Action Recognition

Part 1: Pretrained ResNet (ImageNet) for single-frame classification

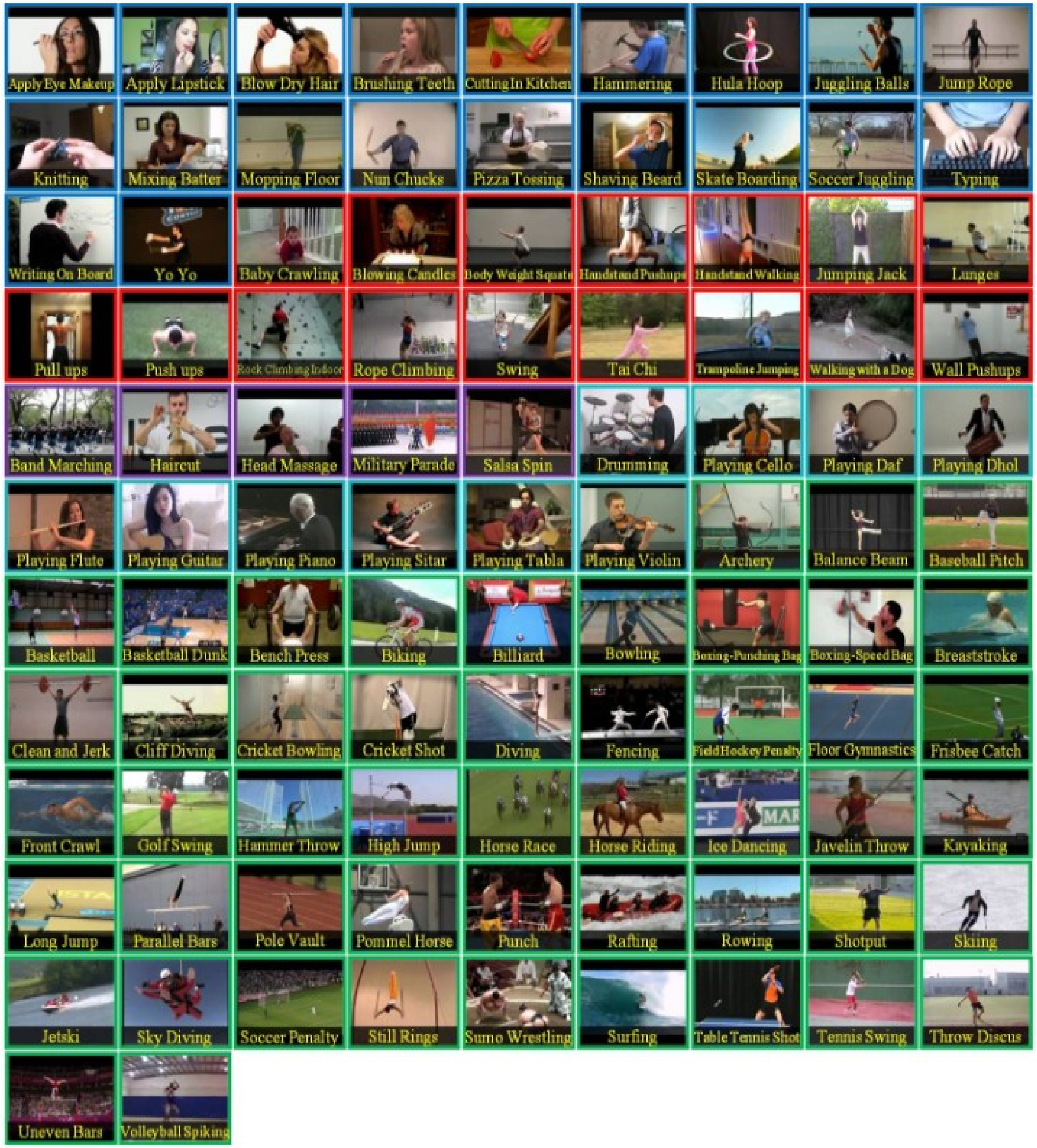
Part 2: Pretrained 3D ResNet (Kinetics) for multi-frame classifcaiton

Due Wed Nov 28th

Learning From Video Sequences

- CNNs seem to work well so far for all image related tasks – Nearly always pretrained on ImageNet
- Audio and NLP work frequently with temporal information
- Videos are sequences of images
- Many ways to combine spatial and temporal information – spatiotemporal features
- UCF101/HMDB51 → small datasets
- Kinetics/ActivityNet → 100x larger
- Can be hard to make sense of the progress over the years due to changing datasets, better CNN architectures, and the combination of techniques

