Jane 15th, 2021

Human Segmentation with Dynamic LiDAR Data

Tao Zhong, Wonjik Kim, Masayuki Tanaka and Masatoshi Okutomi

Tokyo Institute of Technology

Problem

Segmenting humans in the dynamic LiDAR Data with the help of motion information

Spatial information

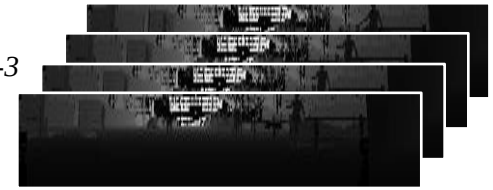
Single frame

Deep Neural Network

Human segmentation

Frame *t*

Frame *t-3*



Sequential frames

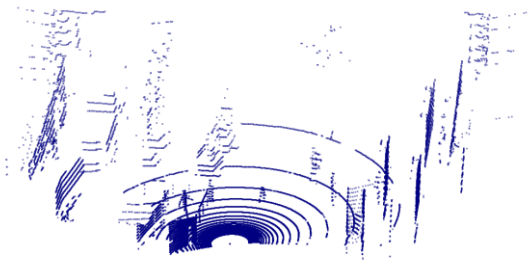
Deep Neural Network

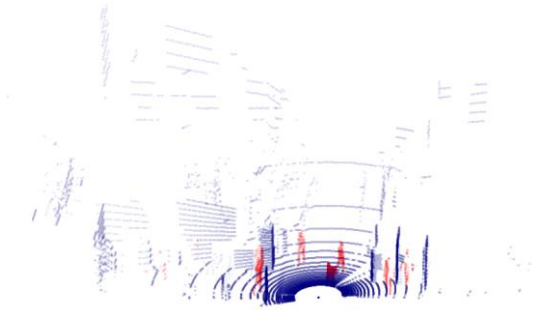
Human segmentation

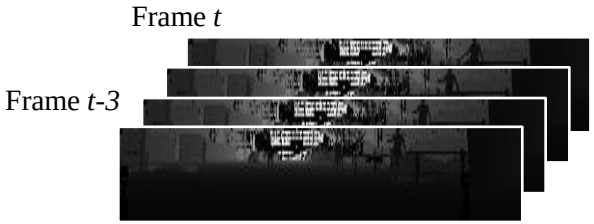
Spatial information
+ Temporal information

Objective

How to extract and leverage temporal features effectively







Sequential frames

Deep Neural Network

Human segmentation

Spatial information
+ Temporal information

Related Works

- **Static (Single frame)**
  - SqueezeSeg [1]
- **Dynamic (Sequential frames)**
  - Kim [2]
  - Meteornet [3]

