PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences

Hehe Fan[1], Xin Yu[2], Yuhang Ding[3], Yi Yang[2] and Mohan Kankanhalli[1]

[1]School of Computing, National University of Singapore

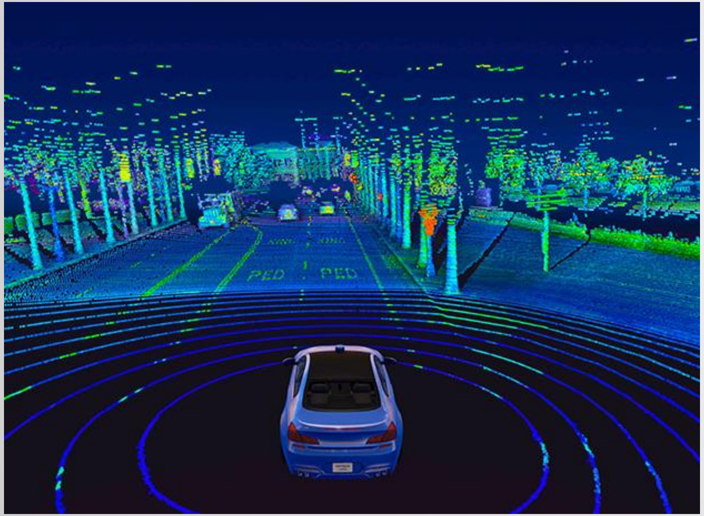
[2]ReLER, University of Technology Sydney

