Expression Snippet Transformer for Robust Video-based Facial Expression Recognition

Yuanyuan Liu[1], Wenbin Wang[1], Chuanxu Feng[1], Haoyu Zhang[1], Zhe Chen[*2], and Yibing Zhan[3]

[1]China University of Geosciences (Wuhan)
[1]{*liuyy, wangwenbin, fcxfcx, zhanghaoyu*}*@cug.edu.cn*
[2]The University of Sydney
[2]*zhe.chen1@sydney.edu.au*
[3]Jingdong
[3]*zhanyibing@jd.com*

February 10, 2023

Abstract

Although Transformer can be powerful for modeling visual relations and describing complicated patterns, it could still perform unsatisfactorily for video-based facial expression recognition, since the expression movements in a video can be too small to reflect meaningful spatial-temporal relations. To this end, we propose to decompose the modeling of expression movements of a video into the modeling of a series of expression snippets, each of which contains a few frames, and then boost the Transformer's ability for intra-snippet and inter-snippet visual modeling, respectively, obtaining the Expression snippet Transformer (EST). For intra-snippet modeling, we devise an attention-augmented snippet feature extractor to enhance the encoding of subtle facial movements of each snippet. For inter-snippet modeling, we introduce a shuffled snippet order prediction head and a corresponding loss to improve the modeling of subtle motion changes across subsequent snippets. The EST obtains state-of-the-art performance, demonstrating its superiority to other CNN-based methods. Our code and the trained model are available at https://github.com/DreamMr/EST

[*]Corresponding author

1 Introduction

Video-based Facial Expression Recognition (FER) is important for understanding human emotions and behaviors. Therefore, FER has a wide range of applications in social life, such as multimedia information processing, driver monitoring, lie detection, etc. [3]. FER aims to classify a video into one of several basic emotions, including happiness, anger, disgust, fear, sadness, neutral, and surprise. The task of FER is difficult due to several challenges, namely, long-range spatial-temporal representation, excessive noises from irrelevant frames, and especially, inherently *small* and *subtle* facial movements in FER videos.

To tackle the issues of FER, existing methods commonly apply convolutional neural networks (CNNs) [21] or long-short term memory (LSTM) [15]. However, most of the existing FER methods usually model spatial-temporal visual information without involving effective visual relation reasoning mechanisms [21]. For example, many methods [17, 33, 36] only use static frames selected from the manually defined peak (apex) frames, neglecting the intrinsic relationships between visual cues of adjacent frames. Sequence-based methods [30, 7, 15] attempt to capture motion cues by encoding spatial-temporal in-

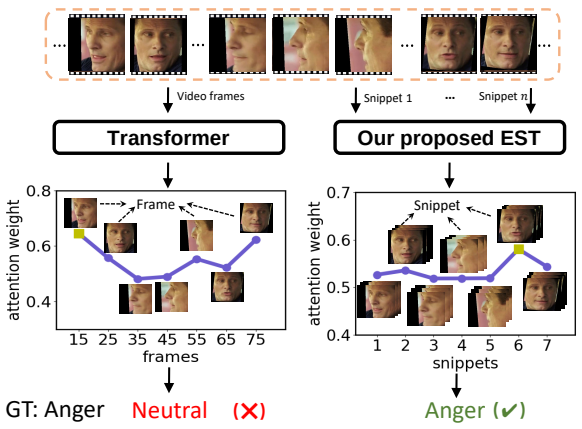

Figure 1: Comparison between a vanilla Transformer method and the proposed expression snippet Transformer (EST) for modeling subtle facial expression movements in facial expression recognition (FER). The vanilla Transformer (left) tends to focus only on the frame with peak expression patterns and can be easily affected by noises such as other non-expression changes, thus obtaining suboptimal results. By decomposing videos into snippets, the EST (right) improves the modeling of intra-snippet and inter-snippet subtle facial changes, respectively, and can achieve more robust FER. The yellow square marker represents the highest attention in the video.

formation within their models, while they still perform weakly in describing subtle expression movements in FER videos if not using overwhelmingly large model capacities [25].

The recent successful Transformer approaches [6, 2, 39] in computer vision has allowed us to take advantage of its powerful relation reasoning ability for understanding FER videos. In general, the Transformer [32] has shown to be particularly effective for translating an input sequence to a target sequence by modeling the relations between features. Accordingly, for video-based FER, we believe that the Transformer has a great potential of describing subtle expression movements more robustly. Despite the potential advantages, it is non-trivial to directly apply a vanilla Transformer on the FER video frames considering that the subtle facial expression movements within videos can be too difficult to be modeled properly by the vanilla Transformer. For example, as shown in Fig. 1, the per-frame visual information (*i.e.* raw pixels on each frame) may contain noises such as non-expression changes (head poses, speaking, and so on) that can easily affect the recognition performance of the Transformer. Furthermore, the subtle expression movements would make the Transformer only focus on the visual cues from frames with peak expression changes and neglect plenty of beneficial spatial-temporal information from other periods of videos. This limits the potential of Transformer to encode the motion information of the entire video comprehensively and achieve more robust expression recognition.

To tackle the above problems for applying Transformer on FER videos, we first propose to decompose the modeling of facial movements of the entire video into the modeling of a series of small expression snippets. Each expression snippet is a video clip with a few adjacent frames of the input video covering a limited amount of expression changes. Then, by employing the Transformer over the snippets, we can augment the modeling of intra-snippet and inter-snippet expression movements, respectively. In particular, we introduce a novel attention-augmented snippet feature extractor (AA-SFE) to improve the modeling of intra-snippet visual changes for the Transformer. In the AA-SFE, we apply a deep convolutional neural network (DCNN) to extract per-frame visual features and develop a novel hierarchical attention-augmentation architecture to obtain the representation of facial movements within each snippet. The snippet representations generated with the AA-SFE are subsequently fed into the encoder-decoder structure of a Transformer to perform recognition based on snippet-level relations. Meanwhile, we devise a shuffled snippet order prediction (SSOP) head with a corresponding loss for the Transformer to improve the modeling of inter-snippet visual changes. By using SSOP, the Transformer can encode the information from all snippets more comprehensively, thereby delivering a more robust expression movement representation of the entire video. Overall, we briefly name our proposed method as expression snippet Transformer (EST).

To sum up, the major contributions of this paper are summarized as follows:

- We propose the expression snippet Transformer (EST) to achieve accurate video-based facial expres-

sion recognition (FER). To the best of our knowledge, our approach is the first effective snippet-based Transformer method for video-based FER.

- To enhance the Transformer's ability to model intra-snippet and inter-snippet expression movements, we propose the attention-augmented snippet feature extractor (AA-SFE) and the shuffled snippet order prediction (SSOP), respectively. Both techniques effectively tackle the problems of Transformer-based FER and substantially improves the recognition performance.

- Evaluations on four challenging video facial expression datasets, *i.e.*, BU-3DFE, MMI, AFEW, and DFEW, demonstrate the superiority of our proposed EST over existing popular methods. State-of-the-art performance can be achieved with EST on the evaluated datasets. We will release our source code upon acceptance.

2 Related Work

Frame-based methods for video-based FER. The frame-based methods can be divided into two groups: frame aggregation methods that strategically fuse deep features learned from static-based FER networks [25] and peak frame extraction methods that focus on recognizing the peak high-intensity expression frame [37]. Meng *et al.* [25] proposed frame attention networks to adaptively aggregate frame features in an end-to-end framework and achieved an accuracy of 51.18% on the AFEW 8.0 dataset. Moreover, Yu *et al.* [37] proposed a deeper cascaded peak-piloted network (DCPN) that enhances the discriminative ability of features in a cascade fine-tuning manner. The DCPN achieved the best accuracies of 99.9% on the CK+ dataset [22]. However, these methods depend only on static frames and lack powerful modeling of the spatial-temporal relationships of expressions in the video.

