The following publication T. Liu, R. Zhao, J. Xiao and K. -M. Lam, "Progressive Motion Representation Distillation With Two-Branch Networks for Egocentric Activity Recognition," in IEEE Signal Processing Letters, vol. 27, pp. 1320-1324, 2020 is available at https://doi.org/10.1109/LSP.2020.3011326.

Progressive Motion Representation Distillation with Two-branch Networks for Egocentric Activity Recognition

Tianshan Liu, Rui Zhao, Jun Xiao, and Kin-Man Lam, *Senior Member, IEEE*

***Abstract*—Video-based egocentric activity recognition involves fine-grained spatio-temporal human-object interactions. State-of-the-art methods, based on the two-branch-based architecture, rely on pre-calculated optical flows to provide motion information. However, this two-stage strategy is computationally intensive, storage demanding, and not task-oriented, which hampers it from being deployed in real-world applications. Albeit there have been numerous attempts to explore other motion representations to replace optical flows, most of the methods were designed for third-person activities, without capturing fine-grained cues. To tackle these issues, in this letter, we propose a progressive motion representation distillation (PMRD) method, based on two-branch networks, for egocentric activity recognition. We exploit a generalized knowledge distillation framework to train a hallucination network, which receives RGB frames as input and produces motion cues guided by the optical-flow network. Specifically, we propose a progressive metric loss, which aims to distill local fine-grained motion patterns in terms of each temporal progress level. To further enforce the proposed distillation framework to concentrate on those informative frames, we integrate a temporal attention mechanism into the metric loss. Moreover, a multi-stage training procedure is employed for the efficient learning of the hallucination network. Experimental results on three egocentric activity benchmarks demonstrate the state-of-the-art performance of the proposed method.**

***Index Terms*—Egocentric activity recognition, knowledge distillation, two-branch networks, motion representation.**

I. Introduction

EGOCENTRIC activity recognition is an emerging research topic in computer vision, owing to the popularization of wearable sensors and its widespread real-world applications [1] [2], such as robot navigation, smart homes, augmented reality, etc. Compared with the current common studies of third-person activity [3] [4], egocentric activity usually contains more complex fine-grained human-object interactions in both spatial and temporal dimension. The performance of a recognition method largely depends on whether the relevant fine-grained spatio-temporal patterns can be extracted and utilized. However, it is a challenging task to capture such information because of many factors, such as the invisibility of the camera wearer and the frequent egocentric motions [5].

Most of the recent studies employ a two-branch-based deep-learning architecture for egocentric activity recognition, which has achieved promising results [6] [7]. The spatial branch extracts appearance cues from RGB frames for noun classification, while the temporal stream is trained to capture motion cues from stacked optical-flow images for verb classification. Then, the two separate branches are combined by late fusion to identify the activities. However, conventional optical-flow estimation approaches are computationally expensive and storage demanding, due to the hundreds of optimization iterations for each frame [8]. In addition, this two-stage strategy that first computes optical flow and then feeds it into Convolutional Neural Networks (ConvNets) for predicting activity classes is potentially sub-optimal. The underlying reason for this is that the conventional optical-flow estimation is not task-oriented [9]. These drawbacks may hamper the two-branch-based methods to be deployed in real-world applications, due to the limitations of processing time and memory capacity.

To address this problem, researchers have made various attempts to seek other motion representations to substitute optical flow. Zhang et al. [10] proposed to utilize motion vectors, in stead of the precise optical flow. However, speed improvement is achieved by sacrificing recognition accuracy. Zhu et al. [9] presented a motionNet to estimate motion based on unsupervised learning, and further cascaded it with a temporal stream CNN to generate activity labels. Piergiovanni et al. [8] proposed a differentiable CNN layer, which implements the iteration procedure of the traditional TV-L1 optical flow approach [11]. Nevertheless, most of these representations are designed for third-person activities, thus may not be applicable for modelling the fine-grained temporal patterns in egocentric activities. Although Crasto et al. [12] trained a 3D CNN that mimics the optical-flow branch to produce motion cues by taking RGB frames as input, this method only takes global features into consideration and inevitably loses local temporal details. Moreover, exploring both spatial and temporal information implicitly through only one stream is insufficient for recognizing fine-grained egocentric activities.

