Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Temporal fidelity enhancement for video action recognition*

Shaowu XU[1], Xibin JIA[‡1], Qianmei SUN[2], Jing CHANG[2]

[1]*Faculty of Information Technology, Beijing Univerisity of Technology, Beijing 100124, China*

[2]*Beijing Chao-yang Hospital, Capital Medical University, Beijing 100020, China*

E-mail: swxu@emails.bjut.edu.cn; jiaxibin@bjut.edu.cn; sunqianmei5825@126.com; cj006006@126.com

Received Mar. 14, 2025; Revision accepted June 4, 2025; Crosschecked

Abstract: Temporal attention mechanisms are essential for video action recognition, enabling models to focus on semantically informative moments. However, these models frequently exhibit temporal infidelity—misaligned attention weights caused by limited training diversity and the absence of fine-grained temporal supervision. While video-level labels provide coarse-grained action guidance, the lack of detailed constraints allows attention noise to persist, especially in complex scenarios with distracting spatial elements. To address this issue, we propose temporal fidelity enhancement (TFE), a competitive learning paradigm based on disentangled information bottleneck (DisenIB) theory. TFE mitigates temporal infidelity by decoupling action-relevant semantics from spurious correlations through adversarial feature disentanglement. Using pre-trained representations for initialization, TFE establishes an adversarial process in which segments with elevated temporal attention compete against contexts with diminished action relevance. This mechanism ensures temporal consistency and enhances the fidelity of attention patterns without requiring explicit fine-grained supervision. Extensive studies on UCF-101, HMDB-51, and Charades benchmarks validate the effectiveness of our method, with significant improvements in action recognition accuracy.

Key words: Action recognition; Disentangled information bottleneck (DisenIB); Temporal modeling; Temporal fidelity

https://doi.org/10.1631/FITEE.2500164 **CLC number:**

Unedited

1 Introduction

Learning paradigms based on physical systems and neural mechanisms have advanced video-based action recognition by enhancing temporal dependency modeling and discriminative motion pattern extraction (Jiao et al., 2024a, 2024b). While traditional 3D convolutional networks and two-stream architectures (Liu et al., 2021, 2024) established the foundation for temporal modeling, their reliance on predefined hierarchies limits adaptability to long-term dynamics. Multi-scale temporal dependency methods (Yu et al., 2020; Zhou et al., 2021; Zhang et al., 2025) partially alleviate this but remain constrained by fixed parameterization. This has driven the use of temporal attention mechanisms, which dynamically prioritize critical temporal segments and improve motion feature extraction (Wu et al., 2022; Jiao et al., 2023; Yamazaki et al., 2023; Zhou et al., 2024; Gao et al., 2024).

However, these attention-based models for video action recognition (Mondal et al., 2023; Wang et al., 2024; Wu et al., 2024) primarily depend on video-level supervision, providing only coarse-grained supervision despite the inherent fine-grained temporal variations in action videos (e.g., eating speed differences or utensil usage patterns). This discrepancy causes temporal infidelity (Liang et al., 2020; Aghaeipoor et al., 2023)—a misalignment between attention distributions and action-relevant temporal segments. For instance, as shown in Fig. 1, for actions such as "eating a sandwich," a model

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 62476015 and 62171298)

ORCID: Xibin JIA, https://orcid.org/0000-0001-8799-8042

© Zhejiang University Press 2025

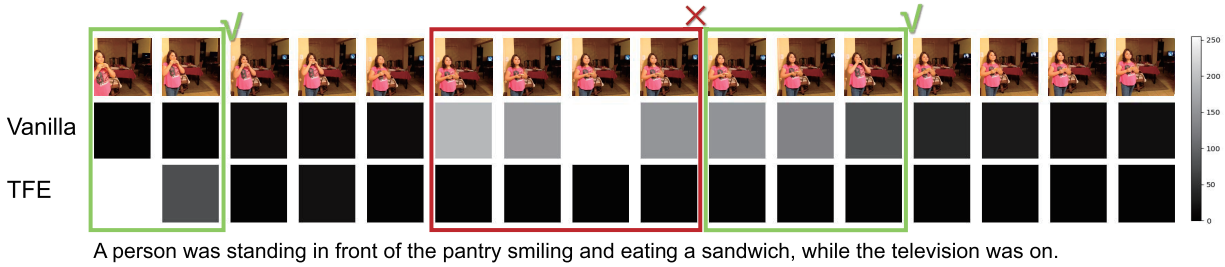

Figure 1 Temporal attention distribution computed by the BIKE ViT-L/14 (Wu et al., 2023) on training and testing samples with similar actions. Row ① shows the video segments, and row ② depicts attention weights, with white indicating higher values and black indicating lower values. Green boxes highlight essential segments based on human cognitive experience, while red boxes indicate non-essential segments. (a) Training sample: model allocates high attention to adjacent segments related to the action "eating a sandwich". (b) Testing sample: model also gives significant attention to adjacent segments, but those that include the action "eating a sandwich" receive low attention

may mistakenly pay significant attention to adjacent segments that are irrelevant while overlooking key discriminative moments in unseen scenarios. This limitation arises from the training paradigm which depends solely on coarse, video-level supervision and fails to capture the fine-grained temporal dynamics intrinsic to action semantics. Although fine-grained annotations could alleviate, their acquisition is prohibitively labor-intensive and resource-consuming.

To address this challenge, we propose a novel TFE framework based on disentangled information bottleneck (DisenIB) theory (Pan et al., 2021). TFE explicitly optimizes temporal attention fidelity by decoupling latent video embedding into action-relevant and semantically redundant components. Unlike conventional attention refinement methods, TFE resolves infidelity through adversarial disentanglement, which maximizes the sufficiency of critical temporal features while suppressing spurious correlations from noisy contexts. The framework first encodes video segments into temporally discriminative embeddings using pre-trained models, followed by an adversarial process to approximate the DisenIB objective. A disentangler separates embeddings into temporal fidelity-preserving and fidelity-deviating ones, while dual approximators compete to retain action-relevant semantics and discard redundant patterns. This mechanism circumvents the combinatorial complexity of direct optimization, allowing for robust temporal attention alignment without requiring fine-grained supervision. The main contributions of this paper are summarized as follows:

1. We introduce TFE, a novel framework based on DisenIB theory that explicitly optimizes temporal attention fidelity by disentangling action-relevant semantics from redundant information.

2. We propose an adversarial disentanglement mechanism that approximates the DisenIB-based objective, enabling effective suppression of spurious correlations without requiring fine-grained supervision.

