LEARNING FROM THE RAW DOMAIN: CROSS MODALITY DISTILLATION FOR COMPRESSED VIDEO ACTION RECOGNITION

Yufan Liu[1,2], *Jiajiong Cao*[3], *Weiming Bai*[1,2], *Bing Li*[1,4*], *Weiming Hu*[1,2]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Ant Financial Service Group [4]PeopleAI, Inc.

ABSTRACT

Video action recognition is faced with the challenges of both huge computation burden and performance requirements. Using compressed domain data, which saves much decoding computation, is a possible solution. Unfortunately, existing compressed-domain-based (CD) methods fail to obtain high performance, compared with state-of-the-art (SOTA) raw-domain-based (RD) methods. In order to solve the problem, we propose a cross-modality knowledge distillation method to force the CD model to learn the knowledge from the RD model. In particular, spatial knowledge and temporal knowledge are first constructed to align feature space between the raw domain and the compressed domain. Then, an adaptively multi-path knowledge learning scheme is presented to help the CD model learn in a more efficient way. Experiments verify the effectiveness of the proposed method in large-scale and small-scale datasets.

Index Terms— Cross modality, knowledge distillation, compressed domain, video action recognition

1. INTRODUCTION

With the development of Internet techniques, a huge amount of videos is processed every day. Though the network speed is increased, the computational speed becomes the bottleneck for video action recognition applications because of cumbersome deep learning models and spatial-temporal redundancy of the video data.

To alleviate this problem, some methods perform video action recognition on compressed domain (CD) instead of raw domain (RD). Since it only needs partial decoding to obtain CD data, as shown in Fig. 1. The decoding time and storage can be significantly reduced. For example, CoViAR [1] simply replaces RD data, *i.e.*, RGB frames, with CD data, *i.e.*, Intra-frame (I frame), motion vector (MV) and residual [2]. DMC-Net [3] and Slow-I-Fast-P [4] utilize optical flow or pseudo optical flow to construct the motion information and enhance the model performance. Some recent works [5,6] explore multi-modal fusion techniques for CD data. However, these CD-based methods are still inferior to state-of-the-art (SOTA) RD-based methods. Although the CD data contains most of the video information, the implicit knowledge is hard to be sufficiently learned. Since it only contains sparse appearance information, *i.e.*, the spare I-frames, rather than the dense RGB frames which are crucial for prediction.

Knowledge Distillation (KD) [7–9] from RD model to CD model is one possible solution to close the performance gap between them. Different from traditional KD pipelines, the CD student model and the RD teacher model receive different modalities as input. There exist large challenges in knowledge definition, feature alignment and learning scheme design for cross-modal KD. Some recent works [10] study on cross-modal KD, but most of them focus on different tasks such as biometric matching and lip reading. The effects of video recognition are still limited. Battash *et al.* [11] proposes a cross-model KD framework to align CD and RD directly in the input space. It first reconstructs the missing Predicted frames (P frames) based on the I frames and the residuals, and then feeds these P frames to the student network. However, the performance gain is marginal since it is difficult to simply align the input space between CD and RD.

In this paper, we propose a novel cross-modal KD method to align the knowledge between CD and RD in the feature space. We first define two types of knowledge including spatial knowledge and temporal knowledge. The former refers to the appearance feature of each video frame. A pseudo-decoder is proposed in the student to decode the CD features to the appearance features of missing P frames. The latter is defined as the temporal relation among different frames, which is constructed by the proposed Temporal Graph (TG). Furthermore, an adaptively multi-path knowledge (AMK) scheme is presented to boost the learning of the above knowledge. Specifically, a multi-path spatial attention gate module is proposed to select the most informative spatial features. And the multi-birth TGs are presented to enrich the temporal knowledge. Our main contributions are summarized below:

- We define two types of video knowledge, *i.e.*, spatial

[*] Corresponding author.

knowledge and temporal knowledge, for cross-modal KD, to align CD and RD in video action recognition.
- We propose an adaptively multi-path knowledge (AMK) learning scheme to boost the learning of knowledge.
- Experiments show that the proposed method is effective. It outperforms 8 CD-based methods and achieves competitive performance with RD-based methods.

