A survey on Adversarial Networks for improving Pseudo-LiDAR

Sourav Sahoo
Subhankar Chakraborty

3D Object Detection

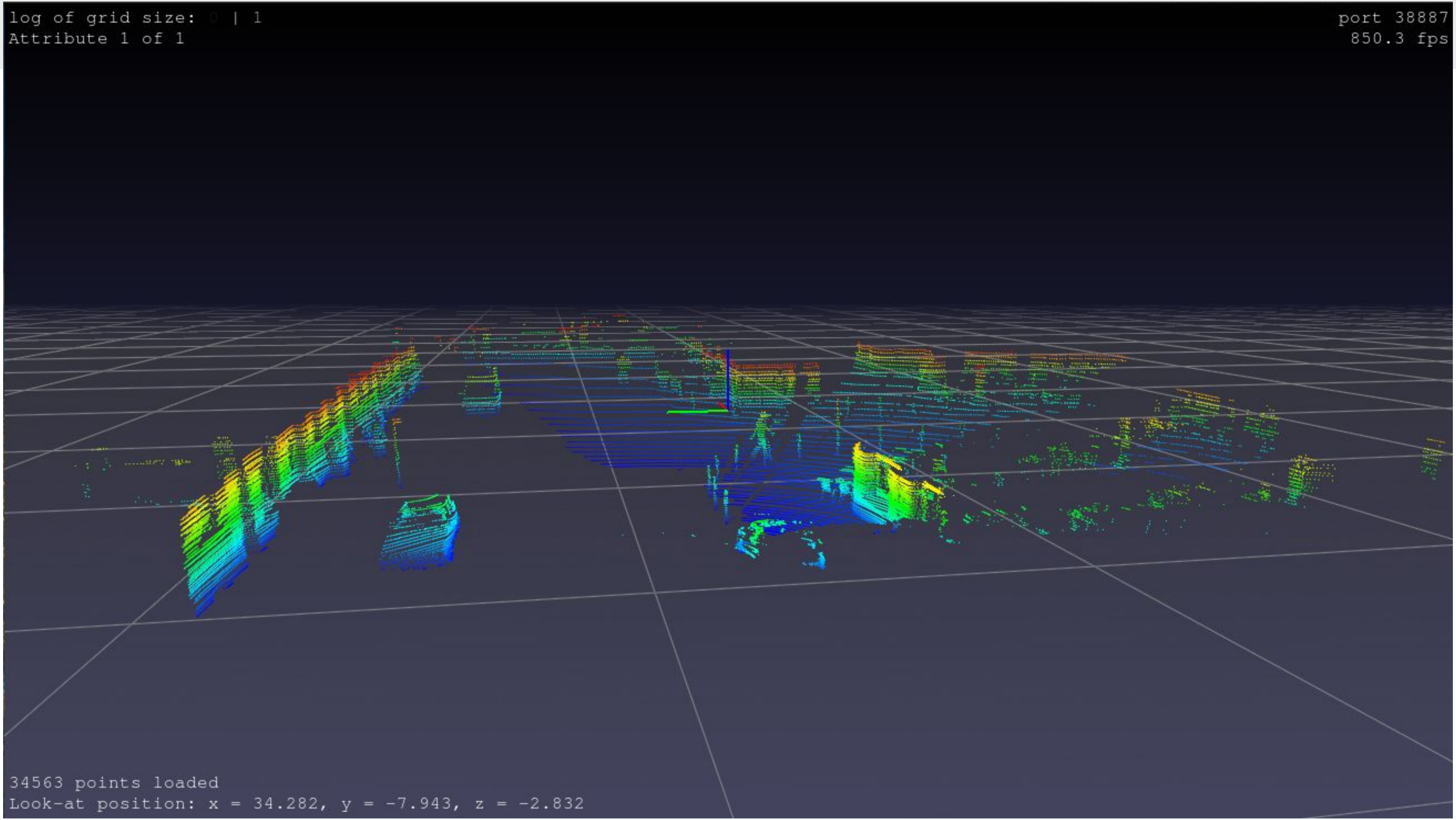
- LiDAR based
- Stereo Image Based
- Monocular Image Based

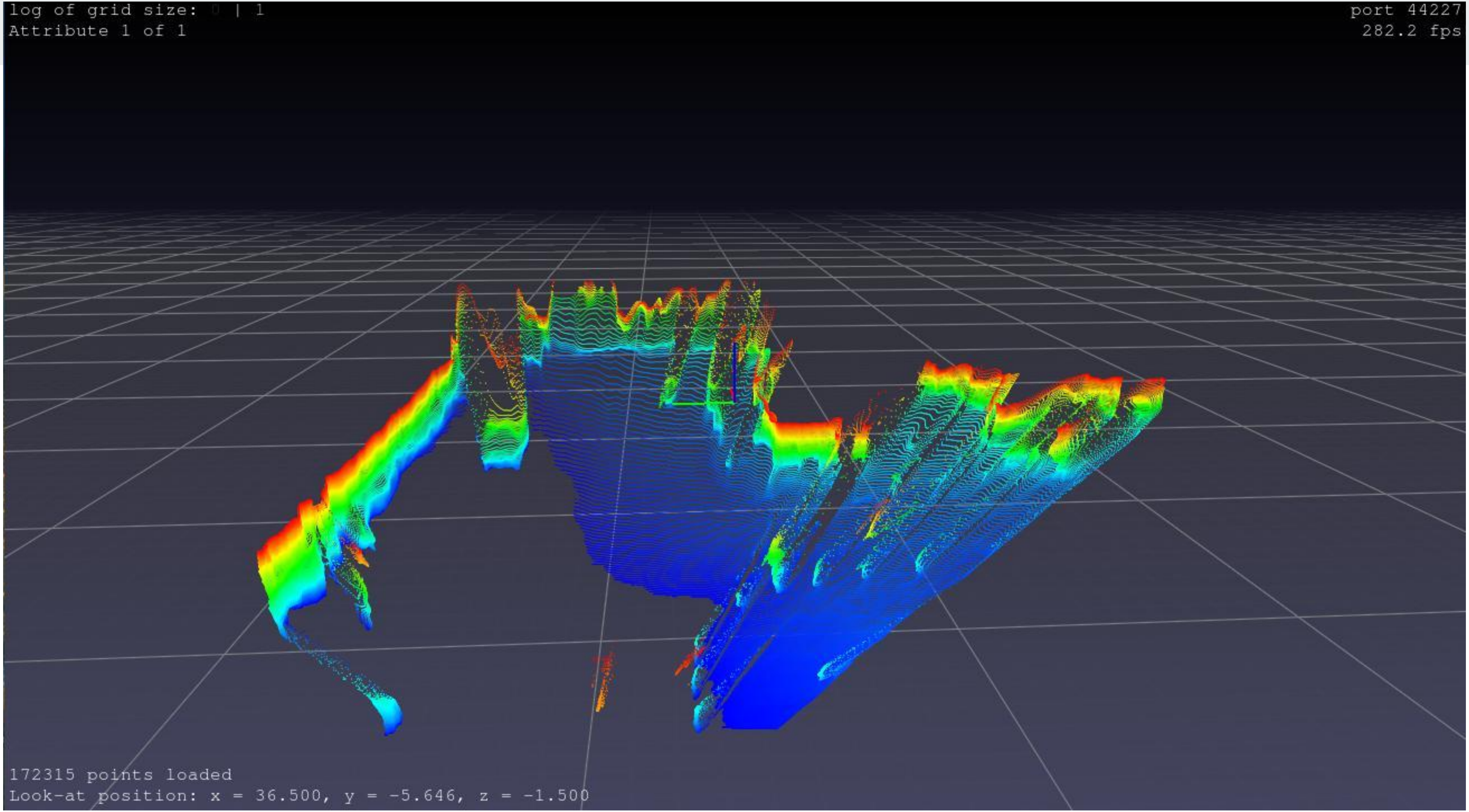
Current Issues

- LiDAR based algorithms perform the best.
- Good quality LiDARs used for this purpose are very expensive. They cost around $75000 (~50L INR).
- Stereo and Monocular images are much cheaper to acquire but their performance is quite inferior.