3. Our method achieves outstanding performances on benchmark datasets, including Charades, UCF-101, and HMDB51.

2 Related works

Video recognition, unlike image recognition, requires understanding of the temporal evolution of objects. The most common methods used convolutional networks, with many adopting 3D CNNs for action recognition (Carreira and Zisserman, 2017; Tran et al., 2018; Xie et al., 2018). Although advanced 3D CNN-based methods have showed promising results on short-term action benchmarks (Kuehne et al., 2011; Soomro et al., 2012), including CF-IIH (Liu et al., 2024) with counterfactual reasoning and KCMM (Liu et al., 2025) with knowledge-driven composition modulation, robust long-term dependency modeling remains crucial for real-world applications (Girdhar et al., 2017; Sigurdsson et al., 2017; Zhou et al., 2018; Feichtenhofer et al., 2019; Feichtenhofer, 2020). CSVR (Zhang et al., 2025) enhances dynamic-static feature learning through generative-discriminative synergy, but it still shares

Unedited

this limitation. Long-term video recognition (Sigurdsson et al., 2016) require identifying long-term actions or their subactions(Yu et al., 2020; Zhou et al., 2021). Timeception (Hussein et al., 2019), TRN (Zhou et al., 2018), VideoGraph (Hussein et al., 2019), RhyRNN (Yu et al., 2020), and GHRM (Zhou et al., 2021) are methods for addressing address long-term dynamics, but they are limited by predetermined patterns.

Recently, Visual Transformer models (Vaswani et al., 2017; Fan et al., 2021) and derivatives (Jiao et al., 2023; Yamazaki et al., 2023; Gao et al., 2024) have gained attention for their adaptability across video scales. UGPT (Guo et al., 2022) uses self-attention for long-term action dynamics, Wang et al. (Wang et al., 2023) use masked video distillation, TwinFormer (Zhou et al., 2024) models fine-to-coarse temporal structures, and MSQNet (Mondal et al., 2023) introduces a multimodal semantic query mechanism. Large-scale pre-trained vision-language models including ViLT-CLIP (Wang et al., 2024), Text4Vis (Wu et al., 2024), and BIKE (Wu et al., 2023) have significantly advanced video action recognition through their respective novel cross-modal interaction designs.

However, these attention-based methods largely adhere to the Information Bottleneck (IB) principle (Tishby et al., 2000; Dimitrov and Miller, 2001) for video embedding refinement. They use IB-based paradigms (Srivastava et al., 2021; Chi et al., 2022; Cen et al., 2023) to learn compressed yet predictive representations by maximizing mutual information between video embeddings and target labels. While this paradigm makes efficient use of coarse-grained (e.g., video-level) supervision, it remains inadequate in separating intra-class dynamic variations when fine-grained (e.g, segment-level) annotations are unavailable, resulting in temporal infidelity.

Our TFE framework addresses temporal infidelity by introducing a novel objective function based on DisenIB (Pan et al., 2021), which resolves temporal infidelity through two key innovations: (1) self-supervised separation of embeddings into action-relevant and redundant components using adversarial optimization, directly targeting the root cause of temporal attention misalignment; (2) dynamic suppression of spurious correlations using competing approximators, outperforming conventional IB's passive compression. TFE achieves temporal fidelity through feature-space purification rather than mere attention redistribution, while maintaining the scalability of coarse-grained supervision paradigms.

3 Method

TFE is a novel learning framework for video action recognition. This section first reviews the general objective function of temporal attention-based models, followed by the formulation of a DisenIB-based learning objective and its associated loss function. We subsequently detail the neural architecture design and conclude with the complete learning scheme.

3.1 Objetive function

3.1.1 Preliminary

Temporal attention-based models (Guo et al., 2022; Mondal et al., 2023; Wang et al., 2023; Zhou et al., 2024) generally comprise three core components: visual encoder, a temporal aggregator, and an approximator. Given input video segments, the encoder maps them to embeddings $\boldsymbol{x} = \{\boldsymbol{x}_t \mid \boldsymbol{x}_t \in \mathbb{R}^d, t = 1, 2, \ldots, T\}$, where T is the sequence length. The aggregator computes temporal attention weights and aggregates the embeddings into a video-level representation $\boldsymbol{z}_S$, termed salient embedding for its focus on high-attention segments. The approximator learns the mapping from $\boldsymbol{z}_S$ to the target action category distribution. This process emphasizes salient segments while suppressing others; it inherently follows the IB principle (Chen et al., 2018; Watson et al., 2024), minimizing:

$$\mathcal{L} = \mathrm{I}(\boldsymbol{x}; \boldsymbol{z}_S) - \mathrm{I}(\boldsymbol{z}_S; \boldsymbol{y}), \tag{1}$$

where $\boldsymbol{y}$ represents action label, and $\mathrm{I}(\cdot;\cdot)$ represents mutual information.

However, this formulation does not explicitly differentiate action-relevant and action-irrelevant information. Due to substantial fine-grained temporal variations in action videos, models relying on video-level supervision often misallocate attention to redundant segments, resulting in temporal infidelity.

3.1.2 Learning objective

TFE aims to explicitly disentangle video embeddings into two distinct semantic components: an

Unedited

action-relevant embedding that preserves recognition fidelity, and an action-irrelevant counterpart that reduces fidelity. This enables TFE to learn discriminative features for action recognition that are robust to intra-class temporal variations. Based on the DisenIB (Pan et al., 2021) principle, TFE introduces a novel objective function for learning the disentanglement of action semantics and redundant information. Let $\boldsymbol{z}_N$ denote non-salient video embedding that captures action-irrelevant information and is semantically complementary to $\boldsymbol{z}_S$. The objective function is formalized as follows:

$$\mathcal{L}_{\mathrm{DIB}} = -\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{y}) + \mathrm{I}(\boldsymbol{z}_N;\boldsymbol{y}) + \mathrm{I}(\boldsymbol{z}_S;\boldsymbol{z}_N). \quad (2)$$

Theorem 1 The DisenIB-based objective function, $\mathcal{L}_{\mathrm{DIB}}$, to be minimized is consistent with maximum compression.

Proof sketch. We first identify the global minimum $\mathcal{L}^*_{\mathrm{DIB}} = -\mathrm{H}(\boldsymbol{y})$. Assuming $\mathcal{L}_{\mathrm{DIB}} - \mathcal{L}^*_{\mathrm{DIB}} < \delta$, we derive $\mathrm{H}(\boldsymbol{y}) - \mathrm{I}(\boldsymbol{z}_S;\boldsymbol{y}) < \delta$, $\mathrm{I}(\boldsymbol{z}_N;\boldsymbol{y}) < \delta$, and $\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{z}_N) < \delta$. Given $\boldsymbol{x} = \boldsymbol{z}_S \cup \boldsymbol{z}_N$, it follows that $\mathrm{H}(\boldsymbol{x}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{z}_N,\boldsymbol{y}) < 3\delta$. Utilizing the Markov chains $\boldsymbol{z}_S \leftrightarrow \boldsymbol{x} \leftrightarrow \boldsymbol{y}$, $\boldsymbol{z}_N \leftrightarrow \boldsymbol{x} \leftrightarrow \boldsymbol{y}$, and $\boldsymbol{z}_S \leftrightarrow \boldsymbol{x} \leftrightarrow \boldsymbol{z}_N$, we apply mutual information properties to show $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{H}(\boldsymbol{y})$. By leveraging inequalities and combining intermediate results, we derive $|\mathrm{I}(\boldsymbol{x};\boldsymbol{z}_S) - \mathrm{H}(\boldsymbol{y})| + |\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{y}) - \mathrm{H}(\boldsymbol{y})| < 25\delta$. Thus, setting $\delta = \epsilon/25$, we show that $\mathcal{L}_{\mathrm{DIB}}$ is consistent on maximum compression according to Definition 1 in (Pan et al., 2021). Detailed proof is shown in the supplementary material.

