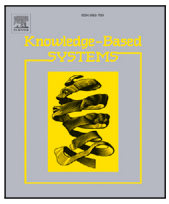
Knowledge-Based Systems 350 (2026) 116590

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Compressed video-driven multimodal modeling and interaction for dynamic expression recognition

Weicheng Xie [a,b,d], Junliang Zhang [a], Haijian Liang [a], Linlin Shen [c,d,*], Zhihui Lai [a], Siyang Song [e], Zitong Yu [f]

[a] School of Computer Science & Software Engineering, Shenzhen University, China
[b] Anhui Provincial Key Laboratory of Humanoid Robots, Hefei, 230000, China
[c] Computer Vision Institute, School of Artificial Intelligence, Shenzhen University, China
[d] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, China
[e] School of Computer Science, University of Exeter, UK
[f] Department of Computing and Information Technology, Great Bay University, China

ARTICLE INFO

Keywords:
Dynamic expression recognition
Compressed video
Modal-parallel temporal modeling
Multimodal interaction

ABSTRACT

Most existing dynamic facial expression recognition (DFER) algorithms are developed based on raw video data, e.g. using 3D CNN-, RNN-, or Transformer-based models to incorporate temporal cues of raw videos, while they overlook that the videos obtained in practice are compressed videos due to their advantage of low storage requirements. Consequently, these algorithms may overfit to redundant patterns in raw videos and fail to exploit the multimodal information available in compressed data. To address these problems, we propose a compressed video-driven framework for DFER. Unlike other schemes that use only decoded RGB frames, our approach is carried out entirely in the compressed domain and uses all readily available modalities, i.e., I-frames, motion vectors, and residuals, for the purpose of making full use of the compressed-video modalities and their interaction in the video. To obtain richer expression features, we design a multimodal interaction attention paradigm to model temporal modality sequences, inter-modal weighting, multimodal interaction, and intra-modal enhancement based on the compressed-video representation. Our algorithm benefits from not only the modal-parallel temporal modeling of the multimodal representation to take into account the characteristics of each modality, but also the inter-modal weighting, multimodal interaction and intra-modal enhancement. Extensive experimental and ablation studies have shown that our approach achieves state-of-the-art (SOTA) results in three benchmark datasets, i.e. Oulu-CASIA, AFEW and DFEW. Our codes are publicly available at https://github.com/JunLiangZ/CompressedVideo-DFER.

1. Introduction

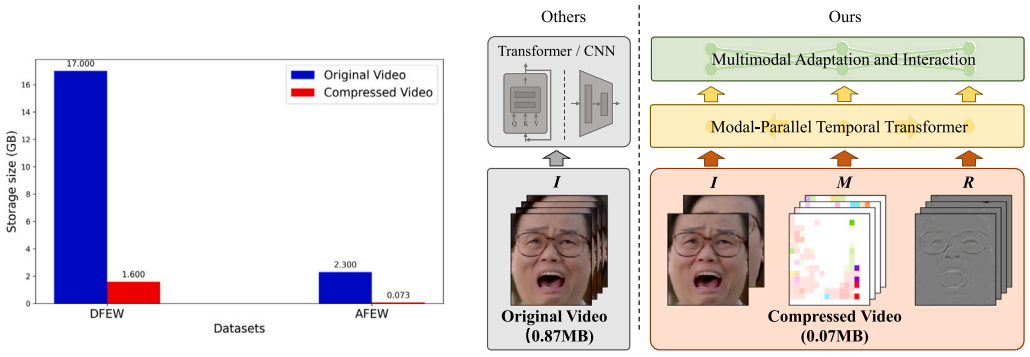
Facial expression is one of the most natural signals for humans to convey their emotions and intentions, which can strengthen the emotional connection between people. Facial Expression Recognition (FER) aims to classify several basic expression categories from images or videos, including ‘happy’, ‘angry’, ‘disgust’, ‘fear’, ‘sad’, ‘neutral’ and ‘surprise’ [1]. Since FER has extraordinary significance for understanding human emotions, it has a wide range of applications in real life, such as human–computer interaction, medical diagnosis, classroom teaching, etc [2].

Expression video carries rich cues of emotions [3] since it can additionally capture temporal dynamics that cannot be conveyed by static images. However, with the rapid growth of video content, storage requirements have increased dramatically, and reducing storage and transmission costs has become a main concern [4]. In order to reduce storage space or enable large-scale video transmission in bandwidth-limited communication networks, videos are frequently converted into the compressed format (i.e., compressed domain) [5]. However, most existing Dynamic Facial Expression Recognition (DFER) methods rely on raw videos. They do not consider the storage burden of raw data or exploit the multimodal cues in compressed videos, often using a single modality [6,7] or processing them separately [8]. This neglect of multimodal information is partly due to the difficulty of reasoning about expressive cross-modal interactions [9].

One of the ongoing challenges for DFER in the compressed domain of videos is how to effectively model the spatio-temporal dynamic representation of each modality. Though video sequences contain temporal information that sheds light on the dynamic of facial expressions, quite

* Corresponding author at: Computer Vision Institute, School of Artificial Intelligence, Shenzhen University, China.
E-mail address: llshen@szu.edu.cn (L. Shen).

https://doi.org/10.1016/j.knosys.2026.116590
Received 20 March 2025; Received in revised form 27 May 2026; Accepted 30 June 2026
Available online 7 July 2026



(a) Storage size before and after video compression **(b) Comparison of our method with other methods**

Fig. 1. (a) Storage space usages before (blue) and after (red) video compression. (b) Comparison between previous methods and ours for compressed video recognition on DFEW. Compared to other methods that operate on the original video (with the storage size of 0.87MB), our method operates directly on the compressed domain (with the storage size of 0.07MB), allowing collaborative representation of I-frames (I), motion vectors (M), and residuals (R) for multimodal interactive learning.

a few CNN-based methods [10] only model the visual information at the frame level, and do not involve visual relationship reasoning between different frames. Sequence-based approaches are modeled by long short-term memory (LSTM) [11] or 3D convolution [12]. However, they still perform mediocrely in describing subtle expressive motions when the scale of model parameters is not large enough. In addition, it is also very common to use optical flow to capture dynamic motion information [8,13], but obtaining optical flow frame by frame is very time-consuming, which is usually the main computational bottleneck of recognition models. Meanwhile, the successive frames of the video are filled with many uninformative and repetitive patterns [6], which may also drown out the truly meaningful information about the dynamic facial expressions. Although several works have attempted to explore the compressed domain (e.g., MIC [6]), they are often limited to single-modality inputs such as residuals and lack explicit inter-modal coordination or temporal modeling. In contrast, our method jointly exploits I-frames, motion vectors, and residuals, and introduces structured modules to capture both inter- and intra-modal dependencies for more comprehensive dynamic expression representation.

Specifically, to address the aforementioned issues, we exploit compressed representations developed for storing and transmitting videos, rather than operating on raw video (see Fig. 1(b)). We employ the compressed representation of expression video from the motivations of three aspects.

(i) Video compression (using H.264, HEVC, etc.) reduces the storage overhead (e.g., as shown in Fig. 1(a), video compression reduces the storage size of DFEW from 17 GB to 1.6 GB, reducing the storage overhead by 90.5%). As shown in Fig. 2, a compressed video contains only a small number of original frames (i.e., I-frames), and the cues associated with the other frames are encoded in motion vectors and residuals occupying much less space than the original I-frames. In this way, the compressed video implicitly provides three types of modality information, where an I-frame denotes a complete RGB image of an video, the motion vectors being similar to optical flow represent pixel shifts between consecutive frames, and the residuals represent the difference between the pixel value of the original frame and that estimated by motion vectors, which are usually treated as the changes in appearance and motion boundaries. Thus, the compressed representation can remove many uninformative and repetitive patterns, and highlight the meaningful signal of the original video [14].

(ii) Since motion vectors can represent similar motion signals for the-same-category expressions from different identities, e.g. skin colors and lighting conditions, they can reduce the influence of identities and expression-irrelevant factors on expression recognition.

(iii) Since residuals can capture subtle variations of the key expression parts (e.g., mouth or eyes) reflecting expression semantics, they mitigate the effect of the inherent large-intra-class gap in FER task. These merits motivate us to act directly in the compressed domain of video expressions (please refer to Fig. 1). We then treat the three types of information representations, i.e., I-frames, motion vectors, and residuals as three different modalities, and perform feature representation learning via multimodal interaction.

Table 1
Comparison of our method with related methods. I represents the original image, M represents the motion vector, and R represents the residual. MITM denotes the Multimodal inter-frame temporal modeling. Mult denotes the weighting, interaction and enhancement between the three modalities, i.e. I, R, M.

Methods	I	M	R	MITM	Mult
MIC [6]'2021	No	No	Yes	No	No
Former-DFER [15]' 2021	Yes	No	No	No	No
DPCNet [16]' 2022	Yes	No	No	No	No
EST [17]' 2023	Yes	No	No	No	No
LOGO-FORMER [18]' 2023	Yes	No	No	No	No
Ours	Yes	Yes	Yes	Yes	Yes

To encode the multimodal features for the compressed video, we use the transformer [19] to complement the ResNet backbone, motivated from its powerful capacity of modeling global relationship in natural language processing (NLP) [19], video [20] and multimodal [21,22]. Based on this, we introduce a modal-parallel temporal transformer for the three modalities to allow the model to capture long-term dependencies in video sequences and represent the temporal characteristics of each modality in a parallel fashion. In addition, we design a multimodal adaptation and interaction module to learn complementary representations and inter-modal interaction cues. In particular, the recent work MIC [6] also recognizes dynamic facial expressions in the compressed domain of a video, while it only uses a single modality (i.e., residuals) and ignores the correlations between modalities. Our approach extracts richer information by fully utilizing the cues from multiple modalities as well as the interactions between modalities, thus representing expressions more accurately and comprehensively.

For clarity, we illustrate the main differences between our compressed video-driven method and existing works in Table 1. It shows that our method fully utilizes the cues from the three compressed-video modalities, instead of relying solely on single-modality information (e.g., I or R), compared with existing methods. In addition to the temporal cues of each modality, the complementarity of the cues between modalities is also considered. Therefore, we design a modal-parallel temporal transformer as well as a multimodal adaptation and interaction module, to better capture the contextual information and inter-modal complementarity in video expressions. Our work's contributions are summarized as:

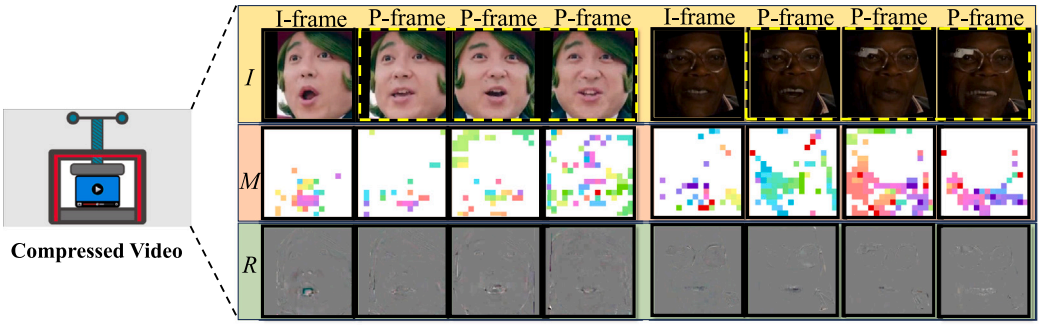


Fig. 2. The visualization of three modalities of two compressed videos under different lighting conditions from the DFEW dataset. The 1st row includes the I-frames (I), as well as the reconstructed frames (labeled with the yellow dotted boxes) using the I-frames, motion vectors and residuals. Motion vectors (M) in the 2nd row capture the motion information of consecutive frames. Residuals (R) in the 3rd row provide subtle changes in the face, such as the mouth and eyes. P-frames are dependent on the reference frame (i.e., I) for decoding, and they store only the differences (i.e., M and R) from the reference frame I.

- We propose a compressed video-driven framework for DFER that operates directly in the compressed domain, utilizing multimodal-based temporal modeling and inter-modal interaction. To the best of our knowledge, this is one of pioneering works to explore multimodal temporal modeling and interaction for the recognition of compressed expression videos.
- We design a DFER-oriented multimodal interaction attention mechanism to encode both the temporal and cross-modal interaction cues by modeling the inconsistency and complementarity among the three compressed-video modalities under an asymmetric target-source interaction strategy.
- Extensive experimental results on three datasets, i.e. Oulu-CASIA, AFEW and DFEW demonstrate the superiority of our multimodal-based paradigm, over existing state-of-the-art (SOTA) methods in terms of WAR and UAR.

