Faster-FCoViAR: Faster Frequency-Domain Compressed Video Action Recognition

Lu Xiong[1]
xionglu24@bupt.edu.cn
Xia Jia[2]
jia.xia@zte.com.cn
Yue Ming[1,*]
yming@bupt.edu.cn
Jiangwan Zhou[1]
zhoujiangwan@bupt.edu.cn
Fan Feng[1]
fan.feng@bupt.edu.cn
Nannan Hu[1]
hunan246@bupt.edu.cn

[1] Beijing Key Laboratory of Work Safety
Intelligent Monitoring, School of
Electronic Engineering, Beijing
University of Posts and
Telecommunications,
Beijing 100876, P.R. China

[2] State Key Laboratory of Mobile Network
and Mobile Multimedia Technology,
Shenzhen 518055, P.R. China

Abstract

Human action recognition (HAR) is an essential task in computer vision, which still faces the critical challenge of reducing the data redundancy of decompressed video frames and extracting identification information. To address this challenge, we propose a novel faster frequency-domain compressed video action recognition framework (termed Faster-FCoViAR), which consists of a frequency-domain partial decompression method (FPDec), a frequency-domain channel selection strategy (FCS), and a spatial-to-frequency domain student-teacher network (S2FNet). The FPDec obtains frequency-domain DCT coefficients of compressed videos directly without inverse discrete cosine transform (IDCT) for decompression. The FCS down-samples frequency-domain data to enhance the saliency of input. The S2FNet transfers spatial semantic knowledge from a spatial teacher network to a light-weight student network in the frequency domain, and it thus improves the spatial feature extraction ability of the frequency-domain network. Experiments on datasets UCF-101, HMDB-51, and Kinetics-400 show that our Faster-FCoViAR is 12.3 times faster than the frame-based methods and 6.7 times faster than other compressed domain methods based on competitive recognition accuracy compared with the state-of-the-art action recognition methods.

1 Introduction

Human action recognition (HAR) plays a vital role in many applications such as intelligent video understanding [17, 43], video surveillance [25], human-computer interaction [2]. Current HAR methods are facing the loss of long-range temporal information and the spatio-temporal aggregation problem, especially for the high computational complexity.

*Corresponding author: yming@bupt.edu.cn

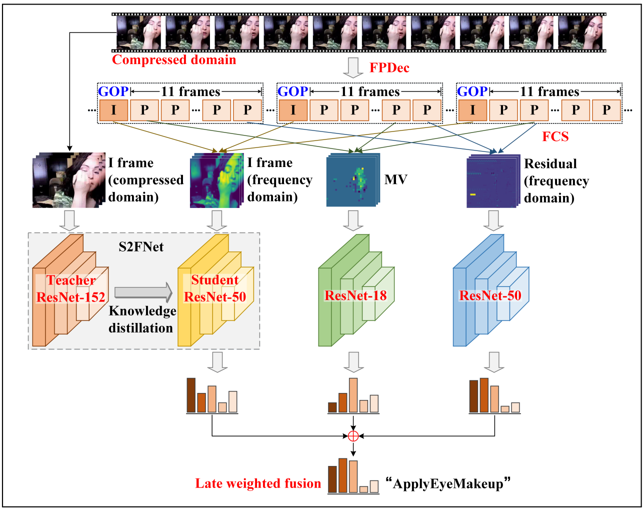

Figure 1: The structure of Faster-FCoViAR. The FPDec gets frequency-domain data from compressed videos. The FCS selects salient channels of frequency-domain data to reduce data redundancy. The S2FNet leverages spatial information of the teacher network in the spatial domain to improve the performance of the student network in the frequency domain.

Most existing HAR methods are based on 2D-CNN [26, 30, 31, 32], 3D-CNN [1, 7, 20, 34, 41] or RNN [3, 18]. However, these methods generally need to completely decode the video into RGB frames, which usually take up huge computing resources. At the same time, the simple down-sampling in the spatial domain ignores the salient information of the video, which can result in accuracy degradation.

Recently, HAR methods based on the compressed domain have received a lot of attention [8, 13, 22, 33, 38]. CoViAR [33] is proposed for partially decoding compressed videos into the RGB domain. The compressed video is decoded into I-frames (I), residuals (R), and motion vectors (MV). Then, the following developed CoViAR methods are presented, such as DMC-Net [22], IP-TSN [13], and IF-TTN [38]. However, this decompression method requires Inverse Discrete Cosine Transform (IDCT), which takes up about 80% of video decompression time [21].

