MULTI-TASK LEARNING OF GENERALIZABLE REPRESENTATIONS FOR VIDEO ACTION RECOGNITION

Zhiyu Yao, Yunbo Wang*, Mingsheng Long (✉), Jianmin Wang, Philip S. Yu, and Jiaguang Sun*

School of Software, BNRist, Tsinghua University, China
Research Center for Big Data, Tsinghua University, China
Beijing Key Laboratory for Industrial Big Data System and Application
{yaozy19,wangyb15}@mails.tsinghua.edu.cn, {mingsheng,jimwang,psyu,sunjg}@tsinghua.edu.cn

ABSTRACT

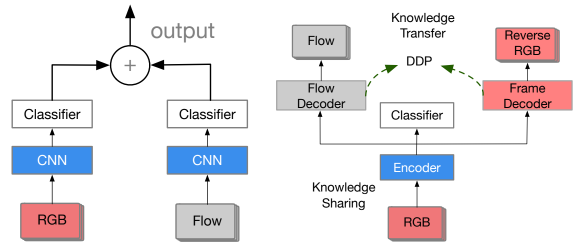
In classic video action recognition, labels may not contain enough information about the diverse video appearance and dynamics, thus, existing models that are trained under the standard supervised learning paradigm may extract less generalizable features. We evaluate these models under a cross-dataset experiment setting, as the above *label bias* problem in video analysis is even more prominent across different data sources. We find that using the optical flows as model inputs harms the generalization ability of most video recognition models.

Based on these findings, we present a multi-task learning paradigm for video classification. Our key idea is to avoid label bias and improve the generalization ability by taking data as its own supervision or supervising constraints on the data. First, we take the optical flows and the RGB frames by taking them as auxiliary supervisions, and thus naming our model as Reversed Two-Stream Networks (**Rev2Net**). Further, we collaborate the auxiliary flow prediction task and the frame reconstruction task by introducing a new training objective to Rev2Net, named Decoding Discrepancy Penalty (**DDP**), which constraints the discrepancy of the multi-task features in a self-supervised manner. Rev2Net is shown to be effective on the classic action recognition task. It specifically shows a strong generalization ability in the cross-dataset experiments.

***Index Terms*—** Video action recognition, self-supervised learning, multi-task learning

1. INTRODUCTION

Learning generalizable representations is a new direction and yet an important problem in video analysis, considering that a good action recognition model should be able to handle various video environments. Different from most previous methods for video data, this paper discusses more about the generalization ability. Because of the *label bias* problem that coarse video-level labels may only express short snippets of the entire untrimmed videos (within a single video), and may not contain enough information for the various video scenarios, including frame appearance and long-term action dynamics (across different videos). Thus, the traditional strongly supervised learning paradigm that is used by most existing video action models [1, 2, 3] may suffer from the label bias problem and leads to less generalizable spatiotemporal features.



(a) Two-Stream 2D/3D CNNs (b) Rev2Net

Fig. 1. We propose a self-supervised multi-task learning framework to learn more generalizable video representations for action recognition. The proposed model, Rev2Net, differs from previous two-stream networks from the perspectives of knowledge transfer and knowledge sharing in the training procedure.

In this paper, we first evaluate the generalization ability of the most widely used video classification models [2, 3, 4] in a cross-dataset setting, where it exists more severe label bias problem caused by different data sources. We find that these two-stream networks that are based on multi-modality inputs degenerate in such experiment settings. Particularly, we observe that the reason of this issue is that the optical flow features are **NOT generalizable**. Therefore, the multi-modality video data is not used properly from the perspective of domain generalization, and so solving this problem requires a different learning paradigm.