2. Related work

2.1. Video-based FER

For video-based FER, although hand-crafted feature methods [23] and STLMBP [24] show reasonable effectiveness, they have limited applicability in specific scenarios and are difficult to optimize. Currently, deep learning techniques of Long Short-Term Memory (LSTM) [25] and Convolutional 3D (C3D) [26] are the two most widely used methods for this task, and these methods usually extract spatio-temporal facial information from video sequences, e.g. Zhang et al. [27] proposed a double-channel weighted mixture CNN–LSTM network, Yu et al. [28] extracted both local–global and spatial–temporal information based on the cascaded CNN–LSTM structure, and Kim et al. [10] integrated C3D and LSTM for spatiotemporal feature representation learning. More recently, Gan et al. [29] explored motion similarity for macro-/micro-expression recognition, and Liu et al. [30] introduced muscle-movement-aware transformer representations for FER.

However, current algorithms rarely take into account the high storage overhead of original videos, which is often tens of times higher than that of the compressed format in actual transmission or storage. In terms of methodology, although current methods are robust to variations in expression intensity, they only consider a single visual modality and fail to fully utilize the multimodal information of video data. In addition, there are many uninformative and repetitive patterns in consecutive frames, and this redundant information may impair the generalization capacity of models on learning truly meaningful expression information.

Furthermore, recent studies have shown that compressed video implicitly contains cues from multiple modalities, i.e., I-frames, motion vectors, and residuals [31]. Meanwhile, video compression can remove up to two orders of magnitude of redundant information, and motion vectors and residuals can provide original RGB images with additional motion cues [14,32]. Liu et al. [6] propose to conduct the FER task for dynamic expressions in the compressed domain, while it uses only the residual information and ignores the complementary cues in the three modalities. To this end, our method employs all the three modalities in the compressed domain, and proposes the multimodal interaction between these modalities for exploring this complementarity.

Different modalities such as visual signals, natural language or acoustic signals are often complementary in terms of semantics, exploring effective integration between different modalities is thus a key aspect of multimodal learning [33]. Compared with simple fusion methods of different modalities, e.g. [34] that tend to ignore the complementary information between modalities, multimodal transformers achieve cross-modal interactions to capture complex relationships and interactions between different modality data [35]. For the interaction between multiple modalities in multimodal transformer, Tsai et al. [36] proposed to align the multimodal language sequences, Lv et al. [37] suggested to progressively enhance modal features, Zhang et al. [38] introduced an unified transformer-based multimodal framework for two tasks. However, these approaches [36–38] mainly focus on the interactive fusion between modalities, ignoring the importance inconsistency among different modalities. For instance, the multimodal cues of I-frames, motion vectors, and residuals show largely inconsistent importance to the compressed expression videos.

To this end, we propose to not only implement interactive learning between modalities in the compressed domain, but also adaptively weigh and enhance each modality to trade off their contributions for the final representation.

2.2. Multimodal interaction attention mechanism

The multimodal attention mechanism enables the model to focus on multiple modalities or levels of features to achieve a more comprehensive representation, which has been widely adopted in the fields of machine translation [39], image detection [40], and emotion recognition [41]. For the task of multimodal emotion representation, the multimodal attention mechanism is frequently used to capture the correlation between modalities, e.g. Yadav and Vishwakarma [42] used it to explore the correlation between image and text modalities, and Akhtar et al. [43] explored the correlation between adjacent utterances and their multimodal information. However, these representation methods tend to ignore the inconsistency between modalities, while this inconsistency exists widely in multimodal-related tasks, e.g. DFER.

For DFER, current works use the attention mechanism to mainly explore the multi-level cues in the temporal or spatial dimension, e.g. Chen et al. [44] modeled spatio-temporal and channel correlations, Xia et al. [45] learned salient facial features at the spatial

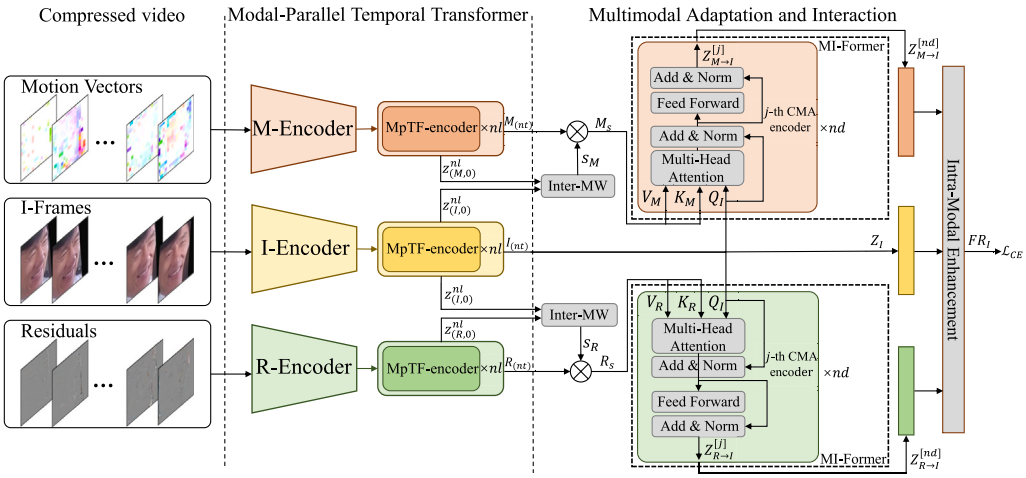


Fig. 3. Overview of our proposed method. Modal-parallel temporal transformer (MpTF) encoder models temporal information of three modalities parallelly; inter-modal weighting (Inter-MW) re-weighs the source modalities to distinguish their different contributions on expression recognition; multimodal interaction transformer (MI-Former) based on cross-modal attention (CMA) encoder learns interactions between modalities; and intra-modal enhancement (Intra-ME) adaptively enhances the target modality for its final representation. ⊗ denotes scalar multiplication. ∗ *nl* denotes that a module is executed for *nl* times. *nl* is also the number of MpTF layers, *nd* is the number of MI-Former layers, M_s/ R_s are the weighted source modality features, and the Z-series variables are the cross-modal attention outputs.

and temporal levels, and Xia and Jiang [46] proposed a transformer-based hierarchical architecture. Since these works are developed for the original video, the inherent multiscale features of compressed video are rarely explored. To this end, we make use of the inherent multiscale characteristic of multiple modalities in the compressed domain of expression video, and design a multimodal interaction attention paradigm to encode these multimodal cues for interaction learning, in addition to the temporal representation of each modality.

2.3. Compressed-domain video understanding

With the increasing demand for efficient video storage and transmission, compressed-domain video analysis has become an emerging research direction in computer vision. Unlike conventional approaches that rely on fully decoded RGB frames, compressed-domain methods operate directly on intermediate signals, i.e., I-frames, motion vectors, and residuals, which are generated during video encoding. These signals inherently retain spatial appearance, motion dynamics, and fine-grained residual information, thereby preserving key semantic cues while significantly reducing storage usage.

Early studies mainly focused on the task of action recognition. Zhang et al. [47] replaced optical flow with motion vectors to accelerate computation, highlighting the efficiency potential of the compressed domain. Wu et al. [14] further demonstrated the feasibility of using motion vectors and residuals jointly for video understanding, and Huang et al. [48] introduced a teacher–student framework that fused these cues to approximate optical-flow supervision. Chen and Ho et al. [31] explored multimodal fusion of I-frames and residuals via a transformer framework, showing improved representation capability. However, these methods were primarily designed for coarse-grained motion recognition and did not explicitly investigate modality complementarity or adaptive weighting among compressed-domain signals.

For the task of FER, the study within the compressed domain remains at an early stage. Liu et al. [6] introduced the MIC model, which verified the feasibility of applying compressed features to dynamic facial expression recognition but was limited to single-modality (residual-based) modeling, lacking explicit cross-modal interaction. More recently, Li et al. [49] employed cross-attention and adaptive gradient modulation to balance multimodal contributions in compressed-domain action recognition. However, their framework primarily addresses large-scale motion dynamics rather than subtle facial deformations that require fine-grained inter- and intra-modal coordination.

To this end, our work extends compressed-domain learning to DFER, which explores subtle emotional information depending on the joint modeling of I-frames, motion vectors, and residuals. Specifically, we design the Modal-Parallel Temporal Transformer (MpTF) and the multimodal Adaptation and Interaction (MAaI), to explicitly capture both inter-modal and intra-modal dependencies, thereby enabling efficient, fine-grained, and semantically consistent emotion representation learning. Unlike previous I/M/R-based compressed-video fusion schemes, our method designs a DFER-oriented asymmetric interaction strategy, where I-frames serve as the semantic target modality and motion vectors/residuals provide complementary source cues.

3. Proposed method

Overview: As illustrated in Fig. 3, our framework is carried out directly on the compressed domain, and comprises three branches that take three modalities of the compressed video, i.e., I-frame, motion vectors, and residuals as input, where I-frames represent complete image information, motion vectors contain motion information between consecutive video frames, and residuals capture the difference between the original frame and the frame predicted based on the reference frame and motion vectors. With three parallel branches, we introduce the modal-parallel temporal transformer (MpTF) to model the temporal characteristics of each modality based on a frame-level attention operator. Furthermore, based on the multimodal representation of compressed videos, we propose a multimodal interaction attention aimed at learning inter-modal, multimodal interaction and intra-modal cues for a more robust expression representation.

3.1. Modeling compressed video representations

Considering that there is redundant information in the video, algorithms for video compression such as MPFG4, H.264 or HEVC can be adopted to distill the key information [50]. They use the contents of a few frames from the video (termed as reference frames) to compress other frames by only storing their differences. In the compression domain, there are two typical types of frames, i.e., I-frames (intra-frame coded frames), which are regular RGB images, and P-frames (predicted frames) [14], which are predicted based on a reference frame (e.g., I-frame), motion vectors and residuals. In our framework, P-frames are introduced only to illustrate the physical principle and feature decomposition of compressed videos, rather than serving as explicit model

inputs. The subsequent multimodal modeling is conducted directly on the three component modalities, i.e., I-frames, motion vectors, and residuals.

Specifically, motion vectors represent the movement of macroblocks (a frame is usually divided into multiple macroblocks, e.g. 16 × 16 in the video compression process) between sequential frames, and the residuals preserve the pixel difference between the P-frame and the original frame. Formally, we use $M_{(t)}$ to represent the motion vector from the original frame to the predicted frame at time t, and $R_{(t)}$ to represent the residual between the original frame and the predicted frame, after that the motion vector $M_{(t)}$ has compensated the predicted frame. Then P-frames at pixel pi are reconstructed based on the recurrence relation [14] as:

$$I^{pi}_{(t)} = I^{pi-M^{pi}_{(t)}}_{(t-1)} + R^{pi}_{(t)} \tag{1}$$

where $I_{(t)}$ denotes the RGB image at time t. The motion vectors and residuals are then passed through discrete cosine transform, and entropy-encoded [14].

Consequently, the I-frames, motion vectors, and residuals in a compressed video implicitly provide a multimodal representation, and each modality can provide unique information. Therefore, our method uses the representations of these three modalities to learn the temporal and modal interaction cues of compressed expression videos. Similar conclusions have also been reported in previous works on compressed video understanding [14,31,51], which demonstrate that I-frames, motion vectors, and residuals convey complementary semantic, motion, and detail information, and their joint modeling yields more discriminative representations.

3.2. Modal-parallel temporal transformer (MpTF)

Based on the reconstructed I-frames modality in Eq. (1), the modalities of motion vectors and residuals, we introduce the MpTF to optimize the temporal transformer of them in a parallel fashion. That is, instead of optimizing the network parameters with respect to (w.r.t.) only a single modality, e.g. the I-frames in current works for temporal video representation, we propose to optimize the temporal representations of the three modalities simultaneously with three transformers.

Specifically, our method uses nt sampled I-frames I, motion vectors M and residuals R from each of n compressed video sequences, i.e. $\{cv_k\}^n_{k=1}$ as input, where $I, M, R \in \mathbb{R}^{nt\times 3\times h\times w}$, h and w represent the height and width of the images, respectively. Specifically, for the training, the original sequence is equally divided into nu segments, and then nv frames are randomly selected from each segment. For the testing, we first divide all frames into nu segments and then select nv frames in the middle of each segment. Thus, the length of the sampled segments is $nt = nu \cdot nv$ for both the training and test sets. Then we extract the cues of three modalities for each frame using the encoder (e.g. pre-trained ResNet18 [52]) to obtain the features, i.e., $\mathbf{x}_i \in \mathbb{R}^{nt\times d}$, $i \in \{I, M, R\}$ with the same dimension.