To improve the computational efficiency, some studies focus on using the frequency-domain DCT coefficients instead of RGB pixels [4, 5, 9, 29, 37]. The frequency-domain data expresses the importance distribution of different frequency components. The low-frequency components contain the most spatial and motion information of the video, and the Y channel of YCbCr contains more spatial and motion information than the others [37].

In this paper, we propose a novel faster frequency-domain compressed video action recognition framework (Faster-FCoViAR) as shown in Figure 1, which can efficiently describe the action discrimination information. Firstly, a frequency-domain partial decompression method (FPDec) is proposed, consisting of entropy decoding, zigzag reordering, and inverse quantization to obtain DCT coefficients of compressed videos. Then, we design a frequency-domain channel selection strategy (FCS) to down-sample the salient frequency components. Thirdly, we build a spatial-to-frequency domain student-teacher network (S2FNet) to improve the classification performance of our frequency-domain I-network. Finally, we adopt the late fusion of I, R, and MV networks, achieving comparable accuracy on UCF-101[24], HMDB-51 [16], and Kinetics-400 [15] with high efficiency.

The contributions of this paper are summarized as follows: (1) We propose a frequency-domain partial decompression method (FPDec) which can reduce the data redundancy and make spatial and motion information more prominent by obtaining frequency-domain data and motion vectors directly from the compressed video stream. (2) We propose a spatial-to-frequency domain student-teacher network (S2FNet) to extract identification information in the spatial and frequency domain simultaneously with a small computational complexity. During training, the light-weight I-network in the frequency domain learns discriminant spatio-temporal features from the complex I-network in the RGB domain. (3) We conduct experiments on UCF-101, HMDB-51, and Kinetics-400. The results show that our method can achieve competitive accuracy compared with state-of-art methods with high efficiency.

2 Related Work

Action Recognition: Representative methods of HAR include Two-stream models [17, 23, 30] and 3D Convolution models [1, 7, 10, 20, 41]. The Two-stream model uses two independent 2D-CNNs to obtain spatial and temporal information, but the complexity of optical flow calculation is high that leads to the difficulty of training in an end-to-end way. To alleviate the problem, several studies [14, 27] exploit CNNs to estimate optical flows directly from RGB sequences. Several attempts [28, 36] simulate 3D convolutions with the combination of 2D spatial and 1D temporal convolutions to reduce the high computational cost. However, the 3D convolutions adopted by these methods still require huge computing resources.

Action Recognition in Compressed Videos: Compressed video action recognition approaches make use of motion vectors to capture dynamic information. Zhang *et al.* [39] first replace the optical flow stream with a motion vector stream, but it still needs to decode RGB images for P-frames and ignores other motion-encoding modalities such as the residuals. CoViAR [33] uses compressed video data, namely I-frames, motion vectors and residuals with three independent networks respectively, which achieves a high efficiency for action recognition. However, the performance is worse than traditional Two-stream methods. DMC-Net [22] improves CoViAR and achieves state-of-the-art results by adding an optical flow generation network, but both models [22, 33] employ the large Resnet-152 [11] as the backbone with high computational cost. More recently, Zhou *et al.* [42] use the motion vector information to select key information sequences for recognition and further to formulate the representation of the selected sequences. Wu *et al.* [35] propose a multi-teacher knowledge distillation framework, to compress the model by transferring the knowledge from multiple teachers to a small student model. All the above works require decoding videos to RGB image sequences, which increases the preprocessing time.

Learning in the frequency domain: Frequency-domain data contains motion cues and appearance changes. When decoding the compressed data to the frequency domain, it only needs to obtain DCT coefficients from Huffman code without complete decompression, which can shorten the decompression time. Matej *et al.* [29] first use CNN to directly learn image classification in the frequency domain instead of spatial domain. Lionel *et al.* [9] train the CNN classifier directly on the DCT coefficients computed by the JPEG codec. Ehrlich *et al.* [6] propose a model conversion algorithm to convert the spatial domain CNN models to the frequency domain. Xu *et al.* [37] propose a learning-based frequency channel selection method to replace the traditional spatial down-sampling method. In this paper, we introduces a method for transferring frequency-domain learning into compressed domain action recognition, obtaining DCT coefficients by partially decoding videos into the frequency domain.

