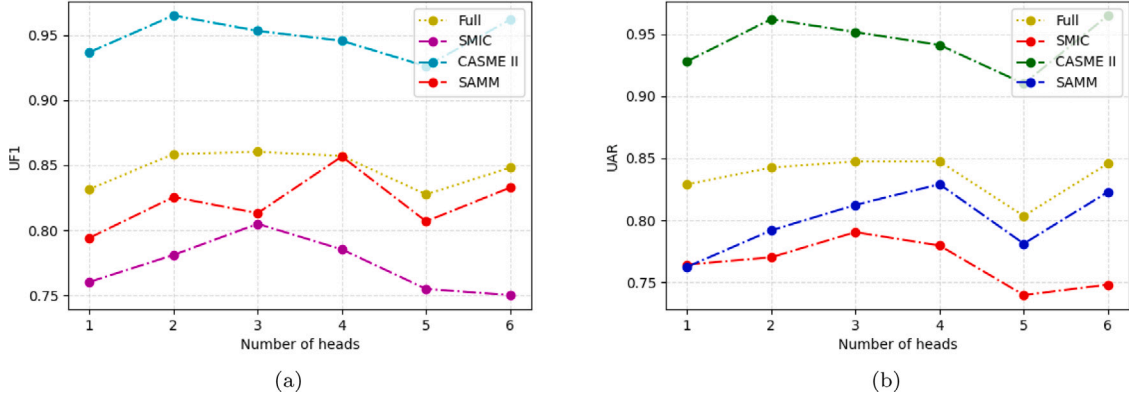
they might miss some crucial facial parts, while larger block sizes generally perform better. However, if the block size becomes too large, some of the facial areas may overlap, such as the left-eye area overlapping with the right-eye area, which can negatively impact the model's performance. Therefore, finding an optimal balance between block sizes is crucial to ensure the best performance in micro-expression recognition tasks.

4.4.2. Impacts of dimensions

We research how the number of dimensions affects the accuracy of composite datasets -SMIC, SAMM and CASME II. The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance of composite datasets are reported in Fig. 5. The smaller hidden dimension has worse performance because the small hidden dimension is hard to encode the optical flow feature map. However, using a larger hidden

dimension will lead to overfitting problems. The full UF1 score for hidden dimension 256 has the best performance, around 0.86.

4.4.3. Impact of number of heads

We investigate the effects of the transformer layer's head count on accuracy in composite datasets- SMIC, SAMM and CASME II. The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance of composite datasets are reported in Fig. 6. In Fig. 6(a), it demonstrates that three heads in transformer layers will have the best performance.

4.4.4. Impact of different number of transformer blocks

Table 5 presents an investigation into the impact of the number of transformer layers on the accuracy of our HTNet model, as evaluated using several datasets. To explore the effect of the number of

Table 5
Study the impacts of the number of transformer layers on composite datasets. The composite datasets' Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) performance are reported.

#Number of transformer layers		(2,2,2)	(2, 2, 4)	(2, 2, 6)	(2, 2, 8)	(2, 2, 10)	(2, 2, 12)
Full	UF1	0.8105	0.8297	0.8316	**0.8603**	0.8464	0.8499
Full	UAR	0.7848	0.8029	0.8183	**0.8475**	0.8376	0.8207
SMIC	UF1	0.748	0.7247	0.744	**0.8049**	0.76	0.7561
SMIC	UAR	0.7363	0.7065	0.734	**0.7905**	0.756	0.7345
SAMM	UF1	0.7984	0.8137	0.7732	0.8131	0.8571	**0.8682**
SAMM	UAR	0.75	0.7849	0.7742	0.8124	**0.8453**	0.8419
CASME II	UF1	0.8845	0.9647	**0.9722**	0.9532	0.9492	0.95
CASME II	UAR	0.8588	0.9554	**0.9658**	0.9516	0.9346	0.9412
#Params(millions)		149.57	245.88	342.20	438.51	534.82	631.13
#Training time (Seconds)		3767	4836	5906	6942	8085	9147

Table 6
The Unweighted F1-score ($UF1$) and Unweighted Average Recall (UAR) are reported for analyze number of hierarchies of HTNet.

