



UNIVERSITY OF  
**OXFORD**

Department of  
**COMPUTER SCIENCE**

# **Towards Semantic Understanding of Urban-Scale 3D Point Clouds: Datasets, Benchmarks and Challenges**

*Qingyong Hu*

*Supervisor: Niki Trigoni & Andrew Markham*

*Department of Computer Science, University of Oxford, UK*

Contact Email:

*[qingyong.hu@cs.ox.ac.uk](mailto:qingyong.hu@cs.ox.ac.uk)*

# Agenda

01

## Introduction

*Background & Literatures*

02

## Urban-Scale Point Cloud Dataset

*SensatUrban, CVPR'21 & IJCV'21*

03

## Urban-Scale Data Generation

*STPLS3D, Arxiv'22*

04

## Urban-Scale Scene Understanding

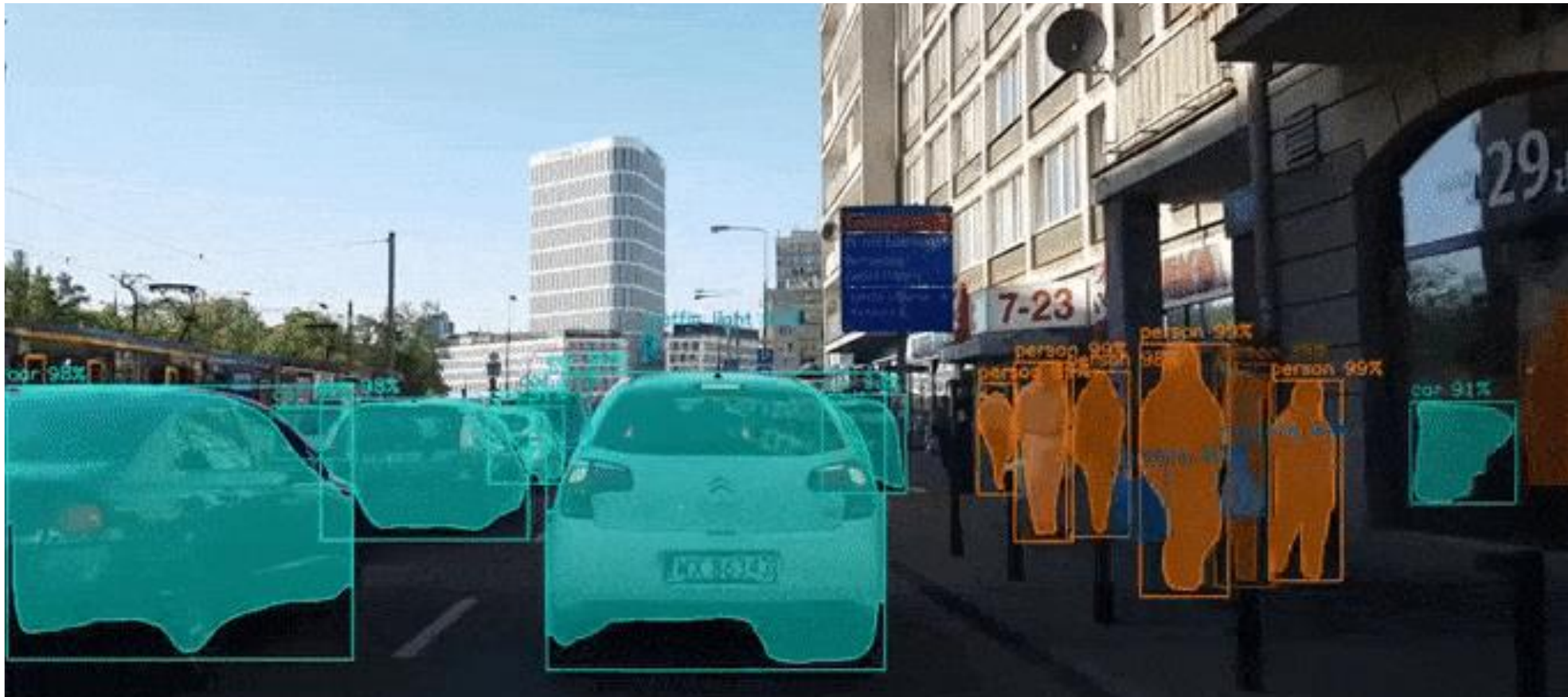
*SQN, Arxiv'22*

05

## Conclusion

*Future works*

Today's AI systems: 2D Recognition / Detection / Segmentation



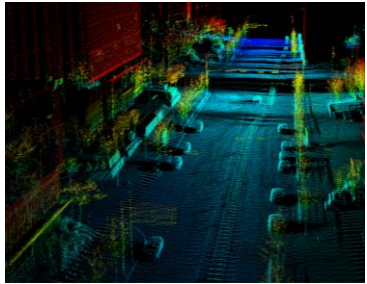
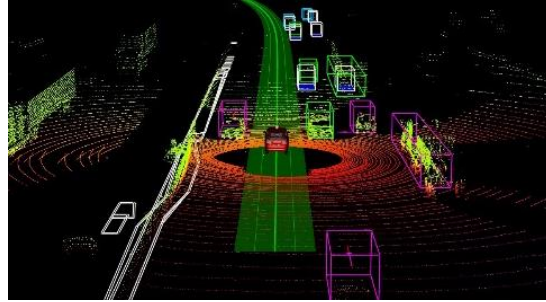
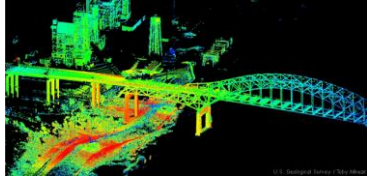
These systems do not really perceive the 3D world !



Future of AI: Intelligent systems that perceive the 3D world



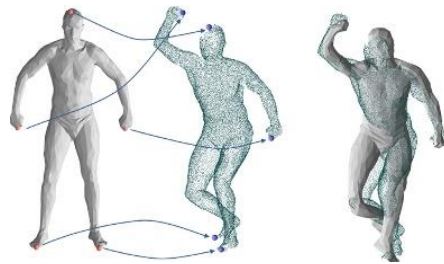




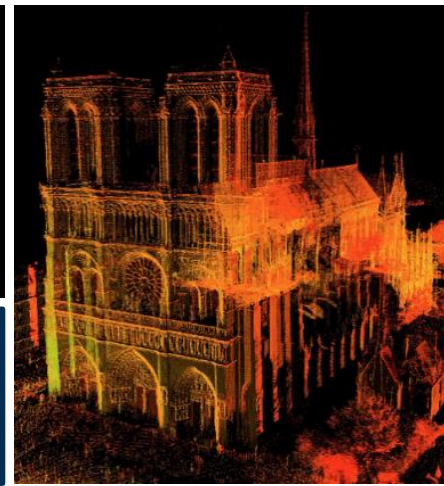
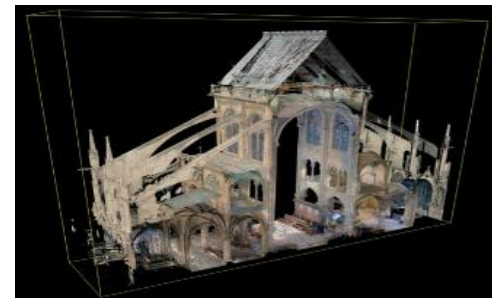
**Autonomous Driving**



**AR/VR**



**Entertainment**



**Cultural Heritages**



Figure from Xiangli et al. "CityNeRF: Building NeRF at City Scale"

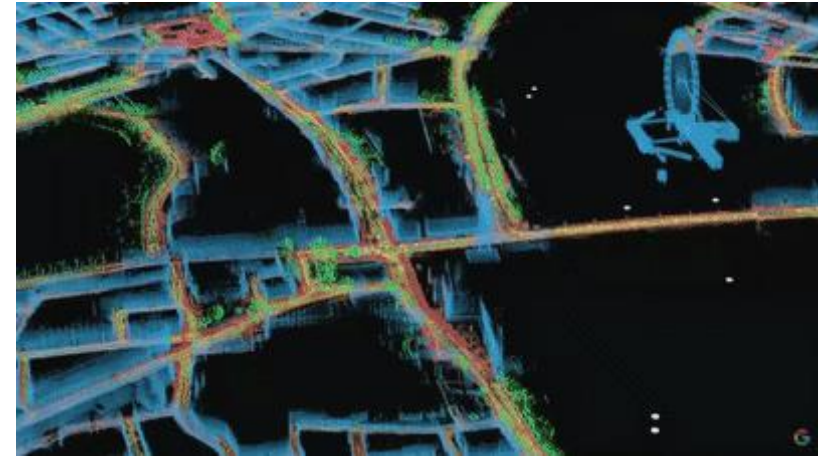


Figure from "Immerse View for Google Maps"

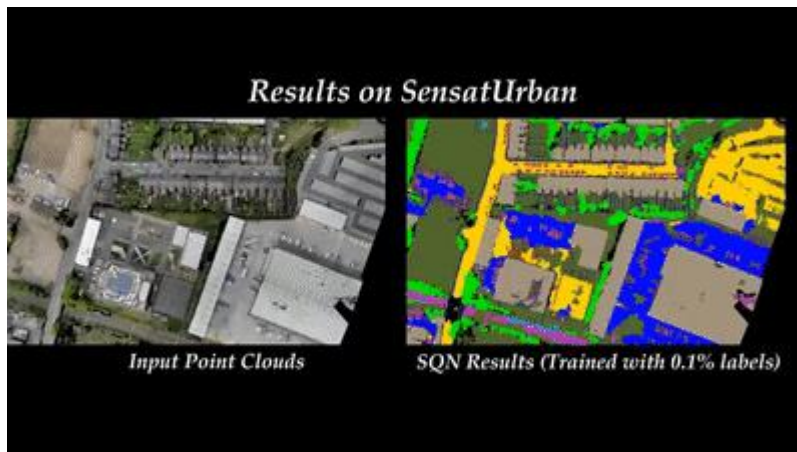


Figure from Hu et al, "SQN: Weakly-Supervised Semantic Segmentation of Large-Scale 3D Point Clouds"



Figure from Liu et al, "UrbanScene 3D: A Large Scale Urban Scene Dataset and Simulator"

## Semantic Understanding of Urban-Scale 3D Scenes

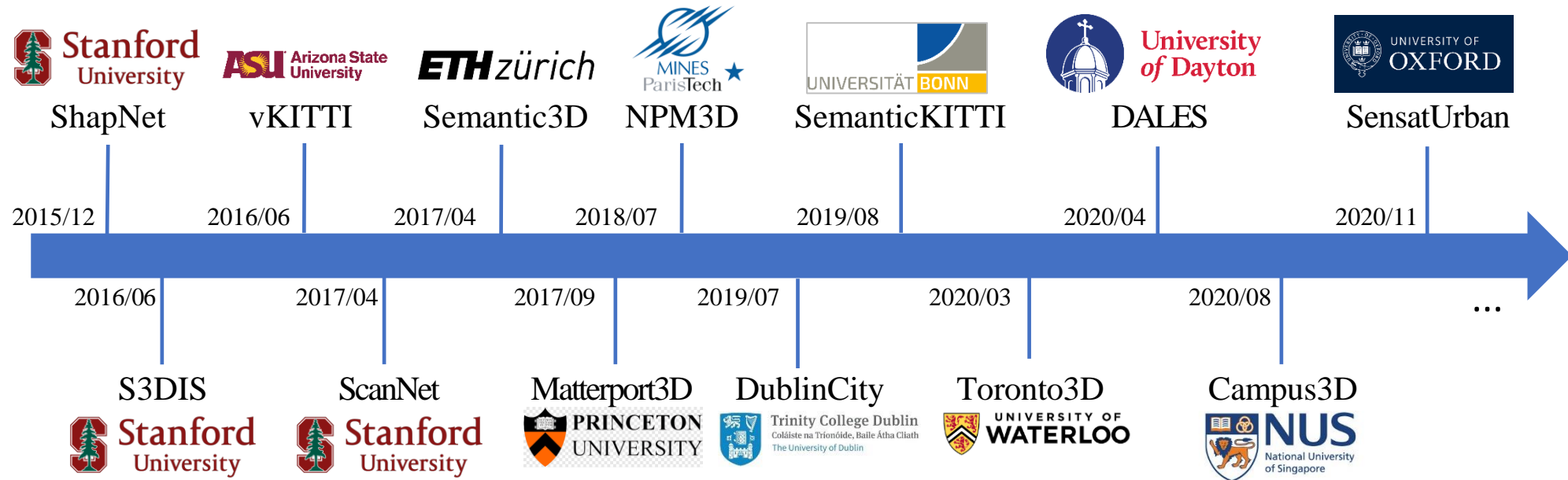
- *How to build urban-scale 3D datasets? What are the main challenges of urban 3D understanding?*
- *How to achieve synthetical generation of urban-scale 3D scenes?*
- *How to achieve label-efficient learning of large-scale 3D scenes?*



## Research Question 1

**How to build urban-scale 3D datasets? What are the main challenges of urban 3D understanding?**

- Large-scale annotated datasets have driven tremendous progress in this field





2015/12

2016/06

2017/04

2018/07

2019/08

2020/04

2016/06

2017/04

2017/09

2019/07

2020/03

2020/08



Figure from Chang et al, "ShapeNet: An Information-Rich 3D Model Repository", Arxiv 2015



Figure from Qi et al, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation", CVPR 2017

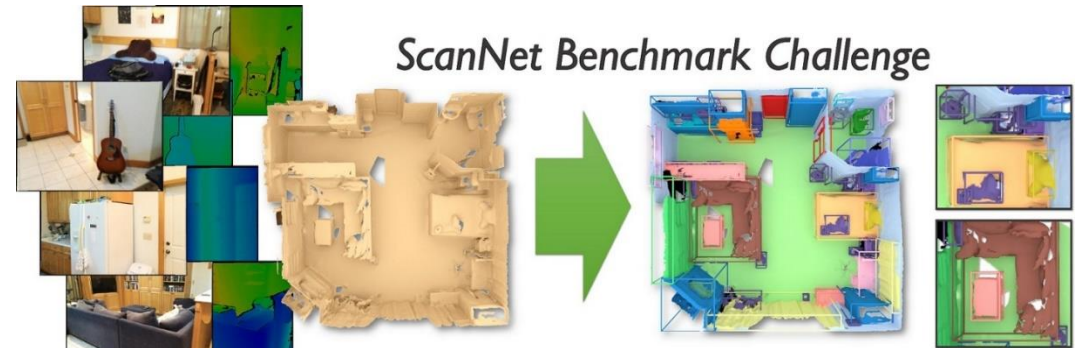


Figure from Dai et al, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes", CVPR 2017

Indoor Scene-Level, RGB-D surface reconstructed Point Clouds



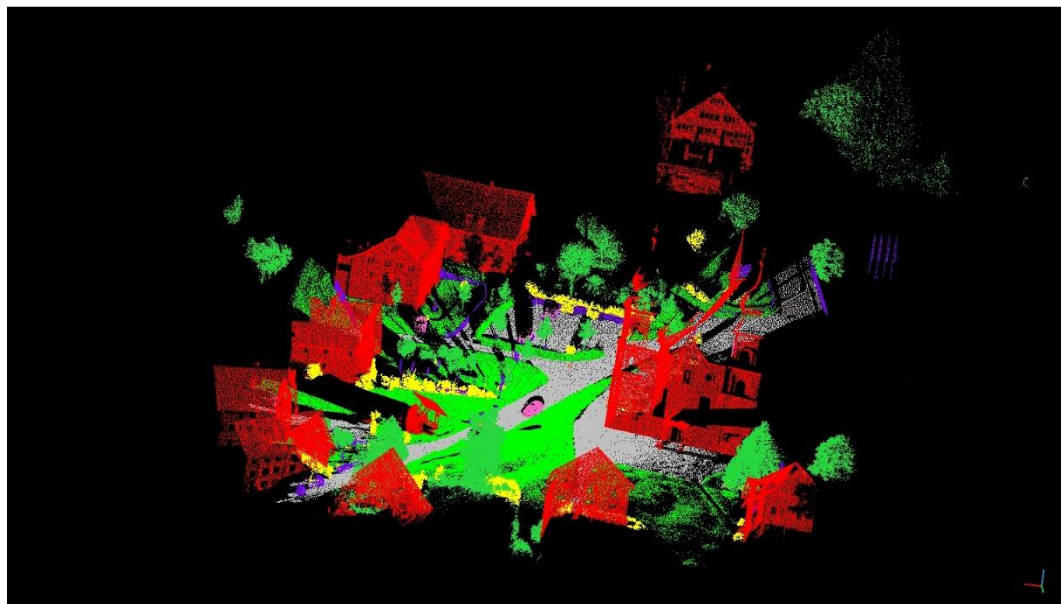
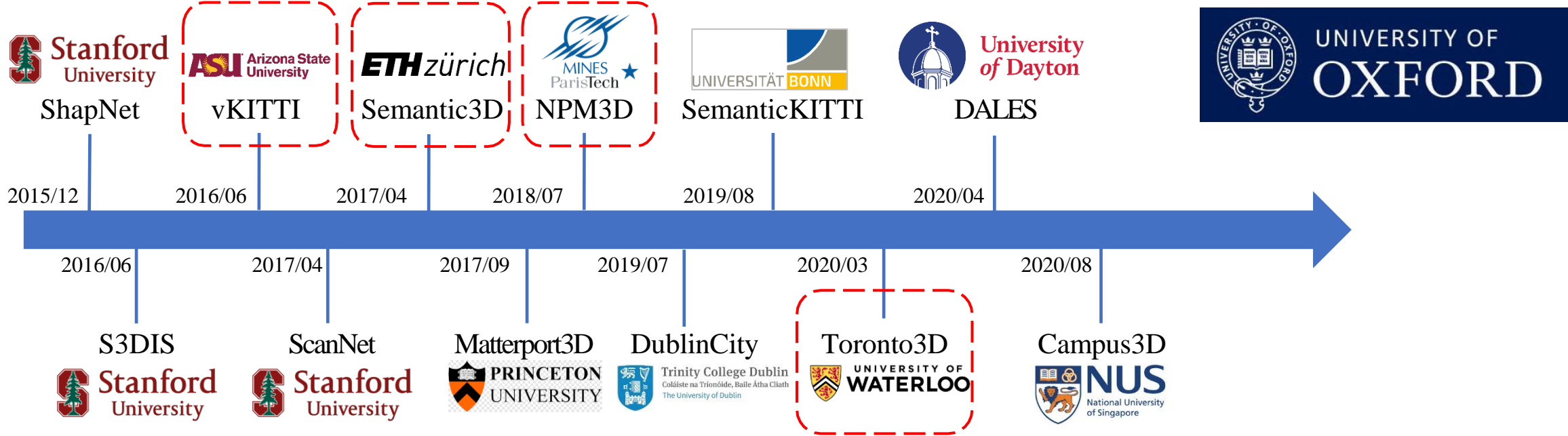


