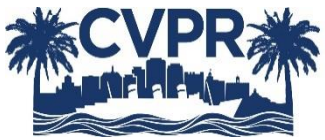


LONG BEACH
CALIFORNIA
June 16-20, 2019

ACTIVITYNET
Large Scale Activity Recognition Challenge

Local and Global Diffusion Networks for Trimmed Action Recognition

Zhaofan Qiu, Yingwei Pan, and Ting Yao
JD AI Reseach, Beijing, China

Presenter: Zhaofan Qiu

Outlines

ACTIVITYNET
Large Scale Activity Recognition Challenge

01 Local and Global Diffusion Networks

02 Backbone Networks

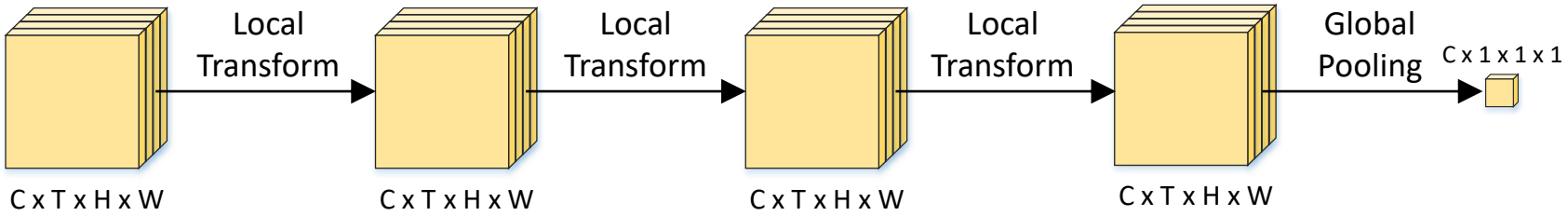
03 Feature Aggregation

04 Experimental Results

05 Take-Home Messages

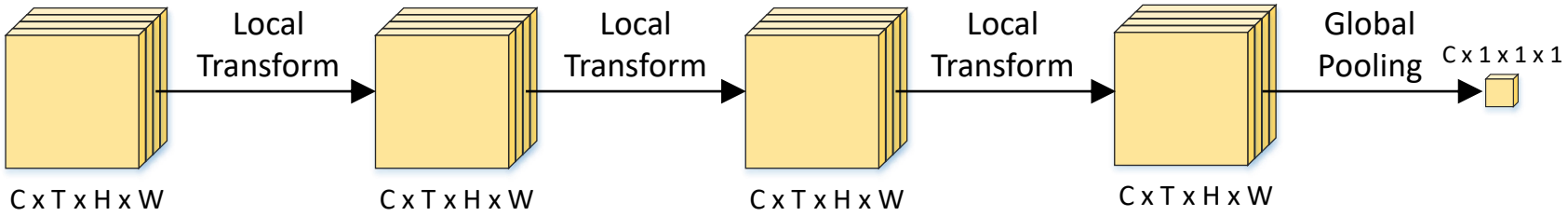
01 Local and Global Diffusion Networks

- Traditional 3D CNN:
  - Local representations + Local transformations + Global pooling



01 Local and Global Diffusion Networks

- Traditional 3D CNN:
  - Local representations + Local transformations + Global pooling

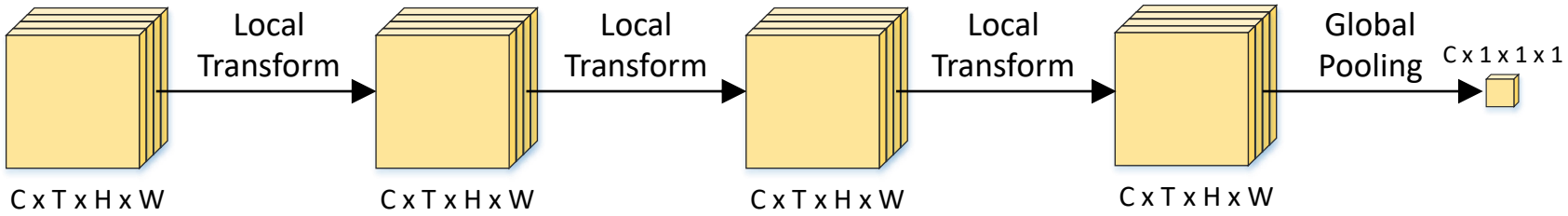


  - Local Transformation (2D/3D convolutions, 2D/3D poolings) ignores the large-range dependency
  - Large number of transformations for large receptive field
  - Holistic view of input video(clip) is only involved after global pooling

01 Local and Global Diffusion Networks

“Learning Spatio-Temporal Representation with Local and Global Diffusion”

[Qiu et al. CVPR 2019]