#Number of hierarchies	Full		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
1	0.7089	0.6747	0.6327	0.6220	0.8567	0.8398	0.5948	0.5490
2	0.8588	0.8375	0.7923	0.7755	0.9607	0.9620	0.8267	0.7938
3	0.8500	0.8510	0.7820	0.7821	0.9526	0.9430	0.8137	0.8251

transformer layers in the top-level, we vary the count from 2 to 12. As observed, an increase in the number of transformer layers leads to a corresponding growth in model parameters and training time. A smaller number of transformer layers may have limited capacity to effectively encode optical flow features. Conversely, a larger number of layers can introduce overfitting issues, impacting the model's performance. Our experiments indicate that employing two transformer layers in the bottom-level and middle-level, along with a moderate number of layers in the top-level, results in the best performance for our HTNet model.

4.4.5. Impact of different number of hierarchies

The ablation study, as indicated by the provided Table 6, will specifically address the impact of varying hierarchical levels on the performance of HTNet. The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) metrics on composite datasets are reported to demonstrate the effectiveness of the hierarchical approach within our model. This provide clear insights into impacts of different number of hierarchies on the accuracy of micro-expression recognition. Without hierachical structure in our HTnet, the model can only obtain UF1 score of 0.7089 and UAR score of 0.6747 on the composite dataset, UF1 score of 0.6327 and UAR score of 0.6222 on the SMIC dataset, UF1 score of 0.8567 and UAR score of 0.8398 on the CASME II dataset and UF1 score of 0.5948 and UAR score of 0.5490 on the SAMM dataset. But after adding one more hierachical structure, the model can improve the UF1 score and UAR score of composite datasets from 0.7089 to 0.8588 and 0.6747 to 0.8375, the UF1 score and UAR score of SMIC datasets from 0.6327 to 0.7923 and 0.6220 to 0.7755, the UF1 score and UAR score of CASME II datasets from 0.8567 to 0.9607 and 0.8398 to 0.9620 and the UF1 score and UAR score of SAMM datasets from 0.5948 to 0.8267 and 0.5490 to 0.7938. This highlights that hierarchical structures will enhance ability to recognize micro-expressions with greater accuracy.

5. Qualitative and quantitative analysis of our proposed method

Fig. 7(a) displays the confusion matrices of HTNet, presenting the accuracy achieved for each emotion category in the full composite datasets. Notably, HTNet attains accuracies of 0.94, 0.81, and 0.80 for the negative, positive, and surprise categories, respectively, in these datasets (SMIC, SAMM, and CASME II). The high accuracy for the negative category can be attributed to the larger number of training samples available for this class in the three datasets. However, challenges arise

in distinguishing between the surprise and negative emotion categories in the SMIC and SAMM datasets. The limited number of training samples in these datasets may result in occasional misclassifications between these two categories. Consequently, the negative, positive, and surprise categories in the CASME II dataset achieve relatively high accuracies, each exceeding 90%. Misclassifications are minimal, further validating the efficacy of HTNet on this dataset. Additionally, the SMIC dataset employs lower frame rates for capturing images, which introduces background noise such as flickering lights, shadows, and variations in illumination. These factors can impact the accuracy of classifying positive and surprise emotions in the SMIC dataset, resulting in accuracies of 76% and 72%, respectively.

Fig. 7(e) presents the confusion matrix obtained for CAS(ME)[3]. From the confusion matrix, we observe that the negative emotion category achieves the highest accuracy among the three emotion categories. This result is consistent with the performance observed in the composite datasets, where the negative category also exhibited the highest accuracy. On the other hand, the positive emotion category obtains the lowest accuracy, approximately 0.15. Most of the positive samples are misclassified as negative. In contrast, in the composite datasets, only a small portion of samples are misclassified as negative.

6. Conclusion

In this paper, we propose a hierarchical transformer architecture to extract important four facial areas features for micro expression recognition. Compared with previous deep learning methods, one of the key strengths is that the hierarchical architecture of the model allows for the extraction of features at multiple scales, which can improve recognition accuracy in the expression. In addition, HTNet model is based on the Transformer architecture, which has demonstrated state-of-the-art performance in natural language processing tasks. This architecture is well-suited for micro-expression recognition because it allows for the modeling of complex dependencies between different parts of the facial expression, while also being computationally efficient.

