

End-to-End Compressed Video Representation Learning for Generic Event Boundary Detection

Congcong Li^{1*}, Xinyao Wang², Longyin Wen², Dexiang Hong^{1*}, Tiejian Luo¹, Libo Zhang^{3†}

¹University of Chinese Academy of Sciences, Beijing, China

²ByteDance Inc., Mountain View, USA

³Institute of Software Chinese Academy of Sciences, Beijing, China

licongcong18@mails.ucas.edu.cn, {xinyao.wang, longyin.wen}@bytedance.com

hongdexiang19@mails.ucas.edu.cn, tjluo@ucas.ac.cn, libo@iscas.ac.cn

Abstract

Generic event boundary detection aims to localize the generic, taxonomy-free event boundaries that segment videos into chunks. Existing methods typically require video frames to be decoded before feeding into the network, which demands considerable computational power and storage space. To that end, we propose a new end-to-end compressed video representation learning for event boundary detection that leverages the rich information in the compressed domain, i.e., RGB, motion vectors, residuals, and the internal group of pictures (GOP) structure, without fully decoding the video. Specifically, we first use the ConvNets to extract features of the I-frames in the GOPs. After that, a light-weight spatial-channel compressed encoder is designed to compute the feature representations of the P-frames based on the motion vectors, residuals and representations of their dependent I-frames. A temporal contrastive module is proposed to determine the event boundaries of video sequences. To remedy the ambiguities of annotations and speed up the training process, we use the Gaussian kernel to preprocess the ground-truth event boundaries. Extensive experiments conducted on the Kinetics-GEBD dataset demonstrate that the proposed method achieves comparable results to the state-of-the-art methods with $4.5\times$ faster running speed.

1. Introduction

Video traffic will account for 82% percent of all internet traffic by 2022, up from 75% in 2017 [11]. Understanding video content using AI technology is an active area of research in recent years. However, it is still a challenging task due to the complex temporal evolution in the enormous size

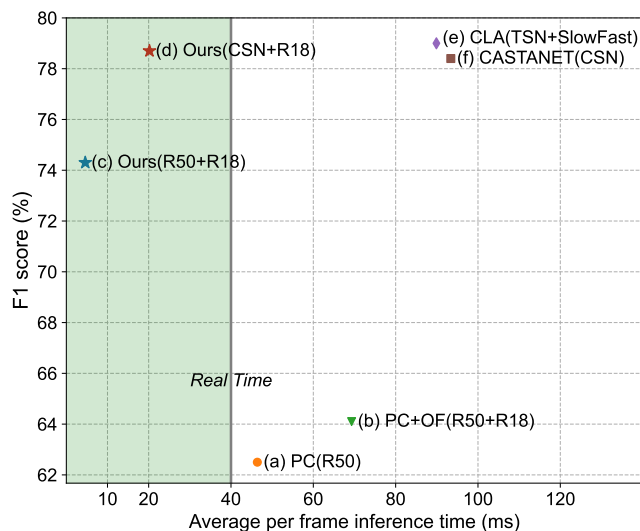


Figure 1. The inference time vs. F1 score of different methods on the Kinetics-GEBD dataset [28]. (a) The previous method [28] PC is relatively fast with inferior results. (b) After integrating the optical flow (OF) module, the accuracy is improved with much slower running speed. (c, d) Our method achieves competitive F1 score with extremely fast running speed by directly leveraging motion vectors and residuals in the compressed domain. (e, f) CLA [18] and CASTANET [14] take fully decoded RGB frames as input, which are much slower than the methods conducted in compressed domain. The green region indicates the methods run in real-time.

of raw video streams with high temporal redundancy.

Video understanding is one of the most fundamental problems in computer vision, which includes video tagging, action recognition, and video boundary detection, etc. In contrast to static images, videos provide rich information involving temporal consistency in consecutive frames which can be additionally utilized. Currently, the two-stream network [8, 9, 30] and 3D convolutional network [17, 33, 34, 37] are two popular network architectures in the video under-

*This work was done during internships at ByteDance Inc.

†Corresponding author (libo@iscas.ac.cn)

standing field. The two-stream network incorporates both the decoded RGB video frames and optical flow to exploit temporal information. However, extracting optical flow is very slow, which dominates the overall pre-processing time in the video understanding tasks. 3D convolutional network is another choice to model temporal information using the spatio-temporal filters. The drawbacks of 3D convolutional network is the massive parameters contained in 3D convolution operations, which slows down the inference speed. Besides the aforementioned methods, the new trend in video understanding is using the transformers, including [1, 5, 6, 24, 51], achieving competitive results.

In recent years, several methods [16, 29, 42, 45, 47, 49] demonstrate the advantages of directly taking videos in compressed domain as input for video understanding. These methods use motion vectors and residuals in the compressed representation that developed for storage and transmission of videos rather than operating on the decoded RGB frames, which run in two orders of magnitude faster than the methods using optical flow while achieving competitive results [29]. Specifically, these methods use the almost compute-free motion vectors and residuals encoded in P-frames as an alternative to the compute-intensive optical flow. For example, CoViAR [45] directly feeds motion vectors and residuals into 2D CNNs for action recognition, and DMC-Net [29] improves the CoViAR method by reconstructing the optical flow based on motion vectors and residuals. Although the aforementioned method achieves promising results, they are still far from satisfactory, which lack effective fusion strategies between different modalities, such as decoded I-frames, motion vectors, and residuals.