UCF-101 Action Recognition Dataset



- 13,320 videos, 101 action categories
- 320x240 dimension frames
- ~2-10 seconds average clip duration (30 fps)
- Original 43.90% classification rate (2012) (has since greatly improved)

Large-scale Video Classification with Convolutional Neural Networks (2014)

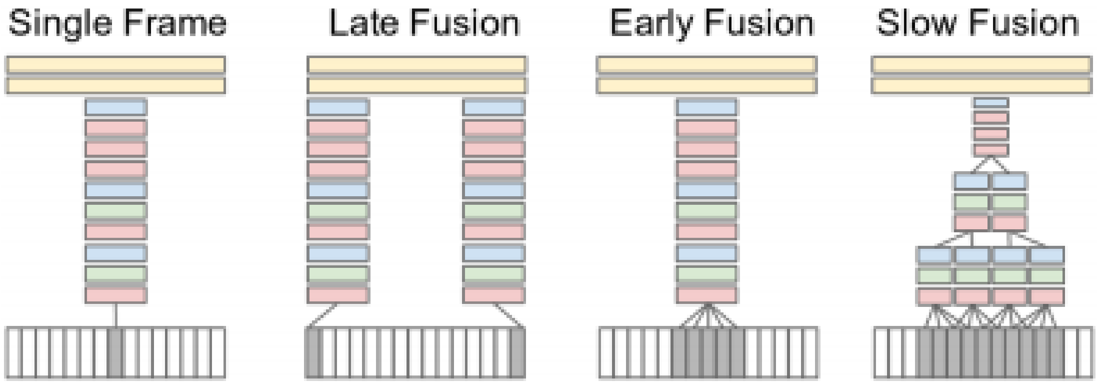
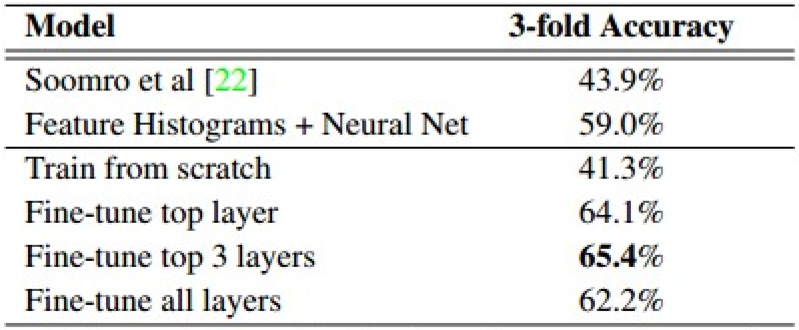


Figure 1: Explored approaches for fusing information over temporal dimension through the network. Red, green and blue boxes indicate convolutional, normalization and pooling layers respectively. In the Slow Fusion model, the depicted columns share parameters.



Model	3-fold Accuracy
Soomro et al [22]	43.9%
Feature Histograms + Neural Net	59.0%
Train from scratch	41.3%
Fine-tune top layer	64.1%
Fine-tune top 3 layers	**65.4%**
Fine-tune all layers	62.2%

Table 3: Results on UCF-101 for various Transfer Learning approaches using the Slow Fusion network.

- Trained on Sports 1M dataset
- Combine inputs for multiple time steps
- 65.4% accuracy on UCF-101

Two-Stream Convolutional Networks for Action Recognition in Videos (2014)