Theorem 1 shows that the $\mathcal{L}_{\mathrm{DIB}}$ adheres to the maximum compression principle, which ensures that the learned salient and non-salient video embeddings fulfill distinct roles to optimally balancing information preservation and compression, thereby preserving recognition fidelity.

3.1.3 Training strategy

This paper proposes Theorem 2 to derive an optimizable training objective of Eq. (2).

Theorem 2 The global optimum for minimizing $\mathcal{L}_{\mathrm{DIB}}$ satisfies:

$$D^* = \arg\min_D \mathbb{E}\left[-\log p(\boldsymbol{y}|\boldsymbol{z}_S) + \log p(\boldsymbol{y}|\boldsymbol{z}_N)\right], \quad (3)$$

where D^* denotes the optimal disentangler in Eq. (2) that divides input segment embeddings $\boldsymbol{x}$ into salient and non-salient video embeddings, $\boldsymbol{z}_S$ and $\boldsymbol{z}_N$.

Proof sketch. We divide $\mathcal{L}_{\mathrm{DIB}}$ into two sections: $\mathcal{L}_1 = -\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{y}) + \mathrm{I}(\boldsymbol{z}_N;\boldsymbol{y})$ and $\mathcal{L}_2 = \mathrm{I}(\boldsymbol{z}_S;\boldsymbol{z}_N)$. First, we show that D^* minimizes $\mathcal{L}_1$ by showing $\mathbb{E}[-\log p(\boldsymbol{y}|\boldsymbol{z}_S) + \log p(\boldsymbol{y}|\boldsymbol{z}_N)] \geq \mathbb{E}[-\log p(\boldsymbol{y}|\boldsymbol{z}^*_S) + \log p(\boldsymbol{y}|\boldsymbol{z}^*_N)]$ for any pair $(\boldsymbol{z}_S, \boldsymbol{z}_N)$, leading to $-\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{y}) + \mathrm{I}(\boldsymbol{z}_N;\boldsymbol{y}) \geq -\mathrm{I}(\boldsymbol{z}^*_S;\boldsymbol{y}) + \mathrm{I}(\boldsymbol{z}^*_N;\boldsymbol{y})$. Second, using a proof by contradiction, we demonstrate that minimizing $\mathcal{L}_1$ also minimizes $\mathcal{L}_2$, ensuring $\mathrm{I}(\boldsymbol{z}_S;\boldsymbol{z}_N)$ is minimized. Thus, D^* provides the global optimum for minimizing $\mathcal{L}_{\mathrm{DIB}}$. Detailed proof can be found in the supplementary material.

Notably, the optimal solution D^* is based on the true conditional distributions $p(\boldsymbol{y}|\boldsymbol{z}_S)$ and $p(\boldsymbol{y}|\boldsymbol{z}_N)$, which are analytically intractable, making direct optimization infeasible. Inspired by (Liang et al., 2020), TFE uses a self-adversarial training strategy to approximate D^*. Specifically, two approximators, A_S and A_N, estimate the variational distributions $q_S(\boldsymbol{y}|\boldsymbol{z}_S)$ and $q_N(\boldsymbol{y}|\boldsymbol{z}_N)$. An adversarial optimization framework then alternates between maximizing $\log q_S(\boldsymbol{y}|\boldsymbol{z}_S)$ and minimizing $\log q_N(\boldsymbol{y}|\boldsymbol{z}_N)$, driving D to the optimal state. This strategy is formalized as follows:

$$D^* = \arg\min_D \mathbb{E}[-\log q_S(\boldsymbol{y}|\boldsymbol{z}_S) + \log q_N(\boldsymbol{y}|\boldsymbol{z}_N)], \quad (4)$$

$$(A^*_S, A^*_N) = \arg\min_{A_S, A_N} \mathbb{E}[-\log q_S(\boldsymbol{y}|\boldsymbol{z}_S) - \log q_N(\boldsymbol{y}|\boldsymbol{z}_N)]. \quad (5)$$

3.2 Network architecture

3.2.1 Overview

The TFE framework contains three components: a visual encoder that extracts visual features and contextual semantics from raw video segments, resulting in segment embeddings; a disentangler that divides them into salient and non-salient video embeddings; and paired approximators that process respective embeddings using the training strategy outlined in Section 3.1.3.

Fig. 2 shows our model architecture. First, the visual encoder maps the input segment sequence to embeddings $\boldsymbol{x} = \{\boldsymbol{x}_t \in \mathbb{R}^d\}_{t=1}^T$, where T represents the sequence length. Our design decouples the encoder from the subsequent disentangler-approximator modules, allowing for seamless integration into SOTA methods by simply replacing their final layer with our components.

Subsequently, the disentangler processes input $\boldsymbol{x}$ to produce salient and non-salient video embed-

Unedited

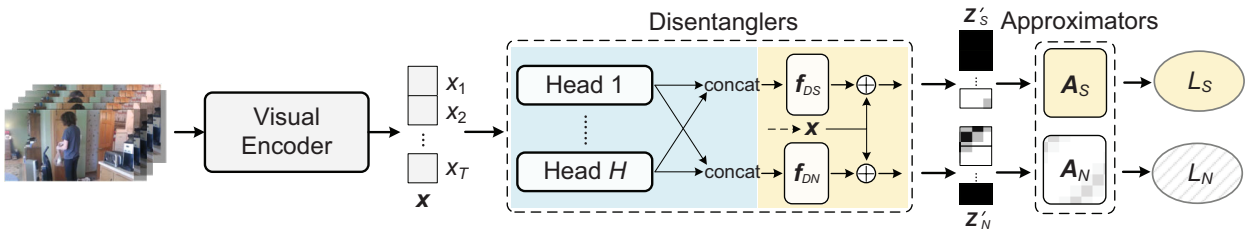

Figure 2 Overview of TFE. We guide our model to learn temporal attention with fidelity for action recognition with the DisenIB objective. The model consists of a visual encoder, a disentangler, and a pair of approximators. DisenIB, disentangled Information Bottleneck; TFE, temporal fidelity enhancement