Dynamic sequence-based methods for video-based FER. In order to explore the spatial-temporal representation of expressions, dynamic sequence-based methods take a video sequence as a single input and utilize both textural information and temporal dependencies in the sequence for more robust expression recognition [15, 25]. Recently, the Long Short-Term Memory (LSTM) and C3D are two widely-used spatial-temporal methods. Kim *et al.* [15] proposed a new spatio-temporal feature representation learning for FER by integrating C3D and LSTM networks, which is robust to expression intensity variations. Although the C3D networks can capture the spatial-temporal change of an expression, the C3D networks introduce expensive space- and computational complexity to learn subtle expression movements more effectively.

A more related study of FAN [25] introduces an attention module to refine the visual features for FER. It first estimates per-frame attention to obtain refined features and then employs two fully-connected layers interpret the attentional features, which is less adaptive to diversified features and is also easier to be affected by trivial information. Different from the motivation of FAN, we propose AA-SFE mainly to explore the intra-snippet relations, so that more valuable information about emotion can be highlighted and trivial information could be depressed. This is beneficial for the later Transformer to better model subtle changes in a video. Meanwhile, the attention implementation strategies of AA-SFE and FAN are very different. In AA-SFE, we follow a self-attention structure for the first level and use cosine-similarity to implement the attention at the second level. The applied self-attention allows us to make it compatible with different popular Transformer networks. Meanwhile, the cosine-similarity used here follows the concept of routing mechanism between capsules [28] and can achieve more appropriate weighting for different snippet features.

Transformer in different tasks. Transformer was introduced by Vaswani *et al.* [32] as a new attention-based building block for machine translation. Transformer included self-attention layers to scan through each token in a sequence and learn the tokens' relationships by aggregating information from the whole sequence, replacing RNNs in many tasks, such as natural language processing (NLP), speech processing, and computer vision [8, 6, 2, 39]. Recently, Nicolas *et al.* expanded the basic Transformer architecture to the field of object detection and proposed the DETR algorithm [2]. Girdhar *et al.* proposed an action Transformer to aggregate features from the spatial-temporal contexts around persons for action recognition in a video [8]. However, applying the vanilla Transformer to capture subtle expression movements in an untrimmed video is still challenging due to

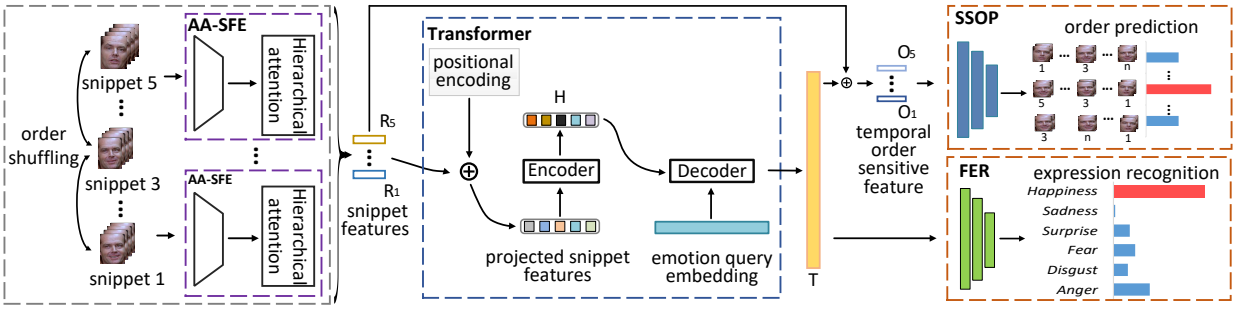

Figure 2: The training pipeline of the EST for video-based FER. Using expression snippets, we apply the AA-SFE and SSOP to improve Transformer's ability for respectively modeling intra-/inter-snippet expression movements and relations, thus achieving robust FER.

the noises and the limited motion variations within input frames. Although Former-DFER [39] has applied Transformers to model spatial and temporal information via frame relations for video-based FER, it is simply based on per-frame feature without explicit and effective mechanisms to tackle the problem of subtle facial changes.

3 Expression Snippet Transformer

3.1 EST Architecture

The overall EST architecture is illustrated in Fig. 2. Firstly, we collect expression snippets from the input video. For each snippet, we apply an attention-augmented snippet feature extractor (AA-SFE) to extract per-snippet features. Then, we employ a Transformer with a shuffled snippet order prediction (SSOP) head to help achieve more robust expression understanding. In the following sections, we will subsequently explain the Expression snippets, Transformer, AA-SFE, and SSOP.

Expression Snippets. We decompose the input video into a series of snippets to augment the Transformer's ability to model subtle visual changes within each snippet and across different snippets, respectively. Formally, given an input FER video $\mathcal{C}$, we decompose it into a series of smaller sub-videos: $\mathcal{C} = \{C_1, C_2, ...C_n\}$, where C_i represents the i-th sub-video and n is the total number of sub-videos. Each sub-video C_i refers to an expression snippet that contains several adjacent frames of the video. All the snippets have the same length, and they follow consecutive orders along time.

Transformer Architecture. We first extract snippet features with AA-SFE, which will be discussed later. With the snippet features, a Transformer is applied here to model the expression movements across snippets and discover a more robust emotion representation for FER. We follow the typical Transformer formulation and apply a multi-head attention-based encoder-decoder pipeline for the processing. In general, the multi-head attention estimates the correlation between a *query* tensor and a *key* tensor and then aggregates a *value* tensor according to correlation results to obtain an attended output. We would like to mention that our paper mainly follows the formulation of DETR[2] to define the Transformer. As a result, the employed Transformer uses several decoder layers with query embeddings to translate features. Rather than using the formulation of standard visual Transformer for image classification like ViT [6], we found that the DETR formulation that attempts to use object query embedding rather than class tokens could be easier to understand in expression recognition, thus we use its implementation strategy to define the Transformer. In fact, we analyze that the emotion query in the decoder may share a similar function with the class token used in ViT [6], both of which would tend to accumulate useful information through Transformer layers. For more details of the Transformer, please refer to [32].

In our approach, we employ the encoder to encode

snippet features and then use the decoder to translate the encoded features into a more robust expression representation. Let $R_i \in \mathbb{R}^d$ denote the extracted snippet feature of C_i with a size of d, and $\mathcal{R} = \{R_1, R_2, ...R_n\}$. We feed $\mathcal{R}$ to the encoder of the Transformer in EST. In the encoder, for each head of the multi-head attention, we perform linear projections on a snippet feature R_i to obtain the corresponding query vector q_i, key vector k_i, and value vector v_i, respectively. All the q_i, k_i, v_i are vectors of size d as well.

Then, we stack different snippets' query vectors, key vectors, and value vectors to obtain a query tensor Q, a key tensor K, and a value tensor V, respectively. $Q, K, V \in \mathbb{R}^{n \times d}$. Afterward, we perform self-attention across the snippets based on the obtained Q, K, and V. In addition, we apply a snippet positional encoding to describe the positions of snippets within a video, following the sine and cosine positional encoding [32]. The output of the encoder is the encoded snippet features $H \in \mathbb{R}^{n \times d}$:

$$H = A(Q, K, V) = softmax(\frac{Q\,K^T}{\sqrt{d}})V, \qquad (1)$$

where $A(\cdot)$ represents the self-attention. In this study, we employ 3 encoder layers, each with 4 attention heads. Our original intention is to reduce the complexity of the employed Transformer while still maintaining its power. Therefore, we experimentally halve the number of encoder layers and the number of attention heads used in the DETR method [2]. More experiments about selecting these meta-parameters are provided in section 4.4.4.