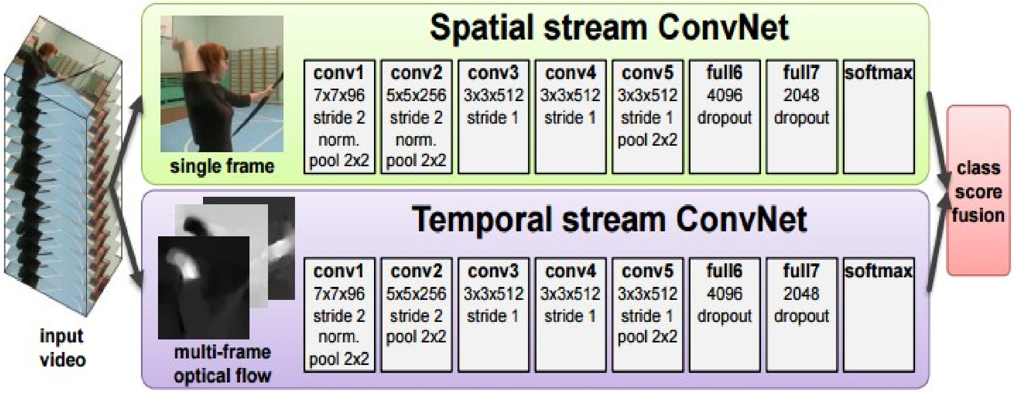
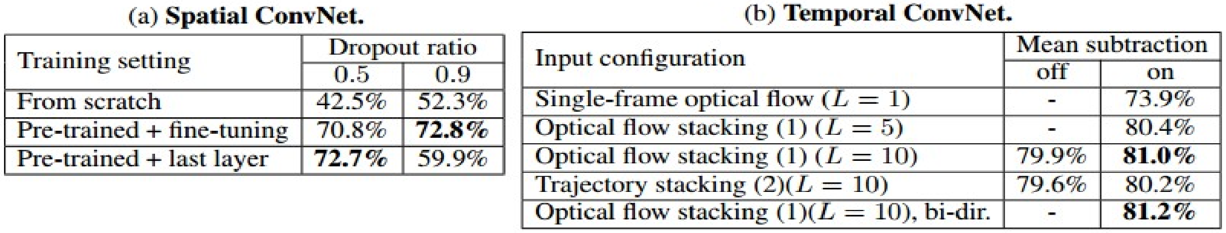


Figure 1: **Two-stream architecture for video classification.**

Table 1: **Individual ConvNets accuracy on UCF-101 (split 1).**



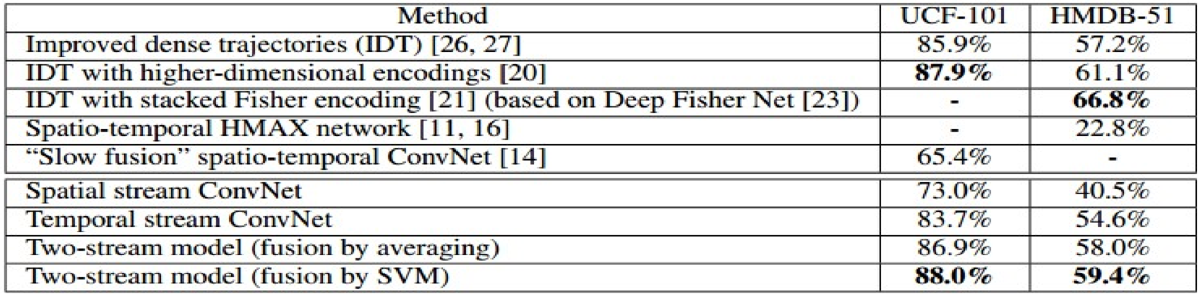
(a) **Spatial ConvNet.**

Training setting	Dropout ratio	
	0.5	0.9
From scratch	42.5%	52.3%
Pre-trained + fine-tuning	70.8%	**72.8%**
Pre-trained + last layer	**72.7%**	59.9%

(b) **Temporal ConvNet.**

Input configuration	Mean subtraction	
	off	on
Single-frame optical flow ($L = 1$)	-	73.9%
Optical flow stacking (1) ($L = 5$)	-	80.4%
Optical flow stacking (1) ($L = 10$)	79.9%	**81.0%**
Trajectory stacking (2)($L = 10$)	79.6%	80.2%
Optical flow stacking (1)($L = 10$), bi-dir.	-	**81.2%**

Table 4: **Mean accuracy (over three splits) on UCF-101 and HMDB-51.**



Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

- Trained on Sports 1M dataset
- Optical Flow features
- Improvements with CNN architectures
- 88.0% accuracy (Spatial stream 73.0%, Optical Flow stream 83.7%)

Beyond Short Snippets: Deep Networks for Video Classification (2015)