In this paper, we focus on the generic event boundary detection (GEBD [28]) task that aims to localize the moments where humans naturally perceive taxonomy-free event boundaries that segment a longer event into shorter temporal segments. The ability to divide a long form video into small meaningful clips makes this task demanding for several downstream video understanding tasks and industry applications that requires high accuracy and low latency. The previous attempt [28] formulate it as a classification task by considering the context information of the candidate boundaries. However, it neglects the temporal relations between consecutive frames and operates inefficiently during feature extraction stage. Inspired by [16, 29, 42, 45, 47, 49], we design an end-to-end trained network to exploit the discriminative features for GEBD in compressed domain, *i.e.*, MPEG-4, which is able to save decoding cost and improve feature extraction efficiency. Specifically, most modern codecs split a video into several group of pictures (GOP), where each GOP is formed by one I-frames and T P-frames. To solve difficulty arised from the long chain of dependency of the P-frames, inspired by [45], we use the back-tracing technique to compute the accumulated motion vectors and

residuals in linear time. In this way, the consecutive P-frames in each GOP are only depending on the reference I-frame, which can be processed in parallel.

In contrast to the I-frame, it is difficult to learn the discriminative features of the P-frames. Refining the features of the reference I-frame based on the motion vectors and residuals becomes an intuitive option. Motion vectors and residuals provide information to reconstruct P-frames by referring the dependent I-frames. In addition to that, they also provide motion information that obtained from the video encoding process. To that end, we design a light-weight spatial-channel compressed encoder to refine the features of the reference I-frame with the guidance of the motion vectors and residuals. In this way, the features of P-frames and I-frames are converted to the same feature space, which benefits the subsequent processing. After that, a temporal contrastive module is proposed to capture the context information in temporal domain to predict the event boundaries of videos. Notably, our temporal contrastive module imitates humans, *i.e.*, look back and forth around the candidate frames to determine event boundaries, by comparing the extracted features before and after the candidate frames. In addition, to remedy the ambiguities of annotations and speed up the training process, we use the Gaussian kernel to preprocess the ground-truth event boundaries instead of using the “hard lables” of boundaries. Extensive experiments conducted on the Kinetics-GEBD dataset to demonstrate the effectiveness of the proposed method. Specifically, the proposed method achieves comparable results to the state-of-the-art method at the CVPR’21 LOVEU Challenge [18] with $4.5\times$ faster running speed, see Figure 1.

The main contributions of this paper are listed as follows.

- (1) We propose an end-to-end compressed video representation learning method to solve the challenging GEBD task.
- (2) We design the spatial-channel compressed encoder to project the features of reference I-frame with the guidance of motion vectors and residuals to compute the features of P-frames with low cost.
- (3) A temporal contrastive module is proposed to determine the event boundaries of videos by exploiting the context information in temporal domain.
- (4) The proposed method achieves comparable results to the state-of-the-art methods at the CVPR’21 LOVEU Challenge [18] with $4.5\times$ faster running speed, demonstrating its effectiveness.

2. Related Work

Video recognition. Over the last decade, video recognition has achieved great progress thanks to the emergence of deep learning. Early methods [20, 26, 38, 39] use hand-crafted features for video recognition. After the arriving of deep learning, the video recognition field is quickly dominated by the CNN-based methods, such as the two-stream network and 3D convolutional network. The two-stream network based methods [8, 9, 30] use additional temporal

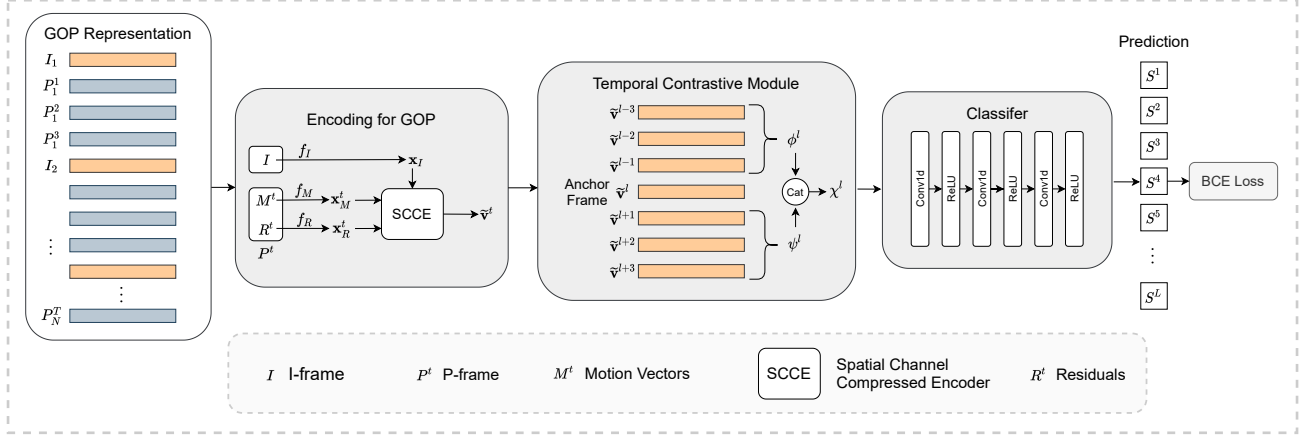


Figure 2. The architecture of the proposed method. The spatial-channel compressed encoder (SCCE) is designed to obtain the refined P-frame representation \tilde{v}^t based on reference I-frame feature x_I , motion vectors M^t and residuals R^t . This module regards each GOP as a process unit, which is efficient and can be paralleled in a large batch size. Then we use temporal contrastive module to capture temporal dependence explicitly based on unified representation \tilde{v}^t , which provides strong cues for boundary detection. After that, a simple classifier is used to make final predictions trained with the Gaussian smoothed soft labels.

stream to learn motion information and design various fusion strategies to combine the information from the image stream and temporal stream, achieving superior results. The optical flow is generally used to describe the motion information, which is computationally expensive. Some other methods [17, 33, 34, 37] attempt to use the 3D convolutional network with the spatio-temporal filters to integrate temporal information. However, these methods are hard to optimize and require large-scale datasets in the training phase. The new trend in recent years is the introduction of transformers [1, 5, 6, 24, 51], which achieving promising results on various datasets in video understanding.