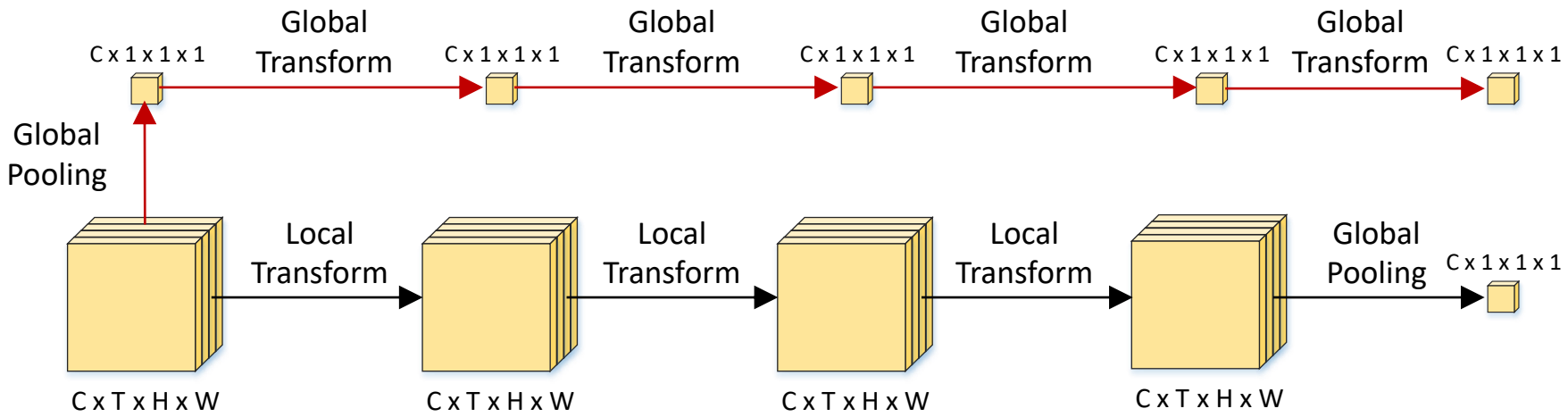


01 Local and Global Diffusion Networks

"Learning Spatio-Temporal Representation with Local and Global Diffusion"

[Qiu et al. CVPR 2019]

1. Local Path + Global Path

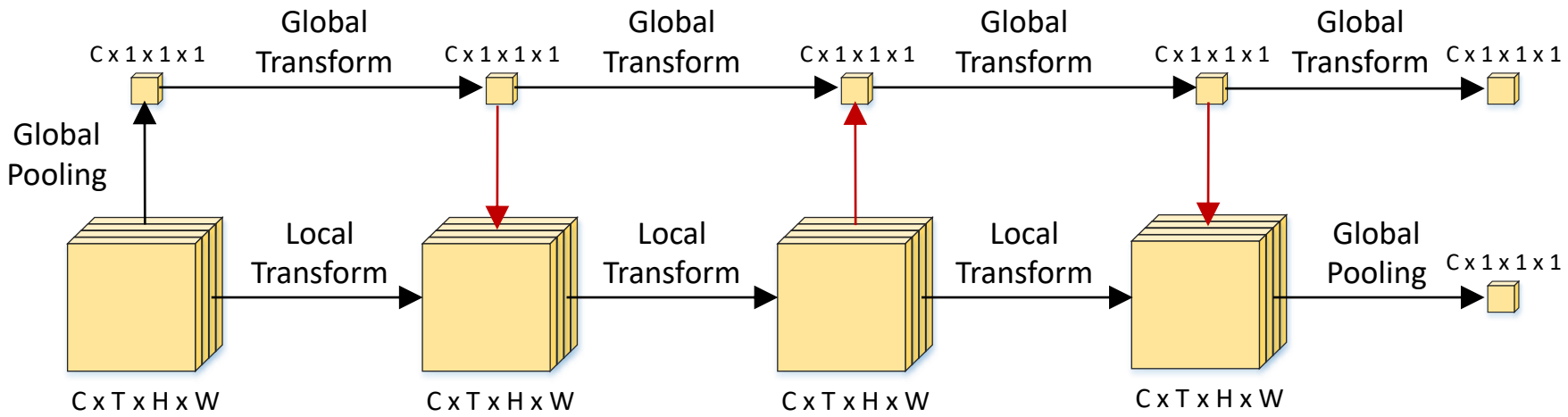


01 Local and Global Diffusion Networks

"Learning Spatio-Temporal Representation with Local and Global Diffusion"

[Qiu et al. CVPR 2019]

