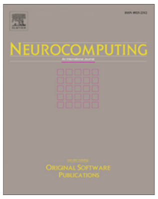
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Cross-modality online distillation for multi-view action recognition

Chao Xu [a,1], Xia Wu [a,1], Yachun Li [a,1], Yining Jin [b], Mengmeng Wang [a,*], Yong Liu [a,*]

[a] *State Key Laboratory of Industrial Control Technology and Institute of Cyber-systems and Control, Zhejiang University, China*
[b] *Department of Electrical and Computer Engineering, University of Alberta, Canada*

ARTICLE INFO

Article history:
Received 13 July 2020
Revised 26 February 2021
Accepted 21 May 2021
Available online 25 May 2021
Communicated by Zidong Wang

Keywords:
Multi-view
Cross-modality
Action recognition
Online distillation

ABSTRACT

Recently, some multi-modality features are introduced to the multi-view action recognition methods in order to obtain a more robust performance. However, it is intuitive that not all modalities are avail- able in real applications. For example, daily scenes lack depth modal data and capture RGB sequences only. Thus comes the challenge of learning critical features from multi-modality data at train time, while still getting robust performance based on RGB sequences at test time. To address this chal- lenge, our paper presents a novel two-stage teacher-student framework. The teacher network takes advantage of multi-view geometry-and-texture features during training, while the student network is given only RGB sequences at test time. Specifically, in the first stage, Cross-modality Aggregated Transfer (CAT) network is proposed to transfer multi-view cross-modality aggregated features from the teacher network to the student network. Moreover, we design a Viewpoint-Aware Attention (VAA) module which captures discriminative information across different views to combine multi-view fea- tures effectively. In the second stage, Multi-view Features Strengthen (MFS) network with the VAA module further strengthens the global view-invariance features of the student network. Besides, both of CAT and MFS learn in an online distillation manner, so that the teacher and the student network can be trained jointly. Extensive experiments on IXMAS and Northwestern-UCLA demonstrate the effectiveness of our proposed method.

1. Introduction

With the explosion of video data nowadays, the need to recognition human action information automatically is growing. Apparently, it is impossible to recognition action in a video by simply processing an isolated frame, especially to recognize human actions. Thus, there are several methods [1–4] which take motion information into consideration to encode the temporal features between adjacent frames. Recently, many remarkable two-stream approaches [5–7] and 3D convolutional based methods [8–10] have achieved great success in action understanding.

However, multi-view action recognition [11,1,12–15] remains a challenging problem. The reasons are as follows. First, the first challenge is that the mannings of visual expressions from different viewpoints would change significantly, which could easily cause ambiguity. In order to extract view-invariance features, Cai et al. [16] design a global descriptor that is composed of relatively distinctive features of each view for recognition tasks. Ji et al. [17] introduce a novel multi-view space hidden Markov model algorithm for view-invariance action recognition. From then on, with the emergence of deep convolution networks and the development of sensors [18], more and more datasets contain depth and skeleton modality, and the 3D motion begins to be view-invariance action representations. Li et al. [19] conduct extensive experiments to prove that multi-modality information can extract the essential features of human action and is not affected by the environment. Li et al. [20] and Shi et al. [21] both represent skeleton data as graphs and take advantage of graph convolutional networks for skeleton-based action recognition. Xiao et al. [22] use multi-view depth videos to extract 3D characteristics. Although the skeleton and depth information could improve multi-view action recognition to some extent, 3D representations would increase the cost of computing resources and the inference time. Besides, it is expensive to deploy depth cameras in actual scenes. Most of the cameras only capture RGB video sequences which are the cheapest available data modality. Considering this limitations, the challenge to use only RGB modalities to learn robust representations at the inference phase comes out. The second challenge is to make full use of multi-view information [15,23,24] to analyze human action. The common solution for multi-view action recognition is to transfer knowledge from one viewpoint to the other viewpoints [25,26], or to learn separated

[*] Corresponding authors.
E-mail addresses: mengmengwang@zju.edu.cn (M. Wang), yongliu@iipc.zju.edu.cn (Y. Liu).
[1] These authors contribute equally to this work.

https://doi.org/10.1016/j.neucom.2021.05.077
0925-2312/© 2021 Elsevier B.V. All rights reserved.

view-specific features first and then aggregate all of the feature for classification [27]. The above methods explore the characteristics of single viewpoints and the correlation between multiple viewpoints to understand human actions. However, they neglect that not all view-specific features are critical. For example, as shown in Fig. 1, it is usually impossible to obtain clear human characteristics from a top view, which means gaining features from some viewpoint may lead to inaccurate 3D representations.

To address the above two problems, we propose a teacher-student learning framework that learns from multi-view RGB and DensePose sequences and could be deployed with a single-view RGB sequence input. Our model is inspired from the knowledge distillation [28]. Knowledge distillation usually refers to a training procedure where a teacher network has been trained previously and then provides supervision for a student network on the same modality. Different from the traditional distillation framework, we propose an online distillation training strategy, in which the teacher and the student networks are trained simultaneously. Another work that inspires us is proposed by Gupta et al. [29], they transfer supervision from one modal to another. We employ these ideas to designing a novel two-stage learning paradigm, where both of the two stages follow the online distillation framework. Specifically, we use DensePose [30] as the 3D representation, which maps the human pixels from a single 2D image to a 3D human surface model. Many previous works [19,31,32] have proved that DensePose is an effective 3D human representation. Our two-stage network plays a different roles during training. In the first stage, Cross-modality Aggregated Transfer (CAT) network is used to transfer the geometry-and-texture information of multi-view DensePose representations from the teacher network to the student network. The teacher network is given multi-view DensePose sequences while the student network is given single-view RGB sequences. To aggregate multi-view information effectively, we build a channel-wise Viewpoint-Aware Attention (VAA) module to capture complementary information across different views. Furthermore, we apply feature supervision on middle layer during training to help the student network to learn extra cross-modality information complementary to the appearance information. After training, the student network is able to capture multi-view geometry-and-texture representations. In the second stage, we design the Multi-view Features Strengthen (MFS) network to further enhance the multi-view features of the student network. The structure of the MFS network is the same as the CAT network, which include I3D backbone and VAA module as well, but the teacher and the student networks are both given RGB sequences. Besides, there is no feature supervision since this stage is not used to learn complementary information from different modalities. More detail information is depicted in Fig. 2. As a result, the student network of the MFS could process single-view RGB sequences more efficiently and could be deployed in real life without increasing cost.