Since RGB and optical flow describe activities from the appearance and motion aspects, respectively, they can be regarded as the information from two different modalities. To achieve task-oriented fine-grained motion information extraction, while avoiding pre-computation of optical flow during the inference time, in this letter, the task is converted as follows. We exploit both RGB and optical-flow data to train the two-branch networks in the training stage, and only utilize the RGB frames in the testing phase. Then, we tackle this issue by exploring the theory of the teacher-student learning model [13]

Tianshan Liu, Rui Zhao, Jun Xiao, and Kin-Man Lam are with Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: tianshan.liu@connect.polyu.hk; rick10.zhao@connect.polyu.hk; jun.xiao@connect.polyu.hk; enkm-lam@polyu.edu.hk). *(Corresponding author: Tianshan Liu)*

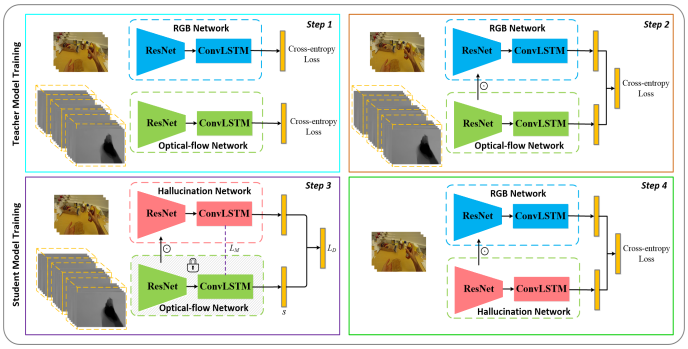

Fig. 1. The overall architecture and training paradigm of the proposed method.

and a generalized distillation framework [14]. Specifically, we propose a two-branch-based distillation framework, which learns to distill motion representations into a hallucination network guided by an optical-flow network. This also implies that the hallucination network is trained to produce motion cues by receiving RGB frames as input. To distill local fine-grained motion patterns regarding each temporal level, we propose a metric loss based on a progressive manner. In addition, a temporal attention strategy is introduced into the metric loss, which aims to force the teacher-student model to pay more attention to conducting motion distillation based on the more informative frames. The proposed metric loss is combined with a generalized distillation loss to distill knowledge from both the local and global perspectives. Furthermore, a multi-stage training paradigm is employed to train the hallucination network. Finally, the RGB network and the hallucination network account for providing appearance and motion representations, respectively, for egocentric activity recognition. The overall network architecture and training paradigm are illustrated in Fig. 1.

The main contributions of this letter can be summarized in three ways. First, we propose a two-branch-based distillation framework, which learns a hallucination network to produce fine-grained motion cues by taking RGB frames as input. Second, a progress-wise metric loss is proposed by integrating temporal attention mechanism, to distill local motion representations considering different temporal levels. Third, experimental results on three public egocentric activity benchmarks highlight that the proposed method achieves promising performance.

II. Methodology

A. Overall Network Architecture

As illustrated in Fig. 1, the overall network is designed based on a two-branch architecture. Each branch is developed by combining ResNets [15] with LSTM cells. To further localize the informative spatio-temporal regions, an attention mechanism based on class activation map (CAM) [16] is introduced into each branch, as in [6]. In addition, we employ multiplicative interactions [17] in the convolutional layers of the ResNets between the two branches to facilitate information sharing. The multiplicative connections are implemented by injecting the optical-flow signal into the residual unit of the RGB or hallucination branch. In terms of the form of input data, we sample a series of RGB frames from a video clip for both the RGB and hallucination branches. For the optical-flow branch, we collect an optical-flow volume by stacking multiple consecutive optical-flow images around the current frame as the input. This strategy ensures the temporal synchronism between the RGB and optical-flow modalities. In other words, the optical-flow volume can provide relevant short-term motion patterns for each sampled RGB frame. The training paradigm can be divided in two parts: the first part (Steps 1 and 2) aims to learn the teacher model by separate training and joint fine-tuning, using RGB and optical-flow data; the second part (Steps 3 and 4) concentrates on learning the student model, i.e., the hallucination network, leveraging the proposed loss function.