Pseudo-LiDAR

- Simultaneously proposed by Wang et. al [1] and Weng and Kitani [2]
- It is a LiDAR like representation that is generated from the depth map of a stereo or monocular image.

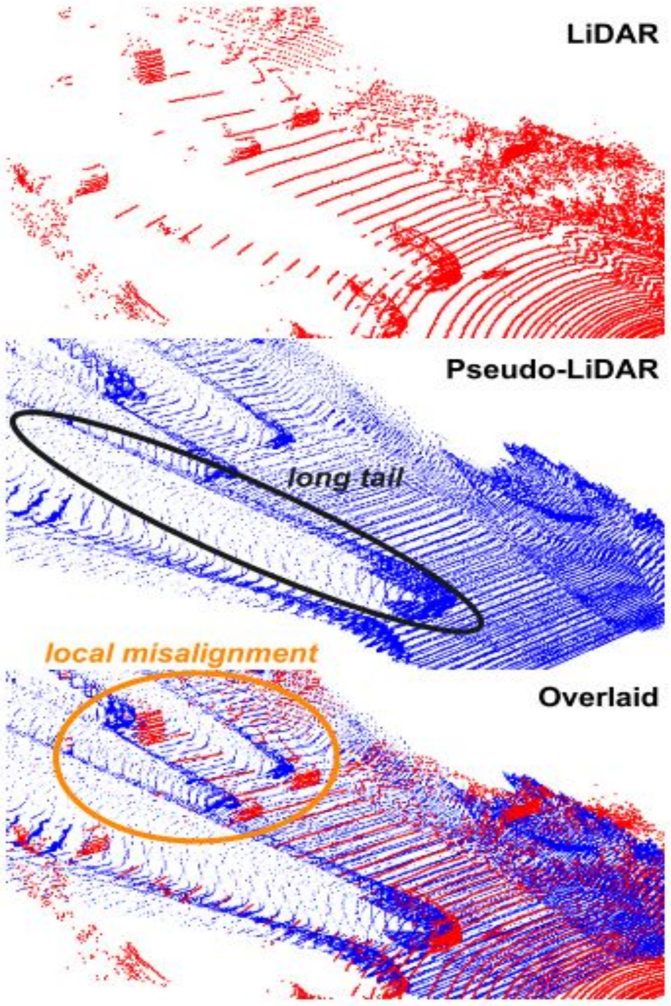

log of grid size: 0 | 1
Attribute 1 of 1
port 38887
850.3 fps
34563 points loaded
Look-at position: x = 34.282, y = -7.943, z = -2.832


log of grid size: 0 | 1
Attribute 1 of 1
port 44227
282.2 fps
172315 points loaded
Look-at position: x = 36.500, y = -5.646, z = -1.500

Detection algorithm	Input signal	IoU = 0.5			IoU = 0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MONO3D [4]	Mono	30.5 / 25.2	22.4 / 18.2	19.2 / 15.5	5.2 / 2.5	5.2 / 2.3	4.1 / 2.3
MLF-MONO [33]	Mono	55.0 / 47.9	36.7 / 29.5	31.3 / 26.4	22.0 / 10.5	13.6 / 5.7	11.6 / 5.4
AVOD	Mono	61.2 / 57.0	45.4 / 42.8	38.3 / 36.3	33.7 / 19.5	24.6 / 17.2	20.1 / 16.2
F-POINTNET	Mono	70.8 / 66.3	49.4 / 42.3	42.7 / 38.5	40.6 / 28.2	26.3 / 18.5	22.9 / 16.4
3DOP [5]	Stereo	55.0 / 46.0	41.3 / 34.6	34.6 / 30.1	12.6 / 6.6	9.5 / 5.1	7.6 / 4.1
MLF-STEREO [33]	Stereo	-	53.7 / 47.4	-	-	19.5 / 9.8	-
AVOD	Stereo	89.0 / 88.5	77.5 / 76.4	68.7 / 61.2	74.9 / 61.9	56.8 / 45.3	49.0 / 39.0
F-POINTNET	Stereo	89.8 / 89.5	77.6 / 75.5	68.2 / 66.3	72.8 / 59.4	51.8 / 39.8	44.0 / 33.5
AVOD [17]	LiDAR + Mono	90.5 / 90.5	89.4 / 89.2	88.5 / 88.2	89.4 / 82.8	86.5 / 73.5	79.3 / 67.1
F-POINTNET [25]	LiDAR + Mono	96.2 / 96.1	89.7 / 89.3	86.8 / 86.2	88.1 / 82.6	82.2 / 68.8	74.0 / 62.0

Drawbacks of Pseudo-LiDAR

- **Local Misalignment:** The extracted point cloud frustum could be largely off from its original location
- **Long Tail artifact:** Depth artifacts around the periphery of the detected object form a tail like structure.



Actual LiDAR representation of a scene

Long tail artifact in Pseudo-LiDAR representation

Local Misalignment

Image Courtesy: Xinshuo Weng and Kris Kitani.Monocular 3d object detection with pseudo-lidar point cloud. arXiv preprint arXiv:1903.09847, 2019

Formalizing the Problem

- 3D Object detection works best on LiDAR point clouds than stereo/monocular images.
- Pseudo-LiDAR aims to bridge the gap.
- Want to improve the Pseudo-LiDAR generation process.

Our Approach

- Difficult to generate a complete representation from scratch.
- We feel that a prior is required for the generation process.
- The existing Pseudo-LiDAR representation is used as a prior.

Challenges in this approach

- Number of points in the actual LiDAR and Pseudo-LiDAR quite different.
- The number of LiDAR data points changes with the scene.
- No existing literature to directly synthesize 3D point clouds accurately given a RGB image.

Bird’s Eye View (BEV) Maps

- This is a 2D representation of the 3D point cloud.
- Like RGB images, contains 3 channels: Height, Density and Reflectance.
- The actual implementation is inspired from Chen et. al [3].

Preprocessing

- The X and Y axes are cropped to be between [0,70]m and [-40,40]m respectively.
- The ground plane is discretized into square cells with a resolution of 10 cm.

Height

- We select points with lying within a height of [-1.5, 1]m from the LiDAR plane.
- The height in a bin is taken to be the height of the highest point in that bin.