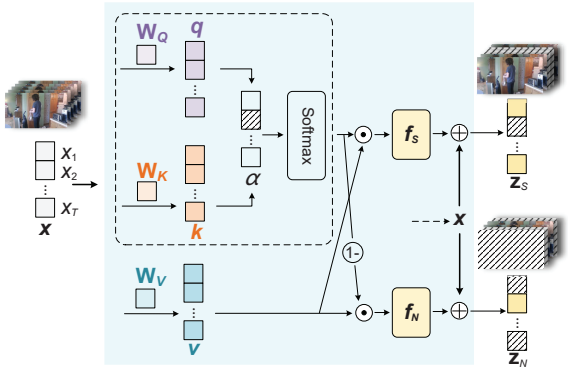

Figure 3 Architecture of the TD module. Given a segment embedding sequence x, this module outputs a pair of semantically distinct video embeddings ($\mathbf{z}_S, \mathbf{z}_N$). TD, temporal disentanglement

dings using H parallel temporal disentanglement (TD) heads. Each head h generates semantically distinct embeddings $\boldsymbol{z}_S^h, \boldsymbol{z}_N^h \in \mathbb{R}^{T\times d}$. Concatenating all heads results in $\boldsymbol{z}_S = [\boldsymbol{z}_S^1; \ldots; \boldsymbol{z}_S^H]$ and $\boldsymbol{z}_N = [\boldsymbol{z}_N^1; \ldots; \boldsymbol{z}_N^H]$. Two feedforward networks f_{DS} and f_{DN}, are then merged to form final embeddings:

$$(\boldsymbol{z}'_S, \boldsymbol{z}'_N) = D(\boldsymbol{x}) = (f_{\mathrm{DS}}(\boldsymbol{z}_S) + \boldsymbol{x}, f_{\mathrm{DN}}(\boldsymbol{z}_N) + \boldsymbol{x}). \tag{6}$$

Finally, TFE incorporates two approximators: the salient approximator A_S and the non-salient approximator A_N. A_S predicts action labels based on salient video embeddings, while A_N predicts from non-salient embeddings. In practice, custom-designed networks can be simple fully connected layers; when integrating into existing models, the original approximator structure can be reused. To compute the losses, these approximators use salient embeddings $\boldsymbol{z}'_S$ and non-salient embeddings $\boldsymbol{z}'_N$ as inputs and output action predictions:

$$\mathcal{L}_S = \ell(\boldsymbol{y}, A_S(\boldsymbol{z}'_S)), \quad \mathcal{L}_N = \ell(\boldsymbol{y}, A_N(\boldsymbol{z}'_N)), \tag{7}$$

where $\ell(\cdot,\cdot)$ represents the cross-entropy loss.

The detailed descriptions of the temporal disentanglement module and learning scheme are elaborated in the following sections.

3.2.2 Temporal disentanglement

TD divides segment embeddings into two distinct representations: salient video embedding which preserves temporal fidelity for action recognition, and non-salient embedding, which encodes redundant temporal contexts. As shown in Fig. 3, our module implements this using a temporal attention mechanism that disentangles input sequences into semantically divergent embeddings.

The module accepts segment embeddings $\boldsymbol{x}$ as input. It first computes a temporal attention mask $\boldsymbol{m}$ to probabilistically separate salient and non-salient temporal segments during feature transformation.

$$\boldsymbol{m} = \text{Gumbel-softmax}(\alpha(\boldsymbol{x}), k) \in \mathbb{R}^{T\times T}, \tag{8}$$

$$\alpha(\boldsymbol{x}) = \frac{(\boldsymbol{x}\otimes \boldsymbol{W}_Q)(\boldsymbol{x}\otimes \boldsymbol{W}_K)^\top}{\sqrt{d}}, \tag{9}$$

where k is an integer satisfying $0 < k \le T$, and $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d\times d}$ are projection matrices. This differentiable mask $\boldsymbol{m}$ quantifies temporal saliency, where higher values indicate action-critical segments while lower values correspond to non-essential patterns. We adopt gumbel-softmax reparameterization (Liang et al., 2020) to enable efficient k-hot sampling for combinatorial optimization of feature disentanglement.

The proposed temporal disentangling module computes two mutually distinct video embeddings as follows:

$$\boldsymbol{z}_S = f_S(\boldsymbol{m}\otimes \boldsymbol{x}\otimes \boldsymbol{W}_V + \boldsymbol{x}) \in \mathbb{R}^{T\times d}, \tag{10}$$

Unedited

$$\boldsymbol{z}_N = f_N((1-\boldsymbol{m}) \otimes \boldsymbol{x} \otimes \boldsymbol{W}_V + \boldsymbol{x}) \in \mathbb{R}^{T\times d}, \quad (11)$$

where $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$ represents value projection matrices, f_S and f_N are feedforward neural networks with the same structure, $\boldsymbol{z}_S$ is the salient video embedding, and $\boldsymbol{z}_N$ is the non-salient video embedding. The complementary mask $1-\boldsymbol{m}$ ensures that $\boldsymbol{z}_N$ captures temporal features neglected by $\boldsymbol{m}$. Since $\boldsymbol{m}$ quantifies saliency, $1-\boldsymbol{m}$ inversely weights segments with lower saliency, which generally encode residual motion patterns (e.g., background hand movements in "eating" actions) or redundant temporal contexts. This explicit inversion ensures mutual exclusivity between $\boldsymbol{z}_S$ and $\boldsymbol{z}_N$, which corresponds to DisenIB's requirement for complementary latent disentanglement.

3.2.3 Learning scheme

The proposed TFE adopts a joint objective in Eq. (12), which derives from the theoretical bounds in Eqs. (4)-(5), to materialize DisenIB principles through adversarial coordination:

$$\min_{(A_S,A_N)} (\mathcal{L}_S + \mathcal{L}_N), \quad \min_{D} (\mathcal{L}_S - \mathcal{L}_N). \quad (12)$$

This is executed via iterative two-phase optimization:

1. Approximator phase: update A_S and A_N using $\nabla(\mathcal{L}_S + \mathcal{L}_N)$ to maximize feature utility from both salient and non-salient embeddings.

2. Disentangler phase: update D with $\nabla(\mathcal{L}_S - \mathcal{L}_N)$ to enforce orthogonality between $\boldsymbol{z}_S$ and $\boldsymbol{z}_N$.

This coordinated optimization addresses temporal infidelity by enhancing temporal discriminability in $\boldsymbol{z}_S$ by $\mathcal{L}_S$ maximization, and suppressing action-irrelevant temporal variations in $\boldsymbol{z}_N$ via $\mathcal{L}_N$ minimization.