Meanwhile, some recent methods attempt to directly take the raw compressed videos as input for different tasks in the video understanding field, such as action recognition [16, 29, 45, 47, 49], object detection [42], and video segmentation [10]. The aforementioned methods use motion vectors and residuals directly obtained from the compressed videos as the alternatives to optical flow and achieve comparable results in terms of both the speed and accuracy.

Generic event boundary detection. Generic event boundary detection (GEBD) [28] aims to localize the moments where humans naturally perceive taxonomy-free event boundaries that break a longer event into shorter temporal segments. The previous method [28] takes 5 video frames before and after the candidate boundaries as input, and separately determines whether each candidate is the event boundary or not. Kang *et al.* [18] propose to use the temporal self-similarity matrix (TSM) as the intermediate representation and use the popular contrastive learning method to exploit the discriminative features for better performance. Hong *et al.* [14] use the cascade classification heads and dynamic sampling strategy to boost both recall and precision.

Meanwhile, Rai *et al.* [27] attempt to learn the spatiotemporal features using a two stream inflated 3D convolutions architecture. To the best of our knowledge, there do not exist any prior work focuses on the GEBD task in the compressed domain.

Attention mechanism. To learn more discriminative features, numerous methods have been proposed, which mainly focus on enhancing the feature representations using the attention mechanisms on the spatial or(and) channel dimensions. SENet [15] develops the “Squeeze-and-Excitation” (SE) block that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. Non-local network [43] capture long-range dependencies by computing the response at a position as a weighted sum of the features at all positions in the input feature maps. SKNet [22] proposes to adaptively adjust the receptive field size of the input feature map by fusing multiple feature maps of different kernel sizes with softmax attention in a weighted manner. CBAM [44] sequentially infers attention maps along both channel and spatial dimensions, and then uses the attention maps to recalibrate the origin input feature. In contrast to the aforementioned methods, we attempt to refine the P-frame feature with the guidance of motion vectors and residuals by considering both spatial and channel dimensions of the features of I-frame, which fully leverages the information of the decoded reference I-frame to enrich the features of P-frames.

3. Method

The existing method [28] formulates the GEBD task as binary classification, which predicts the boundary labels of each frame by considering the temporal contextual information. That is, the preceding and succeeding frames of each video frame are feed into a neural network to detect

the boundaries. It is inefficient due to the duplicated computation is conducted of consecutive frames. To remedy this, we propose an end-to-end compressed video representation method for GEBD, which regards each video clip as a whole. Specifically, we use MPEG-4 encoded videos as our input. Each video clip \mathcal{V} is formed by N groups of pictures (GOPs), and each GOP contains one I-frame and T P-frames, *i.e.*,

$$\mathcal{V} = \{I_i, P_i^1, P_i^2, \dots, P_i^T\}_{i=1}^N, \quad (1)$$

where $I_i \in \mathbb{R}^{3 \times \mathcal{H} \times \mathcal{W}}$ denotes the reference I-frame and P_i^t denotes the t -th P-frame of the i -th GOP, and \mathcal{H} and \mathcal{W} are the height and width of the video frame. For simplicity, we assume that there exists the same number of P-frames in all GOPs. The P-frame P_i^t in the i -th GOP is formed by the motion vector $M_i^t \in \mathbb{R}^{2 \times \mathcal{H} \times \mathcal{W}}$ and residual $R_i^t \in \mathbb{R}^{3 \times \mathcal{H} \times \mathcal{W}}$, which can be obtained nearly cost-free from the compressed video stream. Notably, the motion vectors and residuals alone do not contain the full information of a P-frame. The P-frame depends on the reference I-frame or other P-frames, making it difficult to learn the discriminative feature representations for P-frames. Following [45], we trace all motion vectors back until to the reference I-frame and accumulate the residual on the way to decouple the dependencies between the consecutive P-frames. In this way, each P-frame only depends on the reference I-frame rather than other P-frames. After that, we build our model based on the back-traced motion vectors and residuals and regard each GOP as a process unit. The overall network architecture is presented in Figure 2. As shown in Figure 2, the GOP is first encoded by the designed spatial-channel compressed encoder (SCCE) to generate the unified video representation. After that, a temporal contrastive module is used to exploit the temporal context information to get the discriminative feature representations. Finally, a classifier is used to generate the accurate event boundaries.

3.1. Spatial-Channel Compressed Encoder

Motion, uncovered regions, and lighting variations frequently happen in video sequences. Modern codecs use macroblock as the basic unit for motion compensated prediction in a number of mainstream visual coding standards such as MPEG-4, H.263, and H.264. Motion vectors record the moving direction of each macroblock with respect to its reference frame(s), describing the motion patterns of videos, which is important for the GEBD task. The residuals can be regarded as the compensations of the motion information, which contains the boundary information of moving objects and plays a crucial role to identify the important regions in the I-frame. Thus, we propose to apply the attention mechanism to different regions of I-frame with the guidance of motion vectors to enrich the features by con-

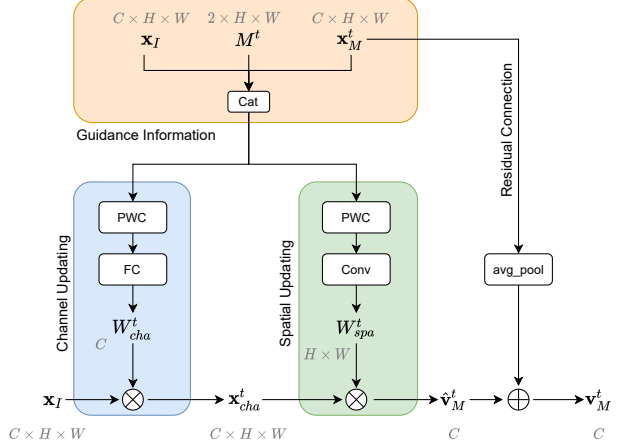


Figure 3. The architecture of the proposed spatial-channel compressed encoder (SCCE) module. We concatenate the features of I-frame \mathbf{x}_I , motion vectors M^t and the features of motion vectors \mathbf{x}_M^t to modulate the features of the reference I-frame \mathbf{x}_I in both channel and spatial dimensions. After that, the modulated features $\hat{\mathbf{v}}_M^t$ is residually added with the features of motion vectors to obtain the refined vector representations \mathbf{v}_M^t .

sidering both channel and spatial dimensions. For simplicity, we omit the index i of the GOP in the following sections.