Figure from Hackel et al, "SEMANTIC3D.NET: A new large-scale point cloud classification benchmark", ISPRS 2017

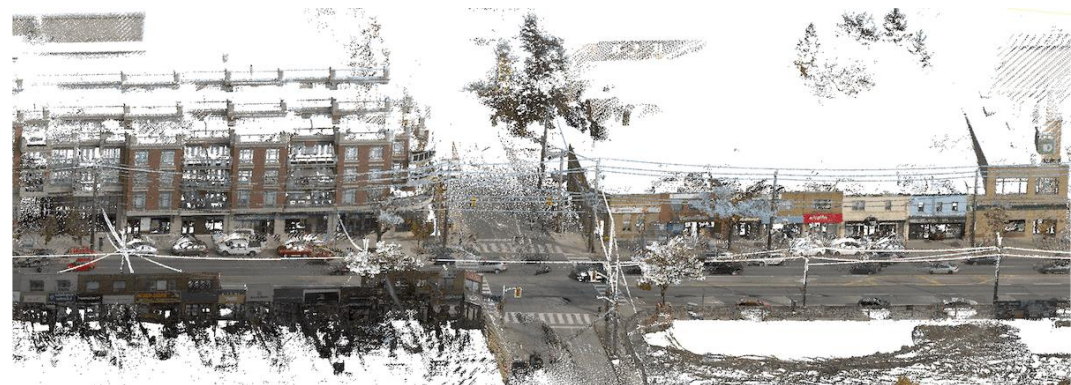


Figure from Tan et al, "Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways", CVPRW 2020

## Outdoor Roadway-Level Point Clouds

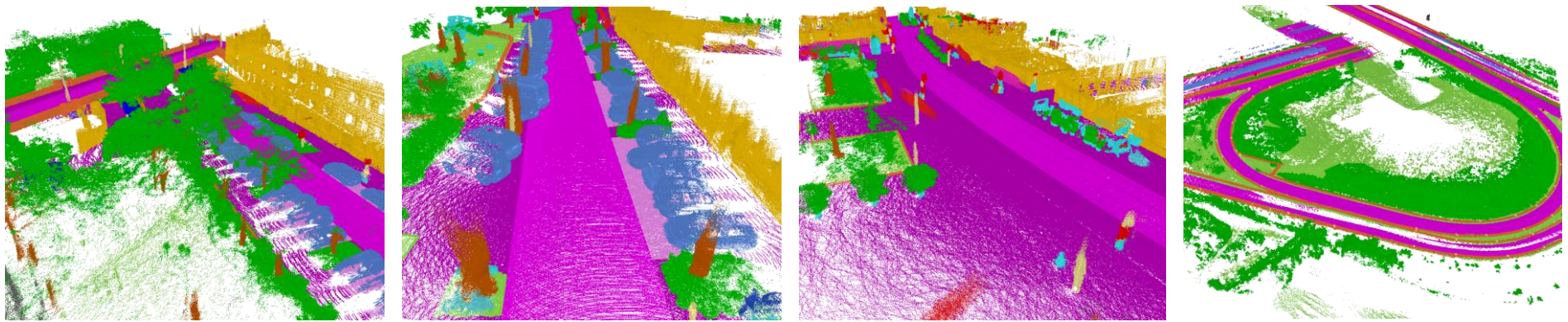
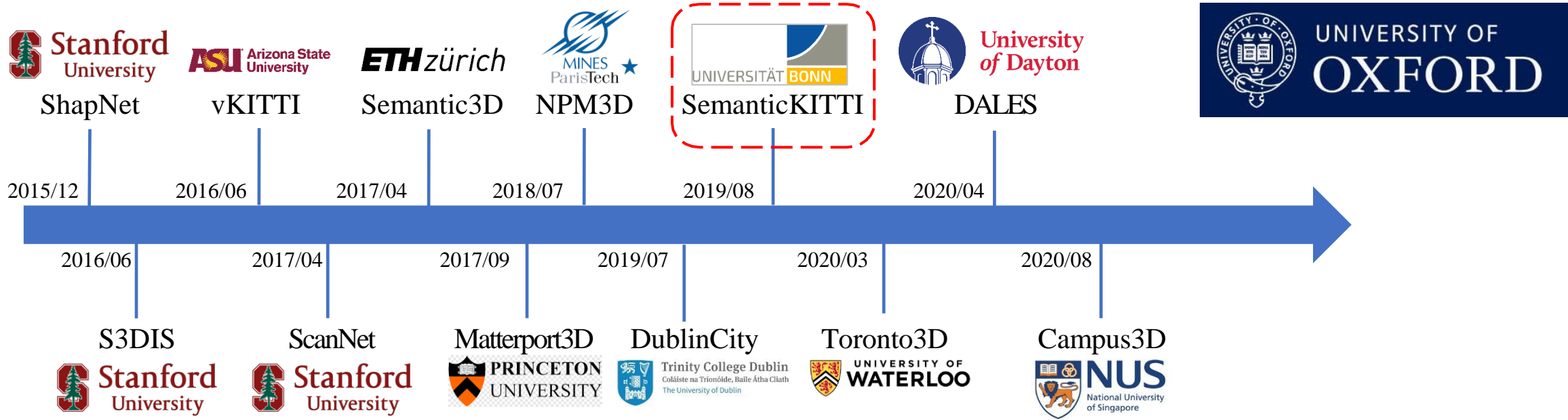


Figure from Behley et al, "Semantickitti: A dataset for semantic scene understanding of LiDAR sequences", ICCV 2019

### Sequential Street-View LiDAR Point Clouds



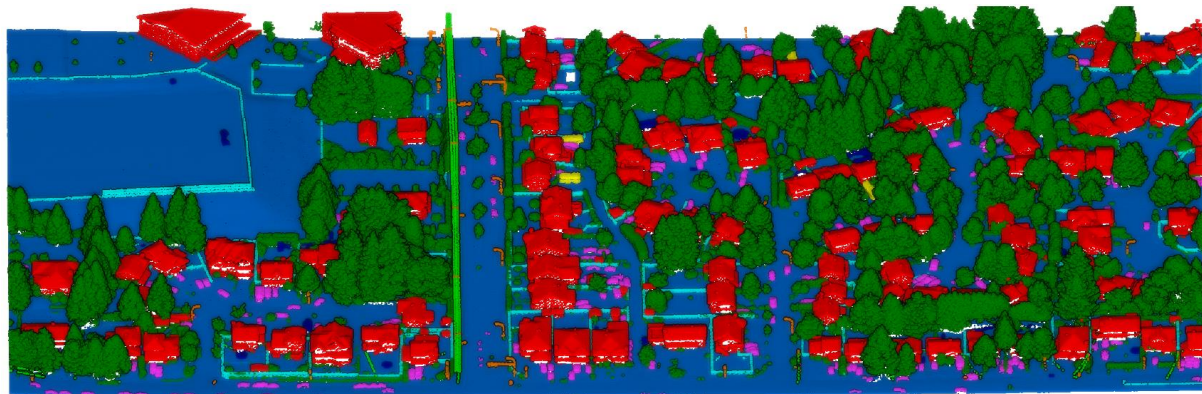
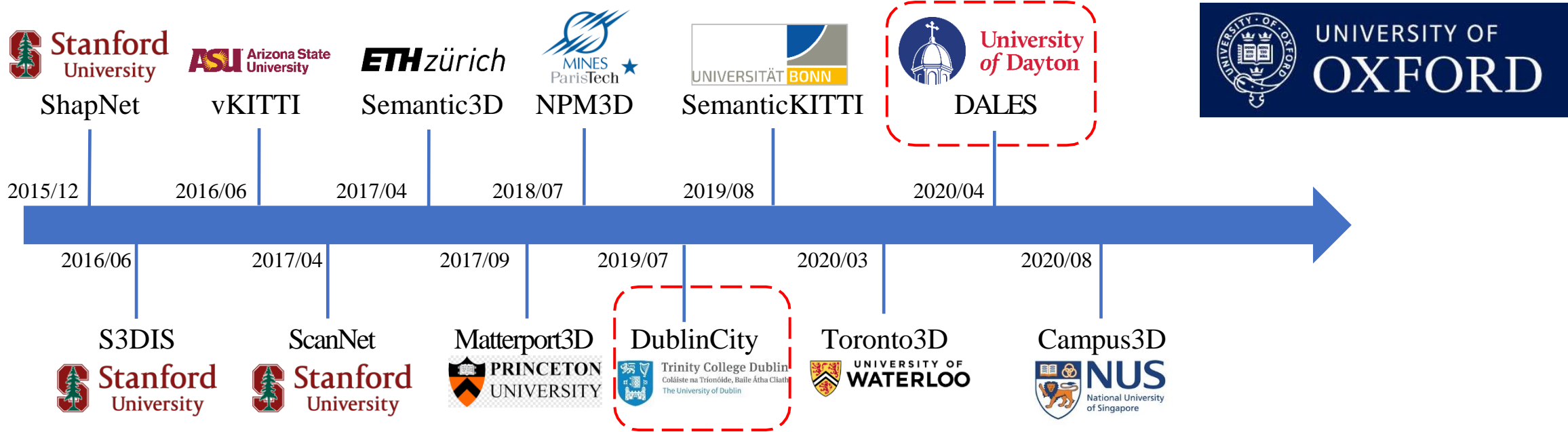


Figure from Varney et al, "DALES: A large-scale aerial LiDAR data set for semantic segmentation", CVPRW 2020

## Aerial Urban-level LiDAR Point Clouds

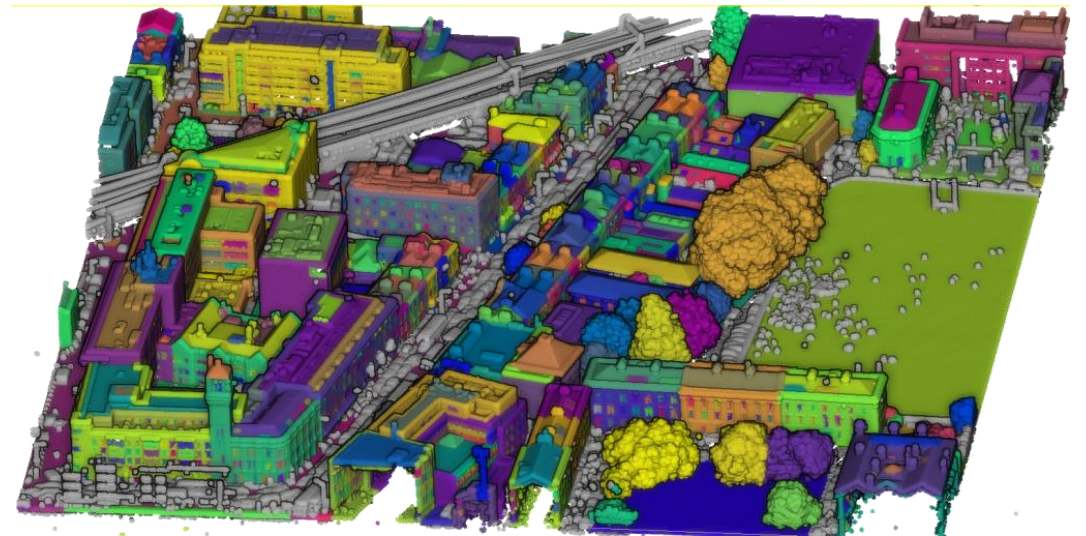
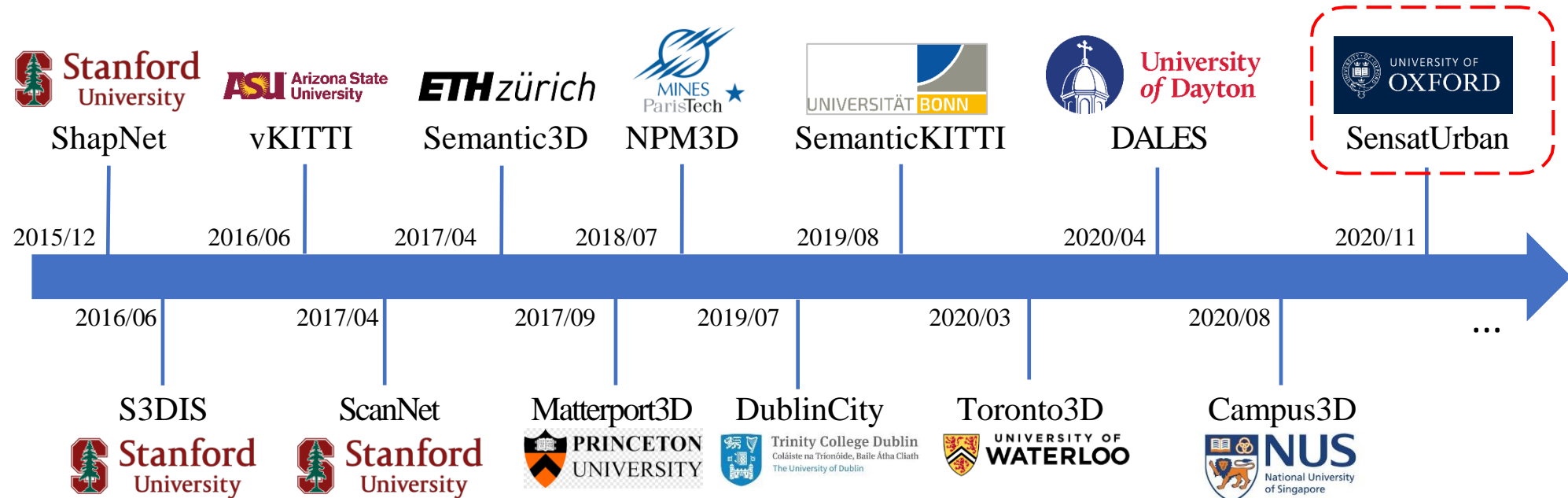
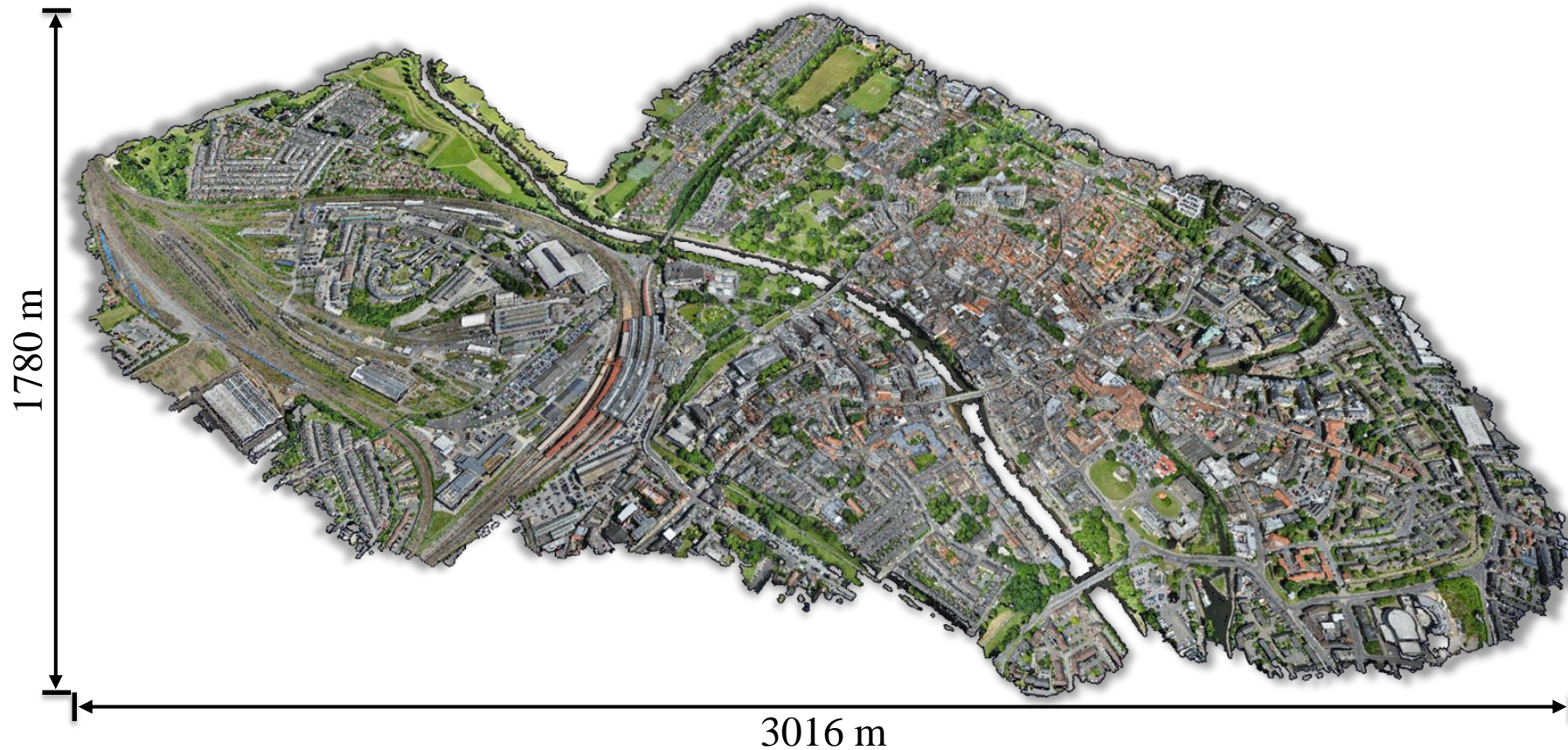


Figure from Zolanvari et al, "DublinCity: Annotated LiDAR Point Cloud and its Applications", BMVC 2019



- Various modalities: indoor RGB-D reconstruction, outdoor LiDAR
- Increasing spatial scale: from object-level -> indoor scene-level -> outdoor roadway-level -> urban city-level
- Richer information: 3D coordinates, RGB color, sequential flow
- Geometrical structure: simple object -> complex structure





- Largest urban-scale photogrammetric point cloud dataset (nearly three billion labeled 3D points )
- Consists of large areas from three UK cities, covering about  $7.6 \text{ km}^2$  of the city land-scape.
- Point is manually-labeled as one of 13 semantic categories such as *ground, vegetation, car, etc*



