VCD: VIEW-CONSTRAINT DISENTANGLEMENT FOR ACTION RECOGNITION

Xian Zhong[1,3], *Zhuo Zhou*[1], *Wenxuan Liu*[1], *Kui Jiang*[4], *Xuemei Jia*[4], *Wenxin Huang*[2,*], *and Zheng Wang*[4]

[1] School of Computer and Artificial Intelligence, Wuhan University of Technology
[2] School of Computer Science and Information Engineering, Hubei University
[3] School of Electronics Engineering and Computer Science, Peking University
[4] School of Computer Science, Wuhan University

ABSTRACT

Action recognition is a hot topic in computer vision due to its wide range of applications in urban surveillance. Although some methods are more advanced from an invariant view perspective, those approaches do not perform well for the viewpoint change. To address this issue, one possible solution is tantamount to track the view-invariant representation as it evolves with the performed action. However, the views' and actions' performance always complement each other, once simply looking for the view-invariant representation may cause some behavior information to be lost. In this paper, we propose the View-Constraint Disentanglement (VCD) framework for cross-view action recognition. Specifically, Constraint Disentanglement Module (CDM) is utilized to learn an action-invariant representation by discretizing view-specific representation and its normal distribution, which resolves the entangled relationship between view and action. Moreover, a novel Adaptive Distribution Module (ADM) is intended to befit enhance the high-correlation viewpoint variation information and refine the suitable weight. Extensive experiments are conducted on public benchmarks, indicating that our approach achieves better performance than other state-of-the-art approaches.

***Index Terms*—** Action recognition, Cross-view, Disentangled representation learning, Adaptive distribution

1. INTRODUCTION

Action recognition is fundamental to accomplishing effective expression and interaction between humans. With the rapid development of urban surveillance, this task has become a hot topic because of its significance in practical application [1, 2]. Some state-of-the-art recognition methods have achieved tremendous success across a range of condition [3, 4, 5]. However, the appearance and dynamics of action are visually different from one viewpoint to another [6, 7, 8]. Viewpoint invariance of a real complex environment is still a stark challenge waiting to be solved.

∗ Corresponding author: wenxinhuang_wh@163.com. This work was supported in part by the Department of Science and Technology, Hubei Provincial People's Government under Grant 2021CFB513 and 2021CFB281.

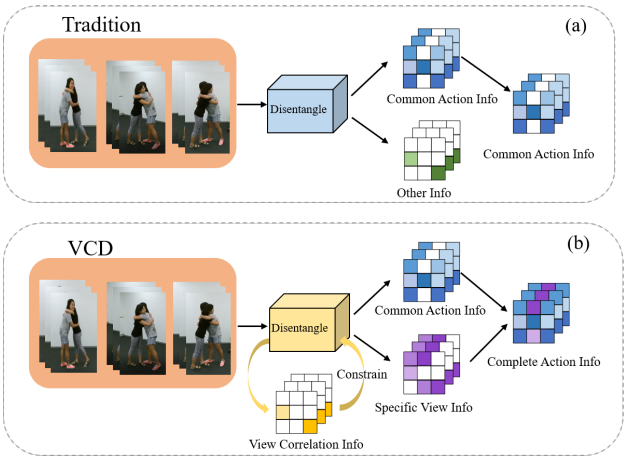


Fig. 1. Problem description. Given a kind of target action, (a) traditional work ignores the special action info in other info. (b) VCD pays attention to the view correlation constraints and finally obtains the feature of not losing important information.

To mitigate the issue, existing cross-view action recognition works can be roughly divided into two categories: 1) A few studies leverage extra modal information to enhance the viewpoint representation [6, 9]. 2) Some researchers purely learn view-invariant features [7, 10]. They completely separate the view-invariant feature from the behavior, but a lot of realistic motion information is ignored.