2. THE PROPOSED METHOD

The proposed cross-modal KD framework is illustrated in Fig. 1, wherea the CD-based model is the student and the RD-based model is the teacher. The teacher takes the fully-decoded raw frames $\mathbf{V}_{\text{raw}} = \{\mathrm{F}_t\}_{t=1}^{T}$ as inputs while the student takes the partially-decoded data $\mathbf{V}_{\text{comp}} = \{\mathrm{I}, \{\mathrm{M}_t, \mathrm{R}_t\}_{t=1}^{T}\}$ as inputs. Note that I denotes the I frame. And $\mathrm{M_t}$ and $\mathrm{R_t}$ represent the t-th MV and the t-th residual, respectively. We first construct the cross-modal knowledge to align the teacher and the student to the same knowledge space in Sec. 2.1. Then, an adaptively multi-path knowledge (AMK) learning scheme introduced in Sec. 2.2 is adopted to make the student better absorb and digest the knowledge.

2.1. Cross-modal knowledge construction

In order to sufficiently distill knowledge from RD to CD, we propose two types of knowledge including multi-path spatial knowledge and multi-path temporal knowledge as follows.

(1) Spatial knowledge. The multi-path spatial knowledge is defined as the features of the dense frame sequence, where each path refers to the feature of one frame. For the RD-based teacher, the t-th path spatial knowledge $\mathbf{h}_t = \mathrm{Backbone}^{\mathrm{T}}(\mathrm{F_t})$ is the feature of the t-th raw frame. For the CD-based student, inspired by the classic video decoder which decodes the P frames based on the relation among I frames, MVs and residuals, a *pseudo-decoder* is proposed to decode the compressed-domain features (*i.e.*, $\{\mathbf{f}^{\mathrm{I}}, \{\mathbf{f}^{\mathrm{M}}{}_t, \mathbf{f}^{\mathrm{R}}{}_t\}_{t=1}^{T}\} = \mathrm{Backbone}^{\mathrm{S}}(\{\mathrm{I}, \{\mathrm{M}_t, \mathrm{R}_t\}_{t=1}^{T}\})$) and obtain the decoded dense feature sequence $\{\hat{\mathbf{h}}_t\}_{t=1}^{T}$. Note that $\hat{\mathbf{h}}_t$ is the aligned spatial knowledge, *i.e.*, the reconstructed feature of the t-th P frame. To fully explore the relation, we adopt two cross-attention transformer blocks. One learns the relation between I frame and MV while the other learns the relation between I frame and residual:

$$\begin{aligned} &\mathbf{h}_t^{\text{rec}} = \mathrm{Attn}(Q^{\mathrm{I}}, K_t^{\mathrm{M}}, V^{\mathrm{I}}) + \mathrm{Attn}(Q^{\mathrm{I}}, K_t^{\mathrm{R}}, V^{\mathrm{I}}), \\ &s.t. \quad Q^{\mathrm{I}} = \mathbf{f}^{\mathrm{I}} \cdot W^{Q}, \qquad V^{\mathrm{I}} = \mathbf{f}^{\mathrm{I}} \cdot W^{V}, \\ &\qquad\quad K_t^{\mathrm{M}} = \mathbf{f}^{\mathrm{M}}{}_t \cdot W^{K^{\mathrm{M}}}, \quad K_t^{\mathrm{R}} = \mathbf{f}^{\mathrm{R}}{}_t \cdot W^{K^{\mathrm{R}}}. \end{aligned} \tag{1}$$

Note that $\mathrm{Attn}(\cdot)$ is the *self-attention* operation of the Transformer encoder. Then, a convolutional bi-directional long-short term model (Bi-ConvLSTM) is adopted to enhance the temporal relation among the reconstructed features. Consequently, the final spatial knowledge $\hat{\mathbf{h}}_t$ for the student is computed as follows:

$$\{\hat{\mathbf{h}}_t\}_{t=1}^{T} = \mathrm{BiConvLSTM}(\{\mathbf{h}_t^{\text{rec}}\}_{t=1}^{T}). \tag{2}$$

(2) Temporal knowledge. The multi-path temporal knowledge is defined as the relation among the frame sequence. We construct a temporal graph (TG) to represent the temporal relation. The TG of the teacher TG^{raw} and that of the student TG^{comp} can be represented as:

