Action Recognition Using High Temporal Resolution 3D Neural Network Based on Dilated Convolution

YONGYANG XU1,2, YAXING FENG1,2, ZHONG XIE1,2, MINGYU XIE3, AND WEI LUO1,2
1School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China
2National Engineering Research Center of Geographic Information System, Wuhan 430074, China
3College of Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106, USA

Corresponding author: Zhong Xie (xiezhong@cug.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505500 and Grant 2018YFB0505504, in part by the National Natural Science Foundation of China under Grant 41671400, and in part by the Key Laboratory of Geological Survey and Evaluation of Ministry of Education under Grant GLAB2020ZR05.

ABSTRACT 3D Convolution Neural Networks (CNNs), an important deep learning model, has good performance in recognizing actions in videos. When recognizing actions from videos, 3D CNNs usually down-sample in temporal dimension, leading to loss of the temporal information. To obtain more temporal information from the videos, this work proposed a new model based on the Inflated 3D ConvNet (I3D), named as I3D-T. Instead of using down-sample in temporal dimension, the proposed model applied the dilated convolution in temporal dimension to enlarge the receptive field. At the same time, a non-local feature gating block was designed in the model to learn the correlations between different feature maps. The experimental results showed that the proposed I3D-T has the state-of-art performance. Using RGB frames as input, the action recognition accuracies are respectively 95% and 74.8% in public dataset of UCF101 and HMDB-51.

INDEX TERMS Action recognition, dilated convolution, receptive field, non-local feature gating.
dilated convolution [22] to enlarge receptive field and keep high temporal resolution while training 3D CNNs.

To improve neural networks’ performance of recognizing actions, methods from two aspects are usually considered. From the aspect of data, models can first be trained by some large datasets such as Kinetics [18] and Sports-1M [2] and then fine-tuning them with small datasets like UCF101 or HMDB51. However, training the models from scratch using large datasets usually take long period [23], especially for 3D CNNs [19]. From the aspect of method, the transfer learning can be extracted from some trained state-of-art models, such as Inflated 3D convolutional networks (I3D) [18] or residual 3D convolutional networks (R3D) [24], which can save a lot of training time. Moreover, some functional blocks like non-local block [25] and Temporal Transition Layer (TTL) [26] were designed to improve the performance of action recognition. However, these methods failed to extract the correlation among different channels of a 3D CNN with respect to temporal and spatial features [27]. Nowadays, novel blocks including ‘spatio-temporal channel correlation’ (STC) [27], spatio-temporal feature gating [27] and context feature gating [28] have been designed to learn correlations among different channels, which are beneficial for the performance of action recognition. Based on the analysis above, this work designed a neural network based on the 3D Inception [29] and designed a novel non-local feature gating (NFG) block that can capture channels’ correlation information throughout network layers efficiently. Compared with the non-local block generating non-local descriptor by extracting correlation of pixels in spatio-temporal dimension or spatial dimension, the NFG can directly capture feature maps’ correlation in spatio-temporal dimension.

High temporal resolution plays an important role in action recognition using 3D CNNs. To keep high temporal resolution during the recognition, this work proposed a new model based on the 3D Inception, named I3D-T. A novel non-local feature gating (NFG) block was designed to capture the correlation of inter channels. Moreover, the block can be combined with any existing deep convolutional neural network architectures. The contributions of this work are summarized as follows:

1) A high temporal resolution 3D neural network based on dilated convolution is used for action recognition and this strategy can improve the recognition accuracy of 3D model.
2) A novel non-local feature gating (NFG) block is proposed which can learn the relationship between channels in the spatial-temporal dimension and improve the stability of the model.
3) Experimental results showed that this work obtain state-of-art performance on both UCF101 and HMDB51, either pre-trained by ImageNet or Kinetics.

II. RELATED WORK
Temporal information is important for action recognition via videos. When considering how to learn the feature of temporal dimension, one natural idea would be expanding 2D network to 3D network. Thus, the spatio-temporal CNN [19] was proposed to learn spatial and temporal dimension features, so as to conduct action recognition. These spatio-temporal neural networks trained their model on a large dataset Sports-1M, and then used SVM to classify the videos. By finetuning their model on small datasets like UCF101 and HMDB51, they got state-of-art results. DenseNet [30] also extended from 2D to 3D for feature extraction and TTL units were introduced to the network to learn features of different time scales [26] for action recognition. By borrowing the state-of-art results obtained by ResNet [31] in image classification, a new 3D residual network was designed. Taking full advantage of the ResNet’s powerful ability in extracting features, the model achieved good performance in action recognizing [32].