Firstly, we use the convolutional neural network taking the decoded RGB image as input to extract the feature representation \mathbf{x}_I of the I-frame I , *i.e.*, $\mathbf{x}_I = f_I(I)$, where $\mathbf{x}_I \in \mathbb{R}^{C \times H \times W}$ is the features of the I-frame I , and C , H and W are the channel, height and width of the features \mathbf{x}_I , respectively. $f_I(\cdot)$ denotes the model used to extract features for I-frame, which is pretrained on the large-scale datasets (*e.g.*, ResNet50 pretrained on ImageNet). Meanwhile, we can similarly compute the features for the P-frames $\{P^1, P^2, \dots, P^T\}$ with a more lightweight model than the one used for I-frame, by directly taking the motion vectors M^t and residuals R^t , *i.e.*, $\mathbf{x}_M^t = f_M(M^t)$, and $\mathbf{x}_R^t = f_R(R^t)$ as input, where $\mathbf{x}_M^t, \mathbf{x}_R^t \in \mathbb{R}^{C \times H \times W}$ denote the features of the motion vectors and residuals, respectively. In this way, a considerable amount of time can be saved on extracting features for the P-frames. This simple strategy can only bring limited performance gain [45]. The method [29] attempts to integrate the optical flow in training phase, which can further improve the accuracy. However, there is still much room for improvement of the aforementioned methods. Specifically, the motion vectors record the motion patterns of both the scenes and objects in videos, and the residuals provide the compensation information. Both of them do not contain the context information of scenes. To that end, we design the spatial-channel compressed encoder module by integrating the features of the reference I-frame \mathbf{x}_I in computing the features of P-frames.

We first compute the features \mathbf{x}_M^t of the motion vectors by refining the features of the reference I-frame \mathbf{x}_I in both

the channel and spatial dimensions. As indicated by [48], different regions on the feature maps focus on different parts of images. Thus, we introduce the attention weight for each feature map of \mathbf{x}_I based on the information of the P-frame \mathbf{x}_M^t . Specifically, we concatenate I-frame feature \mathbf{x}_I , motion vector feature \mathbf{x}_M^t and motion vectors M^t together in the channel dimension to compute the channel weight W_{cha}^t using a lightweight PWC-Net [32], *i.e.*,

$$\begin{aligned} W_{\text{cha}}^t &= \sigma(W_2 \cdot \zeta(W_1 h_{\text{cha}}^t + b_1) + b_2) \\ \mathbf{h}_{\text{cha}}^t &= \text{AvgPool}(\mathbf{z}_{\text{cha}}^t) \\ \mathbf{z}_{\text{cha}}^t &= \text{PWC}([\mathbf{x}_I; \mathbf{x}_M^t; M^t]) \end{aligned} \quad (2)$$

where σ is the sigmoid function, ζ is the ReLU function, and W_1, b_1, W_2, b_2 are the learnable weights of the FC layers. After that, the features of the I-frame \mathbf{x}_I are updated based on W_{cha}^t as follows.

$$\mathbf{x}_{\text{cha}}^t = \mathbf{x}_I \otimes W_{\text{cha}}^t, \quad (3)$$

where \otimes is the channel-wise multiplication. In this way, we can compute the channel-weighted feature $\mathbf{x}_{\text{cha}}^t$ by updating \mathbf{x}_I in channel dimension, with the guidance of the motion vectors. Meanwhile, the channel-weighted feature $\mathbf{x}_{\text{cha}}^t$ is further updated in the spatial dimension and the spatial dimension is reduced. That is, given the features \mathbf{x}_I of the reference I-frame, motion vector features \mathbf{x}_M^t and motion vectors M^t , we compute the 2D weight map W_{spa}^t , *i.e.*,

$$\begin{aligned} W_{\text{spa}}^t &= \text{softmax}(\mathbf{h}_{\text{spa}}^t) \\ \mathbf{h}_{\text{spa}}^t &= \text{2DConv}(\mathbf{z}_{\text{spa}}^t) \\ \mathbf{z}_{\text{spa}}^t &= \text{PWC}([\mathbf{x}_I; \mathbf{x}_M^t; M^t]) \end{aligned} \quad (4)$$

where $W_{\text{spa}}^t \in \mathbb{R}^{H \times W}$ is the spatial weight map. After that, we use W_{spa}^t to weight the features $\mathbf{x}_{\text{cha}}^t$ in the spatial dimension to compute the enriched features of the motion vectors $\hat{\mathbf{v}}_M^t \in \mathbb{R}^C$, *i.e.*,

$$\hat{\mathbf{v}}_M^t = \sum_{p=1}^{H \cdot W} \mathbf{x}_{\text{cha}}^t \cdot W_{\text{spa}}^t, \quad (5)$$

where p enumerates all spatial positions of $\mathbf{x}_{\text{cha}}^t \cdot W_{\text{spa}}^t$. Finally, we add $\hat{\mathbf{v}}_M^t$ to the original features \mathbf{x}_M^t of the P-frame to obtain the refined features of the motion vectors $\mathbf{v}_M^t \in \mathbb{R}^C$, *i.e.*,

$$\mathbf{v}_M^t = \hat{\mathbf{v}}_M^t + \text{AvgPool}(\mathbf{x}_M^t). \quad (6)$$