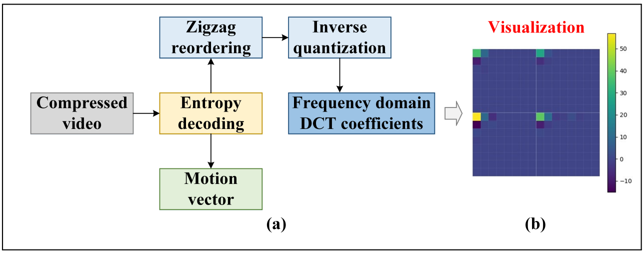

Figure 2: The process of FPDec: (a) the steps of FPDec; (b) visualized image of four macroblocks in the frequency domain.

3 Methods

Our Faster-FCoViAR is an efficient compressed video HAR framework in the frequency domain. First, DCT coefficients of the compressed video are obtained by the frequency-domain partial decompression (FPDec) with high efficiency. Second, we use a frequency-domain channel selection (FCS) strategy to down-sampling the frequency-domain data. Third, to fully learn spatial and frequency semantics, we propose a spatial-to-frequency domain student-teacher network (S2FNet) for the frequency-domain I-network.

3.1 Frequency-domain Partial Decompression Method (FPDec)

In our FPDec, compressed videos are decoded into the frequency domain by entropy decoding, zigzag reordering, and inverse quantization, as shown in Figure 2. For frequency-domain I-frames and residuals, first, the entropy decoding decodes the bitstream to frequency-domain data. Second, the zigzag reordering restores the sorting of DCT coefficients. Finally, the inverse quantization regains the DCT coefficients according to the quantizer scale parameter without IDCT. The compression standard we use in this paper is MPEG-4 Simple Profile/Level 1 [21], which performs DCT and IDCT on macroblocks of 8×8 pixels in the YCbCr color space, thus the frequency-domain data is in a unit of 8×8 block, as shown in Figure 2(b). Noting that our method can be also applicable for other compression standards such as H.264, with the same encode and decode processes consisting of the entropy decoding, zigzag reordering, inverse quantization, and IDCT [21]. DC, low-frequency, and high-frequency components are sorted inside the macroblocks from top to bottom and left to right. The DC and low-frequency areas contain the salient information, while most high-frequency areas are zeros which represent redundant and non-salient information. The ordering of frequency components indicates the different significance of frequency components. For videos with a resolution of $H \times W$, the size of frequency-domain I-frames and residuals is $H \times W \times 3$.

For motion vectors, we obtain them by entropy decoding. According to the MPEG-4 Simple Profile/Level 1, the motion vector is in a unit of 16×16 with the same value. For videos with a resolution of $H \times W$, the size of MV is $H \times W \times 2$. In the following experiments, we choose $H = W = 448$.

3.2 Frequency-domain Channel Selection (FCS)

To alleviate the non-salient high-frequency data, we develop the frequency-domain channel selection (FCS). For frequency-domain I-frames and residuals of $448 \times 448 \times 3$, we first

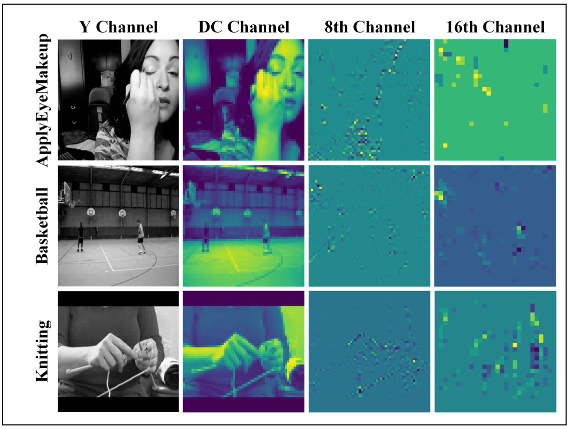

Figure 3: Visualization of raw Y channel and channels after FCS. The columns are raw Y channel, the first DC channel, the 8th channel and the 16th channel.

group all components in 8 × 8 blocks of the same frequency into one channel, maintaining the spatial relationships. Therefore, Y, Cb, and Cr have $8 \times 8 = 64$ channels separately. The size of frequency-domain I-frames and residuals is $56 \times 56 \times 192$. Secondly, we down-sample the channels in upper left corner out of 8 × 8 channels [37]. Finally, the input size of frequency-domain I-frames and residuals is $56 \times 56 \times K$, where K means the number of selected channels. The visualization of the DC channel is similar to the raw Y channel, as demonstrated in Figure 3, because the DC channel contains the salient information of raw data. The FCS of I-frames and residuals remains the most important information with a smaller size compared with the common input size of $224 \times 224 \times 3$ by spatial down-sampling. Thus, the FCS reduces the computing complexity of networks.