After encoding snippet features with self-attention, the decoder phase then applies cross-attention to decode the encoded features H into an emotion representation T and $T \in \mathbb{R}^d$. Our introduced emotion query embedding represents the query embedding in a Transformer network, which is similar to how the object query embedding is defined in DETR [2]. It is defined as a 512-dimentional weight vector in the Transformer. To obtain a proper emotion query embedding, we follow the training procedure of the DETR [2] and optimize its weight values together with the Transformer. During the training, we randomly initialize the emotion query embedding at first and then optimize its weights according to the final objective function. We use the encoded feature H to calculate both key and value tensors in the decoder. In practice, we stack 3 decoder layers, each with 4 attention heads, to progressively refine the decoding results.

After the encoder-decoder processing, we make the Transformer provide two outputs, forming two prediction heads. The first head, built upon a 3-layer perception network, is the expression recognition prediction, classifying the T into different expression types. The second head is the SSOP, which estimates the correct snippet order since snippets are shuffled. We will discuss the details of SSOP later.

3.2 Attention-augmented Snippet Feature Extraction

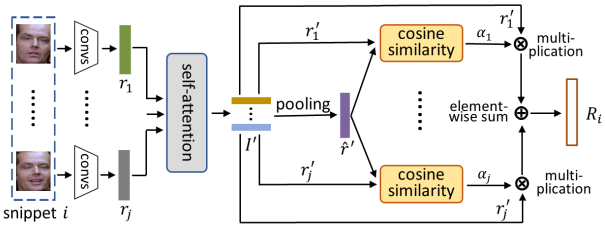

Figure 3: The detailed architecture of the AA-SFE.

Directly applying the Transformer on raw frames can be sub-optimal due to visual noises within pixels, making it difficult to obtain robust expression representation. Using snippets, we boost the Transformer to better model intra-snippet expression movements by introducing the AA-SFE which improves the encoding of spatial-temporal information across frames within a snippet.

Fig. 3 shows the structure of an AA-SFE. In particular, with the help of normal DCNNs, such as a pre-trained ResNet-18, the AA-SFE applies a *hierarchical attention augmentation* for modeling intra-snippet information. The hierarchical attention aims to gradually extract a more representative feature of a snippet, progressively filtering out less meaningful non-expression information to reduce the negative impacts of noises within per-frame features. We mainly apply the attention from *two-level* hierarchy to model subtle visual changes. The first level extracts frame-level attention, and the second focuses on extracting snippet-level global attention.

For the *first-level* hierarchy, we investigate frame-level relation to obtaining attention. Similar to the Transformer,

we apply self-attention here for relation modeling. Mathematically, we use $r_{j,i}$ to represent the feature vector of the j-th frame in the i-th snippet. We extract the global average pooling output of a DCNN as the per-frame feature: $r_{j,i} \in \mathbb{R}^d$. Suppose each snippet has J frames. By stacking all the features $r_{j,i}$ from the i-th snippet, we obtain the tensor $I_i \in \mathbb{R}^{J \times d}$. Since we only consider frames of a single snippet at this stage, we drop the symbol i here for simplicity, *i.e.*, $I = I_i$, $r_j = r_{j,i}$ in this part. Using linear projections, we transform I into three tensors: query tensor I_Q, key tensor I_K, and value tensor I_V. Then, we apply self-attention described in Eq.1 on I_Q, I_K, I_V to obtain the attended feature $I' \in \mathbb{R}^{J \times d}$.

In the *second-level* hierarchy, we introduce the snippet-level global information to further refine representations of snippets. Firstly, we summarize the I' into a unified general feature vector. Then, we estimate the relations between the general feature and per-frame features. The obtained relations are later used to re-weight per-frame features for refinement. Lastly, the refined features are reduced to a single representation to describe the whole snippet. More specifically, we denote the symbol $\hat{r}'$ as the general feature vector of I', with a size of d. It is obtained by performing max pooling on I' across frames. Then, we estimate the relation between $\hat{r}'$ and per-frame features using cosine similarity. We denote r'_j as the feature in I' correspond to the j-th frame. We compute cosine similarity α_j between $\hat{r}'$ and each r'_j:

$$\alpha_j = cos(r'_j, \hat{r}') = \frac{r'_j \cdot \hat{r}'}{\|r'_j\| \cdot \|\hat{r}'\|}, \tag{2}$$

where $\|\cdot\|$ means Euclidean norm. With the relation estimated by α_j, we can identify which frame contains more deviated information that could be more likely to contain noise with non-expression. Thus, we have the summarized snippet feature by re-weighting and aggregating per-frame features based on:

$$R_i = \frac{\sum_j \alpha_j \cdot r'_j}{\sum_j \alpha_j}. \tag{3}$$

To sum up, the self-attention of AA-SFE first provides powerful relation modeling to facilitate the encoding of frame-level spatial-temporal information. Then, we introduce the second hierarchy with cosine similarity-based attention modeling to consider the global motion information of a snippet to help further resist noises existing in per frame. According to Eq. $2 \sim 3$, the attention can identify the more useful intra-snippet visual change information and facilitate the computation of a more focused snippet feature $R_i \in \mathbb{R}^d$. We experimentally prove that the AA-SFE delivers better snippet features comparing to a normal self-attention-based Transformer.

3.3 Shuffled Snippet Order Prediction

With the snippet features $\mathcal{R}$ and the Transformer, we can estimate expressions of videos. However, we observe that the Transformer with AA-SFE could still focus only on the snippet with peak expression changes and neglects the rest of the parts of a video. This situation happens because the cross-attention modeling mechanism of Transformer probably easily overlooks the slight motion changes across subsequent snippets. In practice, the Transformer usually fails to deliver the comprehensive inter-snippet relation modeling for all the snippets and thus can be easily distracted by noisy information in the peak snippet. To make the Transformer model expression motions more comprehensively and avoid the negligence of subtle visual changes from off-peak snippets, we further introduce a shuffled snippet order prediction (SSOP) head with corresponding loss to assist training the EST. The algorithm of the SSOP is shown in Algorithm 1.

Algorithm 1 The Pseudo Code of the SSOP.

Input: FER video $\mathcal{C}$;Permuted order S
Output: Predicted permuted order probability $p(S|O)$

$\{C_1, ..., C_n\} \leftarrow \mathcal{C}$
Shuffle the snippet order according to the S
Collect featuers: $O_i = R_i + T$
Concat features: $O = Concat([O_1, ..., O_n])$
Predict the snippet order probability $p(S|O)$
Update SSOP head and emotion information T by Eq. 6

To train the Transformer with SSOP, we mainly shuffle the snippets randomly and make the Transformer predict this shuffled snippet order. For example, we have 7 snippets of a video and assign an index to each snippet,

i.e., $S = (1, 2, 3, 4, 5, 6, 7)$. We can randomly re-arrange the 7 snippets' indexes according to a specific permutation, such as $S1 = (2, 4, 5, 7, 1, 3, 6)$, which is 1 of the 7!=5,040 possible permutation orders. For simplicity, we randomly select 10 permutation orders for training the SSOP. Then, among all the generated orders, we sample one order and re-arrange the snippets accordingly. The snippets with a shuffled order are later sent to the EST.

It is worth mentioning that we fuse the T with the $\mathcal{R} \in \mathbb{R}^{7\times d}$ to obtain the features used for predicting shuffled orders, obtaining a temporal order sensitive feature $O \in \mathbb{R}^{7\times d}$ as described in Algorithm 1. This is because, after extracting the general emotion information T, we find that it is difficult for T to encode detailed order information via the Transformer. Meanwhile, R contains more information on the shuffled snippet order and can complement the T in the SSOP and facilitate the EST to learn better on the inter-snippet relations (achieving an relative increase of **6.8%** on AFEW in the experiments). We further apply three fully connected layers on the O to define the SSOP head which predicts the current permuted shuffling order. The prediction is obtained according to a classification output. Therefore, training the Transformer with the SSOP involves maximizing a posterior probability (MAP) estimate, where the related conditional probability density function is:

$$p(S|C_1, C_2, ..., C_n) = p(S|O_1, ..., O_n) \prod_{i=1}^{n} p(O_i|C_i), \tag{4}$$

where O_i is the feature vector in O for the i-th snippet. C_i represents i-th snippet.