In summary, the main contributions of this paper are as follows:

- We propose a Cross-modality Aggregated Transfer (CAT) network and a Multi-view Features Strengthen (MFS) network, which could transfer view-invariance and multi-modality features to the student network in the training stage while using only RGB data in inference phase.
- We build a simple yet effective Viewpoint-Aware Attention (VAA) module, which could capture complementary information across different views.
- We propose a novel online distillation training strategy to guide the teacher-student learning.
- We make extensive experimental analyses based on two datasets, IXMAS and N-UCLA, which shows that our method achieves competitive performances.

2. Related work

2.1. Multi-view cross-modality action recognition

Human action sequences in RGB modal usually vary with the camera view, which brings challenges for multi-view action recognition. The key to this problem is to learn discriminative view-invariance features. Thus, it relies heavily on diverse multi-modality data. Holte et al. [11] generate 3D optical flow in each view and combines them into enhanced 3D motion vector fields for human action recognition. Kong et al. [12] introduce an automatic encoder to learn view-specific representations and view-shared representations in the hidden layer, thus achieving robust multi-view action recognition. Gupta et al. [2] recover 3D pose sequences and generate multiple motion projections to transfer knowledge across views. Besides, dense trajectories [3,4] have been proved to be effective to promote recognition performance. Subsequently, Zhang et al. [13] use a 3D trajectory created by synthetic data and applies dictionary learning to project real-world 2D video into a view-invariance sparse representation. In [33], Rahmani et al. propose a non-linear unsupervised model that transfers knowledge from multiple views to a canonical view. In addition to 3D trajectories, other forms of 3D modal representations are also widely used. Cheng et al. [1] take the depth information into consideration and designs a new descriptor of depth information. Rahmani et al. [34] process point clouds directly for cross-view action recognition from unknown and unseen views. Wang et al. [35] model the geometry, appearance and motion variations into a multi-view spatio-temporal AND-OR graph, which projects the 3D skeleton to different views, and combines different views with apparent features for classification. Compared with the existing methods, we use RGB and DensePose sequences as the cross-modality inputs, which could benefit from the rich geometry-and-texture features of DensePose.

2.2. Knowledge distillation

Knowledge distillation is originally proposed in [28] to transfer knowledge from a strong model to a weak one. The key idea is to transfer a soft target, which is learned by a large model, to a small model. Recently, knowledge distillation has been applied in action recognition, which can be classified into three categories in general. In the first category, the teacher is given multi-modality data to guide the student to learn. Luo et al. [36] introduce a graph distillation method that can incorporate privileged information from a multi-modality dataset to enrich the target domain where training data are scarce. Stroud et al. [37] train the teacher network with the motion input and the student network with regular images. Thus, the student network could recognize actions based on both appearance and motion. In the second category, multi-teacher single-student networks and multi-student single-teacher networks are proposed. Wu et al. [38] hold the view that multi-teacher knowledge distillation can help a single small student model to learn comprehensive knowledge better than that with the single-teacher network. Thoker et al. [39] use mutual learning to train two or more student networks together so that each student learns from the supervision of the teacher and other students. In the third category, the student learns global spatio-temporal information under the guidance of the teacher network. Bhardwaj et al. [40] achieve a memory-efficient video classification model, they feed a compute-heavy teacher network into all the frames to train a compute-efficient student which only processes a small fraction of frames. Similarly, in [41], Wang et al. use full frames to recognize actions in the teacher networks and predicts early actions from partial videos in the student network. All the above

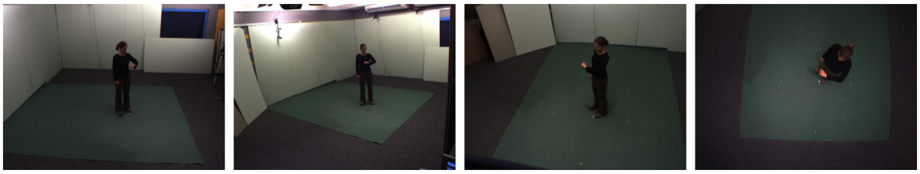


Fig. 1. *Four sampled images from different views of the same action.* The last image is obtained from the overhead camera, which does not clearly express the meaning of the action.

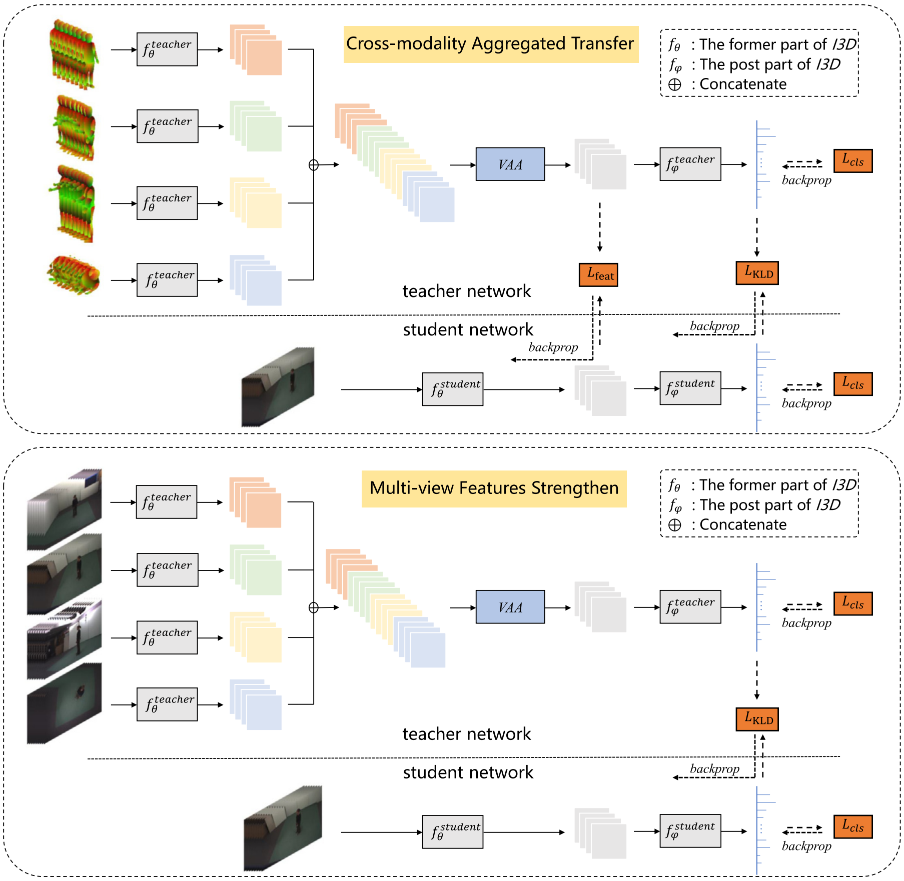


Fig. 2. *Overview of our method that consists of two stages.* In the first stage, the teacher network of our CAT is given multi-view DensePose sequences, while the student network is given single-view RGB sequences. In the second stage, both the teacher network and the student network of our MFS are given RGB sequences. Our method is designed in online distillation manner.

methods follow the pattern that the teacher network is trained in advance and provides the supervision for the student network. In contrast, we propose a strategy called online distillation in which the teacher and the student network are in the same structure and parameters are updated simultaneously.

