Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST) *CVPR18*

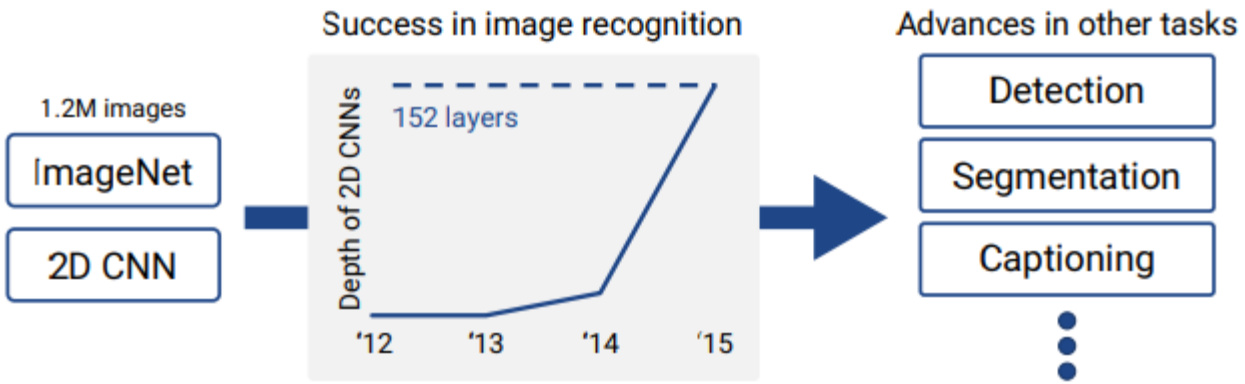
Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles
Dahun Kim, Donghyeon Cho, In So Kweon
Dept. of Electrical Engineering, KAIST, Daejeon, Korea *AAAI19*

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)

Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles
Dahun Kim, Donghyeon Cho, In So Kweon
Dept. of Electrical Engineering, KAIST, Daejeon, Korea

