Shaowu XU, Xibin JIA, Qianmei SUN, Jing CHANG, 2025. Temporal fidelity enhancement for video action recognition. *Frontiers of Information Technology & Electronic Engineering*, 26(8):1293-1304. https://doi.org/10.1631/FITEE.2500164

Temporal fidelity enhancement for video action recognition

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

Corresponding author: Xibin JIA
E-mail: jiaxibin@bjut.edu.cn
ORCID: https://orcid.org/0000-0001-8799-8042

Motivation

Current temporal attention mechanisms for video action recognition suffer from temporal infidelity, where attention weights misalign with semantically critical moments. This fundamental limitation stems from two key challenges in existing approaches:

1. **Inadequate training signal diversity**: Models are constrained by the limited variability in available training data, thus failing to capture the full spectrum of temporal variations in real-world actions (e.g., differences in utensil usage patterns).

2. **Over-reliance on coarse supervision**: The exclusive dependence on video-level labels forces models to make suboptimal generalizations, often leading to attention misallocation where irrelevant segments receive high attention weights while critical action moments are overlooked.

Main idea

The proposed temporal fidelity enhancement (TFE) framework addresses temporal infidelity via:

1. disentangled embeddings, which explicitly split video embeddings into action-relevant embeddings preserving recognition fidelity and action-irrelevant embeddings encoding redundant contexts;

2. self-adversarial learning, which uses competing approximators to enforce semantic divergence without fine-grained annotations;

3. theoretical guarantee——the DisenIB-based objective ensures maximum compression, optimizing the trade-off between information preservation and redundancy suppression.

Method

1. The TFE framework comprises:

(1) a visual encoder for segment embeddings;

(2) a disentangler with multiple TD (temporal disentanglement) modules to split embeddings into salient/non-salient pairs via Gumbel-Softmax masks;

(3) competing approximators trained under a two-phase adversarial strategy to approximate the DisenIB objective.

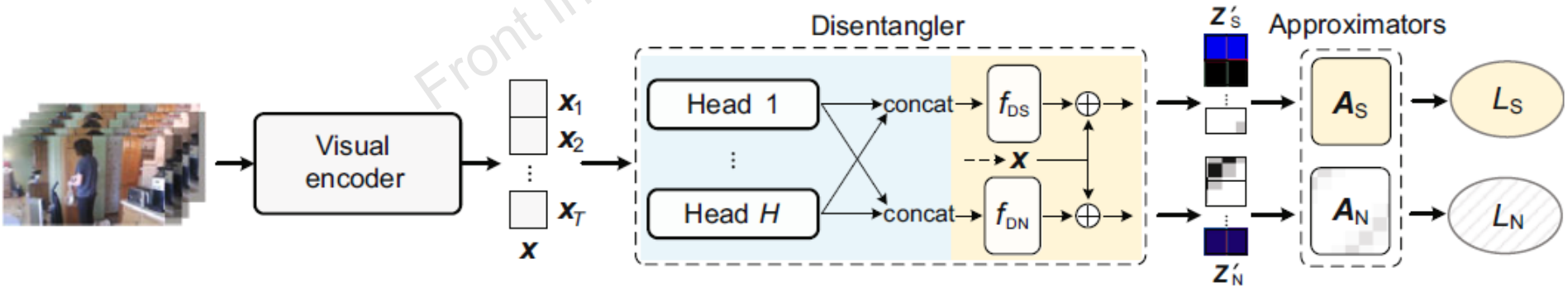

Fig. 2 Overview of TFE. We guide our model to learn temporal attention with fidelity for action recognition with the DisenIB objective. The model consists of a visual encoder, a disentangler, and a pair of approximators

Method (Cont'd)

2. 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.

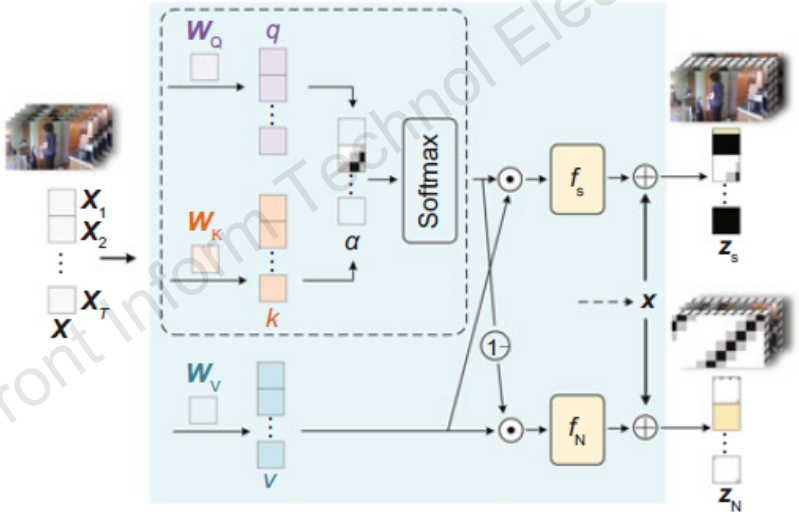

Fig. 3 Architecture of the TD module. Given a segment embedding sequence x, this module outputs a pair of semantically distinct video embeddings (z_S, z_N). W_Q, W_K, and W_V represent the mapping matrices of query, key, and value, respectively. q, k, and v represent the results obtained from the input x after mapping transformations