2.3. Multi-modal data capturing

Diverse input data modalities could provide complementary cues for action recognition to achieve more robust performance. However, it is expensive to capture all modalities with sensors in intelligent monitoring. To reduce the cost of multi-modal information capturing, one solution is to design the transfer network for learning the mapping from RGB to the target modality when target modal data is unavailable. Recently, DTMMN [42] first employs U-Net to synthesize depth modality data from RGB modality and then utilizes multi-metric learning to extract discriminative multi-modality features to classify actions. PM-GANs [43] generates infrared features from RGB images for action recognition. Another solution is to train the model using all modalities and exploit RGB only at test time. This learning paradigm is generally known as learning with side information or cross-modal distillation. In the domain of action recognition, D3D [44] distills knowledge from the temporal stream that given optical flow information into the spatial stream that given RGB sequences to improve motion representations. Thoker et al. [39] extract the source modal information of the trained teacher network and transfer it to a small ensemble of student networks. Both solutions do not require any annotated modal data at test time, which reduce the cost of multi-modal data acquisition. Since cross-modal distillation is more computation-friendly, our design follows the structure of the distillation framework.

3. The proposed method

3.1. Problem overview

For the multi-view action recognition dataset $\mathcal{D}$, it can be divided into $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$ according to different views. The train test $\mathcal{D}_{train}$ contains video sequences of the same actor's action in several different views $\{v_1, v_2, \cdots, v_K\}$, and the validation set $\mathcal{D}_{val}$ includes video sequences of a new view v_l that is not in $\mathcal{D}_{train}$. Therefore, in multi-view action recognition tasks, the more global action representations learn from the train test, the better results achieved in the validation set.

Let us denote the available training data as $\{(X_{v_1}, X_{v_2}, \cdots, X_{v_K}); y\}$, where X is the training sample from the v_k view, y is the score vector of classification. After learning, the network needs to map video sequences $\{(X_{v_1}, X_{v_2}, \cdots, X_{v_K})\}$ into y, defined as $\mathcal{F}$. In our method, in addition to the original RGB sequences $\{(X_{v_1}, X_{v_2}, \cdots, X_{v_K})\}$, we use pre-trained DensePose-RCNN [30] model to extract the corresponding DensePose sequences $\{(X\prime_{v_1}, X\prime_{v_2}, \cdots, X\prime_{v_K})\}$.

In this work, we propose a two-stage network that is developed in a teacher-student learning framework. Our model uses RGB and DensePose data at training time and uses exclusively RGB data as input at test time. In the following content, we will describe our viewpoint-aware attention (VAA) module, online distillation strategy and the two-stage network structure.

3.2. Viewpoint-aware attention module

Intuitively, to effectively aggregate multi-view geometry-and-texture features, the views that are full of key information need to be emphasized, otherwise should be ignored or weakened. However, a naive implementation for multi-view feature aggregation is usually concatenating them along the channel dimension, which treats the features of different view equally.

In our viewpoint-aware attention (VAA) module, we employ a structure similar to the CBAM [45,46] to reweight the feature maps of different views. Given the feature maps $F_{v_k} \in \mathbb{R}^{B\times C\times H\times W}$ from v_k, we firstly concatenate multi-view features along channel dimension $F \in \mathbb{R}^{B\times V\times C\times H\times W}$ and feed them into 3d average pooling and 3d max pooling operations separately to generate F_{avg} and F_{max} descriptions. After that, we sum the two descriptions directly and forward them to a multi-layer perceptron, which serve as a bottleneck module. To reduce the parameters, the fully connected layer is applied to compress the number of channels C into C/r, where r is a scaling factor, the next fully connected layer is applied to restore the number back to C. After the aggregated features going through a sigmoid layer, the attention module generates combined weights of different channels. Finally, we multiply F by the weights, the output F_c of VAA can be described as:

$$F_c = F \cdot \sigma(\theta(3DAvgPool(F) + 3DMaxPool(F))) \tag{1}$$

where θ denotes multi-layer perceptron and σ denotes the sigmoid function. Details of the attention module is shown in Fig. 3

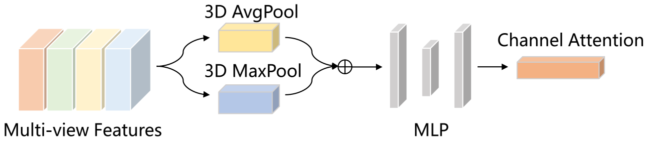


Fig. 3. *Diagram of Viewpoint-aware Attention (VAA) Module.* As illustrated, the attention module utilizes both 3d-max-pooling outputs and 3d-average-pooling outputs. A bottleneck network is followed to get channel-wise attention.

3.3. Online distillation

The previous works [38,47–49] usually train a large-scale teacher network in advance and then guide the small-scale student network to learn offline. Unlike the above works, we propose an online distillation strategy to simultaneously train two networks with similar structures, so that the student network gradually follows the teacher network, and the teacher network optimizes itself during training. The online knowledge distillation still follows the framework proposed by Hinton et al. [28]. The output denotes $z^{(i)}$ before the softmax layer is used to calculate the probability of each category $p_s^{(i)}$. Complete formula is as follows:

$$p_s^{(i)} = \frac{\exp(z^{(i)}/T)}{\sum_j \exp(z^{(j)}/T)} \tag{2}$$

where T represents temperature, which is an adjustable hyperparameter. The larger of T, the smoother of each category probability distribution.