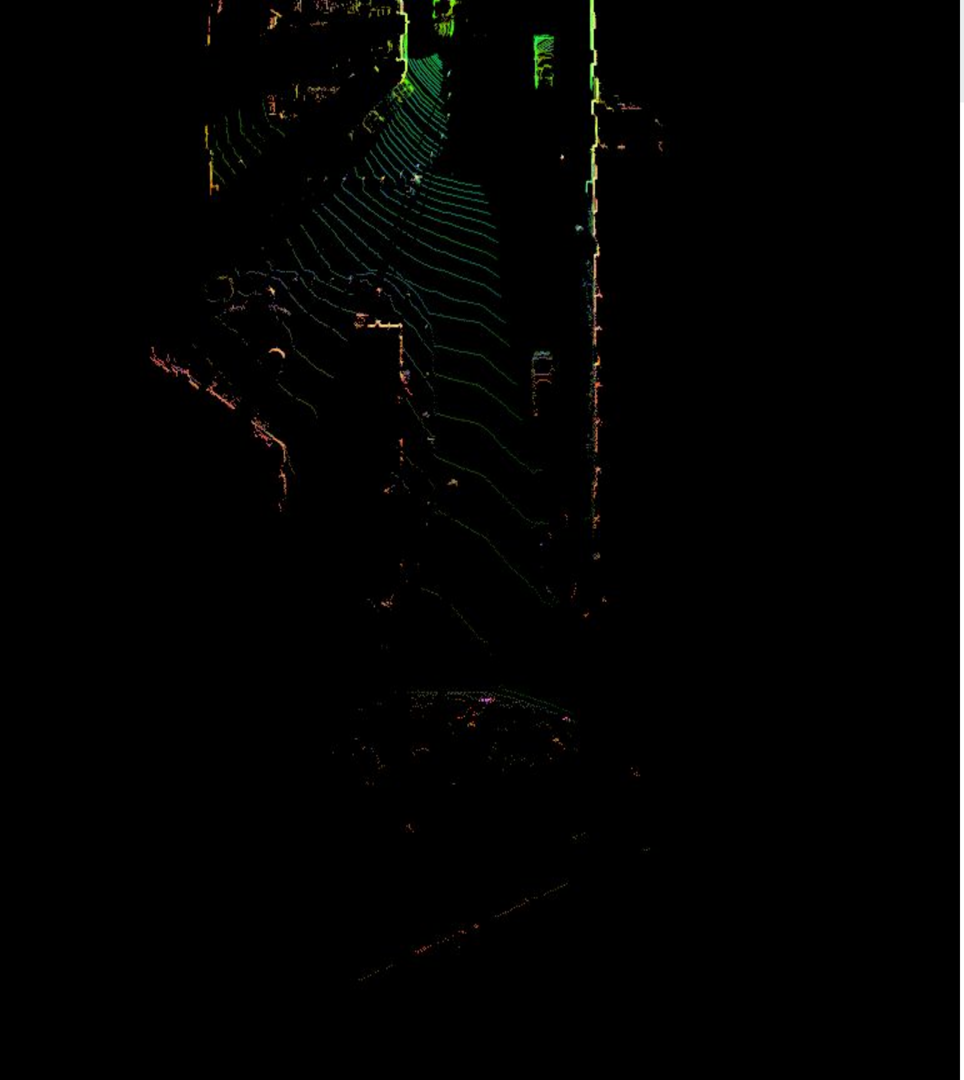
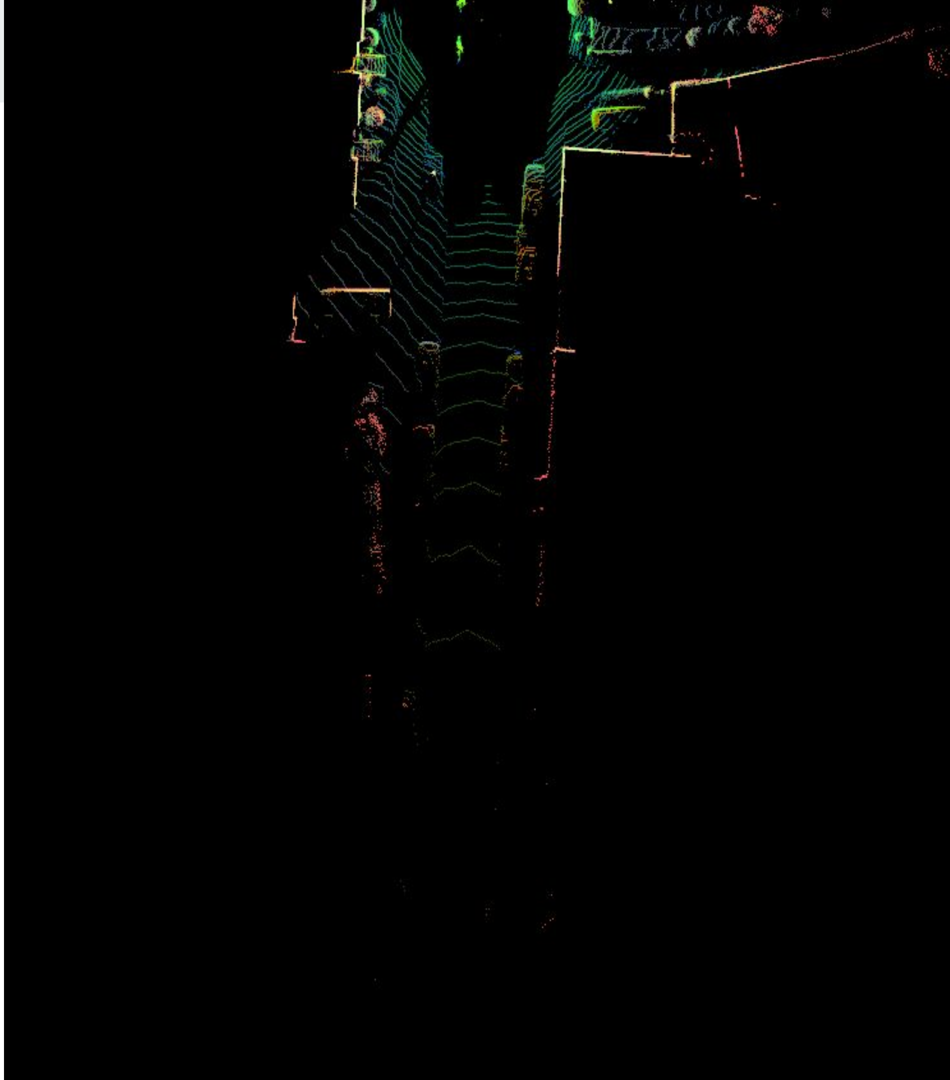
Density

- Density corresponds to the number of points in a bin.
- If there are N points in a cell, the density value assigned is min(1.0, log(N+1)/log(T)).
- T is threshold, that is set to 16 and 64.

Reflectance

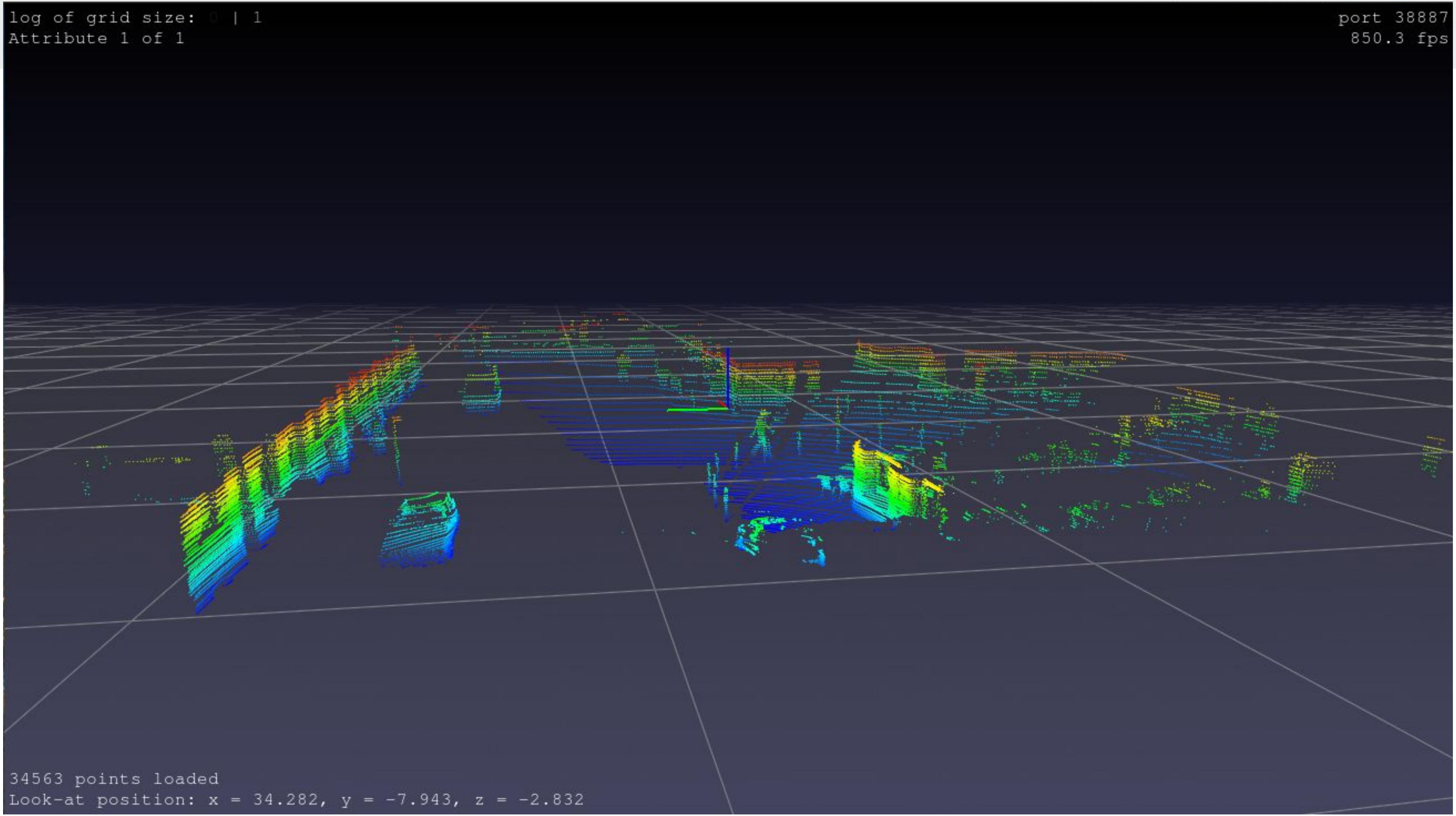
- Reflectance is defined as the measure of the fraction of light or other radiation striking a surface which is reflected off it.
- The reflectance values lie between 0 and 1.