4 Experiments

4.1 Datasets

UCF-101 (Soomro et al., 2012) is an action recognition dataset of realistic action videos collected from YouTube, with 101 action categories and 13320 videos. The videos in 101 action categories are grouped into 25 groups, with each group consisting of 4–7 videos of an action. All videos in this dataset are single-label, with an average duration of 5 s and ranging from 4 s to 7 s in length. We split the dataset into 9537 videos for training and 3783 videos for testing.

HMDB-51 (Kuehne et al., 2011) is a collection of realistic videos from various sources, including movies and web videos. The dataset comprises 5100 video clips from 51 action categories. This dataset contains only single-label videos, with an average duration of 7 s and video lengths ranging from 4 s to 22 s. We split the dataset into 3570 videos for training and 1530 videos for testing.

Charades (Sigurdsson et al., 2016) is a dataset composed of 9848 videos of daily indoor activities. The dataset contains 66,500 temporal annotations for 157 action classes, 41,104 labels for 46 object classes, and 27,847 textual descriptions of the videos. Each video in this dataset has multiple labels, and the average duration is 29 s, with video lengths ranging from 2 s to 187 s. We divided the dataset into 7985 videos for training and 1863 videos for testing.

4.2 Implement details

In experiments, the TFE framework integrates established architectures from SOTA methods (Mondal et al., 2023; Wu et al., 2023, 2024) as backbone components, preserving their visual encoders while replacing final classification layers with our disentangler-approximator modules. Hyperparameters align with original implementations, with 16 uniformly sampled segments per video with frame resolutions of 336×336 (Wu et al., 2023, 2024) or 224×224 (Mondal et al., 2023). The network is trained using AdamW (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9, \beta_2 = 0.999$), initial learning rate 5e-5 (cosine decay with 5-epoch warmup), weight decay of 0.2 over 100 epochs. Implemented in PyTorch (Paszke et al., 2019) on an NVIDIA RTX 3090 GPU with 24GB memory.

4.3 Main results

In temporal attention models, temporal fidelity quantifies how effectively attention mechanisms prioritize action-critical segments. Enhancing temporal fidelity directly correlates with improved recognition accuracy. To validate TFE's fidelity enhancement capability, we integrate it into three SOTA baselines: MSQNet (Mondal et al., 2023), BIKE (Wu et al., 2023), and Text4Vis (Wu et al., 2024). Experiments cover UCF-101 (Soomro et al., 2012), HMDB-

Unedited

Table 1 Comparisons with state-of-the-art on Charades using mAP values of subactions within long-term actions as evaluation metrics

Method	mAP(%)
w/o Temporal Attention	
Temporal Fields(Sigurdsson et al., 2017)	22.4
ResNet-152(He et al., 2016)	22.8
TRN(Zhou et al., 2018)	25.2
Rhyrnn(Yu et al., 2020)	25.4
I3D(Carreira and Zisserman, 2017)	32.9
STM(Jiang et al., 2019)	35.3
Timeception(Hussein et al., 2019)	37.2
VideoGraph(Hussein et al., 2019)	37.8
GHRM(Zhou et al., 2021)	38.3
SlowFast R101(Feichtenhofer et al., 2019)	43.4
X3D-XL(Feichtenhofer, 2020)	44.3
w Temporal Attention	
UGPT(Guo et al., 2022)	42.4
TwinFormer(Zhou et al., 2024)	43.6
MViT-B(Fan et al., 2021)	43.9
ActionCLIP(Wang et al., 2023)	44.3
AdaFocus(Li et al., 2023)	47.8
MSQNet(Mondal et al., 2023)	48.5
MSQNet + TFE	49.5
BIKE(Wu et al., 2023)	49.4
BIKE + TFE	**50.1**

51 (Kuehne et al., 2011), and Charades (Sigurdsson et al., 2016). For multi-label Charades, evaluation uses mean Average Precision (mAP), while single-label datasets (UCF-101/HMDB-51) are assessed via classification accuracy.

Long-term action recognition: table 1 compares state-of-the-art approaches on the Charades dataset for long-term video understanding. While 3D convolution-based methods (e.g., ResNet-152 (He et al., 2016)) and short-term temporal approaches (e.g., Temporal Fields (Sigurdsson et al., 2017)) show limited effectiveness in capturing long-range dependencies, architectures modeling complex temporal relationships achieve superior performance. Specifically, GHRM (Zhou et al., 2021) uses graph-based reasoning while SlowFast R101 (Feichtenhofer et al., 2019) uses dual-pathway modeling, though both remain constrained by fixed architectural priors. Temporal attention mechanisms address this limitation through dynamic importance weighting of video segments, thereby simultaneously enhancing local action discrimination and global temporal context modeling. This explains the superior performance of TwinFormer (Zhou et al., 2024), ActionCLIP (Wang et al., 2023), and BIKE (Wu et al., 2023) over non-attentive counterparts.

Our proposed TFE framework achieves the best results, 50.1% mAP with the BIKE backbone and 49.5% with the MSQNet backbone. Notably, TFE shows smaller gains with BIKE than MSQNet. This is due to their supervision granularity: BIKE leverages segment-level labels by splitting long videos into clips during training, while MSQNet uses untrimmed videos with only video-level supervision. This indicates TFE's superior capability to approximate DisenIB's optimal solution through self-adversarial learning, particularly under coarse-grained supervision without explicit temporal fidelity constraints.

Table 2 Comparisons with the state-of-the-arts on short-term action datasets UCF-101 and HMDB-51, where the recognition accuracies of single-label videos are used as evaluation metrics

Method	**UCF**(%)	**HMDB**(%)
w/o Temporal Attention		
ARTNet(Wang et al., 2018)	94.3	74.8
CSVR(Zhang et al., 2025)	94.5	65.5
I3D(Carreira and Zisserman, 2017)	95.6	70.9
CoViFocus(Zheng et al., 2024)	95.8	74.8
TSM(Lin et al., 2019)	95.9	73.5
STM(Jiang et al., 2019)	96.2	72.2
R(2+1)D(Tran et al., 2018)	96.8	74.5
MVFNet(Wu et al., 2021)	96.6	75.7
S3D-G(Xie et al., 2018)	96.8	75.9
CF-IIH(Liu et al., 2024)	96.9	76.7
TDN(Wang et al., 2021)	97.4	76.4
KCMM(Liu et al., 2025)	97.4	77.1
STANet(Li et al., 2023)	97.6	77.7
w Temporal Attention		
LCVE(Ishikawa et al., 2024)	95.6	70.0
VideoMAE(Tong et al., 2022)	96.1	73.3
ActionCLIP(Wang et al., 2023)	97.1	76.2
ViLT-CLIP(Wang et al., 2024)	97.5	73.3
MVD-B(Wang et al., 2023)	97.5	79.7
ZeroI2V(Li et al., 2023)	98.6	83.4
MSQNet(Mondal et al., 2023)	96.0	72.8
MSQNet + TFE	96.5	73.3
Text4Vis(Wu et al., 2024)	98.1	81.3
Text4Vis + TFE	98.3	81.4
BIKE(Wu et al., 2023)	98.9	83.1
BIKE + TFE	**99.1**	**84.3**