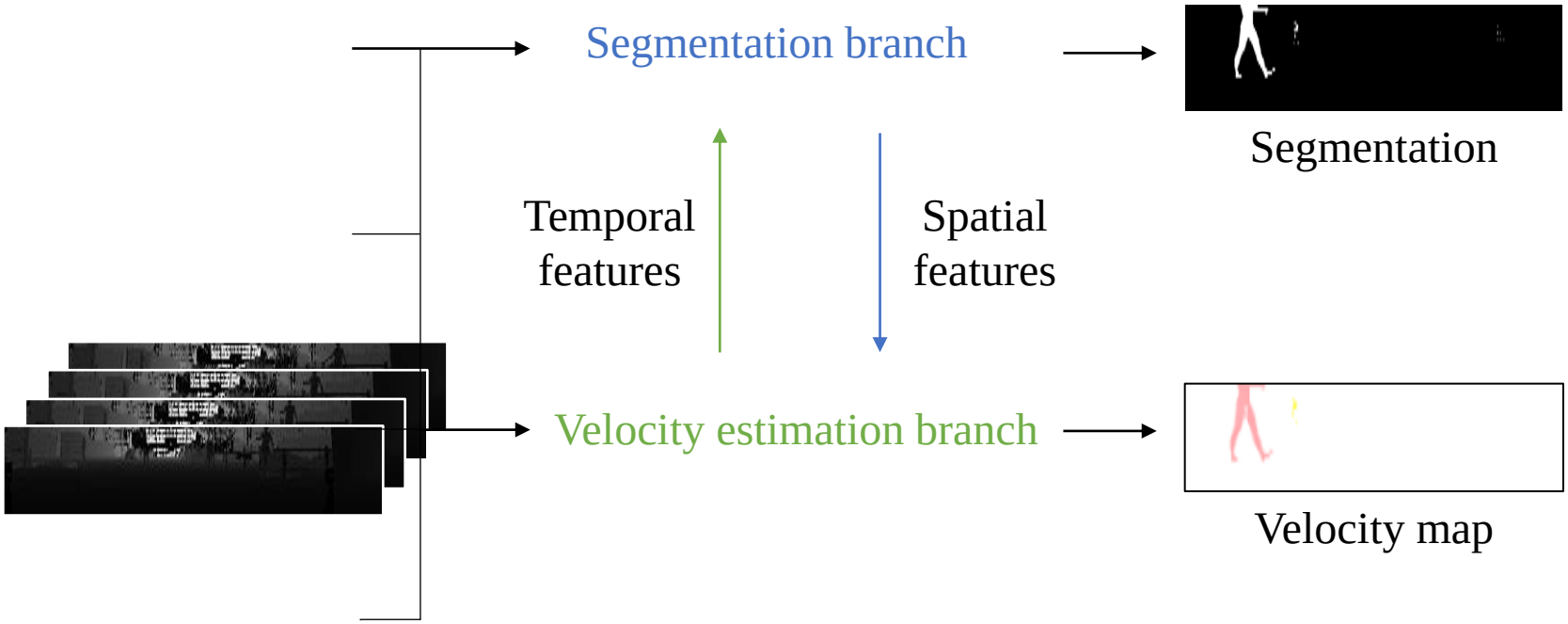
The temporal information were only implicitly used.

[1] B. Wu, A. Wan, X. Yue, and K. Keutzer, “Squeezeseg: Convolutionalneural nets with recurrent crf for real-time road-object segmentationfrom 3d lidar point cloud,” in2018 IEEE International Conference onRobotics and Automation (ICRA). IEEE, 2018, pp. 1887–1893
[2] W. Kim, M. Tanaka, M. Okutomi, and Y. Sasaki, “Learning-based human segmentation and velocity estimation using automatic labeled lidar sequence for training,” IEEE Access, vol. 8, pp. 88 443–88 452, 2020.
[3] Liu X, Yan M, Bohg J. MeteorNet: Deep learning on dynamic 3D point cloud sequences[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 9246-9255.