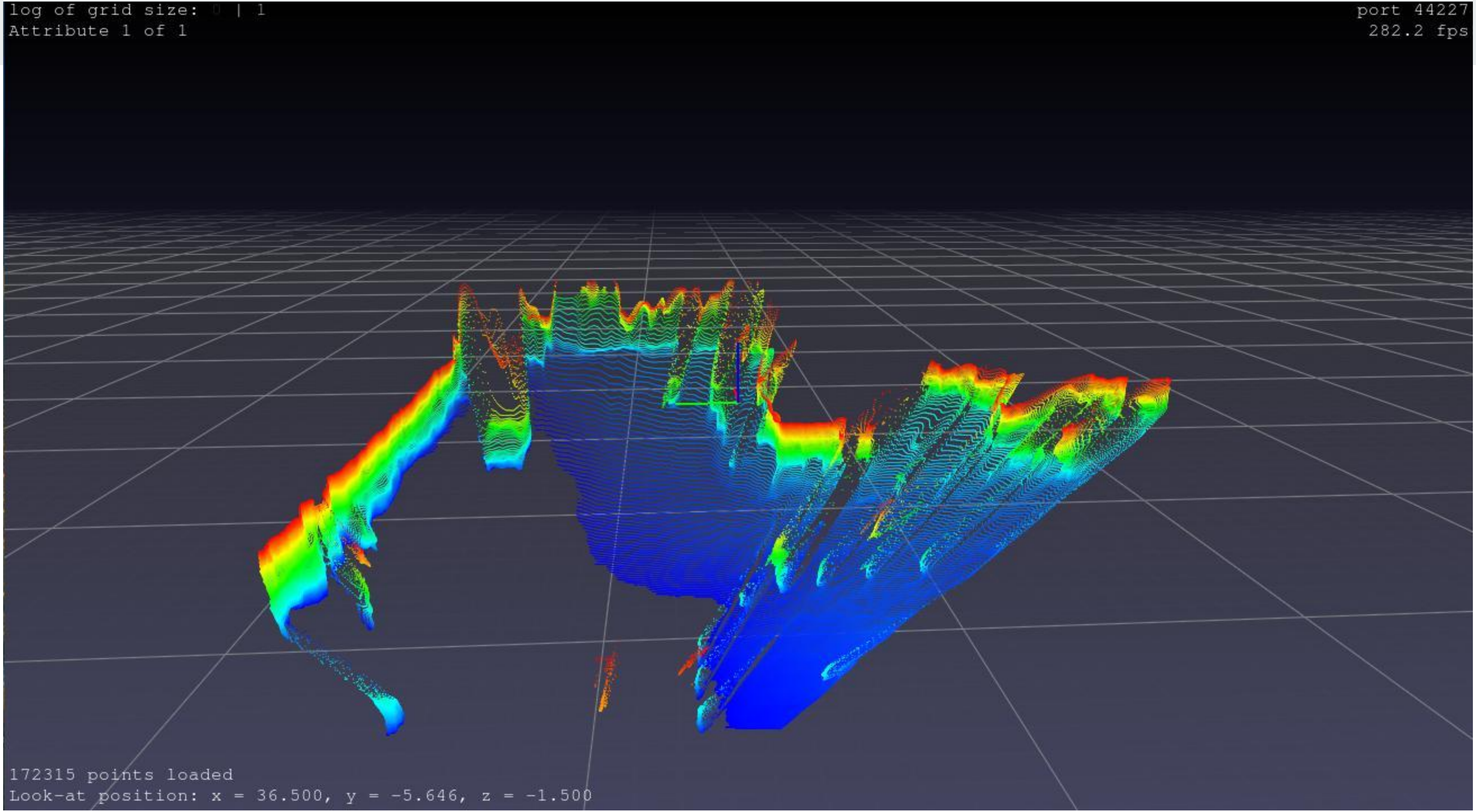
Some Examples of BEV Maps ...

Going from BEV to 3D point clouds

- Necessary as SOTA 3D object detection models work on point clouds.
- This reconstruction is obviously not 100% accurate due to quantization errors in the BEV map generation process.


log of grid size: 0 | 1
Attribute 1 of 1
port 38887
850.3 fps
34563 points loaded
Look-at position: x = 34.282, y = -7.943, z = -2.832


log of grid size: 0 | 1
Attribute 1 of 1
port 44227
282.2 fps
172315 points loaded
Look-at position: x = 36.500, y = -5.646, z = -1.500