The overall computing process of \mathbf{v}_M^t is presented in Figure 3. Similarly, we can compute the refined features for the residuals $\mathbf{v}_R^t \in \mathbb{R}^C$. The final feature representations for the P-frame is further computed as

$$\tilde{\mathbf{v}}^t = \mathbf{v}_M^t + \mathbf{v}_R^t. \quad (7)$$

In this way, we can compute the features of the P-frames $\{\tilde{\mathbf{v}}^1, \tilde{\mathbf{v}}^2, \dots, \tilde{\mathbf{v}}^T\}$ in the GOP by considering the reference

I-frame I in both channel and spatial dimensions. The overall process is very efficient and can be processed in parallel in GOPs. After extracting the discriminative features for both the I-frames and P-frames in the same feature space, we can predict the event boundaries efficiently and accurately.

3.2. Temporal Contrastive Module

Based on the extracted features of the video \mathcal{V} , we design the temporal contrastive module to predict the event boundaries. Inspired by humans, *i.e.*, look back and forth around the candidate boundary frames to determine event boundaries, we compute the contrastive features before and after the candidate boundary frames in the temporal domain. Specifically, given feature representations $\{\tilde{\mathbf{v}}^{l-k}, \tilde{\mathbf{v}}^{l-(k-1)}, \dots, \tilde{\mathbf{v}}^{l-1}\}$ before k frames of candidate boundary frame l , we compute the left features ϕ^l of the candidate boundary frame l using the simple linear weighted summation strategy, *i.e.*,

$$\phi^l = \sum_{j=1}^k W_j \cdot \tilde{\mathbf{v}}^{l-j}, \quad (8)$$

where $W_j \in \mathbb{R}^C$ is the learnable weights and shared at different position l . The simple linear weighted summation can be efficiently implemented using the 1D convolutional operation. Meanwhile, the right features ψ^l can be similarly computed, *i.e.*, weighted summing the feature representations of the k features after the candidate boundary frame l . After that, the contrastive feature χ^l is computed as the concatenation of ϕ^l and ψ^l , *i.e.*, $\chi^l = [\phi^l; \psi^l]$. Then for the final classification, we use the contrastive representations $\{\chi^1, \chi^2, \dots, \chi^L\}$ to make the event boundary predictions.

3.3. Loss Function

Given the feature representations $\{\chi^1, \chi^2, \dots, \chi^L\}$ of each video frame and the corresponding ground-truth labels, the event boundary detection task is intuitively formulated as the binary classification task. However, the ambiguities of annotations disrupt the learning process, which leads to poor convergence. To solve this issue, we use the Gaussian kernel to preprocess the ground-truth event boundaries to obtain the soft labels instead of using the “hard labels” of boundaries. Specifically, for each annotated boundary, the intermediate label of the neighboring position i is computed as:

$$g_i^l = \exp\left(-\frac{(l-i)^2}{2\alpha^2}\right) \quad (9)$$

where g_i^l indicates the intermediate label at time i corresponding to the annotated boundaries at time l . We set $\alpha = 1$ in all our experiments. The final soft labels are computed as the summation of all intermediate labels. Finally, a simple nonlinear Conv1D classifier is applied to predict the boundary score S^l and the binary cross-entropy loss is used to guide the training process.

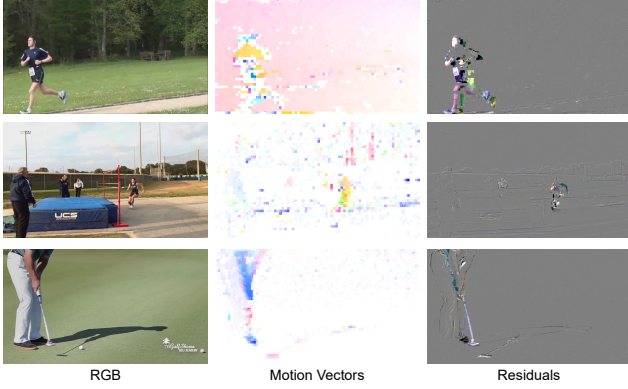


Figure 4. Visualization of the compressed information. The decoded RGB frames, motion vectors and residuals are presented in different columns. Best view in color.

4. Experiments

Implementation detail. ResNet50 and ResNet18 [13] pre-trained on ImageNet [4] are used to extract the features for I-frames and P-frames in all experiments if not particularly indicated. Our method is implemented based on the MPEG-4 Part 2 specifications [12], where each GOP contains 1 I-frame and 11 P-frames. We sample 3 P-frames in each GOP to reduce the redundancy, *i.e.*, $T = 3$ in (1). We use the standard SGD with momentum set to 0.9, weight decay set to 10^{-4} , and learning rate set to 10^{-2} . We set the batch size to 4 for each GPU and train the network on 8 NVIDIA Tesla V100 GPUs, resulting in a total batch size of 32. The network is trained for 30 epochs with a learning rate drop by a factor of 10 after 16 epochs and 24 epochs, respectively. We test the running speed of all methods on 1 NVIDIA Tesla V100 GPU. All the source code of our method will be made publicly available after the paper is accepted.

Datasets. We conduct our experiments on the Kinetics-GEBD dataset [28], which contains the largest number of temporal boundaries. The Kinetics-GEBD dataset includes 54691 videos and 1,290,000 event boundaries, spans a broad spectrum of video domains in the wild and is open-vocabulary rather than building on a pre-defined taxonomy. Besides, to verify the generality and effectiveness of our method, we also conduct experiments on the popular action recognition datasets UCF101 [31] and HMDB51 [19]. UCF101 consists of 101 action classes over 13,320 videos and HMDB51 contains 51 distinct action categories with a total of 6,766 video clips.