Short-term action recognition: table 2 compares the recognition results of different models on the short-term video action datasets UCF-101 and HMDB-51. While some models not relying on temporal attention, such as TDN (Wang et al., 2021),

Unedited

CF-IIH (Liu et al., 2024), and STANet (Li et al., 2023), achieve good recognition accuracy, attention-based models such as ViLT-CLIP(Wang et al., 2024), Text4Vis (Wu et al., 2024), ZeroI2V (Li et al., 2023), and BIKE (Wu et al., 2023) generally achieve higher recognition rates. The temporal attention mechanism enhances the model's sensitivity to short-term actions and optimizes information extraction, enabling the model to extract action-relevant features

The TFE framework consistently enhances short-term action recognition across backbones by suppressing non-salient temporal embeddings. The limited gains with Text4Vis originate from its classification-layer text-visual alignment, which diminishes temporal attention reliance when compared to BIKE's segment-level optimization. MSQNet's smaller improvements versus BIKE reflect short-term videos' inherent low intra-class variation, which reduces the need for segment-level supervision. This performance ceiling highlights that TFE's attention refinement is secondary to base feature quality in scenarios where BIKE's visual–textual interaction dominates temporal modeling.

4.4 Ablation study

We evaluate three key components: TD, optimization via Eq. (4), and optimization via Eq. (5). Table 3 shows ablation results using MSQNet, which was selected based on section 4.3 findings where it exhibited more discernible performance variations than other backbones, enabling clearer ablation insights.

Variants 1 and 2 show that using dual-stream architectures with identical optimization objectives (Equation (4) or (5)) is insufficient for effective embedding differentiation, and structural complexity degrades overall accuracy. Variant 3 also shows that simply combining both objectives without TD leads to even lower performance, suggesting that without TD's structural guidance, the interaction between the two objectives may introduce conflicting gradients that hinder effective representation learning. Variant 5 also shows that using Eq. (5) alone—focused on approximator distribution matching-fails to induce mutually exclusive salient/non-salient segment embeddings, resulting in no significant improvement and even slight performance degradation compared to Variant 1. This indicates that Eq. (5) is insufficient on its own without the complementary effect of adversarially driven embedding separation. In contrast, the opposing prediction requirement of Eq. (4) (i.e., Variant 4) shows a clearer improvement when combined with TD, indicating its effectiveness in leveraging TD to enforce discriminative learning, though gains remain limited due to its non-adversarial formulation.

Table 3 Ablation study of TFE on Charades, UCF-101, and HMDB-51 datasets

Method	TD	Opt w Eq. (4)	Opt w Eq. (5)	mAP(%)
Variant 1	-	-	✓	48.3
Variant 2	-	✓	-	48.2
Variant 3	-	✓	✓	48.1
Variant 4	✓	✓	-	48.7
Variant 5	✓	-	✓	48.2
TFE	✓	✓	✓	**49.5**

The comprehensive comparison in Table 3 shows that only the disentangler-approximators framework with adversarial learning enables TFE, thereby improving recognition accuracy.

4.5 Visualization

As shown in Fig. 4, we visualize the temporal attention calculated by the vanilla model and the model with TFE for comparison. This visualization highlights a benefits of TFE: removal of non-essential temporal variation. For the video shown in Fig. 4, TFE assigns high attention to the first two segments containing the essential information "eating" and "opening television", avoiding interference from irrelevant information that could disrupt action recognition.

4.6 Computational cost

While the proposed TFE framework introduces additional computational overhead due to its dual-stream structure within a single attention layer, it achieves overall computational efficiency by reducing the total number of attention layers in the network, resulting in improved performance without increasing resource consumption. As shown in Table 4, taking MSQNet as an example, the vanilla model uses 12 Transformer layers for segment embedding refinement following the front-end vision-language model. Our approach reduces the total number of layers to 7 by replacing the final layer with a disentangler and a pair of approximators to implement TFE. This results in a reduction of 2.5×10^9 FLOPs and 4.2×10^6 parameters compared to the vanilla model.

Unedited

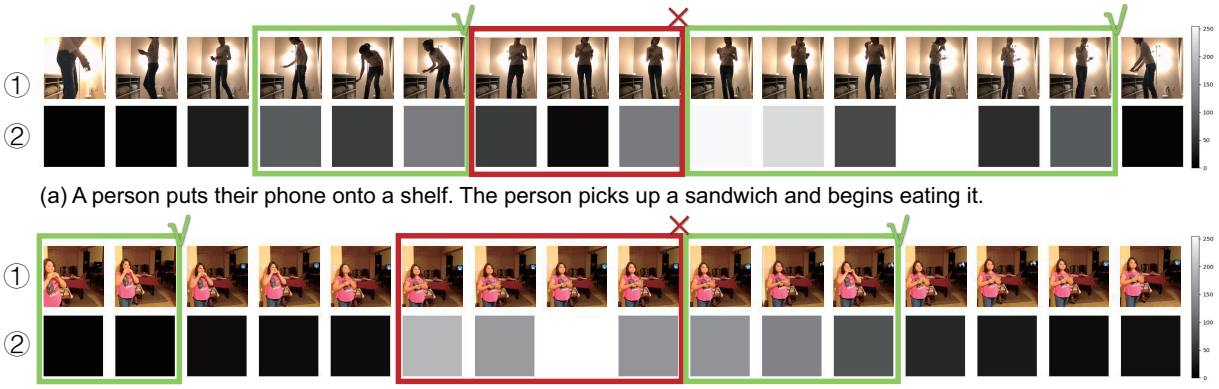

Figure 4 Visualization of temporal attentions calculated by vanilla model and TFE. In each subfigure, the first row is the original video segments. The second and third rows are the temporal attentions computed by the vanilla model and TFE, respectively, with white indicating higher values and black indicating lower values. The green box highlights essential segments based on human cognitive experience, while the red box indicates non-essential ones. TFE, temporal fidelity enhancement

Table 4 Performance and efficiency comparison with-/without TFE on Charades dataset. FLOPs (Giga, 10^9) and parameters (Mega, 10^6) are measured for a single inference pass with the same input. Layer counts refer exclusively to segment embedding refinement modules, excluding the visual encoder

Method	Layers	FLOPs	Params	mAP
MSQNet(Mondal et al., 2023)	12	1303.9	220.4	48.5
MSQNet + TFE	7	1301.4	216.2	49.5
BIKE(Wu et al., 2023)	6	1864.7	230.4	49.4
BIKE + TFE	2	1864.4	211.5	50.1

4.7 Discussion

Although TFE improves temporal attention models, there are several limitations to consider. The framework's TD module specifically targets segment-level infidelity, successfully aligning attention learning with segment saliency to improve overall performance. However, this approach fails to address finer-grained infidelity at the patch level within segments. This limitation becomes particularly evident in short-term action recognition, where TFE shows more modest improvements on MSQNet compared to BIKE across short-term datasets. The performance gap stems from rapid motion changes creating semantically similar but visually heterogeneous patch dynamics. While BIKE's stronger visual-textual interaction reduces sensitivity to these dynamics, MSQNet's visual-only segment embedding refinement remains affected. As a result, TFE's visual self-adversarial mechanism, without considering patch-level infidelity, achieves suboptimal results for short-term action recognition compared to long-term scenarios. Future work should incorporate patch-level fidelity considerations to better handle highly dynamic scenes.

