Frequency-Temporal Feature Integration for Compressed Video Action Recognition

Jiangwan Zhou, Yue Ming

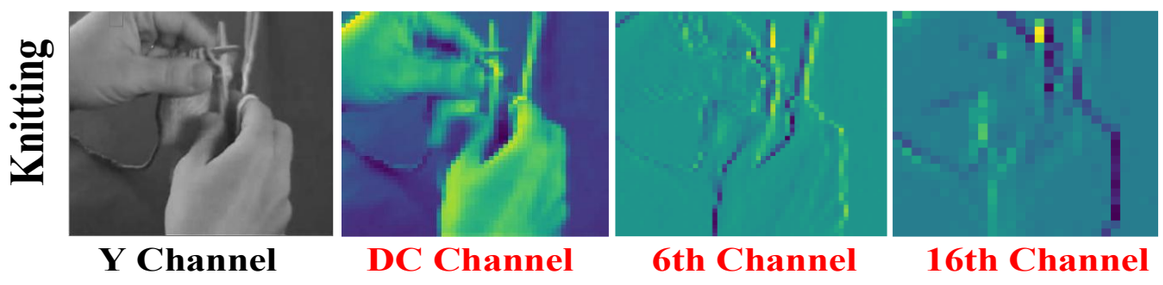
Beijing Key Laboratory of Work Safety Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications

BMVC 2025

Motivations

Challenges:

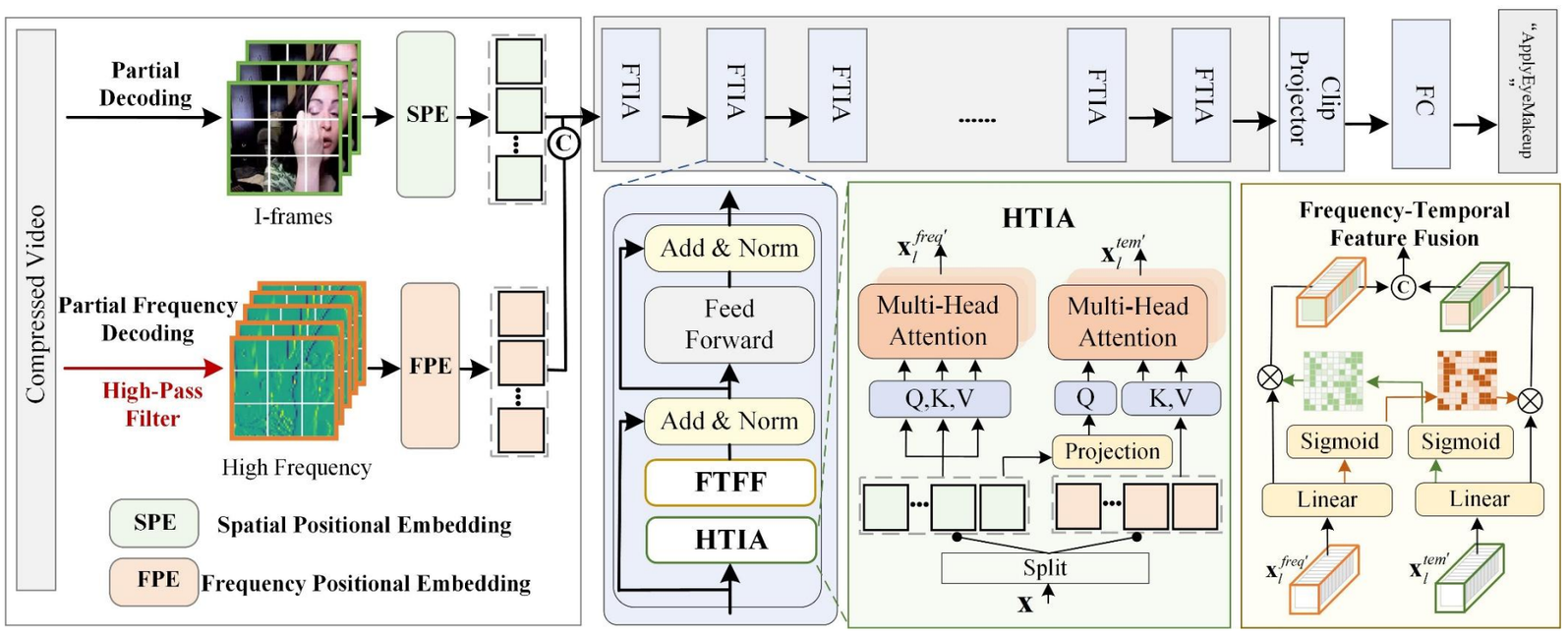
- Transformer-based action recognition methods often ignore **high-frequency motion cues** (e.g., edge textures, motion boundaries), leading to **loss of fine-grained temporal details.**
- Existing compressed video approaches mainly focus on low-frequency global context, **overlooking local high-frequency dynamics** that are crucial to distinguish similar actions (e.g., writing vs. typing).
- There is a need for a model that balances temporal and frequency information, efficiently integrating both global motion and fine edge cues.