1. Local Path + Global Path
2. Local and Global Diffusion (two-direction)

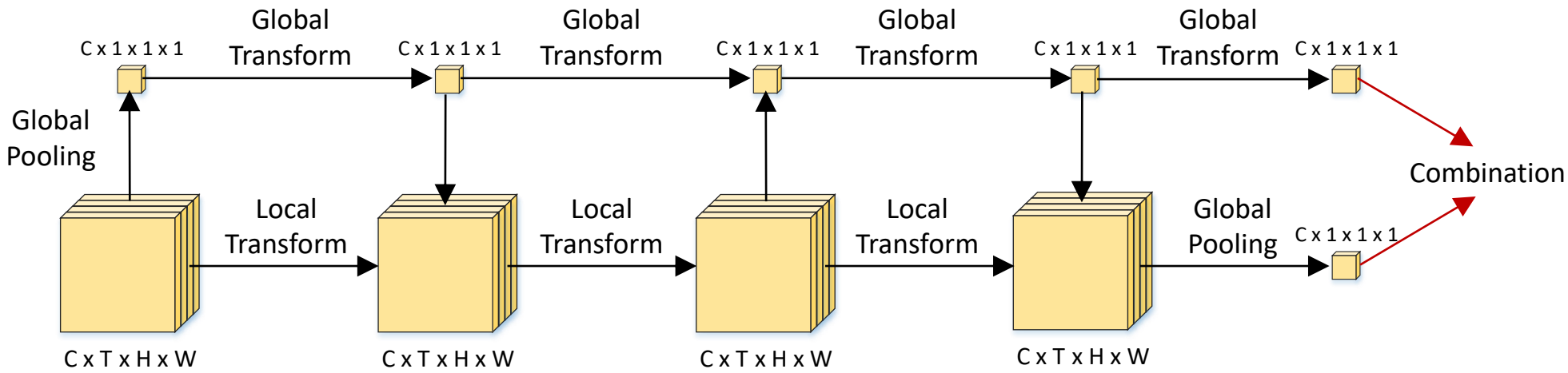


01 Local and Global Diffusion Networks

"Learning Spatio-Temporal Representation with Local and Global Diffusion"

[Qiu et al. CVPR 2019]

1. Local Path + Global Path
2. Local and Global Diffusion (two-direction)
3. Local and Global Combination



01 Local and Global Diffusion Networks

“Learning Spatio-Temporal Representation with Local and Global Diffusion”

[Qiu et al. CVPR 2019]

- Local transformation: 2D/3D convolutions, 2D/3D poolings
- Global transformation: non-linear mapping

01 Local and Global Diffusion Networks

"Learning Spatio-Temporal Representation with Local and Global Diffusion"

[Qiu et al. CVPR 2019]

- Local transformation: 2D/3D/P3D convolutions, 2D/3D poolings
- Global transformation: non-linear mapping
- Diffusion:
  - (1) Global-to-local diffusion:
  - The global residual from global path is broadcasted to each local position in local path:

$$\mathbf{x}_l = \mathrm{ReLU}(\mathcal{F}(\mathbf{x}_{l-1}) + \mathcal{US}(\mathbf{W}^{x,g}\mathbf{g}_{l-1}))$$

  - (2) Local-to-global diffusion:
  - The global-average-pooled local representation is linearly embedded into global path:

$$\mathbf{g}_l = \mathrm{ReLU}(\mathbf{W}^{g,x}\mathcal{P}(\mathbf{x}_l) + \mathbf{W}^{g,g}\mathbf{g}_{l-1})$$

01 Local and Global Diffusion Networks

“Learning Spatio-Temporal Representation with Local and Global Diffusion”

[Qiu et al. CVPR 2019]

- In this paper, the experiments are conducted on Kinetics-400, Kinetics-600, UCF101, HMDB51 for action recognition and J-HMDB, UCF101D for spatio-temporal action detection.

01 Local and Global Diffusion Networks

"Learning Spatio-Temporal Representation with Local and Global Diffusion"

[Qiu et al. CVPR 2019]

- In this paper, the experiments are conducted on Kinetics-400, Kinetics-600, UCF101, HMDB51 for action recognition and J-HMDB, UCF101D for spatio-temporal action detection.
- For more details and experimental results, please check our paper:

  "Learning Spatio-Temporal Representation with Local and Global Diffusion"

- Or come to our poster:

  Jun 20, 15:20, poster #128

- The code & model in this paper will be released (after conference) at:

  https://github.com/ZhaofanQiu/local-and-global-diffusion-networks

Outlines

ACTIVITYNET
Large Scale Activity Recognition Challenge

01 Local and Global Diffusion Networks

02 Backbone Networks

03 Feature Aggregation

04 Experimental Results

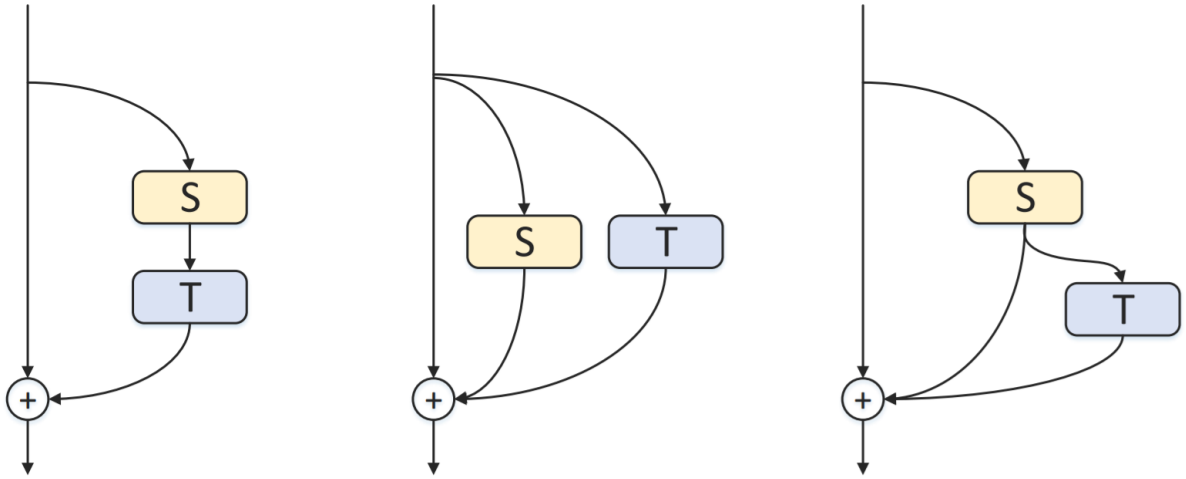
05 Take-Home Messages

02 Backbone Networks

- **Local and Global Diffusion** is a general component that can be injected into any existing 3D CNN backbones.