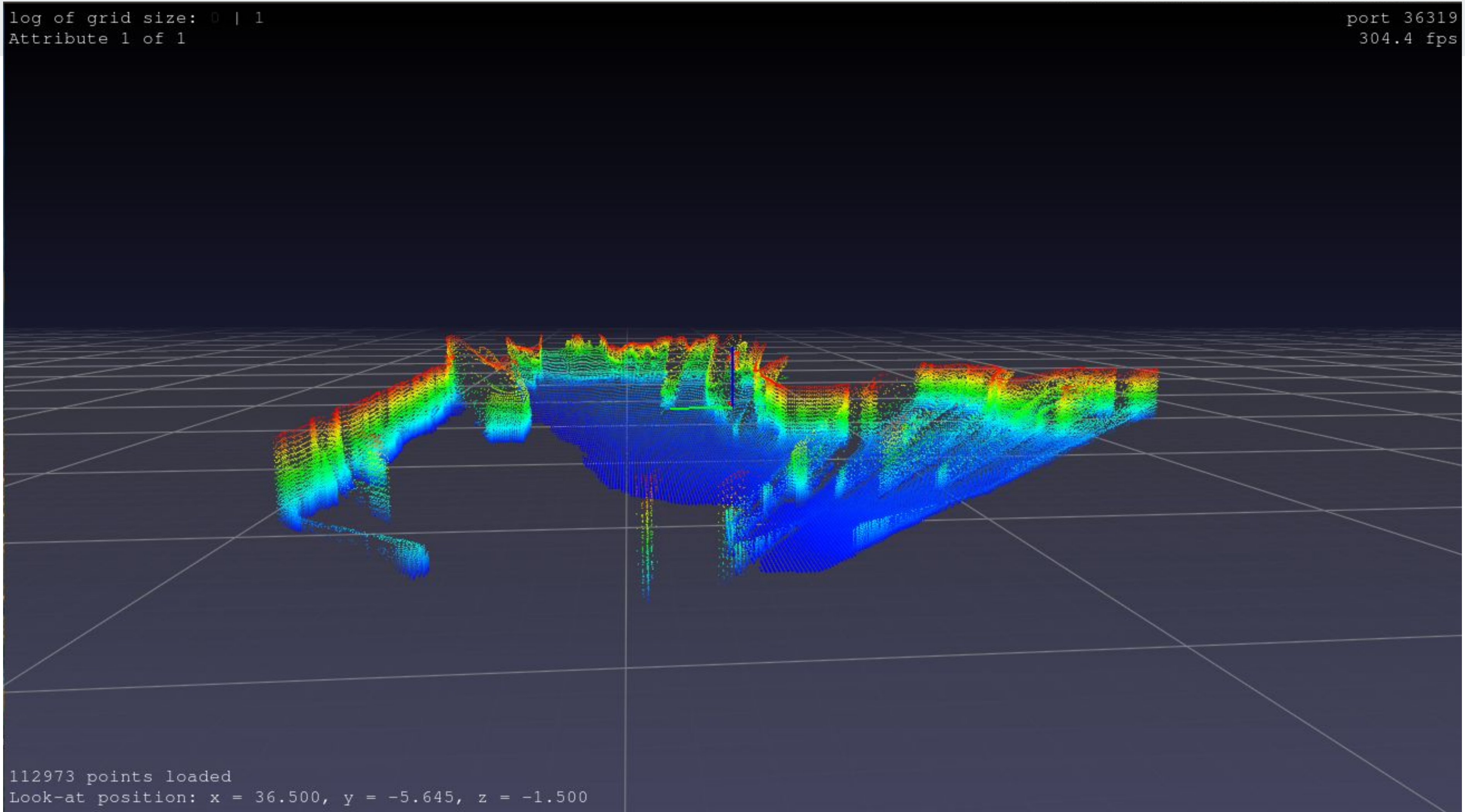

log of grid size: 0 | 1
Attribute 1 of 1
port 36319
304.4 fps
112973 points loaded
Look-at position: x = 36.500, y = -5.645, z = -1.500

Image-to-Image Translation

- This is an Image to Image translation problem where the BEV maps of the pseudo-LiDAR is like a degraded image and the LiDAR representation is the desired one.
- Used pix2pix architecture with our dataset.

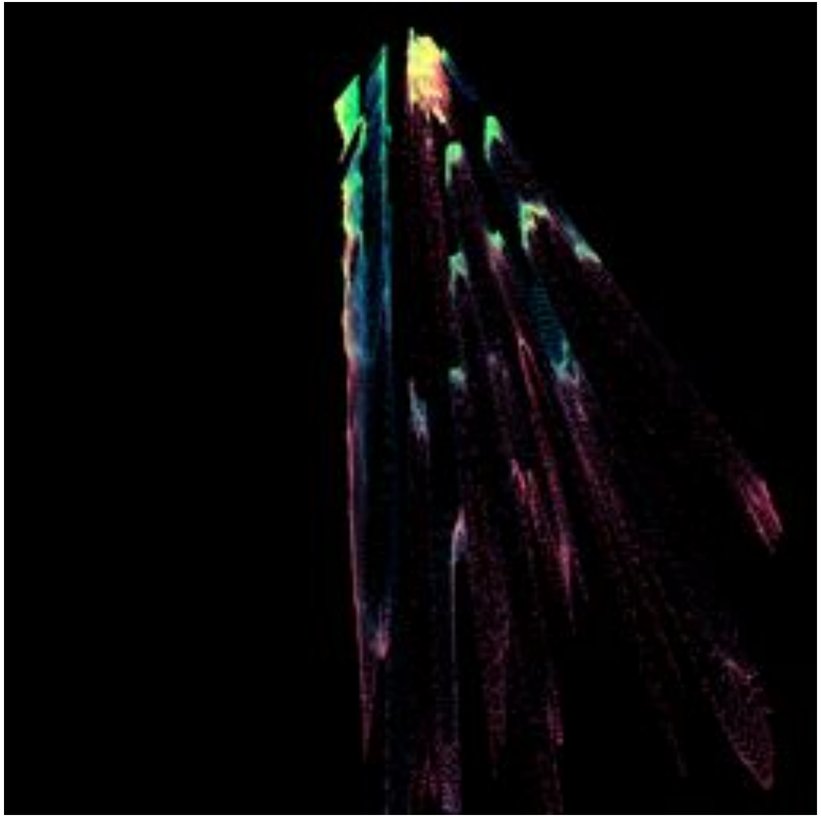
pix2pix

- GAN based Image to Image Translation network.
- Two variants for the Generator
  - Resnet based(~11.5M parameters)
  - UNet based(~54.5M parameters) [We used this network]
- The discriminator is a CNN (~2.8M parameters)

Approach I

- T = 16
- BEV Maps are resized as 256 X 256 patches
- Supervised Method

Actual LiDAR



Pseudo-LiDAR

Actual LiDAR | **Modified Pseudo-LiDAR**

Results

	Easy	Medium	Hard
Car	47.7/28.2	29.7/16.8	24.5/15.7
Pedestrian	22.9/16.5	18.9/14.1	17.1/12.3
Cyclist	18.2/12.1	10.8/8.1	10.7/7.2

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird’s Eye View Map

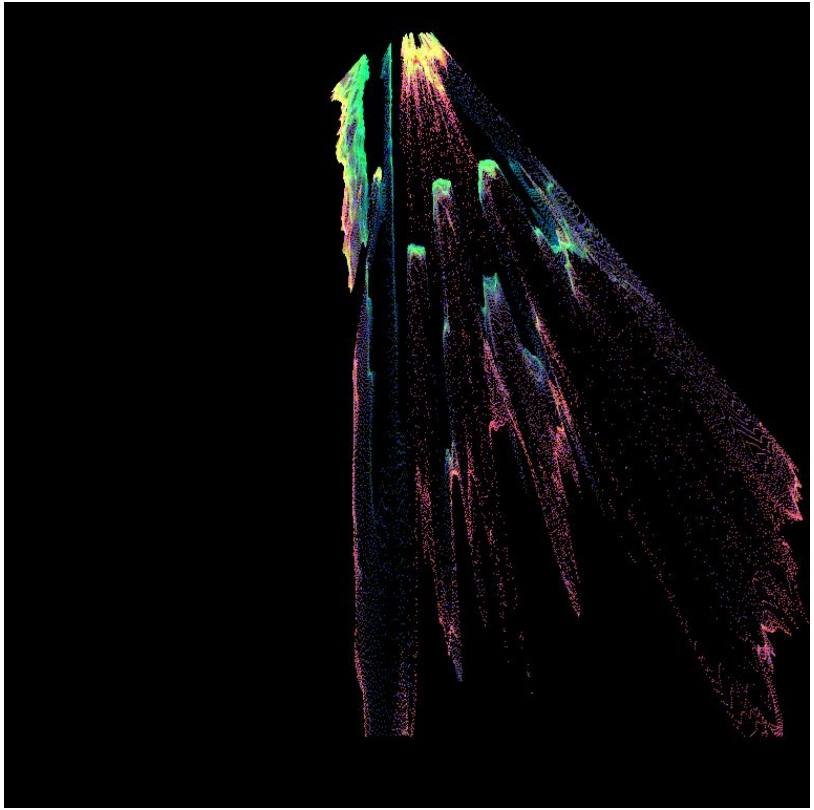
Drawbacks of Approach I

- Works only for 256 X 256 crops.
- Requires the training of an additional super-resolution network.
- Requires a training label for translation task.

Approach II

- T = 16
- 768 X 768 crops of the BEV Maps are used
- Supervised Method


Actual LiDAR


Pseudo-LiDAR

Actual LiDAR

Modified Pseudo-LiDAR

Results

	Easy	Medium	Hard
Car	55.1/34.7	35.4/21.9	29.4/17.7
Pedestrian	17.6/7.7	14.0/6.1	12.4/5.4
Cyclist	20.7/13.8	14.8/11.3	14.1/11.3