02

# SensatUrban

Data Examples

## ■ Dataset





02

# SensatUrban

Data Examples

## ■ Dataset





02

# SensatUrban

Data Examples

## ■ Dataset

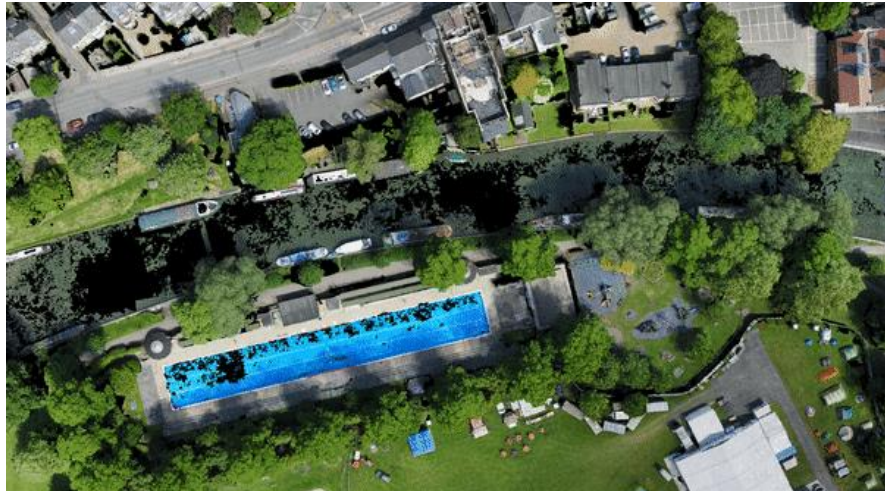




02

# SensatUrban

Visualization



## ■ Comparison with existing datasets

	#Name and Reference	#Year	#Spatial size <sup>1</sup>	#Classes <sup>2</sup>	#Points	#RGB	#Sensors
Object-level	ShapeNet [8]	2015	-	55	-	No	Synthetic
	PartNet [34]	2019	-	24	-	No	Synthetic
Indoor Scene-level	S3DIS [3]	2017	$6 \times 10^3 m^2$	13 (13)	273M	Yes	Matterport
	ScanNet [13]	2017	$1.13 \times 10^5 m^2$	20 (20)	242M	Yes	RGB-D
Outdoor Roadway-level	Paris-rue-Madame [44]	2014	$0.16 \times 10^3 m$	17	20M	No	MLS
	IQmulus [54]	2015	$10 \times 10^3 m$	8 (22)	300M	No	MLS
	Semantic3D [21]	2017	-	8 (9)	4000M	Yes	TLS
	Paris-Lille-3D [43]	2018	$1.94 \times 10^3 m$	9 (50)	143M	No	MLS
	SemanticKITTI [5]	2019	$39.2 \times 10^3 m$	25 (28)	4549M	No	MLS
	Toronto-3D [48]	2020	$1 \times 10^3 m$	8 (9)	78.3M	Yes	MLS
Urban-level	ISPRS [42]	2012	-	9	1.2M	No	ALS
	DublinCity [67]	2019	$2 \times 10^6 m^2$	13	260M	No	ALS
	DALES [55]	2020	$10 \times 10^6 m^2$	8 (9)	505M	No	ALS
	LASDU [61]	2020	$1.02 \times 10^6 m^2$	5	3.12M	No	ALS
	Campus3D [26]	2020	$1.58 \times 10^6 m^2$	24	937.1M	Yes	UAV Photogrammetry
	<b>SensatUrban (Ours)</b>	2020	$7.64 \times 10^6 m^2$	13 (31)	2847M	Yes	UAV Photogrammetry

Table 1: Comparison with the representative datasets for segmentation of 3D point clouds. <sup>1</sup>The spatial size (Area/Length) in the dataset, m: meter, <sup>2</sup> The number of classes used for evaluation and the number of sub-classes annotated in brackets. MLS: Mobile Laser Scanning system, TLS: Terrestrial Laser Scanning system, ALS: Aerial Laser Scanning system.



## ■ Acquisition Equipment



eBee X Fixed-Wing Mapping Drone



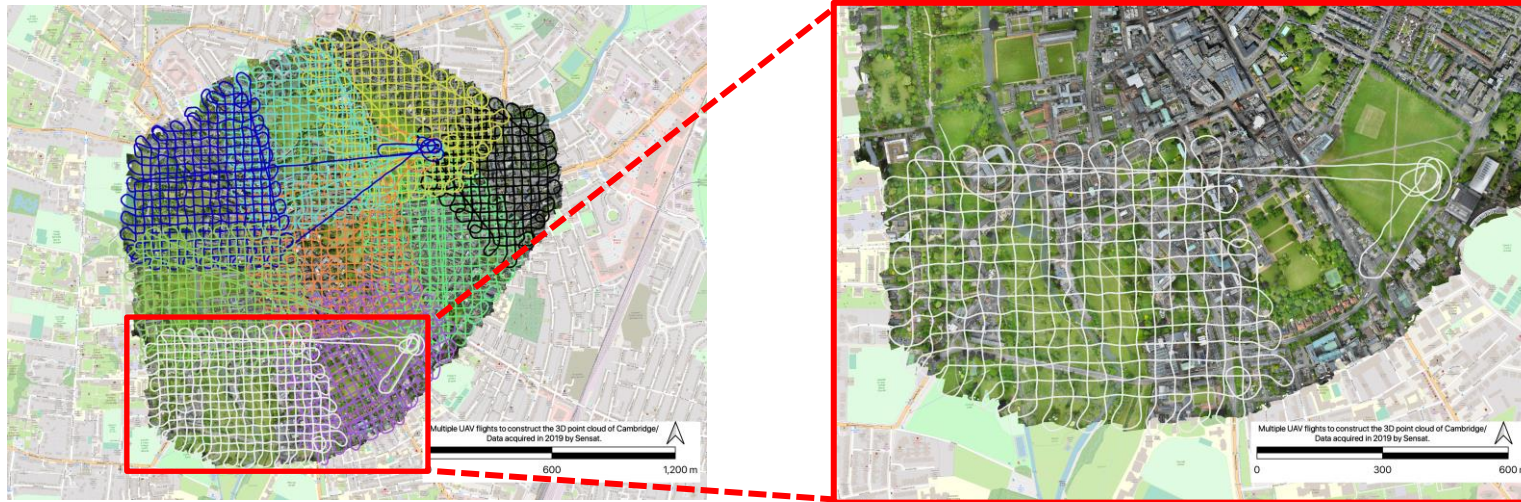
senseFly S.O.D.A. 3D  
photogrammetry camera

**Fig. 2** The drones and cameras we used in the urban survey.

**Table 2** Detailed specifications of the camera (*i.e.*, Sensefly SODA 3D camera) used in our survey.

		Specification
Sensor size		1 inch
RGB Lens	F/2.8-11, 10.6 mm (35 mm equivalent: 29 mm)	
RGB Resolution		5,472 x 3,648 px (3:2)
Exposure compensation		±2.0 (1/3 increments)
Shutter		Global Shutter 1/30 – 1/2000s
White balance		Auto, sunny, cloudy, shady
ISO range		125-6400
RGB FOV	Total FOV: 154°, 64° optical, 90° mechanical	
GNSS		RTK/PPK

## Sequential Aerial Imagery Acquisition



(a) Multi-flights survey

(b) Zoomed-in single  
flight survey

Figure 2: The survey of a region in Cambridge. All 9 flight plans (*left*) are collated together to cover the site. Lines with different colors represent different flight paths of UAVs. The circular path is the takeoff and landing pattern.



### Point-wise Semantic Annotations

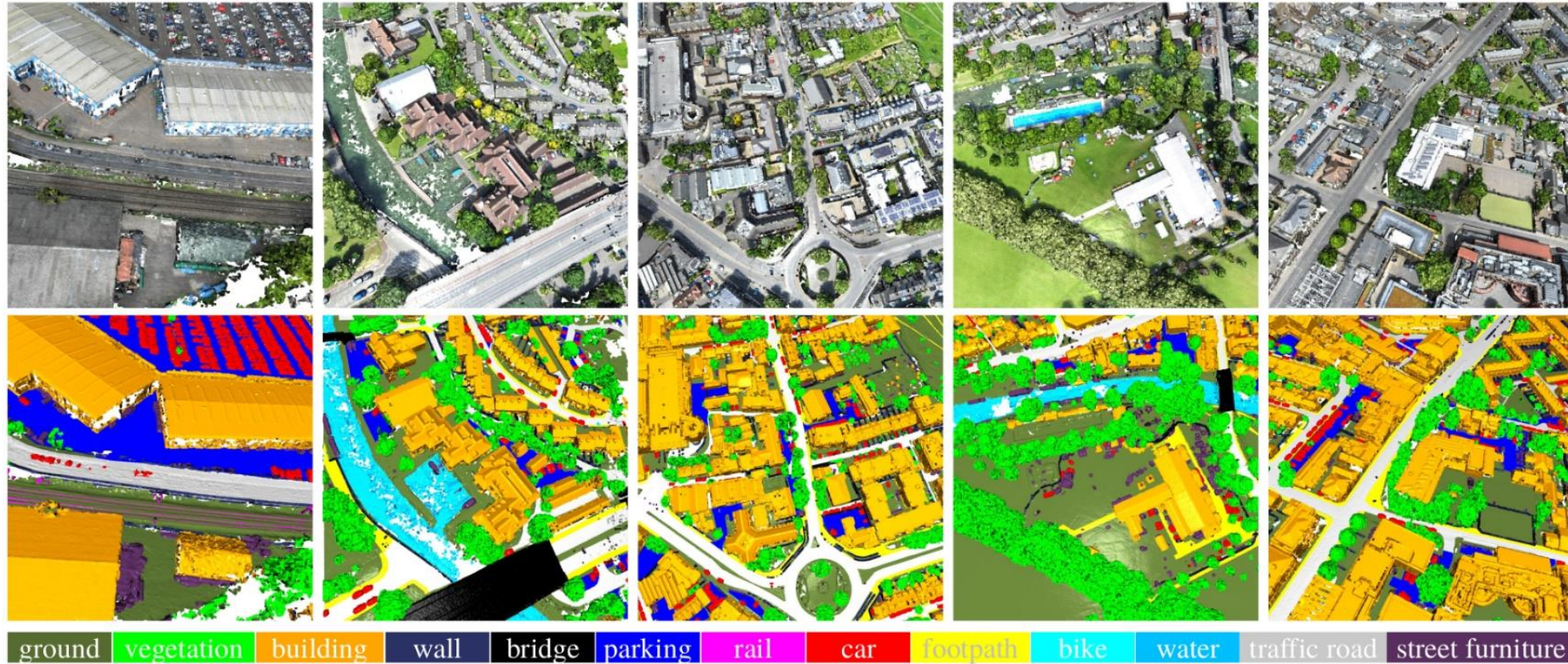
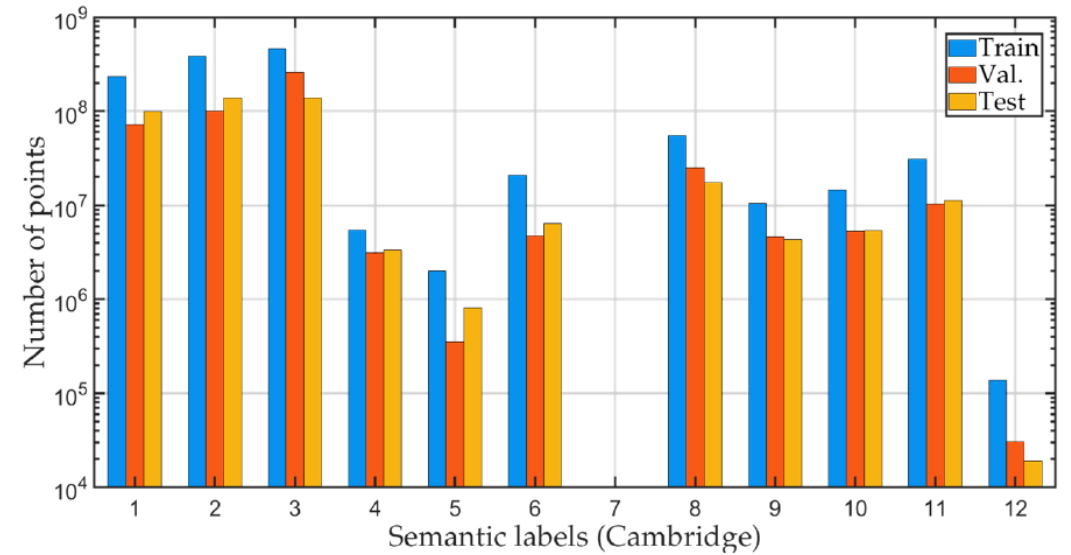
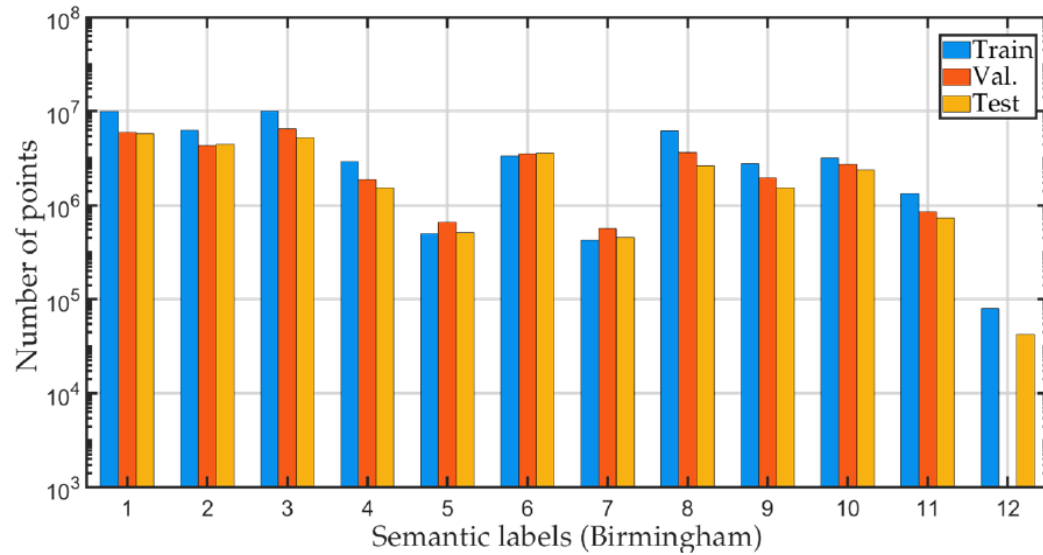


Figure 3: Examples of our SensatUrban dataset. Different semantic classes are labeled by different colors.



## Data Distribution



**Fig. 6** Statistics of our SensatUrban dataset. The number of points in different semantic categories is reported. Please note that the vertical axis is on the logarithmic scale. Additionally, there are no points annotated as *rail* in Cambridge.

## ■ Evaluation of 7 representative methods

	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37]	80.78	30.32	23.71	67.96	89.52	80.05	0.00	0.00	3.95	0.00	31.55	0.00	35.14	0.00	0.00	0.00
PointNet++ [38]	84.30	39.97	32.92	72.46	94.24	84.77	2.72	2.09	25.79	0.00	31.54	11.42	38.84	7.12	0.00	56.93
TagentConv [50]	76.97	43.71	33.30	71.54	91.38	75.90	35.22	0.00	<u>45.34</u>	0.00	26.69	19.24	67.58	0.01	0.00	0.00
SPGraph [24]	85.27	44.39	37.29	69.93	94.55	88.87	32.83	12.58	15.77	<b>15.48</b>	30.63	22.96	56.42	0.54	0.00	44.24
SparseConv [19]	88.66	63.28	42.66	74.10	97.90	<u>94.20</u>	<u>63.30</u>	7.50	24.20	0.00	30.10	<u>34.00</u>	74.40	0.00	0.00	54.80
KPConv [51]	<b>93.20</b>	<u>63.76</u>	<b>57.58</b>	<b>87.10</b>	<b>98.91</b>	<b>95.33</b>	<b>74.40</b>	<u>28.69</u>	41.38	0.00	<u>55.99</u>	<b>54.43</b>	<b>85.67</b>	<b>40.39</b>	0.00	<b>86.30</b>
RandLA-Net [23]	<u>89.78</u>	<b>69.64</b>	<u>52.69</u>	<u>80.11</u>	<u>98.07</u>	91.58	48.88	<b>40.75</b>	<b>51.62</b>	0.00	<b>56.67</b>	33.23	<u>80.14</u>	<u>32.63</u>	0.00	<u>71.31</u>

Table 2: Benchmark results of the baselines on our SensatUrban. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) scores are reported.

- Carefully selected 7 representative baselines, including projection, volumetric, point based methods
- KPConv achieves the highest mIoU scores
- A number of key categories such as bridge, rail, street, footpath, bike that are poorly segmented.