In order to make the output distribution of the student network as close as possible to the soft target of the teacher network, we minimize the KL divergence between the p_s of two networks, the $\mathcal{L}_{KLD}$ loss is defined by:

$$\mathcal{L}_{KLD}(p_s^{teacher}||p^{student}) = \sum p_s^{teacher} \log \frac{p_s^{teacher}}{p_s^{student}} \tag{3}$$

3.4. Two-stage teacher-student learning

Cross-modality Aggregated Transfer (CAT) network. To make the student network full of multi-view 3D information, we design a novel cross-modality aggregation transfer approach to transfer the multi-view geometry-and-texture representations to the student network. The two networks with different modalities can be described as:

$$\hat{y}^{teacher} = \mathcal{F}^{teacher}(X\prime_{v_1}, \cdots, X\prime_{v_K}) \tag{4}$$

$$\hat{y}^{student} = \mathcal{F}^{student}(X_{v_k}) \tag{5}$$

In other words, multi-view DensePose sequences are fed into the teacher network to learn global features. In contrast, the student network is only given a single-view RGB sequence. Considering the features of the middle layer have more geometry-and-texture representations, we apply VAA module after the third residual block, the output channel of the attention mechanism has been expanded by a factor V, which represents the number of training viewpoints, we restore the feature channels to C as the input feature. In short, the mapping of the teacher network is described as:

$$\mathcal{F}^{teacher}(\cdot) = f_{\varphi}^{teacher}\left(g(Att(f_{\phi}^{teacher}(\cdot)))\right) \tag{6}$$

where g denotes the 3D convolutional bottleneck network and Att denotes VAA module.

Since the student network only has single-view input data, no extra aggregated function is needed. The mapping of $\mathcal{F}^{student}$ is described as:

$$\mathcal{F}^{student}(\cdot) = f_{\varphi}^{student}\left(f_{\phi}^{student}(\cdot)\right) \tag{7}$$

Besides, the original distillation method only rely on $\mathcal{L}_{KLD}$ loss, and the transfer happens at the last prediction layer, where only contains classification distribution but loses detailed semantic information of humans. We add auxiliary feature supervision after the viewpoint-aware attention module. Such choice is consistent with observations from [29] that feature loss is an effective means to learn the complementary information from different modalities, the lower layers are modality-specific while the features of mid-level layers are semantic. We use ℓ_1 loss to minimize the regression loss between the corresponding features of the two modalities.

$$\mathcal{L}_{feat}^{CAT} = \sum_{X \in \mathcal{D}_{train}} \| g(Att(f_{\phi}^{teacher}(X'_{v_1}, \cdots, X'_{v_k}))) - f_{\phi}^{student}(X_{v_k}) \|_1 \tag{8}$$

During training, we also optimize the classification loss $\mathcal{L}_{cls}$ of the category probabilities relative to the ground truth label y.

$$\mathcal{L}_{cls}^{CAT:teacher} = -\sum \hat{y}^{teacher} \log p^{teacher} \tag{9}$$

$$\mathcal{L}_{cls}^{CAT:student} = -\sum \hat{y}^{student} \log p^{student} \tag{10}$$

In summary, we train the teacher network by optimizing its own classification loss:

$$\mathcal{L}^{CAT:teacher} = \mathcal{L}_{cls}^{CAT:teacher} \tag{11}$$

As for the student network, we employ a loss function which consists of three terms:

$$\mathcal{L}^{CAT:student} = \lambda_{dt}^{CAT} \mathcal{L}_{KLD}^{CAT} \cdot T^2 + \lambda_{feat}^{CAT} \mathcal{L}_{feat}^{CAT} + \lambda_{cls}^{CAT} \mathcal{L}_{cls}^{CAT:student} \tag{12}$$

where $\lambda_{dt}^{CAT}, \lambda_{cls}^{CAT}$ and λ_{feat}^{CAT} is a set of adjustable parameters measuring the importance between three loss terms.

Multi-view Features Strengthen (MFS) network. In stage two, we use the same structure of CAT to strengthen the multi-view features in the student network. So the mapping of the teacher network and the student network are consistent with that of CAT. But MFS uses only RGB sequences both in the teacher network and the student network. The two networks can be described as:

$$\hat{y}^{teacher} = \mathcal{F}^{teacher}(X_{v_1}, \cdots, X_{v_K}) \tag{13}$$

$$\hat{y}^{student} = \mathcal{F}^{student}(X_{v_k}) \tag{14}$$

Another difference between CAT and MFS is that we remove feature supervision, because the input of the teacher and the student network are both RGB modal, the scene of video sequences are the same as well. The $\mathcal{L}_{KLD}$ loss is enough for the student network to learn multi-view features from the teacher network.

Similar to the loss function in CAT, the full loss function of MFS is defined as follow:

$$\mathcal{L}^{MFS:teacher} = \mathcal{L}_{cls}^{MFS:teacher} \tag{15}$$

$$\mathcal{L}^{MFS:student} = \lambda_{cls}^{MFS} \mathcal{L}_{cls}^{MFS:student} + \lambda_{dt}^{MFS} \mathcal{L}_{KLD}^{MFS} \tag{16}$$

where λ_{cls}^{MFS} is the weight of classification loss and λ_{dt}^{MFS} is the weight of online knowledge distillation loss.

4. Network architecture

In order to improve the operating efficiency and reduce the model scale, the network structure is readjusted. We use I3D [9] as the backbone. The first convolution kernel in ResNet [50] is expanded to 5 × 7 × 7. In addition, some 1 × 1 convolutions in the residual module are expanded to 3 × 1 × 1. After the last residual block, a global average pooling and a fully connected layer are followed to get classification output.

As described in Section 3.3, we apply VAA module after the third residual block. In the meantime, we add feature supervision of the reweighted features output from VAA module, where the multi-view cross-modality aggregated transfer has the best performance according to our experimental results.

5. Experiments

We implement our multi-view action recognition architecture in Pytorch. Our model is evaluated based on two benchmark datasets, IXMAS [51] and N-UCLA [35]. In the following, we will describe details of the training process, experimental results and the corresponding analysis. We train and evaluate on a server with four NVIDIA GeForce GTX 1080 Ti GPUs.

5.1. Dataset

IXMAS is a multi-view human action recognition dataset. It contains 11 actions, such as checking a watch, getting up, waving, kicking, etc. The dataset is relatively small in size and has a total of 1695 labeled video clips. Each action is performed by 10 actors for three times, which is captured synchronously by four horizontal cameras and one vertical camera.

N-UCLA is another multi-view action recognition benchmark dataset. It contains 10 daily actions, which are captured by three static cameras and are performed by 10 subjects for several times. In total, the dataset consists of 1475 RGB videos and the correlated depth frames and skeleton information. We only use the RGB videos in our experiments.

