The above experiments were all trained from scratch on UCF101.

II. NETWORK

Video contains a lot of time and space information. We need frame-level feature representation and temporal modeling to fuse these information. SlowFast can be used to extract dynamic information with 3D ConvNets. In addition, when treating each frame of a video as a single image, TSM can be used as a supplement to spatial information.

The SlowFast model includes a slow pathway, working at a low frame rate, to capture spatial semantics. And a Fast pathway, working at a high frame rate, to capture motion at fine temporal resolution. The Fast pathway can be made very lightweight by reducing its channel capacity, yet it can learn useful temporal information for video recognition.

In SlowFast, we tried different backbones including ResNet-50 and ResNet-101, optionally augmented with non-local (NL) blocks. We combined the network structure of 2+1D network, and one k × k × k kernel can be separated into three parts, k × 1 × 1, 1 × k × k and k * k * k. Meanwhile, we tried the SE module and the TAM module. And attempt to add Spatio-temporal attention to 3D convolution.

In contrast to existing 3D convolution-based methods that use ReLU as the activation function, replacing ReLU with ELU improved the accuracy. And we replace all the downsampling convolution layers with max-pooling layers(POOL). For the input of SlowFast, we combined residual frames and RGB frames to form a two-path solution, which can get the best effect.

Because residual frames contain little information of object appearance, we further use a 2D convolution TSM to extract appearance features.

Temporal Shift Module (TSM) are variants of 2D ConvNets, which can be seen in Figure 3. TSM shifts part of the channels along with the temporal dimension, facilitating information exchange among neighboring frames. It can be inserted into 2D CNNs to achieve temporal modeling at zero computation and zero parameters. Shift operation is actually used to integrate the frame features coming from the front and rear frames to increase the time perception field. We choose the embedding method of residual TSM to ensure the extraction of spatiotemporal features without increasing the computation.
III. EXPERIMENT

A. Data augmentation

In this competition, we take 64 consecutive frames as the input of SlowFast. Since training from scratch, we further enhance the data diversity. We do multi-scale enhancements by randomly scale the data to 0.8x-1.25x of 112, then randomly crop each frame to 112*112. In addition to lighting, horizontal flipping, corner cropping, contrasting in image enhancement, we play the video in reverse order as new training samples. And randomly extracting frames in one or double step. The method of data enhancement described above applies to the residual-frame-input path. The residual-frame-input path is a good supplement for existing RGB-frame-input models. The input of TSM is 16 consecutive frames, and the same method as above is used for data enhancement.

| Method | Acc(%) |
|--------|--------|
| SlowFast 3DResNet 50 | 70.0 |
| SlowFast +non local | 70.6 |
| SlowFast +2plus1D | 71.2 |
| SlowFast +ELU+POOL | 73.1 |
| SlowFast +ELU+POOL+non local | 72.8 |
| SlowFast +ELU+POOL+TAM | 71.2 |
| SlowFast +ELU+POOL+(3*3+3*1+1*3) | 73.5 |
| SlowFast +attention | 70.5 |

B. Training

In the experiment, we tested MFnet, R2plus1D, ResNetR3D, SlowFast network, and TSM networks, and finally chose to use SlowFast network and TSM for model fusion. The structure of SlowFast and 2+1D can be seen in Figure 1 and Figure 2 respectively. In experiments, MFnet, R2plus1D, and ResNetR3D can only get low accuracy compared to SlowFast. It is very difficult to train TSM from scratch because of the extra shift module which can be shown in Figure 3. But TSM can make full use of single-frame information and the fusion between channels can perfectly cooperate with SlowFast multi-frame fusion information.

For it is difficult to train TSM. We first used single-frame as input to train TSM, when the accuracy of the single-frame training model reaches 53%, we finetune TSM by using 16-frames as input.