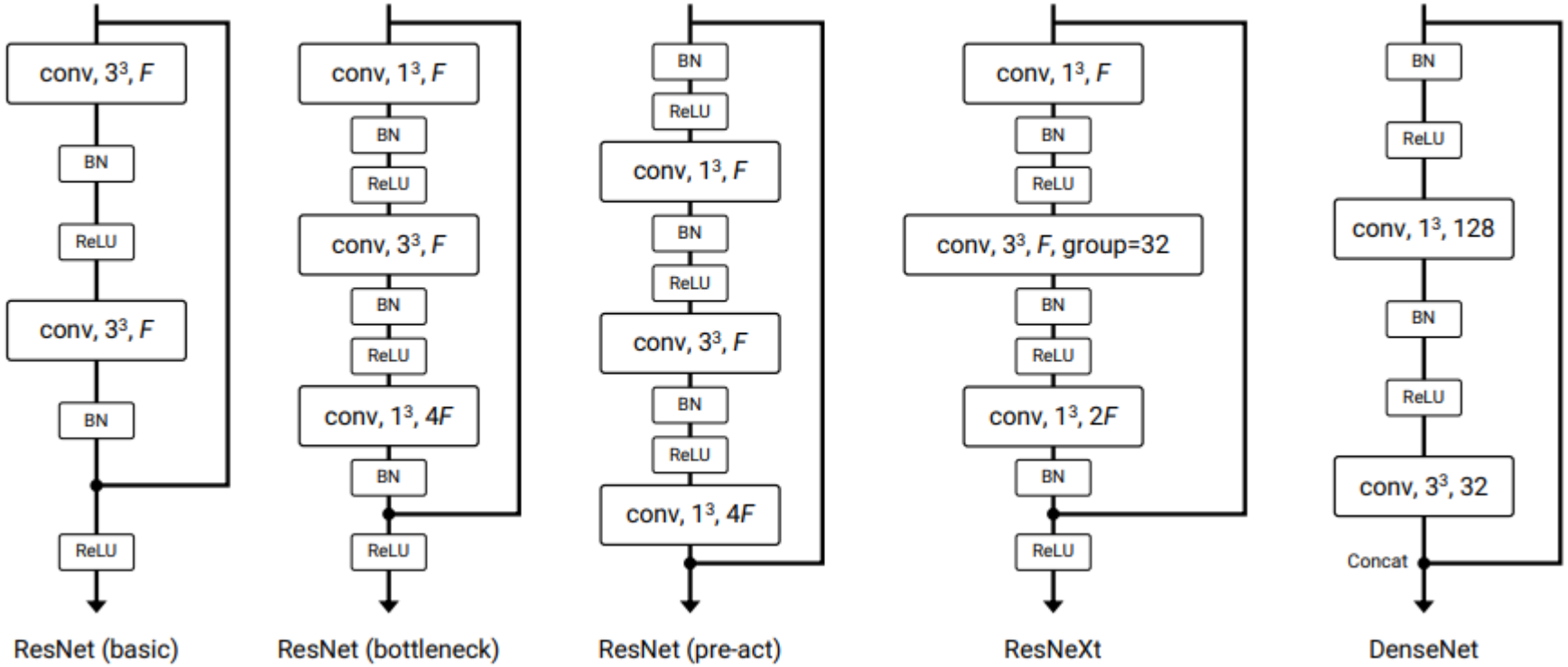
Problem Background

- The use of very deep 2D CNNs pretrained on ImageNet generate outstanding performance



Whether current video datasets have sufficient data for training very deep 3D CNNs?

Related Work: Network Architecture[1]



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

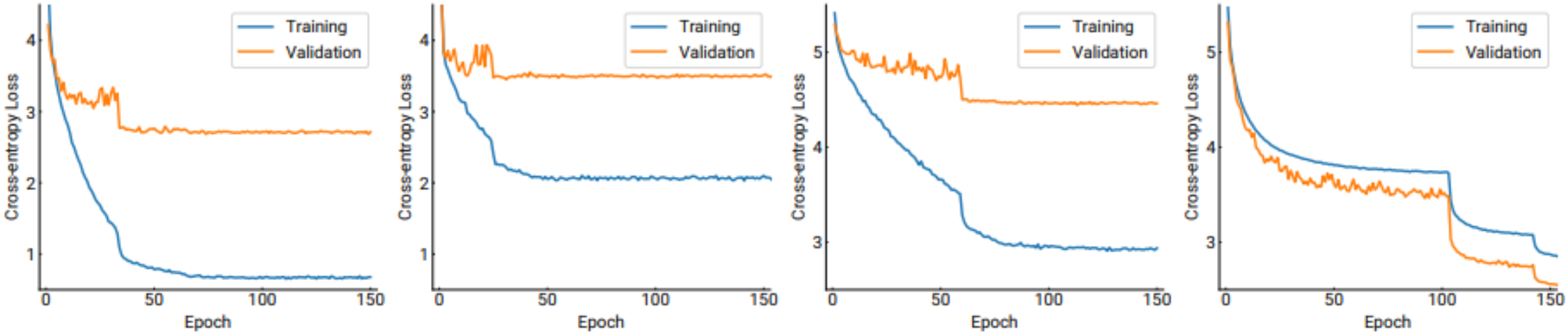
Experiment

- Datasets

  UCF101[1]: 101 action categories, 13320 videos.
  HMDB-51[2]: 51 human-action classes, 6766 videos.
  Activity-Net[3]:200 human-action classes, 27400 videos.
  Kinetics[4]: 400 action classes, 306245 videos.
  Sports-1M[5]
  YouTube-8M[6]

[1] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, 2012.
[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. ICCV, pages 2556–2563, 2011.
[3] B. G. Fabian Caba Heilbron, Victor Escorcia. ActivityNet: A large-scale video benchmark for human activity understanding. CVPR, pages 961–970, 2015.
[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. arXiv preprint, arXiv:1705.06950, 2017.
[5] A. Karpathy, G. Toderici, S. Shetty, R. Sukthankar, L. Fei-Fei. Large-scale video classification with convolutional neural networks. CVPR, pages 1725–1732, 2014.
[6] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. arXiv preprint, arXiv:1609.08675, 2016.

Experiment



(a) UCF-101 (split 1). (b) HMDB-51 (split 1). (c) ActivityNet. (d) Kinetics.

Figure 4: ResNet-18 training and validation losses. The validation losses on UCF-101, HMDB-51, and ActivityNet quickly converged to high values and were clearly higher than their corresponding training losses. The validation losses on Kinetics were slightly higher than the corresponding training losses, significantly different than those on the other datasets.

Experiment

Table 2: Accuracies on the Kinetics validation set. *Average* is averaged accuracy over *Top-1* and *Top-5*.

Method	Top-1	Top-5	Average
ResNet-18	54.2	78.1	66.1
ResNet-34	60.1	81.9	71.0
ResNet-50	61.3	83.1	72.2
ResNet-101	62.8	83.9	73.3
ResNet-152	63.0	**84.4**	**73.7**
ResNet-200	**63.1**	**84.4**	**73.7**
ResNet-200 (pre-act)	63.0	83.7	73.4
Wide ResNet-50	64.1	85.3	74.7
ResNeXt-101	**65.1**	**85.7**	**75.4**
DenseNet-101	59.7	81.9	70.8
DenseNet-201	61.3	83.3	72.3

Table 3: Accuracies on the Kinetics test set. *Average* is averaged accuracy over *Top-1* and *Top-5*. Here, we refer the results of RGB- and Two-stream I3D trained from scratch [3] for fair comparison.