Despite these strengths, HTNet model requires a large amount of training data, which may be difficult to obtain in some scenarios. The model may also be sensitive to variations in lighting, pose, and other environmental factors that can affect the appearance of micro-expressions.
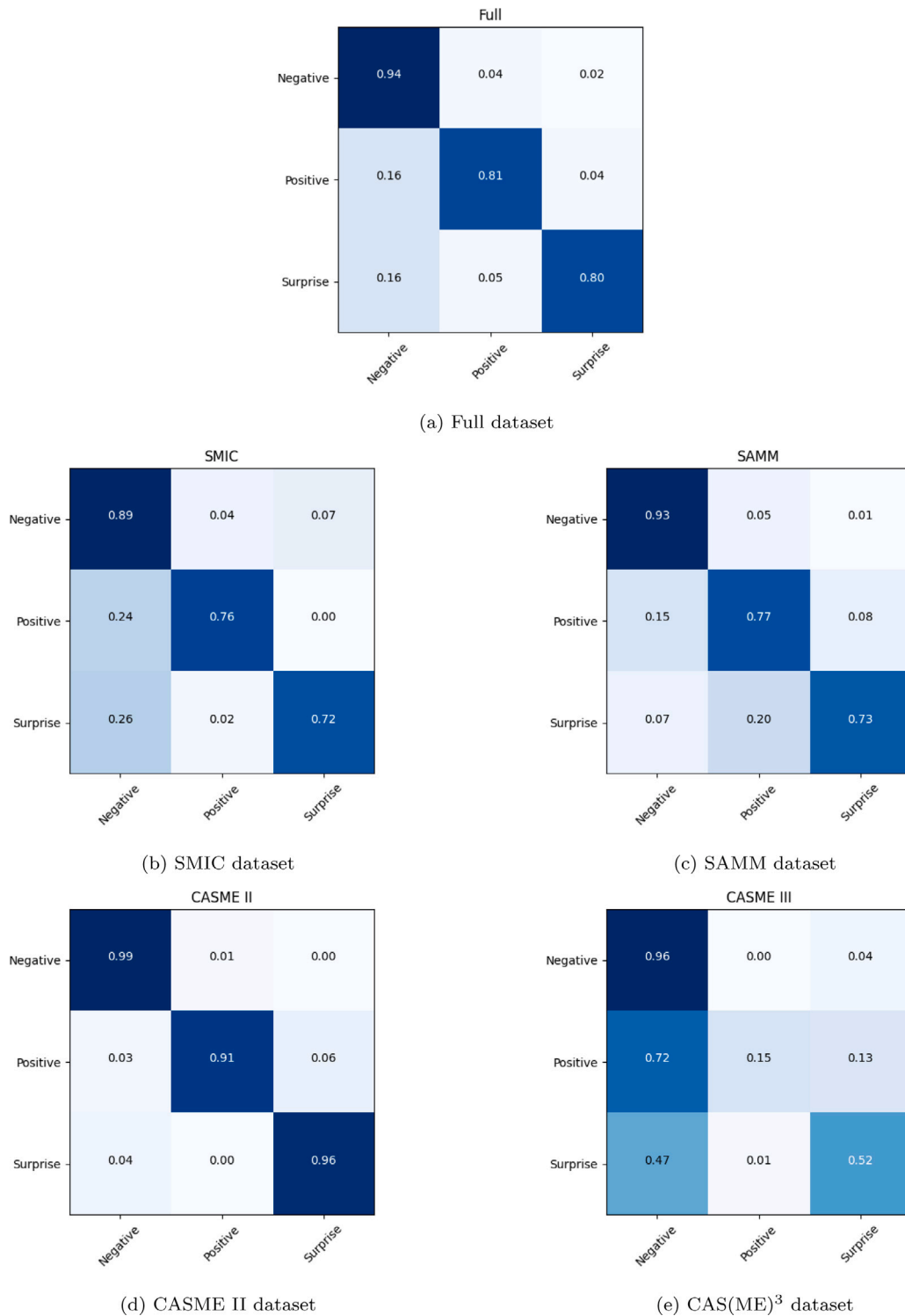
(a) Full dataset

(b) SMIC dataset

(c) SAMM dataset

(d) CASME II dataset

(e) CAS(ME)^3 dataset

Fig. 7. Confusion matrices for the proposed HTNet on the composite database (SMIC, SAMM, CASME II) and CAS(ME)^3 dataset using 3 classes.