We adopted the same train strategy as TSM train SlowFast. After the accuracy of SlowFast reaches 50%, we saved it as a pre-trained model. Based on this pre-trained model we do more research, such as add 3D convolution modules, non local and 2plus1D modules. And adopted data augmentation to fin-tune it. Also fin-tune it by using normal and reverse contacted video. As shown in Table 1. Finally, we choose SlowFast with ELU, POOL, and 2+1D network with k=3 as our modified SlowFast.

In addition to adding modules to the model, we've tried to use 3DResNet50 or 3DResNet101 as the backbone of the model respectively. The result as shown in Table 2, and the performance of 3DResNet101 is not as good as that of 3DResNet50, so we choose 3DResNet50 as the backbone.

We use the standard SGD with momentum set to 0.9 in all cases. The batch size is set to 128. In the experiment, the learning rate was set to 0.001, and weight decay is set to 0.0005 at the beginning. The learning rate is reduced to 1/10 for every twenty epochs.

Finally, by using Residual frames or RGB frames respectively as input, we trained four models with the best SlowFast structure through the data augmentation mentioned above.

| Method | Input type | Acc(%) |
|--------|------------|--------|
| SlowFast+ELU+POOL+(3*3+3*1+1*3) 3D ResNet 50 | RGB | 73.5 |
| SlowFast+ELU+POOL+(3*3+3*1+1*3) 3D ResNet 101 | RGB | 72.1 |
| SlowFast+ELU+POOL+(3*3+3*1+1*3) 3D ResNet 50 | Diff | 72.8 |
| SlowFast+ELU+POOL+(3*3+3*1+1*3) 3D ResNet 101 | Diff | 71.9 |
| TSM(with shift) 2D ResNet 50 | RGB | 72.1 |
| TSM(with shift) 2D ResNet 101 | RGB | 70.8 |
| TSM(no shift) 2D ResNet 50 | RGB | 71.9 |
| TSM(no shift) 2D ResNet 101 | RGB | 71.2 |
C. Test

After getting trained model based on above method. In test stage, we adopted test-time augmentation to get better result. We conducted different test fusion methods based on the above models. In all tests, we randomly select 64 consecutive frames from the video as input, and take ten consecutive times to calculate the average as the final score of the entire video. Note: we use the data before Softmax for ensemble. As shown in Table 3.

Table 3: The results when using different test augmentation.

| Network  | Input type | Test augmentation | Acc(%) | ID |
|----------|------------|-------------------|--------|----|
| SlowFast | RGB Interval 1 | Center crop | 75.5   | 1  |
| SlowFast | Diff Interval 1 | Center crop | 74.6   | 2  |
| SlowFast | RGB Interval 1 | Horizontal Flip | 75.6   | 3  |
| SlowFast | Diff Interval 1 | Horizontal Flip | 74.3   | 4  |
| SlowFast | RGB Interval 1 | Random crop | 73.3   | 5  |
| SlowFast | RGB Interval 1 | Normal and reverse contacted video | 75.1   | 6  |
| SlowFast | RGB Interval 1 | Reverse order | 74.6   | 7  |
| SlowFast | Diff Interval 1 | Normal and reverse contacted video | 72.1   | 8  |
| SlowFast | Diff Interval 1 | Reverse order | 71.6   | 9  |
| TSM      | RGB Interval 1 | Center crop | 70.8   | 10 |
| TSM      | RGB Interval 2 | Center crop | 71.2   | 11 |
| TSM      | RGB Interval 1 | Random crop | 71.3   | 12 |
| TSM      | RGB Interval 1 | Horizontal Flip | 72.2   | 13 |
| TSM      | RGB Interval 1 | Reverse order | 70.8   | 14 |
| SlowFast | RGB Interval 2 | Center crop | 71.3   | 15 |
| SlowFast | Diff Interval 2 | Center crop | 74.5   | 16 |
| SlowFast | RGB Interval 2 | Horizontal Flip | 75.4   | 17 |
| SlowFast | Diff Interval 2 | Horizontal Flip | 74.0   | 18 |
| SlowFast | RGB Interval 2 | Random crop | 73.2   | 19 |
| SlowFast | RGB Interval 2 | Normal and reverse contacted video | 75.0   | 20 |
| SlowFast | RGB Interval 2 | Reverse order | 73.6   | 21 |
| SlowFast | Diff Interval 2 | Normal and reverse contacted video | 72.0   | 22 |

To get the final result, we ensemble results that got above in different methods. The ID n in Table 4 correspond to ID n in Table 3. Finally, we got our best score of 87.6%, as shown in Table 4.

IV. CONCLUSION

In this paper, we mainly focused on extracting motion features without optical flow. 3D ConvNets are believed to be capable of capturing motion features when combining RGB and residual frames as the network input. The overhead for this computation was negligibly small. Residual frames can be a fast but effective way for a network to capture motion features, and they are a good choice for avoiding complex computation for optical flow. Finally, We combined modified SlowFast and TSM, using both action information and space-time information for optimal results.

Table 4: The result of different ensemble method.

| Ensemble method | Acc(%) |
|-----------------|--------|
| (SlowFast)(Interval 1) RGB+Diff center crop + Horizontal Flip + Random crop + Concat normal and reverse order | 84.5 |
| (SlowFast + TSM)(Interval 1)RGB+Diff center crop + Horizontal Flip + Random crop + Concat normal and reverse order + Reverse order | 86.3 |
| (SlowFast + TSM)(Interval 2)RGB+Diff center crop + Horizontal Flip + Random crop + Concat normal and reverse order + Reverse order | 87.6 |
| (SlowFast + TSM)(Interval 2)RGB+Diff center crop + Horizontal Flip + Random crop + Concat normal and reverse order + Reverse order | 85.8 |

References

[1] Wang H, Schmid C. Action Recognition with Improved Trajectories[C]// IEEE International Conference on Computer Vision. IEEE, 2014.
[2] Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[3] Chen Y, Kalantidis Y, Li J, et al. Multi-Fiber Networks for Video Recognition[J]. 2018.
[4] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond Short Snippets: Deep Networks for Video Classification[J]. 2015.
[5] Ravanbakhsh M, Moussavi H, Rastegari M, et al. Action Recognition with Image Based CNN Features[J]. 2015.
[6] Tran D, Bourdev L, Fergus R, et al. Deep End2End Voxel2Voxel Prediction[C]// Computer Vision & Pattern Recognition Workshops. IEEE, 2016.
[7] Baker S, Roth S, Scharstein D, et al. A Database and Evaluation Methodology for Optical Flow[C]// IEEE International Conference on Computer Vision. IEEE, 2007.
[8] Condell J, Scotney B, Morrow P. Adaptive Grid Refinement Procedures for Efficient Optical Flow Computation[J]. International Journal of Computer Vision, 2005, 61(1):31-54.

[9] Kajo I, Malik A S, Kamel N. An evaluation of optical flow algorithms for crowd analytics in surveillance system[C]/ International Conference on Intelligent & Advanced Systems. IEEE, 2017.

[10] Wang L, Xiong Y, Wang Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[C]/ Ecv. Springer International Publishing, 2016.

[11] Liu J, Shahroudy A, Xu D, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition[J]. 2016.

[12] Zhou B, Andonian A, Torralba A. Temporal Relational Reasoning in Videos[J]. 2017.

[13] Zhao L, Wang J, Li X, et al. Deep Convolutional Neural Networks with Merge-and-Run Mappings[J]. 2016.

[14] Girdhar R, Ramanan D, Gupta A, et al. ActionVLAD: Learning spatio-temporal aggregation for action classification[C]/ IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017.

[15] Chen Y, Kalantidis Y, Li J, et al. Multi-Fiber Networks for Video Recognition[J]. 2018.