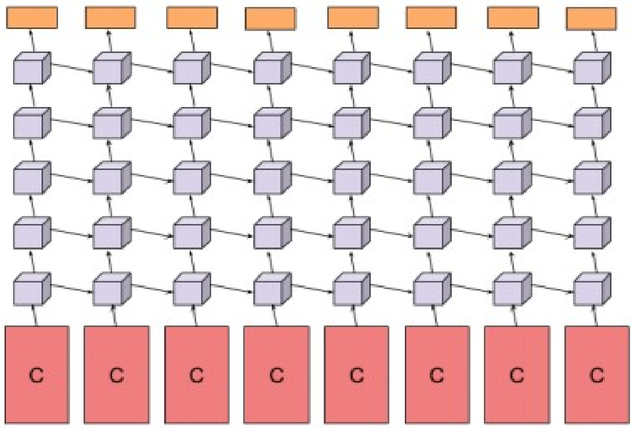
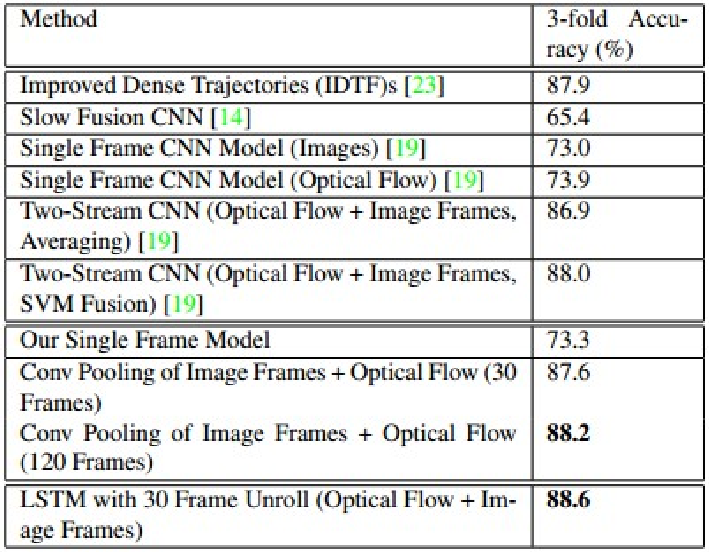


Figure 4: Deep Video LSTM takes input the output from the final CNN layer at each consecutive video frame. CNN outputs are processed forward through time and upwards through five layers of stacked LSTMs. A softmax layer predicts the class at each time step. The parameters of the convolutional networks (pink) and softmax classifier (orange) are shared across time steps.

- Trained on Sports-1M dataset
- Feed CNN output (image and optical flow) into stacked LSTMs
- 88.60% accuracy
- 73.0% single frame accuracy and 82.6% LSTM accuracy (no optical flow)



Method	3-fold Accuracy (%)
Improved Dense Trajectories (IDTF)s [23]	87.9
Slow Fusion CNN [14]	65.4
Single Frame CNN Model (Images) [19]	73.0
Single Frame CNN Model (Optical Flow) [19]	73.9
Two-Stream CNN (Optical Flow + Image Frames, Averaging) [19]	86.9
Two-Stream CNN (Optical Flow + Image Frames, SVM Fusion) [19]	88.0
Our Single Frame Model	73.3
Conv Pooling of Image Frames + Optical Flow (30 Frames)	87.6
Conv Pooling of Image Frames + Optical Flow (120 Frames)	**88.2**
LSTM with 30 Frame Unroll (Optical Flow + Image Frames)	**88.6**

Table 7: UCF-101 results. The bold-face numbers represent results that are higher than previously reported results.

Learning Spatiotemporal Features with 3D Convolutional Networks (2015)

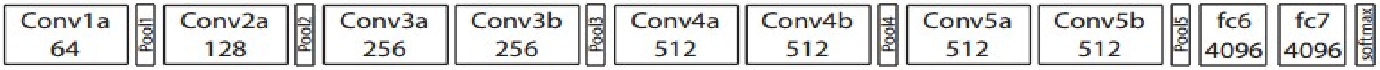
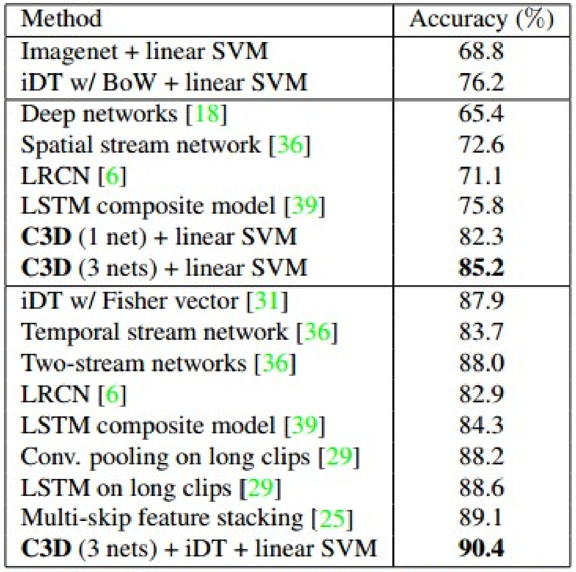


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are 3 × 3 × 3 with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are 2 × 2 × 2, except for `pool1` is 1 × 2 × 2. Each fully connected layer has 4096 output units.