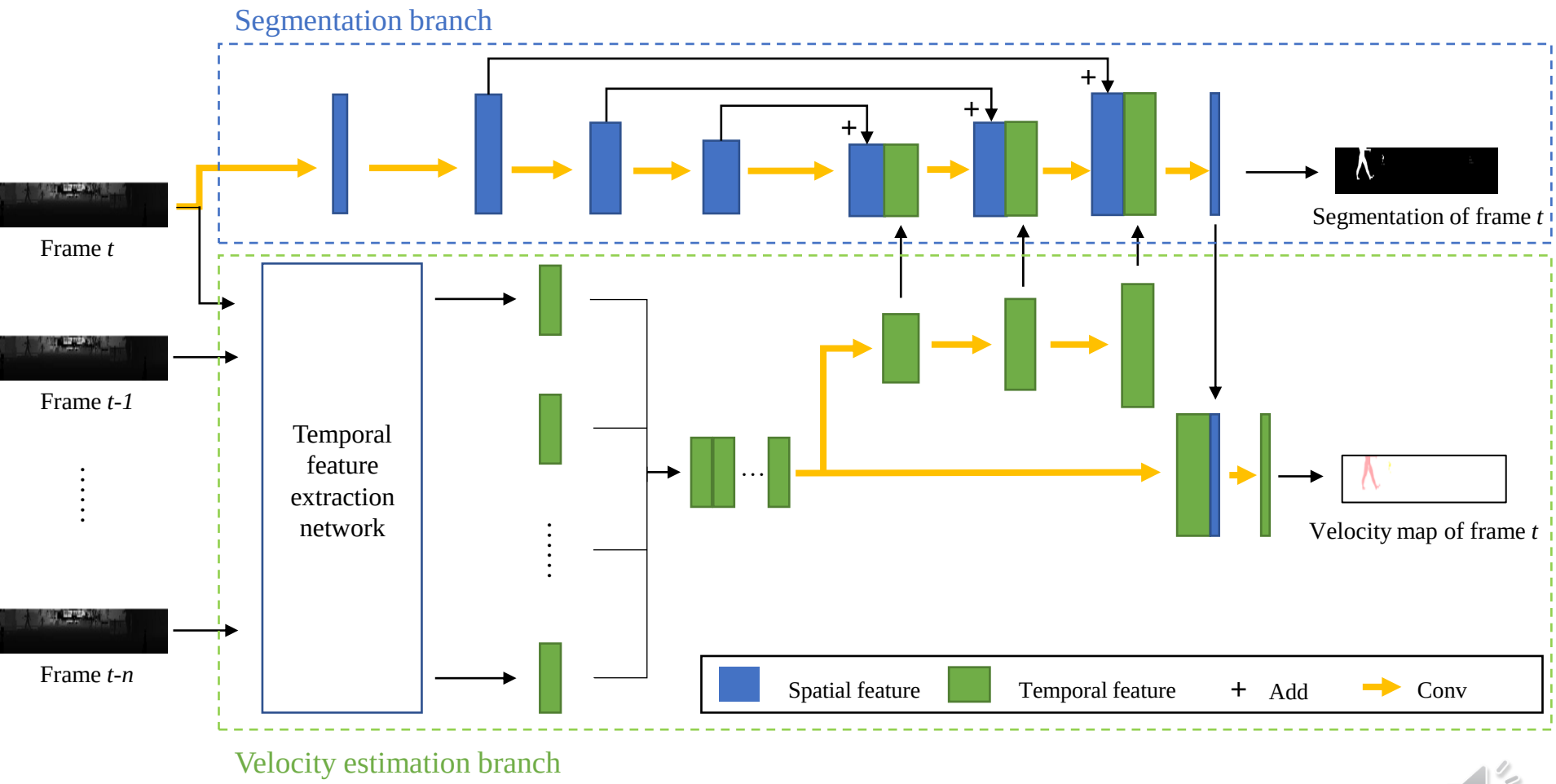
Proposed Network Architecture

Joint learning of segmentation and velocity estimation

Extract temporal features explicitly and leverage it in the segmentation task



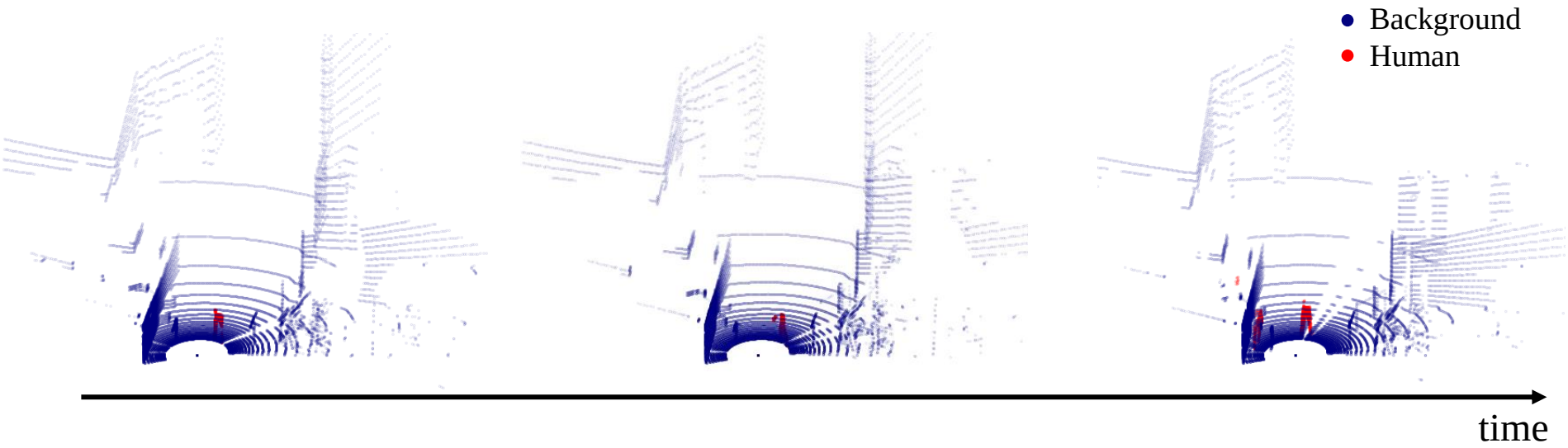
Proposed Network Architecture



Data Generation[2]

Real background LiDAR data + Human walking model

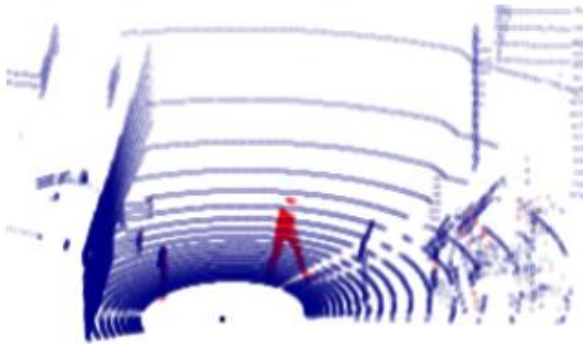
Synthesis



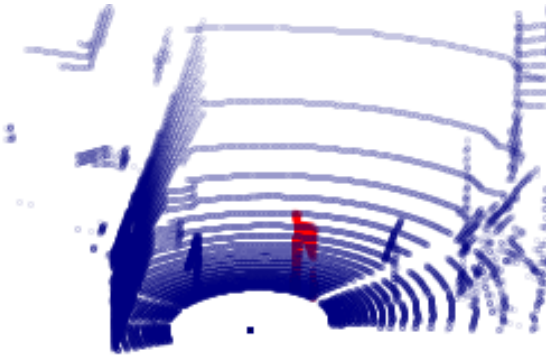