The MpTF consists of nl temporal transformer encoders, which takes the frame feature $\mathbf{x}_i$ as input, aiming to model the temporal information of three modalities parallelly. The detailed framework is shown in Fig. 3. MpTF-encoder is responsible for extracting temporal features of the three modalities; Inter-MW adaptively adjusts the modality weights; MI-Former models inter-modal interactions; Intra-ME performs adaptive enhancement for the target modality. In each transformer encoder, the time position code $\mathbf{e}_{(i,t)} \in \mathbb{R}^d$ is embedded to preserve the position information as:

$$\mathbf{z}^0_{(i,t)} = \mathbf{x}_{(i,t)} + \mathbf{e}_{(i,t)} \tag{2}$$

where $\mathbf{z}^0_{(i,t)}, i \in \{I, M, R\}, t \in \{1, 2, \dots, nt\}$ denotes the input embedding representation of the transformer encoder, where each transformer encoder further consists of nl encoders. In the $l \in \{1, 2, \dots, nl\}$-th encoder, the query ($\mathbf{q}$), key ($\mathbf{k}$) and value ($\mathbf{v}$) matrices are specifically computed for each frame based on the representation $\mathbf{z}^{l-1}_{(i,t)}$ produced from preceding blocks. This process can be formulated as:

$$\begin{aligned}
\mathbf{q}^l_{(i,t)} &= W^l_{(i,q)} LN\left(\mathbf{z}^{l-1}_{(i,t)}\right) \in \mathbb{R}^d \\
\mathbf{k}^l_{(i,t)} &= W^l_{(i,k)} LN\left(\mathbf{z}^{l-1}_{(i,t)}\right) \in \mathbb{R}^d \\
\mathbf{v}^l_{(i,t)} &= W^l_{(i,v)} LN\left(\mathbf{z}^{l-1}_{(i,t)}\right) \in \mathbb{R}^d
\end{aligned} \tag{3}$$

where $W^l_{(i,q)}$, $W^l_{(i,k)}$, $W^l_{(i,v)}$ are projection matrices used in the attention operation [19], $LN(\cdot)$ represents layer normalization [53]. Then, the self-attention weights for query matrix $\mathbf{q}^l_{(i,t)}$ are given by:

$$\alpha^l_{(i,t)} = SM\left(\frac{\left(\mathbf{q}^l_{(i,t)}\right)^T}{\sqrt{d}} \cdot [\mathbf{k}^l_{(i,1)}, \dots, \mathbf{k}^l_{(i,nt)}]\right) \tag{4}$$

where $SM(\cdot)$ denotes the Softmax activation function, $[\cdot]$ denotes the concatenation operator. The output $\mathbf{z}^l_{(i,t)}$ at the l-th layer is formulated as:

$$\begin{aligned}
\mathbf{z}^l_{(i,t)} &= MLP(LN(\mathbf{z'}^l_{(i,t)})) + \mathbf{z'}^l_{(i,t)} \\
\mathbf{z'}^l_{(i,t)} &= W_i \mathbf{s}^l_{(i,t)} + \mathbf{z}^{l-1}_{(i,t)} \\
\mathbf{s}^l_{(i,t)} &= \alpha^l_{(i,t)} [\mathbf{v}^l_{(i,1)}, \dots, \mathbf{v}^l_{(i,nt)}]
\end{aligned} \tag{5}$$

Finally, we concatenate different frame features $\{\mathbf{z}^{nl}_{(i,t)} \in \mathbb{R}^d, t \in \{1, \dots, nt\}\}$ to obtain the modality feature representations with temporal information, i.e., $I_{(nt)}$, $M_{(nt)}$, $R_{(nt)} \in \mathbb{R}^{nt\times d}$, e.g., $I_{(nt)} = [\mathbf{z}^{nl}_{(I,1)}, \dots, \mathbf{z}^{nl}_{(I,nt)}]^T$. With the above parallel feature representations for the three modalities, we can take into account the characteristics of each modality during the temporal modeling of them, based on the inherent multimodality of a compressed video.

3.3. Multimodal adaptation and interaction (MAaI)

In addition to the temporal dynamics encoding of three modalities with MpTF, we further propose a multimodal adaptation and interaction (MAaI) module to learn complementary cues among different modalities. As shown in Fig. 3, it consists of the modules of inter-modal weighting, multimodal interaction, and intra-modal enhancement.

Considering the fact that the information contained in different modalities in the compressed video domain is quite unbalanced, i.e., modality I contains more useful information than modalities M and R [14], we classify the three modalities into source modality M, R and target modality I, and utilize the source modality to interact with the target modality.

Inter-Modal Weighting (Inter-MW). In the compressed domain, let the modalities M and R represent non-original RGB image that may contain much noise, while modality I contains much more useful information than M and R [14]. Therefore, we use only the dynamic global features of the target modality I as the anchor to obtain importance weights w.r.t. the source modalities of M and R, based on the similarity computation. Here, the term 'anchor' indicates that the target modality I serves as a semantic reference and weighting baseline to guide the adaptive fusion of the auxiliary modalities M and R.

Specifically, we first embed a global token (similar to the class token [17]) $\mathbf{x}_{(i,0)} \in \mathbb{R}^d$ in the MpTF to obtain the global feature representation for each modality, i.e., $\mathbf{z}^{nl}_{(I,0)}$, $\mathbf{z}^{nl}_{(M,0)}$, $\mathbf{z}^{nl}_{(R,0)} \in \mathbb{R}^d$ via Eqs. (2)∼(5). Then, we compute the cosine similarities between modalities I and M or R, i.e., s_M or s_R as follows:

$$s_M = \cos\left(\mathbf{z}^{nl}_{(I,0)}, \mathbf{z}^{nl}_{(M,0)}\right) = \frac{\left(\mathbf{z}^{nl}_{(I,0)}\right)^T \mathbf{z}^{nl}_{(M,0)}}{\left\|\mathbf{z}^{nl}_{(I,0)}\right\| \cdot \left\|\mathbf{z}^{nl}_{(M,0)}\right\|}, \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm. We then normalize s_M and s_R by a linear normalization, i.e. $s_M \leftarrow (s_M + 1)/2$, to use them as the importance of the corresponding modality. Based on this, we obtain the final feature representations of the source modalities as

$M_s = M_{(nt)} \cdot s_M \in \mathbb{R}^{nt\times d}$, $R_s = R_{(nt)} \cdot s_R \in \mathbb{R}^{nt\times d}$, and that of the target modality as $I_s = I_{(nt)} \in \mathbb{R}^{nt\times d}$.

Multimodal Interaction.

Each modality provides distinct and complementary expression cues, while simple concatenation ignores their representational inconsistencies [14]. Therefore, on the basis of the feature representation via Inter-MW, we further propose the multimodal interaction transformer (MI-Former), to explore the potential adaptation process from the source modality to the target one, in order to fully utilize the inconsistent complementary information of the modalities in the compressed domain.

Specifically, to encourage interactions between different modalities, we use cross-modal attention (CMA) to help the target modality learn cues from the source modalities, where a CMA in the case that I_s and M_s interact is defined as:

$$\mathrm{CMA}_{M\to I}\left(I_s, M_s\right) = SM\left(\frac{Q_I K_M^T}{\sqrt{d_k}}\right) V_M \in \mathbb{R}^{nt\times d_v} \tag{7}$$

where the query $Q_I = I_s W_{Q_I}$ with $W_{Q_I} \in \mathbb{R}^{d\times d_k}$, the key $K_M = M_s W_{K_M}$ with $W_{K_M} \in \mathbb{R}^{d\times d_k}$, and the value $V_M = M_s W_{V_M}$ with $W_{V_M} \in \mathbb{R}^{d\times d_v}$. $SM(\cdot)$ is conducted row by row. For the interaction between I_s and R_s, $\mathrm{CMA}_{R\to I}\left(I_s, R_s\right)$ can be similarly obtained. Based on multiple stages of CMA, we then design a multimodal interaction transformer (MI-Former) to enable the modality I to learn information from the modalities of M and R. Compared with common feature interaction methods, our interaction strategy is modality-asymmetric, and the cross-attention strategy provides more targeted and adaptive fusion. It enables the I-frame features to selectively focus on useful cues from motion and residual modalities, dynamically aligns complementary information across modalities, and highlights salient expression regions through adaptive weighting, leading to more robust and discriminative representations.

First, to enable the feature representation to carry temporal information, we add positional embedding (PE) enhancement [36] to the three modalities I, M and R. For instance:

$$\begin{aligned} Z_I^{[0]} &= I_s + \mathrm{PE}_I(nt, d) \\ Z_M^{[0]} &= M_s + \mathrm{PE}_M(nt, d) \end{aligned} \tag{8}$$

where $\mathrm{PE}_I(nt,d), \mathrm{PE}_M(nt,d) \in \mathbb{R}^{nt\times d}$ represent the embedding of the corresponding position index [36].

Formally, the proposed MI-Former consists of nd CMA encoders, which are progressively computed as:

$$\begin{aligned} Z_{M\to I}^{[0]} &= Z_I^{[0]} \\ \hat{Z}_{M\to I}^{[j]} &= \mathrm{CMA}_{M\to I}^{[j]}\left(LN\left(Z_{M\to I}^{[j-1]}\right), LN\left(Z_M^{[0]}\right)\right) \\ &\quad + LN\left(Z_{M\to I}^{[j-1]}\right) \\ Z_{M\to I}^{[j]} &= f_{\theta_{M\to I}^{[j]}}\left(LN\left(\hat{Z}_{M\to I}^{[j]}\right)\right) + LN\left(\hat{Z}_{M\to I}^{[j]}\right) \end{aligned} \tag{9}$$

where $f_{\theta_{M\to I}^{[j]}}$ is the jth feed-forward sublayer parameterized by $\theta_{M\to I}^{[j]}$ $(1 \le j \le nd)$. In this way, the target modality I can continuously receive information from the source modality M through a MI-Former, which enables network to exchange cues between modalities and learn their correlation.

As a result, we obtain the representations $Z_{M\to I}^{[nd]} \in \mathbb{R}^{nt\times d}$ corresponding to the target modality I after interaction with the source modality M. Similarly, we get the representation $Z_{R\to I}^{[nd]} \in \mathbb{R}^{nt\times d}$ of I after the interaction with another source modality R, for further learning meaningful information about cross-modal association.

Intra-Modal Enhancement (Intra-ME). To enable the target modality to learn the information of the source modalities after inter-modal weighting and cross-modal interaction, we weigh each representation of the target modality via the proposed Intra-ME. First, based on the three representations of the target modality I, i.e., $Z_{M\to I}^{[nd]}, Z_I$ and $Z_{R\to I}^{[nd]} \in \mathbb{R}^{nt\times d}$ through the preceding two stages, we flatten them into the vector representations, i.e. $FZ_{M\to I}, FZ_I$ and $FZ_{R\to I} \in \mathbb{R}^{d_t\times 1}$, where $d_t = nt \cdot d$. Then, for each of the vector representation, we use a shared attention vector $sq \in \mathbb{R}^{d_t\times 1}$ to obtain the attention weights $\mu_{M\to I}, \mu_I$ and $\mu_{R\to I}$. Here, sq denotes a learnable global query vector that is randomly initialized and optimized during training to adaptively extract importance information from the three-modality representations. Although Intra-ME involves linear projection and fusion operations, it performs adaptive weighting guided by sq to emphasize semantically relevant modalities and suppress redundant information, thereby enhancing cross-modal feature representation. Specifically,

$$\mu_{M\to I} = sq^T \tanh\left(\boldsymbol{W}_{M\to I} FZ_{M\to I} + \boldsymbol{b}_{M\to I}\right) \tag{10}$$

where $W_{M\to I} \in \mathbb{R}^{d_t\times d_t}$, $b_{M\to I} \in \mathbb{R}^{d_t\times 1}$. Similar to Eq. (10), $\mu_I, \mu_{R\to I}$ can also be obtained, and these weights are further normalized using the Softmax function as:

$$[\psi_{M\to I}, \psi_I, \psi_{R\to I}] = SM([\mu_{M\to I}, \mu_I, \mu_{R\to I}]) \tag{11}$$

The final representation of modality I, i.e., $FR_I \in \mathbb{R}^{3d_t}$ is obtained by the concatenation of weighted summation as:

$$FR_I = \left[\psi_{M\to I} \otimes FZ_{M\to I}, \psi_I \otimes FZ_I, \psi_{R\to I} \otimes FZ_{R\to I}\right] \tag{12}$$

where $\otimes$ represents the scalar multiplication operator.