Method	Accuracy (%)
Imagenet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [18]	65.4
Spatial stream network [36]	72.6
LRCN [6]	71.1
LSTM composite model [39]	75.8
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	**85.2**
iDT w/ Fisher vector [31]	87.9
Temporal stream network [36]	83.7
Two-stream networks [36]	88.0
LRCN [6]	82.9
LSTM composite model [39]	84.3
Conv. pooling on long clips [29]	88.2
LSTM on long clips [29]	88.6
Multi-skip feature stacking [25]	89.1
C3D (3 nets) + iDT + linear SVM	**90.4**

Table 3. **Action recognition results on UCF101**. C3D compared with baselines and current state-of-the-art methods. Top: simple features with linear SVM; Middle: methods taking only RGB frames as inputs; Bottom: methods using multiple feature combinations.

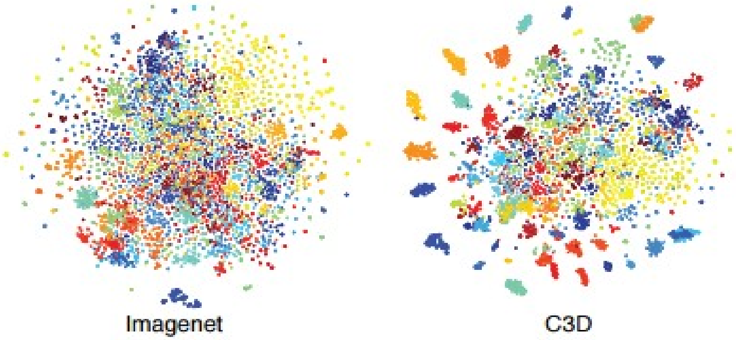
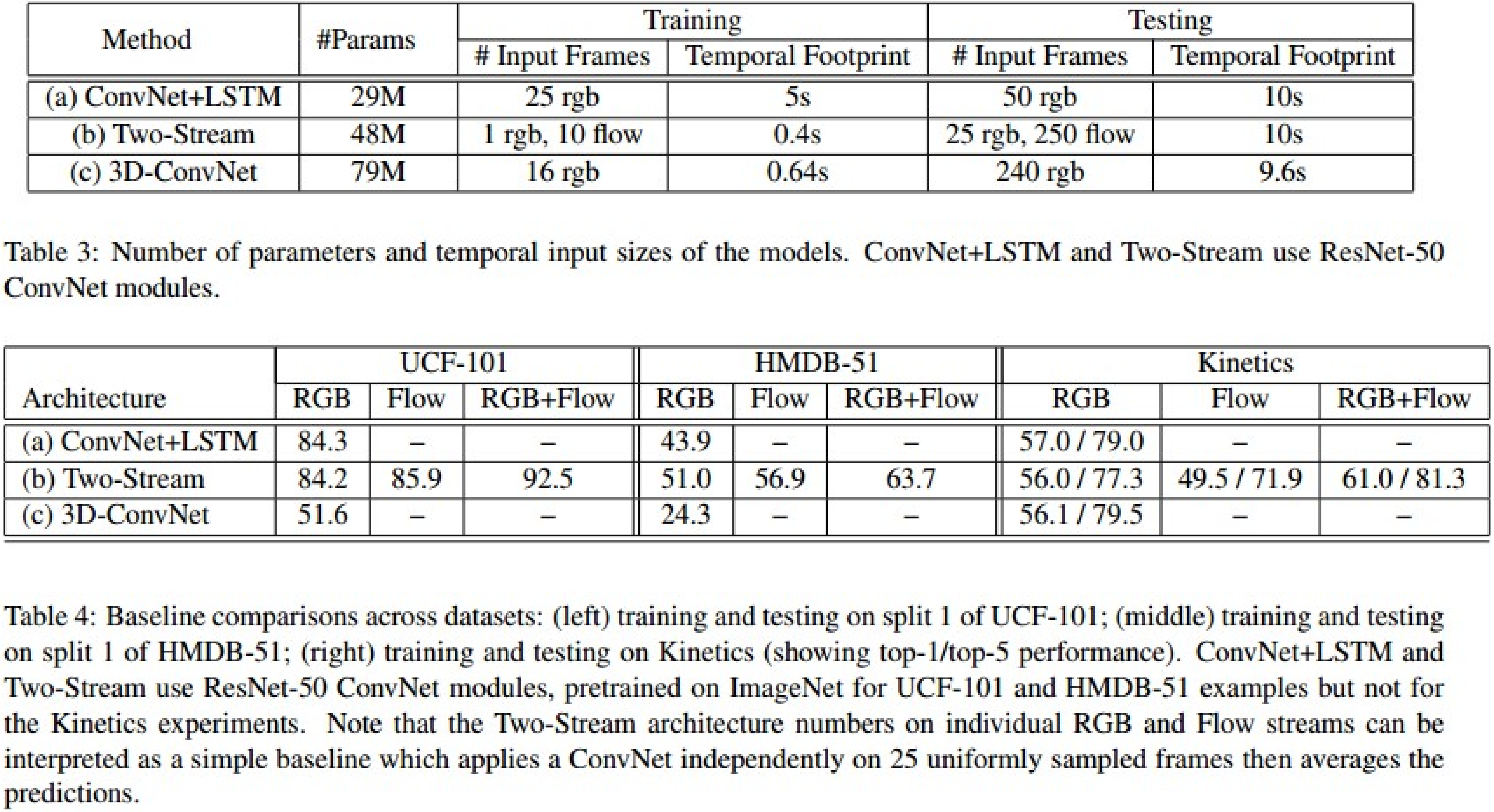