$$\begin{aligned} &TG^{\text{raw}} = (\mathcal{V}^{\text{raw}}, \mathcal{E}^{\text{raw}}) = (\{\mathbf{h}_t\}_{t=1}^{T}, \mathbf{A}^{\text{raw}}), \\ &TG^{\text{comp}} = (\mathcal{V}^{\text{comp}}, \mathcal{E}^{\text{comp}}) = (\{\hat{\mathbf{h}}_t\}_{t=1}^{T}, \mathbf{A}^{\text{comp}}), \\ &s.t. \quad \mathbf{A}^{\text{raw}}(i,j) = ||\mathbf{h}_i - \mathbf{h}_j||_2^2, \\ &\qquad \mathbf{A}^{\text{comp}}(i,j) = ||\hat{\mathbf{h}}_i - \hat{\mathbf{h}}_j||_2^2, \ i,j = 1, ..., T, \end{aligned} \tag{3}$$

where $\mathcal{V}$ denotes the vertex of the TG and $\mathcal{E}$ is the edge of the TG. Each vertex represents the feature (or reconstructed feature) of a single frame. Furthermore, we extend a single TG to multi-birth TGs as described in Sec. 2.2. In this way, original single-path temporal knowledge is enriched to be multi-path temporal knowledge.

2.2. Adaptively multi-path knowledge learning

After constructing the cross-modal knowledge, an adaptively multi-path knowledge learning scheme is proposed to boost the learning of this knowledge. For spatial knowledge learning, an *attention gate* is introduced to selectively learn the multi-path appearance features of the teacher. Among these features, useful information is actually not uniformly distributed. Thus, simply learning these multi-path features may introduce redundant information and harm performance. The compressed domain natively contains hints for knowledge importance. For example, the MVs contain temporal and motion cues reflecting the importance of each frame. Taking advantage of the CD data, the proposed *attention gate* controls the supervision intensity of the spatial knowledge at different frames. In particular, it takes compressed features $\{\{\mathbf{f}^{\mathrm{M}}{}_t\}_{t=1}^{T}, \{\mathbf{f}^{\mathrm{R}}{}_t\}_{t=1}^{T}\}$ as input and adopts a squeeze-and-excitation (SE) block to obtain the importance weights $\boldsymbol{\alpha} = \{\alpha_t\}_{t=1}^{T}$ of the spatial knowledge at different frames. The loss function of learning the spatial knowledge is:

$$\mathcal{L}_{\text{spatial}} = \sum_{t=1}^{T} \alpha_t ||\mathbf{h}_t - \hat{\mathbf{h}}_t||_2^2, \tag{4}$$

where α_t generated from the *attention gate* controls the supervision intensity at the t-th frame.

For the temporal knowledge learning, *multi-birth TGs* are proposed to enrich temporal knowledge. In particular, the original vertexes are transformed into different feature spaces by different mapping functions, to construct different TGs. We select 3 mapping functions to obtain the multi-birth TGs:

$$\begin{aligned} &TG_1 = (\mathcal{V}_1, \mathcal{E}_1) = (\{\exp(-\frac{||\mathbf{h}_t - \boldsymbol{\mu}||^2}{2\sigma^2})\}_{t=1}^{T}, \mathbf{A}_1), \\ &TG_2 = (\mathcal{V}_2, \mathcal{E}_2) = (\{\exp(-\frac{||\mathbf{h}_t - \boldsymbol{\mu}||}{2\sigma^2})\}_{t=1}^{T}, \mathbf{A}_2), \\ &TG_3 = (\mathcal{V}_3, \mathcal{E}_3) = (\{a(\mathbf{h}_t)^2 + b(\mathbf{h}_t) + c\}_{t=1}^{T}, \mathbf{A}_3). \end{aligned} \tag{5}$$