B. Progressive Motion Representation Distillation

The hallucination network takes RGB frames as input and is trained to distill motion information from the optical-flow network at multiple levels, including both local and global perspectives. The schematic diagram of the proposed progressive motion representation distillation framework is illustrated in Fig. 2. Specifically, given a convolutional feature tensor extracted by ResNet as $\mathbf{x}_t$, where t is the time index ($t = 1, 2, ..., T$). Then, the corresponding latent embeddings calculated by the LSTM cells for hallucination and optical-flow networks are denoted as $\mathbf{E}_t^h$ and $\mathbf{E}_t^o$, respectively.

To capture local knowledge at each time step, we adopt a Euclidean distance-based metric loss in a progress-wise manner, which can be defined as follows:

$$L_M = \left\| \frac{1}{T} \sum_{t=1}^{T} \alpha_t \left(\mathbf{E}_t^o - \mathbf{E}_t^h \right) \right\|_2^2, \tag{1}$$

where α_t is a weight indicating the contribution of the metric loss at different time steps. We generate α_t by employing a self-attention-based strategy [18]. The latent embeddings $\mathbf{E}_t^h$ are first fed into a multilayer perceptron. Then we utilize a non-linear operation, e.g., the $sigmoid$ activation function, to obtain the temporal importance weights α_t regarding the whole video clip. This temporal attention-based strategy aims to assign higher weights to those frames that contribute more to the activity classification. The proposed metric loss explicitly considers temporal ordering. Therefore, minimizing L_M implies distilling motion representations at each temporal progress level.

For a global distribution perspective, motivated by the teacher-student learning model explored in multimodal tasks, we utilize a generalized distillation loss [19] for the prediction layer. The specific distillation loss is defined as follows:

$$L_D = \lambda g\left(y, \eta\left(f_h\left(\mathbf{E}_T^h\right)\right)\right) + (1-\lambda)\, g\left(s, \eta\left(f_h\left(\mathbf{E}_T^h\right)\right)\right), \tag{2}$$

where η is a softmax function and $g(\cdot)$ indicates a cross-entropy loss. $f_h(\cdot)$ maps the latent embeddings of the last time step $\mathbf{E}_T^h$ to the logits of the hallucination branch. The parameter $\lambda \in [0, 1]$ balances the importance between hard labels y and the soft labels s in the distillation loss. The soft labels s

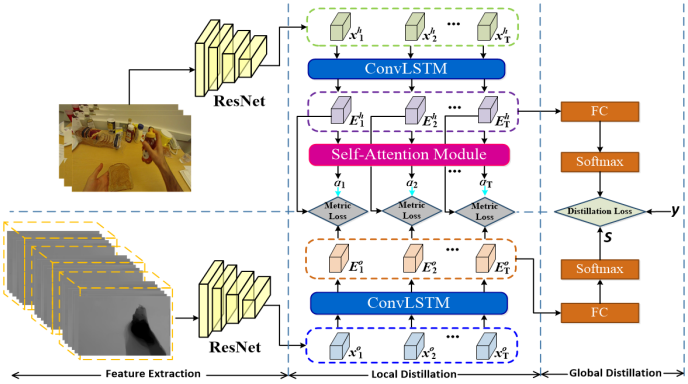

Fig. 2. Illustration of the proposed progressive motion representation distillation framework.

of the optical-flow network are calculated by $\eta\left(f_o\left(\mathbf{E}_T^o\right)/\delta\right)$, where the temperature parameter δ is employed to smooth the logits vector. $f_o\left(\cdot\right)$ transforms the latent embeddings $\mathbf{E}_T^o$ to the logits of the optical-flow network.