## ■ Challenge 1: Data preparation

	Sampling	Input sets	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet	Grid	Constant Number	<b>90.57</b>	<b>56.30</b>	<b>49.69</b>	<b>83.55</b>	<b>97.67</b>	90.66	<b>22.56</b>	<b>43.54</b>	40.35	9.29	<b>50.74</b>	<b>29.58</b>	68.24	<b>29.27</b>	0.00	<b>80.55</b>
PointNet	Grid	Constant Volume	88.27	49.80	42.44	80.20	96.43	87.88	8.45	35.14	32.52	0.00	43.03	19.26	54.66	18.26	0.00	75.87
PointNet	Random	Constant Number	90.34	55.17	48.49	83.47	97.51	<b>90.89</b>	18.55	33.31	<b>42.82</b>	<b>11.85</b>	47.95	26.83	<b>68.37</b>	29.12	0.00	79.71
PointNet	Random	Constant Volume	88.09	48.45	41.68	79.82	96.24	87.64	5.69	27.70	34.98	0.00	42.85	13.81	54.29	20.64	0.00	78.24
RandLA-Net	Grid	Constant Number	<b>91.55</b>	<b>74.87</b>	<b>58.64</b>	<b>82.99</b>	<b>98.43</b>	<b>93.41</b>	<b>57.43</b>	<b>49.47</b>	<b>55.12</b>	<b>27.33</b>	<b>60.65</b>	<b>39.43</b>	<b>84.57</b>	<b>39.48</b>	0.00	73.97
RandLA-Net	Grid	Constant Volume	88.11	64.91	49.18	78.18	97.92	90.87	45.02	30.89	35.82	0.00	45.73	31.96	77.78	29.90	0.00	75.30
RandLA-Net	Random	Constant Number	91.14	74.14	57.55	82.25	98.33	92.37	54.20	43.10	54.74	25.02	60.40	39.17	82.77	37.59	0.00	<b>78.25</b>
RandLA-Net	Random	Constant Volume	88.37	60.84	47.27	81.16	97.52	90.45	44.75	16.36	37.18	0.00	42.19	26.28	76.76	30.46	0.00	71.39

Table 3: Quantitative results achieved by PointNet [23] and RandLA-Net [23] with different input preparation steps. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

- Step 1: Downsample the raw point clouds at the very beginning (Random sampling vs. Grid sampling)
- Step 2: To obtain individual input set of points to feed into the networks. (Constant-number vs. Constant-volume)



## ■ Challenge 2: Geometry vs. Appearance

	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37] (w/o RGB)	83.50	33.52	28.85	67.35	92.66	84.72	16.02	0.00	13.65	2.68	17.09	0.33	54.54	0.00	0.00	26.04
PointNet [37] (w/ RGB)	90.57	56.30	49.69	83.55	97.67	90.66	22.56	43.54	40.35	9.29	50.74	29.58	68.24	29.27	0.00	80.55
PointNet++ [38] (w/o RGB)	90.85	56.94	50.71	79.05	98.37	94.22	66.76	39.74	37.51	0.00	51.53	38.82	81.71	5.80	0.00	65.68
PointNet++ [38] (w RGB)	93.10	64.96	58.13	86.38	98.76	94.72	65.91	50.41	50.53	0.00	58.40	46.95	82.31	38.40	0.00	<b>82.88</b>
SPGraph [24] (w/o RGB)	84.81	42.12	35.29	69.60	94.18	88.15	34.55	20.53	15.83	16.34	31.44	10.54	55.01	0.98	0.00	21.57
SPGraph [24] (w RGB)	85.27	44.39	37.29	69.93	94.55	88.87	32.83	12.58	15.77	15.48	30.63	22.96	56.42	0.54	0.00	44.24
KPConv [51] (w/o RGB)	91.47	57.43	51.79	80.43	98.82	94.93	74.17	44.53	32.11	0.00	54.32	37.83	84.88	14.48	0.00	56.79
KPConv [51] (w RGB)	<b>93.92</b>	71.44	<b>64.50</b>	<b>87.04</b>	<b>99.01</b>	<b>96.31</b>	<b>77.73</b>	<b>58.87</b>	49.88	<b>37.84</b>	<b>62.74</b>	<b>56.60</b>	<b>86.55</b>	<b>44.86</b>	0.00	81.01
RandLA-Net [23] (w/o RGB)	88.90	67.96	51.53	77.30	97.92	91.24	51.94	47.46	45.04	9.71	49.79	34.21	79.97	21.13	0.00	64.18
RandLA-Net [23] (w RGB)	91.24	<b>74.68</b>	58.14	82.23	98.39	92.69	56.62	49.00	<b>54.19</b>	25.10	60.98	38.69	83.42	38.74	0.00	75.80

Table 4: Quantitative results of five selected baselines on our SensatUrban dataset. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

### ■ Challenge 3: Extremely imbalanced distribution

	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet+ce	<b>90.57</b>	56.30	49.69	<b>83.55</b>	<b>97.67</b>	<b>90.66</b>	22.56	43.54	40.35	9.29	<b>50.74</b>	29.58	<b>68.24</b>	29.27	0.00	<b>80.55</b>
PointNet+wce [23]	88.13	<b>68.05</b>	51.24	81.01	97.12	87.87	24.46	45.76	47.78	<b>34.93</b>	49.82	29.58	61.28	31.78	0.00	74.67
PointNet+wce+sqrt [2]	89.72	67.97	52.35	82.87	97.33	90.42	28.32	44.94	48.39	32.07	49.58	<b>32.63</b>	65.11	32.59	<b>2.60</b>	73.71
PointNet+lovas [6]	89.58	67.50	<b>52.53</b>	82.74	97.27	90.28	28.11	43.89	<b>48.53</b>	33.58	49.68	32.21	64.01	<b>33.05</b>	1.46	78.13
PointNet+focal [28]	89.46	67.33	52.37	82.47	97.34	90.25	<b>28.36</b>	<b>51.87</b>	46.40	30.50	48.62	32.43	65.00	32.23	1.21	74.10
RandLA-Net+ce	<b>93.10</b>	64.30	57.77	<b>85.39</b>	<b>98.63</b>	<b>95.40</b>	62.55	54.85	56.49	0.00	58.13	<b>45.90</b>	82.24	30.68	0.00	80.70
RandLA-Net+wce [23]	91.24	74.68	58.14	82.23	98.39	92.69	56.62	49.00	54.19	25.10	<b>60.98</b>	38.69	83.42	38.74	0.00	75.80
RandLA-Net+wce+sqrt [2]	92.51	<b>79.92</b>	<b>62.80</b>	84.94	98.47	95.07	59.01	<b>62.18</b>	56.76	28.96	57.36	44.47	<b>84.67</b>	41.67	<b>24.31</b>	78.49
RandLA-Net+lovas [6]	92.56	76.99	61.51	84.92	98.55	94.64	<b>63.17</b>	52.37	55.43	<b>36.37</b>	59.35	45.79	84.28	41.24	2.66	<b>80.89</b>
RandLA-Net+focal [28]	92.49	77.26	60.41	85.03	98.38	94.74	59.49	58.70	<b>57.11</b>	25.97	58.19	42.74	82.26	<b>42.00</b>	2.71	77.97

Table 5: Quantitative results achieved by PointNet [37] and RandLA-Net [23] with different loss functions. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.



## ■ Challenge 4: Cross-city generalization

	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37]	87.33	54.76	48.73	80.91	94.58	87.40	33.69	0.51	66.23	16.98	49.55	36.08	74.59	1.49	0.00	91.51
PointNet++ [38]	89.85	64.24	57.39	84.34	<u>97.11</u>	89.74	<u>61.56</u>	<b>3.78</b>	68.08	41.95	54.43	<u>51.54</u>	84.73	14.43	0.00	<b>94.34</b>
SPGraph [24]	80.13	42.87	36.95	65.75	93.33	87.24	41.28	0.00	42.69	20.94	2.28	32.05	64.06	0.00	0.00	30.76
KPConv [51]	<b>91.44</b>	<u>68.41</u>	<b>61.65</b>	<b>86.00</b>	<b>97.66</b>	<b>92.90</b>	<b>75.07</b>	0.91	<u>69.74</u>	<b>55.50</b>	<u>57.94</u>	<b>60.73</b>	<b>89.48</b>	<u>21.44</u>	0.00	<u>94.13</u>
RandLA-Net [23]	<u>90.77</u>	<b>72.11</b>	<u>59.72</u>	<u>85.14</u>	96.89	<u>90.77</u>	59.45	<u>1.52</u>	<b>75.83</b>	<u>48.88</u>	<b>62.58</b>	48.65	<u>86.31</u>	<b>28.82</b>	0.00	91.51
PointNet [37]	86.06	38.56	29.70	74.94	94.57	85.38	8.62	<u>13.42</u>	<u>16.47</u>	0.00	38.64	14.27	36.96	0.09	0.00	2.75
PointNet++ [38]	<u>89.46</u>	44.64	36.93	<u>77.68</u>	<u>97.28</u>	<u>91.95</u>	<u>54.59</u>	0.52	15.84	0.00	<u>42.08</u>	<u>29.00</u>	<u>67.71</u>	0.24	0.00	3.16
SPGraph [24]	<u>82.02</u>	24.83	20.70	<u>61.72</u>	88.26	<u>78.27</u>	8.29	0.00	0.00	0.00	0.64	1.87	30.00	0.00	0.00	0.00
KPConv [51]	<b>90.62</b>	<u>48.71</u>	<b>40.51</b>	<b>78.88</b>	<b>98.33</b>	<b>94.24</b>	<b>76.20</b>	0.01	14.70	0.00	41.77	<b>39.32</b>	<b>74.22</b>	<u>0.39</u>	0.00	<b>8.61</b>
RandLA-Net [23]	88.92	<b>51.57</b>	<u>40.29</u>	78.46	97.12	89.93	46.77	<b>28.76</b>	<b>20.03</b>	0.00	<b>46.98</b>	18.70	65.99	<b>24.91</b>	0.00	<u>6.15</u>

Table 6: All baselines are trained on the Birmingham split. The top five records show the testing results on the testing split of Birmingham, while the bottom five rows show the scores on the testing split of Cambridge. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

## ■ Challenge 5: Cross-dataset generalization

**Table 14** Quantitative cross-dataset generalization results were achieved by the selected baseline approaches on the proposed SensatUrban dataset and the DALES dataset.

Methods	Settings	OA(%)	mIoU(%)	Ground	Vegetation	Cars	Street furniture	Fences	Buildings
PointNet <a href="#">Qi et al. (2017a)</a>	DALES → DALES	94.10	59.72	94.68	86.69	16.48	73.62	0.00	86.87
	DALES → SensatUrban	74.25	30.75	89.44	55.69	0.02	0.03	0.00	39.32
	SensatUrban → SensatUrban	92.46	56.27	92.90	92.14	52.69	0.33	14.25	85.33
	SensatUrban → DALES	87.45	41.98	92.64	72.15	2.77	11.79	8.31	64.23
RandLA-Net <a href="#">Hu et al. (2020)</a>	DALES → DALES	96.98	84.31	96.99	92.71	80.54	89.08	50.09	96.47
	DALES → SensatUrban	83.69	40.69	93.02	64.03	0.25	0.23	16.63	69.96
	SensatUrban → SensatUrban	96.55	79.47	96.87	98.28	80.44	45.18	60.92	95.16
	SensatUrban → DALES	84.25	43.57	92.63	66.26	27.33	2.27	8.89	64.07

## ■ Challenge 5: Pre-Training

	OA(%)	mAcc(%)	mIoU(%)	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet-Rand [37]	86.29	53.33	45.10	80.05	93.98	87.05	<u>23.05</u>	19.52	41.80	3.38	43.47	<u>24.20</u>	63.43	26.86	0.00	<u>79.53</u>
PointNet-Jigsaw [44]	<u>87.38</u>	<b>56.97</b>	<u>47.90</u>	<u>83.36</u>	<u>94.72</u>	<u>88.48</u>	22.87	<u>30.19</u>	<u>47.43</u>	<u>15.62</u>	<u>44.49</u>	22.91	<u>64.14</u>	<b>30.33</b>	0.00	77.88
PointNet-OcCo [57]	<b>87.87</b>	<u>56.14</u>	<b>48.50</b>	<b>83.76</b>	<b>94.81</b>	<b>89.24</b>	<b>23.29</b>	<b>33.38</b>	<b>48.04</b>	<b>15.84</b>	<b>45.38</b>	<b>24.99</b>	<b>65.00</b>	<u>27.13</u>	0.00	<b>79.58</b>
PCN-Rand [66]	86.79	<u>57.66</u>	47.91	<u>82.61</u>	<b>94.82</b>	<u>89.04</u>	<b>26.66</b>	21.96	34.96	<u>28.39</u>	43.32	<u>27.13</u>	62.97	30.87	0.00	<u>80.06</u>
PCN-Jigsaw [44]	<b>87.32</b>	57.01	<u>48.44</u>	<b>83.20</b>	<u>94.79</u>	<b>89.25</b>	<u>25.89</u>	19.69	<b>40.90</b>	<b>28.52</b>	43.46	24.78	<u>63.08</u>	<u>31.74</u>	0.00	<b>84.42</b>
PCN-OcCo [57]	<u>86.90</u>	<b>58.15</b>	<b>48.54</b>	81.64	94.37	88.21	25.43	<b>31.54</b>	<u>39.39</u>	22.02	<b>45.47</b>	<b>27.60</b>	<b>65.33</b>	<b>32.07</b>	0.00	77.99
DGCNN-Rand [58]	87.54	60.27	51.96	83.12	95.43	89.58	<b>31.84</b>	35.49	45.11	<u>38.57</u>	45.66	<u>32.97</u>	64.88	30.48	0.00	<b>82.34</b>
DGCNN-Jigsaw [44]	<u>88.65</u>	<u>60.80</u>	<u>53.01</u>	<b>83.95</b>	<b>95.92</b>	89.85	<u>30.05</u>	<b>43.59</b>	<u>46.40</u>	35.28	<u>49.60</u>	31.46	<u>69.41</u>	<b>34.38</b>	0.00	80.55
DGCNN-OcCo [57]	<b>88.67</b>	<b>61.35</b>	<b>53.31</b>	<u>83.64</u>	<u>95.75</u>	<b>89.96</b>	29.22	<u>41.47</u>	<b>46.89</b>	<b>40.64</b>	<b>49.72</b>	<b>33.57</b>	<b>70.11</b>	<u>32.35</u>	0.00	79.74

Table 8: Quantitative results achieved by using OcCo [57], Jigsaw [44] and Random (Rand) initialization on the SensatUrban dataset, based on PointNet [37], PCN [66] and DGCNN [58] encoders. Note that, all the initialized weights are obtained by pre-training on the ModelNet40 [60], since these techniques are mainly designed for object-level classification and segmentation. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.



02

SensatUrban

ICCV 2021 workshop



# Urban3D 2021



1<sup>st</sup> International Workshop on Urban-Scale Point Clouds Understanding, at ICCV 2021



## Organizers



Qingyong Hu  
University of Oxford



Bo Yang  
The Hong Kong Polytechnic University



Sheikh Khalid  
Sensat Inc.



Ronald Clark  
Imperial College London



Wen Xiao  
Newcastle University



Yulan Guo  
National University of Defense Technology



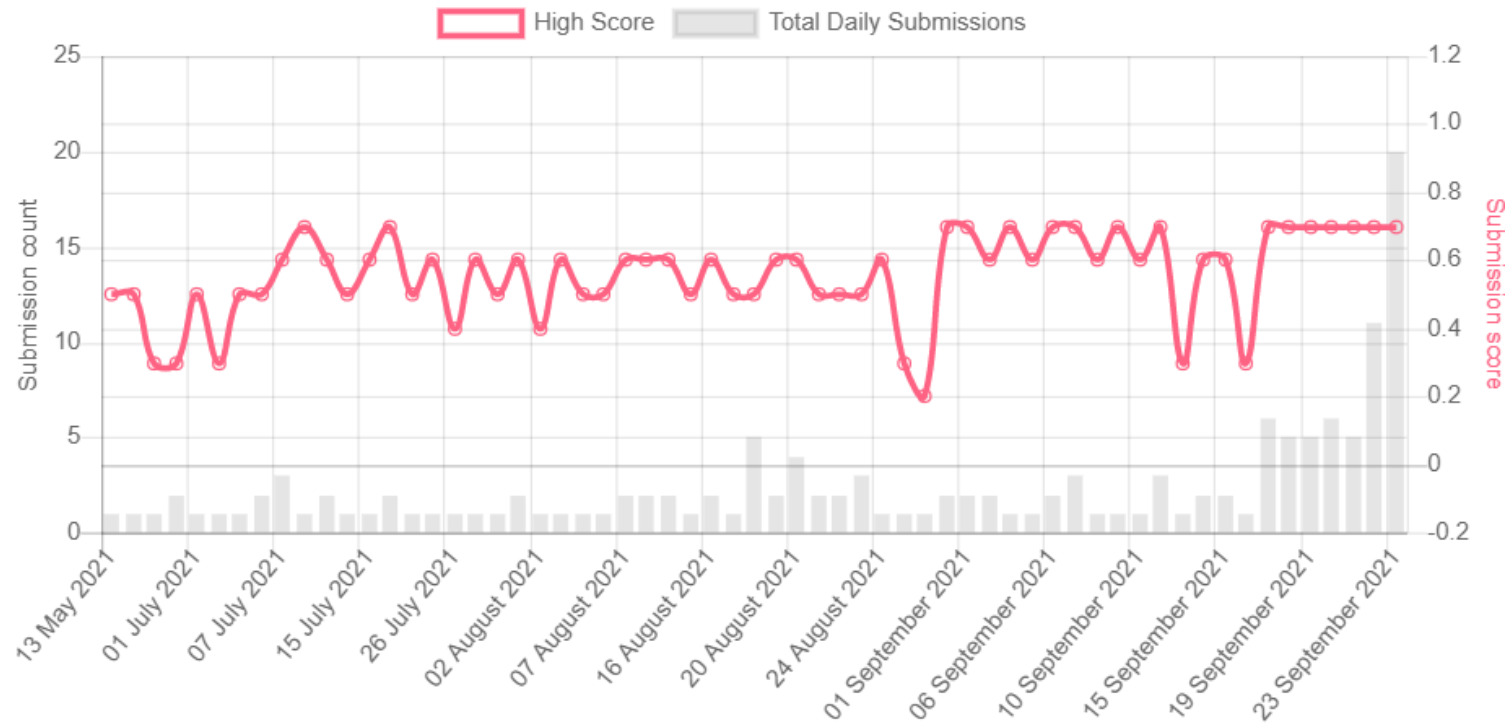
Aleš Leonardis  
University of Birmingham



Niki Trigoni  
University of Oxford



Andrew Markham  
University of Oxford



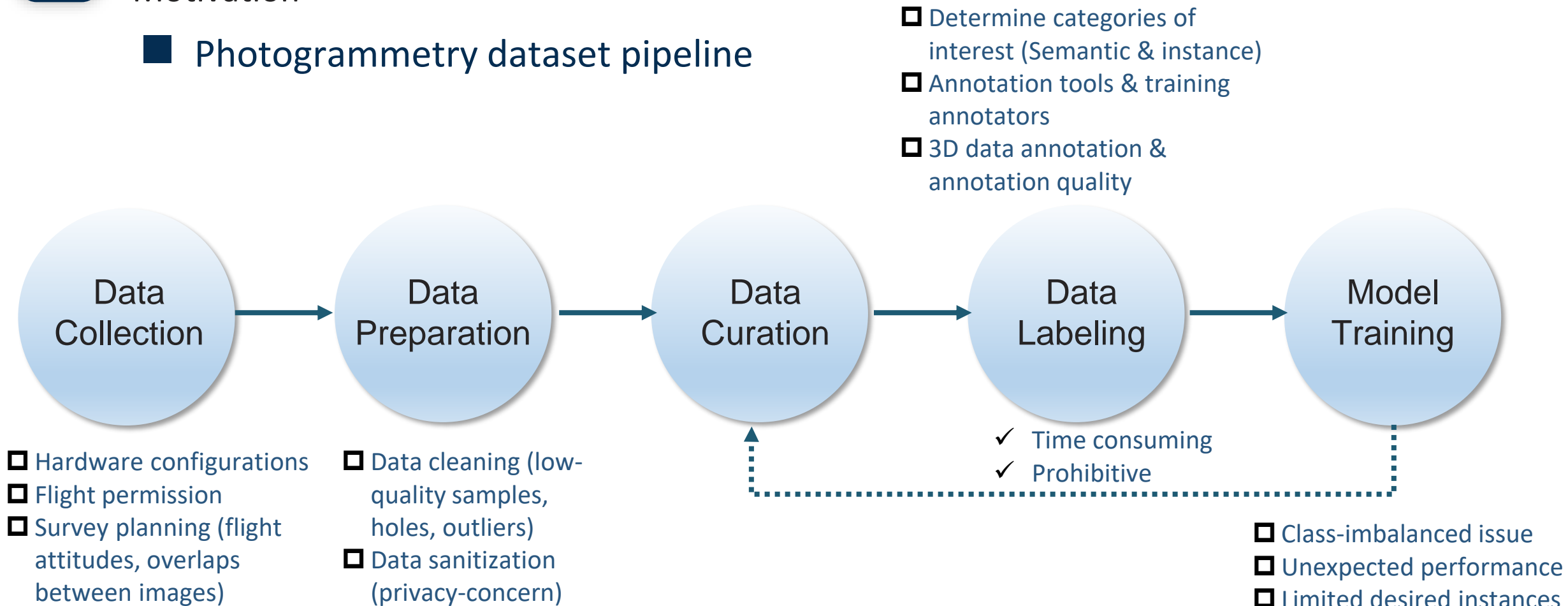
- 446 researchers from different institutes have registered to download our SensatUrban dataset
- 111/154 teams have successfully participated in our challenge in CodaLab
- Nearly 200 valid submissions are reported during the competition phase
- The top performed method has surpassed the baseline methods (KPCConv, RandLA-Net) by more than 15% in terms of mIoU