As we can see from Fig. 1(a), just as the perception of human beings, we can infer activity categories from various views based on summarizing the most typical feature of this motion. The category of each action is isolated, which leads [7, 10] to focus on the maximum common visual information of all views in order to obtain view-invariant representation. But in the actual scene, it is impossible to completely strip view-invariant features, causing losses of some essential clues. Inspired by [11], we find that the unique visual information of each view is often continuous and has different effects on the same behavior. Therefore, we propose the novel View-Constraint Disentanglement (VCD) framework to address the cross-view action recognition problem, as shown in Fig. 1(b). Our purpose is to utilize the inter- and intra-view correlations

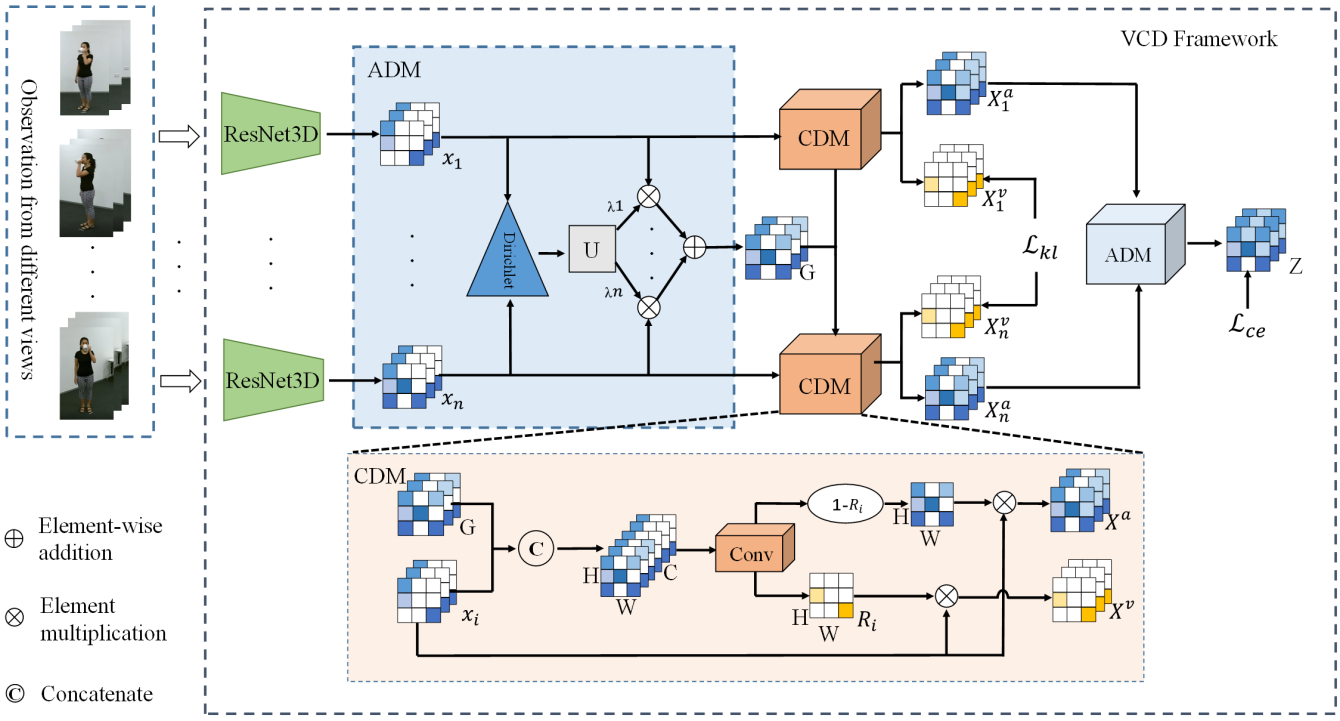


Fig. 2. Overview of View-Constraint Disentanglement (VCD) framework. For a sequence of input videos, Resnet3D generates a sequence of representations. ADM adaptively assigns weights for these representations to gain a view correlation info. U stands for the classification uncertainty. CDM disentangle those representations to get view-specific features ($\boldsymbol{X}_i^v$) and action-invariant features ($\boldsymbol{X}_i^a$), and the $\boldsymbol{X}_i^a$ can learn complete action information. Then ADM uses the action-invariant features of each view to achieve a complete action-invariant feature for action recognition from the invisible view.

among different views to form a positive intervention in the entire disentangling process. Namely, disentangling under different views considers both behavior characteristics and potential connections among multiple perspectives. Specifically, Constraint Disentanglement Module (CDM) explores the potential connections of the same action at different views and makes full use of the maximum common visual information under the constraint of each view to get the action-invariant feature and the view-specific feature. CDM replenishes the missing information caused by relying solely on the view-invariant representation of perspective.