To address these limitations, future work can explore ways to improve the efficiency and robustness of the HTNet model. Data augmentation techniques can be used to increase the amount of training data and improve the model's ability to generalize to different environmental conditions. Transfer learning can also be used to leverage pre-trained models on related tasks and improve the efficiency of the HTNet model. With further development and optimization, the HTNet

model may prove to be a valuable tool for accurately recognizing micro-expressions in various applications, including lie detection, emotion recognition and mental health assessment.

CRediT authorship contribution statement

Zhifeng Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology,

Investigation, Formal analysis, Conceptualization. **Kaihao Zhang:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Wenhan Luo:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Ramesh Sankaranarayana:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

Data availability

Data will be made available on request.

Acknowledgments

References

[1] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, W. Zhang, FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20922–20931.

[2] S. Thuseethan, S. Rajasegarar, J. Yearwood, EmoSeC: Emotion recognition from scene context, Neurocomputing 492 (2022) 174–187.

[3] Y.S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, OFF-ApexNet on micro-expression recognition system, Signal Process., Image Commun. 74 (2019) 129–139.

[4] Y. Gan, J. See, H.-Q. Khor, K.-H. Liu, S.-T. Liong, Needle in a haystack: Spotting and recognising micro-expressions in the wild, Neurocomputing 503 (2022) 283–298.

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

[6] S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Process., Image Commun. 62 (2018) 82–92.

[7] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Trans. Affect. Comput. 7 (4) (2015) 299–310.

[8] S. Happy, A. Routray, Fuzzy histogram of optical flow orientations for micro-expression recognition, IEEE Trans. Affect. Comput. 10 (3) (2017) 394–406.

[9] S.-T. Liong, R.C.-W. Phan, J. See, Y.-H. Oh, K. Wong, Optical strain based recognition of subtle emotions, in: 2014 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS, IEEE, 2014, pp. 180–184.

[10] A. Sengupta, Y. Ye, R. Wang, C. Liu, K. Roy, Going deeper in spiking neural networks: VGG and residual architectures, Front. Neurosci. 13 (2019) 95.

[11] P. Ballester, R.M. Araujo, On the performance of GoogLeNet and AlexNet applied to sketches, in: Thirtieth AAAI conference on artificial intelligence, 2016.

[12] H. Zhang, H. Zhang, A review of micro-expression recognition based on deep learning, in: 2022 International Joint Conference on Neural Networks, IJCNN, IEEE, 2022, pp. 01–08.

[13] L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition, Pattern Recognit. 122 (2022) 108275.

[14] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing (2023) 126866.

[15] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, Neurocomputing 391 (2020) 42–51.

[16] A. Bhandari, N.R. Pal, Can edges help convolution neural networks in emotion recognition? Neurocomputing 433 (2021) 162–168.

[17] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[18] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, X. Fu, Micro-expression recognition using dynamic textures on tensor independent color space, in: 2014 22nd international conference on pattern recognition, IEEE, 2014, pp. 4678–4683.

[19] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, LBP with six intersection points: Reducing redundant information in LBP-top for micro-expression recognition, in: Asian conference on computer vision, Springer, 2014, pp. 525–537.

[20] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, X. Fu, Micro-expression recognition using dynamic textures on tensor independent color space, in: 2014 22nd international conference on pattern recognition, IEEE, 2014, pp. 4678–4683.

[21] X. Li, J. Yu, S. Zhan, Spontaneous facial micro-expression detection based on deep learning, in: 2016 IEEE 13th International Conference on Signal Processing, ICSP, IEEE, 2016, pp. 1130–1134.

[22] Z. Lu, Z. Luo, H. Zheng, J. Chen, W. Li, A delaunay-based temporal coding model for micro-expression recognition, in: Computer Vision-ACCV 2014 Workshops: Singapore, Singapore, November (2014) 1-2, Revised Selected Papers, Part II 12, Springer, 2015, pp. 698–711.

[23] L. Lei, T. Chen, S. Li, 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.

[24] K. Zhang, Y. Huang, H. Wu, L. Wang, Facial smile detection based on deep learning features, in: 2015 3rd IAPR Asian Conference on Pattern Recognition, ACPR, IEEE, 2015, pp. 534–538.