## Research Question 2

**How to achieve synthetic generation of  
urban-scale 3D scenes?**



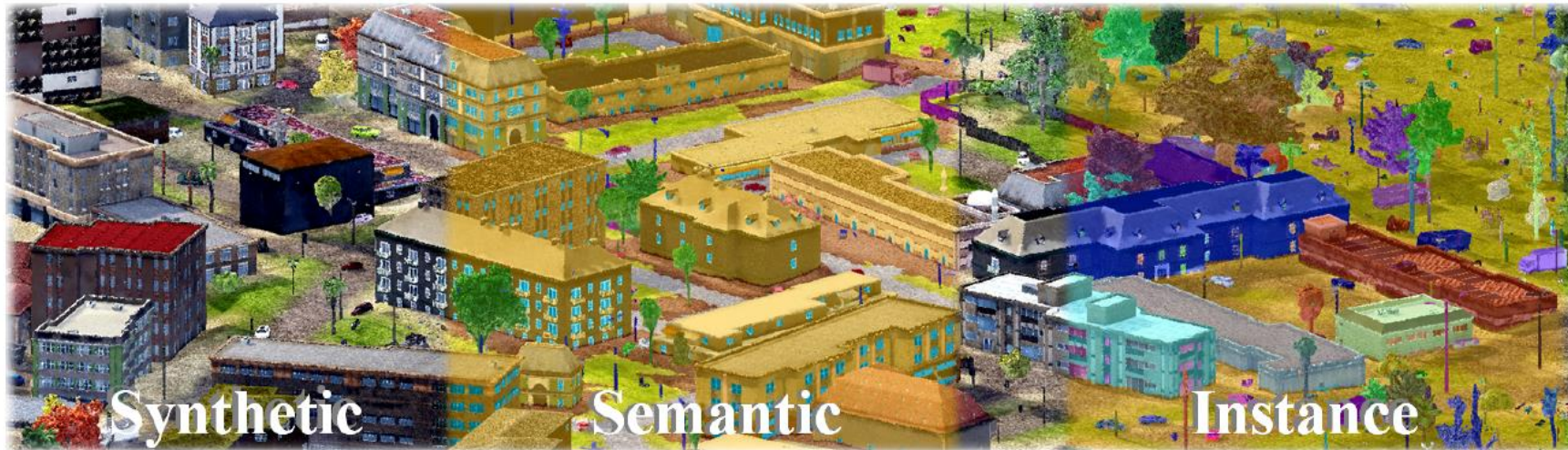
## ■ Photogrammetry dataset pipeline



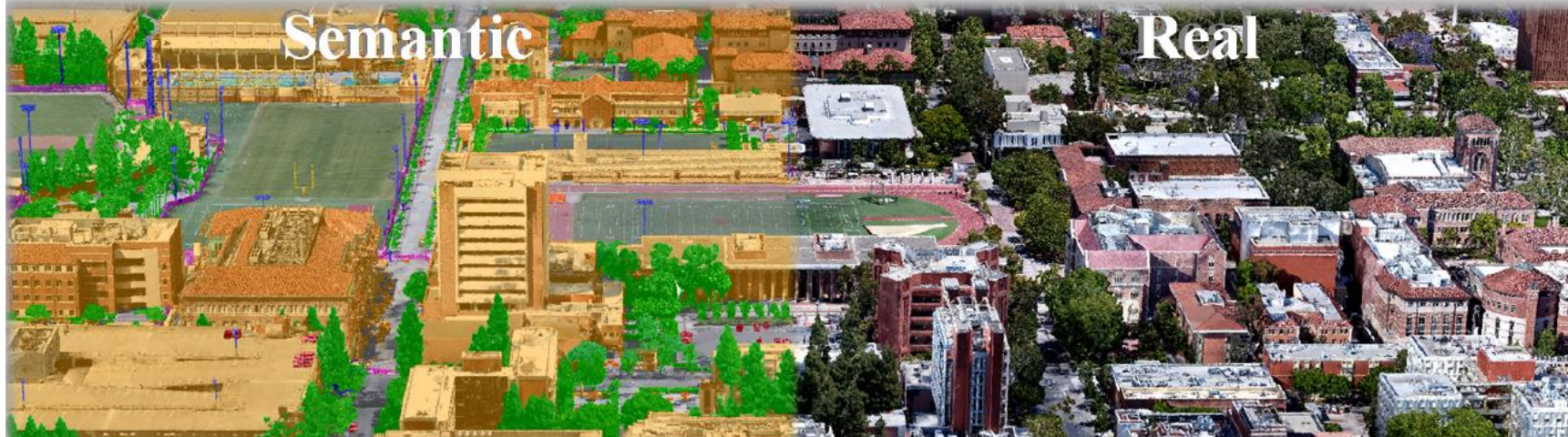
It is highly challenging for individuals to complete the whole pipeline of dataset creation!

- Equip individuals with the full capability of:
  - **Creating large-scale annotated photogrammetry datasets**
    - *Exploring open geospatial data sources*
    - *Leveraging off-the-shelf commercial packages*
  - **Controllable & Efficient & Low cost**
    - *Procedurally synthetic 3D data generation*
    - *Automatic annotation generation, avoiding time-consuming manual annotation*
  - **Realistic & Effective**
    - *Simulates the reconstruction process of the real environment*
    - *Following the same UAV flight pattern, ensure similar quality, noise pattern, and diversity*





VS.





Name and Reference	# Semantic	# Instance <sup>1</sup>	# Views / scenes	2D Annotations	Area <sup>2</sup> (km <sup>2</sup> )	Sensor
DublinCity [107]	13	No	8,504 / 2	No	2	
DALES [93]	8	No	1 large scene	-	10	Aerial LiDAR
LASDU [102]	5	No	1 scene	-	1.02	
Swiss3DCities [6]	5	No	3 scenes	No	2.7	quadcopter + photogrammetry
Campus3D [52]	14	4 classes	6 scenes	No	1.58	quadcopter + photogrammetry
SensatUrban [40]	13	No	3 scenes	No	4.4	fixed wing + photogrammetry
<b>STPLS3D - Real</b>	<b>6</b>	<b>No</b>	<b>16,376 / 4</b>	<b>Yes</b>	<b>1.27</b>	<b>quadcopter + photogrammetry</b>
<b>STPLS3D - SyntheticV1</b>	<b>5</b>	<b>No</b>	<b>17,164 / 14</b>	<b>Yes</b>	<b>4.22</b>	<b>Synthetic Aerial photogrammetry</b>
<b>STPLS3D - SyntheticV2</b>	<b>17</b>	<b>14 classes</b>	<b>13,229 / 24</b>	<b>Yes</b>	<b>5.76</b>	<b>Synthetic Aerial photogrammetry</b>
<b>STPLS3D - SyntheticV3</b>	<b>18</b>	<b>14 classes</b>	<b>15,888 / 25</b>	<b>Yes</b>	<b>6</b>	<b>Synthetic Aerial photogrammetry</b>

- Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset
- Synthetic V1-V3: **16 km<sup>2</sup>** of the city landscape, with up to **18** semantic classes and **14** instance classes
- Real Datasets: **1.27 km<sup>2</sup>** landscape, **6** semantic classes

## ■ Dataset Highlights

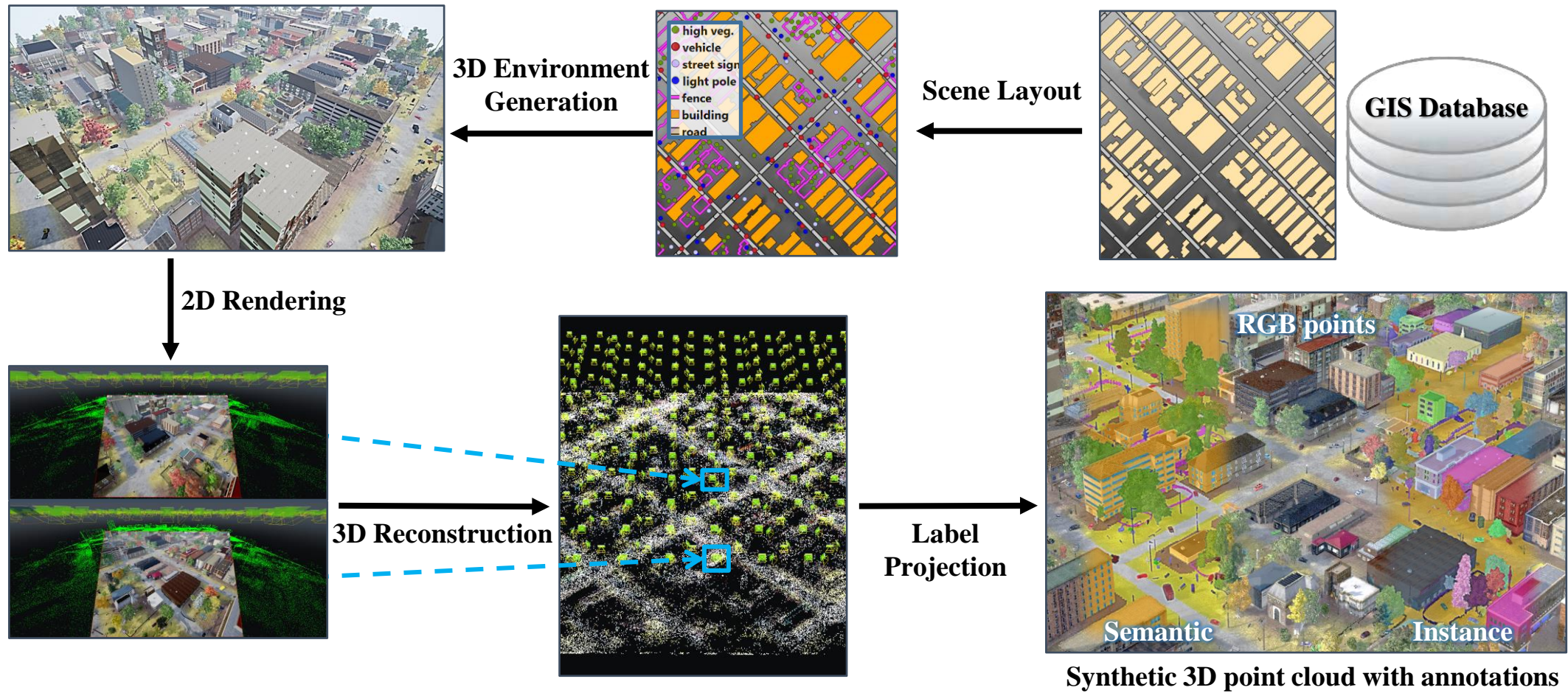
### ● Quality

- Fully exploits existing open geo-spatial data sources, compared with *limited gaming environments in virtual gaming engine-based generation*
- Leverage procedural modeling tools to create building models with variations and adapted different *material* databases to enrich the *diversity* for building appearances
- *Simulate similar UAV paths* over the virtual terrain as the real-world survey
- Up to *18* different semantic annotations + point-wise *instance labels*

### ● Scalability

- Synthetic environments were procedurally generated with *great flexibility and scalability*
- Freely changing scene layouts, object materials, architectural models
- Explicitly balance the class distribution by heuristically placing 3D models





03

## STPLS3D

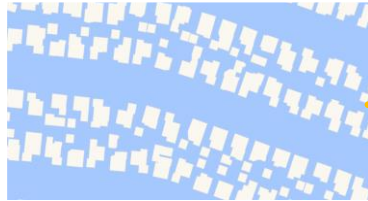
Data Generation Pipeline

■ Demo

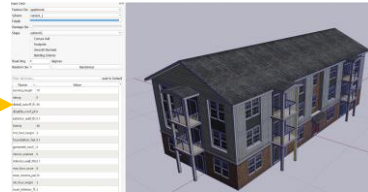
*Synthetic aerial photogrammetry  
point cloud generation pipeline*



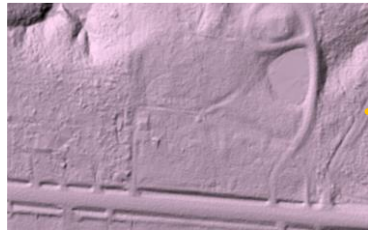
## ■ 3D Scene Generator



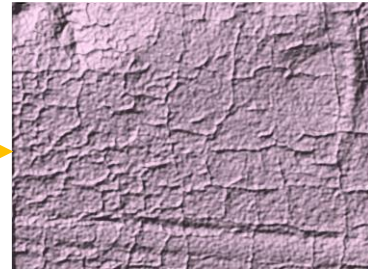
OSM building footprints



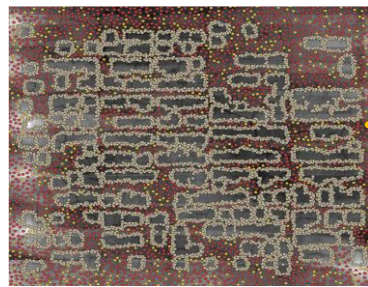
PMGS



DSM



Adding details

Generated  
Object Positions

Game Objects



Procedural City

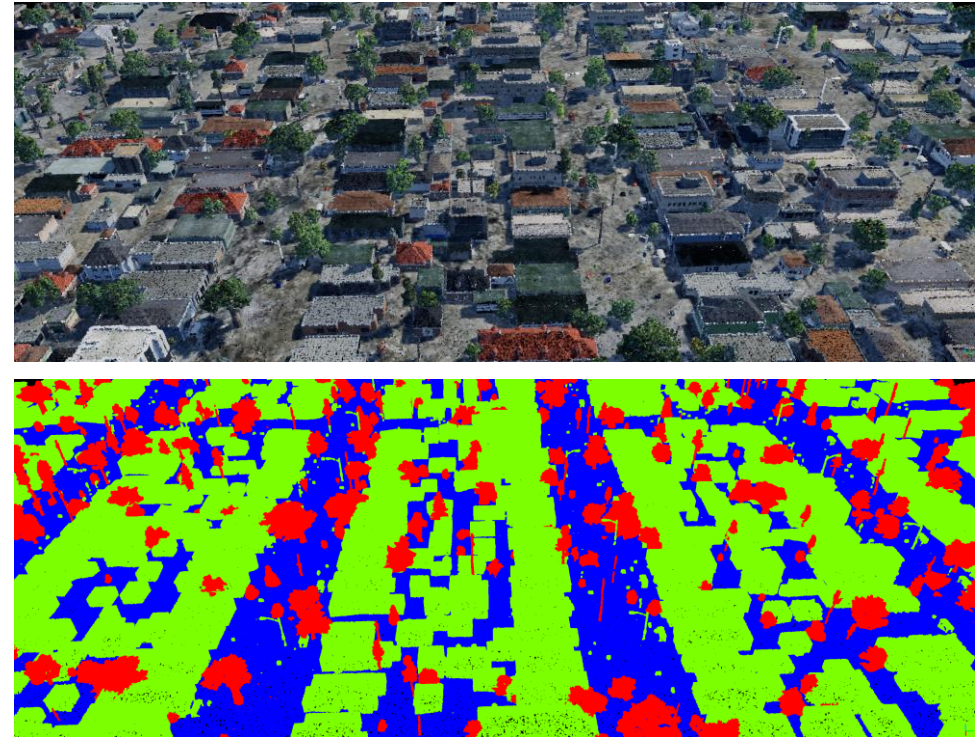
- 2D Rendering Engine/Simulator (AirSim)



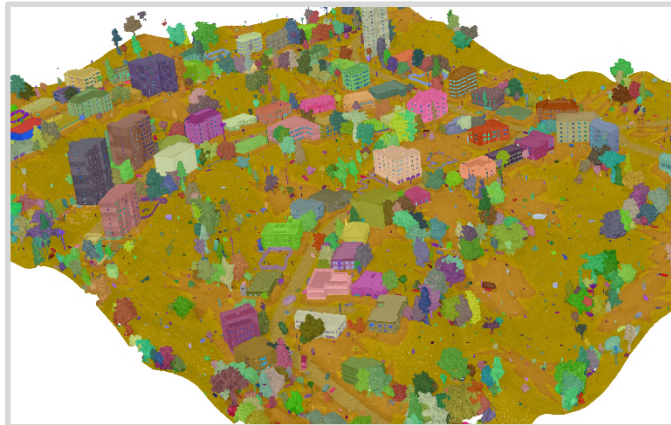
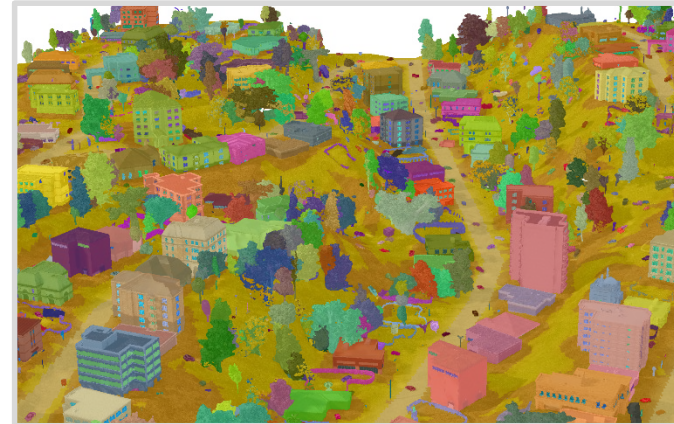
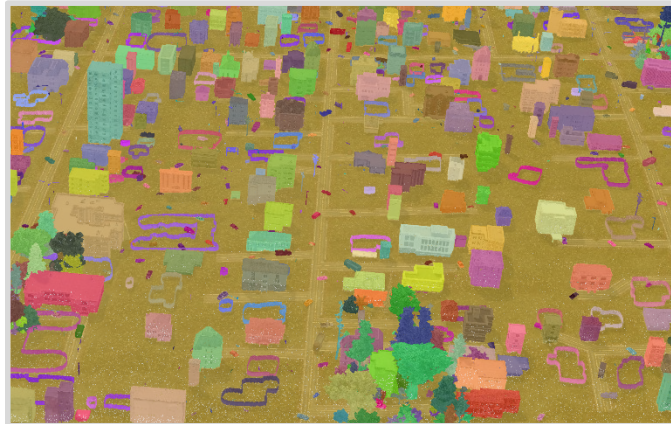


- Photogrammetric point clouds with annotations

Rendered RGB  
images  
+  
ContextCapture  
+  
Ray casted point  
cloud (for label)

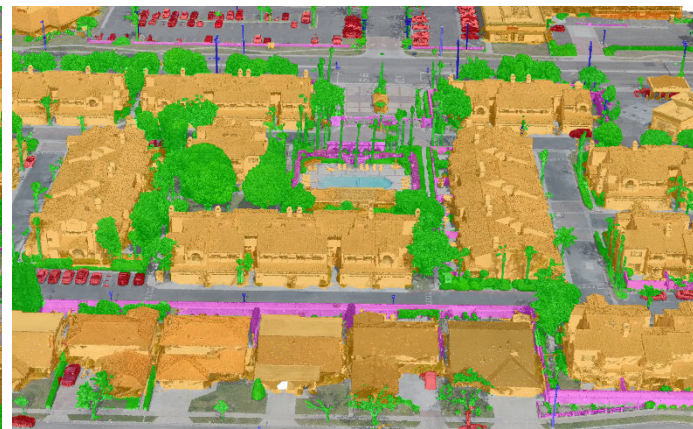
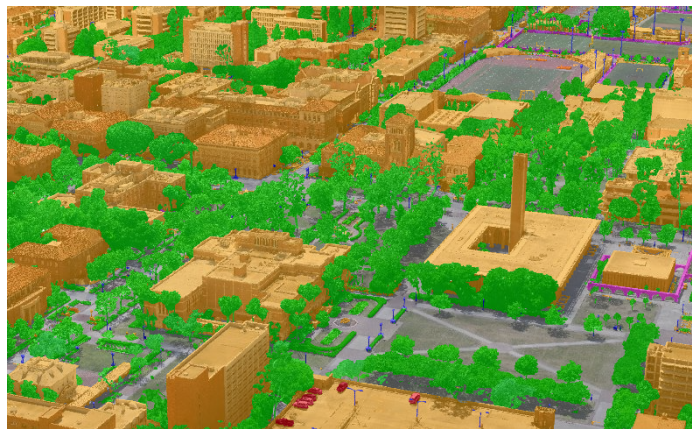
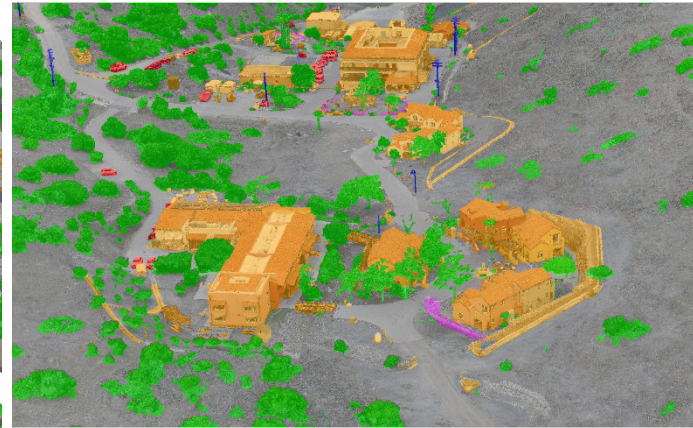
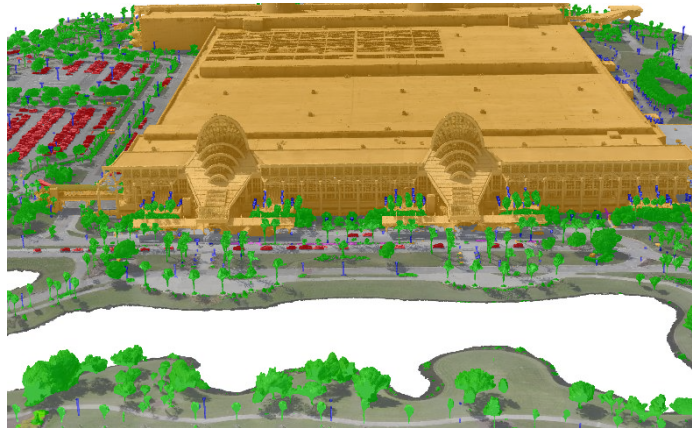


## ■ Synthetic Subsets

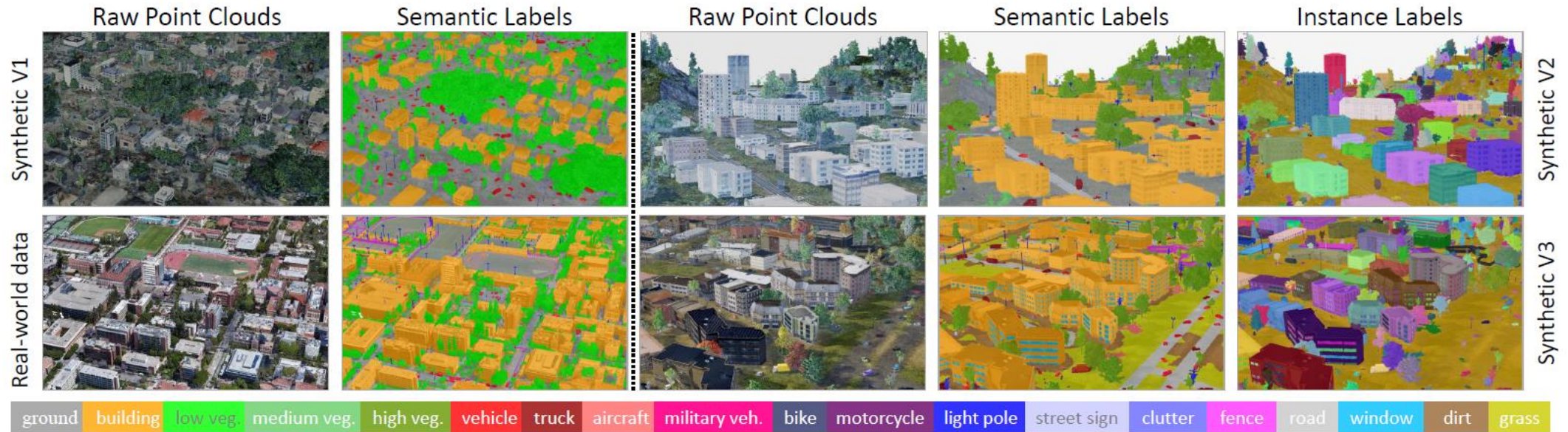




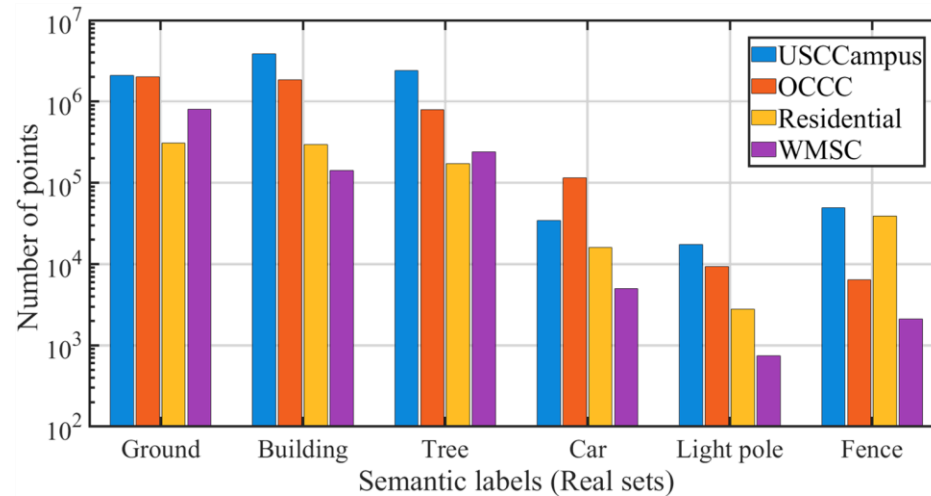
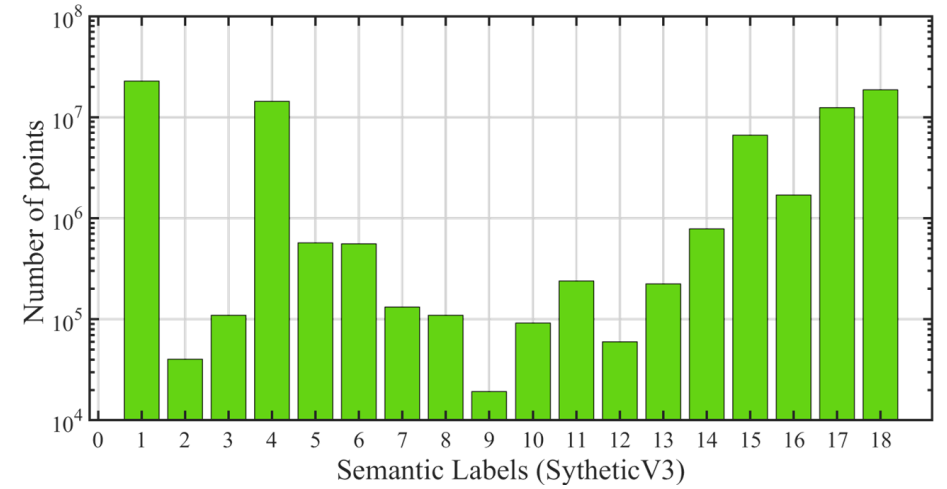
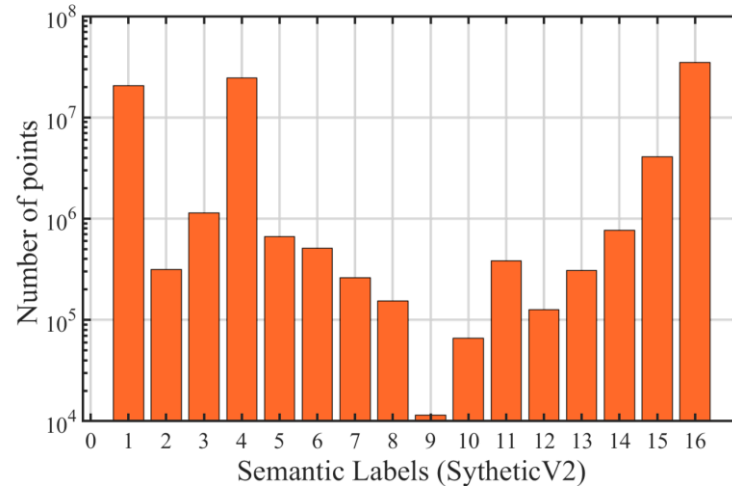
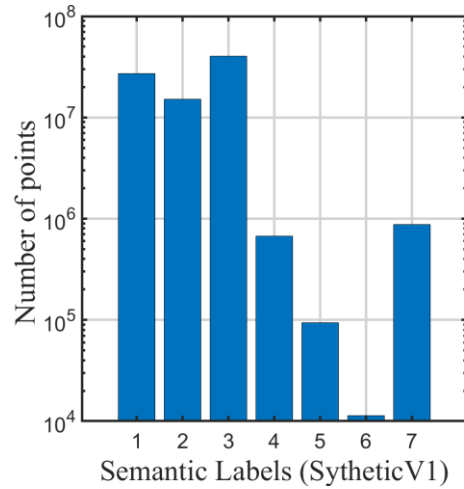
## ■ Real-World Subsets



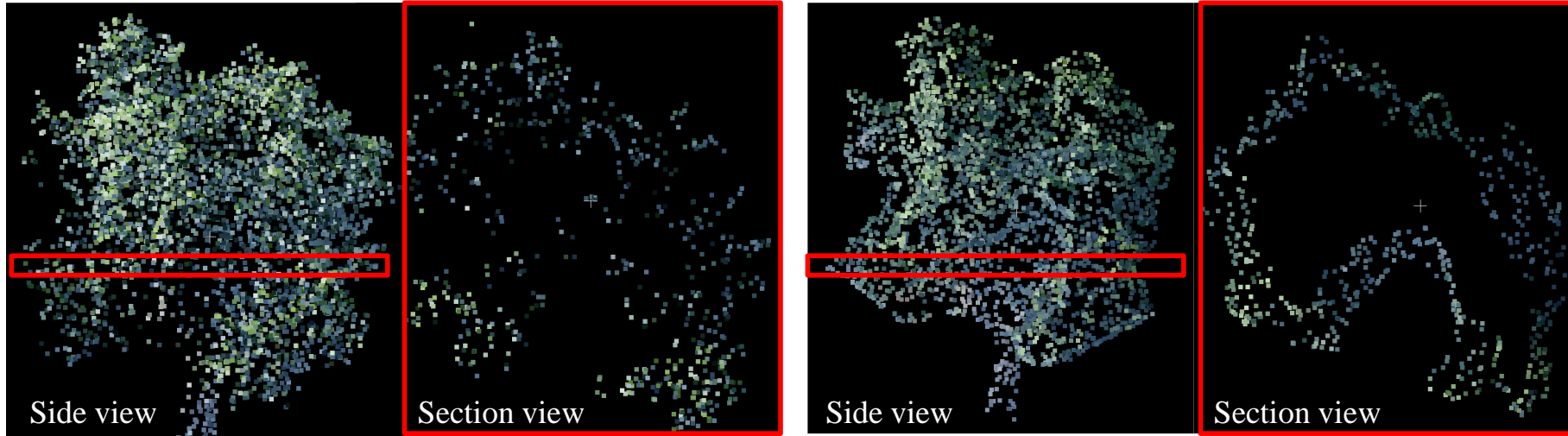




## Data Distribution







Ray casted point clouds

Photogrammetric reconstructed  
point clouds

*Comparison of Data Quality*



**Real-World Subset:**

- Spatial area:  $1.27 \text{ km}^2$
- Cost: **Over four months of team efforts**
  - ✓ Getting flight permits
  - ✓ Planning
  - ✓ Repeatedly executing the data collection process
  - ✓ Data cleaning, sanitization
  - ✓ Semantic & instance labeling

**Synthetic Subset:**

- Spatial area: over  $16 \text{ km}^2$
- Cost: **Single person within a month efforts**
  - ✓ A desktop PC
  - ✓ Intel Core™ i9-10900X CPU
  - ✓ NVIDIA RTX 3090
  - ✓ Can be parallel accelerated
  - ✓ Not constrained by workforce talent

## ■ Semantic Segmentation Results

Training sets	Methods	mIoU (%)	oAcc (%)	Per Class IoU (%)					
				Ground	Building	Tree	Car	Light pole	Fence
Real subsets	RandLA-Net [42]	42.33	60.19	46.13	24.23	<b>72.46</b>	53.37	44.82	12.95
	SCF-Net [25]	<b>45.93</b>	<b>75.75</b>	<b>68.77</b>	<b>37.27</b>	65.49	51.50	31.22	<b>21.34</b>
	KPConv [89]	45.22	70.67	60.87	32.13	69.05	<b>53.80</b>	<b>52.08</b>	3.40
Synthetic subsets	RandLA-Net [42]	45.03	81.30	76.78	57.74	56.08	28.44	40.36	10.78
	SCF-Net [25]	47.82	82.69	77.51	68.68	56.81	<b>29.87</b>	<b>42.53</b>	11.52
	KPConv [89]	<b>49.16</b>	<b>88.08</b>	<b>85.50</b>	<b>70.65</b>	<b>63.84</b>	28.75	32.97	<b>13.22</b>
Real+Synthetic	RandLA-Net [42]	50.53	86.25	82.90	66.59	63.77	33.91	<b>41.84</b>	14.19
	SCF-Net [25]	50.65	83.32	77.80	58.98	64.86	<b>46.37</b>	40.50	<b>15.41</b>
	KPConv [89]	<b>53.73</b>	<b>89.87</b>	<b>87.40</b>	<b>78.51</b>	<b>66.18</b>	39.63	41.30	9.34

- Baselines: RandLA-Net, SCF-Net, KPConv
- Mapping to 6 unified semantic classes
- Testing set: real-world test set (WMSC)

## ■ Instance Segmentation Results

	Metric	mean (%)	Build.	LowVeg.	MediumVeg.	HighVeg.	Vehicle	Truck	Aircraft	MilitaryVeh.	Bike	Motorcycle	LightPole	StreetSign	Clutter	Fence
HAIS[15]	AP	<b>40.4</b>	68.7	29.8	23.9	25.4	78.8	59.2	47.4	37.9	13.0	59.3	56.4	10.2	23.0	32.4
	AP50	<b>51.9</b>	73.2	46.4	34.5	29.8	89.0	69.3	66.7	48.1	24.3	76.4	70.5	16.7	28.4	53.0
	AP25	<b>57.3</b>	74.2	56.0	42.8	32.0	91.2	76.1	73.8	51.9	26.4	82.2	75.6	18.3	32.1	69.0
PointGroup[46]	AP	27.4	62.3	17.3	18.2	20.4	62.8	47.0	31.0	24.3	4.5	19.1	27.7	9.2	15.1	24.1
	AP50	44.2	71.1	36.8	30.3	26.0	87.4	67.3	50.0	40.7	12.6	60.3	54.2	15.6	20.8	45.9
	AP25	54.2	73.9	50.6	37.9	29.9	91.4	71.9	61.9	50.6	21.7	80.2	75.2	18.0	23.8	72.0

- Baselines: HAIS, PointGroup
- Selected 14 instance classes
- Training set: 20 synthetic data from V3; Testing set: 5 synthetic data from V3





# Urban3D 2022

2<sup>nd</sup> International Workshop on Urban-Scale Point Clouds Understanding, at ECCV 2022



ECCV 2022

Tel-Aviv  
Oct. 23-27 2022



- 1<sup>st</sup> place: \$1500
- 2<sup>nd</sup> place: \$1000
- 3<sup>rd</sup> place: \$500
- Invited presentation at ECCVW 2022

## Organizers



Qingyong Hu  
University of Oxford



Meida Chen  
University of Southern California - Institute for Creative Technologies



Ta-Ying Cheng  
University of Oxford



Sheikh Khalid  
Sensat LTD.



Bo Yang  
The Hong Kong Polytechnic University



Ronald Clark  
Imperial College London



Jiahui Chen  
Sun Yat-sen University



Leying Zhang  
Sun Yat-sen University



Rongkun Yang  
Sun Yat-sen University



Yulan Guo  
National University of Defense Technology



Aleš Leonardis  
University of Birmingham



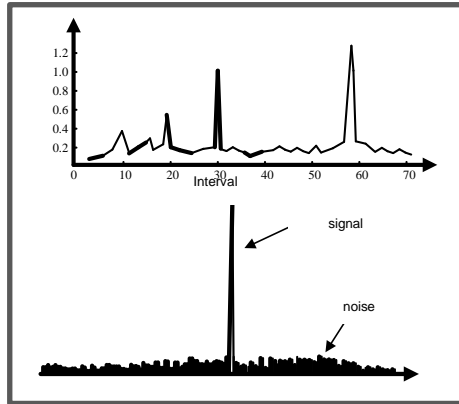
Niki Trigoni  
University of Oxford



Andrew Markham  
University of Oxford

## Research Question 3

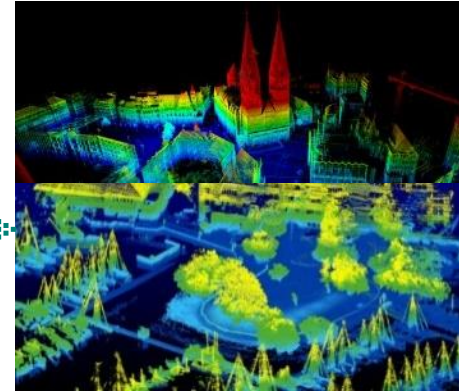
**How to achieve label-efficient learning of urban-scale 3D scenes?**



1D signal: KB



2D image: MB



3D point clouds: GB

Higher Resolution  
Richer Information



Massive Data  
Expensive Labeling Efforts



### ■ Statistics

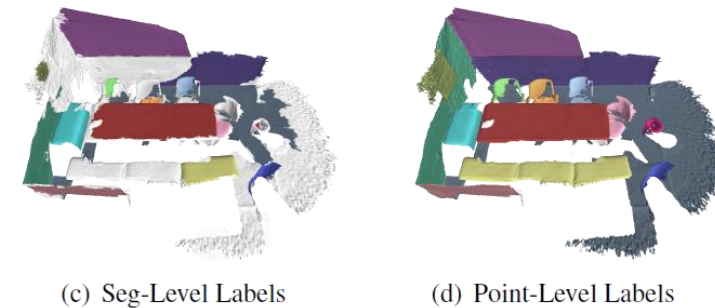
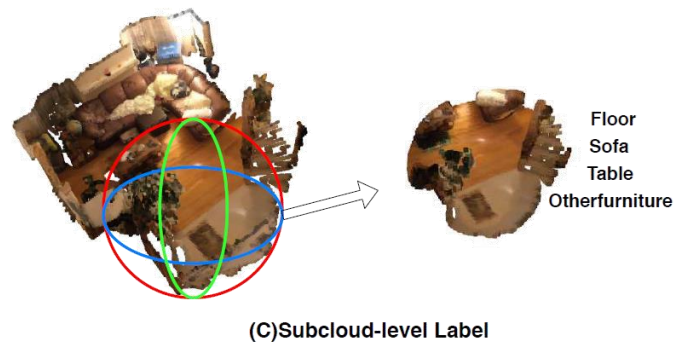
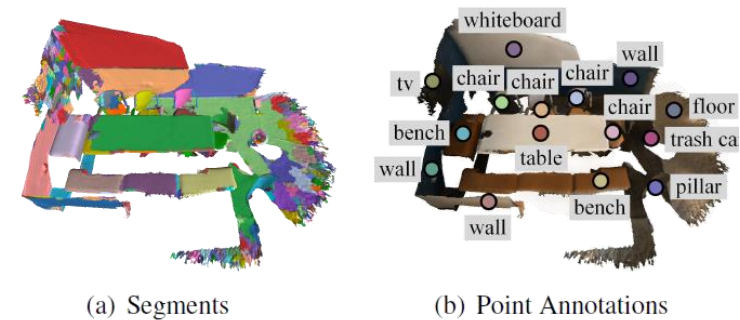
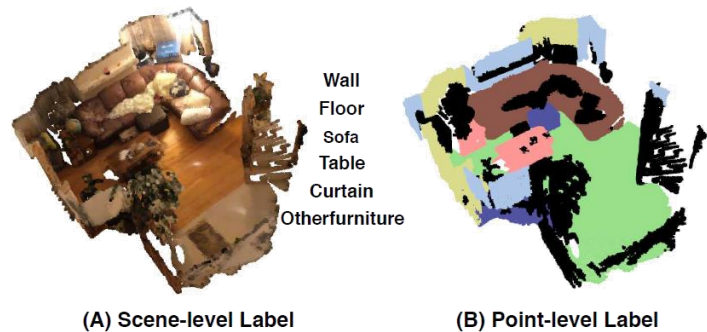
- A single vehicle with LiDAR captures **84 billion** points per day
- It takes more than **1700 hours** to annotate the SemanticKITTI dataset (4 billion points)
- It takes around **22.3 minutes** to annotate **a single indoor scene** (5m×5m×2m) in ScanNet, even with the oversegmentation preprocessing to reduce labeling cost

### ■ Goal

- Reducing the **labeling efforts for large-scale point clouds with billions of points**

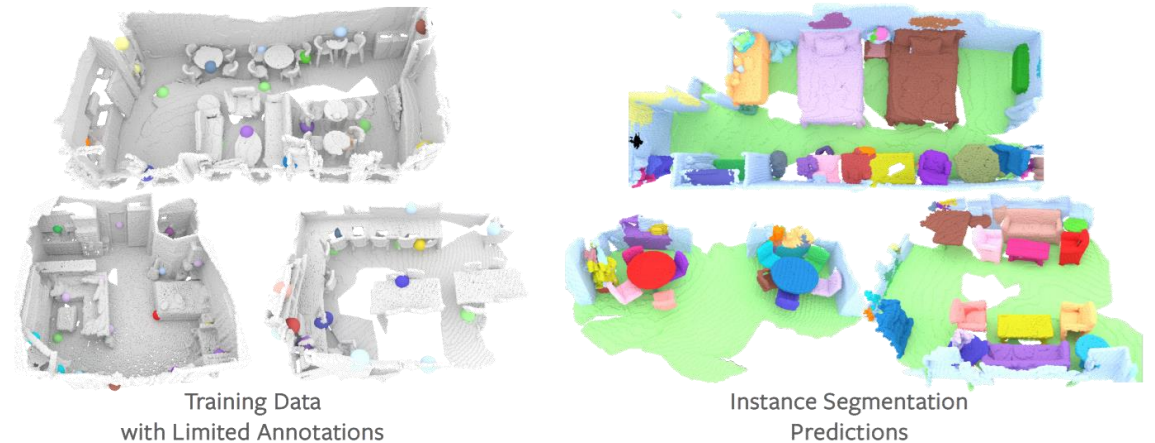
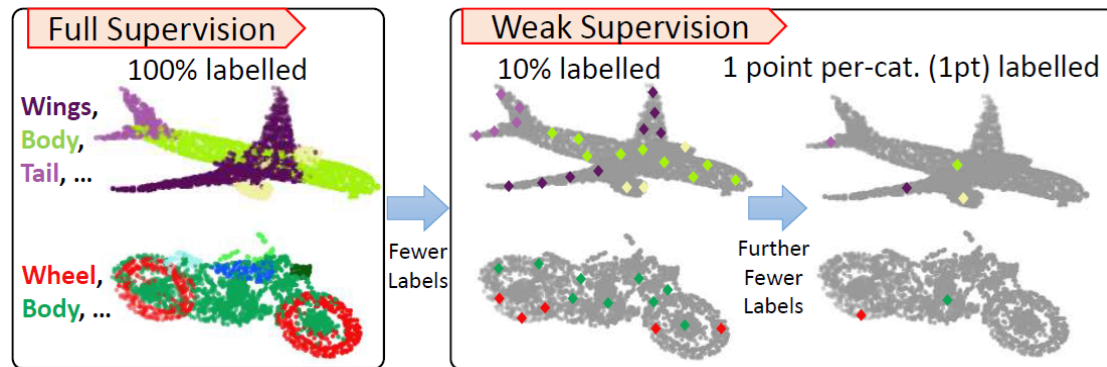
### ■ Limited Indirect Annotations

- Generating pseudo labels from **indirect scene-level tags/seg-level/sub-cloud labels/2D image labels** (MPRM'CVPR20, SegGroup'Arxiv21, BMVC19)



### ■ Limited Point Annotations

- **Approximating gradients with fewer 3D labels** (10x Fewer labels'CVPR20)
- **Contrastive pretraining followed by fine-tuning with fewer labels**  
(PointContrast'ECCV20, DepthContrast'Arxiv21, P4Contrast'Arxiv21)





### ■ Limitations

- Existing approaches adopt **custom methods and proportions of labels** for training (10%/5%/1% of raw points or superpoints), making fair comparison infeasible.
- Existing pipelines usually involve **multiple stages** including careful data augmentation, self-pretraining, fine-tuning, and/or post-processing such as the use of dense CRF.
- The **strong local semantic homogeneity of point neighbors** in large-scale point clouds is not fully exploited yet.

### ■ Questions

- Whether, and how, do existing fully-supervised methods perform **given different amounts of annotated data for training?**
- Given fewer and fewer labels, **where the weakly supervised regime actually begins?**

### ■ Weakly-Supervised Setting

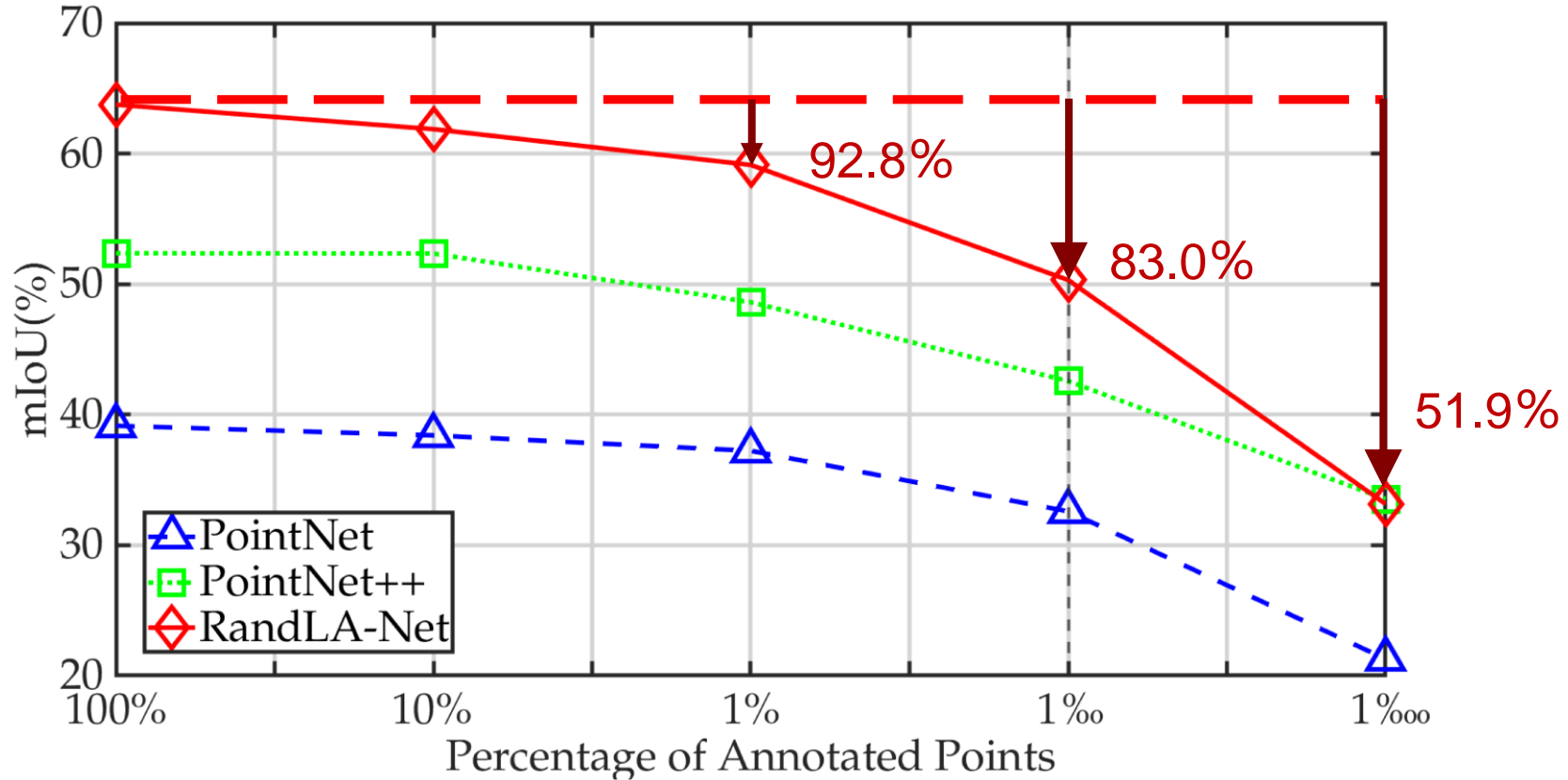


Input Point Clouds

Random Sparse Annotation

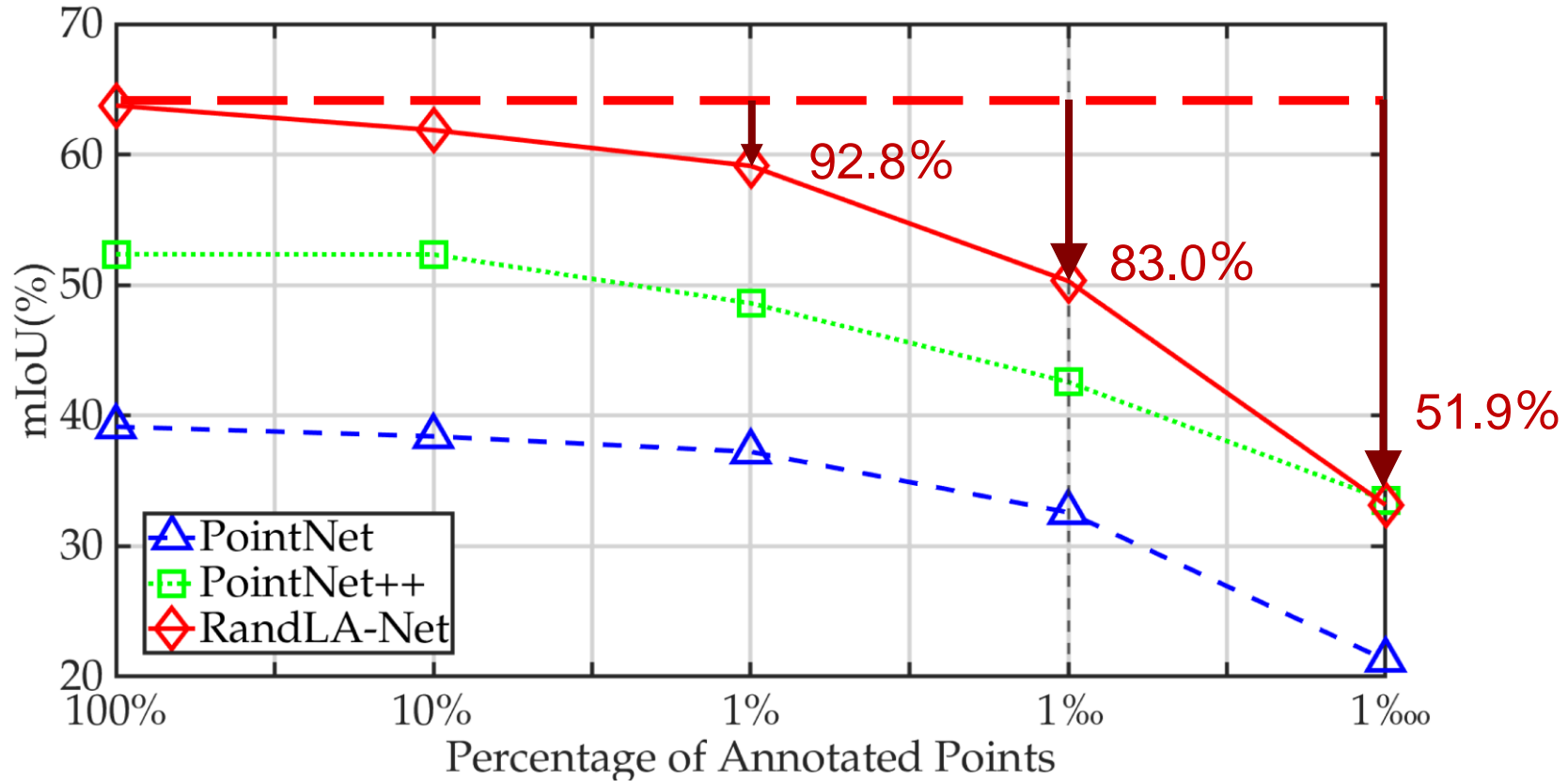


### ■ Benchmarking on the S3DIS dataset



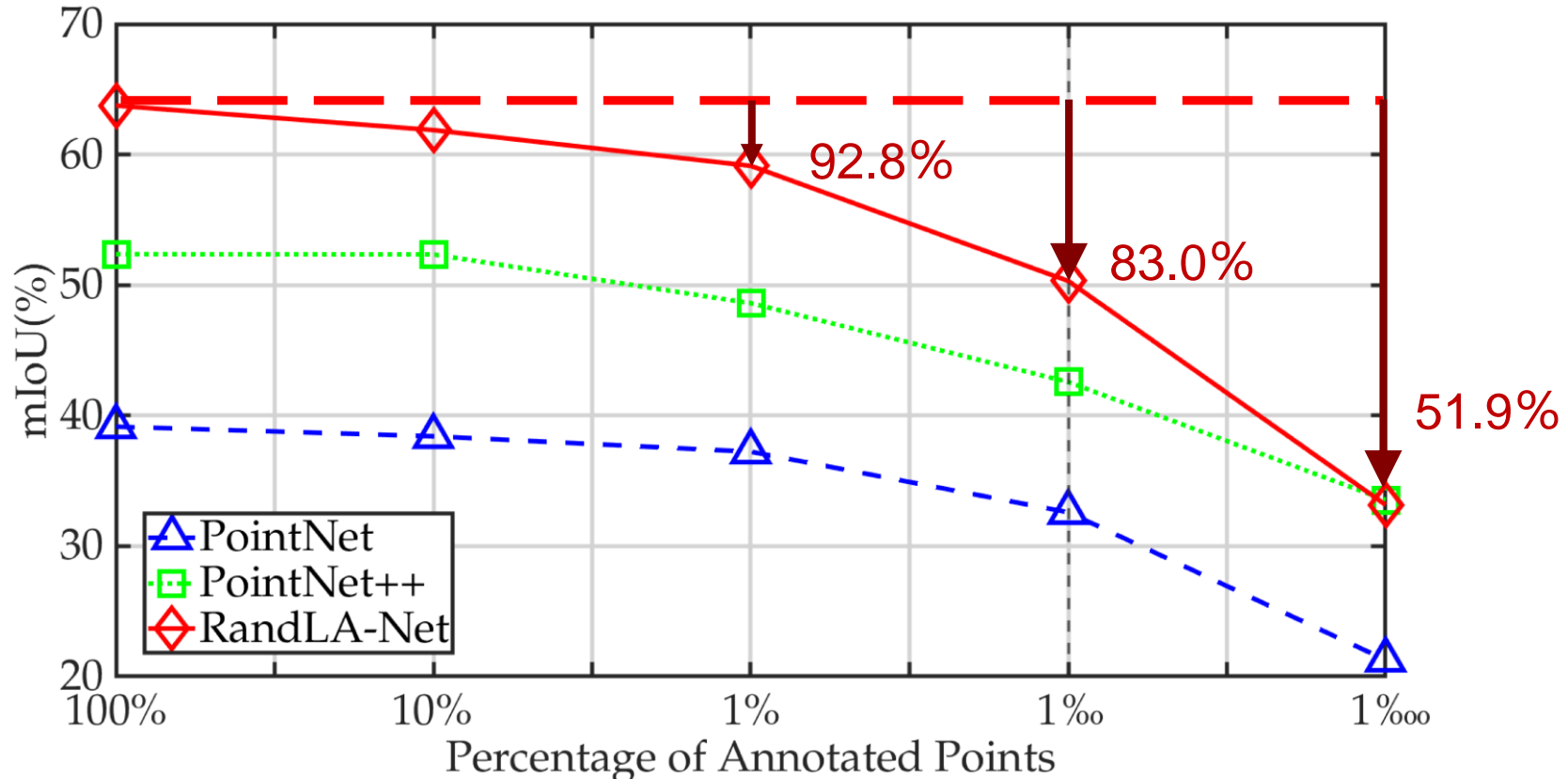
- **Dense annotations** are actually unnecessary to obtain a comparable and favorable segmentation accuracy.