5.2. Implementation details

Detailed training process. In the first stage, the teacher network and the student network are trained end-to-end simultaneously. The teacher network is optimized based on the classification loss $\mathcal{L}_{cls}^{CAT:teacher}$, while the student network is optimized based on the feature regression loss $\mathcal{L}_{feat}^{CAT}$, classification loss $\mathcal{L}_{cls}^{CAT:student}$ and online knowledge distillation loss $\mathcal{L}_{KLD}^{CAT}$. During our experiments, we attempt to calculate the regression loss $\mathcal{L}_{feat}^{CAT}$ in ℓ_2 constraint and supervise the middle feature maps of multiple layers at the same time. However, the test results are not as good as that of the single layer with the ℓ_1 loss. Both the teacher and the student network are initialized with ImageNet weights. Besides, we explore the position of the feature loss. As shown in Fig. 4, we sample one sequence from the N-UCLA dataset, and visualize the feature maps output from the first residual block and the third residual block of both the student and the teacher network. Consistent with the observations from previous works [29], the lower layers are modality-specific and thus harder to transfer across modalities. While the visualizations from higher layers are abstract, and the two modalities share one semantic space and semantic representation. As a result, it is reasonable to add a feature loss after VAA.

The second stage, which is still end-to-end jointly training, is designed to enhance multi-view global information through further online distillation. The teacher is optimized based on the classification loss $\mathcal{L}_{cls}^{MFS:teacher}$ and the student network is optimized based on the classification loss $\mathcal{L}_{cls}^{MFS:student}$ and distillation loss $\mathcal{L}_{KLD}^{MFS}$. We initialize the student network in MFS based on the student network in CAT, while the teacher network is initialized by the weights that are trained with a set of RGB video sequences separately with a certain iteration.

Detailed training parameters. Since the action fragments in these two benchmark datasets are relatively short, we sample 1 frame from every 4 frames, thus, a total of 8 frames are taken from a fragment. During the training stage, we use Adam [62] optimizer for all the modules and set $\beta_1 = 0.99, \beta_2 = 0.999$. For CAT, the learning rate is set to $2e^{-3}$, we set loss weights $\lambda_{dt}^{CAT}, \lambda_{feat}^{CAT}, \lambda_{cls}^{CAT}$ to 1, 2, and 1 respectively. Since both teacher and student networks are optimized by a certain number of iterations, the learning rate of MFS is initialized to $2e^{-4}$, and the λ_{dt}^{MFS} and λ_{cls}^{MFS} are both set to 1. The learning rate decays 10 times in every 100 epochs. We train CAT for 200 epochs and MFS for 100 epochs, with batch size of 12. Besides, the temperature value T is set to 2 empirically.

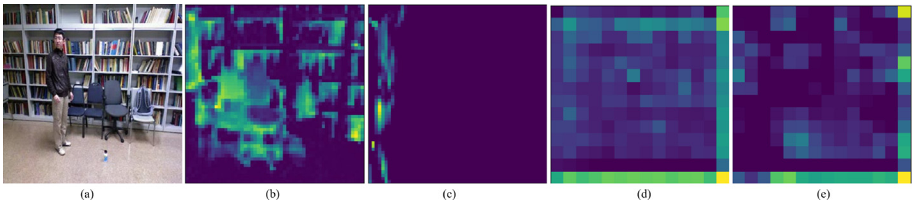


Fig. 4. *Visualization of the original frame and corresponding feature maps.* (a) visualizes the first frame from the sampled sequence on the N-UCLA dataset. (b), (c) visualize the same channel output from the first residual block of the student network and the teacher network, respectively. (d), (e) visualize the same channel- output from the third residual block of the student network and the teacher network, respectively.

5.3. Comparison with state-of-the-arts

We conduct experiments on *leave-one-view-out* protocol, in which we use video sequences from one view as the test set, and employ video sequences from the remaining views to train our model. During training, for the student network, we randomly select a video sequence and get the corresponding actor ID, while for the teacher network, we input the video sequences from the same actor.

For the IXMAS dataset, we conduct fivefold cross-validation. The video sequences from the unseen view are not available during training. We report our results in Table 1. Some previous methods trained with one source view and test with another target view. In contrast, our method is trained with three source views and test with the remaining one target view. For a fair comparison, we average the accuracy of them when a certain camera is used for testing. The upper part of the table shows the results of non-CNN based methods, while the methods of lower part introduce CNN for feature learning. Our model performs better than both CNN based and non-CNN based methods. Compared with I3D, the average accuracy increases from 83.0% to 87.8% for five views and from 92.9% to 97.3% for four views. However, the average accuracy of the top view as the test set is generally poor. There are two reasons that account for these results. First, the backgrounds and the actors' pose captured from the horizontal cameras are different from that of the vertical cameras. Second, it is difficult to obtain a reliable human representations from the vertical cameras. Besides, we note that the TDL and Temporal 3D gain the competitive results under camera 4. The TDL process pre-training phase that utilizes a large number of automatically synthesized multi-view 2D and 3D videos to learn 3D dense trajectories, and then train a view-invariant action classifier using 2D videos of IXMAS. After such training phase, the TDL could capture robust view-invariant representations under camera 4. For the temporal 3D, it use the TCN to lift from the 2D space to the 3D space. In TCN, temporal information of the skeleton is considered, the temporal coherence reduces the noisy joints estimates. For our method, the 3D DensePose representations extracted from the pre-trained DensePose-RCNN, and the pre-training dataset is the DensePose-COCO, which has little overhead view data compared with the synthesized dataset of TDL. Moreover, DensePose-RCNN estimate DensePose per frame without temporal consistency in Temporal 3D. The above two reasons may account for the lower accuracy of our method under camera 4, but it doesn't make sense because of its generally low accuracy.

For the N-UCLA dataset, we conduct threefold cross-validation. From the comparison results shown in Table 2, our method achieves the best performance and improves the average accuracy by 5.9% compared to I3D. In detail, when cam_1 and cam_3 are used as test set separately, our method achieves an accuracy of 85.2% and 91.1%, which is 1.8% and 1% higher than the Glimpse Clouds, respectively. However, as shown in Table 2, the Glimpse Clouds gain more competitive results than our method under camera 2. We also find that the accuracy under camera 3 is usually higher than that under camera 2, and the accuracy under camera 2 is usually higher than that under camera 1. It is obvious that the accuracy and scene are highly correlated. The possible reason for the comparative results between the proposed method and the Glimpse Clouds under camera 2 lies in that the Glimpse Clouds focus on view-invariant spatial features of the scene captured from camera 1 and camera 3 due to its visual attention module, which is crucial to recognize activities under camera 2. While our method that without a spatial attention module could not effectively obtain discriminative features from camera 1 and camera 3. In general, the average accuracy of our method achieves much higher than any other method.