Finally, FR_I is used in the classification with the cross entropy loss:

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{\#class} 1_{y_k=c} \log\left(\hat{y}_{(k,c)}\right) \\ \hat{y} &= FC\left(FR_I\right) \in \mathbb{R}^{\#class} \end{aligned} \tag{13}$$

where FC denotes a fully connected layer, $\#class$ denotes the number of classes, n denotes the number of samples (or video sequences), $\hat{y}_{(k,c)}$ denotes the probability that the kth sample is predicted to be its ground-truth category, i.e. y_k. For clarity, the pseudo code of the proposed algorithm is presented in Algorithm 1, and the detailed flow of the module MAaI is shown in Algorithm 2.

Complexity analysis. We then analyze the time complexities of our proposed MpTF and MAaI. Firstly, the MpTF consists of nl single-head attention transformer encoders for the three modal features $\mathbf{x}_I$, $\mathbf{x}_M$ and $\mathbf{x}_R \in \mathbb{R}^{nt\times d}$. Therefore, the complexity of the MpTF module is $O\left(nl \cdot nt^2\right)$. For the MAaI module, it consists of an Inter-MW, a MI-Former, and an Intra-ME. The time complexity of the Inter-MW, mainly arises from the computation of the cosine similarity between $\mathbf{z}_{(I,0)}^{nl} \in \mathbb{R}^d$ and $\mathbf{z}_{(M,0)}^{nl}, \mathbf{z}_{(R,0)}^{nl} \in \mathbb{R}^d$, which is $O(d)$. In MI-Former, nd cross-modal attention encoders are utilized for multimodal interaction, hence the time complexity of this module is $O\left(nd \cdot nt^2 \cdot nh\right)$. The time complexity of the Intra-ME, mainly arises from the differentiated enhancement of three representations of the target modality I, i.e., $FZ_{M\to I}$, FZ_I and $FZ_{R\to I} \in \mathbb{R}^{d_t\times 1}$, which is $O(d_t^2)$. Thus, the complexity of the MAaI module is approximate to $O(d) + O\left(nd \cdot nt^2 \cdot nh\right) + O(d_t^2)$. In summary, the total complexity of our proposed method is $O\left(nl \cdot nt^2\right) + O(d) + O\left(nd \cdot nt^2 \cdot nh\right) + O(d_t^2)$.

4. Experiments

4.1. Datasets

Oulu-CASIA: The Oulu-CASIA [54] dataset contains six basic expressions of 80 subjects, including 'angry', 'disgust', 'fear', 'happy', 'sad' and 'surprise'. Each video can be divided into three parts: normal, weak and dark according to lighting conditions. Each part contains 480 sequences, and each expression sequence starts from the neutral stage and ends with the emotional peak. We employ the same 10-fold cross-validation strategy as current state of the arts.

AFEW: The AFEW [55] dataset is composed of video clips from films, including spontaneous expressions, lighting changes, various head poses and occlusions. It contains 1809 video clips with seven expressions: 'angry', 'disgust', 'fear', 'happy', 'sad', 'surprise', and 'neutral'.

Algorithm 1: The training of the proposed algorithm

Input: Compressed video training dataset $\{cv_k, y_k\}_{k=1}^{n}$, model parameters θ.
Output: Optimized model.
Initialize θ with a pre-trained model;
while *not converged* **do**
  Randomly sample nt frames in compressed video of facial expressions
  Get the three modalities I, M, R for each frame by Eq. (1)
  #Employ MpTF module
  Extract each modality feature $\mathbf{x}_I, \mathbf{x}_M, \mathbf{x}_R$
  Get the spatio-temporal features $I_{(nt)}$, $M_{(nt)}$, $R_{(nt)}$ parallelly by Eqs. (2)∼(5).
  #Employ Inter-MW module
  Get the global information $\mathbf{z}^{nl}_{(I,0)}$, $\mathbf{z}^{nl}_{(M,0)}$, $\mathbf{z}^{nl}_{(R,0)}$ of each modality based on global tokens $\{\mathbf{x}_{(i,0)}\}$
  Calculate the weights s_M and s_R of the modalities M and R by Eq. (6)
  Get the weighted features I_s, M_s, R_s.
  #Employ MI-Former module
  Get the modal interaction features $Z^{[nd]}_{M\to I}$, $Z^{[nd]}_{R\to I}$ by Eqs. (7)∼(9).
  #Employ Intra-ME module
  Obtain the final representation FR_I with fusion of three representations of the target modality by Eqs. (10)∼(12).
  #Perform forward calculation and backpropagation
  Get the prediction probabilities $\hat{y}$ of videos and calculate the cross-entropy loss $\mathcal{L}_{CE}$ by Eq. (13)
  Update θ with back propagation according to $\mathcal{L}_{CE}$.
end

Algorithm 2: The Cross-Modal Interaction in MAaI

Input: Temporal features of three modalities I, M, R from MpTF.
Output: Final fused representation FR_I of the target modality I.
Step 1: Inter-Modal Weighting (Inter-MW)
Obtain the global information $\mathbf{z}^{nl}_{(I,0)}$, $\mathbf{z}^{nl}_{(M,0)}$, and $\mathbf{z}^{nl}_{(R,0)}$ of each modality by aggregating global tokens $\mathbf{x}_{(i,0)}$ via Eqs. (2)∼(5).
Compute cosine similarities between I and M, R to generate importance weights s_M, s_R by Eq. (6).
Derive the weighted modality features: $I_s = I_{(nt)} \in \mathbb{R}^{nt\times d}$, $M_s = M_{(nt)} \cdot s_M \in \mathbb{R}^{nt\times d}$, and $R_s = R_{(nt)} \cdot s_R \in \mathbb{R}^{nt\times d}$.
Step 2: Multimodal Interaction (MI-Former)
Use $\mathbf{I}_s$ as the Query and $\mathbf{M}_s, \mathbf{R}_s$ as the Key/Value inputs in the cross-attention mechanism.
Compute cross-modal attention to obtain the complementary interaction outputs: $Z^{[nd]}_{M\to I}$ and $Z^{[nd]}_{R\to I}$, representing motion- and residual-enhanced features of the target modality by Eqs. (7)–(9).
Step 3: Intra-Modal Enhancement (Intra-ME)
Take the three target-side representations, $Z^{[nd]}_{M\to I}$, Z_I, and $Z^{[nd]}_{R\to I}$ (obtained from Step 2 and Step 1), and flatten them into vectors $FZ_{M\to I}$, FZ_I, and $FZ_{R\to I}$, respectively.
Compute attention weights $\mu_{M\to I}$ with the shared attention vector $\mathbf{s}_q$ by Eq. (10); obtain μ_I and $\mu_{R\to I}$ analogously.
Compute adaptive attention weights $\psi_{M\to I}$, ψ_I, and $\psi_{R\to I}$ via the normalization function defined in Eq. (11).
Fuse these weighted representations to form the final target representation FR_I by Eq. (12).

All video clips are divided into three parts: training (773 video clips), validation (383 video clips) and testing (653 video clips). Since its test set is not publicly available, we train our model on the training set and evaluate the performances on the validation set.

DFEW: The DFEW [56] dataset comes from more than 1500 movies around the world, covering various challenging interferences, and is a large-scale unconstrained facial expression database. It contains 12,059 single-label video clips and also includes 7 emotional labels, namely 'angry', 'disgust', 'fear', 'happy', 'sad', 'surprise' and 'neutral'. All samples were divided into 5 equal-sized and non-overlapping folds, i.e. (fd1∼fd5), one of which was used for testing and the rest for training. We used the 5-fold cross-validation setup provided by DFEW to ensure fair comparisons among different methods.

4.2. Implementation details

Data Pre-processing: In our experiments, we convert the original video into MPEG-4 encoded compressed video and set the Group Of Pictures (GOP) to 1. Then we select a sequence of $nt = 8$ frames from the compressed video and transform each frame into the three modalities, i.e. I-frames, motion vectors and residuals as the input of our model. In addition, we employ only simple data augmentation techniques, i.e., image flipping and random erasure to avoid over-fitting. To reduce the dependence on computational resources, we preprocessed all the original images according to facial landmarks to generate face images and resized the images to 112×112 before acquiring the modalities.

Training Setting: We perform our experiments with GeForce RTX 3090 GPUs to build the entire framework on PyTorch. For feature extraction of the three modalities, we use ResNet18 [52] pre-trained on the face recognition dataset, i.e. MS-Celeb-1M (MS1M) [57]. For the AFEW and DFEW datasets, we set the learning rate to 3e-5 and the batch size to 32. For the Oulu-CASIA dataset, we set the learning rate to 3e-4 and the batch size to 4. For all the models, we use the Adam optimizer with a weight decay of 0.0001 and the learning rate scheduler of ExponentialLR with the gamma value of 0.9. Training ends at epoch 120. In our approach, we set the numbers of encoders in the MpTF and MI-Former to $nl = 1$ and $nd = 3$, respectively.

In all experiments, we use weighted average recall (WAR) and unweighted average recall (UAR) as evaluation metrics. In the following experiments, we conduct the analysis and discussion mainly on DFEW.

4.3. Ablation studies

For the baseline model, the features of the three modalities are extracted and directly concatenated for classification. In this section, we perform ablation studies to evaluate the performance of each module, the importance of different modalities, the performance sensitivity against the numbers of encoders and frames, and the performance under different target modalities.

Ablation study of proposed modules: We first investigate the effectiveness of each module of our approach in Table 2. It shows that MpTF and MI-Former bring large improvements, e.g. using MpTF can increase UAR and WAR by 1.33% and 1.12%, respectively, and integrated MI-Former can further improve UAR by 0.65% and WAR by 0.59%. The effectiveness of MpTF maybe due to that it can take into account the characteristics of each modality, when providing temporal features for dynamic FER tasks. MI-Former achieves good performance since it enables inter-modal interactions and provides correlation information between modalities. In addition, the use of Inter-MW reduces the influence of modality-specific noise, bringing further improvements (UAR/0.17%, WAR/0.31%). Finally, using Intra-ME for differentially enhancing the three representations of the target modality, our method improves the baseline by 2.65% in terms of UAR and 2.26% in terms of WAR.

Evaluation of the modal importance: To evaluate the importance of each modality, we conducted an ablation study of our model on DFEW based on different combinations of modalities, and the results are shown in Table 3. For individual modalities, I-frames obtained the best accuracy of 63.95%, indicating that this modality contains the

Table 2
Ablation study of our method on DFEW.

MpTF	MAaI			Params(MB)	UAR(%)	WAR(%)
	Inter-MW	MI-Former	Intra-ME			
×	×	×	×	33.61	53.44	65.79
✓	×	×	×	37.16	54.77	66.91
✓	✓	×	×	37.16	54.94	67.22
✓	✓	✓	×	37.57	55.59	67.81
✓	✓	✓	✓	38.60	**56.09**	**68.05**

Table 3
Comparison of modality importance on DFEW.

I	M	R	Params(MB)	UAR(%)	WAR(%)
✓	×	×	12.39	51.98	63.95
×	✓	×	12.38	34.39	44.23
×	×	✓	12.39	43.11	53.59
×	✓	✓	25.99	48.45	60.09
✓	×	✓	25.99	54.42	66.91
✓	✓	×	25.99	54.88	67.39
✓	✓	✓	38.60	**56.09**	**68.05**

Table 4
Performance sensitivity against different numbers of encoders, i.e. the hyper-parameters of *nl* and *nd* on DFEW.

nl	*nd*	Params(MB)	UAR(%)	WAR(%)
1	1	38.08	54.56	67.33
3	1	45.61	54.58	66.92
1[a]	3[a]	38.60	**56.09**	**68.05**
3	3	45.67	55.03	67.39
1	6	38.83	55.18	67.48

[a] Denotes the default setting used in our method.

Table 5
Performances of different target modality settings on DFEW.