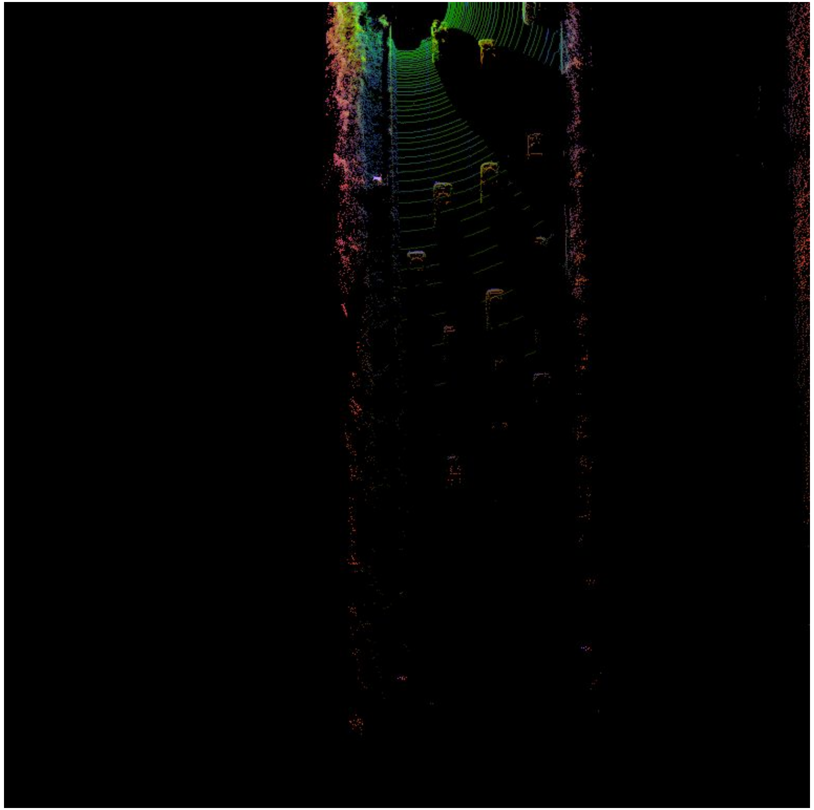
For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird's Eye View Map

Drawbacks of Approach II

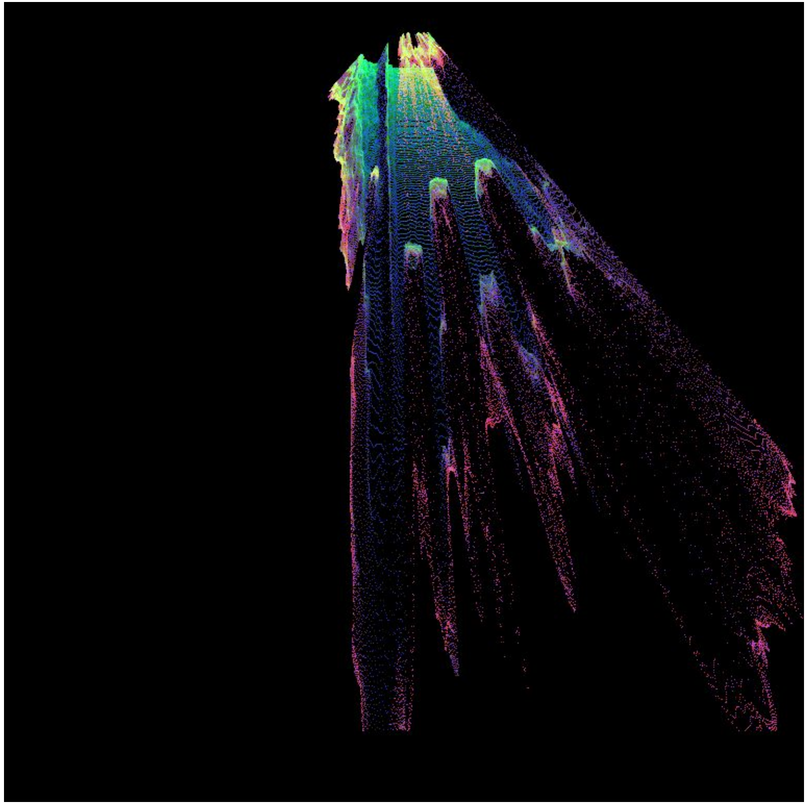
- Fails to generate finer details.
- When T = 16, a very small fraction of the entire point cloud is considered.
- Requires a training label for translation task.

Approach III

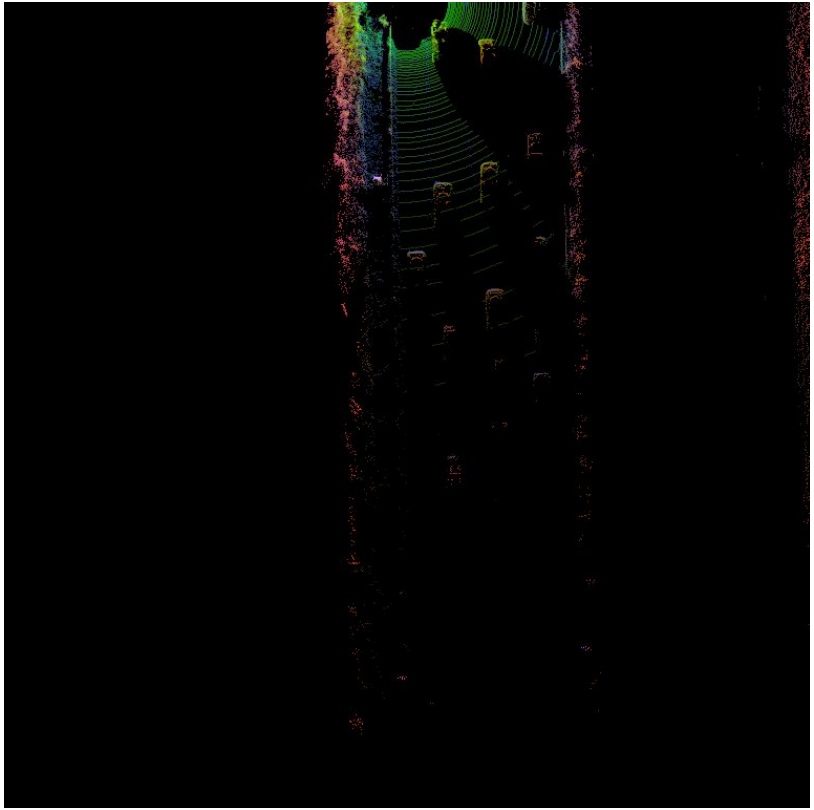
- T = 64 [Earlier it was fixed at 16]
- 768 X 768 crops of the BEV Maps are used
- Supervised Method



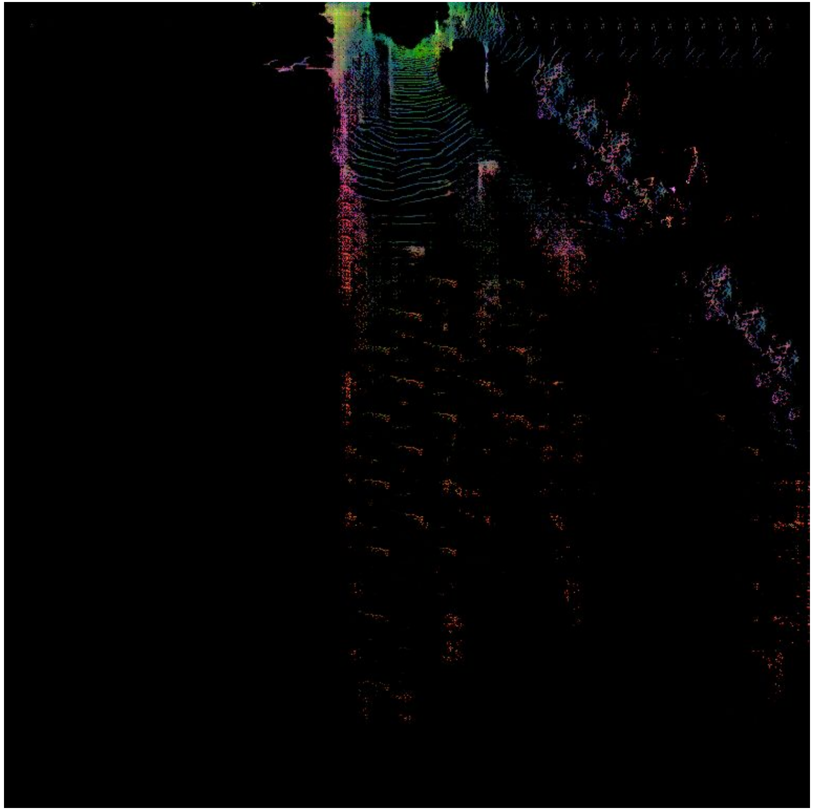
Actual LiDAR



Pseudo-LiDAR



Actual LiDAR



Modified Pseudo-LiDAR

Results

	Easy	Medium	Hard
Car	31.8/15.4	22.5/11.4	20.4/10.3
Pedestrian	8.3/6.5	8.0/5.9	7.4/5.7
Cyclist	8.3/4.4	5.1/2.9	4.6/2.7