4.1. Discussion

Kinetics-GEBD. We first train and evaluate the proposed method on the Kinetics-GEBD [28] train-validation split. The evaluation protocol presented in [28] uses Relative Distance (*i.e.*, **Rel.Dis.**, the error between the predicted and ground truth timestamps) to determine whether a prediction

Table 1. Accuracy on the HMDB-51 and UCF-101 datasets for both decoded video based methods and compressed video based methods. Our spatial-channel compressed encoder (SCCE) performs favorably against the state-of-the-art compressed video based methods.

	HMDB-51	UCF-101
Decoded video based methods (RGB only)		
ResNet-50 [13]	48.9	82.3
ResNet-152 [13]	46.7	83.4
ActionFlowNet (2-frames) [25]	42.6	71.0
ActionFlowNet [25]	56.4	83.9
PWC-Net (ResNet-18) + CoViAR [32]	62.2	90.6
TVNet [7]	71.0	94.5
C3D [34]	51.6	82.3
Res3D [35]	54.9	85.8
ARTNet [40]	70.9	94.3
MF-Net [3]	74.6	96.0
S3D [46]	75.9	96.8
I3D RGB [2]	74.8	95.6
Compressed video based methods		
EMV-CNN [49]	51.2 (split1)	86.4
DTMV-CNN [50]	55.3	87.5
CoViAR [45]	59.1	90.4
DMC-Net(ResNet-18) [29]	62.8	90.9
DMC-Net(I3D) [29]	71.8	92.3
Ours (ResNet-18)	63.3	91.0
Ours (I3D)	72.1	92.5

is correct or not and then use the precision, recall, and F1 scores as the evaluation metrics. The results are shown in Table 2. Compared to the previous method PC [28], our method achieves 11.8% absolute improvement while running $10\times$ faster. Meanwhile, we also add an additional optical flow input stream to PC. A slight improvement is observed after integrating optical flow, which indicates that the motion information (*i.e.*, optical flow) alone can only provide limit temporal information for the generic event boundary detection task. Using motion vectors and residuals, our method provides more information features of the compressed P-frames by considering both the spatial and channel dimensions. The performance gap between the PC with optical flow and the proposed method demonstrates that the proposed method provides strong temporal cues for GEBD explicitly.

UCF101 and HMDB51. To validate the effectiveness of our method, we also conduct experiments on the action recognition datasets UCF-101 and HMDB-51. We follow the same settings as CoViAR [45] except that we use spatial-channel compressed encoder to process the motion vectors and residuals. Note that our temporal contrastive module is designed to capture temporal dependency, which is more suitable for event boundary detection. Thus, it is not applied in the action recognition task. The results are shown in Table 1. Our method achieves competitive results comparing with the state-of-the-art methods in com-

Table 2. The evaluation results on the Kinetics-GEBD validation set with different Rel.Dis. threshold. Our method improves the F1 score over all thresholds by a large margin.

Rel.Dis. Threshold	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg
BMN [23]	0.186	0.204	0.213	0.220	0.226	0.230	0.233	0.237	0.239	0.241	0.223
BMN-StartEnd [28]	0.491	0.589	0.627	0.648	0.660	0.668	0.674	0.678	0.681	0.683	0.640
TCN-TAPOS [28]	0.464	0.560	0.602	0.628	0.645	0.659	0.669	0.676	0.682	0.687	0.627
TCN [21]	0.588	0.657	0.679	0.691	0.698	0.703	0.706	0.708	0.710	0.712	0.685
PC [28]	0.625	0.758	0.804	0.829	0.844	0.853	0.859	0.864	0.867	0.870	0.817
PC + Optical Flow	0.646	0.776	0.818	0.842	0.856	0.864	0.868	0.874	0.877	0.879	0.830
Ours	0.743	0.830	0.857	0.872	0.880	0.886	0.890	0.893	0.896	0.898	0.865

pressed domain, *i.e.*, EMV-CNN [49], DTMV-CNN [50], CoViAR [45] and DMC-Net [29]. We believe that this is because our method is able to generate more discriminative P-frame representations with the help of the proposed SCCE module. In contrast to other methods that process motion vectors and residuals in separate branches from the reference I-frame, we integrate the features of I-frame with the guidance of motion vectors and residuals on both spatial and channel dimensions. In this way, the rich information from the features of I-frame, motion vectors and residuals could be effectively fused together to generate high quality P-frame features with little overhead. It’s worth noting that DMC-Net [29] needs extra optical flow as supervision during training phase while our method can directly learn discriminative features with the spatial-channel compressed encoder.

4.2. Ablation Study

We conduct several ablation studies to demonstrate the effectiveness of different components in the proposed method. All experiments are conducted on the Kinetics-GEBD train split with ResNet50 backbone and tested on a local minval split to reduce the computation cost. The local minval split is constructed from the Kinetics-GEBD validation split by randomly sampling 2000 videos.

Table 3. The effectiveness of our proposed end-to-end architecture. “E2E” indicates end-to-end training strategy and “GS” indicates using soft labels generated by the Gaussian kernel strategy. To study the influence of these modules, we simply replace the inputs of the PC method by down-sampling each video into a succession of frames and use a Gaussian kernel to smooth the labels. This strategy improves the accuracy of the PC [28] method with a much faster running speed by reducing redundant computations.