In general, the SSOP mainly helps the Transformer learn to focus on snippets with more meaningful visual cues rather than only looking at the central snippet. Without SSOP, although we have positional encoding in the Transformer, the Transformer tends generate the highest attention weights only for the snippet covering the middle of a video and motion information from off-peak snippets is usually not well encoded due to very subtle facial changes. This can affect the recognition if the salient emotion cues occur in other parts of the input video. We analyze that the reason for this phenomenon is that most of the emotion videos are normalized so that the largest emotion changes occur during the middle. To avoid this, we introduce the SSOP aiming to break the dependence of emotion recognition to the middle of the input video by randomly shuffling the video content orders, so that salient emotion changes can occur at different periods of the video. Meanwhile, only shuffling snippet orders may also disturb the learning on inter-snippet relations, thus we only sample a few shuffling orders before training and make the network predict the shuffling order during training to learn the inter-snippet relations better. Besides, the SSOP also enriches the number of expression change patterns for training without requiring additional manual annotation.

3.4 Optimization Objectives

For training, the EST has two objectives. The first one is a FER classification loss L_{cls}, and the second one is a shuffled snippet order prediction loss L_S. We use the cross-entropy loss for optimization. Formally, the FER loss L_{cls} can be written as:

$$L_{cls} = -E_{\hat{Y}_\mathcal{C}, Y_\mathcal{C}}[Y_\mathcal{C} log \hat{Y}_\mathcal{C}], \tag{5}$$

where $Y_\mathcal{C}$ denotes the facial expression label for each video, $\mathcal{C}$ indexes a training video, and $\hat{Y}_\mathcal{C}$ denotes the probabilities of facial expressions predicted by the EST.

To identity the shuffled snippet order, the order prediction loss function L_S for SSOP is based on:

$$L_S = -E_{\hat{S}_\mathcal{C}, S_\mathcal{C}}[S_\mathcal{C} log \hat{S}_\mathcal{C}], \tag{6}$$

where $\hat{S}_\mathcal{C}$ denotes the permutation type of the shuffled order predicted by the EST, and $S_\mathcal{C}$ is the ground truth one-vs-all label indicating the correct permutation type.

The overall objective function L of the EST is the sum of a classification loss L_{cls} and SSOP loss L_S. Mathematically, the L can be written as:

$$L = L_{cls} + \frac{1}{n} \cdot L_S, \tag{7}$$

where n is the number of snippets. The n is used to average over snippets and avoid the excessive impact of L_S during training.

4 Experimental Results

4.1 Datasets

To evaluate our approach, four face expression datasets were used: BU-3DFE dataset [35], MMI dataset [31], AFEW8.0 dataset [5], and DFEW dataset [14]. The more detailed information about the four datasets is shown in Table. 1.

Table 1: The detailed information of the used datasets.

Dataset	Frame rates	Mean clip lengths	Std. of clip lengths
BU-3DFE	24∼25	99.67	9.95
MMI	25	89.24	35.38
AFEW	24∼25	50.09	26.26
DFEW	23.98	79.26	44.72

BU-3DFE [35]: 3D facial expressions annotated with 6 emotion labels, *i.e.*, anger, disgust, happiness, fear, sadness, and surprise. BU-3DFE contains 606 3D facial expression sequences captured from 101 subjects. Each expression sequence contains nearly 100 frames.

MMI [31]: A total of 205 expression sequences were collected from 30 subjects. The expression sequences were recorded at a temporal resolution of 24 fps. Each expression sequence of the dataset was labeled with one of the six basic expression classes (*i.e.*, anger, disgust, fear, happiness, sadness, and surprise).

AFEW [5]: The AFEW serves as an evaluation platform for the annual EmotiW since 2013. Seven emotion labels are included in AFEW, *i.e.* anger, disgust, fear, happiness, sadness, surprise, and neutral. AFEW contains videos collected from different movies and TV serials with spontaneous expressions, and also contain lots of aforementioned non-expression noises, such as head pose changes, speaking-related mouth movements, occlusions, and illuminations. AFEW is divided into three splits: Train (738 videos), Val (352 videos), and Test (653 videos).

DFEW [14]: The DFEW is a large-scale unconstrained dynamic facial expression database, containing 16,372 video clips extracted from over 1,500 different movies. It contains 12,059 single-label video clips and also includes seven emotion labels, *i.e.* anger, disgust, fear, happiness, sadness, surprise, and neutral. Similar to AFEW, the DFEW also contains the non-expression visual cues that could distract recognition, *e.g.*, head pose changes and speaking-related mouth movements.

4.2 Snippet Extraction and Implementation Details

Snippet Extraction. We unified the input video length to 105 frames via interpolation and clipping operation and detected face regions of each frame to the size of 224×224 via the Retinaface [4]. Then, we randomly selected one of the first 30 frames as the starting frame, and extracted the following 75 consecutive frames to form a video. Next, we split the 75 frames into 7 sub-videos, each of which had 15 frames, with five frames overlapping between each sub-video. To enhance expression movement variation, 5 frames were randomly sampled from each sub-video to form a new sub-video which is an expression snippet. In particular, if encountering over-length videos like 715-frame videos, we still follow the above scheme to obtain snippets. we do have 7*5=35 frames in total for training and testing all the considered videos. However, these 35 frames do not create new contents. They represent 7 5-frame snippets rather than 1 35-frame clip. We show detailed ablation study of snippet settings in later ablation study.

Experimental setting. We used the Pytorch for implementing the EST. The key training parameters include initial learning rate (0.0001), cosine annealing schedule to adjust the learning rate, mini-batch size (8), and warm up. The experiments were conducted on a PC with Intel(R) Xeon(R) Gold 6240C CPU at 2.60GHz and 128GB memory, and NVIDIA GeForce RTX 3090. Following the setting of other compared methods, we conducted a 10-fold person-independent validation on the BU-3DFE and MMI, a Train/Val set validation on the AFEW, and a 5-fold validation on DFEW dataset. The BU-3DFE dataset was used for cross-validation during parameter selection. We kept these parameters unchanged and followed the experimental setup the same as [33, 25, 39] to perform experiments on the other datasets, which have very different data domains.

Baseline structure. In this study, we mainly follow the typical definition of the Transformer structure [32], which is a multi-head attention-based encoder-decoder pipeline, to define our network. In particular, for the baseline structure, we first use a pre-trained ResNet18 [10] to extract

Table 2: Comparison results on the BU-3DFE dataset. Note: the best result is highlighted in bold while * indicates the result is reproduced by author.

Methods	#Input frame amount	Feature setting	WAR/UAR(%)
FAN [25]*	35	frame-based	84.17
DeRL [33]	1	peak frame-based	84.17
C3D [30]	75∼100	sequence-based	82.18
ICNP [40]	75∼100	sequence-based	83.20
C3D-LSTM [26]*	35	sequence-based	79.17
Transformer(Baseline)*	35	sequence-based	85.60
Our EST	35	**snippet-based**	**88.17**

Table 3: Comparison results on the MMI dataset. Note: the best result is highlighted in bold while * indicates the result is reproduced by author.