For MV of $448 \times 448 \times 2$, we down-sample it to $224 \times 224 \times 2$ without loss of information because the values inside a 16×16 block of MV are the same. Our MV input keeps the common size of $224 \times 224 \times 2$ while contains more motion cues of high-resolution videos.

3.3 Spatial-to-frequency Domain Student-teacher Network (S2FNet)

Our frequency-domain I-network takes DCT coefficients as input after FCS processing, which ensures that useful low-frequency information is learned by the network. In addition, frequency-domain and spatial information are complementary to their different representations of videos. To learn the spatial and frequency semantics simultaneously, we design a spatial-to-frequency domain student-teacher network (S2FNet) as shown in Figure 4. A spatial domain frame-based I-network of Resnet-152 [33] is built as the teacher network and a frequency-domain I-network of Resnet-50 [11] is built as the student network. To accommodate the input size of I-network, we skip its first convolution layer and the subsequent max-pooling layer. We first train the teacher I-network in the spatial domain, then freeze its weights, and take its output logits as the soft labels. Secondly, we train the student I-network in the frequency domain. The loss function is composed of two parts. The first part is the classification loss function of the student I-network,

$$L_1(W) = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} (y_{true})_i^c \ln(q_s)_i^c \tag{1}$$

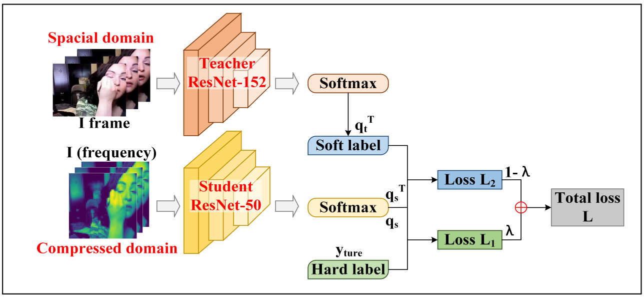

Figure 4: The structure of S2FNet.

where q_s is the output logit of the student I-network, (y_{true}) is the hard label of ground truth.

The second part is shown as follows,

$$L_2(W) = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}(q_t^T)_i^c \ln(q_s^T)_i^c \tag{2}$$

where $(q_s^T)_i^c$ is the output soft logit of the student I-network,

$$(q_s^T)_i^c = \frac{exp((z_s)_i^c/T)}{\sum_k exp((z_s)_i^k)/T)}, \quad for \quad c = 1,...,C \tag{3}$$

and $(q_t^T)_i^c$ is the soft label of the teacher I-network,

$$(q_t^T)_i^c = \frac{exp((z_t)_i^c/T)}{\sum_k exp((z_t)_i^k)/T)}, \quad for \quad c = 1,...,C \tag{4}$$

where z_s and z_t are the logits vector produced by the student network and the teacher network respectively. T is a temperature that a higher value for T produces a softer probability distribution over classes [12]. The total loss function can be written as,

$$L = \lambda L_1 + (1-\lambda)L_2 \tag{5}$$

where λ is the weighted parameter. The S2FNet can transfer the spatial knowledge from the teacher network to the student network while retaining the frequency-domain learning ability, so the S2FNet can aggregate the spatial and frequency semantic information.

The architecture of Faster-FCoViAR is shown in Figure 1. For our frequency-domain R-network, Resnet-50 the same as our frequency-domain I-network is adopted as the backbone network, as I and R data both consist of DCT coefficients. For the motion vectors, we use Resnet-18 [11] as the backbone network. Finally, we adopt the late fusion of I, R, and MV-networks and get the recognition results.

4 Experiments

To validate our method, we conduct experiments on UCF-101 [24], HMDB-51 [16], and Kinetics-400 [15]. We have also done the ablation studies of FCS, FPDec, and S2FNet, the speed and efficiency of our Faster-FCoViAR and the comparison with state-of-art methods.