02 Backbone Networks

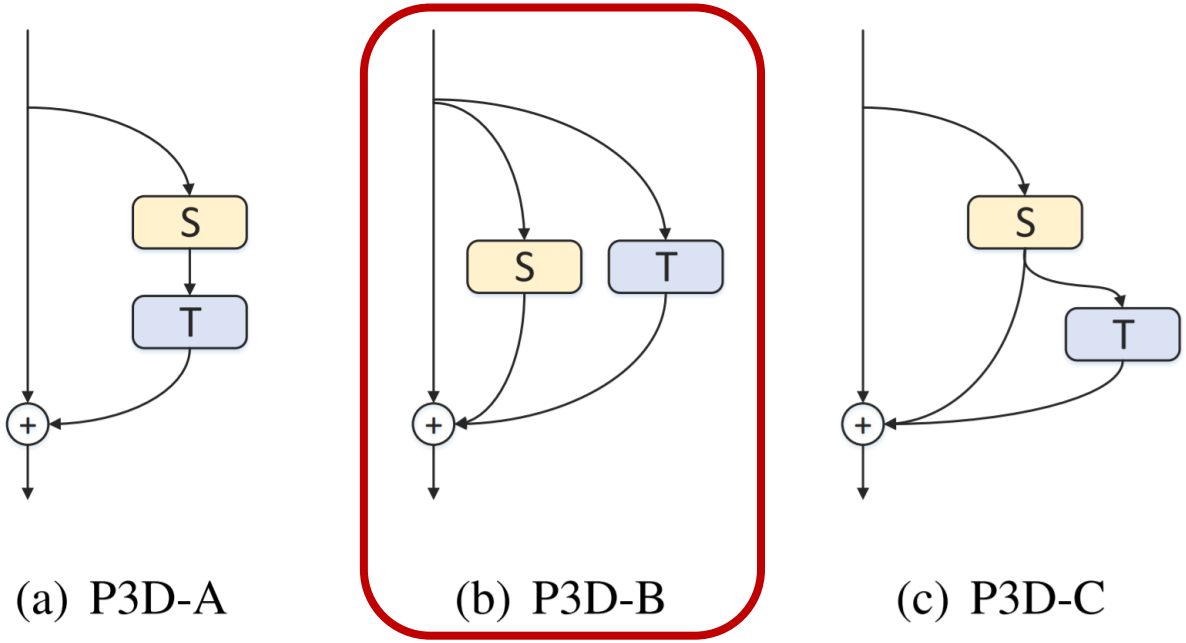
- **Local and Global Diffusion** is a general component that can be injected into any existing 3D CNN backbones.
- Pseudo-3D (P3D) Networks [Qiu et al. ICCV 2017]



(a) P3D-A (b) P3D-B (c) P3D-C

02 Backbone Networks

- **Local and Global Diffusion** is a general component that can be injected into any existing 3D CNN backbones.
- Pseudo-3D (P3D) Networks [Qiu et al. ICCV 2017]