Target modality	Params(MB)	UAR(%)	WAR(%)
I[a]	38.60	**56.09**	**68.05**
M	38.60	52.06	63.96
R	38.60	54.19	66.17
I, M, R	38.58	55.51	67.51

[a] Denotes the default setting used in our method.

Table 6
End-to-end runtime breakdown for different data setups on DFEW.

Data mode	Full-frame decoding	Data loading / modality extraction	Model inference	Total time
Raw RGB video	1.64 s	0.45 s	2.20 s	4.29 s
Compressed video	N/A	1.10 s	2.55 s	**3.65 s**

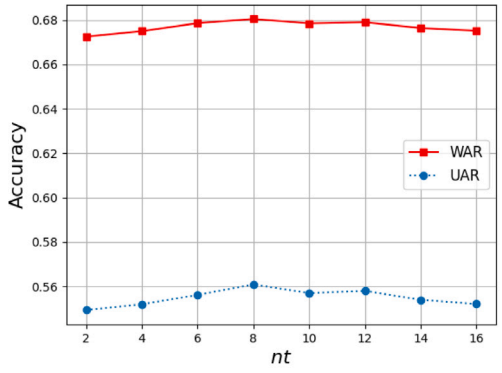


Fig. 4. Accuracy of different numbers of frames (*nt*) on DFEW. Our method uses the setting of $nt = 8$.

richest expression cues, that is why we set I-frame as the target modality and the other two modalities as the source modalities. Although individual motion vectors and residuals achieve less competitive performances, they can bring further improvement over I-frames, i.e., they helped I-frames improve by 2.90%/2.44% and 3.44%/2.96% in terms of UAR and WAR, respectively.

Evaluation of the number of encoders: In our approach, we set the numbers of encoders, i.e. *nl* for MpTF and *nd* for MI-Former as 1 and 3, respectively. In this section, we study the performance sensitivity against these parameters, and present the results in Table 4. Table 4 shows that a shallow model, corresponding to small n_l and n_d, performs poorly due to its limited parameters. From another aspect, a deep network with large *nl* and *nd* may lead to overfitting, since the patterns of the motion vectors and residuals are much simpler than those of the I-frames in the compression domain. Therefore, increasing the number of model encoders maybe not be always helpful to the performance.

Evaluation of frame number: In the proposed method, a small number of frames (*nt*) is employed to obtain the cues associated with the three modalities, we further study the performance when a larger *nt* is used, i.e., an ablation study of this parameter is conducted in Fig. 4. It shows that our method obtains the best results in terms of WAR and UAR at $nt = 8$, which is only 0.4% higher than the average performance of $nt = 2 \sim 16$ (except $nt = 8$), indicating that our method is robust to the variation of this parameter. In addition, a too large *nt* may impair the generalization of the model due to possible invalid frames, while a smaller *nt* is more helpful to the robustness of our method on capturing the temporal cues of frame-level modal features, and it can also reduce the computational overhead.

Evaluation of different target modalities: Considering that the I-frames (I) often contain the richest cues in the compressed domain (a detailed description is introduced in Section 4.5), we use them as our target modality, while the motion vectors (M) and residuals (R) as the source modalities for interaction with them in the representation learning. To study the performance when the target modality is set to M, R or all the three modalities (I, M, R) are used as target modalities (when one modality is used as target one, the other two modalities are used as source ones), we perform an ablation study on the setting of the target modality in Table 5, where the performances are compared with that of the default setting, i.e. the target modality is set as I.

Table 5 shows that the performance is degraded when either M or R is used as the target modality. Specifically, when M is used as the target modality, the performance is decreased by 4.03% and 4.09% in terms of UAR and WAR, respectively. Especially, when using all the three modalities (I, M, R) as the target modalities, the performances of UAR and WAR are reduced by margins of 0.58% and 0.54%, compared with the default setting of using I as the target modality. This performance decline maybe due to the fact that the modalities of motion vectors and residuals also introduce noise or redundant information to the I-frame, besides of the complementary cues.

Comparison of computational complexity: Our method operates directly in the compressed domain of videos. To validate its efficiency, we compared it with an approach that decompresses the original video first. To provide a detailed efficiency analysis, we further report an end-to-end runtime breakdown under two data setups, including full-frame decoding, data loading/modality extraction, actual model inference, and total processing time. All measurements were conducted on the DFEW dataset using a workstation equipped with an NVIDIA GeForce RTX 3090 GPU.

As shown in Table 6, the raw RGB video mode requires 1.64 s for full-frame decoding, 0.45 s for data loading, and 2.20 s for model inference, resulting in a total processing time of 4.29 s per video. In contrast, the compressed-video mode does not require full-frame RGB decoding, and requires 1.10 s for data loading and compressed-domain modality extraction, including I-frames, motion vectors, and residuals. Although its model inference time is slightly higher due

Table 7
Robustness under different compression settings on DFEW.

Compression factor	Setting	UAR(%)	WAR(%)
Codec	H.264[a]	56.09	68.05
Codec	H.265/HEVC	55.95	67.92
Codec	AV1	55.82	67.79
GOP size	1[a]	56.09	68.05
GOP size	3	55.78	67.68
GOP size	5	55.45	67.31
QP	25	56.00	68.00
QP	30[a]	56.09	68.05
QP	40	55.97	67.93

[a] Denotes the default setting used in our method.

to the three-branch multimodal design, the total processing time is reduced from 4.29 s to 3.65 s per video. These results indicate that the proposed compressed-domain framework reduces the end-to-end runtime by avoiding full-frame RGB decoding while preserving the advantages of multimodal compressed-domain representation.

Robustness under different compression settings: Since the proposed method directly relies on compressed-domain signals, we further evaluate its robustness under different compression configurations on DFEW, including video codecs, Group of Pictures (GOP) sizes, and quantization parameters (QP). Unless otherwise specified, the default setting is H.264 with GOP = 1 and QP = 30.

As shown in Table 7, the proposed method remains stable under different codec settings. When replacing the default H.264 codec with H.265/HEVC and AV1, the largest performance drops are only 0.27 percentage points in UAR and 0.26 percentage points in WAR. When increasing the GOP size from 1 to 5, UAR and WAR decrease by 0.64 and 0.74 percentage points, respectively. This indicates that the method can tolerate different temporal coding structures, although a larger GOP size slightly changes the distribution of motion vectors and residuals. In addition, varying QP from 25 to 40 leads to only minor performance fluctuations, with the maximum drop being 0.12 percentage points in both UAR and WAR compared with the default QP = 30. These results demonstrate that the proposed compressed-domain framework is not overly sensitive to variations in codec type, GOP size, and quantization parameter (QP), and can maintain stable performance under realistic compression settings.

Comparison with key-frame-based methods: To further validate the effectiveness of dynamic temporal modeling, we conducted an additional comparison between our dynamic sequence-based framework and static key-frame recognition approaches on the Oulu-CASIA dataset. Specifically, the last three frames of each sequence were selected as key frames to represent peak expressions, following the standard protocol used in static facial expression recognition. As shown in Table 8, our dynamic facial expression recognition framework achieves a 0.57% improvement in WAR over the best-performing static key-frame-based method (DICE-FER [58]), while using the same backbone (ResNet18). This result confirms that static key-frame approaches tend to lose essential temporal variation cues, such as sequential eye-gaze and mouth-motion changes, which are crucial for distinguishing visually similar expressions. By contrast, dynamic modeling effectively captures these temporal dependencies, yielding more robust and discriminative representations.

Evaluation of inter-modal weighting sensitivity: To further analyze the sensitivity of inter-modal weighting, we conducted additional experiments on the DFEW dataset to evaluate the contribution of each modality weighting factor. As shown in Table 9, the motion vector weighting S_M improves UAR by 0.15% and WAR by 0.27% by comparing the 2nd and 3rd rows, while the residual weighting S_R improves UAR by 0.13% and WAR by 0.13% by comparing the 2nd and 4th rows. These results indicate that both weighting factors effectively suppress modality-specific noise and enhance overall feature representation, demonstrating the robustness of the inter-modal weighting mechanism.

Table 8
Comparison of static recognition methods and ours on the Oulu-CASIA. Bold denotes the best. Underline denotes the second best.

Methods	Training Data	Feature	WAR(%)
ATT [59]' 2024	The last three frames	Static	89.03
HNCL [60]' 2025	The last three frames	Static	87.31
DICE-FER [58]' 2025	The last three frames	Static	91.10
Ours	The last three frames	Static	<u>91.23</u>
Ours	Sequence	Dynamic	**91.67**

Table 9
Comparison of inter-modal weight effects on the DFEW.

S_M	S_R	UAR(%)	WAR(%)
–	–	55.84	67.65
✓	–	55.99	67.92
–	✓	55.97	67.78
✓	✓	56.09	68.05

Table 10
Comparison of SOTA methods and ours on Oulu-CASIA. Bold denotes the best. Underline denotes the second best.

Methods	Training Data	Feature	WAR(%)
STM-Exp [61]' 2014	All frames	Dynamic	74.59
DTAGN [62]' 2015	All frames	Dynamic	81.46
LOMo [63]' 2016	All frames	Dynamic	82.10
Traj. on S+(2, n) [64]' 2017	All frames	Dynamic	83.13
PHRNN-MSCNN [65]' 2017	All frames	Dynamic	86.25
SAANet [66]' 2020	All frames	Dynamic	88.33
MGLN [28]' 2020	All frames	Dynamic	90.40
CAN [67]' 2021	Sixteen frames	Dynamic	88.33
STANER [68]' 2022	All frames	Dynamic	89.52
STAFER [69]' 2023	All frames	Dynamic	88.34
CSTAN [70]' 2023	Sixteen frames	Dynamic	<u>90.67</u>
ESTLNet [71]' 2024	Eight frames	Dynamic	89.38
Baseline	Eight frames	Dynamic	88.75
Ours	Eight frames	Dynamic	**91.67**

4.4. Comparison with SOTA methods

Results on in-the-laboratory dataset. To study the performance of our algorithm on in-the-lab datasets, Table 10 reports the performances of different state of the arts on Oulu-CASIA, where only the metric WAR is used for the evaluation and comparison following the works [68,70]. It shows that our proposed method achieves WAR of 91.67% in recognizing six basic expressions, outperforming current SOTA methods. CSTAN [70] proposes to learn and integrate spatiotemporal emotional information during expression changing, while it only considers spatiotemporal characteristics and relies on the original video modality, without utilizing the multimodal information of the video. In addition, it uses 16 frames of a single video as input, which demands larger overhead of network training compared with ours that uses only 8 frames. Meanwhile, our method not only considers the temporal inconsistency among the employed three modalities, but also makes full use of the complementary information among these modalities, and achieves the best performance with 1% higher WAR than CSTAN.

Results on in-the-wild dataset. Compared with data collected in laboratory scenarios, the data in the wild conditions are more complex and challenging. In this section, we evaluate the performance of our model on in-the-wild datasets of DFEW and AFEW.

For the DFEW dataset, the experiments are performed under 5-fold cross-validation following the previous work [56], and the results are shown in Table 11. The results show that our method outperforms current SOTA method, i.e., DPCNet [16] by 1.07% in terms of UAR.

Table 11
Comparison of our method with SOTA methods on DFEW. Bold denotes the best. Underline denotes the second best. ‡ denotes that the results were computed based on the confusion matrix reported in the corresponding paper.

Methods	Feature	Backbone	FLOPs(G)	Param(MB)	UAR(%)	WAR(%)
C3D [26]' 2015	Dynamic	C3D	77.23	78	42.74	53.54
C3D,EC-STFL [56]' 2020			77.23	78	45.10	55.50
R3D18 [72]' 2018	Dynamic	3D ResNet18	40.86	33	42.79	53.22
R3D18,EC-STFL [56]' 2020			81.73	33	45.05	56.19
VGG11-LSTM [56]' 2020	Dynamic	VGG11	63.87	153	42.39	53.70
VGG11-LSTM,EC-STFL [56]' 2020			63.87	153	44.78	56.25
ResNet18-LSTM [56]' 2020	Dynamic	ResNet18	15.71	31	42.86	53.08
ResNet18-LSTM,EC-STFL [56]' 2020			15.71	31	43.60	54.72
Former-DFER [15]' 2021			9.11	146	53.69	65.70
ResNet50-LSTM [16]' 2022	Dynamic	ResNet50	35.67	37	50.85	62.26
DPCNet [16]' 2022			9.52	51	55.02‡	66.32
RS-DFER [73]' 2023	Dynamic	3D Resnet18	4.78	42.52	51.79	63.31
Transformer [17]' 2023	Dynamic	Transformer	N/A	34.37	50.39	63.85
EST [17]' 2023			N/A	42.78	53.94	65.85
LOGO-FORMER [18]' 2023	Dynamic	ResNet18	10.27	N/A	54.21	66.98
ESTLNet [71]' 2024			N/A	N/A	N/A	**68.78**
D2SP [74]' 2025	Dynamic	ResNet18+LSTM	N/A	N/A	51.40	64.72
MS-CRFNet [75]' 2025	Dynamic	ResNet18	N/A	N/A	54.00	66.81
ResNet18(baseline)	Dynamic	ResNet18	11.62	33.61	53.44	65.79
Ours			11.66	38.60	**56.09**	68.05

Table 12
Comparison of our method with SOTA methods on AFEW. Bold denotes the best. Underline denotes the second best.