Net	Backbone	Input	LR	Steps	Epochs	Batch	Views	Pretrain
I-net (S)	Res152/Res50	RGB	0.0003	[55, 110, 160]	220	16	-	RGB
I-net (F)	Res50	DCT	0.0006	[55, 110]	150	16	25	DCT
I-net$_{Lite}$ (F)	Res18	DCT	0.01	[150, 270, 350]	380	16	25	DCT
R-net	Rse50	DCT	0.0003	[55, 110, 160]	220	16	25	DCT
R-net$_{Lite}$	Rse18	DCT	0.01	[150, 270, 350]	380	16	25	DCT
MV-net	Res18	-	0.01	[150, 270, 390]	510	256	25	RGB

Table 1: Training and testing details. "LR" means the learning rate; "Steps" means epochs to decay learning rate; "Epochs" means the number of training epochs; "S" means networks in the spatial domain; "F" means networks in the frequency domain; "Lite" means I and R networks that take Resnet-18 as their backbone.

In addition, we also evaluate I and R networks that use Resnet-18 as their backbones to get a more light-weight model, named Faster-FCoViAR$_{Lite}$.

Video data augmentation. To increase the variations of the data while avoiding complete decompression, we propose a new video data augmentation strategy for compressed videos. Without decompression, we randomly flips and crops the videos following the ratio of CoViAR [33].

Training and testing details. The training and testing schedule is shown in Table 1. The teacher I-network is trained first, then, its weights are frozen to train the student I-network. The R and MV networks are trained separately. Finally, we take the weighted average of the scores of I, MV, and R-network as the final classification result. For the S2FNet of Kinetics-400, we use the I-network of CoViAR as the teacher network, but the backbone of which is Resnet-50 rather than Resnet-152 for higher computational efficiency in training. Expriments of preprocessing time and VPS are all evaluated on an NVIDIA RTX-2080Ti GPU with Intel Xeon E5-2620 v4 CPU. Other experiments of UCF-101 and HMDB-51 are conducted on four NVIDIA RTX-2080Ti GPUs with Intel Xeon E5-2620 v4 CPU, while other experiments of Kinetics-400 are conducted on eight NVIDIA RTX-3090 GPUs with Intel Xeon Silver 4216 CPU.

4.1 Ablation Study

We first conduct experiments of hyperparameter K in FCS on UCF-101. As shown in Table 2, when we select $K = 24$ channels out of 192 channels, the frequency-domain network achieves the best performance. Thus, K is set to 24 in the following experiments.

K	Dataset	I-net
16	UCF-101	68.0
24	UCF-101	**73.9**
32	UCF-101	69.3
64	UCF-101	66.7
192	UCF-101	59.2

Table 2: Analysis for the hyperparameter K in FCS.

T	λ	Dataset	I-net
1.0	0.2	UCF-101	72.6
2.0	0.2	UCF-101	75.4
2.0	0.1	UCF-101	**77.8**
5.0	0.1	UCF-101	77.2
1.0	0.2	HMDB-51	41.3
2.0	0.2	HMDB-51	42.2
2.0	0.1	HMDB-51	**44.7**
5.0	0.1	HMDB-51	42.5

Table 3: Analysis for the hyperparameters T and λ in S2FNet.

We also conduct ablation experiments to evaluate the performance of FCS, DCT pretrained model, and video data augmentation on UCF-101 and HMDB51. They can improve the accuracy of both frequency-domain I and R networks, as shown in Table 4. Among them,

Dataset	Pretrain	Data aug	FCS	I-net	R-net
UCF-101	×	✓	×	57.6	58.0
	×	✓	✓	62.0	72.6
	✓	×	✓	60.8	61.9
	✓	✓	×	59.2	55.4
	✓	✓	✓	**73.9**	**75.4**
HMDB-51	×	✓	×	27.9	16.1
	×	✓	✓	38.1	43.6
	✓	×	✓	24.6	25.1
	✓	✓	×	26.5	19.2
	✓	✓	✓	**40.0**	**45.9**

Table 4: Ablation studies on pretrained model, FCS, and video data augmentation.