The distance-based metric loss and the generalized distillation loss account for distilling motion cues from the local and global aspects, respectively. Based on the proposed progressive strategy, the final loss function is formulated by combining these two loss terms as follows:

$$L = \gamma L_M + \left(1-\gamma\right) L_D, \tag{3}$$

where the parameter $\gamma \in [0, 1]$ is used to balance the weights of the two losses in the training phase.

C. Multi-stage Training Procedure

To avoid excessive heuristic-based searching of hyper-parameters, we employ a multi-stage learning procedure.

Step 1. The RGB and optical-flow networks are first trained separately, as in the classical two-branch-based architecture. The ResNet, used in the RGB branch, is initialized with the weights pre-trained on ImageNet. We follow the cross-modality initialization strategy [20] to initialize the optical-flow branch. Both these branches are trained by minimizing the cross-entropy loss.

Step 2. To facilitate knowledge sharing between different modalities, we introduce cross-branch connections into the architecture. Specifically, after initializing the RGB and the optical-flow networks with the weights learned in Step 1, we jointly train the whole model with multiplicative connections between the ResNets of these two branches.

Step 3. The hallucination network is trained to distill motion patterns from the teacher model, i.e., the optical-flow network. The optical-flow network learned in Step 2 receives optical-flow data, and is frozen to provide a stable target for the hallucination network. We initialize the hallucination network with the weights of the optical-flow network and feed RGB frames into it. Then, it receives interactions from the optical-flow signal and is trained by optimizing the proposed loss function defined in Eq. (3).

Step 4. We initialize the RGB network with the weights learned from Step 2. The hallucination network is initialized with the weights learned from Step 3. Then, we jointly fine-tune these two branches by only taking RGB frames as input. The hallucination network acts as the optical-flow network to provide motion representations for activity recognition.

III. Experiments

A. Datasets

We evaluate the proposed method on three public egocentric activity-recognition data sets, including GTEA 61, GTEA 71 and EGTEA Gaze+ [21]. Both GTEA 61 and GTEA 71 contain 457 samples performed by 4 subjects. GTEA 61 and GTEA 71 consist of 61 and 71 activity classes, respectively. EGTEA Gaze+ is one of the large-scale data sets, involving 10,325 samples and 106 activity classes. These activity instances lie in a long-tailed distribution. For the GTEA 61 and GTEA 71 data sets, we adopt the experimental settings of leave-one-subject-out cross-validation as in [7]. We report the averaged accuracy over three splits for the EGTEA Gaze+ data set.

B. Implementation Details

We choose ResNet-34 as the backbone CNN to extract spatial features for each branch. The standard convLSTM block with 512 hidden units is adopted for temporal encoding. We uniformly sample 20 RGB frames from each video clip and feed them into the RGB and hallucination networks, in both training and testing phases. For each sampled RGB frame, we construct an optical-flow volume by stacking 3 consecutive optical-flow images, i.e., 6 channels. Then we feed the generated 20 optical-flow volumes into the optical-flow network. For the training of the hallucination network in Step 3, we adopt the SGD optimization algorithm [22] with an initial learning rate of 0.001 and a momentum of 0.9 for 750 epochs. The weight parameters λ and γ in Eq. (2) and Eq. (3) are set as 0.6 and 0.5, respectively. The batch size is set to be 32. We utilize random horizontal flipping and multi-scale corner cropping techniques for data augmentation during the training stage. In the inference phase, we use the center crop of each frame for activity classification.

C. Ablation Study

To evaluate the effectiveness of the proposed motion distillation method, we conduct ablation experiments using different loss terms and configurations on the GTEA 61 and EGTEA Gaze+ data sets. We report the results after the learning of each stage, as presented in Table I.

1) Effect of the loss function: To investigate the contribution of each part of the proposed loss function, we conduct experiments using the distance-based metric loss (Eq. (1)), generalized distillation loss (Eq. (2)), and the combination of them (Eq. (3)). The detailed results are presented in rows 4-6 of Table I. We can find that the generalized distillation loss improves the performance over the metric loss. Although the proposed metric loss is time-aware and explores local knowledge, it is still insufficient to force the hallucination

TABLE I. Ablation recognition accuracies (%) on the GTEA 61 and EGTEA Gaze+ data sets.