Figure 6. **Feature embedding**. Feature embedding visualizations of Imagenet and C3D on UCF101 dataset using t-SNE [43]. C3D features are semantically separable compared to Imagenet suggesting that it is a better feature for videos. Each clip is visualized as a point and clips belonging to the same action have the same color. Best viewed in color.

- Trained on Sports-1M dataset
- Include temporal dimension with convolution
- No optical flow
- 85.2%
- 90.4% accuracy with improved dense trajectories

The Kinetics Human Action Video Dataset (2017)

- ~300k videos, 10s, 400 action classes
- Baseline for popular models



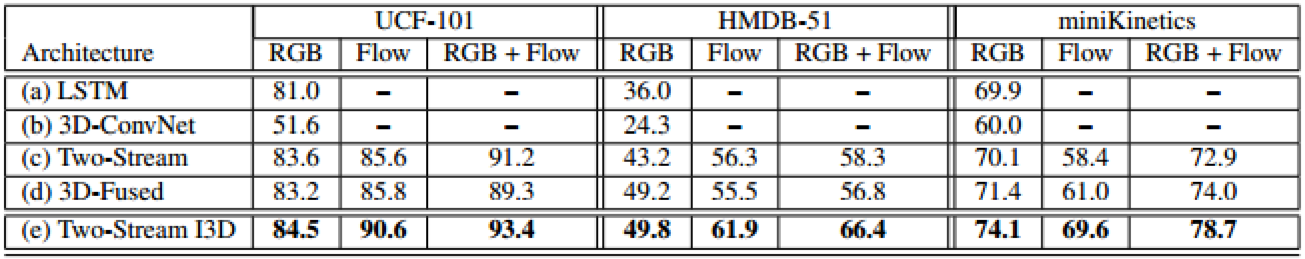
Method	#Params	Training		Testing	
		# Input Frames	Temporal Footprint	# Input Frames	Temporal Footprint
(a) ConvNet+LSTM	29M	25 rgb	5s	50 rgb	10s
(b) Two-Stream	48M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
(c) 3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s

Table 3: Number of parameters and temporal input sizes of the models. ConvNet+LSTM and Two-Stream use ResNet-50 ConvNet modules.