Input	Pre	I-net	MV-net	R-net	I+MV	I+R	I+MV+R
Frames	9.55	83.2	67.1	80.2	87.6	86.3	88.6
DCT	775	77.7	73.1	67.2	85.7	81.0	86.3
Partial	4.71	73.9	81.6	75.4	85.6	80.9	87.8

Table 5: The validation of our frequency-domain data obtained by FPDec on UCF-101. "Pre" means preprocessing time(ms) for decoding videos; "Frames" means RGB frames obtaining by complete decompression; "DCT" means frequency-domain data obtained by DCT; "Partial" means frequency-domain data obtained by FPDec.

FCS and video data augmentation can significantly improve the accuracy on both datasets, in which FCS can improve the accuracy of the R-network by more than 20% on HMDB-51.

To validate the effectiveness of frequency-domain data by FPDec, we compare our method with CoViAR [33] and DCT coefficients converted by manual DCT. The frequency-domain data obtained by either FPDec or DCT is effective input for action recognition, but the former is 165 times faster, as shown in Table 5. Compared with CoViAR, our method has higher accuracy, and the preprocessing speed is 1.5 times faster. The accuracy of our MV-network is higher than CoViAR, because even though our MV-network has the same size input of 224×224 as frame-based networks, our FCS down-samples the 16×16 blocks to 8×8 of high-resolution videos which contain more motion cues. Moreover, the fusion of I, R and MV networks can improve the final accuracy.

The analysis for the hyperparameters T and λ in S2FNet is shown in Table 3. It shows that when T is set to 2 and λ is set to 0.1, the S2FNet achieves the best performance. We follow these settings in the subsequent experiments. Experiments showing the effectiveness of S2FNet are conducted in Table 6. For Resnet-50 as backbone (I-net), frequency-domain I-network is improved by 3.9% and 4.7% on UCF-101 and HMDB-51, and the Faster-FCoViAR is improved by 3.4% and 5.3% on UCF-101 and HMDB-51; for Resnet-18 as backbone (I-net$_{Lite}$), frequency-domain I-network is improved by 6.2% and 3.5% on UCF-101 and HMDB-51, and the Faster-FCoViAR$_{Lite}$ is improved by 3.8% and 7.1% on UCF-101 and HMDB-51. It indicates that the joint learning of spatial and frequency-domain semantics can improve the results.

4.2 Speed and Efficiency

We compare the efficiency and accuracy of our method with TSN-based methods, including CoViAR [33], Fast-CoViAR [4], Multi-teacher [35], and TSN [30] in Table 7. Our method has few parameters and small $GFLOPs \times Views$ while reaching a higher accuracy of 91.2% and 63.2% on UCF-101 and HMDB-51. For the preprocessing time and VPS, our method is also much faster. As shown in Table 7, Faster-FCoViAR achieves higher accuracy than

Dataset	S2FNet	Backbone	I	Fusion
UCF-101	✗	Res50	73.9	87.8
	✓	Res50	77.8	91.2
	✗	Res18	64.1	86.0
	✓	Res18	70.3	89.8
HMDB-51	✗	Res50	40.0	56.9
	✓	Res50	44.7	63.2
	✗	Res18	34.3	51.3
	✓	Res18	37.8	58.4

Table 6: Ablation study of S2FNet. When the backbone of student I-network is Resnet-50, "Fusion" means Faster-FCoViAR; when the backbone of student I-network is Resnet-18, "Fusion" means Faster-FCoViAR$_{Lite}$.

Method	$G \times V$	Params	Pre	VPS	I	MV	R	Fusion
TSN (RGB-only) [30]	33.0×25	58.3	24.5	2.2	-	-	-	87.7
TSN (OF) [30]	66.2×25	116.6	>100	<0.1	-	-	-	94.0
CoViAR [33] (our impl.)	15.1×25	80.8	7.23	4.0	84.8	67.1	77.6	89.6
Multi-teacher [35]	5.5×25	35.0	7.23	7.5	84.2	70.8	83.9	88.5
Fast-CoViAR [4]	5.7×250	37.4	-	-	78.8	67.6	-	85.5
Faster-FCoViAR$_{Lite}$	4.7×25	33.6	4.71	42.4	70.3	81.6	63.7	89.8
Faster-FCoViAR (I+MV)	5.8×25	34.9	3.16	37.6	77.8	81.6	-	90.2
Faster-FCoViAR (I+MV+R)	9.6×25	58.6	4.71	26.9	77.8	81.6	75.4	91.2

Table 7: Comparison of speed and accuracy with TSN-based methods on UCF-101. "$G \times V$" means $GFLOPs \times Views$; "Params" means the number of trainable parameters(M); "Pre" means preprocessing time (ms) for decoding videos; VPS means videos per second in inference. For the approaches in this table except Fast-CoViAR, we consider one crop per sample to calculate preprocessing time and VPS. The results of CoViAR are the reproduced results on our device and the results of other methods are reported by their papers.