[3]Baidu Research

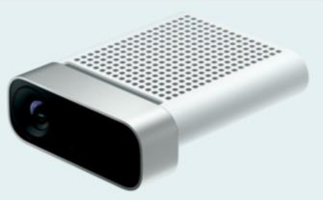
NUS National University of Singapore

LiDAR is Becoming a Real Business

Self-driving car



2020



Azure Kinect
Microsoft

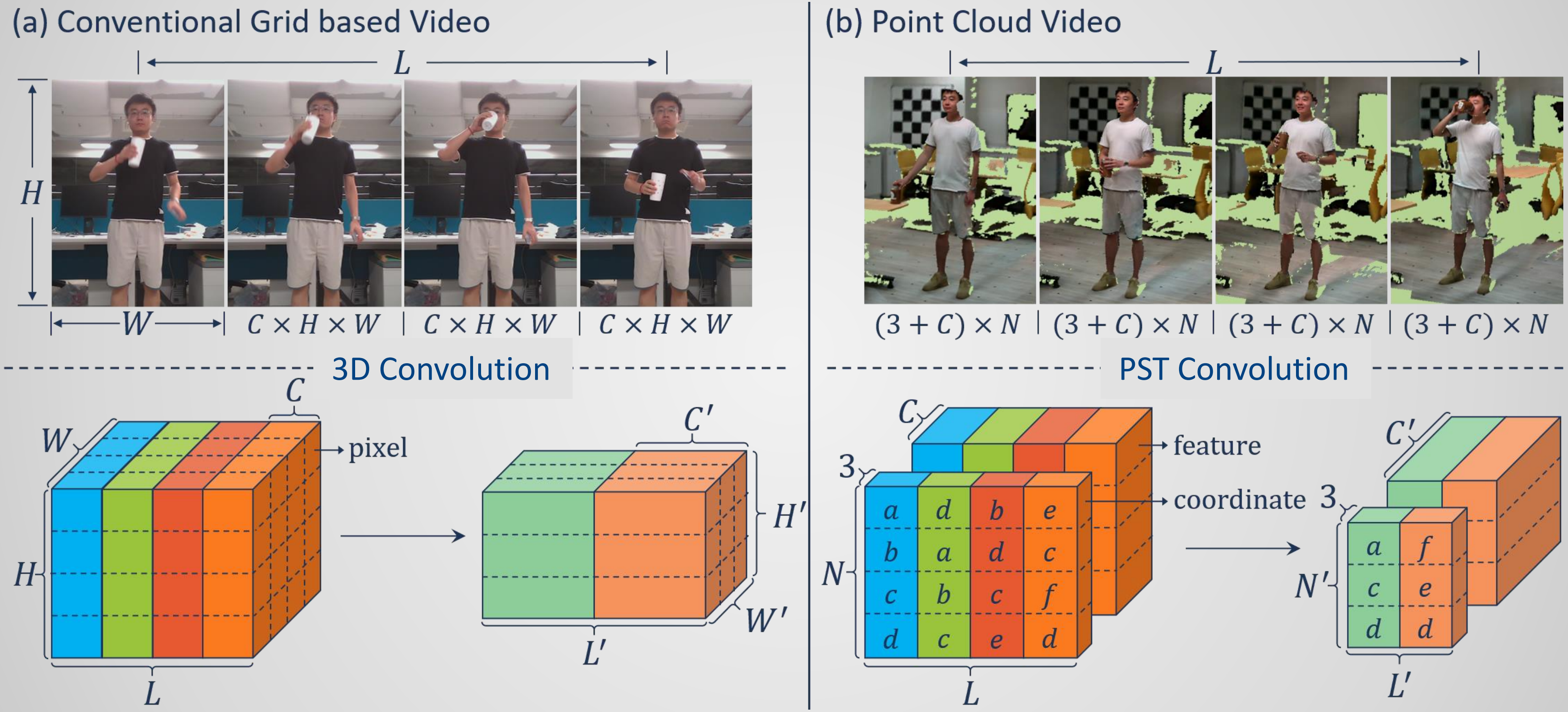
RealSense L515
Intel

iPad/iPhone Pro
Apple

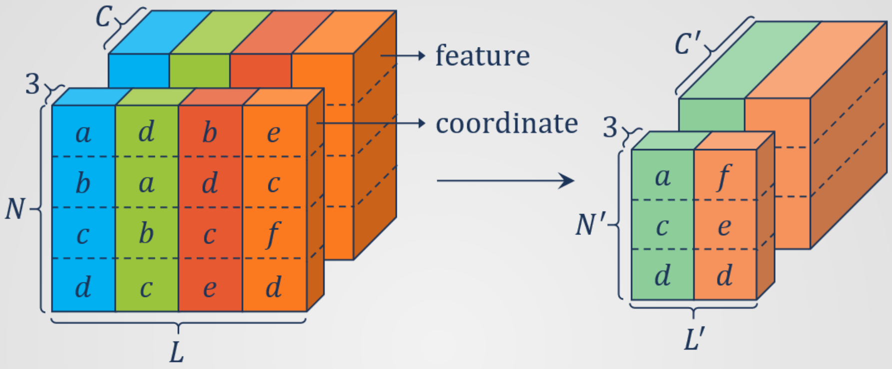
3D Point Cloud Sequence/Video

NUS National University of Singapore

Conventional Grid-based Video vs. Point Cloud Video

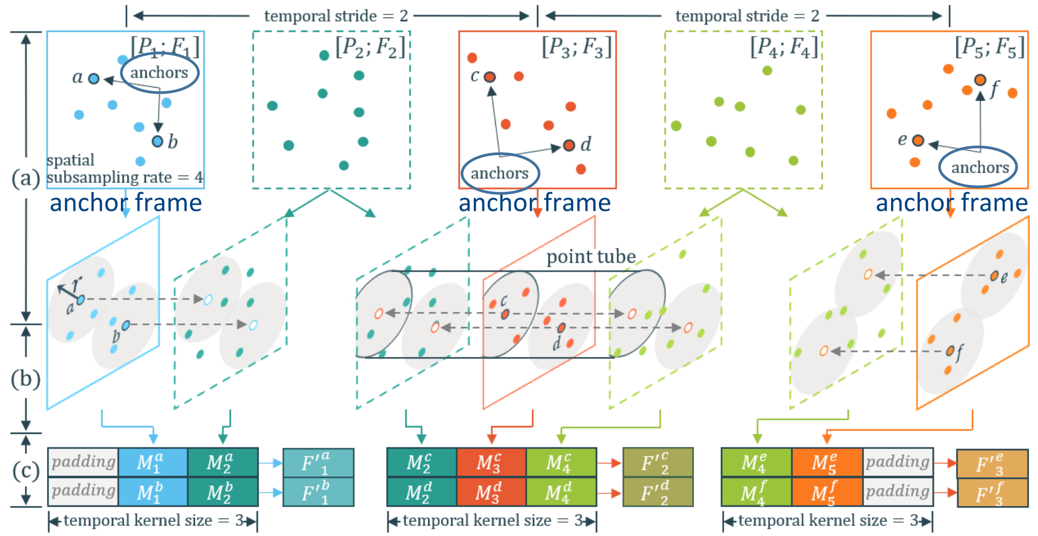


Challenges in Point Cloud Video Modeling



- Points emerge inconsistently across different frames.
  - Tracking points (✗).
  - Spatio-temporal hierarchy (✓).
- Point cloud videos are **spatially** unordered and irregular but **temporally** ordered and regular.
  - Decompose the spatio-temporal modeling in point cloud videos.