Method	Top-1	Top-5	Average
ResNeXt-101	–	–	74.5
ResNeXt-101 (64f)	–	–	**78.4**
CNN+LSTM [16]	57.0	79.0	68.0
Two-stream CNN [16]	61.0	81.3	71.2
C3D w/ BN [16]	56.1	79.5	67.8
RGB-I3D [3]	68.4	88.0	78.2
Two-stream I3D [3]	71.6	90.0	**80.8**

Experiment

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**

Table 5: Top-1 accuracies on UCF-101 and HMDB-51 comared with the state-of-the-art methods. All accuracies are averaged over three splits. *Dim* indicate the dimension of convolution kernel.

Method	Dim	UCF-101	HMDB-51
ResNeXt-101	3D	90.7	63.8
ResNeXt-101 (64f)	3D	**94.5**	**70.2**
C3D [23]	3D	82.3	–
P3D [19]	3D	88.6	–
Two-stream I3D [3]	3D	**98.0**	**80.7**
Two-stream CNN [20]	2D	88.0	59.4
TDD [27]	2D	90.3	63.2
ST Multiplier Net [7]	2D	94.2	68.9
TSN [29]	2D	94.2	69.4

Conclusion

- Examined the architectures of various CNNs with 3D kernels on current video dataset.
- Kinetics dataset has sufficient data for training deep 3D CNNs similar to 2D CNNs on ImageNet.

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)

Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles
Dahun Kim, Donghyeon Cho, In So Kweon
Dept. of Electrical Engineering, KAIST, Daejeon, Korea

Problem Background

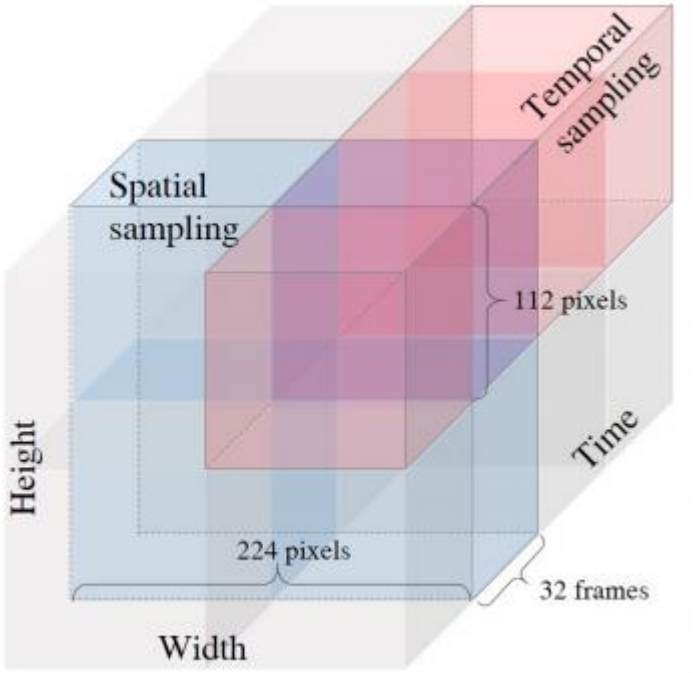
- Self-Supervised Learning
  Defines an annotation-free pretext task from raw data.
  e.g. Image patches permutation, motion frame ordering, etc.

Conclusion

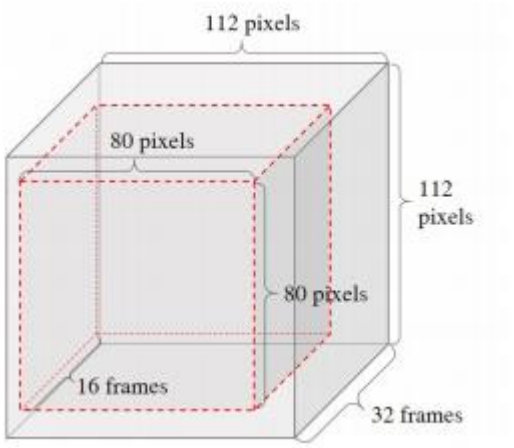
- Proposed a novel pretext task for 3D CNNs.
- Achieved good performance on video recognition datasets(UCF101, HMDB51).