- LiDAR scanning frequency is 10 Hz.
- Each sequence consists of 32 frames.
- Each point is annotated with class and velocity information.

Segmentation Results on Generated Data

- Networks are trained with 900 generated sequences.
- The length of the input sequence is 4 frame except for SqeezeSeg.

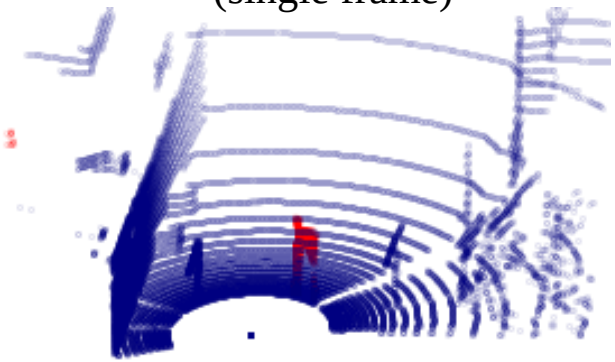


SqueezeSeg[1]
(single frame)

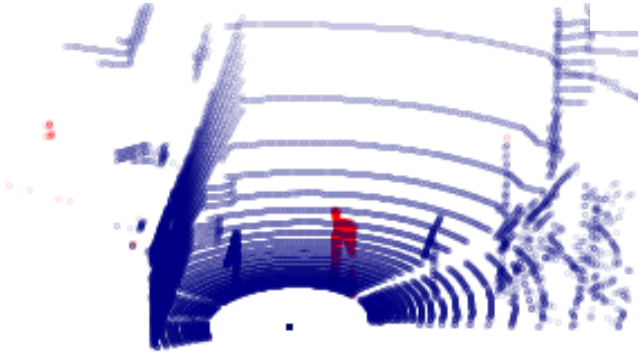


Kim[2]