Moreover, traditional algorithms generally learn multi-view representations with a fixed and equal value for each view. The underlying assumption is that the qualities or significance of these views are stable for all samples, which is practically impossible. A conscious model to assign adaptive weights for different views is necessarily needed. Thus, we design an Adaptive Distribution Module (ADM), where different weights rather than equal values are allocated to multiple views in a more adequate manner.

The contributions of this work are threefold:

• We propose a novel VCD framework with CDM mining each view's peculiar and interaction without any favorable information deficiency.

• To address the unfair feature fusion problem, we add ADM that can integrate to obtain a holistic representation from multiple views to our VCD. This allows the network to preserve the uniqueness of view, which facilitates better modeling of view information.

• We conduct extensive experiments that validate our model's superior accuracy, robustness, and reliability thanks to our view-constraint methods and multi-view adaptive selection strategy.

2. PROPOSED METHOD

We address cross-view action recognition by constraining the complement relationship between action and view information. Disposing of the interference caused by viewpoint variations, action-invariant features could achieve more precise activity representation.

2.1. Problem Definition

Our VCD framework is shown in Fig. 2. The input of the framework are multiple videos of different views, denoted as $V_i = \{v_1, \ldots, v_i\}$. The feature extraction network uses multi-stream ResNet3D learning spatio-temporal features $\boldsymbol{x}_i$ of each view independently. ADM refines the suitable weight for v_i to obtain a view correlation info G. Then G and x_i are

disentangled into view-specific and action-invariant features. Finally, ADM gains complete action-invariant features for prediction.

2.2. Constraint Disentanglement Module

We use the constraint disentanglement module to constrain view-specific features, and then action-invariant features can learn complete action representations. By utilizing G and x_i, we gain disentangled view-specific and action-invariant features of each view. We calculate the similarity between the G and x_i to obtain the corresponding correlation map. The correlation map $\boldsymbol{R}_i \in \mathbb{R}^{1\times H\times W}$ related to $\boldsymbol{x}_i$ under the constraint can be computed by:

$$\boldsymbol{R}_i \;=\; \sigma(W_i([\boldsymbol{x}_i, \boldsymbol{G}])) \tag{1}$$

where $[\cdot,\cdot]$ means to connect the two operands, W_i is a learnable weight of three 1×1 3D convolutional layers with Batch Normalization (BN) and Rectified Linear Unit (Relu) activation, and σ denotes the sigmoid activation function. After calculating the correlation map, we multiply the correlation map and G to get $\boldsymbol{X}_i^v$, view-specific representation of each view. Then, we perform the inversion operation on correlation maps to obtain the irrelevant maps. Then we multiply the irrelevant graphs with G to get the action-invariant representation $\boldsymbol{X}_i^a$.

$$\begin{aligned} \boldsymbol{X}_i^v &= \boldsymbol{x}_i \odot \boldsymbol{R}_i \\ \boldsymbol{X}_i^a &= \boldsymbol{x}_i \odot (1 - \boldsymbol{R}_i) \end{aligned} \tag{2}$$

where $\odot$ indicates element-wise multiplication. Finally, we constrain $\boldsymbol{X}_i^v$, extract the specific view information contained in $\boldsymbol{X}_i^v$ and back to the framework, then the $\boldsymbol{X}_i^a$ can learn complete action information.