Proposed Approach: Space-Time Cubic Puzzles

- Given a randomly permuted sequence of 3D spatio-temporal pieces cropped from a video clip, we train a network to predict their original arrangement.



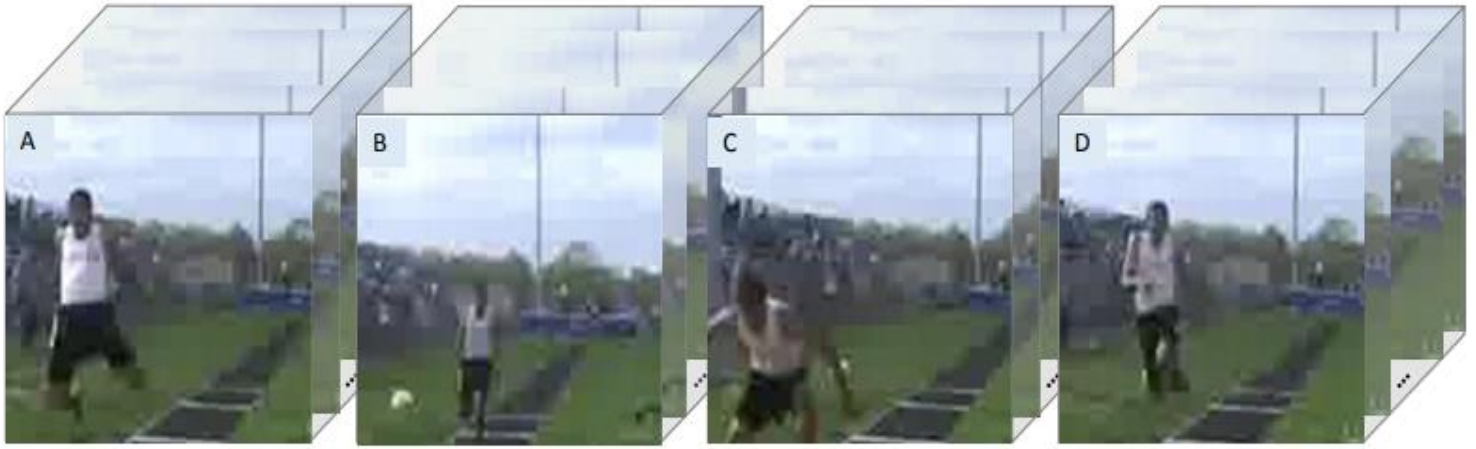
2 x 2 x 1 : spatio dimension
1 x 1 x 4 : temporal dimension



Proposed Approach: Space-Time Cubic Puzzles

(Spatial)