Major results

Table 1 Comparisons of the state-of-the-art approaches on Charades in terms of mAP values of sub-actions within long-term actions

Method*	mAP (%)
Temporal Fields (Sigurdsson et al., 2017)	22.4
ResNet-152 (He et al., 2016)	22.8
TRN (Zhou BL 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., 2019a)	37.2
VideoGraph (Hussein et al., 2019b)	37.8
GHRM (Zhou JM et al., 2021)	38.3
SlowFast R101 (Feichtenhofer et al., 2019)	43.4
X3D-XL (Feichtenhofer, 2020)	44.3
Method**	**mAP (%)**
UGPT (Guo et al., 2022)	42.4
TwinFormer (Zhou JM et al., 2024)	43.6
MViT-B (Fan et al., 2021)	43.9
ActionCLIP (Wang MM et al., 2023)	44.3
AdaFocus (Li XH et al., 2023)	47.8
MSQNet (Mondal et al., 2023)	48.5
MSQNet+TFE	49.5
BIKE (Wu WH et al., 2023)	49.4
BIKE+TFE	**50.1**

The best result is in bold. * Without temporal attention; ** with temporal attention

Performance of long-term action recognition compared to SOTA methods

Major results (Cont'd)

Table 2 Comparisons of the recognition results of different models on short-term action datasets UCF101 and HMDB-51, where the recognition accuracy of single-label videos is used as the evaluation metric

Method*	RA (%)	
	UCF	HMDB
ARTNet (Wang LM 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 WH et al., 2021)	96.6	75.7
S3D-G (Xie et al., 2018)	96.8	75.9
CF-IIH (Liu Y et al., 2024)	96.9	76.7
TDN (Wang LM et al., 2021)	97.4	76.4
KCMM (Liu Y et al., 2025)	97.4	77.1
STANet (Li XH et al., 2023)	97.6	77.7
Method**	RA (%)	
	UCF	HMDB
LCVE (Ishikawa et al., 2024)	95.6	70.0
VideoMAE (Tong et al., 2022)	96.1	73.3
ActionCLIP (Wang MM et al., 2023)	97.1	76.2
ViLT-CLIP (Wang H et al., 2024)	97.5	73.3
MVD-B (Wang R et al., 2023)	97.5	79.7
ZeroI2V (Li XH et al., 2023)	98.6	83.4
MSQNet (Mondal et al., 2023)	96.0	72.8
MSQNet+TFE	96.5	73.3
Text4Vis (Wu WH et al., 2024)	98.1	81.3
Text4Vis+TFE	98.3	81.4
BIKE (Wu WH et al., 2023)	98.9	83.1
BIKE+TFE	**99.1**	**84.3**

The best result is in bold. RA: recognition accuracy. * Without temporal attention; ** with temporal attention

Performance of short-term action recognition compared to SOTA methods

Major results (Cont'd)

Visualization of TFE's effectiveness in mitigating temporal infidelity compared to the vanilla model

Fig. 4 Visualization of temporal attention calculated by the vanilla model and TFE. In each subfigure, the first row represents the original video segments showing that a person is standing in front of the pantry smiling and eating a sandwich, while the television is on. The second and third rows are the temporal attention 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. References to color refer to the online version of this figure

Conclusions

1. With the TFE framework, a novel temporal fidelity enhancement approach has been proposed for video action recognition.

2. By decoupling salient and non-salient embeddings via adversarial disentanglement, the model achieved accurate temporal attention alignment without fine-grained supervision.

3. The proposed method demonstrated state-of-the-art performance on multiple benchmarks, with a significant improvement in temporal localization fidelity and recognition accuracy.

Shaowu XU is a Ph.D. candidate at Beijing University of Technology. His research focuses on multimodal fusion and causal disentanglement in computer vision, aiming to enhance model interpretability and robustness. His work explores the integration of heterogeneous visual and textual data while uncovering causal structures for more reliable AI systems.

Xibin JIA is a Professor in the Information Faculty at Beijing University of Technology. She received a Ph.D. degree in Computer Science and Technology from Beijing University of Technology in 2007. Her current main research interests include visual information cognition and computing, sentimental analysis, affection computing, and medical image analysis and diagnosis.