CoViAR trained in the spatial domain, illustrating its effectiveness in learning spatial and frequency semantics.

4.3 Compare with State-of-art

We finally compare our Faster-FCoViAR network with state-of-the-art methods. To make a fair comparison, we compare the efficiency at the video level by VPS. Table 8 compares our method with RGB frame-based methods without explicit optical flow, RGB frame-based methods with explicit optical flow, and compressed domain methods. Our method achieves a comparable accuracy of 91.2% and 63.2% on UCF-101 and HMDB-51 with a high VPS

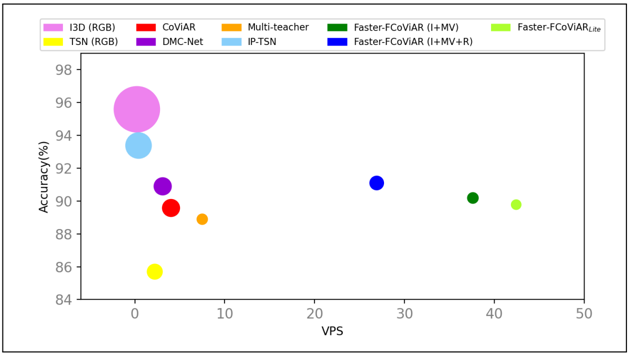

Figure 5: Comparison of the classification accuracy (%) and VPS for Faster-FCoViAR and other state-of-art methods on UCF-101. Node size denotes the GFLOPs.

Methods	VPS	GFLOPs	UCF-101	HMDB-51	Kinetics-400
RGB frame-based methods (without explicit optical flow)					
TSN (RGB-only) [30]	2.2	33.0	87.7	51.0	69.1
I3D (RGB-only) [1]	0.2	107.9	95.6	74.8	72.1
Res3D [11]	0.2	36.7	85.8	54.9	65.1
ECO$_{en}$ [43]	3.6	-	94.8	72.4	70.0
TSM[19]	2.1	33.0	95.9	73.5	72.5
TEA[17]	2.0	35.0	96.9	73.3	**75.0**
RGB frame-based methods (with explicit optical flow)					
Two-Stream (OF) [23]	<0.1	-	88.0	59.4	-
TSN (OF) [30]	<0.1	66.2	94.0	68.5	73.9
R(2+1)D [28]	<0.1	65.5	<u>97.3</u>	<u>78.7</u>	<u>74.9</u>
I3D (OF) [1]	<0.1	215.8	**98.0**	**80.7**	74.3
Compressed domain methods					
EMV-CNN [39]	2.3	-	86.4	51.2	-
DTMV-CNN [40]	2.3	-	87.5	55.3	-
CoViAR [33] (our impl.)	4.0	15.1	89.6	59.1	-
DMC-Net [22]	3.1	15.3	90.9	61.8	-
IP-TSN [13]	0.4	34.0	93.4	69.1	-
Multi-teacher [35]	7.5	5.5	88.9	56.2	-
MVR-AR [42]	2.5	-	92.1	-	-
Fast-CoViAR [4]	-	5.7	85.5	55.8	-
Faster-FCoViAR$_{Lite}$	**42.4**	4.7	89.8	58.4	-
Faster-FCoViAR (I+MV)	<u>37.6</u>	5.8	90.2	57.6	-
Faster-FCoViAR (I+MV+R)	26.9	9.6	91.2	63.2	69.3

Table 8: Compare with state-of-art. "OF" means optical flow.

of 26.9, which is 6.7 times faster than CoViAR. For Kinetics-400, our Faster-FCoViAR achieve a preliminary accuracy of 69.3% with the highest speed, and there are no reports of other compressed domain method on Kinetics-400. Moreover, Our Faster-FCoViAR$_{Lite}$ reaches a VPS of 42.4, which is 5.7 times faster than Multi-teacher (its backbones of I, R, and MV are Resnet-18, the same with our Faster-FCoViAR$_{Lite}$) and have a higher accuracy. As shown in Table 8, methods with explicit optical flow have the highest accuracy, but their speed is rather slow. In other RGB frame-based methods, 3D network methods have high accuracy, but the inference speed is also slow. In the compressed domain methods, our Faster-FCoViAR achieves a competitive accuracy with the highest speed. Figure 5 compares the classification accuracy, GFLOPs, and VPS of our Faster-FCoViAR with other methods.