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird's Eye View Map

Drawbacks of Approach III

- When T = 64, vanilla pix2pix fails badly due to high resolution and higher number of “active” points as compared to T = 16 (~5x).
- Requires a training label for translation task.

pix2pix HD

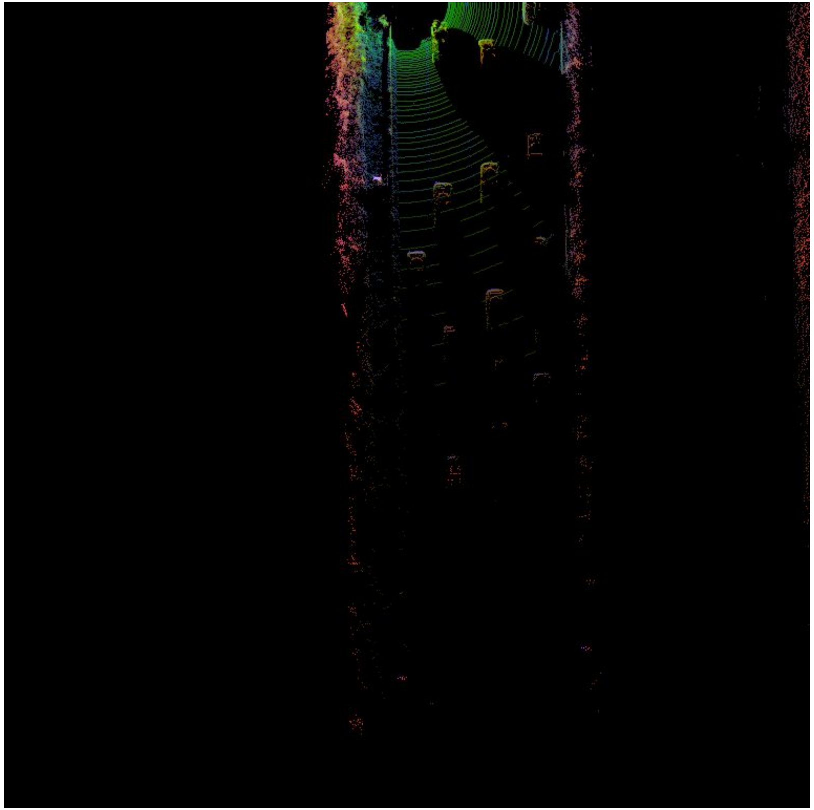
- Pix2pix is ideal for small (256 X 256) images, fails miserably on high resolution images.
- So, NVIDIA came up with pix2pix HD [4], specifically for HR image-to-image translation tasks.

pix2pix HD

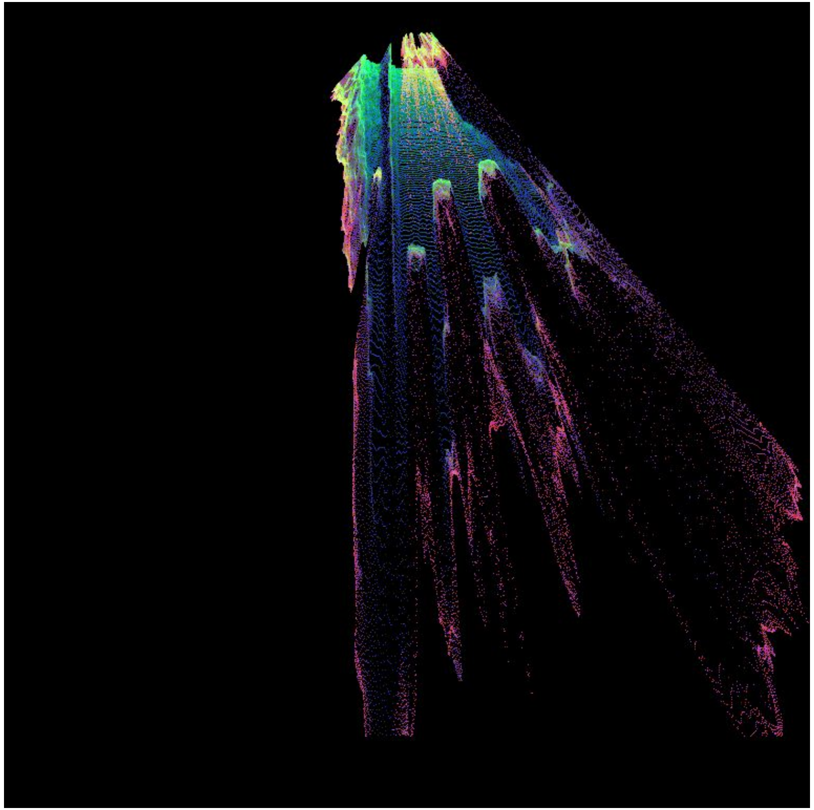
- Two Generators:
  - Global Generator Network
  - Local Enhancer Network
- Three Discriminators:
  - Discriminator for generated image
  - Discriminator for downsampled generated image by factor of 2
  - Discriminator for downsampled generated image by factor of 4

Approach IV

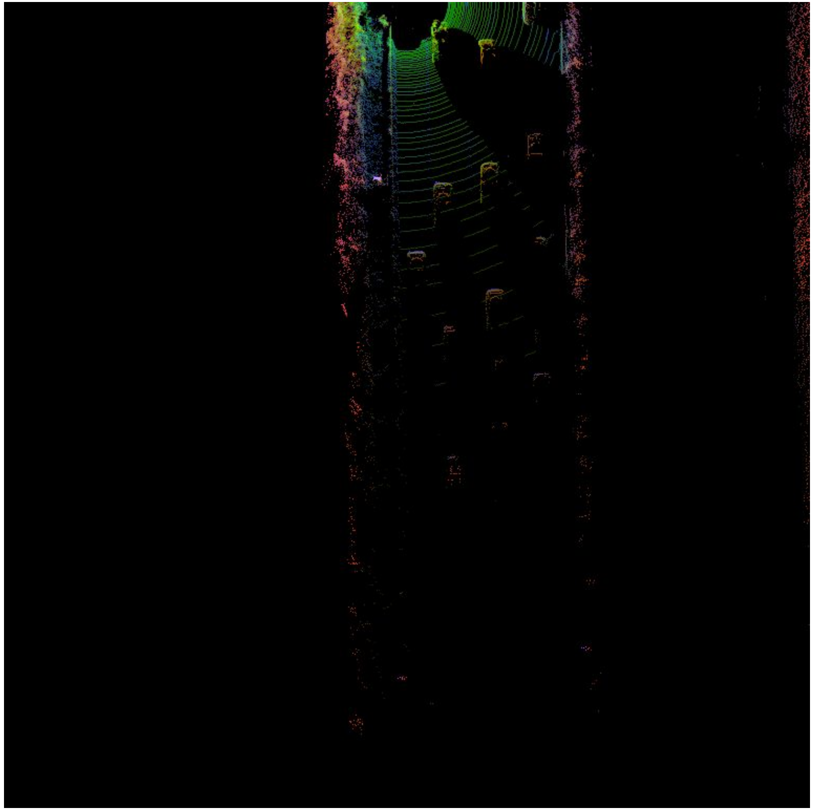
- T = 64
- 768 X 768 crops of the BEV Maps are used
- Supervised Method