In summary, the results on two benchmark datasets demonstrate the effectiveness of our method. We perform better than the other state-of-the-art methods by a large margin, especially on the IXMAS dataset.

5.4. Ablation study

In this section, we conduct several ablation studies on the IXMAS dataset and the N-UCLA dataset in the *leave-one-view-out* protocol to analyze the contributions of the two-stage structure and VAA module, as well as the effectiveness of different loss terms. First, we give descriptions of different methods.

- 1: $\mathcal{L}_{cls}$. We optimize the student network for single-view RGB sequences. This method only uses classification loss for training.
- 2: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}, fixed$. Method 2 denotes the original knowledge distillation method. The teacher network has been trained in advance, the parameters are fixed to provide supervision for the student network.
- 3: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}, online$. We replace the original knowledge distillation with the online distillation based on method 2.
- 4: $\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$. On the basis of method 3, we add feature supervision during training, which use the features obtained by the specific layer in the teacher network to supervise the corresponding features of the student network. The above four methods are only given RGB video sequences.
- 5: $CAT : \mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}, MFS : \mathcal{L}_{cls} + \mathcal{L}_{KLD}$. This method introduces our two-stage model. In stage one, we use cross-modality video sequences, apply online distillation strategy and feature supervision during training, but without VAA module. In stage two, we further argument multi-view features in the student network, attention module is not applied either.
- 6: $CAT : \mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat} + VAA, \; MFS : \mathcal{L}_{cls} + \mathcal{L}_{KLD} + VAA$. The complete implementation of our method. We add VAA module based on method 5.

Influence of online distillation. To evaluate the effectiveness of online distillation, we compare the first three rows in Tables 3 and 4. For the IXMAS dataset, we observe that using the distillation training strategy proposed by Hinton et al. [28] does not improve

Table 1
Average accuracy comparison to the state-of-the-art action classification on IXMAS. cam_0 indicates that cam_0 is used as the test, while the other camera views are used as the train test, and so on. $cam_0 - cam_3$ are the horizontal views, and cam_4 is the top view. The seventh column of the table is the average accuracy values of the 5 camera views ($cam_0 - cam_4$), while the rightmost column is the average accuracy values except for the horizontal camera views ($cam_0 - cam_3$).

Method	cam_0	cam_1	cam_2	cam_3	cam_4	Avg_{0-4}	Avg_{0-3}
VKT [52]	11.6	10.8	10.3	10.3	11.4	10.9	10.7
DVV [53]	47.5	49.1	23.3	43.3	27.6	38.2	40.8
CVP [54]	52.0	53.5	25.9	47.6	31.9	42.2	44.7
Hankelets [55]	60.9	60.8	65.0	57.4	38.0	56.4	61.0
nCTE [2]	73.5	75.4	72.5	72.4	42.6	67.3	73.4
NKTM [33]	79.4	75.4	80.8	76.8	50.2	72.5	78.1
TDL [13]	72.9	79.8	74.3	77.0	51.8	71.1	76.0
Avola et al. [56]	80.5	76.4	73.2	78.0	–	77.0	–
Temporal 3D [57]	80.0	76.5	80.0	79.2	**60.2**	78.9	75.2
I3D [9]	95.3	94.1	91.4	90.9	43.4	83.0	92.9
Ours	**98.5**	**96.5**	**96.5**	**97.6**	50.1	**87.8**	**97.3**

Table 2
Average accuracy comparison to the state-of-the-art action classification on N-UCLA dataset. cam_1 indicates that cam_1 is used as the test, while the other camera views are used as the train test, and so on. The rightmost column of the table are the average accuracy values of 3 camera views ($cam_1 - cam3$).

Method	cam_1	cam_2	cam_3	Avg_{1-3}
DVV [53]	39.3	55.2	58.5	51.0
CVP [54]	39.5	55.8	60.6	52.0
H-RNN [58]	–	–	–	78.5
motion+STD [59]	83.4	88.2	84.5	85.4
DA-Net [27]	83.1	82.7	86.5	84.1
Enhanced viz. [60]	-	-	-	86.1
Glimpse Clouds [61]	83.4	**89.5**	90.1	87.6
I3D [9]	75.1	81.2	89.3	81.9
Ours	**85.2**	87.0	**91.1**	**87.8**

Table 3
Ablation study of our approach based on the IXMAS dataset. The evaluation is performed in the case when cam_0 serves as the test set and the others as the train tests.

	Modality	Accuracy
$\mathcal{L}_{cls}$	RGB	95.28
$\mathcal{L}_{cls} + \mathcal{L}_{KLD}$, *fixed*	RGB	95.28
$\mathcal{L}_{cls} + \mathcal{L}_{KLD}$, *online*	RGB	95.58
$\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$	RGB	95.87
CAT: $\mathcal{L}_{cls} + \mathcal{L}_{KLD} + L_{feat}$	Mix	97.05
MFS: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}$	Mix	97.35
CAT: $\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$*+VAA*	Mix	97.94
MFS: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}$*+VAA*	Mix	**98.53**

Table 4
Ablation study of our approach based on the N-UCLA dataset. The evaluation is performed in the case when cam_1 serves as the test set and the other ones as the train tests.

	Modality	Accuracy
$\mathcal{L}_{cls}$	RGB	75.10
$\mathcal{L}_{cls} + \mathcal{L}_{KLD}$, *fixed*	RGB	83.07
$\mathcal{L}_{cls} + \mathcal{L}_{KLD}$, *online*	RGB	83.27
$\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$	RGB	84.05
CAT: $\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$	Mix	84.24
MFS: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}$	Mix	84.82
CAT: $\mathcal{L}_{cls} + \mathcal{L}_{KLD} + \mathcal{L}_{feat}$*+VAA*	Mix	84.44
MFS: $\mathcal{L}_{cls} + \mathcal{L}_{KLD}$*+VAA*	Mix	**85.21**

the accuracy compared with the baseline. The reason could be the relatively small size of the dataset, which makes the trained model easily overfitted. In this case, the prediction result output from the teacher network could not properly guide the training of the student network. By contrast, the online distillation method trains the two networks simultaneously so that the student network can obtain multi-view prediction results in real-time, and update the parameters based on multiple iterations of the teacher network. The teacher network is able to gradually guide the student network, reaching a recognition accuracy of 95.58%, outperforming method 2 by 0.3%. For the N-UCLA dataset, when adding distillation strategy, the accuracy of method 2 is 7.97% higher than the baseline, which illustrates that the distillation strategy works well. Furthermore, after applying online distillation, compared with method 2, the average accuracy increases from 83.07% to 83.27%, which is an improvement of 0.2%. The improvement shows the superior of the proposed online distillation.