Contributions

- FreqTNet: A ViT-based framework achieving temporal-frequency balance for compressed video action recognition.
- FTE: Frequency-Aware and Temporal-Spatial Embedding compensates for high-frequency loss in ViT.
- FTIA: Frequency-Temporal Interaction Attention enables hierarchical fusion of temporal and frequency features.
- Achieves **state-of-the-art** accuracy in compressed video action recognition with **low computational cost.**

Method



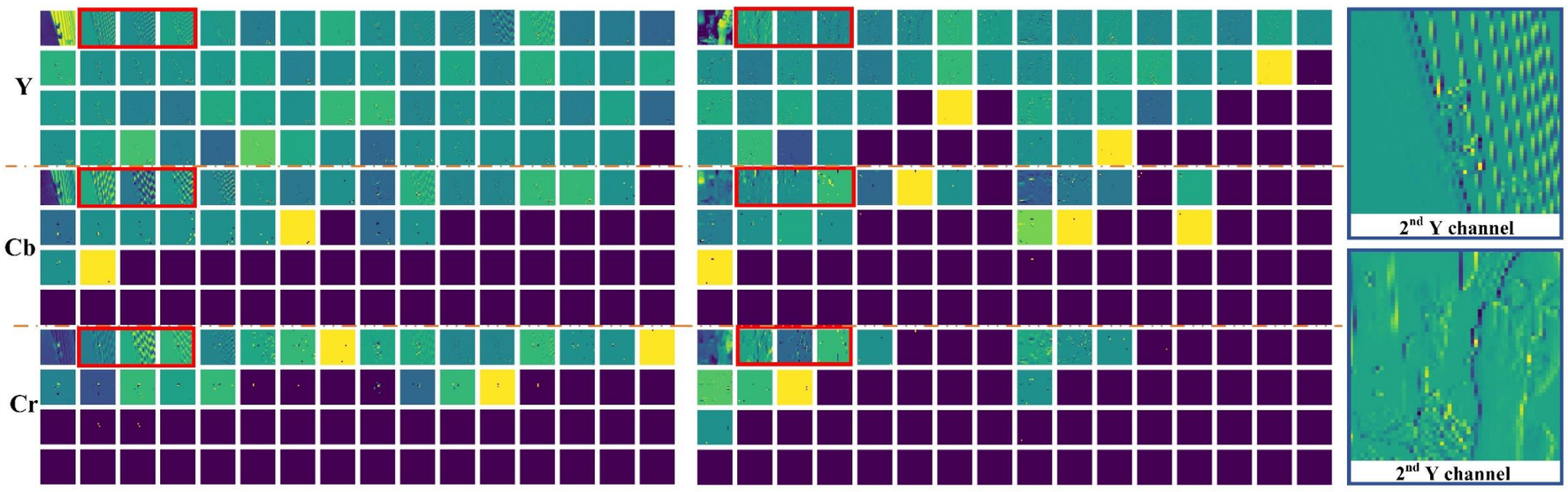
Overview of frequency-temporal feature integration for compressed video action recognition

Frequency-Aware and Temporal-Spatial Embedding (FTE)

- Extracts temporal features from I-frames and high-frequency DCT features from P-frames.
- Encodes both into unified tokens, enabling the model to perceive global motion and fine edge details simultaneously.

Frequency-Temporal Interaction Attention (FTIA)

- High-frequency & Temporal Interaction Attention for cross-domain correlation
- Feature Fusion for adaptive integration of temporal and frequency information



Visualization of 192 channels from Y, Cb, and Cr components.

Experimental Results

Quantitative results:

Methods	Published on	Modality	GFLOPs	Kinetics-400	UCF-101	HMDB-51
Video action recognition						
Vita-CLIP B/16 [35]	CVPR'23	RGB	1164	81.8	-	-
AIM B/16 [39]	ICLR'23	RGB	2316	83.9	-	-
BIKE B/16 [37]	CVPR'23	RGB	1104	83.9	-	-
M^2-CLIP B/16 [31]	AAAI'24	RGB	2453	83.4	-	-
OmniViD [29]	CVPR'24	RGB	-	83.6	-	-
EPK-CLIP B/16 [41]	ESWA'24	RGB	1732	84.3	97.4	77.7
GBC B/16 [40]	TCSVT'24	RGB	1882	84.4	**97.7**	76.8
InternVideo2 1B [33]	ECCV'24	RGB	-	91.3	-	-
OmniVec2 [28]	CVPR'24	RGB	-	**93.6**	-	-
Compressed video action recognition						
CoViAR [36]	CVPR'18	RGB	3775	69.1	90.4	59.1
Faster-FCoViAR [38]	BMVC'21	Frequency	2400	69.3	91.2	63.2
MM-ViT [6]	WACV'22	RGB	2460	-	93.3	-
LEA-Net [13]	ICMM'23	RGB	-	-	94.8	72.2
CoViFocus [42]	TCSVT'23	RGB	296	72.0	95.8	74.8
Fenet [22]	ICIP'23	Frequency	234	72.6	94.5	73.8
DSDMT [24]	TCSVT'23	RGB	141	74.1	95.8	74.9
F2D-SIFPNet [23]	APIN'24	Frequency	192	74.0	95.8	74.1
PKD [27]	BMVC'24	RGB	612	-	93.3	66.3
CFM-Net [19]	ICASSP'25	RGB	48	-	94.8	75.7
Biswas et al. [4]	Arxiv'25	RGB	765	74.2	95.6	74.6
FreqTNet (ours)		Freq + RGB	916	85.1	**97.7**	**79.4**

Preprocessing time (ms):

Input	Modality	Frames	Pre Time	Top1
Coviar frames	RGB	8	76.40	84.1
Fenet frames	Frequency	8	**37.68**	83.1
DCT	Frequency	8	2325	83.9
FreqTNet (DCT)	Frequency + RGB	3 + 5	1481	**85.3**
FreqTNet	Frequency + RGB	3 + 5	52.20	85.1

Qualitative results: