Video Classification

Lin Zhu(s232291), Cong Jin(s232254), Dong Yun(s232293), Yu Zhang(s230000)

Problem description

1. Develop and compare **different video classification models** (per-frame models, late fusion, early fusion, and 3D CNNs) on the **UCF-101 subset** dataset consisting of 10 workout-related classes.
2. Investigate **information leakage** in train/validation/test splits by retraining models using an updated dataset (**ucf101_noleakage**) that ensures no subject appears in multiple splits
3. Implement **dual-stream networks** incorporating optical flow for improved action recognition.

Data description

UCF-101 subset - a **workout action dataset subset**, comprising 720 videos across 10 balanced classes. The dataset includes 10 uniformly sampled frames per video and CSV files containing video metadata.

UCF101_noleakage - a reorganized version of the original dataset that ensures no subject appears in multiple splits, **preventing information leakage**. Additionally includes **pre-computed optical flow data** between the 10 sampled frames using the RAFT model, provided in both raw format and PNG visualizations for developing dual-stream networks. **TRAIN : VAL : TEST** = 500 : 120 : 120

Video classification models

1.For the original **UCF-101 subset** dataset, we found that Per frame with aggression model works better, with the highest accuracy of **90.33%**.

2.While for the **UCF101_noleakage** dataset, all models **performed poorly**, which may indicate that data leakage has a large impact on the performance of the models.

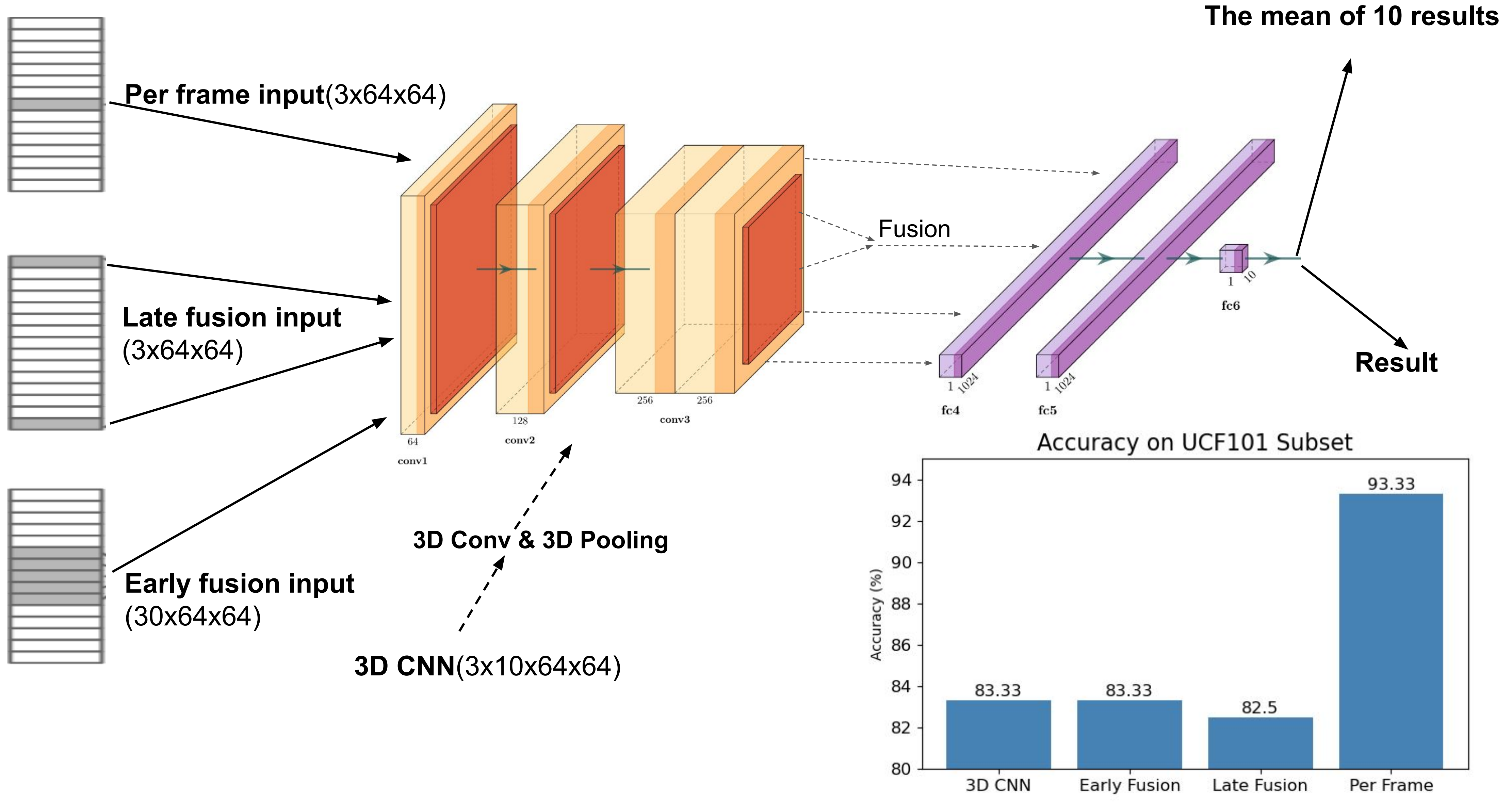

Table1:The performance of different models on two datasets

Models	3DCNN	Early fusion	Late fusion	Per frame with aggression
Accuracy(leakage) Val / Test	85.83% / 83.33%	84.17% / 83.33%	80.83% / 82.50%	**90.00% / 93.33%**
Accuracy(no leakage) Val / Test	22.50% / 25.83%	18.67% / 21.67%	23.33% / 24.17%	27.50% / 27.43%

Dual-Stream ConvNet

1. We use three different architectures in this part. The first one is the CNN mentioned in the original paper, the second is ResNet18 pretrained in ImageNet and the last one is ResNet50 pretrained in ImageNet.
2. The optimizer used was Adam with a learning rate of 0.0001. To mitigate overfitting, different values of L2 regularization were tested, and the best performance was achieved with a weight of 0.001
3. We improved the model performance by increasing the input image size from (64,64) to (128,128), applying data augmentation, and using channel concatenation. As a result, the test accuracy was enhanced to 70.33%.

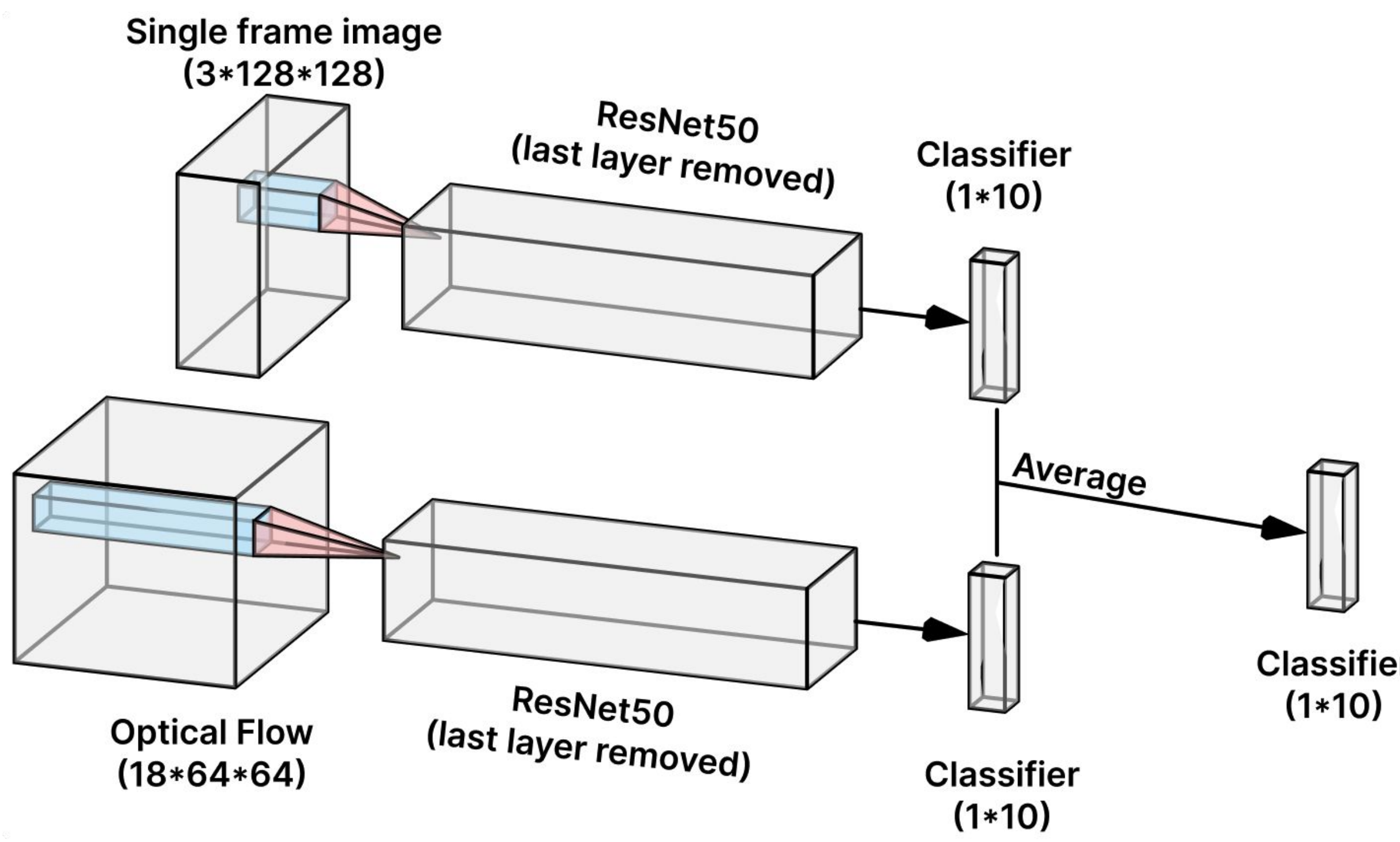

Fig2: Late fusion: (0.5*frame+0.5*flow)

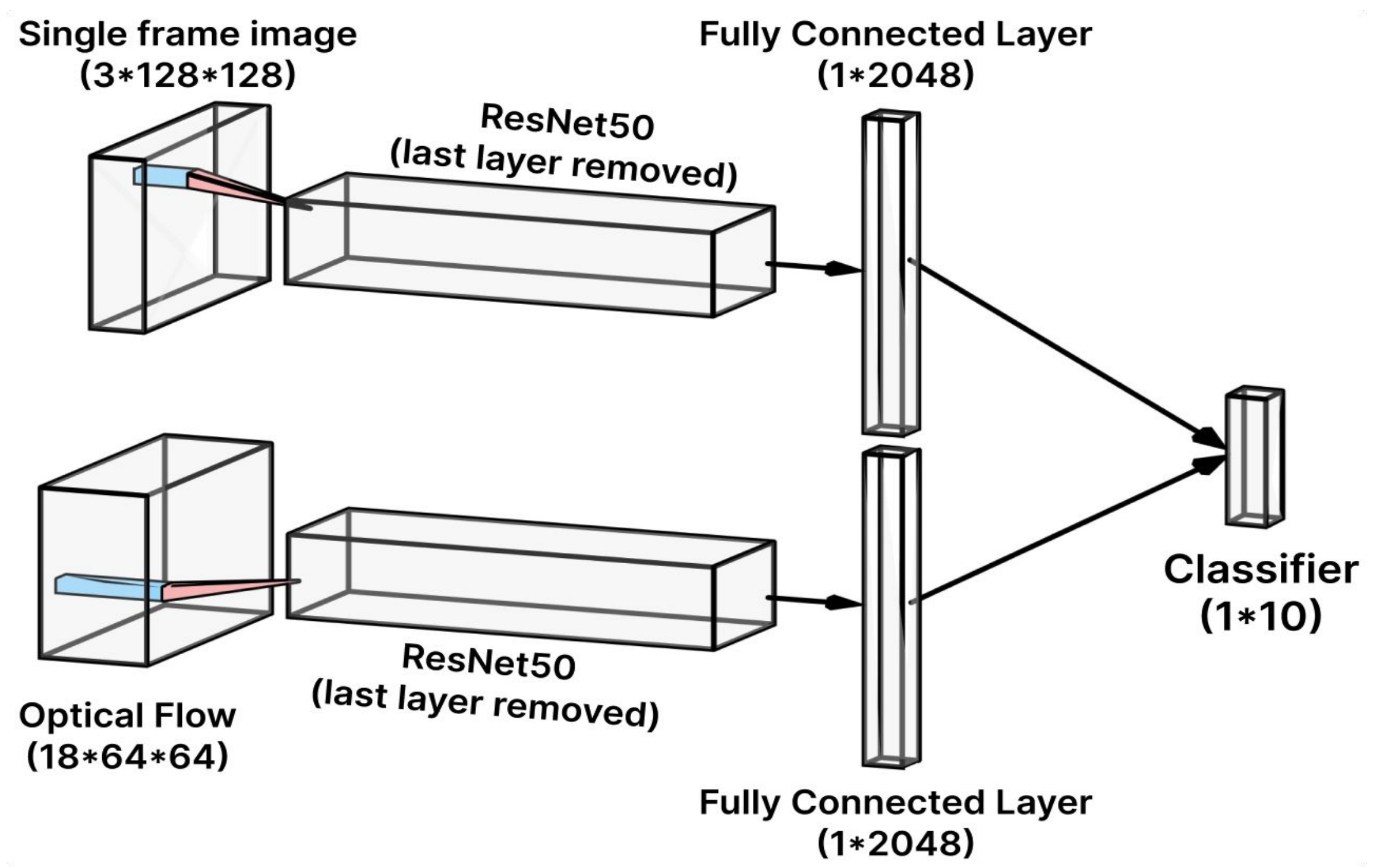

Fig3: Late fusion: (channel concatenation)

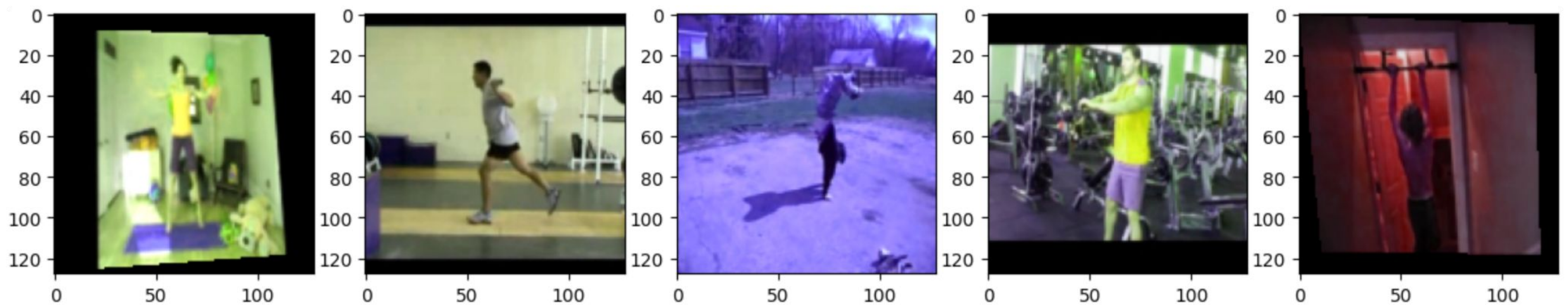

Fig4: Data augmentation

Table2: The performance of different models

architecture	CNN	ResNet18	ResNet50				
Image size	(64,64)				(128,128)		(128,128) with augmentation
weight decay (L2 regularization)	0.001	0.001	0.001	0.01	0.001	0.001	**0.001**
Best val accuracy	49.17%	63.33%	63.33%	54.17%	71.67%	71.67%	**75.0%**
Average val accuracy	43.37%	55.34%	56.1%	47.39%	63.36%	64.39%	**67.25%**
Test accuracy	42.5%	57.5%	59.6%	40.83%	68.33%	69.17%	**70.33%**
Late fusion	0.5frame+0.5flow					0.4frame+0.6flow	channel concatenation

References

[1] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[2] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in neural information processing systems* 27 (2014).

[3] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012): 221-231.