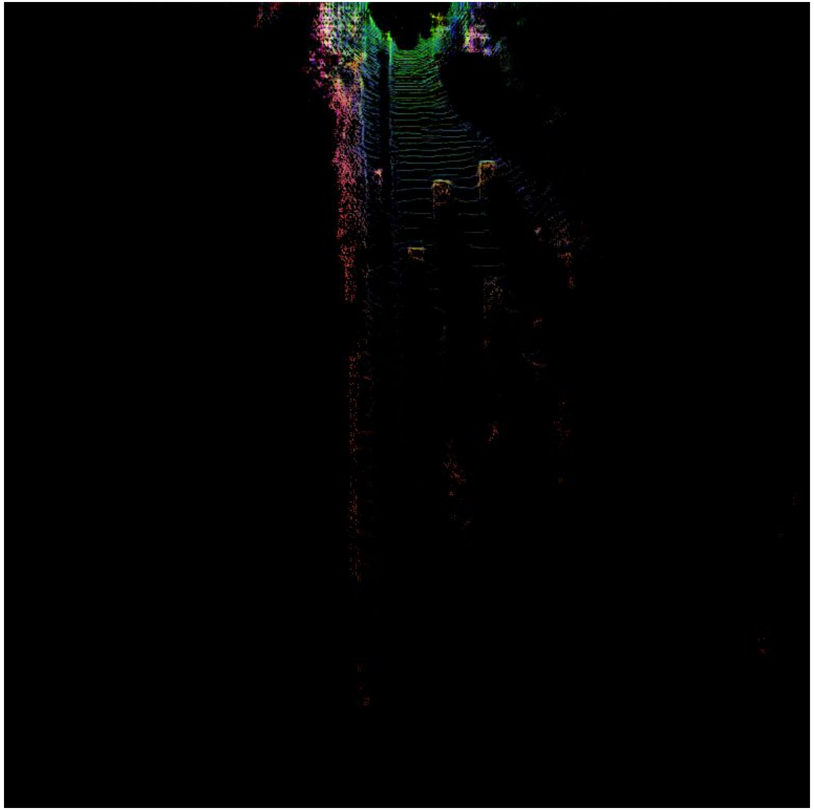
Actual LiDAR



Pseudo-LiDAR



Actual LiDAR



Modified Pseudo-LiDAR

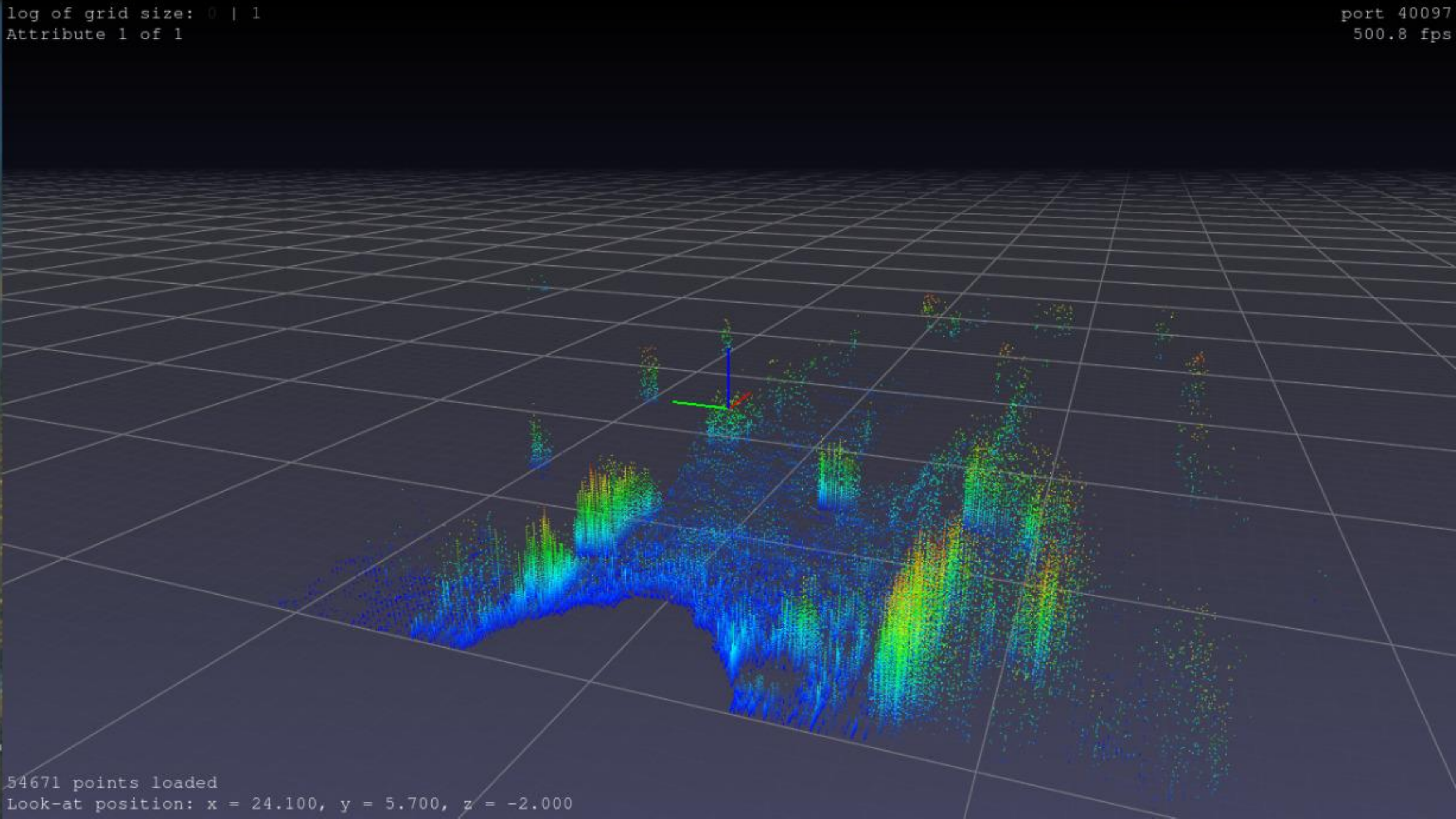
Results

	Easy	Medium	Hard
Car	61.2/42.9	37.5/24.7	31.2/21.4
Pedestrian	27.7/18.9	23.8/15.9	20.8/12.9
Cyclist	41.7/35.1	25.8/23.4	25.3/21.4

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird's Eye View Map

Drawbacks of Approach IV

- Although the BEV maps generated are much better than others, the improvements in 3D object detection is not as much as expected.
- The reconstruction algorithm from BEV maps to point clouds is not satisfactory.


log of grid size: 0 | 1
Attribute 1 of 1
port 40097
500.8 fps
54671 points loaded
Look-at position: x = 24.100, y = 5.700, z = -2.000

Comparison (Car detection)

	Easy	Medium	Hard
Approach I	47.7/28.2	29.7/16.8	24.5/15.7
Approach II	55.1/34.7	35.4/21.9	29.4/17.7
Approach III	31.8/15.4	22.5/11.4	20.4/10.3
Approach IV	61.2/42.9	37.5/24.7	31.2/21.4

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird’s Eye View Map

Comparison (Pedestrian detection)

	Easy	Medium	Hard
Approach I	22.9/16.5	18.9/14.1	17.1/12.3
Approach II	17.6/7.7	14.0/6.1	12.4/5.4
Approach III	8.3/6.5	8.0/5.9	7.4/5.7
Approach IV	27.7/18.9	23.8/15.9	20.8/12.9

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird's Eye View Map

Comparison (Cyclist detection)

	Easy	Medium	Hard
Approach I	18.2/12.1	10.8/8.1	10.7/7.2
Approach II	20.7/13.8	14.8/11.3	14.1/11.3
Approach III	8.3/4.4	5.1/2.9	4.6/2.7
Approach IV	41.7/35.1	25.8/23.4	25.3/21.4

For IoU = 0.7, AP_{BEV}/AP_{3D}, BEV: Bird's Eye View Map

Future Work

- Work on translation directly on point clouds.
- Work on a better reconstruction algorithm to go back from BEV maps to point clouds, especially for higher density.

References

[1]Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8445–8453, 2019.

[2] Xinshuo Weng and Kris Kitani.Monocular 3d object detection with pseudo-lidar point cloud. arXiv preprint arXiv:1903.09847, 2019

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1907–1915,2017.

[4] Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.