Methods	Feature	UAR(%)	WAR(%)
C3D [26]' 2015	Dynamic	43.75	46.72
I3D-RGB [76]' 2017	Dynamic	41.86	45.41
R(2+1)D [72]' 2018	Dynamic	42.89	46.19
3D ResNet18 [77]' 2018	Dynamic	42.14	45.67
Multimodal-CNN [78]' 2018	Dynamic	N/A	51.44
Emotion-BEEU [79]' 2020	Dynamic	N/A	52.49
MIC [6]' 2021	Dynamic	N/A	53.18
Former-DFER [15]' 2021	Dynamic	47.42	50.92
DPCNet [16]' 2022	Dynamic	47.86	51.67
CEFNet [80]' 2022	Dynamic	N/A	53.98
RS-DFER [73]' 2023	Dynamic	48.88	49.76
EST [17]' 2023	Dynamic	49.57	54.26
ESTLNet [71]' 2024	Dynamic	N/A	53.79
baseline	Dynamic	47.58	52.36
Ours	Dynamic	**49.59**	**54.30**

In terms of WAR, our algorithm ranks the second, which is only inferior to current SOTA method ESTLNet [71] by 0.73%. It is probably because that the large-scale in-the-wild dataset DFEW contains more sequences with tiny expression variations, placing a large challenge to our compressed video-driven algorithm. It is noteworthy that, our approach substantially reduces data storage requirements, i.e., from 17G to 1.6G, achieving a 90.5% reduction compared to SOTA methods, validating the substantial merit of our method in real application.

For the AFEW dataset, our model is mainly compared with SOTA methods that are also trained based on videos. Table 12 shows that our method achieves the best results in terms of both UAR and WAR, and improves the baseline by 2.01% and 1.94%, respectively. Compared with MIC [6], which only uses single-modality information, i.e., residuals, in the compressed domain, our method achieves a WAR value that is 1.12% higher, indicating that the multiple modalities we employed can provide complementary cues for the feature representation.

Additionally, we show the performance of our algorithm for each category of expressions from DFEW, i.e. the confusion matrices in Fig. 5, it demonstrates that our method shows relatively consistent accuracies across the five folds of DFEW, and achieves the highest average accuracy of 90.14% for the 'happy' expression, while has difficulty in predicting the expressions of 'disgust' and 'fear'. This maybe due to the severe label imbalance of DFEW, in which 'disgust' and 'fear' account for only the proportions of 1.22% and 8.14%, that are significantly smaller than those of the other categories.

4.5. Visualization

Visualization of the results of the modal-parallel temporal transformer (MpTF). To shed light on the effect of MpTF, we visualize the modification brought by this module on target modality, i.e. the I-frame (since it contains the richest expression information), by studying the heatmaps generated with/without the MpTF, and the results are shown in Fig. 6.

For the setting of baseline (the second row), the model takes no account of the temporal cues and treats each frame as an independent frame. As a result, as shown in Fig. 6, heat maps of frames from the same expression sequence focus on different positions, resulting in predicting a 'angry' expression to be 'sad'. This result shows that MpTF provides the indispensable temporal information for dynamic FER in the compressed domain. In addition, Fig. 6 shows that the three modalities also show a large inconsistency in temporal representations.

Furthermore, to evaluate the performance of parallel optimization of three modalities, that is aware of the inconsistency of different modalities, we conduct a tiny ablation study in Table 13, where a non-modal-parallel setting ('w/o'), i.e., the three network branches of the three modalities share weights, is used for the comparison. Table 13 shows that the proposed modal-parallel setting can achieve improvements of 0.69% and 0.39% on DFEW in terms of UAR and WAR, respectively, with the sacrifice of slightly increased number of model parameters. These results show that the modal-parallel setting is beneficial to the following modal interaction learning, thanks to the exploration of the unique features of each modality.

Visualization of class attention maps (CAM). To explore the contributions of different modalities in the compressed domain, we visualized the representation of each modality and the corresponding CAM distribution [81] in Fig. 8.

As shown in the CAM specific to the I-frame modality (2nd row), our method pays more attention to the regions with the most discriminative cues (e.g., mouth, eyes, etc.). However, as shown in the red boxes, I-frames may focus only on the eye region of the 'happy' expression, in

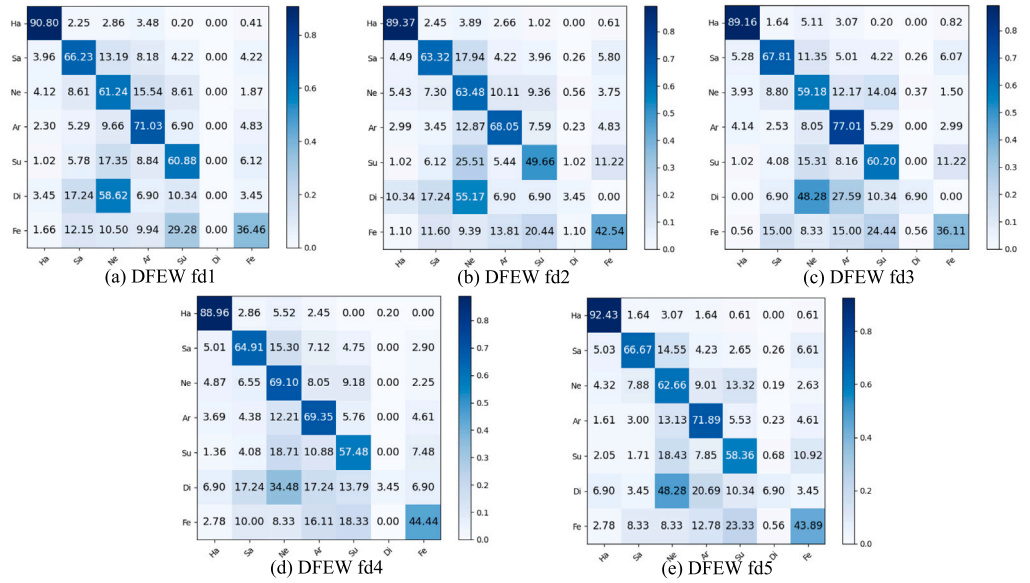


Fig. 5. The confusion matrices of our method on the fd1~fd5 of DFEW.

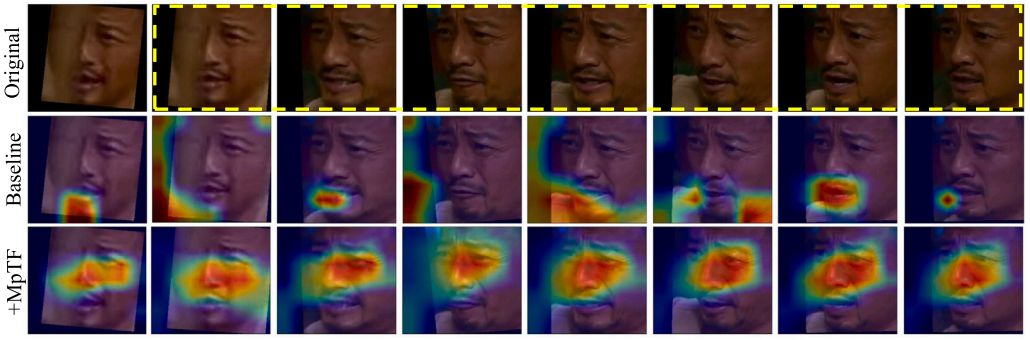


Fig. 6. Visualization of the effect of the modal-parallel temporal transformer (MpTF). The first row shows the original face images of 'angry' from DFEW; the second row shows the heatmaps generated by the baseline (predicted to be 'sad'); and the third row shows the heatmaps generated by the variant with the MpTF added on the baseline (predicted to be 'angry'). The yellow dotted box indicates the reconstructed frames (P-frames) in compressed videos.

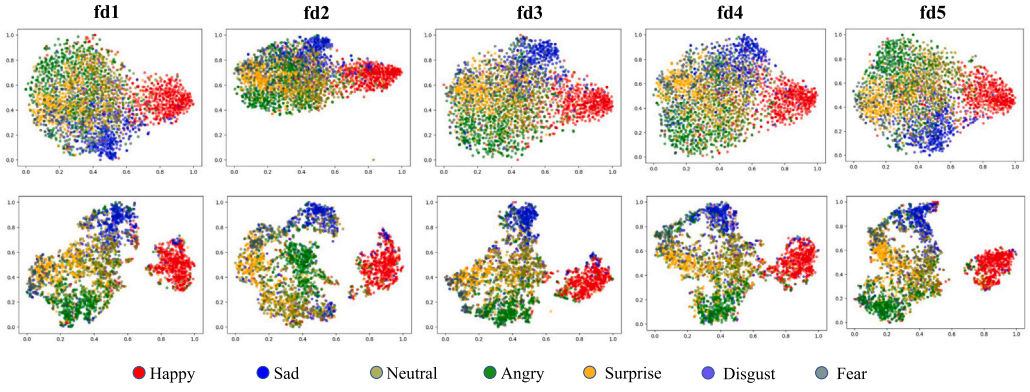


Fig. 7. Feature distributions learned by the baseline (top) and our method (bottom) on fd1~fd5 of DFEW.

this case, motion vectors and residuals help the model to transfer the attention on the cheek and mouth. That is, motion vectors and residuals can provide the movement of facial features and subtle expression changes, to compensate for the weaker temporal variation in I-frames, reflecting the usefulness of the proposed multimodal interactions in video expression representation.

Visualization of weight distributions specific to the target modality. We use box plots in Fig. 9 to further analyze the weights $\psi_{M\to I}$, ψ_I, $\psi_{R\to I}$ of the three representations $Z^{[nd]}_{M\to I}$, Z_I, $Z^{[nd]}_{R\to I}$ of the target modality obtained in Intra-ME (Eqs. (10)~(12)). As shown in Fig. 9, the weight of Z_I is much larger than those of the other two representations, which means that the information provided by I-frame is the most

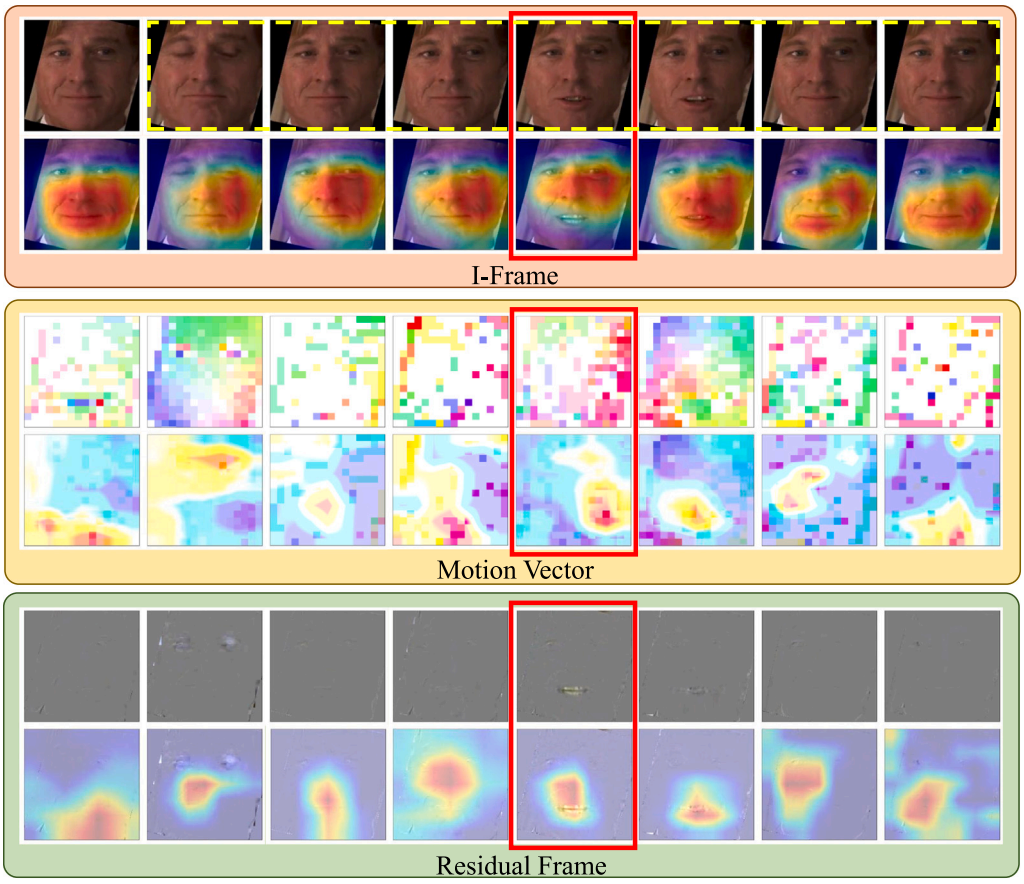


Fig. 8. Visualization of class attention maps (CAM) of three modalities. The modality cues in the compressed domain (odd rows) and the heatmaps (even rows) specific to the three modalities, i.e., I-frames (1st, 2nd rows), motion vectors (3rd, 4th rows), residuals (5th, 6th rows), are generated for a video with the 'happy' category from DFEW. The yellow dotted box indicates the reconstructed frames in compressed videos.