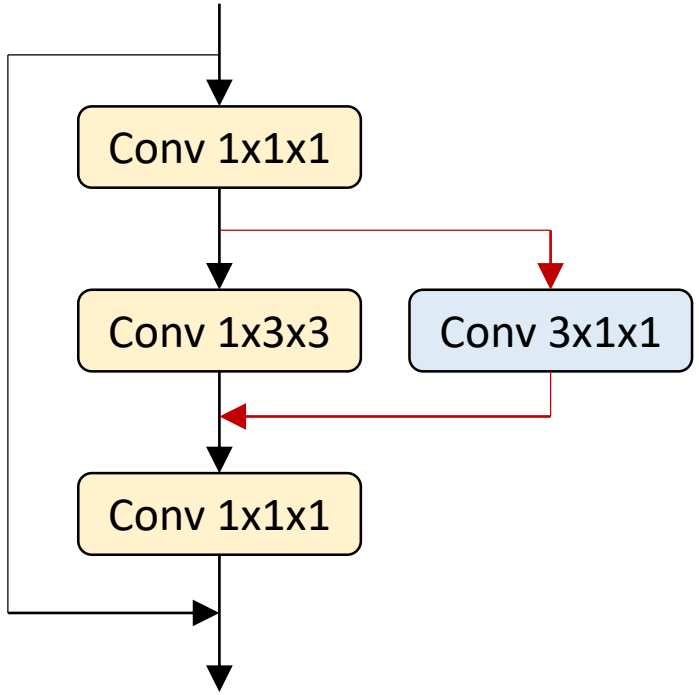


(a) P3D-A (b) P3D-B (c) P3D-C

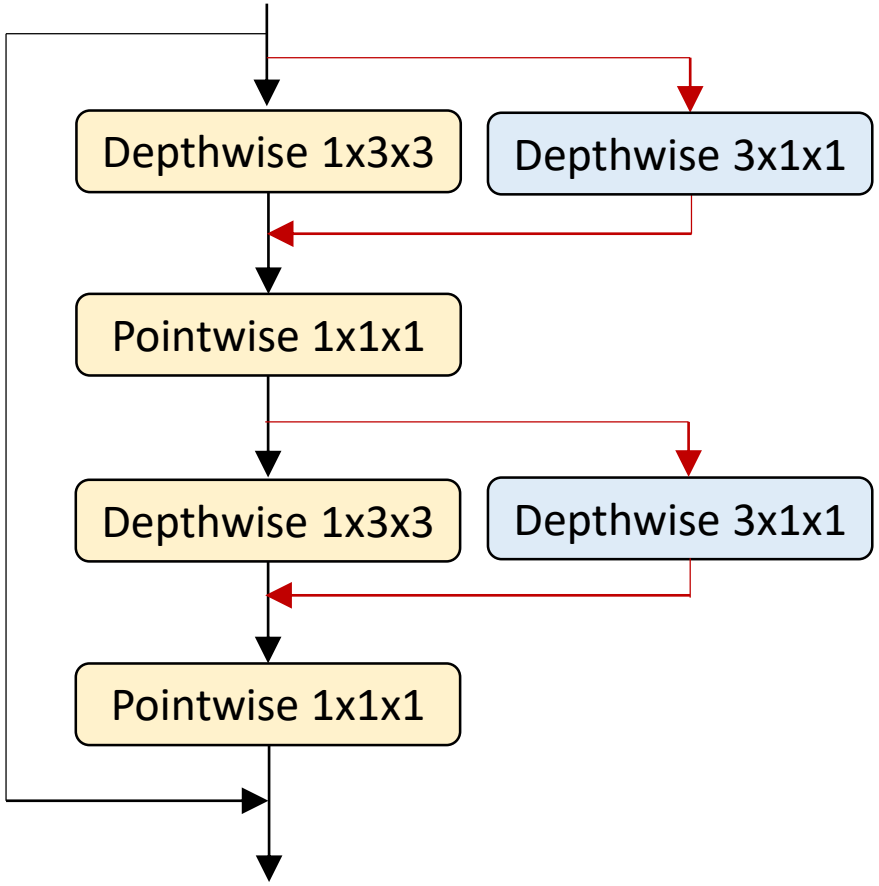
02 Backbone Networks

- Pseudo-3D (P3D) Networks [Qiu et al. ICCV 2017]