Ablation configurations	Input Modality	Loss	GTEA 61	EGTEA Gaze+
1 Step 1, optical-flow branch	Optical flow	Cross-entropy	46.23	37.12
2 Step 1, RGB branch	RGB	Cross-entropy	65.52	50.38
3 Step 2, two-branch	RGB+ Optical flow	Cross-entropy	**79.86**	**61.72**
4 Step 3, hallucination branch	RGB	Eq. (1)	46.46	37.56
5 Step 3, hallucination branch	RGB	Eq. (2)	47.14	38.18
6 Step 3, hallucination branch	RGB	Eq. (3) w/ temporal attention	47.45	38.75
7 Step 3, hallucination branch	RGB	Eq. (3) w/o temporal attention	47.18	38.23
8 Step 4, two-branch	RGB	Cross-entropy	79.78	61.55

TABLE II. Comparison results on three egocentric activity data sets.

Methods	GTEA 61	GTEA 71	EGTEA Gaze+
DEA [21]	64.00	62.10	46.50
Action + object-Net [23]	73.02	73.24	-
Two-stream model [24]	51.58	49.65	41.84
Hidden two-stream [9]	54.62	51.83	45.21
TSN [20]	69.33	67.23	55.93
EleAttG [25]	66.67	60.83	57.01
Ego-RNN [6]	79.00	77.00	60.76
LSTA-two stream [7]	**80.01**	78.14	**61.86**
PMRD-RGB + hall (Ours)	79.78	**78.18**	61.55

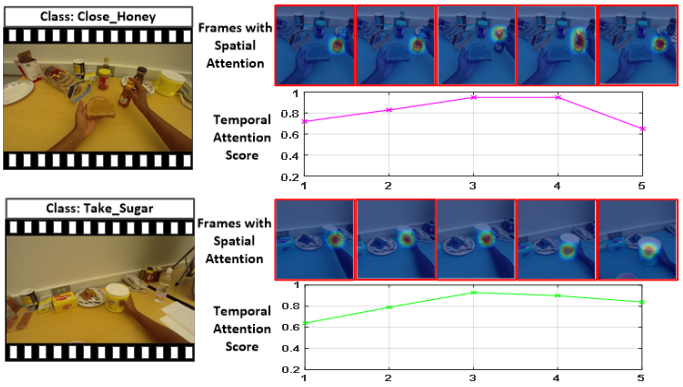

Fig. 3. Visualization of the spatial-temporal attention on two video sequences.

network to generate discriminative representations. The distillation loss exploits both soft and hard labels, thus leading to better performance for the classification task. Generally, the combination of these two loss terms outperforms the individual loss. Since metric loss and distillation loss are designed for distilling motion information from local and global perspectives, they have complementary contributions to the learning of the hallucination network.

2) Effectiveness of the temporal attention in motion distillation: We evaluate the performance of the proposed motion representation distillation method, with and without temporal attention. The comparison results are presented in rows 6-7 of Table I. The performance gap demonstrates that the temporal attention mechanism introduced into the metric loss can guide the hallucination network to concentrate more on the informative frames, thus resulting in more effective learning of the model. We also choose several video clips to visually validate the effectiveness of the temporal attention-based motion distillation strategy. Figure 3 shows the frames with spatial attention maps and the corresponding temporal attention scores. These fine-grained egocentric activity sequences usually contain cluttered backgrounds and multiple objects. The temporal attention is introduced to assign higher scores to the frames, which involve human-object interactions.