Based on the above intuitive and empirical motivations, we present a new self-supervised, multi-task learning approach for video action recognition. The key idea of this approach is to improve the generalization ability of the learned features by leveraging the multi-modality data as its own supervisions, and supervising the specific constraints on the data. Concretely, we propose a new multi-task network architecture named Reversed Two-Stream Networks (**Rev2Net**). As shown in Figure

* Equal contribution

1(b), we train this network to classify video sequence, predict optical flows from RGB frames, and reconstruct the frames in a reverse order simultaneously. Rev2Net enables knowledge transfer from the auxiliary self-supervised tasks to the core supervised task through a common frame encoder. To further enhance the knowledge sharing between flow prediction and frame reconstruction task, we attempt to constrain the discrepancy of the two flow/frame decoders by applying a new training objective named Decoding Discrepancy Penalty (**DDP**) to both high-level and low-level features of Rev2Net. Both the ideas of knowledge sharing and knowledge transfer are shown to benefit the generalization ability. Rev2Net achieves competitive results on classic action recognition tasks within each of the UCF-101, HMDB-51, and Kinetics datasets. It also achieves the best performance in our cross-dataset experiment.

To sum up, this paper has the following contributions:

- We design a cross-dataset experiment to evaluate the generalization ability of video recognition models, and observe that the features extracted from the optical flow data are less generalizable.
- We present a new model named Rev2Net that can learn more generalizable features through a new multi-task learning framework with self-supervisions.
- We propose the DDP loss to encourage the collaboration of multiple auxiliary self-supervised tasks.

2. RELATED WORK

Ever since the significant impact of CNNs upon image classification, many researchers have been trying out various CNN architectures for video action recognition, including 2D CNNs, 3D CNNs, and non-local modules [5]. For all these models, the two-stream network framework with multi-modality inputs has been most widely explored and proved to be effective.

Two-Stream CNN Models. The two-stream networks were first introduced to video analysis by Simonyan *et al.* [1], which averages the classification results of a sub-network with frame inputs and another sub-network with pre-computed optical flow inputs. The optical flow sub-network brings in significant performance gains due to the capability of capturing short-term video dynamics. Since then, two-stream networks have been widely employed by many action recognition models [6, 2, 7, 8], including 3D CNN models.

3D CNN Models. More recently, 3D CNNs have been widely explored [9, 10, 3, 11, 12]. These models realized unified modeling of the spatiotemporal features. But the 3D-Convs bring an inevitable increase in the number of network parameters, making these models hard to train. Carreira *et al.* proposed the I3D model [3] that inflates the filters of the 2D-Convs into 3D, making this model implicitly pretrained on ImageNet. Note that our proposed Rev2Net is mainly based on I3D. In Section 3, we will show that the above two-stream CNNs and 3D CNNs learn less generalizable features in a cross-dataset experiment. This paper attempts to solve this problem.

Optical Flow Prediction and Multi-task Models. Using optical flows as model inputs has been proved to be crucial for the performance of two-stream networks. However, some recent methods [13, 14, 15] showed that training to generate optical flows with some deep networks, e.g., FlowNet [16] and SpyNet [17], could further improve the recognition performance. In this paper, we go beyond optical flow estimation and propose a new method to collaborate multiple auxiliary tasks.

3. ANALYSIS OF GENERALIZATION ABILITY

We evaluate the generalization ability of video action recognition models with the cross-dataset experiments, for the reason that in such settings, the *label bias* problem is more prominent. The videos from different datasets may have distinct scenes and action dynamics. Robust action recognition models should be easily adapted from one dataset to another.

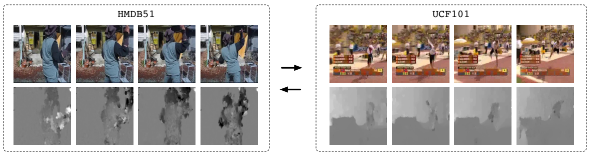


Fig. 2. Videos under the same category ("Archery" in this case) from different datasets have diverse frame scenes and action dynamics, which can better evaluate the generalization ability of the action recognition models. Good video representations should be generalizable to the cross-dataset variations.