P3D-ResNet

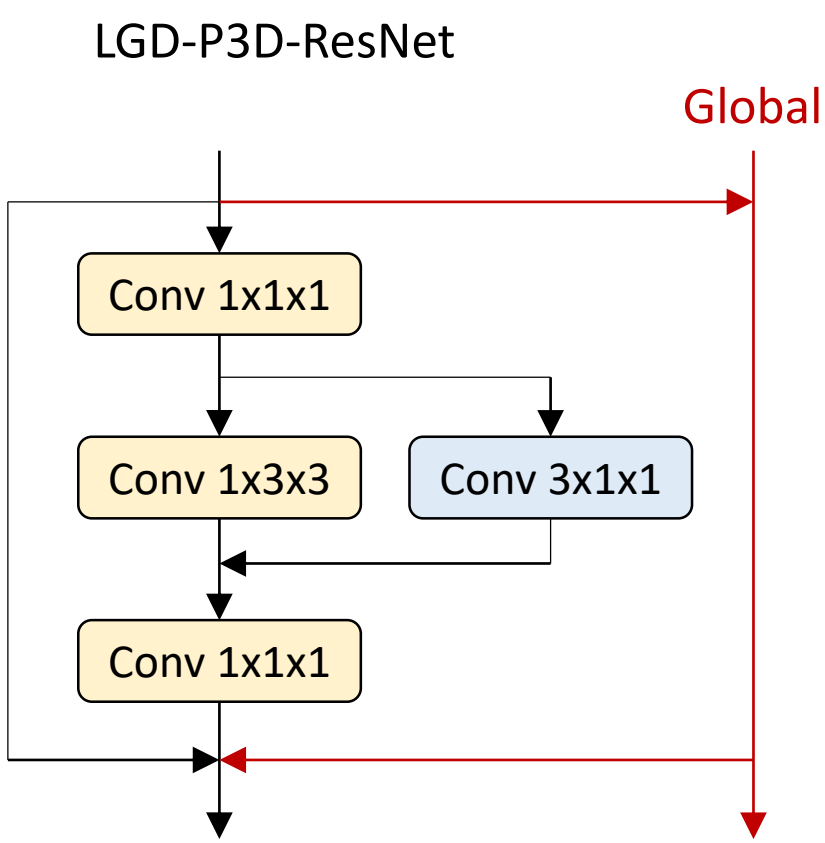


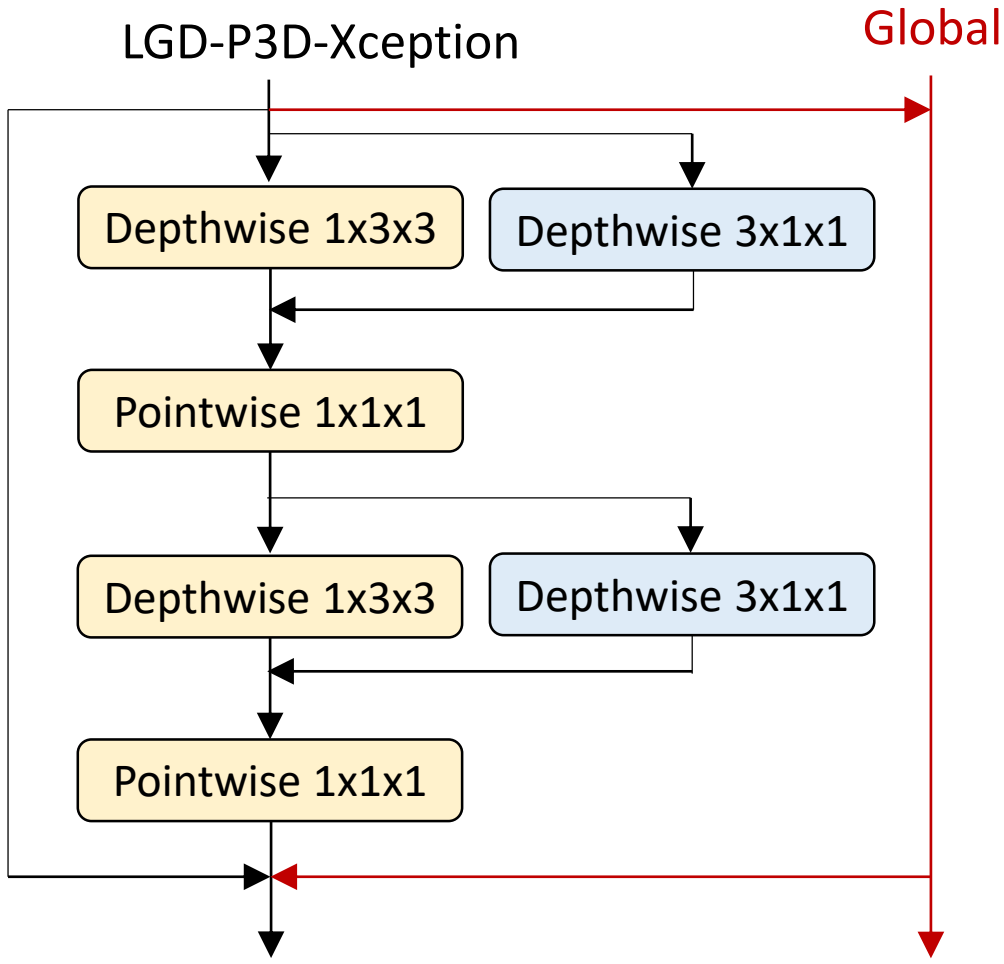
P3D-Xception



02 Backbone Networks

- Pseudo-3D (P3D) Networks [Qiu et al. ICCV 2017]





Outlines

03 Feature Aggregation

- The backbones (LGD-P3D-ResNet, LGD-P3D-Xception) are trained on short/long clips (16-frame and 128-frame), the video-level prediction is produced by clip feature aggregation

03 Feature Aggregation

- The backbones (LGD-P3D-ResNet, LGD-P3D-Xception) are trained on short/long clips (16-frame and 128-frame), the video-level prediction is produced by clip feature aggregation
- **Average Pooling** (AP): Score extraction + average pooling. No additional training.

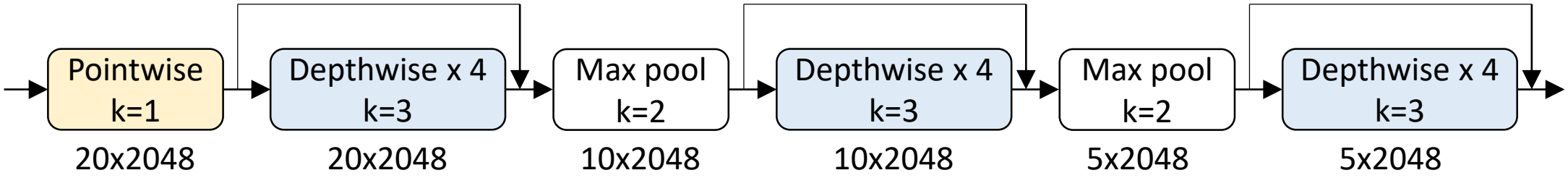
$$R_{AP} = \frac{1}{N} \sum_{i:f_i \in F} f_i$$

- **Temporal Convolutional Pooling** (TCP): Feature extraction + temporal 1D convolution. Need additional training.