2.3. Adaptive Distribution Module

To avoid treating each view equally, we design ADM to adaptively assign the weight of the view according to its probability. We use the Dirichlet distribution [12] to quantify the overall uncertainty of the classification of each view. For each view, we obtain the score y_i by using a Linear layer and a ReLU layer, and then use the conjugate distribution to obtain the Dirichlet distribution d_i induced from y_i, $(d_i = y_i + 1)$, so as to obtain the classification uncertainty u_i of each view.

$$\begin{aligned} u_i &= K/D_i \\ D_i &= \sum_{i=1}^{n}(y_i + 1) \;=\; \sum_{i=1}^{n} d_i \end{aligned} \tag{3}$$

where D_i represents the Dirichlet strength, K denotes the number of classes, u_i indicates the classification uncertainty. It describes the phenomenon that the more accurate predictions in the i-th view, the greater the probability of being assigned to the i-th view. Accordingly, we need to determine the concentration parameters, which are closely related to the uncertainty. Specifically, we assign different weights λ_i to different views and then accumulate them to get the complete action-invariant feature Z, formulated as:

$$\begin{aligned} \boldsymbol{Z} &= \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \\ \lambda_i &= \frac{1 - u_i}{n - \sum_{i=1}^{n} u_i} \end{aligned} \tag{4}$$

2.4. Loss Function

We use classification cross-entropy loss and Kullback-Leibler (KL) divergence loss, denoted as $\mathcal{L}_{ce}$ and $\mathcal{L}_{kl}$, to supervise the model training. The KL divergence loss makes the view-specific representation contain more viewpoint information and then reverse the constraint disentanglement method.

KL divergence loss is employed to normalize the distribution of the view-specific features, making it close to the normal distribution $p(x_i^v) \sim \mathcal{N}(0, 1)$, formulated as:

$$\mathcal{L}_{kl} \;=\; \frac{1}{N}\sum_{i}^{n}\frac{1}{2}(q(\boldsymbol{X}_i^v)^2 + e^{q(\boldsymbol{X}_i^v)} - q(\boldsymbol{X}_i^v) - 1) \tag{5}$$

where $q(\boldsymbol{X}_i^v)$ represents predicted score of view-specific features $\boldsymbol{X}_i^v$. The full objective function is a weighted sum of all the losses:

$$\mathcal{L} \;=\; \beta_1\mathcal{L}_{ce} + \beta_2\mathcal{L}_{kl} \tag{6}$$

where β_1, β_2 are the distribution weights of $\mathcal{L}_{ce}$ and $\mathcal{L}_{kl}$.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Experiments Setup

N-UCLA [13] is captured by three cameras from different views and contains 1,493 videos of 10 action categories. We follow [13] and use videos from the first two views for training and the third for testing.

NTU-RGB+D [14] has RGB videos, depth and skeleton data. It contains more than 56k videos with 60 different action classes and 40 different actors. Our experiments result over the complex actions with interaction classes of this dataset following [15].

The performance of a method is evaluated by computing an accuracy score based on whether it predicts the correct action class or not. Following [7], we adopt Top-1 of Cross-View setting (CV) and Top-1 of Cross-Subject setting (CS) as our evaluation standard.

In previous work, they ignore that Resnet3D [16] can accept the entire video and better retain the action information. The traditional single ResNet3D cannot complete our multi-stream tasks, so we use multi-stream ResNet3D (MR-Net) by reducing the frame rate to obtain features with rich action information as our baseline. For each video, we sample 16 frames and resize them to $3\times112\times112$. During training, we

Table 1. Performance comparison of CV (%) and CS (%) with the state-of-the-arts on **N-UCLA** and **NTU-RGB+D**.