PST Convolution & PSTNet



PST Convolution:

a) Point tube construction

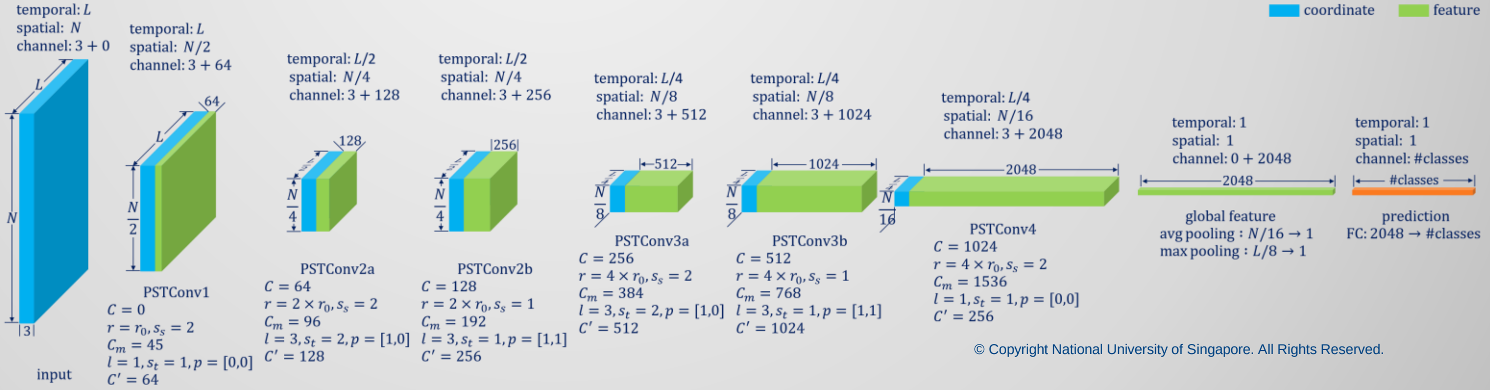
b) Spatial Convolution

$$M_t^{(x,y,z)} = \sum_{\|(\delta_x,\delta_y,\delta_z)\|\leq r} \mathbf{S}^{(\delta_x,\delta_y,\delta_z)} \cdot F_t^{(x+\delta_x,y+\delta_y,z+\delta_z)}$$

c) Temporal Convolution

$$F'^{(x,y,z)}_t = \sum_{k=-\lfloor l/2 \rfloor}^{\lfloor l/2 \rfloor} \mathbf{T}_k \cdot M_{t+k}^{(x,y,z)}$$

Spatio-temporal Hierarchical PSTNet



National University of Singapore

Experiments

3D Action Recognition

Table 2: Action recognition accuracy (%) on the NTU RGB+D 60 and NTU RGB+D 120 datasets.

Method	Input	NTU RGB+D 60 Subject	NTU RGB+D 60 View	NTU RGB+D 120 Subject	NTU RGB+D 120 Setup
SkeleMotion (Caetano et al., 2019)	skeleton	69.6	80.1	67.7	66.9
GCA-LSTM (Liu et al., 2017)	skeleton	74.4	82.8	58.3	59.3
FSNet (Liu et al., 2019b)	skeleton	-	-	59.9	62.4
Two Stream Attention LSTM (Liu et al., 2018)	skeleton	77.1	85.1	61.2	63.3
Body Pose Evolution Map (Liu & Yuan, 2018)	skeleton	-	-	64.6	66.9
AGC-LSTM (Si et al., 2019)	skeleton	89.2	95.0	-	-
AS-GCN (Li et al., 2019)	skeleton	86.8	94.2	-	-
VA-fusion (Zhang et al., 2019)	skeleton	89.4	95.0	-	-
2s-AGCN (Shi et al., 2019b)	skeleton	88.5	95.1	-	-
DGNN (Shi et al., 2019a)	skeleton	89.9	96.1	-	-
HON4D (Oreifej & Liu, 2013)	depth	30.6	7.3	-	-
SNV (Yang & Tian, 2014)	depth	31.8	13.6	-	-
HOG² (Ohn-Bar & Trivedi, 2013)	depth	32.2	22.3	-	-
Li et al. (2018a)	depth	68.1	83.4	-	-
Wang et al. (2018a)	depth	87.1	84.2	-	-
MVDI (Xiao et al., 2019)	depth	84.6	87.3	-	-
NTU RGB+D 120 Baseline (Liu et al., 2019a)	depth	-	-	48.7	40.1
PointNet++ (appearance) (Qi et al., 2017b)	point	80.1	85.1	72.1	79.4
3DV (motion) (Wang et al., 2020)	voxel	84.5	95.4	76.9	92.5
3DV-PointNet++ (Wang et al., 2020)	voxel + point	88.8	96.3	82.4	93.5
PSTNet (ours)	point	**90.5**	**96.5**	**87.0**	**93.8**

4D Semantic Segmentation

Table 3: Semantic segmentation results on the Synthia 4D dataset.

Method	Input	#Frms	#Params (M)	mIoU (%)
3D MinkNet14 (Choy et al., 2019)	voxel	1	19.31	76.24
4D MinkNet14 (Choy et al., 2019)	voxel	3	23.72	77.46
PointNet++ (Qi et al., 2017b)	point	1	0.88	79.35
MeteorNet (Liu et al., 2019e)	point	3	1.78	81.80
PSTNet ($l = 1$)	point	3	1.42	80.79
PSTNet ($l = 3$)	point	3	1.67	**82.24**