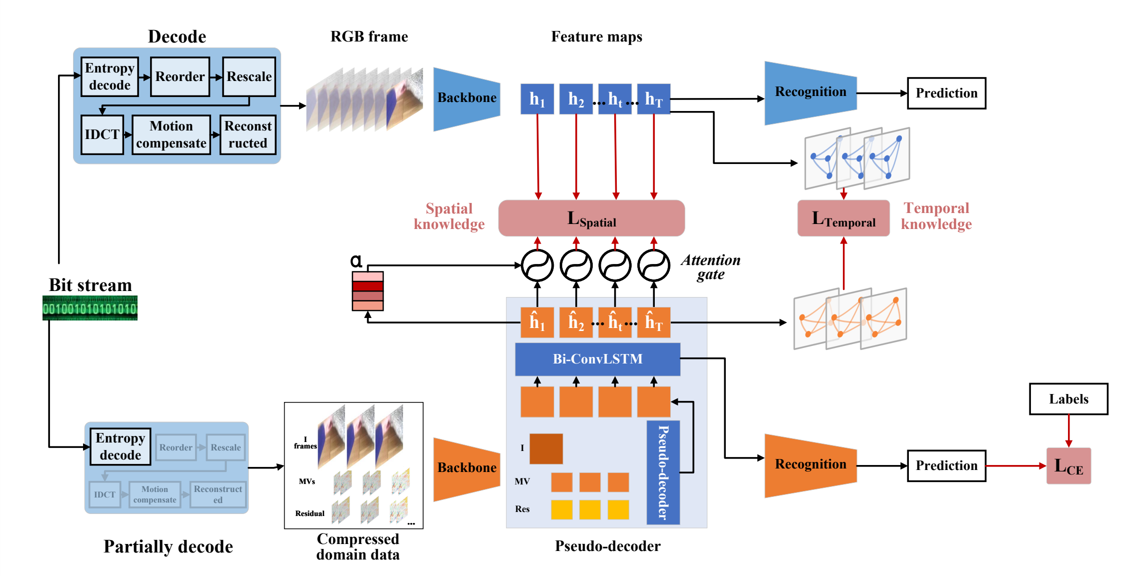

Fig. 1. Overall framework of the proposed method.

Note that the coefficients in this paper are $\sigma = a = b = c = 1$. And $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are the corresponding adjacent matrixes of multi-birth TGs. Under the AMK learning scheme, the student mimics the temporal relation in the multi-birth TGs of the teacher and the loss function is:

$$\mathcal{L}_{\text{temporal}} = \sum_{i=1}^{3}(||\mathcal{E}_i^{\text{raw}} - \mathcal{E}_i^{\text{comp}}||_2^2). \tag{6}$$

Note that $\mathcal{E}_i^{\text{raw}}$ and $\mathcal{E}_i^{\text{comp}}$ are the corresponding edges of the i-th multi-birth TG in the raw domain (teacher) and that in the compressed domain (student). In summary, the overall loss function can be formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1\mathcal{L}_{\text{spatial}} + \lambda_2\mathcal{L}_{\text{temporal}}, \tag{7}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss and λ_1 and λ_2 are the penalty coefficients.

3. EXPERIMENTS

3.1. Setings

Datasets: We evaluate the proposed methods on four popular video action recognition datasets including UCF-101 [12], HMDB-51 [13], Kinetics-400 [14] and Something-Something V2 (SSV2) [15]. UCF-101 contains 13,320 trimmed short videos from 101 action categories, while HMDB-51 contains around 7,000 video clips from 51 action categories. Kinetics-400 is a large-scale YouTube video dataset and has around 300k trimmed videos covering 400 categories, while SSV2 consists of about 220K videos with a time span from 2 to 6 seconds for 174 action categories. In our experiments, all videos are decoded to I-frames and P-frames with an MPEG-4 codec [2].

Implementation: All training videos are first resized to 320 × 256. Then, for data augmentation, random horizontal flipping and random cropping are applied to I-frames, MVs and residuals. For fair comparison with SOTAs, the backbone is ResNet-50 and is pre-trained on ImageNet [16]. The hyper-parameters are set to $\lambda_1 = 0.1, \lambda_2 = 1$. On the other hand, we choose TSM [17] or TDN [18] as the RD teacher. The training details for the teacher strictly follow the original paper. For testing, we resize all the videos to 256 × 256.

Table 1. Accuracies (%) on UCF-101 and HMDB-51.

Method	Backbone	Modality	Year	UCF-101	HMDB-51
C3D [19]	C3D	RD	2015	82.3	51.6
I3D [14]	InceptionV1	RD	2017	95.6	74.8
TSM [17]	ResNet-50	RD	2019	96.0	73.2
TDN [18]	ResNet-50	RD	2021	97.4	76.3
CoViAR [1]	ResNet-50	CD	2018	91.0	73.2
DMC-Net [3]	ResNet-152	CD	2019	92.3	71.8
MFCD-Net [11]	MF-Net	CD	2020	93.2	66.9
IMRNet [5]	ResNet-50	CD	2021	92.6	67.8
TEAM-Net [6]	ResNet-50	CD	2021	94.3	73.8
MM-ViT [20]	ViT-B/16	CD+Audio	2022	95.4	-
TSM+ours	ResNet-50	CD	-	95.8	73.5
TDN+ours	ResNet-50	CD	-	96.9	75.9

Table 2. Accuracies (%) on SSVV2. We adopt TSM [17] and TDN [18] to the RD with three-parallel-backbone (for I-frames, motion vectors and residuals, respectively) and a fully-connected (FC) feature fusion layer.

Method	Backbone	Modality	Year	Top-1	Top-5	GFLOPs
TRN [21]	BNInception	RD	2018	48.8	77.6	33
TSM [17]	ResNet-50	RD	2019	63.4	88.5	390
SmallBigNet [22]	ResNet-50	RD	2020	63.3	88.8	157
TEINet [23]	ResNet-50	RD	2020	65.5	89.8	98
TANet [24]	ResNet-50	RD	2020	66.0	90.1	297
TDN [18]	ResNet-50	RD	2021	67.0	90.3	108
TSM [17]	ResNet-50	CD	2019	60.1	86.2	315
TDN [18]	ResNet-50	CD	2021	62.7	88.1	86
MM-ViT [20]	ViT-B/16	CD	2022	64.9	89.7	2250
TSM+ours	ResNet-50	CD	-	62.8	88.2	348
TDN+ours	ResNet-50	CD	-	66.9	90.2	92

3.2. Comparison with SOTA

Results on UCF-101 & HMDB-51: The results on UCF-101 & HMDB-51 are presented in Table 1. It can be observed

Table 3. Accuracies (%) on Kinetics-400.

Method	Backbone	Modality	Year	Top-1	Top-5	GFLOPs
I3D [14]	InceptionV1	RD	2017	72.1	90.3	-
TSM [17]	ResNet-50	RD	2019	74.7	91.4	1950
SmallBigNet [22]	ResNet-50	RD	2020	76.3	92.5	1710
CorrNet [25]	ResNet-50	RD	2020	77.2	-	6720
TDN [18]	ResNet-50	RD	2021	78.4	93.6	3240
CoViAR [1]	ResNet-50	CD	2018	69.1	-	3615
MFCD-Net [11]	MF-Net	CD	2020	68.3	-	128
TEAM-Net [6]	ResNet-50	CD	2021	72.2	-	-
TSM+ours	ResNet-50	CD	-	73.5	90.8	1740
TDN+ours	ResNet-50	CD	-	77.4	93.1	2760

that our method outperforms the best CD-based methods by a large margin (1.5% on UCF-101 and 2.1% on HMDB-51). It indicates that the proposed method extracts more discriminative features from the compressed domain. On the other hand, the proposed method even outperforms some RD-based methods, which demonstrates that the RD knowledge is well learned by the model via our cross-modality KD pipeline.

Results on SSV2: The detailed results including top-1&top-5 accuracies and computational cost (in FLOPs) are reported in Table 2. The proposed method consistently outperforms all of the counterparts using CD data and most of the counterparts using RGB modality (except TDN). Though the proposed method slightly underperforms TDN by 0.1%, it reduces 14.8% FLOPs. Considering the proposed method only requires partial decoding, it is a more efficient method to be deployed in real-world applications.

Results on Kinetics-400: Kinetics-400 is a large action recognition dataset that shall further verify the generalization of the method. As shown in Table 3, Similar to the results on SSV2, the proposed method outperforms almost all the competing methods. The performance gap is even larger (from 5.2% to 8.3%) compared with its counterparts operating on the compressed domain. It significantly verifies the effectiveness and generalization of the proposed method.

3.3. Ablation Studies

Effectiveness of each component: We use two popular models including TSM and TDN as the baselines to verify spatial KD and temporal KD. In particular, TSM and TDN are adopted to the RD with three-parallel-backbone (for I-frames, motion vectors and residuals, respectively) and a fully-connected (FC) feature fusion layer. The results on UCF-101 and SSV2 are reported in Table 4. Both spatial KD and temporal KD significantly improve the performance of the baseline. And the combination of them obtains a performance gain of around 4% compared with the baseline. The results demonstrate the effectiveness of each component.

Sensitivity of hyper-parameters: We adopt simple grid search strategy to determine the best hyper-parameter values, namely, the values of λ_1 and λ_2. We first conduct experiments to obtain the value of λ_1. Then, λ_2 is searched with λ_1 fixed. The results are shown in Table 5. TSM is adopted as the baseline for hyper-parameter searching. And we find that the best results for TSM also work well on TDN-based frameworks. Consequently, we set λ_1 to 0.1 and λ_2 to 1.0 for all the proposed models.

Table 4. Performance comparison of the proposed method variants on UCF-101 and SSV2.

Method	Spatial KD	Temporal KD	UCF-101	SSV2-top1	SSV2-top5
TSM [17]	No	No	92.2	60.1	86.2
TSM+ours	Yes	No	94.5	61.9	87.6
TSM +ours	No	Yes	94.0	61.5	87.3
TSM+ours	Yes	Yes	95.8	62.8	88.2
TDN [18]	No	No	93.3	62.7	88.1
TDN+ours	Yes	No	95.6	64.9	89.2
TDN +ours	No	Yes	95.3	64.1	88.8
TDN+ours	Yes	Yes	96.9	66.9	90.2

Table 5. Hyper-parameter search and logits distillation results on UCF-101 and SSV2.

Method	λ_1	λ_2	UCF-101	STSTV2-1	STSTV2-5
TSM [17]	0.0	0.0	92.2	60.1	86.2
TSM+ours	0.01	0.0	93.7	61.3	87.1
TSM+ours	**0.1**	0.0	94.5	61.9	87.6
TSM+ours	1.0	0.0	93.6	61.1	87.0
TSM+ours+logits	**0.1**	0.0	94.6	61.9	87.7
TSM+ours	**0.1**	0.1	94.9	62.3	87.8
TSM+ours	**0.1**	**1.0**	95.8	62.8	88.2
TSM+ours	**0.1**	5.0	95.3	62.4	87.8
TSM+ours+logits	**0.1**	**1.0**	95.9	62.8	88.3

Logits KD: Logits distillation is one of the most commonly used KD techniques. Logits distillation is added to the proposed method as shown in Table 5. We find that the addition of logits distillation does not significantly improve the performance. But it introduces one more hyper-parameter and makes the framework more complex. Therefore, we do not utilize logits distillation in the proposed framework.

4. CONCLUSION

In this paper, a cross-modality knowledge distillation method for video action recognition is proposed. It helps the compressed domain based model learn useful knowledge from the raw-domain-based model and obtain a better performance. In particular, multi-path spatial and temporal knowledge are first defined. And an adaptively multi-path knowledge learning scheme is then presented for efficient learning of the above knowledge. Experiments on three popular action recognition datasets verify the superiority of the proposed method.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106800), the National Natural Science Foundation of China (No. 62192785, No. 61972071, No. U1936204, No. 62122086, No. 62036011, No. 62192782 and No. 61721004), the Beijing Natural Science Foundation No. M22005 and L223003, the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (Grant: 2017KZDXM081, 2018KZDXM066), Guangdong Provincial University Innovation Team Project (Project No. 2020KCXTD045). The work of Bing Li was also supported by the Youth Innovation Promotion Association, CAS.

5. REFERENCES

[1] 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*, 2018, pp. 6026–6035.

[2] Didier Le Gall, "Mpeg: A video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 46–58, 1991.

[3] 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*, 2019, pp. 1268–1277.

[4] Jiapeng Li, Ping Wei, Yongchi Zhang, and Nanning Zheng, "A slow-i-fast-p architecture for compressed video action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2039–2047.

[5] Youngjae Yu, Sangho Lee, Gunhee Kim, and Yale Song, "Self-supervised learning of compressed video representations," in *International Conference on Learning Representations*, 2020.

[6] Zhengwei Wang, Qi She, and Aljosa Smolic, "Team-net: Multi-modal learning for video action recognition with partial decoding," *arXiv preprint arXiv:2110.08814*, 2021.

[7] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan, "Knowledge distillation via instance relationship graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7096–7104.

[8] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, and Stephen Maybank, "Learning to explore distillability and sparsability: a joint framework for model compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting D-ing, and Liang Li, "Cross-architecture knowledge distillation," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3396–3411.

[10] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2143–2147.

[11] Barak Battash, Haim Barad, Hanlin Tang, and Amit Bleiweiss, "Mimic the raw domain: Accelerating action recognition in the compressed domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 684–685.

[12] 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.

[13] 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*. IEEE, 2011, pp. 2556–2563.

[14] 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*, 2017, pp. 6299–6308.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al., "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[17] 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*, 2019, pp. 7083–7093.

[18] 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*, 2021, pp. 1895–1904.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[20] Jiawei Chen and Chiu Man Ho, "Mm-vit: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1910–1921.

[21] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[22] Xianhang Li, Yali Wang, Zhipeng Zhou, and Yu Qiao, "S-mallbignet: Integrating core and contextual views for video classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1092–1101.

[23] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Y-ing Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu, "Teinet: Towards an efficient architecture for video recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11669–11676.

[24] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13708–13718.

[25] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli, "Video modeling with correlation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 352–361.