Q. Can you arrange these?

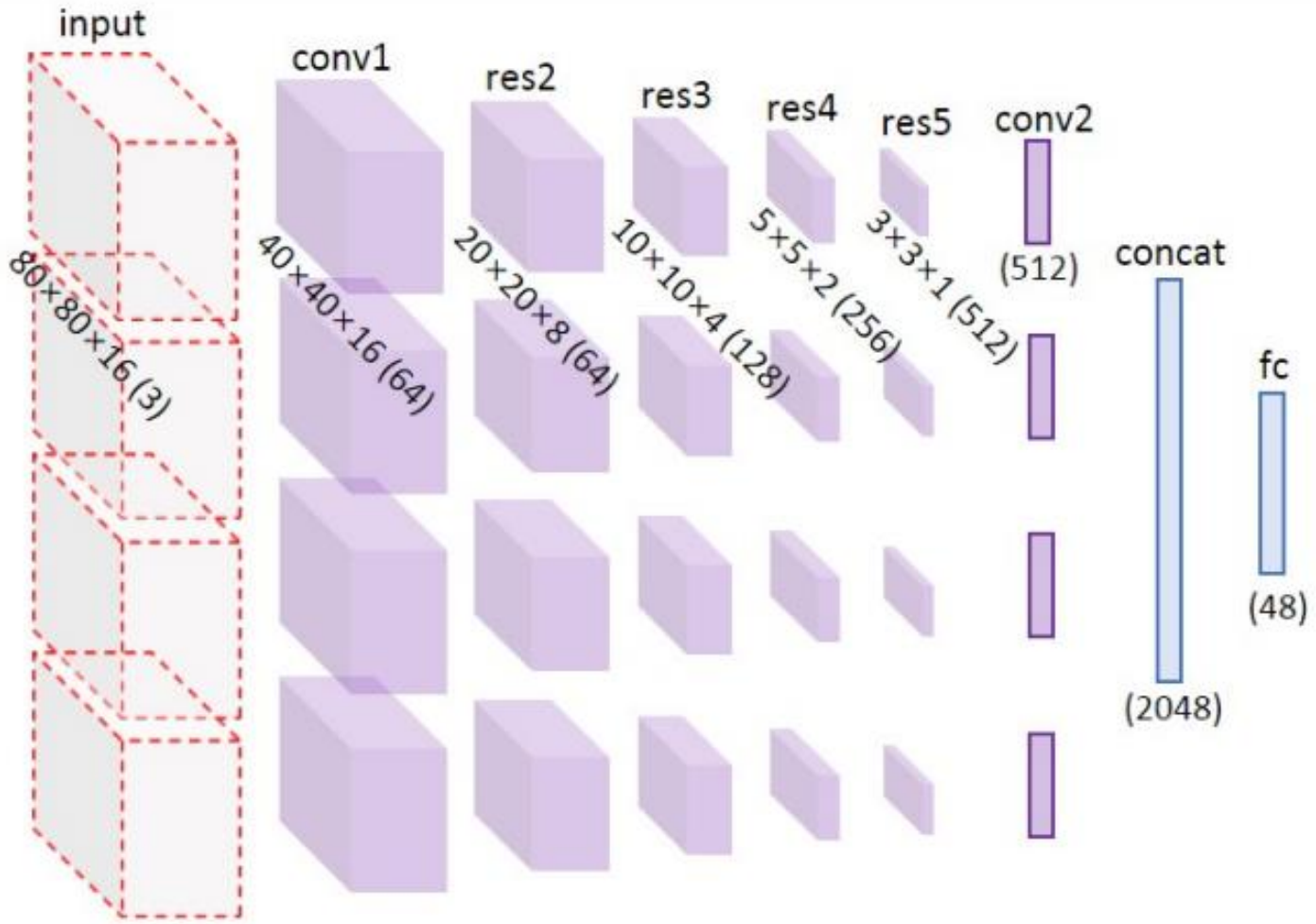


A. Spatial: A-D-B-C

A. Temporal: B-D-A-C

(Temporal)

Proposed Approach: Space-Time Cubic Puzzles



3D Resnet-18

48 = 4! * 2

Experiment

- Datasets

  UCF101[1]: 101 action categories, 13320 videos.
  HMDB-51[2]: 51 human-action classes, 6766 videos.
  Kinetics[3]: 400 action classes, 306245 videos.

[1] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, 2012.
[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. ICCV, pages 2556–2563, 2011.
[3] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. arXiv preprint, arXiv:1705.06950, 2017.

Experiment

Initialization	UCF101(%)	HMDB51(%)
Random init.	42.4	17.1
3D ST-puzzle (ours)	65.8	33.7
Kinetics 1/8	64.2	33.2
Kinetics 1/4	71.1	41.1
Kinetics 1/2	78.0	48.6
Kinetics full	84.4	56.4
ImageNet-inflated	60.3	30.7

Method	UCF101(%)
3D AE	48.7
3D AE + future	50.1
3D inpainting	50.9
3D S-puzzle	58.5
3D T-puzzle	59.3
3D ST score ensemble	61.3
3D ST-puzzle (full)	**65.8**

Table 2: **comparison with alternative methods.** Top-1 accuracies on UCF10. All methods use 3D ResNet-18, and the accuracies are averaged over three splits.

Experiment

Ablation Study

Method	UCF101(%)
with no regularizations	58.7
+ channel replication	61.5
+ random jittering	63.9
+ rotation with classification	65.8

Table 4: **Ablation studies.** Top-1 accuracies on UCF101. Each methods are accumulated down from the top and use 3D ResNet-18. The accuracies are averaged over three splits.

Experiment

Method	Backbone	UCF101(%)	HMDB51(%)
Random initialization	3D ResNet-18	42.4	17.1
Random initialization	AlexNet	38.4	13.4
Temporal Coherency (Mobahi, Collobert, and Weston 2009)	AlexNet	45.4	15.9
Object Patch (Wang and Gupta 2015)	AlexNet	42.7	15.6
Sequence Verification (Misra, Zitnick, and Hebert 2016)	AlexNet	50.9	19.8
OPN (Lee et al. 2017)	AlexNet	<u>56.3</u>	22.1
Geometry (Gan et al. 2018)	AlexNet	54.1	<u>22.6</u>
Time Arrow (Wei et al. 2018)	AlexNet	55.3	-
Video Generation (Vondrick, Pirsiavash, and Torralba 2016)	C3D	52.1	-
3D ST-puzzle (ours)	C3D	60.6	28.3
	3D ResNet-10	63.4	30.8
	3D ResNet-18	**65.8**	**33.7**