5 Conclusion

In this paper, a faster frequency-domain compressed video action recognition framework is proposed (Faster-FCoViAR). In particular, the proposed frequency-domain partial decompression method (FPDec) can directly obtain the frequency-domain DCT coefficients from compressed videos. Then, we design a frequency-domain channel selection strategy (FCS), to enhance the saliency of input. To further improve the learning of spatial-frequency semantic for the frequency-domain network, we propose a spatial-to-frequency-domain student-teacher network (S2FNet). Experiments show that our Faster-FCoViAR framework is 12.3 times faster than the RGB frame-based methods and 6.7 times faster than the compressed domain methods. Our Faster-FCoViAR achieves a comparable accuracy of 91.2%, 63.2%, and 69.3% on UCF-101, HMDB-51, and Kinetics-400 respectively and higher efficiency than any other state-of-the-art methods.

6 Acknowledgements

The work presented in this paper was partly supported by Natural Science Foundation of China (Grant No. 62076030), Beijing Natural Science Foundation of China (Grant No. L201023), together with ZTE Corporation.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Biplab Ketan Chakraborty, Debajit Sarma, Manas Kamal Bhuyan, and Karl F MacDorman. Review of constraints on vision-based gesture recognition for human–computer interaction. *IET Computer Vision*, 12(1):3–15, 2018.

[3] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[4] Samuel Felipe dos Santos and Jurandy Almeida. Faster and accurate compressed video action recognition straight from the frequency domain. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 62–68. IEEE, 2020.

[5] Samuel Felipe dos Santos, Nicu Sebe, and Jurandy Almeida. The good, the bad, and the ugly: Neural networks straight from jpeg. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1896–1900. IEEE, 2020.

[6] Max Ehrlich and Larry S Davis. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3484–3493, 2019.

[7] Linxi Fan, Shyamal Buch, Guanzhi Wang, Ryan Cao, Yuke Zhu, Juan Carlos Niebles, and Li Fei-Fei. Rubiksnet: Learnable 3d-shift for efficient video action recognition. In *European Conference on Computer Vision*, pages 505–521. Springer, 2020.

[8] Yuming Fang and Xiaoqiang Zhang. Visual attention modeling in compressed domain: From image saliency detection to video saliency detection. *ZTE COMMUNICATIONS*, 17(1), 2019.

[9] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems*, 31:3933–3944, 2018.

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

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

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Shiyuan Huang, Xudong Lin, Svebor Karaman, and Shih-Fu Chang. Flow-distilled ip two-stream networks for compressed video action recognition. *arXiv preprint arXiv:1912.04462*, 2019.

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[17] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020.

[18] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[20] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[21] Iain E Richardson. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.

[22] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1268–1277, 2019.

[23] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[25] G Sreenu and MA Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):1–27, 2019.

[26] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020.

[27] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[28] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[29] Matej Ulicny and Rozenn Dahyot. On using cnn with dct based image data. In *Proceedings of the 19th Irish Machine Vision and Image Processing conference IMVIP*, volume 2, 2017.

[30] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[31] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, June 2021.

[32] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13214–13223, 2021.

[33] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6026–6035, 2018.

[34] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[35] Meng-Chieh Wu and Ching-Te Chiu. Multi-teacher knowledge distillation for compressed video action recognition based on deep learning. *Journal of Systems Architecture*, 103:101695, 2020.

[36] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*, pages 305–321, 2018.

[37] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020.

[38] Ke Yang, Peng Qiao, Dongsheng Li, and Yong Dou. If-ttn: Information fused temporal transformation network for video action recognition. *arXiv preprint arXiv:1902.09928*, 2019.

[39] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2718–2726, 2016.

[40] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018.

[41] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R. Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. In *International Conference on Learning Representations*, 2020.

[42] Chenghui Zhou, Xiaolei Chen, Pei Sun, Guanwen Zhang, and Wei Zhou. Compressed video action recognition using motion vector representation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 701–713. Springer International Publishing, 2021.

[43] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision*, pages 695–712, 2018.