Unedited

5 Conclusions

We propose the TFE framework for video action recognition, which approximates the DisenIB through self-adversarial learning. TFE uses a temporal disentangler to divide input videos into two semantically distinct embeddings: an action-relevant one preserving temporal fidelity, and an action-irrelevant counterpart that encodes residual temporal contexts. TFE enforces strict separation between these components by adversarial coordination of dual approximators, enabling robust recognition against intra-class temporal variations without requiring fine-grained supervision. Experiments using Charades, UCF-101, and HMDB-51 validate TFE's effectiveness.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Aghaeipoor F, Sabokrou M, Fernández A, 2023. Fuzzy rule-based explainer systems for deep neural networks: from local explainability to global understanding. *IEEE Trans. Fuzzy Syst.*, 31(9):3069-3080. https://doi.org/10.1109/TFUZZ.2023.3243935

Carreira J, Zisserman A, 2017. Quo Vadis, action recognition? A new model and the kinetics dataset. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4724-4733. https://doi.org/10.1109/CVPR.2017.502

Cen J, Zhang SW, Wang X, et al., 2023. Enlarging instance-specific and class-specific information for open-set action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15295-15304. https://doi.org/10.1109/CVPR52729.2023.01468

Chen JB, Song L, Wainwright MJ, et al., 2018. Learning to explain: an information-theoretic perspective on model interpretation. Proc 35th Int Conf on Machine Learning, p.882-891.

Chi HG, Ha MH, Chi S, et al., 2022. InfoGCN: representation learning for human skeleton-based action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.20154-20164. https://doi.org/10.1109/CVPR52688.2022.01955

Dimitrov AG, Miller JP, 2001. Neural coding and decoding: communication channels and quantization. *Network: Comput Neural Syst*, 12(4):441-472. https://doi.org/10.1080/net.12.4.441.472

Fan HQ, Xiong B, Mangalam K, et al., 2021. Multiscale vision transformers. Proc IEEE/CVF Int Conf on Computer Vision, p.6804-6815. https://doi.org/10.1109/ICCV48922.2021.00675

Feichtenhofer C, 2020. X3D: expanding architectures for efficient video recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.200-210. https://doi.org/10.1109/CVPR42600.2020.00028

Feichtenhofer C, Fan HQ, Malik J, et al., 2019. SlowFast networks for video recognition. Proc IEEE/CVF Int Conf on Computer Vision, p.6201-6210. https://doi.org/10.1109/ICCV.2019.00630

Gao SY, Chen Z, Chen G, et al., 2024. AVSegFormer: audio-visual segmentation with transformer. Proc 38th AAAI Conf on Artificial Intelligence, p.12155-12163. https://doi.org/10.1609/aaai.v38i11.29104

Girdhar R, Ramanan D, Gupta A, et al., 2017. ActionVLAD: learning spatio-temporal aggregation for action classification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3165-3174. https://doi.org/10.1109/CVPR.2017.337

Guo HJ, Wang HJ, Ji Q, 2022. Uncertainty-guided probabilistic transformer for complex action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.20020-20029. https://doi.org/10.1109/CVPR52688.2022.01942

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Hussein N, Gavves E, Smeulders AWM, 2019a. Timeception for complex action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.254-263. https://doi.org/10.1109/CVPR.2019.00034

Hussein N, Gavves E, Smeulders AWM, 2019b. VideoGraph: recognizing minutes-long human activities in videos. https://arxiv.org/abs/1905.05143

Ishikawa Y, Kondo M, Kataoka H, 2024. Learnable cube-based video encryption for privacy-preserving action recognition. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.6988-6998. https://doi.org/10.1109/WACV57701.2024.00685

Jiang BY, Wang MM, Gan WH, et al., 2019. STM: SpatioTemporal and motion encoding for action recognition. Proc IEEE/CVF Int Conf on Computer Vision, p.2000-2009. https://doi.org/10.1109/ICCV.2019.00209

Jiao JY, Tang YM, Lin KY, et al., 2023. DilateFormer: multi-scale dilated transformer for visual recognition. *IEEE Trans Multimedia*, 25:8906-8919. https://doi.org/10.1109/TMM.2023.3243616

Jiao LC, Song X, You C, et al., 2024. AI meets physics: a comprehensive survey. *Artif Intell Rev*, 57(9):256. https://doi.org/10.1007/s10462-024-10874-4

Jiao LC, Ma MR, He P, et al., 2024. Brain-inspired learning, perception, and cognition: a comprehensive review. *IEEE Trans Neural Netw Learn Syst*, 36(4):5921-5941. https://doi.org/10.1109/TNNLS.2024.3401711

Kuehne H, Jhuang H, Garrote E, et al., 2011. HMDB: a large video database for human motion recognition. Proc Int Conf on Computer Vision, p.2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

Li XH, Zhu YH, Wang LM, 2023. Zeroi2V: zero-cost adaptation of pre-trained transformers from image to video. Proc 18th European Conf on Computer Vision, p.425-443. https://doi.org/10.1007/978-3-031-73010-8_25

Li Z, Zhang RQ, Zou DQ, et al., 2023. Robin: a novel method to produce robust interpreters for deep learning-based code classifiers. Proc 38th IEEE/ACM Int Conf on Automated Software Engineering, p.27-39. https://doi.org/10.1109/ASE56229.2023.00164

Liang J, Bai B, Cao YR, et al., 2020. Adversarial infidelity learning for model interpretation. Proc 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.286-296. https://doi.org/10.1145/3394486.3403071

Lin J, Gan C, Han S, 2019. TSM: temporal shift module for efficient video understanding. Proc IEEE/CVF Int Conf on Computer Vision, p.7082-7092. https://doi.org/10.1109/ICCV.2019.00718

Liu Y, Liu F, Jiao LC, et al., 2024. A knowledge-based hierarchical causal inference network for video action recognition. *IEEE Trans Multimedia*, 26:9135-9149. https://doi.org/10.1109/TMM.2024.3386339

Liu Y, Liu F, Jiao LC, et al., 2025. Knowledge-driven compositional action recognition. *Pattern Recognit*, 163:111452. https://doi.org/10.1016/j.patcog.2025.111452

Liu ZY, Wang LM, Wu W, et al., 2021. TAM: temporal adaptive module for video recognition. Proc IEEE/CVF Int Conf on Computer Vision, p.13688-13698. https://doi.org/10.1109/ICCV48922.2021.01345

Loshchilov I, Hutter F, 2019. Decoupled weight decay regularization. Proc 7th Int Conf on Learning Representations.

Mondal A, Nag S, Prada JM, et al., 2023. Actor-agnostic multi-label action recognition with multi-modal query. Proc IEEE/CVF Int Conf on Computer Vision Workshops, p.784-794. https://doi.org/10.1109/ICCVW60793.2023.00086

Unedited

Pan ZQ, Niu L, Zhang JF, et al., 2021. Disentangled information bottleneck. Proc 35th AAAI Conf on Artificial Intelligence, p.9285-9293. https://doi.org/10.1609/aaai.v35i10.17120

Paszke A, Gross S, Massa F, et al., 2019. PyTorch: an imperative style, high-performance deep learning library. Proc 33rd Int Conf on Neural Information Processing Systems, article 721.

Sigurdsson GA, Varol G, Wang XL, et al., 2016. Hollywood in homes: crowdsourcing data collection for activity understanding. Proc 14th European Conf on Computer Vision, p.510-526. https://doi.org/10.1007/978-3-319-46448-0_31

Sigurdsson GA, Divvala S, Farhadi A, et al., 2017. Asynchronous temporal fields for action recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5650-5659. https://doi.org/10.1109/CVPR.2017.599

Soomro K, Zamir AR, Shah M, 2012. UCF101: a dataset of 101 human actions classes from videos in the wild. https://arxiv.org/abs/1212.0402

Srivastava A, Dutta O, Gupta J, et al., 2021. A variational information bottleneck based method to compress sequential networks for human action recognition. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.2744-2753. https://doi.org/10.1109/WACV48630.2021.00279

Tishby N, Pereira FC, Bialek W, 2000. The information bottleneck method. https://arxiv.org/abs/physics/0004057

Tong Z, Song YB, Wang J, et al., 2022. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. Proc 36th Int Conf on Neural Information Processing Systems, article 732.

Tran D, Wang H, Torresani L, et al., 2018. A closer look at spatiotemporal convolutions for action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6450-6459. https://doi.org/10.1109/CVPR.2018.00675

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Wang H, Liu F, Jiao LC, et al., 2024. ViLT-CLIP: video and language tuning clip with multimodal prompt learning and scenario-guided optimization. Proc 38th AAAI Conf on Artificial Intelligence, p.5390-5400. https://doi.org/10.1609/aaai.v38i6.28347

Wang LM, Li W, Li W, et al., 2018. Appearance-and-relation networks for video classification. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1430-1439. https://doi.org/10.1109/CVPR.2018.00155

Wang LM, Tong Z, Ji B, et al., 2021. TDN: temporal difference networks for efficient action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1895-1904. https://doi.org/10.1109/CVPR46437.2021.00193

Wang MM, Xing JZ, Mei JB, et al., 2023a. ActionCLIP: adapting language-image pretrained models for video action recognition. *IEEE Trans Neural Netw Learn Syst*, 36(1):625-637. https://doi.org/10.1109/TNNLS.2023.3331841

Wang R, Chen DD, Wu ZX, et al., 2023b. Masked video distillation: rethinking masked feature modeling for self-supervised video representation learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6312-6322. https://doi.org/10.1109/CVPR52729.2023.00611

Watson DS, O'Hara J, Tax N, et al., 2024. Explaining predictive uncertainty with information theoretic shapley values. Proc 37th Int Conf on Neural Information Processing Systems, article 320.

Wu CY, Li YH, Mangalam K, et al., 2022. MeMViT: memory-augmented multiscale vision transformer for efficient long-term video recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13577-13587. https://doi.org/10.1109/CVPR52688.2022.01322

Wu WH, He DL, Lin TW, et al., 2021. MVFNet: multi-view fusion network for efficient video recognition. Proc 35th AAAI Conf on Artificial Intelligence, p.2943-2951. https://doi.org/10.1609/aaai.v35i4.16401

Wu WH, Wang XH, Luo HP, et al., 2023. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6620-6630. https://doi.org/10.1109/CVPR52729.2023.00640

Wu WH, Sun Z, Song YX, et al., 2024. Transferring vision-language models for visual recognition: a classifier perspective. *Int J Comput Vis*, 132(2):392-409. https://doi.org/10.1007/s11263-023-01876-w

Xie SN, Sun C, Huang J, et al., 2018. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. Proc 15th European Conf on Computer Vision, p.318-335. https://doi.org/10.1007/978-3-030-01267-0_19

Yamazaki K, Vo K, Truong QS, et al., 2023. VLTinT: visual-linguistic transformer-in-transformer for coherent video paragraph captioning. Proc 37th AAAI Conf on Artificial Intelligence, p.3081-3090. https://doi.org/10.1609/aaai.v37i3.25412

Yu TS, Li YK, Li BX, 2020. RhyRNN: rhythmic RNN for recognizing events in long and complex videos. Proc 16th European Conf on Computer Vision, p.127-144. https://doi.org/10.1007/978-3-030-58607-2_8

Zhang J, Wan ZF, Hu LQ, et al., 2025. Collaboratively self-supervised video representation learning for action recognition. *IEEE Trans Inf Forens Secur*, 20:1895-1907. https://doi.org/10.1109/TIFS.2025.3531772

Zheng ZW, Yang L, Wang YL, et al., 2024. Dynamic spatial focus for efficient compressed video action recognition. *IEEE Trans Circuits Syst Video Technol*, 34(2):695-708. https://doi.org/10.1109/TCSVT.2023.3287201

Zhou BL, Andonian A, Oliva A, et al., 2018. Temporal relational reasoning in videos. Proc 15th European Conf on Computer Vision, p.831-846. https://doi.org/10.1007/978-3-030-01246-5_49

Zhou JM, Lin KY, Li HX, et al., 2021. Graph-based high-order relation modeling for long-term action recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8980-8989. https://doi.org/10.1109/CVPR46437.2021.00887

Zhou JM, Lin KY, Qiu YK, et al., 2024. TwinFormer: fine-to-coarse temporal modeling for long-term action recognition. *IEEE Trans Multimedia*, 26:2715-2728. https://doi.org/10.1109/TMM.2023.3302471

Unedited