3D CNNs also have disadvantages for action recognition. The training requires large amount of data, and it is time-consuming. However, Benefited from the good results of the neural network in image classification, some 3D CNNs can be initialized by flattening the corresponding 2D model [18]. The operation reduces the time of training the 3D model greatly.

The network structure of this work is somewhat like that of, but different from the down sampling operation in time dimension. Our approach improves the down sampling operation in time dimension to keep a high temporal resolution throughout the process. To obtain the same perception field as the original network in time dimension, dilated convolution will be used in the network.

Recognizing actions of videos need to extract not only the spatial information, but also the temporal features. To extract spatial information effectively, a simple model is used to incorporate attention in action recognition by combining second-order pooling and top-down attention, which was treated as spatial attention mechanism [33]. At the same time, top-down attention mechanism was introduced for weighting spatial regions and improve the recognition accuracy [34]. A new non-local block for CNNs was designed to capture long range dependencies [25]. It can be applied in spatial and temporal information at the same time to improve the recognition accuracy by inserting the non-local block to 2D CNNs or 3D CNNs. These methods improve the accuracy of the model to some extent, but they neglect the interaction among channels. Recently, Kmiec et al. [28] introduced the context gating method, which applies gating to the features of the output layer. Based on the context gating mechanism, Diba et al. [27] used pooling operations as the gating mechanism to improve the accuracy. Diba et al. [26] introduced a block by combining pooling operations as the gating mechanism to improve the accuracy. Diba et al. [26] introduced a block by combining pooling operations as the gating mechanism to improve the accuracy.
With the help of non-local algorithm capturing long-time series data, this work proposed non-local feature gating as the channel-wise (Fig. 1), which extends non-local block to channels and uses non-local mechanism to calculate channels correlations. In the Fig. 1, the sequence on top shows the input of frames, the one in the middle shows the filters from the I3D-T network, and the bottom row shows the filters after channel-wise by using non-local feature gating. Note that there is an obvious change before and after the channel correlation between different filters.

**III. PROPOSED METHOD**

**A. NETWORK ARCHITECTURE**

This work used the 3D BN-Inception [29] as the base network. Given a video $X$ with $t$ frames, after processing by the model, the value of temporal dimension for output layer before the last global pooling layer is $\sigma \times t$. In the 3D BN-Inception, $\sigma$ was set as $1/8$. For example, if the input video has 32 frames, then the output layer is 4 in temporal dimension. To obtain high temporal resolution, we do not perform temporal down-sampling in the proposed model I3D-T. The proposed model has five stages (Fig. 2), where Stage1, Stage2 and Stage3 use the structure as the 3D BN-Inception. In Stage4 and Stage5, the 3D Inception blocks (Fig. 3a) are replaced by the designed 3D Inception-T blocks (Fig. 3b). By doing so, the receptive field would be smaller than before, thus this paper introduced dilated convolution to enlarge the receptive field. In detail, the stride of max-pooling after Stage 3 and Stage 4 in temporal dimension were set as 1, while the dilated convolution rate $\beta$ in 3D Inception-T block in Stage 4 and Stage 5 were set as 2 and 4 correspondingly.
B. NON-LOCAL FEATURE GATING BLOCK

To model interdependencies among network activations [28] and allow distant pixels to contribute to the filter at a location [40], this work proposed a novel block named as non-local feature gating (NFG, Fig. 4). The block is a channel-wise design for video classification and can also capture long-range dependencies in the spatial-temporal dimension. At the same time, the block has sufficient flexibility to be inserted into any block in the 3D neural network.

For a tensor $X \in \mathbb{R}^{T \times W \times H \times C}$ out of the 3D convolutional and pooling operations, where $T$, $W$, $H$, and $C$ respectively represent the temporal depth, width, height, and number of channels of the feature maps. To capture long-range dependencies in the spatial-temporal dimension, three steps were made in this work. Firstly, the tensor $X$ is reshaped into $X_1 \in \mathbb{R}^{N \times C}$, where $N = T \times W \times H$. Then, the tensor $X_1$ and its transposition $X_1^T$ are multiplied to obtain the channel correlation map $A \in \mathbb{R}^{C \times C}$. At last, a softmax layer is applied to $A$ to obtain the channel correlation map $B \in \mathbb{R}^{C \times C}$ as follows:

$$B_{ij} = \frac{\exp (A_i * A_j)}{\sum_{i=0}^{C} \exp (A_i * A_j)} \quad (1)$$

where $B_{ij}$ represents the correlation between the $i$-th channel and the $j$-th channel. To make sure the output of the proposed block has the same shape and size with the input, matrix multiplication is used between $B$ and $X_1$ and get the $Y \in \mathbb{R}^{T \times W \times H \times C}$. The context feature gating mechanism can be seen as a new channel-wise method for video classification [28], it can help create dependencies between features. Therefore, it is introduced at the end of this block. The module transforms the feature representation into a new representation as follows:

$$Z = \sigma(Y) \odot X \quad (2)$$

where $\odot$ represents elementwise multiplication and $\sigma$ is the sigmoid activation. This mechanism allows the model to upweight certain dimensions of $X$ if the context model $\sigma(Y)$ predicts that they are important, and downweigh irrelevant dimensions. The multiplication between $Y$ and $X$ make sure the output $Z \in \mathbb{R}^{T \times W \times H \times C}$ has the same size as the input $X$.

The non-local feature gating block can be inserted to any stage of the 3D neural network. In this work, it was inserted to Stage 4.

IV. EXPERIMENTS

A. DATASETS AND MODEL PRE-TRAINING

In the experiment, two public datasets UCF101 and HMDB51 were used to validate the proposed method. The UCF101 dataset owns 13,320 videos with 101 action classes, while HMDB51 dataset owns 6,766 videos with 51 action classes. Standard 3 training/testing splits are used for both datasets in the experiment. To make sure the results are objective and accurate, the average value of the three tests’ results were shown in the experiments.

In order to train the model with less time and improve its performance, the model was pre-trained by two public datasets ImageNet [41] and Kinetics [18] and fine-tuned on UCF101 and HMDB51. Since ImageNet is a static image dataset, it is mostly used to initialize 2D CNNs. To solve the problem and make sure the pre-trained model can be used for 3D CNNs, the 2D CNNs were inflated into corresponding 3D CNNs and 2D convolutional filters of $n \times n$ were converted.
into 3D convolutional filters of $n \times n \times n$. Since Kinetics is a large dataset, using it to train a neural network will be time consuming. In this work, a pre-trained Inception 3D model was used to initialize the corresponding layers in the designed model.

**B. IMPLEMENTATION DETAILS**

The proposed deep learning of the I3D-T is implemented using Tensorflow in the Linux platform with four TITAN X GPUs (12 GB RAM). When training the model, an excellent optimization is required to minimize the energy function and update the parameters of the model algorithm. SGD (Stochastic Gradient Descent), one of the most commonly used algorithms [42], was treated as the network optimizer to minimize the losses and update the parameters including weights, biases, and so on. While training the network, to obtain a better performance and speed up the processing, the learning rate of the I3D-T model was set to 0.001 and the dropout was set to 0.5. SGD is easily trapped in the case of ravines, which means that one direction of the surface is steeper than the others, leading to the SGD oscillating and delaying the approach to the minimum value. To solve this problem, momentum was introduced, and was set to 0.9 in the experiments. Compared with 64 frames used in [18], 32 frames were used as the input size to train the model as recommend in [26]. Because applying 32 frames to the model will take less time which is benefiting for practical application. To obtain 32 frames from each video, the video must first be evenly divided into 32 clips and then pick out one frame from each clip. In this way, a video can be sampled by a group of frames. As to some short videos, some frames will be picked out repeatedly to meet the input requirement. Batch size was set to 8 and all the models are trained with 40 epochs for fine-tuning and we choose the highest test accuracy to report. When training the model, to make sure that the input frames contain most of the views in the video no matter what size it is, the videos were resized to $256 \times 256$ pixels. Then, randomly cropping $224 \times 224$ pixels. To enhance the dataset, random left-right flipping is also used for the frames in training. When testing, each video was first decomposed into of 32 frames as described in training data preparation and then all the frames were clipped $224 \times 224$ pixels center-crop.

**C. ABLATION STUDIES**

In this section, experiments are done to investigate how our approach can learn the different aspects of the video level representation. These experiments are all conducted on the UCF101 and HMDB51 dataset. First experiment examined how the $\sigma$ mentioned in section 3.1 affect the accuracy of the model. The second experiment measured the accuracy of the model in the stages where non-local feature gating block was inserted. The third experiment examined how different number of frames in training and testing can influence the model. To show the efficiency of the model, theoretical FLOPS based on a single input video sequence of 32 frames and spatial size $224 \times 224$ was calculated.

**TABLE 1. The accuracy (%) of different $\sigma$ on the split1 of UCF101 and HMDB51.**

|        | $\sigma=1/8$ | $\sigma=1/4$ | $\sigma=1/2$ |
|--------|--------------|--------------|--------------|
| HMDB-51  | 45.0         | 45.7         | 46.6         |
| UCF-101  | 79.4         | 79.8         | 80.3         |
| FLOPS(G) | 55.61        | 56.55        | 67.66        |

**FIGURE 5. The results of different numbers of NFG blocks in stages 4.**

**TABLE 2. The accuracy (%) of inception 3D model on the split1 of HMDB51 by using two non-local feature gating blocks insert to different stage.**

|        | Stage5 | Stage4 | Stage3 |
|--------|--------|--------|--------|
| Accuracy | 45.5   | 45.7   | 45.3   |

**TABLE 3. The accuracy (%) of non-local feature gating block on the split1 of UCF-101 and HMDB-51.**

|        | 0 NFG | 5 NFG |
|--------|-------|-------|
| HMDB-51 | 45.0  | 46.7  |
| UCF-101 | 79.4  | 80.4  |

Temporal resolution plays an important role in action recognition. In the experiments, this work calculated the accuracy of different $\sigma$ on UCF101 and HMDB51. The models were initialized by corresponding 2D CNNs that are pre-trained by the ImageNet and 32 frames were treated as the input for training models. As shown in Table 1, the accuracy increases as the value of $\sigma$ increases, that is to say, the higher temporal resolution of the model has, the better performance of the model presents. In detail, compared with I3D ($\sigma = 1/8$), the accuracy has been improved 1.6% on HMDB51 and 0.9% on UCF101 when $\sigma = 1/2$. As the temporal resolution increases, GFLOPS also increases. The GFLOPS increased about 12 from $\sigma = 1/8$ to $\sigma = 1/2$ since there will be more parameters learnt if the model has a higher temporal resolution.

To examine the performance of the proposed non-local feature gating block inserted into different stages. This work compared the performance (Table 2) of two blocks that were added in Stage 3, Stage 4, and Stage 5 of Inception 3D ($\sigma = 1/8$).
From Table 2 we can see that the non-local feature gating block performed the best in Stage 4 with accuracy of 45.7%. Therefore, this work added non-local feature gating blocks in Stage 4 in the following experiments.

The number of NFG blocks in the neural network plays an important role in the action recognition. Using fewer NFG blocks will be less beneficial for the result, but too many blocks will use up the memory. To obtain a better number of NFG blocks, this work compared the results based on different numbers of NFG blocks in Stages 4 (Fig. 5), that is, we inserted a NFG block after every 3D Inc-T block. All the comparisons were conducted at $\sigma = 1/8$ and the model was trained by 32-frame snippets. It is clear that the five NFG blocks have improved in accuracy of 1.7% increase on HMDB5 and 1% increase on UCF101.

The more the frames are used to train and valid the model, the better the results are. At the same time, the accuracy will increase as the $\sigma$ increase. To validate the hypothesis, a serial of experiments were conducted (Table 4). To make the comparison fair, the setting was the same as I3D [18], where the model was trained by 64-frame snippets and tested by using 250 frames. All the models performed better than ones trained by 32-frame snippets (Table 3). Moreover, the accuracies have increased on both HMDB-51 and UCF-101 when $\sigma$ increased from $1/8$ to $1/2$, which corresponds to that high temporal resolution is helpful for the performance of 3D CNNs during action recognizing. However, when $\sigma = 1/8$, the accuracy in this work is less than the that in [18]. The reason may be that this work used the mini-batch, however, larger mini-batch size was used in [18] using 64 GPUs. But when $\sigma = 1/2$, this work obtained better results than these in [18] which pre-trained by ImageNet and the accuracies increased 0.4% on HMDB51 and 0.2% on UCF101.

### D. COMPARISON WITH THE STATE-OF-ART

Having analysed the effect of temporal resolution and non-local feature gating block on the performance of action recognition, final experiments on all three testing splits of UCF101 and HMDB51 are implemented with our proposed methods and other state-of-the-art methods’ pre-trained by Kinetics (Table 5). These methods recognized action using spatial and spatio-temporal information, including Two-Stream Convolutional Networks (TSCN) [43], Improved Trajectories (iDT) [16], Factorized Spatio-Temporal Convolution Network (FstCN) [44], Long-term Temporal Convolutions (LTC) [45], ActionVLAD [33], Spatiotemporal Residual Networks (ST-ResNet) [23], 3D Convolutional Networks (C3D) [19], Asymmetric 3D Convolutional Neural Networks (Asymmetric 3D-CNN) [46], Temporal 3D ConvNets (T3D) [26] and Pseudo-3D Residual Networks (P3D) [47], Spatio-Temporal Channel Correlation Networks (STC) [27], Semantic Image Networks (SemIN) [48]. To make the comparison more comprehensible and fairer, $\sigma$ was set as $1/2$ and five NFG blocks were added into the designed I3D-T. Moreover, 32 frames and 64 frames were used as the input to train the model. From the table we can see that the I3D-T-NFG (64 frames) model has an outstanding performance, and it achieved 95.0% and 74.8% accuracy on HMDB51 and UCF101. In fact, the accuracies STC and SemIN were higher than I3D-T-NFG on UCF101, however, SemIN used static RGB images, multiple semantic images, warped optical flows and semantic optical flows as inputs, our network only uses RGB images as inputs; STC only used RGB images as input, but its model takes higher memory during training. Moreover I3D-T-NFG achieved a better result than STC and SemIN on HMDB51.

Furthermore, some state-of-art networks, like the Two-Stream I3D [18], fused two streams including RGB images and optical flow, which have obtained good performance. However, making the optical flow using videos is time-consuming and it is more difficult to train the neural networks with optical flow, which is not conducive to the practical application. Therefore, this work only compared the performance with RGB images. In reality, I3D-T also achieved a better performance using the model by Kinetics pre-training, even though dense optical-flow maps were not used to train the model.

### V. CONCLUSION

This work has proved that high temporal resolution plays an important role in recognizing action using 3D CNNs.

| $\sigma=1/8$ | $\sigma=1/4$ | $\sigma=1/2$ | $\sigma=1/2+5$ NFG |
|-------------|-------------|-------------|-------------------|
| HMDB-51     |             |             |                   |
| 48.2        | 49.3        | 50.2        | 50.8              |
| UCF-101     | 84.0        | 84.5        | 84.7              | 85.0              |
Moreover, a new block named non-local feature gating block was designed for the correlations between channels of a 3D CNN. Experiments showed that the designed I3D-T with non-local feature gating block achieved the state-of-the-art performance on UCF101 and HMDB51, either pretrained by ImageNet or Kinetics. A set of problems were studied, including the effects on different temporal resolution, the number and stage of non-local feature gating block, the effect of training frame number and testing frame number per video. Furthermore, using the non-local feature gating block as a simple addition provides solid improvement over baselines. Moreover, it is flexible enough and it can be added in any stage of the network by end-to-end training. A unified spatial-temporal video-level representation could also be obtained by the fusion optical flow. In the future, we will investigate how to train the I3D-T with two streams to improve the performance. Moreover, the relationship between space and time is a mysterious, the model proposed in this paper is a 3D model, which is time consuming, however, the accuracy is better compared to other models. We will optimize the network structure in our next step.

REFERENCES
[1] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, “T-C3D: Temporal convolutional 3D network for real-time action recognition,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–8.
[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732.
[3] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 568–576.
[4] Z. He, M. Deng, J. Cai, Z. Xie, Q. Guan, and C. Yang, “Mining spatiotemporal association patterns from complex geographic phenomena,” Int. J. Geog. Inf. Sci., vol. 34, no. 6, pp. 1162–1187, Jun. 2020.
[5] X. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961–2969.
[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556
[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–11.
[8] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, Apr. 2017.
[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2881–2890.
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.
[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.
[12] Y.-J. Cha, W. Choi, and O. Büyükköztürk, “Deep learning-based crack damage detection using convolutional neural networks,” Comput.-Aided Civil Infrastruct. Eng., vol. 32, no. 5, pp. 361–378, May 2017.
[13] Y. Xu, L. Wu, Z. Xie, and Z. Chen, “Building extraction in very high resolution remote sensing imagery using deep learning and guided filters,” Remote Sens., vol. 10, no. 1, p. 144, Jan. 2018.
[14] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, “Road extraction from high-resolution remote sensing imagery using deep learning,” Remote Sens., vol. 10, no. 9, p. 1461, Sep. 2018.
[15] Y. Xu, Z. Chen, Z. Xie, and L. Wu, “Quality assessment of building footprint data using a deep autoencoder network,” Int. J. Geograp. Inf. Sci., vol. 31, no. 10, pp. 1929–1951, Oct. 2017.
[16] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3551–3558.
[17] L. Wang, Y. Xiong, Z. Yang, Q. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 20–36.
[18] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new perspective including temporal segment networks,” in Proc. 32nd AAAI Conf. Artif. Intell., Jan. 2018, pp. 305–321.
[19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167
[20] S. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” 2015, arXiv:1511.07122. [Online]. Available: http://arxiv.org/abs/1511.07122
[21] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 4768–4777.
[22] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167
[23] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool “Temporal 3D ConvNets: New architecture and transfer learning for video classification,” 2017, arXiv:1711.08200. [Online]. Available: https://arxiv.org/abs/1711.08200
[24] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new perspective including temporal segment networks,” in Proc. 32nd AAAI Conf. Artif. Intell., Jan. 2018, pp. 305–321.
[25] H. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “End-to-end training on dense trajectories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 4700–4708.
[26] J. Lin, M. M. Karmi, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, “Spatio-temporal channel correlation networks for action classification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 284–299.
[27] S. Kmiec, J. Bae, and R. An, “Learnable pooling methods for video classification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 1–11.
[28] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167
[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[31] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[32] S. Riedholz and D. Ramanan, “Attentional pooling for action recognition,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 34–44.
[33] S. Sudhakaran and O. Laza, “Top-down attention recurrent VLAD encoding for action recognition in videos,” in Proc. Int. Conf. Italian Assoc. Artif. Intell. Cham, Switzerland: Springer, 2018, pp. 375–386.
[34] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 305–321.
[35] S. Wang, L. Wang, W. Li, W. Li, and L. Van Gool, “Appearance-and-relation networks for video classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1430–1439.
[39] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, “Multi-fiber networks for video recognition,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 352–367.

[40] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 60–65.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[43] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1933–1941.

[44] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4597–4605.

[45] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[46] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, “Asymmetric 3D convolutional neural networks for action recognition,” Pattern Recognit., vol. 85, pp. 1–12, Jan. 2019.

[47] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5533–5541.

[48] S. A. Khowaja and S.-L. Lee, “Semantic image networks for human action recognition,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 393–419, Feb. 2020.

YONGYANG XU received the B.S. degree in computer science and technology and the Ph.D. degree in geographic information engineering from the China University of Geosciences, Wuhan, in 2014 and 2019, respectively. He is currently an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences. His main interests include deep learning, and vector data rendering and processing.

YAXING FENG received the B.S. degree from Southwest University in 2017. He is currently pursuing the master’s degree with the School of Geography and Information Engineering, China University of Geosciences, Wuhan.

His research interests include action recognition and image processing.

ZHONG XIE received the B.S., M.S., and Ph.D. degrees in cartography and geographic information engineering from the China University of Geosciences, Wuhan, China, in 1990, 1998, and 2002, respectively.

He is currently a Professor with the Faculty of Information Engineering, China University of Geosciences. His research interests include 3-D rebuilding and spatial analysis, and image processing.

MINGYU XIE is currently pursuing the bachelor’s degree in mathematical sciences with the University of California at Santa Barbara, Santa Barbara, USA.

His research interests include algorithms for bot in video games and the mechanism of game controllers.

WEI LUO received the B.S. degree in surveying and mapping engineering from Guizhou University in 2019. He is currently pursuing the master’s degree with the School of Geography and Information Engineering, China University of Geosciences, Wuhan.

His research interests include deep learning, and vector data rendering and processing.