Influence of feature supervision. Compared with the third row and the fourth row in Tables 3 and 4, it is obvious that with the help of feature supervision, the performance of method 4 achieves an improvement of 0.29% for the IXMAS dataset, and 0.78% for the N-UCLA dataset. Results demonstrate that it has more details in feature maps than in the final probability distribution of categories.

Influence of cross-modality data. In the fourth row and the fifth row in Tables 3 and 4, RGB means only the RGB video sequences are used, and Mix means RGB and DensePose are both used in the experiments. When the teacher network is given the multi-view DensePose sequences, the student network achieves an accuracy of 97.05% for the IXMAS dataset and 84.24% for the N-UCLA dataset, which is 1.18% and 0.19% higher than method 4, respectively. This group of comparative experiments proves that multi-modality input can really improve the performance of multi-view action recognition. Our method utilizes the geometric information in DensePose human body representation and captures essential view-invariance features, which alleviates the over-fitting of the model when data size is quite small.

Influence of two-stage structure. By comparing the fifth row and the sixth row or the seventh row and the eighth row of Tables 3 and 4, the accuracies of the methods using the two-stage structure are all improved. For the IXMAS dataset, the model with the two-stage structure performs better than that with the single-stage structure by 0.3% without VAA and 0.59% with VAA, respectively. For the N-UCLA dataset, the two-stage-based method has an improvement of 0.58% without VAA and 0.77% with VAA, respectively. It is worth noting that our two-stage based method further brings a significant boost in multi-view action recognition.

Influence of VAA. As shown in Tables 3 and 4, it is obvious that the VAA module contributes to a much better performance than without it. In detail, the train test of the IXMAS dataset includes a cam_4 view, in which the DensePose sequences are inaccurate. Our attention module can weaken the useless information and strengthen the useful information. The average accuracy increases from 97.35% to 98.53%, with an improvement of 1.18%. Similarly, for the N-UCLA dataset, the features of the three views are equally critical. The attention module could fine-tune the weights of different views to get a much higher average accuracy. In our test, the average accuracy is improved by 0.39% with VAA.

6. Conclusion

In this paper, we address the challenge of learning critical representations leveraging multi modalities at train time, while missing modalities at test time. Specifically, our method takes DensePose sequences as 3D modal data. The entire training process is divided into two stages and are both optimized in an online distillation manner. In the first stage, we propose a Cross-modality Aggregated Transfer (CAT) network that contains a Viewpoint-Aware Attention (VAA) module to effectively aggregate multi-view features. The teacher network of CAT learns multi-view geometry-and-texture features and transfers to the student network by feature loss and online distillation loss, the student network is only fed with single-view RGB sequences. In the second stage, a Multi-view Features Strengthen (MFS) network that contains VAA as well is designed to further strengthen the multi-view features of the student network with classified loss and online distillation loss. Through our proposed learning paradigm, the student network captures more discriminative features and could be deployed in real life without increasing cost. Extensive experiments based on two multi-view datasets demonstrate the superiority of our proposed method. Besides, we conduct ablation study to verify that all parts of our paradigm contribute to the model performance.

CRediT authorship contribution statement

Chao Xu: Conceptualization, Writing - original draft, Methodology, Software. **Xia Wu:** Data curation, Software, Validation. **Yachun Li:** Methodology, Software. **Yining Jin:** Writing - review & editing. **Mengmeng Wang:** Writing - review & editing. **Yong Liu:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 61836015 and No. 61771193.

References

[1] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian, Human daily action analysis with multi-view and color-depth data, in: European Conference on Computer Vision, Springer, 2012, pp. 52–61.

[2] A. Gupta, J. Martinez, J.J. Little, R.J. Woodham, 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2601–2608.

[3] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin, Action recognition by dense trajectories, 2011.

[4] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[5] Y. Zhu, Z. Lan, S. Newsam, A. Hauptmann, Hidden two-stream convolutional networks for action recognition, in: Asian Conference on Computer Vision, Springer, 2018, pp. 363–378.

[6] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[7] M. Tong, M. Zhao, Y. Chen, H. Wang, D3-lnd: A two-stream framework with discriminant deep descriptor, linear cmdt and nonlinear kcmdt descriptors for action recognition, Neurocomputing 325 (2019) 90–100.

[8] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.

[9] J. Carreira, A. 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.

[10] Z. Zheng, G. An, D. Wu, Q. Ruan, Spatial-temporal pyramid based convolutional neural network for action recognition, Neurocomputing 358 (2019) 446–455.

[11] M.B. Holte, T.B. Moeslund, N. Nikolaidis, I. Pitas, 3d human action recognition for multi-view camera systems, in: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, IEEE, 2011, pp. 342–349.

[12] Y. Kong, Z. Ding, J. Li, Y. Fu, Deeply learned view-invariant features for cross-view action recognition, IEEE Transactions on Image Processing 26 (6) (2017) 3028–3037.

[13] J. Zhang, H.P. Shum, J. Han, L. Shao, Action recognition from arbitrary views using transferable dictionary learning, IEEE Transactions on Image Processing 27 (10) (2018) 4709–4723.

[14] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[15] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, K.-K.R. Choo, Adaptive fusion and category-level dictionary learning model for multi-view human action recognition, IEEE Internet of Things Journal.

[16] Z. Cai, L. Wang, X. Peng, Y. Qiao, Multi-view super vector for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 596–603.

[17] X. Ji, C. Wang, Y. Li, A view-invariant action recognition based on multi-view space hidden markov models, International Journal of Humanoid Robotics 11 (01) (2014) 1450011.

[18] Y. Liu, L. Nie, L. Liu, D.S. Rosenblum, From action to activity: sensor-based activity recognition, Neurocomputing 181 (2016) 108–115.

[19] Y. Li, Y. Liu, C. Zhang, What elements are essential to recognize human actions?.

[20] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.

[21] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with directed graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7912–7921.