D. Comparison with State-of-the-Art Methods

The proposed method is compared with several state-of-the-art approaches, and the results are presented in Table II. These comparison methods include both traditional representations such as DEA [21], and deep learning architectures, such as the two-stream model [24] and the temporal segment networks (TSN) [20]. The first two methods in Table II require additional annotations, such as gaze information [21], object location and hand segmentation [23]. EleAttG [25], Ego-RNN [6], and LSTA-two stream [7] combine attention mechanisms with RNN networks from different perspectives, to locate the relevant spatio-temporal regions. Hidden two-stream architecture [9] is based on a motionNet, which is proposed to take consecutive RGB frames as input and estimates motion. It can be observed that the proposed hallucination network, combined with the RGB branch, achieves satisfactory results on the three egocentric data sets. The performance of our method is also comparable with the LSTA-two stream, which exploits both RGB and optical flow modalities. This suggests that our hallucination network can learn effectively, via the progressive motion pattern distillation strategy. The appearance cues extracted by the RGB branch, and the hallucinated motion representations carry complementary information for egocentric activity recognition.

IV. Conclusion

In this letter, we propose a two-branch-based progressive motion representation distillation method for egocentric activity recognition. Different from the current common practice of using pre-calculated optical flow to provide the motion information for ConvNets, we leverage a generalized knowledge distillation framework to teach a hallucination network to produce discriminative motion cues, while only taking RGB frames as input. Specifically, we propose a progress-wise metric loss, which is integrated into a deep recurrent architecture, i.e., LSTM cells, to distill local fine-grained motion patterns. In addition, we introduce a temporal attention mechanism into the teacher-student framework to focus on distilling motion information on the informative frames. Furthermore, to avoid inefficient heuristic-based searching, we utilize a multi-stage learning paradigm for the training of the hallucination network. Evaluation results on three egocentric activity data sets validate that the proposed method can achieve favorable performance.

REFERENCES

[1] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and Observer: Joint Modeling of First and Third-Person Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7396–7404.

[2] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic Attention with Privileged Information for Egocentric Action Recognition," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[3] D. Purwanto, R. R. A. Pramono, Y. Chen, and W. Fang, "Three-Stream Network With Bidirectional Self-Attention for Action Recognition in Extreme Low Resolution Videos," *IEEE Signal Processing Letters*, vol. 26, no. 8, pp. 1187–1191, 2019.

[4] Y. Wang, W. Li, and R. Tao,, "Multi-Branch Spatial-Temporal Network for Action Recognition," *IEEE Signal Processing Letters*, vol. 26, no. 10, pp. 1556–1560, 2019.

[5] H. Li, Y. Cai, and W. Zheng, "Deep Dual Relation Modeling for Egocentric Interaction Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7924–7933.

[6] S. Sudhakaran and O. Lanz, "Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition," in *British Machine Vision Conference (BMVC)*, 2018, pp. 1–12.

[7] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long Short-Term Attention for Egocentric Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9946–9955.

[8] A. Piergiovanni and M. S. Ryoo, "Representation Flow for Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9937–9945.

[9] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden Two-Stream Convolutional Networks for Action Recognition," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2019, pp. 363-378.

[10] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-Time Action Recognition With Deeply Transferred Motion Vector CNNs," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2326–2339, 2018.

[11] C. Zach, T. Pock, and H. Bischof, "A Duality Based Approach for Realtime TV-L1 Optical Flow," in *Joint Pattern Recognition Symposium*, 2007, pp. 214–223.

[12] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-Augmented RGB Stream for Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7874–7883.

[13] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[14] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph Distillation for Action Detection with Privileged Modalities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 174–192.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[17] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7445–7454.

[18] Y. Peng, Y. Zhao, and J. Zhang, "Two-Stream Collaborative Learning With Spatial-Temporal Attention for Video Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773–786, 2019.

[19] D. Lopez-Paz, L. Bottou, B. Schlkopf, and V. Vapnik, "Unifying distillation and privileged information," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool , "Temporal Segment Networks for Action Recognition in Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2019.

[21] Y. Li, Y. Zhefan, and J. M. Rehg, "Delving into egocentric actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 287–295.

[22] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, 2010, pp. 177–186.

[23] M. Ma, H. Fan, and K. M. Kitani, "Going Deeper into First-Person Activity Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1894–1903.

[24] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems (NIPS)*, 2014, pp. 568–576.

[25] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: Adding Attentiveness to Neurons in Recurrent Neural Networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1061–1073, 2020.