Method	Venue	Modality	N-UCLA		NTU-RGB+D	
			CV	CS	CV	CS
Att-LSTM [10]	ECCV '18	Skeleton	70.6	63.3	-	-
CrosSCLR [17]	CVPR '21	Skeleton	-	-	83.4	77.8
CV-MIM [9]	CVPR '21	Skeleton	-	-	89.5	77.8
MST-AOG [13]	CVPR '14	RGB-S	73.7	81.6	87.3	83.0
GCA-LSTM [18]	CVPR '17	RGB-S	-	-	89.0	85.9
R-NKTM [19]	TPAMI '18	RGB-S	78.1	-	-	-
ST-GCN [20]	AAAI '18	RGB-S	-	-	87.1	83.3
DV-Views [21]	CVPR '12	RGB	58.5	50.7	58.5	50.7
TSN [22]	ECCV '16	RGB	85.4	84.9	80.6	84.9
VIFL-SAM [23]	TIP '17	RGB	77.2	81.1	77.2	81.1
DA-NET [24]	ECCV '18	RGB	75.3	-	84.2	88.1
CVAM [7]	ECCV '20	RGB	83.1	87.5	86.3	82.3
Conflux LSTM [25]	NEUCOM '21	RGB	88.9	-	-	-
VCD (w/o ADM + CDM)		RGB	88.7	87.8	87.6	87.3
VCD (w ADM + CDM)		RGB	91.8	90.3	90.6	89.2
VCD (w ADM + CDM + CL)		RGB	93.6	90.5	91.9	90.4
VCD (w ADM + CDM + CL + MU)		RGB	**93.8**	**91.8**	**92.3**	**90.8**

use a mini-batch of size 16. We train the model using the Adam optimizer with an initial learning rate of 1e-3 and a weight decay of 5e-4. Our implementation is in Pytorch, and the model is trained with Tesla P100 GPUs.

3.2. Comparison with State-of-the-arts

N-UCLA. Table 1 shows the classification accuracy for cross-subject and cross-view with two input modalities. It is observed that the proposed method significantly improves CS and CV evaluation for both RGB (2.8% and 2.9%) and Skeleton/RGB-S (8.7% and 13.7%), which demonstrates the effectiveness of exploring the invariant representation of viewpoint variations. It is worth mentioning that our CV evaluation outperforms previous works by a large margin because of our perceptive insight of investigating view-specific disentanglement. Thus more attention is needed to be paid to action-invariant information, rather than mutative views and actors' appearances.

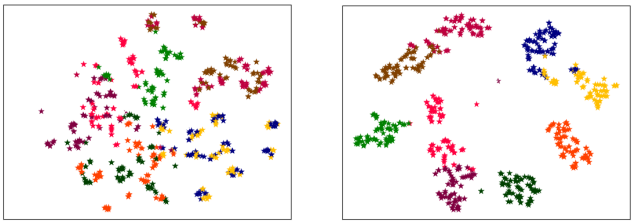
NTU-RGB+D. In Table 1, the methods of the first column take RGB+Skeleton as input, and the methods of the second take RGB-Only as input. Our method meets the best performance over the two kinds of methods under both cross-view and cross-subject settings. It provides solid evidence that distills action-invariant and disposes of view-specific information benefits multi-view action recognition.

3.3. Ablation Study

Effectiveness of Each Component. In Table 2, "B/L" indicates baseline, "LR" indicates the low frame rate sampling, "ADM" indicates distribution module, and "CDM" indicates constraint disentanglement module adaptively respectively. On

Table 2. Performance comparison of CV (%) with different variants of our method on **N-UCLA**. The videos from two views are used for training, and videos from the remaining view are used for testing, denoted as {Source}|Target.

B/L	LR	ADM	CDM	{1,2}|3	{1,3}|2	{2,3}|1	AVG
✓	×	×	×	87.1	83.9	86.3	85.7
✓	✓	×	×	89.9	87.4	89.3	88.7
✓	✓	✓	×	90.6	87.8	90.2	89.6
✓	✓	×	✓	91.2	88.0	90.8	90.0
✓	✓	✓	✓	**93.3**	**89.3**	**92.8**	**91.8**



(a) Initial training data (b) Best training data

Fig. 3. Visualization of feature distributions of videos in the first two prominent dimensions of the initial and the results of the best training data. A total of ten actions are randomly selected from **N-UCLA**. Here, each color represents an action category.