[22] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J.T. Zhou, X. Bai, Action recognition for depth video using multi-view dynamic images, Information Sciences 480 (2019) 287–304.

[23] B. Sheng, J. Li, F. Xiaoc, Q. Li, W. Yang, J. Han, Discriminative multi-view subspace feature learning for action recognition, IEEE Transactions on Circuits and Systems for Video Technology.

[24] L. Wang, Z. Ding, Z. Tao, Y. Liu, Y. Fu, Generative multi-view human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6212–6221.

[25] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, M. Kankanhalli, Benchmarking a multimodal and multiview and interactive dataset for human action recognition, IEEE Transactions on Cybernetics 47 (7) (2016) 1781–1794.

[26] Z. Gao, S. Li, Y. Zhu, C. Wang, H. Zhang, Collaborative sparse representation leaning model for rgbd action recognition, Journal of Visual Communication and Image Representation 48 (2017) 442–452.

[27] D. Wang, W. Ouyang, W. Li, D. Xu, Dividing and aggregating network for multi-view action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 451–467.

[28] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531.

[29] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2827–2836.

[30] R. Alp Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.

[31] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, T. Xiang, Exemplar-based recognition of human–object interactions, IEEE Transactions on Circuits and Systems for Video Technology 26 (4) (2015) 647–660.
[32] Y. Ji, H. Cheng, Y. Zheng, H. Li, Learning contrastive feature distribution model for interaction recognition, Journal of Visual Communication and Image Representation 33 (2015) 340–349.
[33] H. Rahmani, A. Mian, Learning a non-linear knowledge transfer model for cross-view action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2458–2466.
[34] H. Rahmani, A. Mahmood, D. Huynh, A. Mian, Histogram of oriented principal components for cross-view action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (12) (2016) 2430–2443.
[35] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.
[36] Z. Luo, J.-T. Hsieh, L. Jiang, J. Carlos Niebles, L. Fei-Fei, Graph distillation for action detection with privileged modalities, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 166–183.
[37] J.C. Stroud, D.A. Ross, C. Sun, J. Deng, R. Sukthankar, D3d: Distilled 3d networks for video action recognition, arXiv preprint arXiv:1812.08249.
[38] M.-C. Wu, C.-T. Chiu, K.-H. Wu, Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2202–2206.
[39] F.M. Thoker, J. Gall, Cross-modal knowledge distillation for action recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 6–10.
[40] S. Bhardwaj, M. Srinivasan, M.M. Khapra, Efficient video classification using fewer frames, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 354–363.
[41] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, W.-S. Zheng, Progressive teacher-student learning for early action prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3556–3565.
[42] X. Qin, Y. Ge, J. Feng, D. Yang, F. Chen, S. Huang, L. Xu, Dtmmn: Deep transfer multi-metric network for rgb-d action recognition, Neurocomputing 406 (2020) 127–134.
[43] L. Wang, C. Gao, L. Yang, Y. Zhao, W. Zuo, D. Meng, Pm-gans: Discriminative representation learning for action recognition using partial-modalities, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 384–401.
[44] J. Stroud, D. Ross, C. Sun, J. Deng, R. Sukthankar, D3d: Distilled 3d networks for video action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 625–634.
[45] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[46] Z. Li, S. Zhang, J. Zhang, K. Huang, Y. Wang, Y. Yu, Mvp-net: Multi-view fpn with position-aware attention for deep universal lesion detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 13–21.
[47] J. Wang, W. Bao, L. Sun, X. Zhu, B. Cao, S.Y. Philip, Private model compression via knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 1190–1197.
[48] A. Mishra, D. Marr, Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy, arXiv preprint arXiv:1711.05852.
[49] G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, Learning efficient object detection models with knowledge distillation, in: Advances in Neural Information Processing Systems, 2017, pp. 742–751.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[51] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2–3) (2006) 249–257.
[52] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, in: CVPR 2011, IEEE, 2011, pp. 3209–3216.
[53] R. Li, T. Zickler, Discriminative virtual views for cross-view action recognition, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2855–2862.
[54] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, C. Shi, Cross-view action recognition via a continuous virtual path, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2690–2697.
[55] B. Li, O.I. Camps, M. Sznaier, Cross-view activity recognition using hankelets, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1362–1369.
[56] D. Avola, M. Cascio, L. Cinque, G.L. Foresti, C. Massaroni, E. Rodolà, 2d skeleton-based action recognition via two-branch stacked lstm-rnns, IEEE Transactions on Multimedia.
[57] M.A. Musallam, R. Baptista, K. Al Ismaeil, D. Aouada, Temporal 3d human pose estimation for action recognition from arbitrary viewpoints, in: 2019 International Conference on Computational Science and Computational Intelligence (CSCI)s, IEEE, 2019, pp. 253–258.
[58] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[59] C. Dhiman, D.K. Vishwakarma, View-invariant deep architecture for human action recognition using late fusion, arXiv preprint arXiv:1912.03632.
[60] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognition 68 (2017) 346–362.
[61] F. Baradel, C. Wolf, J. Mille, G.W. Taylor, Glimpse clouds: Human activity recognition from unstructured feature points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 469–478.
[62] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Chao Xu received the B.S. degree in electrical engineering and its automation from Nanchang University, Nanchang, China, in 2018.

He is currently working toward the M.S. degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China.

His major research interests include video understanding and video generation.

Xia Wu received the B.S. degree in electronic information from China Agricultural University, Beijing, China, in 2018. She is currently working toward the M.S. degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China.

Her major research interests include computer vision and temporal action recognition.

Yachun Li received the B.S. degree in automation from Zhejiang University, in 2016, the M.S. degree in Visual Computing and Robotics from Imperial College London, in 2017, and the M.Eng. degree in control engineering from Zhejiang University, in 2019.

Her research interests include computer vision and video understanding.

Yining Jin is currently working toward the BSc in Electrical Engineering Co-op at the University of Alberta. Her major research interests include human action recognition, AI chips and operator development.

Mengmeng Wang received the B.S. degree and M.S. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2015 and 2018.
She is currently working toward the doctor degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China.
Her major research interests include computer vision and deep learning.

Yong Liu received the B.S. degree in computer science and engineering and the Ph.D degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively. He is currently a professor of Institute of Cyber-Systems and Control at Zhejiang University.
His main research interests include: intelligent robot systems, robot perception and vision, deep learning, big data analysis, and multi-sensor fusion. He has published over 30 research papers on machine learning, computer vision, information fusion, and robotics.