Methods	Feature setting	WAR(%)	UAR(%)
DeRL [33]	frame-based	73.23	-
WMDCNN [38]	frame-based	78.20	-
FAN [25]*	frame-based	86.49	85.56
AUDN [20]	peak frame-based	75.85	-
CER [17]	peak frame-based	70.12	-
Ensemble Network [29]	peak+neutral frame	91.46	-
LSTM [15]*	sequence-based	70.27	70.48
LPQ-TOP+SRC [13]	sequence-based	64.11	-
SAANet [19]	sequence-based	87.06	-
WMCNN-LSTM [38]	sequence-based	87.10	-
Transformer(Baseline)*	sequence-based	90.50	90.03
Our EST	**snippet-based**	**92.50**	**90.31**

per-frame feature, and then employ 3 encoder layers followed by another 3 decoder layers to progressively translate the extracted features into final expression recognition results. Each of the employed encoder layer consists a series of self-attention operations and multi-layer perception operations, and each of the decoder layer consists a series of cross-attention operations and multi-layer perception operations. Both of the self- and cross-attention operations use 4 head to split the processing. For more details, we refer readers to [32].

In addition, in our EST, the Transformer structure is the same as the baseline but with different inputs. As described in Sec. 3.2, we use AA-SFE to extract features that are later fed into the Transformer for processing. The AA-SFE collects different groups of features from adjacent frames to form different snippets. In each snippet, we apply a hierarchical attention augmentation to compute a 512-dimensional feature vector to represent this snippet. Later, the output feature vectors from different snippets are stacked together and then sent to the Transformer for later processing.

Evaluation metrics. Additionally, consistent with the previous research [39, 14], we choose two different validation metrics, *i.e.,* the unweighted average recall (UAR) and weighted average recall (WAR) to evaluate FER performance on class-uneven datasets like MMI, AFEW and DFEW, while on the class-uniform BU-3DFE dataset, since the results of UAR and WAR are the same, we only use the average accuracy (*i.e.*, WAR) to evaluate our model. The weights are the number of instances in each class. We use Multiply–Accumulate Operations (MACs) to evaluate model complexity [27].

4.3 Overall Performance

4.3.1 Experiments on the BU-3DFE Dataset

The average FER accuracy (i.e., WAR) of the EST was compared with the state-of-the-art methods, including DeRL [33], FAN [25], ICNP [40], C3D [30], FER-Att+Rep+Cls [24], and C3D-LSTM [26] in Table 2. Compared to the best sequence-based result (ICNP) and the baseline Transformer, the proposed EST improved the WAR over 4.97% and 2.57%, respectively. This reveals that our method can effectively discover the more beneficial emotion-related cues by modeling the long-range emotion movement relation in videos.

4.3.2 Experiments on the MMI Dataset

In comparison with the state-of-the-art video-based FER methods, Table 3 lists the WAR on the MMI dataset using deep learning-based methods with spatial feature representation (*i.e.*, AUDN [20], DeRL [33], LSTM [15], Ensemble Network [29], SAANet [19], WMCNN-LSTM [38], WMDCNN [38]), hand-crafted feature based methods (*i.e.*, collaborative expression representation (CER) extracted from the apex frames and LPQ-TOP [13] extracted from the whole sequence), and our EST. As shown in the table, the proposed EST outperformed existing state-of-the-art FER methods. Since most of the existing methods do not use the UAR metric on MMI, we can only reproduce some of the methods that provide source code for UAR comparisons. Compared to

Table 4: Comparison results on AFEW 8.0 dataset. Note: the highest result is highlighted in bold while * indicates the result is reproduced by author.

Methods	Feature setting	WAR(%)	UAR(%)
FAN [25]	frame-based	51.18	-
HoloNet [34]	frame-based	44.57	-
DSN-HoloNet [12]	frame-based	46.47	-
DSN-VGGFace [7]	frame-based	48.04	-
C3D [30]	sequence-based	46.72	43.75
DenseNet-161 [18]	sequence-based	51.44	-
ResNet18+GRU [10]	sequence-based	49.34	45.12
Emotion-BEEU [16]	sequence-based	52.49	-
3D ResNet18 [9]	sequence-based	45.67	42.14
Former-DFER [39]	sequence-based	50.92	47.42
Transformer(Baseline)*	sequence-based	49.72	43.95
Our EST	**snippet-based**	**54.26**	**49.57**

the second best method, Ensemble Network [29], the EST improved the WAR of 1.04%. Compared to the frame-based method FAN [25], our EST achieved a 4.75% and 6.01% boost on the UAR and WAR, respectively.

4.3.3 Experiments on the AFEW Dataset

Although accuracies of Disgust and Fear are relatively lower than the other categories, our method still outperforms other methods in recognizing both emotions. This shows that our proposed EST method which contains AA-SFE for extracting visual features and SSOP for refining the estimation of attention weights is more effective for difficult emotion recognition classes.This may be caused by better modeling the relations of subtle expression movements between snippets. Table 4 reports the WAR and UAR using the EST and state-of-the-art methods. It demonstrates that our method achieves the best performance with great robustness on both WAR and UAR metrics, meanwhile, has obvious advantages over other algorithms (e.g. for the second best method Former-DFER [39], achieving relative increase of 6.16% on WAR and 4.34% on UAR) on the in-the-wild expression dataset. Additionally, our EST improved the baseline (Transformer), achieving a relative WAR increase of 8.37%, which can validate the robustness of our method.

Table 5: Comparison results on DFEW dataset. Note: the highest result is highlighted in bold while * indicates the result is reproduced by author

Methods	Feature setting	WAR(%)	UAR(%)
3D ResNet-18,EC-STFL [14]	sequence-based	56.51	44.73
C3D,EC-STFL [14]	sequence-based	55.50	45.10
P3D,EC-STFL [14]	sequence-based	56.48	45.22
R3D18,EC-STFL [14]	sequence-based	56.19	45.05
VGG11+LSTM,EC-STFL [14]	sequence-based	56.25	44.78
Former-DFER [39]	sequence-based	65.70	53.69
Transformer(Baseline)*	sequence-based	63.85	50.39
Our EST	**snippet-based**	**65.85**	**53.94**

4.3.4 Experiments on the DFEW Dataset

The highest accuracy is 86.87% of Happiness followed by Anger, which achieves 71.84%. Although we only achieved 5.52% accuracy in the Disgust category due to the huge imbalance of categories in the DFEW (only occupies 1.22% in the DFEW dataset), the compared results in Table 5 show that our method is still far superior to other algorithms. More detailed comparison results can be shown in Table 5. Compared to the second best method Former-DFER [39], both the WAR and UAR of our EST achieved significant improvement, e.g., having an relative increase of 0.23% on WAR and 0.46% on UAR).

4.3.5 Confusion Matrices

Fig. 4(a) shows the confusion matrix of BU-3DFE for video FER by using our method. Among the six expressions, the highest accuracy are 95.0% of Happiness and Surprise, while the lowest accuracy is 80.0% for Fear, which has the least amount of facial expression movement and is difficult to distinguish with Disgust. The average FER accuracy is 88.17%. Fig. 4(b) depicts the confusion matrix of MMI for video FER by using our method. We achieved 100% accuracy in Surprise category. The average accuracy of FER is 92.5%. Fig. 4(c) shows the confusion matrix on the challenging AFEW dataset. The average accuracy of FER achieved 54.26%. The highest accuracy is 87.04% of Happiness followed by Anger and Neutral, which respectively reach 78.69% and 75.81%. Fig. 4(d) shows the confusion matrix of FER on the large-scale DFEW dataset. The average accuracy of FER achieved 65.85% by using our EST.