	Rec	Prec	F1	Speed(ms)
PC [28]	0.611	0.631	0.621	46.4
+ E2E	0.629	0.640	0.634	9.3
+ GS	0.665	0.643	0.654	9.3

The end-to-end architecture. The previous PC method [28] formulate GEBD as the classification task, which feeds preceding and succeeding frames as inputs to provide temporal context information. To verify the effectiveness and the efficiency of the proposed end-to-end architecture, we

conduct experiments with the identical architecture as PC [28] except the feature inputs and target labels, shown in Table 3. Simply replacing the feature inputs of PC [28] with continuous video frames gives 1.3% absolute performance gain while increasing inference speed by a large margin, which indicates that sharing features between nearby frames is beneficial for the GEBD task. Besides, using the soft labels generated by Gaussian kernel provides further 2.0% absolute improvements. Using the ambiguous “hard labels” disrupt the learning process, which leads to poor convergence. Our soft label strategy effectively solve this issue and speeds up the training process.

Table 4. Ablation study of different compressed representation. “OF” indicates the optical flow and “Vanilla” indicates using the vanilla ResNet-18 to extract the features of motion vectors and residuals. We observe that both the methods improved from PC [28] and our method benefit from optical flow and motion vectors and residuals. The proposed spatial-channel compressed encoder (SCCE) module further improve the accuracy with similar running speed.

Method	Repre.	Rec	Prec	F1	Speed(ms)
PC [28]	-	0.611	0.631	0.621	46.4
	OF	0.635	0.658	0.646	69.3
	Vanilla	0.643	0.641	0.642	33.2
	SCCE	0.709	0.638	0.669	34.5
Ours	-	0.665	0.643	0.654	9.3
	OF	0.649	0.673	0.661	15.7
	Vanilla	0.659	0.656	0.657	4.1
	SCCE	0.725	0.651	0.686	4.5

Compressed representation. We conduct the ablation studies on various strategies of using the compressed representations, *i.e.*, (1) use optical flow (OF) in PC [28], (2) replace RGB images of P-frames in PC [28] with compressed representation (*i.e.*, motion vectors and residuals), (3) use our spatial-channel compressed encoder in PC [28], (4) remove compressed representation in our method, (5) use optical flow in our method, and (6) replace spatial-channel compressed encoder in our method with a vanilla encoder. The results are shown in Table 4. Visualization exemplars of the compressed information are shown in Figure 4. Both PC [28] and our method are benefit from optical flow, compressed representation and the spatial-channel

Table 5. Ablation study of our temporal contrastive module by varying the window size k . $k = 0$ means we remove the temporal contrastive module. This study shows that it’s critical to learn the temporal dependency explicitly for event boundary detection. However, the value of window size gives limited influence to the performance.

Window size	Rec	Prec	F1
$k = 0$	0.725	0.651	0.686
$k = 2$	0.675	0.749	0.710
$k = 4$	0.697	0.745	0.720
$k = 6$	0.729	0.744	0.736
$k = 8$	0.757	0.736	0.746
$k = 10$	0.725	0.750	0.737
$k = 12$	0.696	0.761	0.727

compressed encoder module respectively. Specifically, the optical flow branch improves F1 score by 2.5% compared to the original PC [28] method, with a much slower inference speed. The compressed information only brings limited improvements to PC [28] and our end-to-end method with a relatively faster inference speed. This phenomenon indicates that the simple usage of motion vectors and residuals cannot fully exploit the rich information contained in compressed domain. The proposed spatial-channel compressed encoder provides significant performance improvements without extra computation cost, indicating that the proposed encoder can learn more discriminative features of P-frames. This module allows our proposed method to fully exploit the compressed representations and capture crucial motion information from the nearly cost-free motion vectors and residuals.

Temporal contrastive module. Besides the discriminative features of P-frames, the temporal dependencies are also important to predict the accurate event boundaries. To validate the effectiveness of the temporal contrastive module, we conduct several experiments, shown in Table 5. As shown in Table 5, without the temporal contrastive module (*i.e.*, $k = 0$), the overall accuracy (F1 score) decreased dramatically. After adapting the proposed temporal module, the F1 score improves sharply, *i.e.*, 0.710 *vs.* 0.686 at $k = 2$. To further analyze the effective of different window size in model accuracy, we also perform several experiments with different k values. Table 5 shows that the recall starts to drop when $k > 8$. We believe that it is because larger window size mixes temporal information cross boundaries, resulting in the combination of multiple different predictions and decreasing the recall value. Considering the performance, we set $k = 8$ in our experiments as the default setting.

Comparisons with the state-of-the-arts. We compare the proposed method with the state-of-the-art methods at CVPR’21 LOnG-form Video Understanding (LOVEU) Challenge¹. CLA [18] uses contrastive learning based ap-

¹<https://sites.google.com/view/loveucvpr21>

Table 6. Comparisons with the state-of-the-art methods. The results are evaluated in validation split. † indicate the results come from our implementations since the test server is unavailable now. † CLA [18] uses the concatenation of pre-trained two-stream TSN [41] and SlowFast [8] features as input, † CASTANET [14] and ours uses pretrained CSN [36] as backbone. The speed is computed by averaging per-frame decoding and inference time.

Method	Rec	Prec	F1	Speed(ms)
† CLA [18]	0.815	0.768	0.791	90.2
† CASTANET [14]	0.838	0.732	0.781	93.9
Ours (CSN+R18)	0.813	0.761	0.786	20.4
Ours (R50+R18)	0.751	0.742	0.746	4.7

proach to deal with the GEBD and utilizes temporal self-similarity matrix (TSM) as an intermediate representation. However, their approach relies on pre-extracted features and uses global similarity matrix, which hurts model’s scalability. CASTANET [14] adapts the identical framework from PC [28] except the feature extractor and thus introduces redundant computations between nearby frames. Our method with ResNet50 runs extremely fast, *i.e.*, $20\times$ faster than CLA [18]. After replacing I-frame feature extractor f_I with a more powerful backbone CSN [36], we achieve competitive result, *i.e.*, 0.787 compared with CLA 0.795 and CASTANET 0.784, while improving inference speed for more than $4\times$. The result shows the efficiency of working on the compressed domain using a lightweight network as the P-frame feature extractor and the effectiveness of our proposed method on high quality representation learning.