N-UCLA, we adopt view {1,2}, view {1,3}, and view {2,3} for training and the rest for testing, respectively. "LR" obtains rich action information features, "VCD" and "ADM" both achieve remarkable performance gains on the baseline under different settings. Specifically, the average accuracy achieves 91.8%, 3.1% higher than the baseline model, proving the benefits of the designed components.

Visualization of Classification Results. To better display the recognition capacity of our proposed method, we evaluate the feature-level views of training sets on **N-UCLA**. We obtain T-SNE distributions of learned feature vectors in Fig. 3. We randomly sample 80 videos from 10 action categories, and different colors denote different classes. The visualization greatly demonstrates the effectiveness of our proposal to classify different activities.

4. CONCLUSION

In this work, we propose a novel learning framework, namely View-Constraint Disentanglement (VCD), for action recognition in the cross-view and cross-subject environment. The VCD constrains the disentanglement process and adapts the number and weight of input views, making it suitable for cross-view action recognition. The generalization ability and its adaptive nature make it suitable for other problem domains in the multi-view environment. The comprehensive experiments have demonstrated that the proposed VCD outperforms the baseline methods.

5. REFERENCES

[1] Ke Yang, Zhiyuan Wang, Huadong Dai, Tianlong Shen, Peng Qiao, Xin Niu, Jie Jiang, Dongsheng Li, and Yong Dou, "Attentional fused temporal transformation network for video action recognition," in *ICASSP*, 2020.

[2] Yuze Tian, Xian Zhong, Wenxuan Liu, Xuemei Jia, Shilei Zhao, and Mang Ye, "Random walk erasing with attention calibration for action recognition," in *PRICAI*, 2021.

[3] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang, "DB-LSTM: densely-connected bi-directional LSTM for human action recognition," *Neurocomputing*, 2021.

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *ICCV*, 2019.

[5] Lu Xu, Xian Zhong, Wenxuan Liu, Shilei Zhao, Zhengwei Yang, and Luo Zhong, "Subspace enhancement and colorization network for infrared video action recognition," in *PRICAI*, 2021.

[6] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli, "Unsupervised learning of view-invariant action representations," in *NeurIPS*, 2018.

[7] Shruti Vyas, Yogesh Singh Rawat, and Mubarak Shah, "Multi-view action recognition using cross-view video prediction," in *ECCV*, 2020.

[8] Federico Angelini, Zeyu Fu, Sergio A. Velastin, Jonathon A. Chambers, and Syed Mohsen Naqvi, "3D-Hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes," in *ICASSP*, 2018.

[9] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J. Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris N. Metaxas, and Ting Liu, "Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization," in *CVPR*, 2021, pp. 12793–12802.

[10] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *ECCV*, 2018.

[11] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He, "Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," *arXiv:2106.11232*, 2021.

[12] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou, "Trusted multi-view classification," in *ICLR*, 2021.

[13] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning, and recognition," in *CVPR*, 2014.

[14] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *CVPR*, 2016.

[15] Mauricio Perez, Jun Liu, and Alex C. Kot, "Interaction relational network for mutual action recognition," *arXiv:1910.04963*, 2019.

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2d CNNs and imagenet?," in *CVPR*, 2018.

[17] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang, "3D human action representation learning via cross-view consistency pursuit," in *CVPR*, 2021.

[18] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *CVPR*, 2017.

[19] Hossein Rahmani, Ajmal S. Mian, and Mubarak Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[20] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[21] Ruonan Li and Todd E. Zickler, "Discriminative virtual views for cross-view action recognition," in *CVPR*, 2012.

[22] 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*, 2016.

[23] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE Trans. Image Process.*, 2017.

[24] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu, "Dividing and aggregating network for multi-view action recognition," in *ECCV*, 2018.

[25] Amin Ullah, Khan Muhammad, Tanveer Hussain, and Sung Wook Baik, "Conflux LSTMs network: A novel approach for multi-view action recognition," *Neurocomputing*, 2021.