$$R_{TCP} = Conv1D(\{f_1, f_2, ..., f_N\})$$

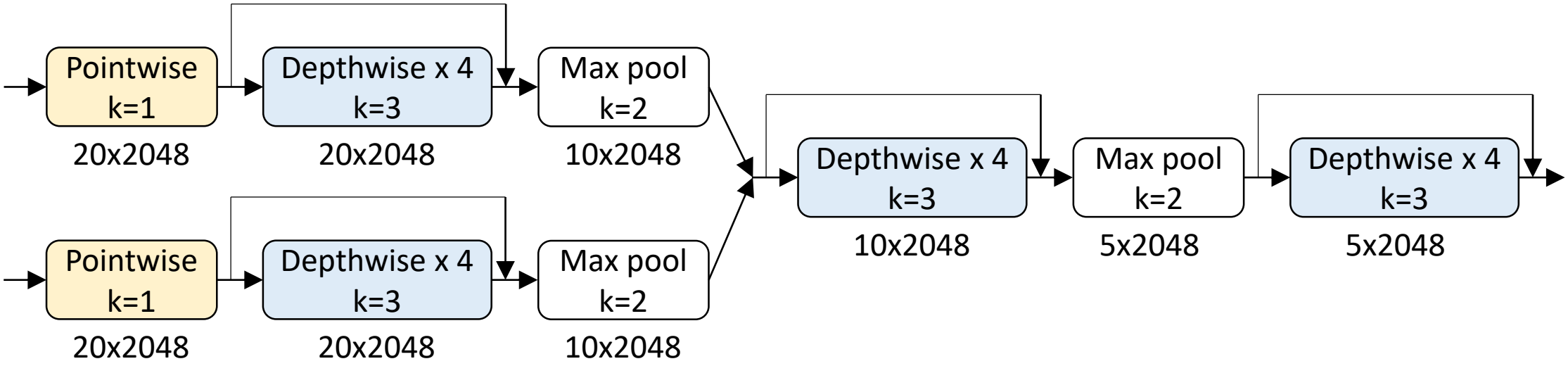
03 Feature Aggregation

- **Temporal Convolutional Pooling** (TCP)
- Conv1D architecture (uniformly sample 20 clips each video)



03 Feature Aggregation

- **Temporal Convolutional Pooling** (TCP)
- Conv1D for multi-stream fusion



Outlines

ACTIVITYNET
Large Scale Activity Recognition Challenge

04 Experimental Results

Validation set

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	**70.6%**	**89.3%**

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	**70.6%**	**89.3%**
Frame	LGD-P3D-ResNet101	128	AP	69.7%	88.8%

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	**70.6%**	**89.3%**
Frame	LGD-P3D-ResNet101	128	AP	69.7%	88.8%

ImageNet pre-training

Kinetics-600 pre-training

04 Experimental Results

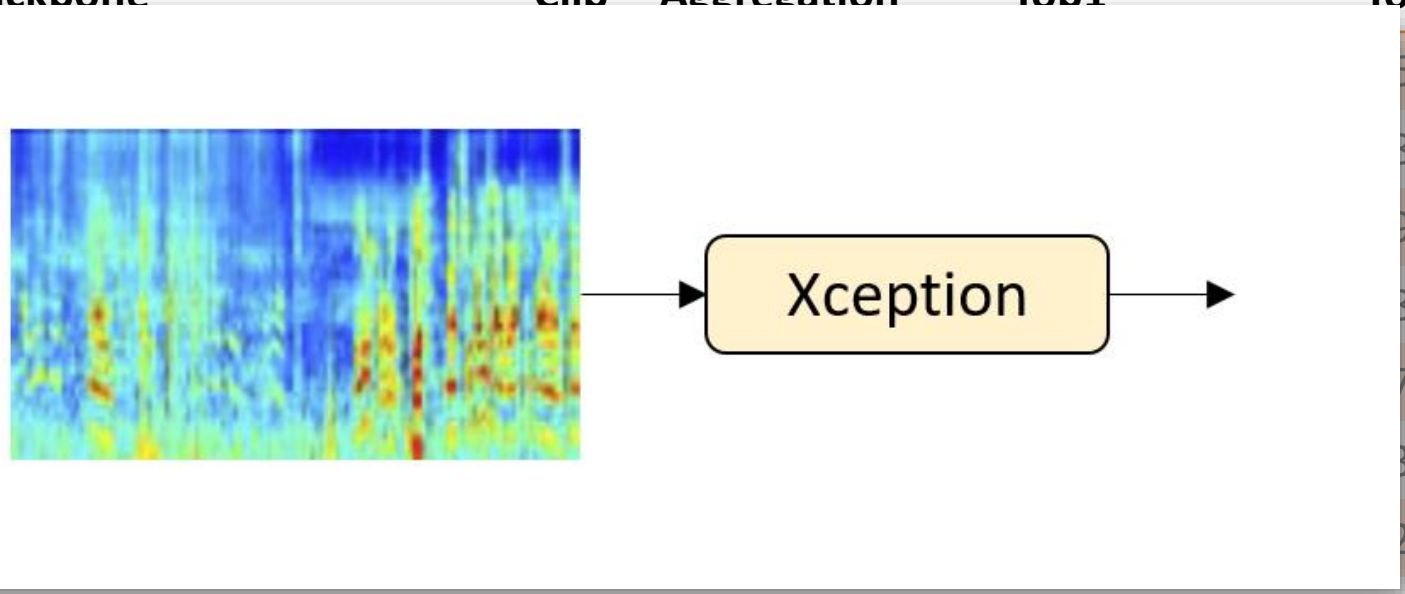
Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	70.6%	89.3%
Frame	LGD-P3D-ResNet101	128	AP	69.7%	88.8%
Flow	LGD-P3D-Xception	16	AP	54.4%	77.2%
Flow	LGD-P3D-Xception	128	AP	61.3%	83.1%
Flow	LGD-P3D-Xception	16	TCP	60.5%	82.4%
Flow	LGD-P3D-ResNet101	128	AP	**64.4%**	**85.5%**