5. Conclusion

In this work, we propose an end-to-end compressed video representation learning method for GEBD. Specifically, we convert the video input into successive frames and use the Gaussian kernel to preprocess the annotations. Meanwhile, we design a spatial-channel compressed encoder to make full use of the motion vectors and residuals to learn discriminative feature representations for P-frames. After that, we propose a temporal contrastive module to model the temporal dependency between frames and generate accurate event boundaries. Extensive experiments have conducted on the Kinetics-GEBD dataset demonstrate that the proposed method performs favorably against the state-of-the-art methods.

6. Acknowledgement

This work was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038. Libo Zhang was supported CAAI-Huawei MindSpore Open Fund and Youth Innovation Promotion Association, CAS (2020111).

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *CoRR*, abs/2103.15691, 2021. 2, 3
- [2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE Computer Society, 2017. 6
- [3] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *ECCV*, volume 11205 of *Lecture Notes in Computer Science*, pages 364–380, 2018. 6
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 6
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 2, 3
- [6] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *CoRR*, abs/2104.11227, 2021. 2, 3
- [7] Lijie Fan, Wen-bing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *CVPR*, pages 6016–6025. Computer Vision Foundation / IEEE Computer Society, 2018. 6
- [8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210. IEEE, 2019. 1, 2, 8
- [9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941. IEEE Computer Society, 2016. 1, 2
- [10] Junyi Feng, Songyuan Li, Xi Li, Fei Wu, Qi Tian, Ming-Hsuan Yang, and Haibin Ling. Taplab: A fast framework for semantic video segmentation tapping into compressed-domain knowledge. *CoRR*, abs/2003.13260, 2020. 3
- [11] CV Forecast. Cisco visual networking index: Forecast and trends, 2017–2022. *White paper, Cisco Public Information*, pages 1–4, 2019. 1
- [12] Didier Le Gall. MPEG: A video compression standard for multimedia applications. *Commun. ACM*, 34(4):46–58, 1991. 6
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 6
- [14] Dexiang Hong, Congcong Li, Longyin Wen, Xinyao Wang, and Libo Zhang. Generic event boundary detection challenge at CVPR 2021 technical report: Cascaded temporal attention network (CASTANET). *CoRR*, abs/2107.00239, 2021. 1, 3, 8
- [15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 3
- [16] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *CVPR*, pages 13886–13895, 2021. 2, 3
- [17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans*, 35(1):221–231, 2013. 1, 3
- [18] Hyolim Kang, Jinwoo Kim, Kyungmin Kim, Taehyun Kim, and Seon Joo Kim. Winning the cvpr’2021 kinetics-ebd challenge: Contrastive learning approach. *CoRR*, abs/2106.11549, 2021. 1, 2, 3, 8
- [19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE Computer Society, 2011. 6
- [20] Zhen-Zhong Lan, Ming Lin, Xuanchong Li, Alexander G. Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *CVPR*, pages 204–212. IEEE Computer Society, 2015. 2
- [21] Colin Lea, Austin Reiter, René Vidal, and Gregory D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 36–52, 2016. 7
- [22] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019. 3
- [23] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3888–3897. IEEE, 2019. 7
- [24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021. 2, 3
- [25] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S. Davis. Actionflownet: Learning motion representation for action recognition. In *WACV*, pages 1616–1624. IEEE Computer Society, 2018. 6
- [26] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *ECCV*, volume 8693, pages 581–595, 2014. 2
- [27] Ayush K. Rai, Tarun Krishna, Julia Dietlmeier, Kevin McGuinness, Alan F. Smeaton, and Noel E. O’Connor. Discerning generic event boundaries in long-form wild videos. *CoRR*, abs/2106.10090, 2021. 3
- [28] Mike Zheng Shou, Deepti Ghadiyaram, Weiyao Wang, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. *CoRR*, abs/2101.10511, 2021. 1, 2, 3, 6, 7, 8
- [29] 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 *CVPR*, pages 1268–1277, 2019. 2, 3, 4, 6, 7
- [30] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 1, 2

- [31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 6
- [32] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943. Computer Vision Foundation / IEEE Computer Society, 2018. 5, 6
- [33] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, volume 6316 of *Lecture Notes in Computer Science*, pages 140–153. Springer, 2010. 1, 3
- [34] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE Computer Society, 2015. 1, 3, 6
- [35] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *CoRR*, abs/1708.05038, 2017. 6
- [36] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5551–5560. IEEE, 2019. 8
- [37] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE Trans*, 40(6):1510–1517, 2018. 1, 3
- [38] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176. IEEE Computer Society, 2011. 2
- [39] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558. IEEE Computer Society, 2013. 2
- [40] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439. Computer Vision Foundation / IEEE Computer Society, 2018. 6
- [41] 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 *ECCV*, volume 9912, pages 20–36, 2016. 8
- [42] Shiyao Wang, Hongchao Lu, and Zhidong Deng. Fast object detection in compressed video. In *ICCV*, pages 7103–7112. IEEE, 2019. 2, 3
- [43] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3
- [44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *ECCV*, pages 3–19, 2018. 3
- [45] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed video action recognition. In *CVPR*, pages 6026–6035, 2018. 2, 3, 4, 6, 7
- [46] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *CoRR*, abs/1712.04851, 2017. 6
- [47] Youngjae Yu, Sangho Lee, Gunhee Kim, and Yale Song. Self-supervised learning of compressed video representations. In *ICLR*, 2021. 2, 3
- [48] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 5
- [49] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *CVPR*, pages 2718–2726. IEEE Computer Society, 2016. 2, 3, 6, 7
- [50] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Trans. Image Process.*, 2018. 6, 7
- [51] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 917–925. ACM, 2021. 2, 3