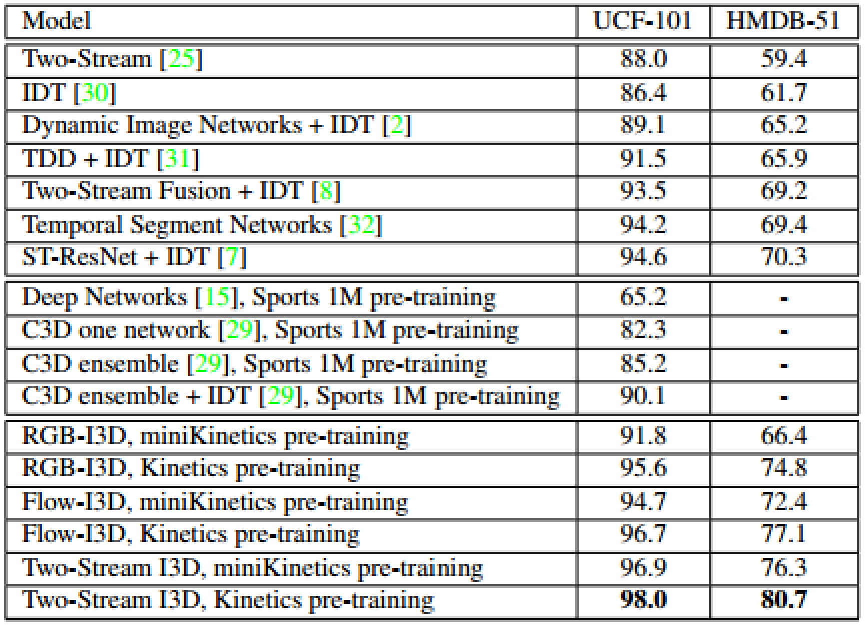
Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB+Flow	RGB	Flow	RGB+Flow	RGB	Flow	RGB+Flow
(a) ConvNet+LSTM	84.3	–	–	43.9	–	–	57.0 / 79.0	–	–
(b) Two-Stream	84.2	85.9	92.5	51.0	56.9	63.7	56.0 / 77.3	49.5 / 71.9	61.0 / 81.3
(c) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1 / 79.5	–	–

Table 4: Baseline comparisons across datasets: (left) training and testing on split 1 of UCF-101; (middle) training and testing on split 1 of HMDB-51; (right) training and testing on Kinetics (showing top-1/top-5 performance). ConvNet+LSTM and Two-Stream use ResNet-50 ConvNet modules, pretrained on ImageNet for UCF-101 and HMDB-51 examples but not for the Kinetics experiments. Note that the Two-Stream architecture numbers on individual RGB and Flow streams can be interpreted as a simple baseline which applies a ConvNet independently on 25 uniformly sampled frames then averages the predictions.

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (2017)



Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**



Model	UCF-101	HMDB-51
Two-Stream [25]	88.0	59.4
IDT [30]	86.4	61.7
Dynamic Image Networks + IDT [2]	89.1	65.2
TDD + IDT [31]	91.5	65.9
Two-Stream Fusion + IDT [8]	93.5	69.2
Temporal Segment Networks [32]	94.2	69.4
ST-ResNet + IDT [7]	94.6	70.3
Deep Networks [15], Sports 1M pre-training	65.2	-
C3D one network [29], Sports 1M pre-training	82.3	-
C3D ensemble [29], Sports 1M pre-training	85.2	-
C3D ensemble + IDT [29], Sports 1M pre-training	90.1	-
RGB-I3D, miniKinetics pre-training	91.8	66.4
RGB-I3D, Kinetics pre-training	95.6	74.8
Flow-I3D, miniKinetics pre-training	94.7	72.4
Flow-I3D, Kinetics pre-training	96.7	77.1
Two-Stream I3D, miniKinetics pre-training	96.9	76.3
Two-Stream I3D, Kinetics pre-training	**98.0**	**80.7**

- 3D Conv model using pretrained 2D Conv model weights
- 98% on UCF-101 after pretraining on Kinetics (95.6% using only RGB)

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? (2017)

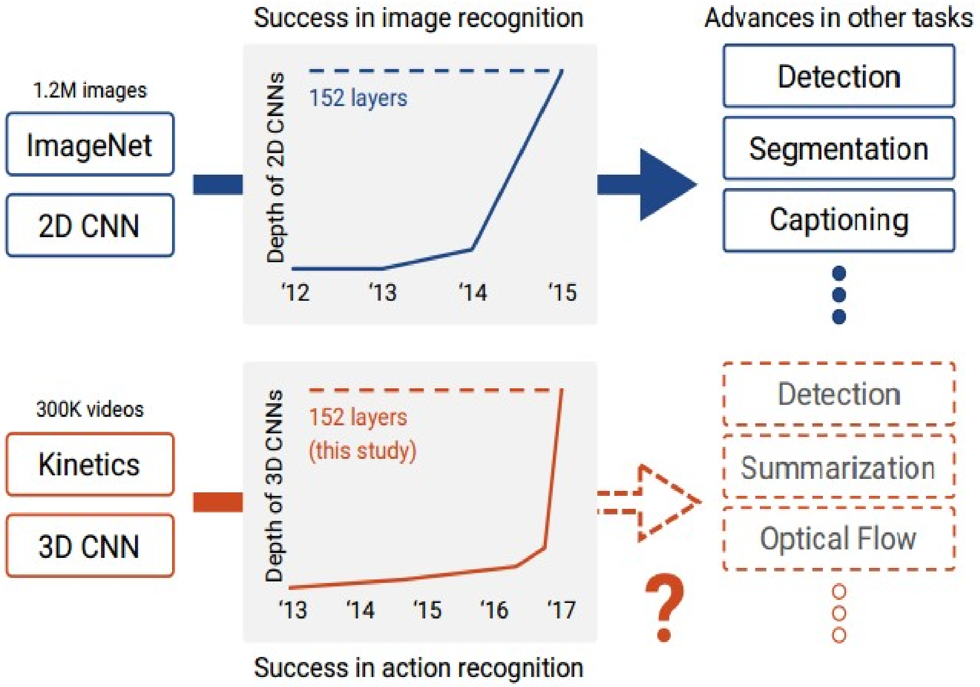
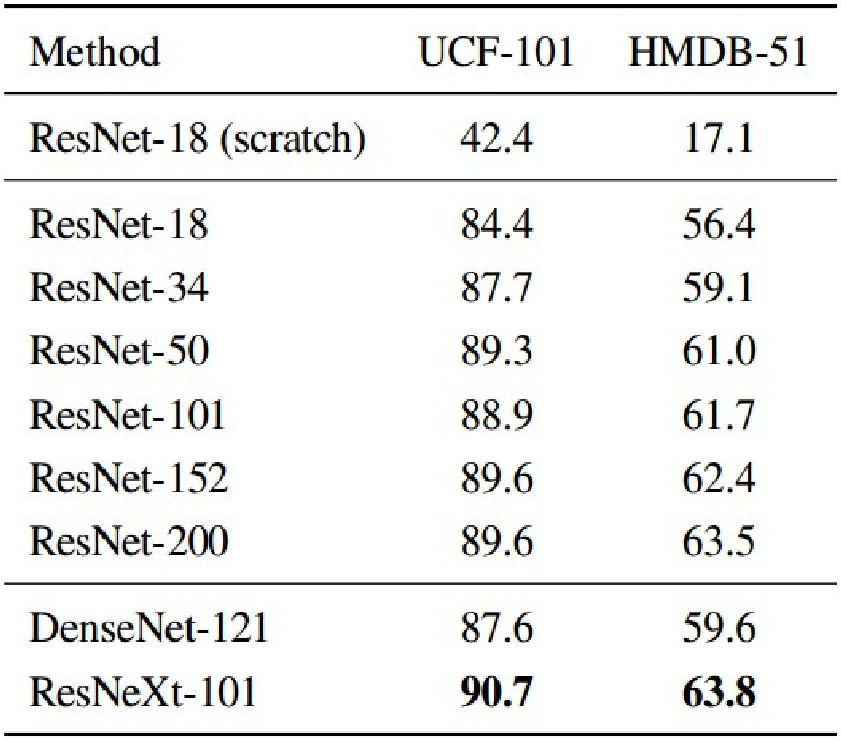


Figure 1: Recent advances in computer vision for images (top) and videos (bottom). The use of very deep 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos?

Table 4: Top-1 accuracies on UCF-101 and HMDB-51. All accuracies are averaged over three splits.



Method	UCF-101	HMDB-51
ResNet-18 (scratch)	42.4	17.1
ResNet-18	84.4	56.4
ResNet-34	87.7	59.1
ResNet-50	89.3	61.0
ResNet-101	88.9	61.7
ResNet-152	89.6	62.4
ResNet-200	89.6	63.5
DenseNet-121	87.6	59.6
ResNeXt-101	**90.7**	**63.8**

https://github.com/kenshohara/3D-ResNets-PyTorch

HW

- Part 1: Fine-tune a 50-layer ResNet model (pretrained on ImageNet) on UCF-101 video frames. Training: 1 Frame, Testing: Average over the video
- Part 2: Fine-tune a 50-layer 3D ResNet model (pretrained on Kinetics) on UCF-101 video sequences. Training: 16 Frames, Testing: Average over video
- Compare top performing and poor performing classes from both models