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	70.6%	89.3%
Frame	LGD-P3D-ResNet101	128	AP	69.7%	88.8%
Flow	LGD-P3D-Xception	16	AP	54.4%	77.2%
Flow	LGD-P3D-Xception	128	AP	61.3%	83.1%
Flow	LGD-P3D-Xception	16	TCP	60.5%	82.4%
Flow	LGD-P3D-ResNet101	128	AP	64.4%	85.5%
Audio	Xception	-	AP	**21.9%**	36.8%
Audio	Xception	-	TCP	21.7%	**36.9%**

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame					6.9%
Frame					3.3%
Frame					9.3%
Frame					3.8%
Flow					7.2%
Flow					3.1%
Flow					2.4%
Flow	LGD-P3D-ResNet101	128	AP	64.4%	85.5%
Audio	Xception	-	AP	**21.9%**	36.8%
Audio	Xception	-	TCP	21.7%	**36.9%**



04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Frame	LGD-P3D-Xception	16	AP	67.5%	86.9%
Frame	LGD-P3D-Xception	128	AP	69.8%	88.3%
Frame	LGD-P3D-Xception	16	TCP	**70.6%**	**89.3%**
Frame	LGD-P3D-ResNet101	128	AP	69.7%	88.8%
Flow	LGD-P3D-Xception	16	AP	54.4%	77.2%
Flow	LGD-P3D-Xception	128	AP	61.3%	83.1%
Flow	LGD-P3D-Xception	16	TCP	60.5%	82.4%
Flow	LGD-P3D-ResNet101	128	AP	64.4%	85.5%
Audio	Xception	-	AP	21.9%	36.8%
Audio	Xception	-	TCP	21.7%	36.9%

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Two-stream	LGD-P3D-Xception	16	AP	69.3%	88.0%
Two-stream	LGD-P3D-Xception	128	AP	71.7%	89.7%
Two-stream	LGD-P3D-Xception	16	TCP	**72.8%**	**90.7%**
Two-stream	LGD-P3D-ResNet101	128	AP	72.3%	90.4%

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Two-stream	LGD-P3D-Xception	16	AP	69.3%	88.0%
Two-stream	LGD-P3D-Xception	128	AP	71.7%	89.7%
Two-stream	LGD-P3D-Xception	16	TCP	72.8%	90.7%
Two-stream	LGD-P3D-ResNet101	128	AP	72.3%	90.4%
+audio	LGD-P3D-Xception	16	AP	70.9%	88.8%
+audio	LGD-P3D-Xception	128	AP	73.9%	90.9%
+audio	LGD-P3D-Xception	16	TCP	**74.8%**	**91.7%**
+audio	LGD-P3D-ResNet101	128	AP	74.1%	91.4%

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Two-stream	LGD-P3D-Xception	16	AP	69.3%	88.0%
Two-stream	LGD-P3D-Xception	128	AP	71.7%	89.7%
Two-stream	LGD-P3D-Xception	16	TCP	72.8%	90.7%
Two-stream	LGD-P3D-ResNet101	128	AP	72.3%	90.4%
+audio	LGD-P3D-Xception	16	AP	70.9%	88.8%
+audio	LGD-P3D-Xception	128	AP	73.9%	90.9%
+audio	LGD-P3D-Xception	16	TCP	74.8%	91.7%
+audio	LGD-P3D-ResNet101	128	AP	74.1%	91.4%
Ensemble	-	-	-	**76.4%**	**92.8%**

04 Experimental Results

Input	Backbone	Clip	Aggregation	Top1	Top5
Two-stream	LGD-P3D-Xception	16	AP	69.3%	88.0%
Two-stream	LGD-P3D-Xception	128	AP	71.7%	89.7%
Two-stream	LGD-P3D-Xception	16	TCP	72.8%	90.7%
Two-stream	LGD-P3D-ResNet101	128	AP	72.3%	90.4%
+audio	LGD-P3D-Xception	16	AP	70.9%	88.8%
+audio	LGD-P3D-Xception	128	AP	73.9%	90.9%
+audio	LGD-P3D-Xception	16	TCP	74.8%	91.7%
+audio	LGD-P3D-ResNet101	128	AP	74.1%	91.4%
Ensemble	-	-	-	**76.4%**	**92.8%**
Ensemble+				**77.1%**	**93.0%**

Ensemble+: other backbones (SENet, ResNeXt) that are not fully accomplished, and some models trained more than once.

Outlines

01 Local and Global Diffusion Networks

02 Backbone Networks

03 Feature Aggregation

04 Experimental Results

05 Take-Home Messages

05 Take-Home Messages

1. Local and Global Diffusion + Pseudo-3D Convolution provides efficient and economic way for discriminative spatio-temporal representation learning (especially with limited time)
2. Temporal Convolutional Pooling is (potentially) more powerful than simple average pooling, however, the additional training may improve the training cost (that is the reason why ResNet + TCP is not accomplished during the challenge)
3. Even the 3D CNNs on RGB frame become more and more powerful, the optical flow and audio streams are still important for video understanding

Future directions:

1. The diffusion functions for LGD
2. The choice of conv1d architecture for TCP
3. The way of multi-stream fusion

Thanks!

Zhaofan Qiu

zhaofanqiu@gmail.com

http://zhaofanqiu.deepfun.club

Poster:

Jun 20, 15:20, poster #128

Resources:

https://github.com/ZhaofanQiu/local-and-global-diffusion-networks