Despite the effectiveness of our method, we can ob-

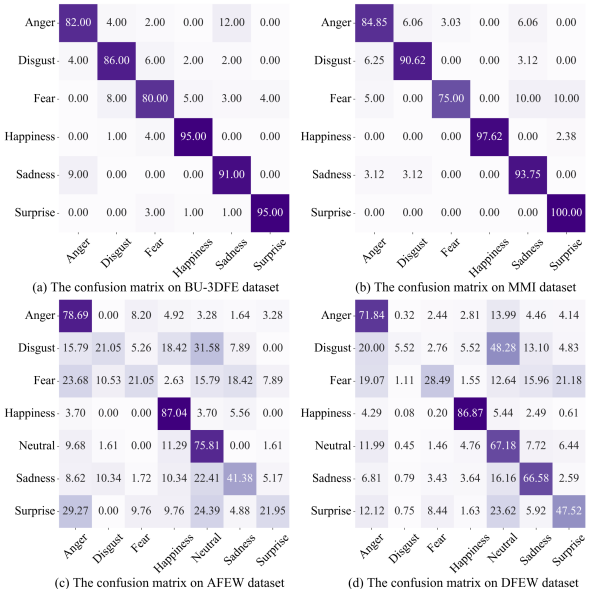

Figure 4: The confusion matrices for video-based FER on the four datasets.

serve that our method has some bad predictions on the DFEW and AFEW datasets. With careful investigation, we found that most of these bad cases of our method is due to the inter-class confusion, *e.g.*, the confusion between Disgust and Neutral or between Fear and Anger. We analyze that this inter-class confusions are mostly related to both extremely subtle visual differences and very unbalanced data distributions in the dataset. Such inter-class confusions are mostly related to both extremely subtle visual differences and unbalanced data distributions in the dataset. For subtle visual differences, although our method can already model plenty of subtle visual differences and achieve state-of-the-art performance, we believe that there are still some cases that require further investigation and research. Regarding the unbalanced distributions, we found that the numbers of occurrences of different emotion classes in a dataset like DFEW and AFEW are quite different, which results in a typical long-tail problem and increases the difficulty of learning on minority classes. For example, in a dataset, the Disgust may have only a few hundreds of examples for training, while

Table 6: Ablation study of the proposed EST. Impact of integrating our different components (AA-SFE and SSOP) into the baseline Transformer on the BU-3DFE dataset.

Transformer	AA-SFE	SSOP	Params(M)	MACs(G)	WAR/UAR(%)
✓			34.37	63.85	85.60
✓	✓		34.37	63.88	87.12
✓	✓	✓	42.78	63.89	**88.17**

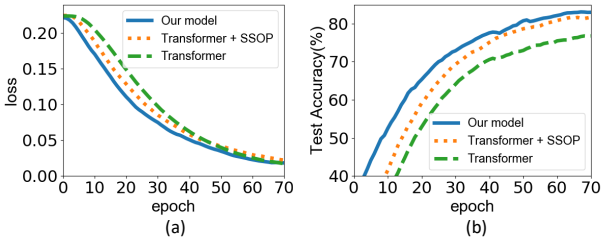

Figure 5: The learning procedure for the EST with different components during training and testing. (a) The training loss variation in terms of epochs, (b) the testing accuracy variation in terms of epochs.

the Neutral may have several thousands. However, we believe that the study of such unbalanced data and the long-tail problem is simply beyond the topic of our paper. We will try to tackle this problem in the future.

4.4 Ablation Experiment and Further Analysis

4.4.1 Effects of Different Components

To better understand the role of each module in the proposed EST, Table 6 presents the ablation results of the gradual addition AA-SFE and SSOP components to the baseline Transformer framework on the BU-3DFE dataset. The Transformer achieved a video-based FER accuracy (see WAR in the Table) of 85.60%. The further integration of AA-SFE improved the accuracy to 87.12%, as the AA-SFE aids in the extraction of snippet-level features via jointly hierarchical attentions. Thanks to learning the order sensitive representation, the addition of SSOP resulted in an increase of 1.05%.

We also show the convergence performance of the EST with different components in Fig. 5. The green dotted curves belong to the baseline Transformer. From the view

Table 7: Ablation study of different attention selection in AA-SFE. The best results are highlighted in bold.

Different attention	Params(M)	WAR/UAR(%)
w/o attention	34.37	85.60
self-attention	42.78	87.63
SE-like attention [11]	43.30	87.46
Our hierarchical attention	42.78	**88.17**

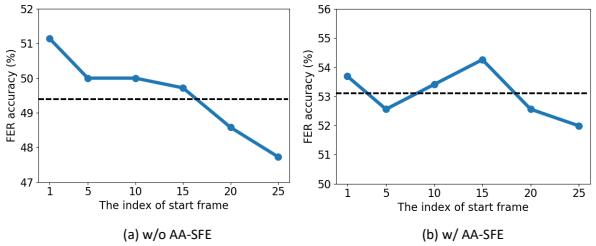

Figure 6: Compare with AA-SFE and without AA-SFE under different start frames. (a) Without AA-SFE. (b) With AA-SFE. The dashed black lines represent the mean FER accuracy.

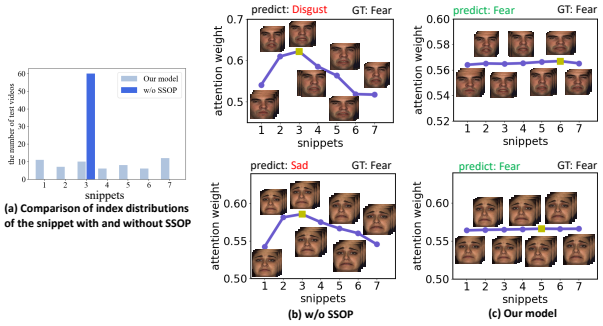

Figure 7: The effects of SSOP. (a) Comparison of index distributions of the snippet with the highest attention weight with and without SSOP in EST, where the horizontal axis shows the seven snippets in videos and the vertical axis shows the number of the highest attention weights received on different snippets. (b) Comparison of modeling subtle facial expression movements without SSOP. (c) Comparison of modeling subtle facial expression movements with SSOP. From the results, we can observe that the EST without SSOP always focused on the 3-rd snippet in each video, while the EST with SSOP helps obtain comprehensively attentional weights according to expression changes.

of the decline rates of training losses, obviously, the gradual addition of AA-SFE and SSOP improved the Transformer performance on both training speed and stability. Meanwhile, the proposed EST with AA-SFE and SSOP is easier to achieve higher accuracy on the test set. In the experimental comparison, we make sure that the training stops once the method converges. In this figure, we only show the statistics with epoch 70 due to the limitation space.

Table 7 lists the recognition results with different attention selection in the AA-SFE. Obviously, two-level hierarchical attention used in AA-SFE achieved the best performance without any computational cost, helping to describe more informative snippet features. In addition, we use different starting frames to obtain snippets for FER, so that the same frame can have different locations in a snippet according to different starting frames. We have studied the final recognition performance with or without using the AA-SFE. Fig. 6 shows that our AA-SFE achieved consistently high emotion recognition performance with a variance of 0.669, while the setting without AA-SFE achieved vastly degraded performance with a variance of 1.2.

4.4.2 Effects of SSOP Head and Permutation Order

Fig. 7 shows more analysis about the effect of the SSOP in the EST. In particular, Fig. 7(a) presents the distribution of the index of the snippet with the highest attention weight in EST with and without SSOP, respectively. Without SSOP (see the dark-blue column in Fig. 7(a)), we can observe that the EST always focused on the 3-rd snippet, which usually contains the peak changes in each test video. Alternatively, the SSOP can make EST distribute similar attention to all the snippets. We further illustrate the SSOP modeling subtle facial expression movements in Fig. 7(b)(c). The results show that the SSOP helps obtain more discriminative representation by comprehensively making the Transformer model inter-snippet visual changes.