Background
Human



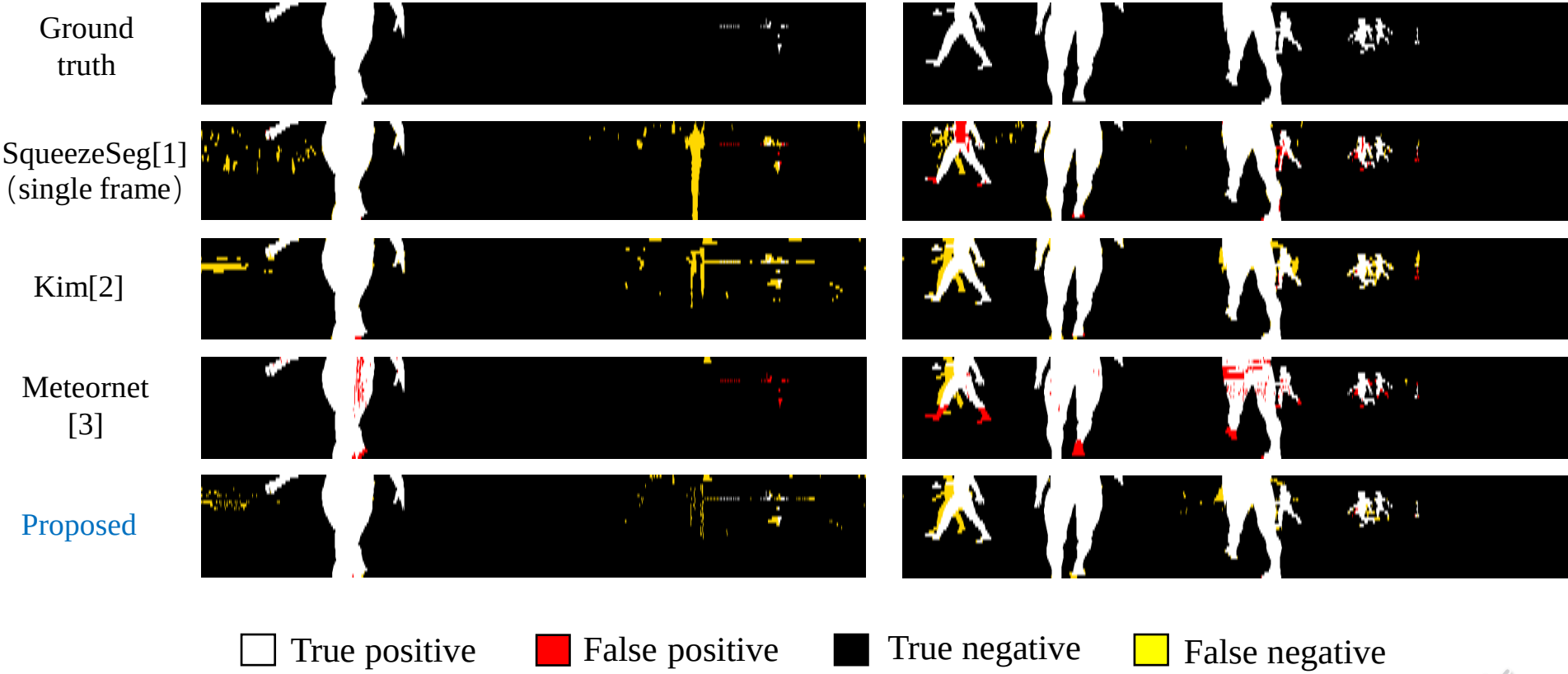
Meteornet[3]