Cross-Dataset Experiment. We design the cross-dataset experiment to verify the generalization ability of video action recognition models, adapting the domain adaptation settings from image data to video data. In general, we need to learn a model from the source dataset and to apply this model on the target dataset. More concretely, we select 16 related categories from the UCF101 [18] and HMDB51 [19] datasets. Note that these two datasets have diverse data patterns: HMDB51 is mostly collected from movies, while UCF101 is collected from YouTube and appears to be closer to real life. Even for the same action category, the video appearances of these two datasets are quite different, e.g. the scene complexity and the camera angles as shown in Figure 2. Under these circumstances, it would be more likely to learn invariant features across datasets from similar motion cues.

We include the most prevalent action recognition networks in our cross-dataset experiment: TSN [2], TSM [4] for 2D CNNs, and I3D [3] for 3D CNNs. To further boost the generalization ability of these models, we incorporate a domain adaptation method that was originally proposed for image data [20], which closes the distributions by matching the mean embeddings in the feature space across different domains.

Table 1. We explore the generalization ability of existing action recognition models using the cross-dataset experiment. We specifically apply a domain adaptation method DANN [20] to further enhance the generalization ability of these models.

Model	Input	HMDB → UCF	UCF → HMDB
TSN	RGB	58.5	40.1
TSN	Flow	33.2	23.0
TSN	RGB + Flow	60.9	40.3
TSM	RGB	58.9	41.2
TSM	Flow	35.1	24.0
TSM	RGB + Flow	61.2	41.5
I3D	RGB	56.4	41.0
I3D	Flow	45.0	31.1
I3D	RGB + Flow	57.9	41.5

Table 1 shows the cross-dataset recognition results of the evaluated models with one or two modalities of input data. Surprisingly, we observe that for both 2D CNNs and 3D CNNs, the frame network consistently outperforms the flow network, which is against our experience on the classic video action recognition. Consequently, the overall recognition accuracy of the two-stream networks yields no further improvement compared with the one-stream RGB network. This observation violates our expectations and our perceptions about the two-stream architecture on the classic video classification task. Under the framework of the traditional two-stream networks that take optical flow as inputs, the only way to improve the overall cross-dataset performance is to improve the generalization ability of the optical flow stream. From these results, we may conclude that neither 2D CNNs nor 3D CNNs shows great generalization ability with optical flow inputs. We conjecture that optical flow data conveys very different action dynamics across datasets, leading to less generalizable features. Later, we will show that our proposed Rev2Net can improve the performance in the same cross-dataset experiment.

4. REVERSED TWO-STREAM NETWORKS

In this section, we first present the Reversed Two-Stream Networks (Rev2Net) trained in a multi-task framework with self-supervisions. We then discuss a new training objective to further constrain the discrepancy and encourage the collaboration between the multiple auxiliary tasks of Rev2Net.

4.1. A Self-Supervised Multi-Task Learning Framework

As the optical flow stream shows limited generalization ability in cross-dataset setting, this stream severely affects the overall performance of the two-stream model. Thus, using optical flow as inputs is not an ideal approach for learning generalizable video features. In contrast, our approach distills more generalizable knowledge by taking data as its own supervision or supervising constraints on the data. Specifically, we present the Reversed Two-Stream Networks (Rev2Net), which is trained in a multi-task learning framework with self-supervision from the multi-modality data. The schematic of Rev2Net is shown in Figure 3. Rev2Net has four components: one encoder stream that only takes RGB frames as inputs, one classifier for action recognition, one decoder stream for optical flow prediction, and another decoder stream for reversed frame reconstruction. The encoder stream operates on consecutive 32 video frames. Along with the classifier, it has the same architecture as I3D. The two decoder streams are composed of 3D transpose convolutions. The flow decoder aims to emphasize learning the **short-term motion** features as well the foreground appearance features by using corresponding flow fields as supervisions. The frame decoder, on the other hand, aims to emphasize learning the **long-term motion** features by reconstructing the input frames in a reversed order from encoded 3D feature volume, which can be viewed as an information bottleneck to force the model to learn high-level video representations. Note that these two decoder streams are only used during the training procedure. At test time, Rev2Net makes decisions without the pre-computed optical flow inputs. In other words, it avoids the disturbance of optical flow data, which has been shown in the cross-dataset experiment to harm the feature transferability.

Why is the multi-task framework useful to learn more generalizable features. Because it encourages the encoder stream to simultaneously learn short-term and long-term features from these two auxiliary tasks. We believe that comprehensively capturing the temporal relations at multiple time-scales can effectively cover various video dynamics across different video data sources. Further, the multi-task framework also enables knowledge transfer from the self-supervisions, from the short-term video dynamics from the optical flow data to the encoded frame features through an end-to-end training. Note that the decoders are removed at test time, and thus Rev2Net can be seen as an I3D model at test time.

4.2. Decoding Discrepancy Penalty

While traditional two-stream networks suffer from the decision discrepancy between the two network streams in the cross-dataset setting, Rev2Net avoids this issue by using only one frame encoder at test time. However, the self-supervised auxiliary tasks may introduce a new problem of decoding discrepancy to our proposed multi-task learning framework, i.e., the auxiliary tasks may not have a collaborative effect for learning features for the core classification task. In other words, though the flow prediction task and the reversed frame reconstruction task will force the encoder to focus on different parts of the input frames, they may not have the common positive and complementary effects on the training the encoder network. To this end, we encourage the knowledge sharing between the two decoder streams in the training procedure, and penalize the distance of their features. We define a new training objective called the Decoding Discrepancy Penalty (DDP). We propose two forms of DDP respectively to the low-level and high-level

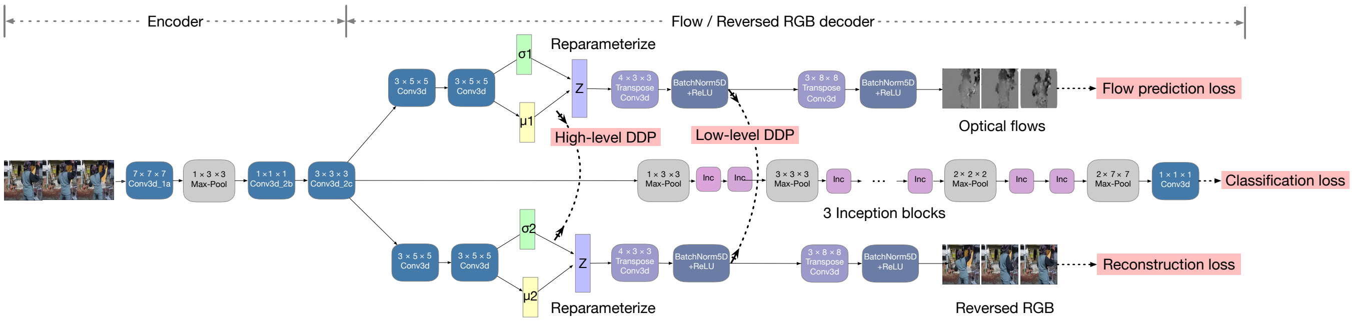


Fig. 3. A schematic of the Rev2Net with decoding discrepancy penalty (DDP). We adopt the inception block from the inflated Inception-V1 [3]. The two decoder streams predicts corresponding optical flow and reconstruct the RGB inputs in a reversed order. DDP is used for overcoming the discrepancy of two decoder streams, allowing them to be trained collaboratively.

features as shown in Figure 3.

DDP Constraints on Low-Level Features. We build the two decoder streams based on the low-level feature maps of the encoder (see Figure 3). In other words, both the decoders and the encoder have only a few convolutional layers. We use the Frobenius norm to penalize the distance of the decoders in the low-level feature space:

$$\mathcal{L}_{\text{DDP}}^{\textbf{low-level}} = \sum_{l \in \mathcal{L}} \| f_{\text{D1}}^{l} - f_{\text{D2}}^{l} \|_F, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, L is a set of convolutional layers included in the DDP, and f_{D1}^{l} and f_{D2}^{l} are the low-level feature maps of the optical flow decoder stream and the reversed frames decoder stream at layer l. As shown in Figure 3. There are two TransposeConv3D layers in the flow decoder and three of them in the frame decoder plus a Conv3D layer for generating the background which is not explicitly shown. These decoders are plugged into the original I3D network, standing on the feature maps of the `Conv3d_2c` layer, taking a feature volume of $16 \times 56 \times 56$ as the inputs. Intuitively, we do not want the decoders to be too strong, since training a good feature encoder may require relatively weak decoders.

DDP Constraints on High-Level Features. We can also mitigate the high-level feature learning discrepancy by allocating more layers into the encoder network and its reversed decoder counterparts. In this method, particularly, the encoder has three parts of outputs: the features that are the inputs of the classifier, the mean μ_1 and the variance σ_1 of the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$ that are used for optical flow prediction, and the mean μ_2 and the variance σ_2 of the Gaussian distribution $\mathcal{N}(\mu_2, \sigma_2)$ that are used for reversed frames reconstruction. We propose to apply KL divergence on the high-level, low-dimensional outputs of the encoder, i.e., $\mu_1, \sigma_1, \mu_2, \sigma_2$, to close the distance of $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$:

$$\mathcal{L}_{\text{DDP}}^{\textbf{high-level}} = D_{\text{KL}}(\mathcal{N}(\mu_1, \sigma_1) \| \mathcal{N}(\mu_2, \sigma_2)). \tag{2}$$

Along with the optical flow prediction objective function or the reversed frames reconstruction objective function, the decoders can be trained similarly as variational autoencoders. By applying DDP to the overall loss function in the training process, including both the Frobenius norm and the KL divergence, we allow the two decoder streams to negotiate and collaborate to improve the consistency of their effects on learning a better encoder.

Final Objective. The final Rev2Net architecture with Frobenius norm DDP and KL divergence is shown in Figure 3. The three tasks, i.e., action recognition, flow prediction, frame reconstruction; along with their corresponding five loss function terms, including the DDP losses, can be jointly trained as

$$\begin{aligned} \mathcal{L}(\mathcal{I}, \mathcal{O}, y) = \mathcal{L}_{\text{CE}}(c, y) + \alpha \mathcal{L}_{\text{DDP}}^{\textbf{high-level}} + \beta \mathcal{L}_{\text{DDP}}^{\textbf{low-level}} \\ + \lambda_{\text{flow}} \|\mathcal{O} - \widehat{\mathcal{O}}\|_F + \lambda_{\text{im}} \|\mathcal{I} - \widehat{\mathcal{I}}\|_F, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss between the softmax output of the classifier c and the ground truth action label y. $\widehat{\mathcal{O}} = \{\widehat{\mathcal{O}}_1, \ldots, \widehat{\mathcal{O}}_\tau\}$ are the generated optical flow, and $\mathcal{O} = \{\mathcal{O}_1, \ldots, \mathcal{O}_\tau\}$ are the target optical flow which are pre-computed using TLV1 [21]. λ_{flow} is the loss weight for the optical flow prediction task. Similarly, $\widehat{\mathcal{I}} = \{\widehat{\mathcal{I}}_1, \ldots, \widehat{\mathcal{I}}_\tau\}$ are generated frames, and $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_\tau\}$ are real input frames. λ_{im} is the loss weight for frames reconstruction task. The hyper-parameters are found using the coarse-to-fine grid-search approach. We first search them using a coarse grid of $\{0, 0.0001, 0.001, 0.1, 1, 10\}$, and then locate the exact values using a fine grid of $\{0, 0.25, \ldots, 1\}$. Finally, we obtain hyper-parameters of $\beta = 0.1$, $\alpha = 0.01$, $\lambda_{\text{im}} = 0.1$, $\lambda_{\text{flow}} = 0.1$.

5. EXPERIMENTS

In this section, we first introduce the datasets and compare Rev2Net with state-of-the-art video classification models for classic action recognition. The compared models include TSN [2], ActionFlowNet [15], Sun-OFF [14], TSM [4], and STM [22]. We also perform ablation studies to explore the respective effectiveness of the multi-task framework and the DDP training objective. Finally, we use the cross-dataset experiments between UCF101 and HMDB51 to validate the generalization ability of Rev2Net.

Table 2. Results on the UCF101 and HMDB51 datasets. All compared models are pretrained on Imagenet.

Model	Input	UCF101	HMDB51
TSN	RGB	86.4	53.7
ActionFlowNet	RGB	83.9	56.4
I3D	RGB	84.5	49.8
Sun-OFF	RGB	93.3	-
Rev2Net w/o frame dec.	RGB	93.8	69.6
Rev2Net w/o flow dec.	RGB	93.1	67.5
Rev2Net w/o DDP	RGB	93.3	69.7
Rev2Net (ours)	RGB	**94.6**	**71.1**
TSN	R + F	94.0	69.4
I3D	R + F	93.4	66.4
Sun-OFF	R + F	96.0	74.2
Rev2Net	R + F	**97.1**	**78.0**

Table 3. Results on the UCF101 and HMDB51 datasets. All compared models are pretrained on Kinetics.

Model	Input	UCF101	HMDB51
TSM	RGB	94.5	70.7
STM	RGB	96.2	72.2
I3D	RGB	95.6	74.8
Rev2Net	RGB	**97.7**	**78.3**
I3D	RGB + Flow	98.0	80.7
Rev2Net	RGB + Flow	**98.7**	**81.9**

5.1. Datasets

We use the following datasets for the classic action recognition, and use the last two of them for the cross-dataset experiments:

- UCF101 [18] that contains 13,320 annotated YouTube video clips from 101 categories. Each video clip lasts 3–10 seconds and consists of 100–300 frames.
- HMDB51 [19] that contains 6,766 video clips and covers 51 action categories. Note that the videos from these datasets are distinct in both scenes and action dynamics.
- Kinetics [23] that contains 306,245 annotated video clips from 400 action categories.

5.2. Classic Action Recognition Results

Comparing with Existing Models. Above all, we show that Rev2Net outperforms the state-of-the-art action recognition deep networks on each of the UCF101 and HMDB51 datasets, with either ImageNet or Kinetics pretrained models (Table 2 and Table 3). It also achieves competitive results on the large scale Kinetics dataset, as shown in Table 4. Rev2Net consistently performs better than other models with the same input modalities, including its base model, I3D. Furthermore, we observe that introducing another I3D stream with optical flow inputs brings additional benefit to Rev2Net.

Table 4. Results on the Kinetics dataset. All compared models are pretrained on ImageNet.

Model	Input	Kinetics
I3D	RGB	71.1
TSM	RGB	72.5
Rev2Net w/o DDP	RGB	72.7
Rev2Net	RGB	**73.3**
I3D	RGB + Flow	74.2
Rev2Net w/o DDP	RGB + Flow	74.1
Rev2Net	RGB + Flow	**74.8**

Table 5. Cross-dataset results. We apply DANN [20] to all compared models to further enhance the transferability.

Model	Input	HMDB → UCF	UCF → HMDB
TSN	RGB	58.5	40.1
TSM	RGB	58.9	41.2
I3D	RGB	56.4	41.0
TSN	R + F	60.9	40.3
TSM	R + F	61.2	41.5
I3D	R + F	57.9	41.5
Rev2Net	RGB	**63.1**	**46.8**

Ablation Studies. We evaluate the Rev2Net models that are trained in different self-supervised, multi-task manners: with or without the frame/flow decoder, and with or without the DDP loss. Table 2 suggests that, first, the frame reconstruction task and the flow prediction task makes individual contributions to the final result, and are complementary to each other. Removing either of the frame decoder or the flow decoder at training time will degenerate the performance of the Rev2Net. Also, a Rev2Net model that is trained without the DDP loss performs even worse than the Rev2Net model that is trained only with the frame decoder (93.3% vs. 93.8% on UCF101), indicating that the multi-task framework may have negative effect without the DDP constraint. In contrast, the DDP strategy could boost the performance of the final Rev2Net which is a multi-task framework (93.3% vs. 94.6% on UCF101).

5.3. Cross-Dataset Action Recognition Results

In Section 3, we have found that the two-stream networks cannot perform well in the cross-dataset experiment setting because the optical flow features are difficult to be transferred. Rev2Net model resolves this problem by enhancing the generalization ability of features from the frame stream, and supervising the network with the optical flows instead of taking them as inputs. From Table 5, Rev2Net remarkably outperforms the TSN, TSM, and I3D models, indicating a strong generalization ability. It improves the base model, I3D, by **11.9%** upon its classification accuracy from UCF101 to HMDB51, and **14.1%** vice versa, with the same inputs of RGB frames.

6. CONCLUSIONS

In this paper, we designed a cross-dataset experiment to evaluate the generalization ability of video recognition models, and observed that the features extracted from the optical flow data are less generalizable. To address the above problems and increase the generalization ability of an action recognition model, we proposed Rev2Net, a new multi-task learning framework for action recognition. Further, we proposed the decoding discrepancy penalty to encourage the collaboration of the training procedures of multiple self-supervised tasks. In the experiments, we showed that our Rev2Net model trained with the DDP constraint outperforms the state-of-the-art methods on three datasets: UCF101, HMDB51, and Kinetics. Our model specifically shows a strong generalization ability in the setting of cross-dataset video action recognition.

7. ACKNOWLEDGE

This work was supported in part by the Natural Science Foundation of China (61772299, 71690231), and the MOE Strategic Research Project on Artificial Intelligence Algorithms for Big Data Analysis.

8. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014. 1, 2

[2] 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. 1, 2, 4

[3] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017. 1, 2, 4

[4] Ji Lin, Chuang Gan, and Song Han, "Temporal shift module for efficient video understanding," in *ICCV*, 2019. 1, 2, 4

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018. 2

[6] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016. 2

[7] Yunbo Wang, Mingsheng Long, Jianmin Wang, and S Yu Philip, "Spatiotemporal pyramid network for video action recognition.," in *CVPR*, 2017. 2

[8] Limin Wang, Wei Li, Wen Li, and Luc Van Gool, "Appearance-and-relation networks for video classification," in *CVPR*, 2018. 2

[9] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," in *Transactions of Pattern Analysis and Machine Intelligence*, 2013. 2

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015. 2

[11] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017. 2

[12] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018. 2

[13] Laura Sevilla-Lara, Yiyi Liao, Fatma Guney, Varun Jampani, Andreas Geiger, and Michael J Black, "On the integration of optical flow and action recognition," in *GCPR*, 2018. 2

[14] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *CVPR*, 2018. 2, 4

[15] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis, "Actionflownet: Learning motion representation for action recognition," in *WACV*, 2018. 2, 4

[16] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015. 2

[17] Anurag Ranjan and Michael J Black, "Optical flow estimation using a spatial pyramid network," in *CVPR*, 2017. 2

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 2, 5

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011. 2, 5

[20] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2014. 2, 3, 5

[21] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint Pattern Recognition Symposium*, 2007. 4

[22] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in *ICCV*, 2019. 4

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 5