To further evaluate the effect of the types of shuffle order in SSOP learning, Fig.8 presents the variation curves of FER accuracy and snippet order prediction accuracy

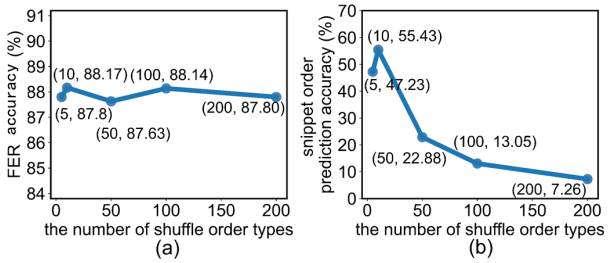

Figure 8: The influence of the types of shuffle order for FER accuracy and order prediction accuracy on the BU-3DFE dataset. (a) FER accuracy, (b) snippet order prediction accuracy.

according to the number of shuffle order types on the BU-3DFE dataset. As shown in Fig.8 (a)(b), when the number of the types is 10, both the FER accuracy and snippet order prediction accuracy reach the highest 88.17% and 55.43%, respectively. Therefore, during the training, we set the types of shuffle order to 10.

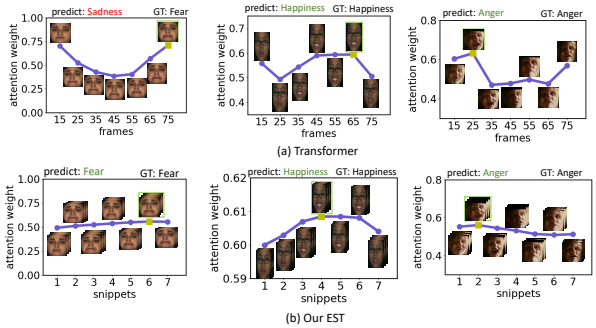

Figure 9: Comparison of modeling subtle facial expression movements in FER on the BU-3DFE, MMI, AFEW, DFEW dataset. (a) vanilla Transformer, (b) our EST. Note: the green square is located at the position of the most informative expression snippet with the most attention weight.

In addition, we present the prediction accuracy of the shuffle order for the SSOP in Table 8. As shown in the Table, even on the challenging DFEW dataset, the SSOP achieves the accuracy of 46.1% for predicting 10 different permutation orders used in the training. We would like to mention that the accuracy is supposed to be not very high because the shuffled order is mainly used to break the discrepancy between the emotion recognition and some specific snippet, and the obtained results align with this goal to some extents.

Table 8: The prediction accuracy of the shuffle order for the SSOP on four datasets.

Datasets	BU-3DFE	MMI	AFEW	DFEW
Accuracy (%)	55.4	45.0	50.9	46.1

Table 9: Study of SSOP with perform zero-padding to the frames in the middle of a video in BU-3DFE.

	w/o SSOP	w/ SSOP
WAR(%)	82.5	**87.5**

To present additional empirical evidence of how SSOP works, we performed an extra experiment. In this experiment, we perform zero-padding to the frames in the middle of a video, so that a model cannot use these frames to perform recognition. Then, we test the results of using or not using the the SSOP are listed in Table. 9. We can find that without frames from middle of the input video, SSOP still performs favorably while normal Transformer without SSOP cannot recognize the emotion correctly.

4.4.3 Effects of Snippet Settings

In Fig. 10, we present the FER accuracy curves, which effected by the number of frames in per snippet and the number of snippets in per video. As shown in the Fig. 10 (a), the accuracy reached the highest 88.17% when we set the number of frames of each snippet to 5. Fig. 10 (b) shows that the accuracy reached the highest when the number of snippets in per video is set to 7. Thus, in our study, $n = 7$ and $j = 5$.

Besides, we also observe that different snippet amounts and snippet lengths only resulted in minor performance changes, suggesting that these hyperparameters are less important to our method. Hence, thanks to the long-range relation modeling ability of the Transformer, our EST can be easily extended to adapt to videos of almost any length upon proper training. According to the snippet generation process as described in Sec.4.2, we would like to further clarify that this procedure is quite different from a sliding window strategy. Our snippet is collected mainly based

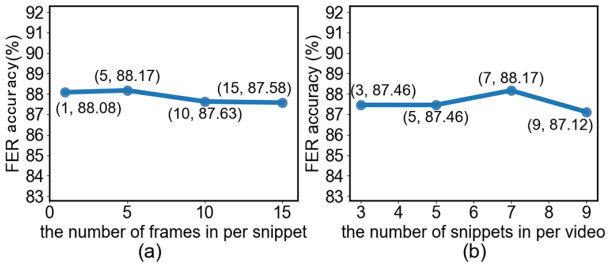

Figure 10: The impact of the number of frames in per snippet and the number of snippets in per video for FER on the BU-3DFE dataset. (a) The effect of the number of frames in per snippet, (b) the effect of the number of expression snippets in per video.

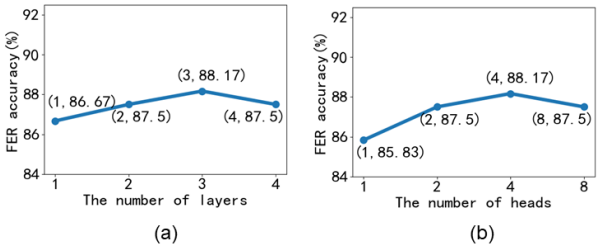

Figure 11: The influence of meta-parameter setting in Transformer on the BU-3DFE dataset.(a) The effect of the number of layers, (b) the effect of the number of heads.

on sampling, while the sliding window processes input in a comprehensive and sequential order. Different from the aforementioned snippet generation process, for a sliding window strategy, it would first collect a few frames within the time window for processing, and then slide by a step of one frame for later processing. The collections of frames between two adjacent sliding steps then highly overlap with each other, and all frames within the window would be exhaustively considered for processing. We believe this could generate plenty of abundant and trivial information that might affect the recognition performance. This hypothesis can be partially supported by the ablation study presented in Figure 10 in which using more frames or more snippets does not contribute to a better performance. For training, the seven snippets were shuffled in a random order (the frame order within each snippet remained unchanged). For test, we only used the normal snippet order as input for robust FER.

4.4.4 Effects of Meta-parameter Settings in Transformer

Fig. 11 presents the FER accuracy curves, which effected by the number of encoder layers and attention head in the Transformer architecture. As shown from the results, the accuracy achieved to the highest 88.17% when we set the number of encoder layers to 3 and the number of attention heads to 4.

4.4.5 Visualization Results of Expression Changes and Representations

Fig. 9 shows the comparison of expression relation curves for modeling subtle facial expression movements between the vanilla Transformer and the proposed EST on four videos from the four datasets. From the Fig. 9(a), the vanilla Transformer shows to only focus on the frames with peak expression patterns, which can be easily affected by noises such as head poses and other non-expression changes. Instead, by decomposing videos into snippets, although the changes of expression movements of intra-/inter-snippets are very subtle in a video, our EST can attend on on all expression snippet changes more comprehensively and can effectively locate the most informative expression snippet by the modelled expression attention weights (see the values of the Ordinate in Fig. 9(b)). This demonstrated that the EST can effectively tackle the problem of the vanilla Transformer and help achieve more robust modeling of subtle facial changes in different videos.

In Fig. 12, we visualized the emotion representations with different settings in a 2D feature space by using the t-SNE [23] on the four datasets. In Fig. 12, we visualized the emotion representations with different settings in a 2D feature space by using the t-SNE [23] on the four datasets. Following existing work [39], we randomly chose one of the folds for visualization on datasets that were conducted cross-validation. The visualizations include the following three cases: video features extracted by FAN [25] (see Fig. 12(a)), sequence-based video features extracted by LSTM [15] (see Fig. 12(b)), emotion-rich representations by our EST (see Fig. 12(c)). Obviously, compared the

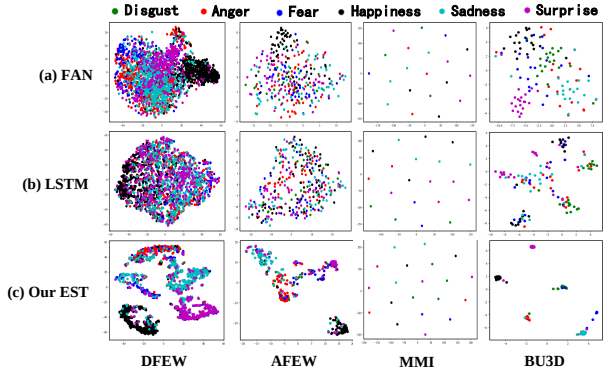

Figure 12: The comparison of different features in 2D space by t-SNE visualization. (a) The frame-based features learned by FAN [25], (b) the sequence-based features learned by LSTM, (c) the unified salient emotion features learned by EST. (Figure best viewed in color)

other emotion features, we can observe that the comprehensive and robust expression representation learned by the EST includes more visual change cues and can significantly be separated according to different expression categories.