Table 13
Performance of the used modal-parallel setting ('w') in MpTF for the DFEW dataset. w/o indicates that the three branches of the modalities share network parameters.

Modal-Parallel	Params(MB)	UAR(%)	WAR(%)
w/o	36.29	55.40	67.66
w	38.60	**56.09**	**68.05**

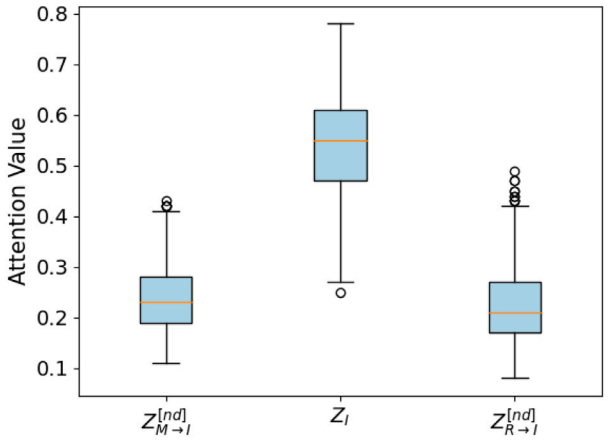


Fig. 9. Visualization of the weight ($\psi_{M\to I}$, ψ_I, $\psi_{R\to I}$) distribution specific to the target modality.

important. Meanwhile, since the motion vectors and residuals can also provide information, such as facial motion and complementary features, to the I-frame via multimodal interaction, the weights of $Z^{[nd]}_{M\to I}$ and $Z^{[nd]}_{R\to I}$ also account for certain proportions.

Visualization of failure cases. As shown in Fig. 10, the proposed method may still fail on expression sequences with very subtle visual changes. Such errors mainly occur when the facial motion is weak, the discriminative cues around the eyes and mouth are not sufficiently salient, or the illumination condition is poor. These challenging cases may produce weak motion-vector and residual responses, thereby limiting the complementary cues that MI-Former can introduce in the I-frame representation of our compressed-domain framework. In addition, degraded appearance cues under poor illumination may also weaken the semantic guidance of the I-frame target modality, making the final multimodal representation ambiguous.

Visualization of feature distribution. In this part, we use t-SNE [82] to visualize the distribution of dynamic expression features represented by the baseline and our method in Fig. 7. It shows that the features obtained by the baseline have significant overlap between categories. In contrast, our method achieves clearer boundaries between categories and more focused clustering centers, verifying that our method can learn more discriminative features for DFER.

4.6. Symbol definitions

To improve the readability and reproducibility of the proposed framework, we provide a summary of all mathematical symbols and their corresponding meanings in Table 14. This table lists the key variables used throughout Eqs. (2)~(13) and Algorithm 1, including

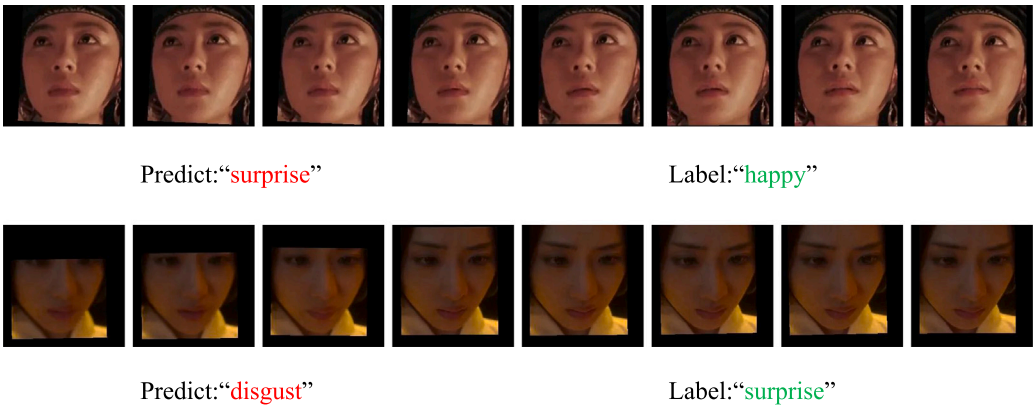


Fig. 10. Representative failure cases with subtle visual changes. The first row shows a 'happy' sequence misclassified as 'surprise', and the second row shows a 'surprise' sequence misclassified as 'disgust'.

Table 14
Definition and description of symbols.

Symbol	Meaning
I, M, R	Three modalities in compressed video: I-frame (full image), Motion Vector (inter-frame displacement), and Residual (pixel difference).
n_t	Number of sampled frames in each video clip.
$x^{(i,t)}$	Feature vector of the tth frame under modality $i \in \{I, M, R\}$, extracted by the encoder (ResNet18).
$e^{(i,t)}$	Temporal positional encoding for modality i, used to preserve frame order information.
$z^{(0)}_{(i,t)}$	Initial input embedding of modality i, obtained by adding feature and positional encoding.
$W^l_{(i,q)}, W^l_{(i,k)}, W^l_{(i,v)}$	Query, Key, and Value projection matrices for modality i in the lth Transformer layer.
$q^l_{(i,t)}, k^l_{(i,t)}, v^l_{(i,t)}$	Query, Key, Value vectors of modality i at frame t.
$\alpha^l_{(i,t)}$	Self-attention weight of modality i at frame t.
$s^l_{(i,t)}$	Aggregated feature based on attention weights.
$z^l_{(i,t)}$	Output feature of modality i from the lth Transformer encoder.
$I^{(n_t)}, M^{(n_t)}, R^{(n_t)}$	Frame-level feature sequences of three modalities after temporal modeling.
$z^{(I,0)}_{n_t}, z^{(M,0)}_{n_t}, z^{(R,0)}_{n_t}$	Global representations (tokens) of each modality.
s_M, s_R	Cosine similarity between I modality and M/R modalities, measuring inter-modal correlation.
M_s, R_s, I_s	Weighted features of target modality I and source modalities M/R.
Q_I, K_M, V_M, K_R, V_R	Query, Key, and Value matrices in the cross-modal attention.
$\mathrm{CMA}_{M\to I}, \mathrm{CMA}_{R\to I}$	Cross-modal attention outputs from source modalities M/R to target modality I.
$Z^{[0]}_I, Z^{[0]}_M, Z^{[0]}_R$	Initial feature representations of each modality after adding positional embeddings.
$\hat{Z}^{[j]}_{M\to I}, \hat{Z}^{[j]}_{R\to I}$	Representations of the jth cross-modal attention encoder layer.
$Z^{[n_d]}_{M\to I}, Z^{[n_d]}_{R\to I}$	Final representations of the I modality after interaction with M/R modalities.
$F_{Z_{M\to I}}, F_{Z_I}, F_{Z_{R\to I}}$	Flattened features of I and its two interacted representations.
s_q	Shared attention query vector in Intra-ME.
$\mu_{M\to I}, \mu_I, \mu_{R\to I}$	Raw attention scores of the three representations.
$\psi_{M\to I}, \psi_I, \psi_{R\to I}$	Normalized attention weights obtained via Softmax.
F_{RI}	Final fused representation of the target modality I, obtained via weighted concatenation.
$\hat{y}$	Predicted class probability distribution output.
$\mathcal{L}_{CE}$	Cross-entropy loss function.
n_l	Number of encoder layers in the Modal-Parallel Temporal Transformer (MpTF).
n_d	Number of encoder layers in the multimodal Interaction Transformer (MI-Former).
d	Dimensionality of modality features.
d_t	Flattened feature dimensionality, $d_t = n_t \times d$.

intermediate feature representations, attention weights, and fusion outputs.

5. Conclusion and discussion

To avoid the high storage requirement of raw videos on which most existing methods were developed, we proposed to recognize videos directly in the compressed domain to reduce the storage overhead. Furthermore, due to the availability of multiple modalities in compressed videos, including I-frames, motion vectors, and residuals, each of which can provide unique facial expression information, we propose a method that combines modal-parallel temporal modeling and multimodal interaction-based attention, to explore facial expression cues in the compressed video domain. In terms of utilizing the multimodal information in the compressed video, our algorithm mainly has two merits. First, the proposed modal-parallel temporal transformer for modeling the temporal information of all the modalities via a parallel

fashion, can well encode the specific characteristics of each modality, or each component of compressed video representation. Second, the multimodal adaptation and interaction module designed based on a multimodal interaction attention, can model the inter-modality, modality interaction and intra-modality cues for a complementary representation. Meanwhile, the inconsistency of the representation intensities of different modalities are well addressed. Our method achieves excellent performances on a laboratory collected dataset, i.e. Oulu-CASIA, as well as two in-the-wild datasets, i.e. DFEW and AFEW.

Although our method has shown effectiveness, there is still room for further improvement. First, our algorithm may fail to deal with the expression sequences with extremely small visual changes. Second, the proposed algorithm may be sensitive to the long-tail problem due to the uneven data distribution. We also note that motion vectors inherently encode temporal variations of body posture, providing a natural signal basis for extending toward posture-driven emotion recognition. Meanwhile, inspired by recent advances in posture- and motion-based emotion recognition, e.g. Zhai et al. [83], Wang et al. [84], and Wang et al. [85], suggesting that posture/gesture features can offer stable and discriminative information to resist facial ambiguity or occlusion, we will explore integrating body posture and gesture information within a compressed-domain framework to build a unified, multi-level emotion understanding model. Finally, although our method is practically meaningful in the compressed-domain setting in terms of reduced storage requirements and free of full-frame RGB decoding, a rigorous significance testing against all compared methods is deserved before its actual deployment.

CRediT authorship contribution statement

Weicheng Xie: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Junliang Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Haijian Liang:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Linlin Shen:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Zhihui Lai:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Siyang Song:** Writing – review & editing, Methodology, Formal analysis. **Zitong Yu:** Writing – review & editing, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive suggestions. The work was supported by the National Natural Science Foundation of China under grants no. 62276170, 62576216, 62576076, 62306061, the Science and Technology Project of Guangdong Province under grant no. 2023A1515010688, the Open Research Fund from Anhui Provincial Key Laboratory of Humanoid Robots under grant no. JH-RX-2025-0103, and the Guangdong Provincial Key Laboratory under grant no. 2023B1212060076. This work was also supported by the Intelligent Computing Center of Shenzhen University.

Data availability

Data will be made available on request.