Proposed

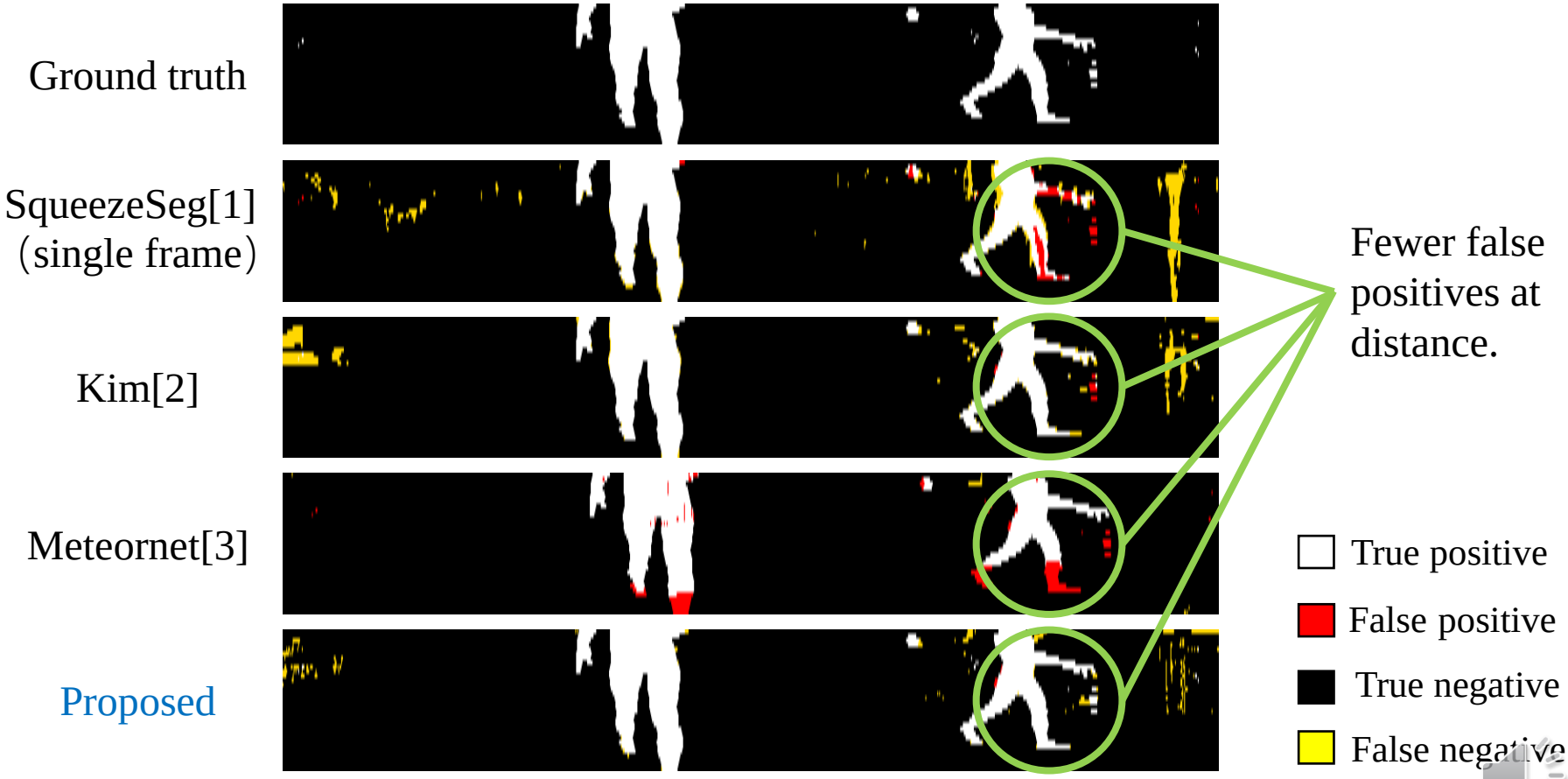
Segmentation Results on Generated Data

- Comparison



Segmentation Results on Generated Data

- Comparison



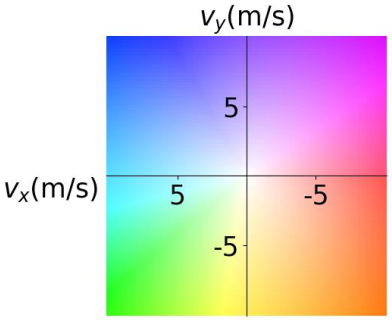
Quantitative Comparison on Generated and Real Data