4.4.6 Analysis on Model Complexity

Table 10 reports model parameters and computational costs of spatial-temporal learning methods on the AFEW dataset. In general, our EST has the best performance (accuracy of 54.26%) with a small computational cost (63.89G MACs) and real-time speed (412 fps), which means that the proposed method exhibits improved accuracy and efficiency, achieving a better trade-off between accuracy and efficiency. In addition, model sizes and computational complexities are decided by different factors. For example, the model size is mainly decided by the factors like the numbers of network layers, the sizes of convolutional kernels, the dimensions of input/output features, and so on. On the other hand, the computational complexity of a neural network can be decided by the factors like input image size, non-parametric operations like self-attention operations, and so on. As a result, it can be possible that the model sizes of two networks are different (*e.g.*, 34.4M vs 42.8M), but their computational complexities are similar (*e.g.*, 63.88G MACs vs. 63.89G MACs).

Table 10: Comparison of model complexity and efficiency.

Methods	Input	Backbone	Params(M)	MACs(G)	fps	WAR(%)
FERAtt [24]	Frame	ResNet-18	67.08	13.56	75	37.22
Dense161 [18]	Video	DenseNet-161	26.52	272.47	47	51.44
VGG16TPSA [1]	Video	VGG16	14.72	537.61	552	49.00
Our EST	Video	ResNet-18	42.78	63.89	412	**54.26**

Table 11: Cross-validation comparison with state-of-the-art methods on DFEW → AFEW and BU-3DFE → AFEW. The best results are in bold.

Cross-validation	Methods	WAR(%)
DFEW → AFEW	Transformer(baseline)	48.86
	EST	**51.14**
BU-3DFE → AFEW	Transformer(baseline)	19.66
	EST	**25.17**

4.4.7 Evaluation on cross-databases

In addition, to verify the generalizability of EST, cross-database validation was conducted on the challenging in-the-wild BU-3DFE, DFEW and AFEW datasets. Images from the BU-3DFE and DFEW dataset were used for training, respectively, whereas images from the AFEW testing set were used for testing without fine-tuning. Table 11 presents the comparison results of the proposed model and state-of-the-art methods, including vanilla Transformer [32]. Although the training and testing datasets have different settings (e.g., scene, pose, lighting, ethnicity, age, etc.), the results of EST demonstrate that it is reusable for facial expression recognition on the AFEW dataset. Our method achieved an accuracy of 51.14%,25.17%, gaining 2.28%, 5.51%, improvements over the recognition accuracy of vanilla Transformer, respectively.

5 Conclusions and Future Works

In this paper, we have carefully explored the challenges and difficulties when applying Transformer to the video-based facial expression recognition task. In particular, we found that the trivial or irrelevant information and extremely subtle visual changes could affect the performance of Transformer to extract useful features and make

accurate predictions. By addressing these issues, we propose to apply snippets to decompose the modeling of facial changes in the whole video into the modeling of a series of sub-videos, so that a Transformer can find it easier to learn useful information for robust expression recognition. As a result, we obtain an expression snippet Transformer (EST). In our proposed EST, we further introduce a novel attention-augmented snippet feature extraction (AA-SFE) module to augment the feature extraction of each snippet by modeling inter-snippet relations effectively. Besides, we also developed a shuffled snippet order prediction (SSOP) module to help break the dependency of expression prediction to some specific snippet locations by shuffling the snippets. In practice, with the AA-SFE and SSOP, our EST achieved state-of-the-art performance on four challenging datasets (BU-3DFE, MMI, AFEW, and DFEW). To the other researchers in the community, we believe that our explorations can bring new insights about how to exploit the impressive relation modeling power of a Transformer more effectively for expression recognition.

Despite the effectiveness of our method, we found that there are still some rooms that our EST could be improved. For example, some extremely minor visual changes are still difficult to be captured by current methods. Besides, the long-tail problem caused by unbalanced data distribution also exist in the video-based facial expression recognition and affected the methods' performance. Therefore, in the future, we will introduce self-supervised learning mechanisms into our Transformers to further capture the emotion-rich features as well as explore information from unlabeled data for better capabilities.

Acknowledgments

This work was partially supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700), the National Natural Science Foundation of China grant (Grant No. 62076227, 62002090) and Wuhan Applied Fundamental Frontier Project under Grant (No. 2020010601012166). Dr Zhe Chen is supported by Australian Research Council Project IH-180100002.

References

[1] Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger. Emotion recognition with spatial attention and temporal softmax pooling. In *International Conference on Image Analysis and Recognition*, pages 323–331. Springer, 2019.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.

[3] Jingying Chen, Lei Yang, Lei Tan, and Ruyi Xu. Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition. *Pattern Recognition*, 129:108753, 2022.

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[5] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*, pages 423–426, 2015.

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

[7] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pages 584–588, 2018.

[8] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[12] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI)*, pages 553–560, 2017.

[13] Bihan Jiang, Michel Valstar, Brais Martinez, and Maja Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE transactions on cybernetics*, 44(2):161–174, 2013.

[14] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, pages 2881–2889, 2020.

[15] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017.

[16] Vikas Kumar, Shivansh Rao, and Li Yu. Noisy student training using body language dataset improves facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 756–773. Springer, 2020.

[17] Seung Ho Lee, Wissam J Baddar, and Yong Man Ro. Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. *Pattern Recognition*, 54:52–67, 2016.

[18] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pages 630–634, 2018.

[19] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.

[20] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159:126–136, 2015.

[21] Xiaofeng Liu, Linghao Jin, Xu Han, and Jane You. Mutual information regularized identity-aware facial expression recognition in compressed video. *Pattern Recognition*, 119:108105, 2021.

[22] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[24] Pedro D Marrero Fernandez, Fidel A Guerrero Pena, Tsang Ren, and Alexandre Cunha. Feratt: Facial

expression recognition with attention net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 837–846, 2019.

[25] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870. IEEE, 2019.

[26] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–28, 2017.

[27] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 367–376, 2021.

[28] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[29] Ning Sun, Qi Li, Ruizhi Huan, Jixin Liu, and Guang Han. Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters*, 119:49–61, 2019.

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 4489–4497, 2015.

[31] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France., 2010.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, page 6000–6010. Curran Associates Inc., 2017.

[33] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2168–2177, 2018.

[34] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen. Holonet: towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 472–478, 2016.

[35] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 211–216. IEEE, 2006.

[36] Wenmeng Yu and Hua Xu. Co-attentive multitask convolutional neural network for facial expression recognition. *Pattern Recognition*, 123:108401, 2022.

[37] Zhenbo Yu, Qinshan Liu, and Guangcan Liu. Deeper cascaded peak-piloted network for weak expression recognition. *The Visual Computer*, 34(12):1691–1699, 2018.

[38] Hepeng Zhang, Bin Huang, and Guohui Tian. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognition Letters*, 131:128–134, 2020.

[39] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the ACM International Conference on Multimedia*, 2021.

[40] Qingkai Zhen, Di Huang, Yunhong Wang, and Liming Chen. Muscular movement model-based automatic 3d/4d facial expression recognition. *IEEE*

Transactions on Multimedia, 18(7):1438–1450, 2016.