References

[1] A.R. Shahid, H. Yan, SqueezExpNet: Dual-stage convolutional neural network for accurate facial expression recognition with attention mechanism, Knowl-Based Syst (KBS) 269 (2023) 110451.
[2] S. Ravindran, S. Rajagopalan, PNasFH-net: Pyramid neural architecture search forward network for facial emotion recognition in uncontrolled and pose variant environment, Knowl-Based Syst (KBS) 310 (2025) 112944.
[3] G.A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2016, pp. 510–526.
[4] J. Yang, C. Yang, F. Xiong, Y. Zhai, R. Wang, Learned video compression with adaptive temporal prior and decoded motion-aided quality enhancement, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 20 (8) (2024) 1–21.
[5] Z. Zheng, L. Yang, Y. Wang, M. Zhang, L. He, G. Huang, F. Li, Dynamic spatial focus for efficient compressed video action recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (2) (2023) 695–708.
[6] X. Liu, L. Jin, X. Han, J. You, Mutual information regularized identity-aware facial expression recognition in compressed video, Pattern Recognit. 119 (2021) 108105.
[7] A.-Q. Bi, X.-Y. Tian, S.-H. Wang, Y.-D. Zhang, Dynamic transfer exemplar based facial emotion recognition model toward online video, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 18 (2s) (2022) 1–17.
[8] D. Poux, B. Allaert, N. Ihaddadene, I.M. Bilasco, C. Djeraba, M. Bennamoun, Dynamic facial expression recognition under partial occlusion with optical flow reconstruction, IEEE Trans. Image. Process. 31 (2021) 446–457.
[9] L.A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, A. Nematzadeh, Decoupling the role of data, attention, and losses in multimodal transformers, Trans. Assoc. Comput. Linguist. 9 (2021) 570–585.
[10] D.H. Kim, W.J. Baddar, J. Jang, Y.M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Trans. Affect. Comput. 10 (2) (2017) 223–236.
[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Adv. Neural Inf. Proces. Syst. 28 (2015) 802–810.
[12] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.
[13] J. Yu, Z. Cai, Z. Liu, G. Xie, P. He, Facial expression spotting based on optical flow features, in: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM), 2022, pp. 7205–7209.
[14] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A.J. Smola, P. Krähenbühl, Compressed video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 6026–6035.
[15] Z. Zhao, Q. Liu, Former-dfer: Dynamic facial expression recognition transformer, in: Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), 2021, pp. 1553–1561.
[16] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, W. Zhang, Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos, in: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM), 2022, pp. 101–110.
[17] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, Y. Zhan, Expression snippet transformer for robust video-based facial expression recognition, Pattern Recognit. 138 (2023) 109368.
[18] F. Ma, B. Sun, S. Li, Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017) 5998–6008.
[20] Q. Jiang, Q. Wang, L. Chi, J. Liu, Cross-scale hierarchical spatio-temporal transformer for video enhancement, Knowl-Based Syst (KBS) 309 (2025) 112773.
[21] J. Liu, S. Guan, Q. Zou, H. Wu, P. Tiwari, Y. Ding, AMDGT: Attention aware multi-modal fusion using a dual graph transformer for drug–disease associations prediction, Knowl-Based Syst (KBS) 284 (2024) 111329.
[22] F. Guo, Y. Wang, H. Qi, W. Jin, L. Zhu, J. Sun, Multi-view distillation based on multi-modal fusion for few-shot action recognition (CLIP-mdmf), Knowl-Based Syst (KBS) 304 (2024) 112539.
[23] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI, 2013, pp. 509–516.
[24] X. Huang, Q. He, X. Hong, G. Zhao, M. Pietikainen, Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild, in: Proceedings of the 16th ACM on International Conference on Multimodal Interaction, ICMI, 2014, pp. 514–520.
[25] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2015, pp. 4489–4497.

[27] H. Zhang, B. Huang, G. Tian, Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture, Pattern Recognit. 131 (2020) 128–134.

[28] M. Yu, H. Zheng, Z. Peng, J. Dong, H. Du, Facial expression recognition based on a multi-task global-local network, Pattern Recognit. 131 (2020) 166–171.

[29] C. Gan, J. Xiao, Q. Zhu, Y. Zhu, Macro-expression-guided micro-expression recognition: A motion similarity perspective, Pattern Recognit. (2025) 112237.

[30] H. Liu, Q. Zhou, C. Zhang, J. Zhu, T. Liu, Z. Zhang, Y.-F. Li, Mmatrans: Muscle movement aware representation learning for facial expression recognition via transformers, IEEE Trans. Ind. Inform. 20 (12) (2024) 13753–13764.

[31] J. Chen, C.M. Ho, MM-vit: Multi-modal video transformer for compressed video action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022, pp. 1910–1921.

[32] S. Song, E. Sánchez-Lozano, M. Kumar Tellamekala, L. Shen, A. Johnston, M. Valstar, Dynamic facial models for video-based dimensional affect estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, CVPRW, 2019.

[33] M. Ma, J. Ren, L. Zhao, D. Testuggine, X. Peng, Are multimodal transformers robust to missing modality? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 18177–18186.

[34] H. Wang, A. Meghawat, L.-P. Morency, E.P. Xing, Select-additive learning: Improving generalization in multimodal sentiment analysis, in: 2017 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2017, pp. 949–954.

[35] J. Wang, X. Tan, Mutually beneficial transformer for multimodal data fusion, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7466–7479.

[36] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proc. Conf. Assoc. Comput. Linguist. Meet., vol. 2019, NIH Public Access, 2019, p. 6558.

[37] F. Lv, X. Chen, Y. Huang, L. Duan, G. Lin, Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 2554–2562.

[38] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, Y. Ding, Transformer-based multimodal information fusion for facial expression analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 2428–2437.

[39] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), 2016, pp. 1480–1489.

[40] H. Wang, S. Qu, Z. Qiao, X. Liu, Kcdnet: Multimodal object detection in modal information imbalance scenes, IEEE Trans. Instrum. Meas. 73 (2024) 1–13.

[41] K. Liu, D. Wang, D. Wu, Y. Liu, J. Feng, Speech emotion recognition via multi-level attention network, IEEE Signal Process. Lett. 29 (2022) 2278–2282.

[42] A. Yadav, D.K. Vishwakarma, A deep multi-level attentive network for multimodal sentiment analysis, ACM Trans. Multimed. Comput. Commun. Appl. 19 (1) (2023) 1–19.

[43] M.S. Akhtar, D.S. Chauhan, A. Ekbal, A deep multi-task contextual attention framework for multi-modal affect analysis, ACM Trans. Knowl. Discov. Data 14 (3) (2020) 1–27.

[44] W. Chen, D. Zhang, M. Li, D.-J. Lee, STCAM: Spatial-temporal and channel attention module for dynamic facial expression recognition, IEEE Trans. Affect. Comput. 14 (1) (2023) 800–810.

[45] X. Xia, L. Yang, X. Wei, H. Sahli, D. Jiang, A multi-scale multi-attention network for dynamic facial expression recognition, Multimedia Syst. (2022) 1–15.

[46] X. Xia, D. Jiang, Hit-MST: Dynamic facial expression recognition with hierarchical transformers and multi-scale spatiotemporal aggregation, Inf. Sci. (2023) 119301.

[47] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2718–2726.

[48] S. Huang, X. Lin, S. Karaman, S.-F. Chang, Flow-distilled ip two-stream networks for compressed video action recognition, 2019, arXiv preprint arXiv:1912.04462.

[49] X. Li, S. Li, M. Ma, Interactive and balanced multimodal learning via cross attention and gradient modulation for compressed video action recognition, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2025, pp. 1–5.

[50] D. Gall, MPEG: A video compression standard for multimedia applications, Commun. ACM 34 (1991) 305–313.

[51] B. Li, J. Chen, D. Zhang, X. Bao, D. Huang, Representation learning for compressed video action recognition via attentive cross-modal interaction with motion enhancement, in: L.D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, ijcai.org, 2022, pp. 1060–1066.

[52] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[53] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.

[54] G. Zhao, X. Huang, M. Taini, S.Z. Li, M. PietikäInen, Facial expression recognition from near-infrared videos, Image Vis. Comput. 29 (9) (2011) 607–619.

[55] A. Dhall, A. Kaur, R. Goecke, T. Gedeon, Emotiw 2018: Audio-video, student engagement and group-level affect prediction, in: Proceedings of the 20th ACM on International Conference on Multimodal Interaction, ICMI, 2018, pp. 653–656.

[56] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, J. Liu, Dfew: A large-scale database for recognizing dynamic facial expressions in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia (ACM MM), 2020, pp. 2881–2889.

[57] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2016, pp. 87–102.

[58] M. Aquib, N.K. Verma, M.J. Akhtar, Decoupling identity confounders for enhanced facial expression recognition: An information-theoretic approach, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 5552–5561.

[59] J. Chen, J. Shi, R. Xu, Dual subspace manifold learning based on GCN for intensity-invariant facial expression recognition, Pattern Recognit. 148 (2024) 110157.

[60] T. Liu, J. Li, J. Wu, B. Du, Y. Zhan, D. Tao, J. Wan, Facial expression recognition with heatmap neighbor contrastive learning, IEEE Trans. Multimed. 27 (2025) 4795–4807.

[61] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2014, pp. 1749–1756.

[62] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2015, pp. 2983–2991.

[63] K. Sikka, G. Sharma, M. Bartlett, Lomo: Latent ordinal model for facial analysis in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 5580–5589.

[64] A. Kacem, M. Daoudi, B. Ben Amor, J. Carlos Alvarez-Paiva, A novel space-time representation on the positive semidefinite cone for facial expression recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2017, pp. 3180–3189.

[65] 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.

[66] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, S. Wen, Saanet: Siamese action-units attention network for improving dynamic facial expression recognition, Neurocomputing 413 (2020) 145–157.

[67] X. Qu, Z. Zou, X. Su, P. Zhou, W. Wei, S. Wen, D. Wu, Attend to where and when: Cascaded attention network for facial expression recognition, IEEE Trans. Emerg. Top. Comput. Intell. 6 (3) (2021) 580–592.

[68] S. Li, X. Zheng, X. Zhang, X. Chen, W. Li, Facial expression recognition based on deep spatio-temporal attention network, in: Int. Conf. Collab. Comput.: Networking, Appl. Worksharing, Springer, 2022, pp. 516–532.

[69] L. Zhang, X. Zheng, X. Chen, X. Ren, C. Ji, Facial expression recognition based on spatial-temporal fusion with attention mechanism, Neural Process. Lett. 55 (5) (2023) 6109–6124.

[70] Y. Ye, Y. Pan, Y. Liang, J. Pan, A cascaded spatiotemporal attention network for dynamic facial expression recognition, Appl. Intell. 53 (5) (2023) 5402–5415.

[71] W. Gong, Y. Qian, W. Zhou, H. Leng, Enhanced spatial-temporal learning network for dynamic facial expression recognition, Biomed. Signal Process. Control. 88 (2024) 105316.

[72] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 6450–6459.

[73] Z. Liu, T. Wang, S. Zhou, M. Shu, Dynamic facial expression recognition in unconstrained real-world scenarios leveraging Dempster-Shafer evidence theory, in: International Conference on Artificial Neural Networks, ICANN, Springer, 2023, pp. 244–258.

[74] H. Wang, X. Mai, Z. Tao, X. Tong, J. Lin, Y. Wang, J. Yu, S. Yan, Z. Zhou, W. Zhang, D2SP: Dynamic dual-stage purification framework for dual noise mitigation in vision-based affective recognition., in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 19218–19229.

[75] M. Xia, H. Zheng, H. Peng, Z. Liu, J. Guo, et al., Enhanced multi-scale dynamic facial expression recognition via conditional random fields, Vis. Comput. (2025) 1–12.

[76] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6299–6308.

[77] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 6546–6555.
[78] C. Liu, T. Tang, K. Lv, M. Wang, Multi-feature based emotion recognition for video clips, in: Proceedings of the 20th ACM on International Conference on Multimodal Interaction, ICMI, 2018, pp. 630–634.
[79] V. Kumar, S. Rao, L. Yu, Noisy student training using body language dataset improves facial expression recognition, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2020, pp. 756–773.
[80] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, Z. Luo, Clip-aware expressive feature learning for video-based facial expression recognition, Inf. Sci. 598 (2022) 182–195.
[81] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2921–2929.
[82] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.
[83] Y. Zhai, G. Jia, Y.-K. Lai, J. Zhang, J. Yang, D. Tao, Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks, IEEE Trans. Affect. Comput. 15 (3) (2024) 1634–1648.
[84] T. Wang, S. Liu, F. He, W. Dai, M. Du, Y. Ke, D. Ming, Emotion recognition from full-body motion using multiscale spatio-temporal network, IEEE Trans. Affect. Comput. 15 (3) (2024) 898–912.
[85] T. Wang, S. Liu, F. He, M. Du, W. Dai, Y. Ke, D. Ming, Affective body expression recognition framework based on temporal and spatial fusion features, Knowl.-Based Syst. 308 (2025) 112744.