[25] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial–temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.

[26] W. Niu, K. Zhang, D. Li, W. Luo, Four-player groupgan for weak expression recognition via latent expression magnification, Knowl.-Based Syst. 251 (2022) 109304.

[27] Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions, IEEE Trans. Multimed. 22 (3) (2019) 626–640.

[28] S.-T. Liong, Y.S. Gan, J. See, H.-Q. Khor, 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, IEEE, 2019, pp. 1–5.

[29] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Trans. Image Process. 29 (2020) 8590–8605.

[30] A.J.R. Kumar, 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.

[31] L. Zhang, X. Hong, O. Arandjelović, G. Zhao, Short and long range relation based spatio-temporal transformer for micro-expression recognition, IEEE Trans. Affect. Comput. 13 (4) (2022) 1973–1985.

[32] S.-T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.-H. Oh, R. Phan, Automatic apex frame spotting in micro-expression database, in: 2015 3rd IAPR Asian conference on pattern recognition, ACPR, IEEE, 2015, pp. 665–669.

[33] E. Jose, M. Greeshma, M.T. Haridas, M. Supriya, Face recognition based surveillance system using facenet and mtcnn on jetson tx2, in: 2019 5th International Conference on Advanced Computing & Communication Systems, ICACCS, IEEE, 2019, pp. 608–613.

[34] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Trans. Affect. Comput. 9 (1) (2016) 116–129.

[35] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition, FG, IEEE, 2013, pp. 1–6.

[36] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PLoS One 9 (1) (2014) 86041.

[37] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Trans. Pattern Anal. Mach. Intell. (01) (2022) 1.

[38] S.-T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.-H. Oh, R. Phan, Automatic apex frame spotting in micro-expression database, in: 2015 3rd IAPR Asian conference on pattern recognition, ACPR, IEEE, 2015, pp. 665–669.

[39] C.-W. Chang, Z.-Q. Zhong, J.-J. Liou, A FPGA Implementation of Farneback Optical Flow By High-LEvel Synthesis, FPGA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 309.

[40] L. Zhou, Q. Mao, L. Xue, Dual-inception network for cross-database micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, IEEE, 2019, pp. 1–5.

[41] N. Van Quang, J. Chun, T. Tokuyama, CapsuleNet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, IEEE, 2019, pp. 1–7.

Zhifeng Wang graduated with an M.S. degree in computer vision and machine learning from the Australian National University in 2021. He is presently pursuing a Ph.D. at the Australian National University's College of Engineering and Computer Science in Canberra, ACT, Australia. His study focuses on computer vision and machine learning, particularly deep learning for facial and emotion identification.

Kaihao Zhang (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia. He has more than 20 referred publications in international conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, AAAI, ACMMM, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), and IEEE TRANSACTIONS ON MULTIMEDIA (TMM). His research interests focuses on computer vision and deep learning.

Wenhan Luo is an associate professor and Ph.D. advisor in Sun Yat-sen University, conducting research on trustworthy AI and creative AI. Before being a faculty member in university, he worked as an applied research scientist at Tencent, solving real-world problems using computer vision and machine learning techniques. Prior to Tencent, he worked for Amazon (A9) in Palo Alto, California, where he developed deep models for better visual search experience. Before that, he worked as a research scientist in Tencent AI Lab. He received the Ph.D. degree from Imperial College London, UK, 2016, M.E. degree from Institute of Automation, Chinese Academy of Sciences, China, 2012 and B.E. degree from Huazhong University of Science and Technology, China, 2009. He have published over 60 peer-reviewed papers, including over 40 of them published in top-tier conferences and journals, like ICML, CVPR, ICCV, ECCV, AAAI, ACL, ACMMM, ICLR, TPAMI, AI, IJCV, TIP, and 2 of them are ESI highly cited papers. He received the CVPR 2019 Best Paper Nominee and was awarded the 2022 ACM China Rising Star Award (Guangzhou Chapter).

Ramesh Sankaranarayana received the M.E. degree from the Indian Institute of Science, India, and the Ph.D. degree from the University of Alberta, Canada. He is currently the associate director (Educational partnerships), Research School of Computer Science, Australian National University, Canberra, Australia. He research interests include information retrieval, human-centered computing and software engineering. He is a member of the ACM.