Method		SqueezeSeg[1] (w/o CRF)	Kim[2]	Meteornet[3]	Proposed
Input		depth	depth	xyz	depth
Num of frames		1	4	4	4
Pedestrian mIoU (%)	Generated data	58.87	68.88	73.87	**86.08**
	Real data	11.35	58.69	52.20	**67.72**
Run time (ms)		6	46	840	**51**

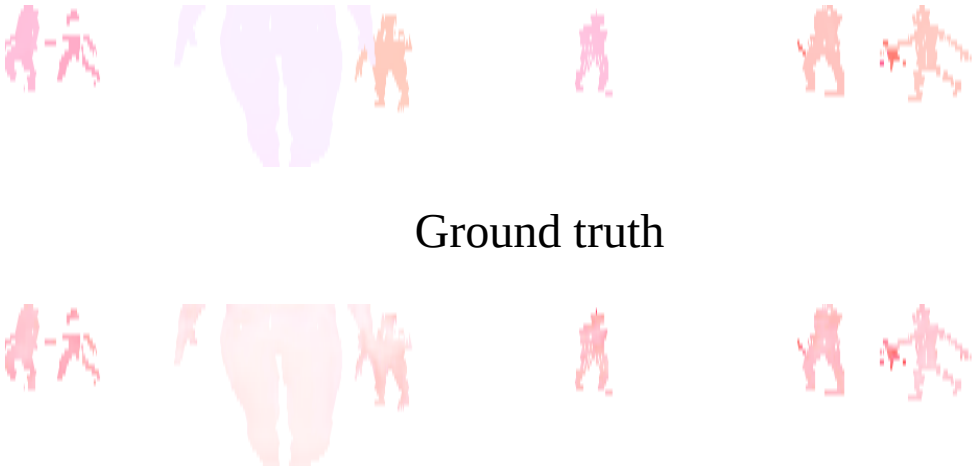
- Proposed network outperforms previous static and dynamic method.

Effect of the velocity estimation

- velocity map of human area



Colormap



Ground truth

Prediction

- The tendency of estimated velocities is similar to that of ground truth.

Effect of the velocity estimation

- Ablation study

	Network		Proposed		
	Number of frames		4		
	Velocity estimation loss			√	√
	Temporal feature propagation		√		√
Generated data	Distance (m)	0 to ∞	81.16	77.11	**86.08**
		0 to 4	92.21	90.98	**94.06**
		4 to 8	63.25	59.07	**72.79**
		8 to ∞	33.65	25.75	**41.44**
Real data	Distance (m)	0 to ∞	64.23	64.87	**67.29**
		0 to 4	74.59	75.73	**76.93**
		4 to 8	49.84	51.83	**53.79**
		8 to ∞	23.43	**25.06**	24.76

- Motion cues help detection in the distance.

Effect of the number of the input frames

	Network		Proposed				
	Number of frames		1	2	4	8	16
Generated data	Distance (m)	0 to ∞	82.01	83.00	86.08	**88.89**	80.73
		0 to 4	91.75	93.17	94.06	**95.10**	90.61
		4 to 8	65.02	67.51	72.79	**78.96**	78.96
		8 to ∞	38.05	34.52	41.44	**45.52**	31.96
Real data	Distance (m)	0 to ∞	65.74	66.22	67.29	**67.31**	63.60
		0 to 4	75.85	77.20	76.93	**77.43**	74.28
		4 to 8	52.50	52.77	**53.79**	51.53	46.01
		8 to ∞	23.39	23.57	**24.76**	22.35	23.61

- When the length is not larger than eight frames, overall, accuracy increases along with the increase in the number of frames

Conclusion

- Proposed a two-branch network for dynamic point cloud segmentation, which achieves high accuracy on a data set for human segmentation.
- Temporal information contained in sequential data is beneficial to segmentation because motion cues can compensate for the sparseness of point cloud.
- An increase in the length of sequence improves the performance by producing more motion cues, but there is a trade-off between accuracy and computation cost.