### ■ Benchmarking on the S3DIS dataset



- This critical point (**1‰**) indicates that keeping a certain amount of training signals is also essential for weak supervision.

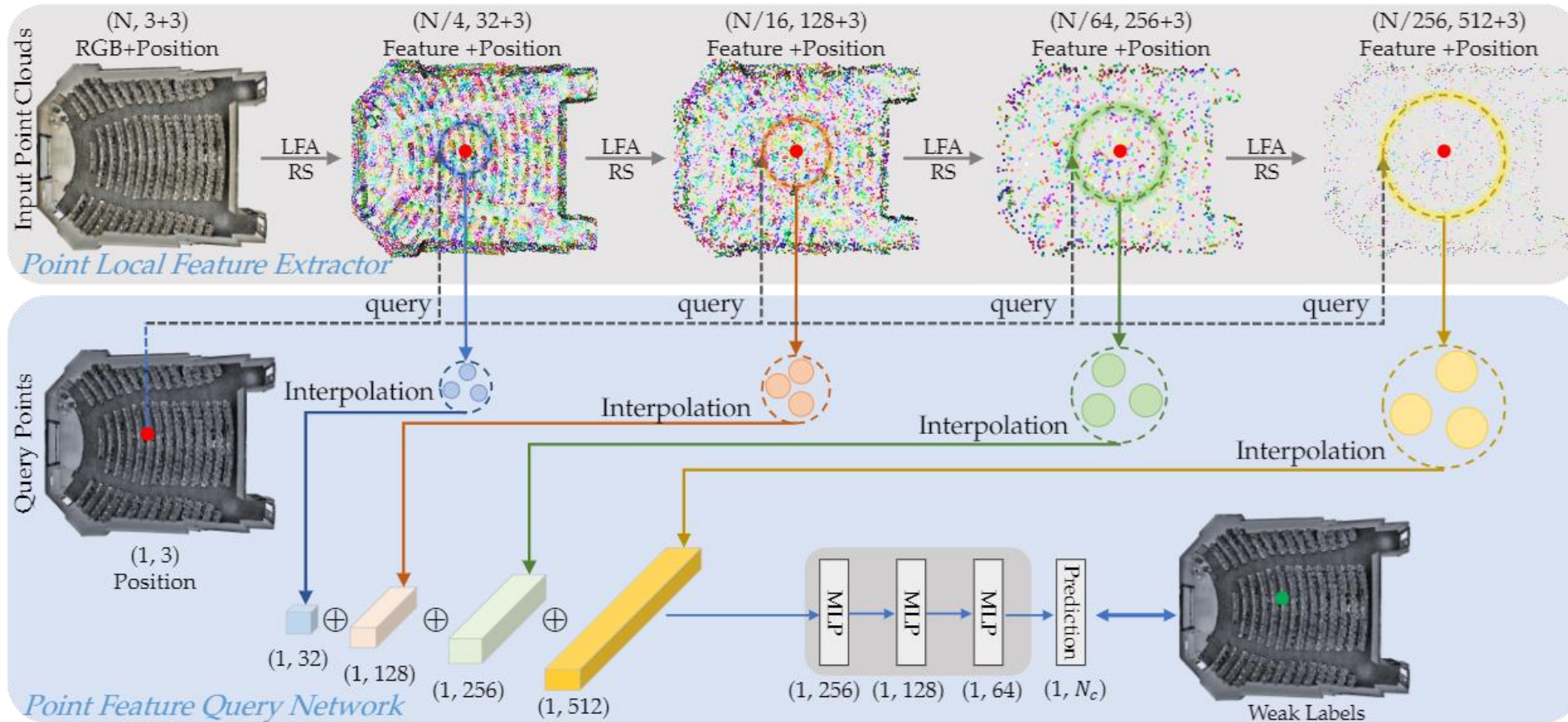
### ■ Benchmarking on the S3DIS dataset



- **The new question:** Given extremely limited point annotations (*e.g.*, 0.1%), how to **fully utilize the sparse yet valuable training signals** to update the network parameters?



### ■ SQN Architecture



### ■ Why Semantic Query Network?

- Training with **limited annotation**
  - The query point is assumed to share similar semantic information with the collected point features, such that the training signals from the query points can be **shared and back-propagated to the relevant points**.
- Flexible
  - The **query point can be arbitrary points in 3D space**, even not within the input point clouds. This allows training in incomplete point clouds, testing in complete point clouds.
- Novel
  - Without using the mature **U-Net architecture and skip connection**
  - Memory & computationally efficient, Lightweight

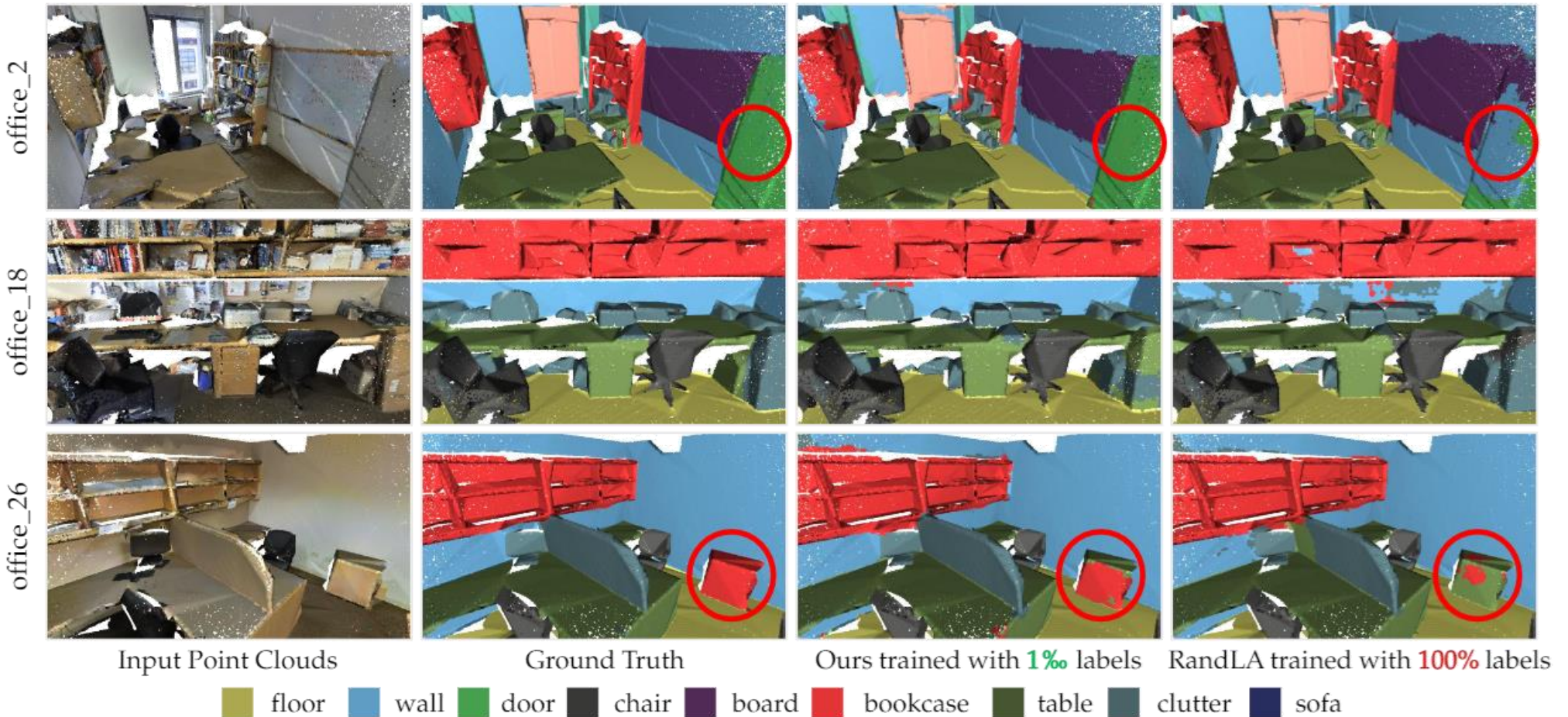
## S3DIS

	Methods	mIoU(%)	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
Full supervision	PointNet [46]	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	58.9	52.6	5.9	40.3	26.4	33.2
	PointCNN [34]	57.3	92.3	98.2	79.4	0.0	17.6	22.8	62.1	74.4	80.6	31.7	66.7	62.1	56.7
	SPGraph [31]	58.0	89.4	96.9	78.1	0.0	<u>42.8</u>	48.9	61.6	<u>84.7</u>	75.4	69.8	52.6	2.1	52.2
	SPH3D [33]	59.5	<u>93.3</u>	97.1	81.1	0.0	33.2	45.8	43.8	<u>79.7</u>	86.9	33.2	71.5	54.1	53.7
	PointWeb [96]	60.3	92.0	<u>98.5</u>	79.4	0.0	21.1	59.7	34.8	76.3	88.3	46.9	69.3	64.9	52.5
	RandLA-Net [24]	63.0	92.4	96.7	80.6	0.0	18.3	<u>61.3</u>	43.3	77.2	85.2	<u>71.5</u>	71.0	<u>69.2</u>	52.3
	KPConv rigid [65]	<u>65.4</u>	92.6	97.3	<u>81.4</u>	0.0	16.5	54.5	<u>69.5</u>	80.2	<u>90.1</u>	66.4	<u>74.6</u>	63.7	<u>58.1</u>
Limited superpoint labels <sup>†</sup>	ITIC (0.02%) [37]	50.1	-	-	-	-	-	-	-	-	-	-	-	-	-
	SSPC-Net (0.01%) [9]	51.5	-	-	-	-	-	-	-	-	-	-	-	-	-
Limited point-wise labels	II Model (10%) [30]	46.3	91.8	97.1	73.8	0.0	5.1	42.0	19.6	67.2	66.7	47.9	19.1	30.6	41.3
	MT (10%) [60]	47.9	92.2	96.8	74.1	0.0	10.4	46.2	17.7	70.7	67.0	50.2	24.4	30.7	42.2
	Xu (10%) [84]	48.0	90.9	97.3	74.8	0.0	8.4	49.3	27.3	71.7	69.0	53.2	16.5	23.3	42.8
	Zhang et al. (1%) [90]	61.8	91.5	96.9	80.6	0.0	18.2	58.1	47.2	75.8	85.7	65.2	68.9	65.0	50.2
	PSD (1%) [91]	63.5	92.3	97.7	80.7	0.0	27.8	56.2	62.5	78.7	84.1	63.1	70.4	58.9	53.2
	II Model (0.2%) [30]	44.3	89.1	97.0	71.5	0.0	3.6	43.2	27.4	63.1	62.1	43.7	14.7	24.0	36.7
	MT (0.2%) [60]	44.4	88.9	96.8	70.1	0.1	3.0	44.3	28.8	63.7	63.6	47.7	15.5	23.0	35.8
	Xu (0.2%) [84]	44.5	90.1	97.1	71.9	0.0	1.9	47.2	29.3	64.0	62.9	42.2	15.9	18.9	37.5
	RandLA-Net (0.1%)	52.9	89.9	95.9	<u>75.3</u>	0.0	<u>7.5</u>	52.4	26.5	62.2	74.5	49.1	60.2	49.3	45.1
	<b>Ours (0.1%)</b>	<b>61.4</b>	<b>91.7</b>	95.6	<b>78.7</b>	0.0	<b>24.2</b>	<b>55.8</b>	<b>63.1</b>	<b>70.5</b>	<b>83.1</b>	<b>60.6</b>	<b>67.8</b>	<b>56.1</b>	<b>50.6</b>

Table 1. Quantitative results of different methods on the *Area-5* of S3DIS dataset. Mean IoU (mIoU, %), and per-class IoU (%) scores are reported. Note that, <sup>†</sup>The ratios of these methods represent the ratio of super-point annotations and cannot be directly compared. Bold represents the best result in weakly setting and underlined represents the best in fully setting.



### S3DIS





### ■ Qualitative results



Input Point Clouds

Semantic Predictions

### ■ ScanNet

Settings	Methods	mIoU(%)
Full supervision	PointNet++ [47]	33.9
	SPLATNet [55]	39.3
	TangentConv [61]	43.8
	PointCNN [34]	45.8
	PointConv [81]	55.6
	SPH3D-GCN [33]	61.0
	KPConv [65]	68.4
	RandLA-Net [24]	64.5
Weak supervision	MPRM* [76]	41.1
	Zhang <i>et al.</i> (1%) [90]	51.1
	PSD (1%) [91]	54.7
	<b>Ours (0.1%)</b>	<b>56.9</b>

### ■ Semantic3D

	Methods	<i>Semantic8</i>		<i>Reduced8</i>	
		OA(%)	mIoU(%)	OA(%)	mIoU(%)
Full sup.	SnapNet [4]	91.0	67.4	88.6	59.1
	PointNet++ [51]	85.7	63.1	-	-
	ShellNet [102]	-	-	93.2	69.3
	GACNet [77]	-	-	91.9	70.8
	RGNet [71]	90.6	72.0	94.5	74.7
	SPG [35]	92.9	76.2	94.0	73.2
	KPConv [70]	-	-	92.9	74.6
	ConvPoint [6]	93.4	76.5	-	-
	WreathProdNet [79]	94.6	<u>77.1</u>	-	-
	RandLA-Net [28]	<u>95.0</u>	<u>75.8</u>	<u>94.8</u>	<u>77.4</u>
Weak sup.	Zhang <i>et al.</i> (1%) [97]	-	-	-	72.6
	PSD (1%) [98]	-	-	-	75.8
	<b>Ours (0.1%)</b>	<b>94.8</b>	<b>72.3</b>	<b>93.7</b>	<b>74.7</b>
	<b>Ours (0.01%)</b>	91.9	58.8	90.3	65.6

- Sub-cloud labels: **Labeling on the fly**
- Sparse annotation: **one-pass labeling at the beginning**, more friendly



## ■ DALES &amp; SensatUrban &amp; Toronto3D &amp; SemanticKITTI

Settings	Methods	DALES [67]		SensatUrban [23]			Toronto3D [57]		SemanticKITTI [3]
		OA(%)	mIoU(%)	OA(%)	mAcc (%)	mIoU(%)	OA(%)	mIoU(%)	mIoU(%)
Full supervision	PointNet [46]	-	-	80.8	30.3	23.7	-	-	14.6
	PointNet++ [47]	95.7	68.3	84.3	40.0	32.9	84.9	41.8	20.1
	PointCNN [34]	97.2	58.4	-	-	-	-	-	-
	TangentConv [61]	-	-	77.0	43.7	33.3	-	-	40.9
	ShellNet [95]	96.4	57.4	-	-	-	-	-	-
	DGCNN [75]	-	-	-	-	-	94.2	61.8	-
	SPG [31]	95.5	60.6	85.3	44.4	37.3	-	-	17.4
	SparseConv [16]	-	-	88.7	63.3	42.7	-	-	-
	KPConv [65]	<u>97.8</u>	<u>81.1</u>	<u>93.2</u>	63.8	<u>57.6</u>	<u>95.4</u>	69.1	<u>58.1</u>
	ConvPoint [4]	97.2	67.4	-	-	-	-	-	-
RandLA-Net [24]	97.1	80.0	89.8	<u>69.6</u>	52.7	92.9	<u>77.7</u>	53.9	
Weak supervision	<b>Ours (0.1%)</b>	97.0	72.0	<b>91.0</b>	<b>70.9</b>	<b>54.0</b>	96.7	<b>77.7</b>	50.8
	<b>Ours (0.01%)</b>	95.9	60.4	85.6	49.4	37.2	94.2	68.2	39.1

### ■ Sensitivity to random sparse annotation

	OA(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board	clut.
Iter1	<b>85.03</b>	<b>57.52</b>	<b>93.28</b>	<u>96.68</u>	75.15	0.00	<b>15.08</b>	46.69	<u>59.81</u>	<b>72.18</b>	<u>81.78</u>	51.25	63.27	<u>44.23</u>	<b>48.37</b>
Iter2	84.04	55.07	<u>91.38</u>	95.30	74.72	0.00	11.87	<b>50.83</b>	48.18	65.28	77.79	41.20	63.89	<b>50.00</b>	45.49
Iter3	<u>84.99</u>	<u>57.10</u>	91.66	<b>96.93</b>	<u>76.64</u>	0.00	13.26	<u>50.08</u>	57.70	67.52	<b>82.15</b>	<b>56.01</b>	<u>64.19</u>	38.88	47.25
Iter4	84.58	55.42	90.48	96.26	<u>75.50</u>	0.00	12.97	47.69	40.51	<u>71.95</u>	81.10	<u>55.26</u>	<b>65.03</b>	35.42	<u>48.24</u>
Iter5	84.54	55.93	89.07	95.43	<b>76.92</b>	0.00	<u>14.34</u>	48.42	<b>62.87</b>	68.47	79.96	41.71	63.16	41.07	45.69
<b>Average</b>	84.64	56.21	91.17	96.12	75.79	0.00	13.50	48.74	53.81	69.08	80.56	49.09	63.91	41.92	47.01
<b>STD</b>	0.40	1.06	1.55	0.73	0.95	0.00	1.24	1.70	9.24	2.96	1.76	7.20	0.76	5.54	1.37

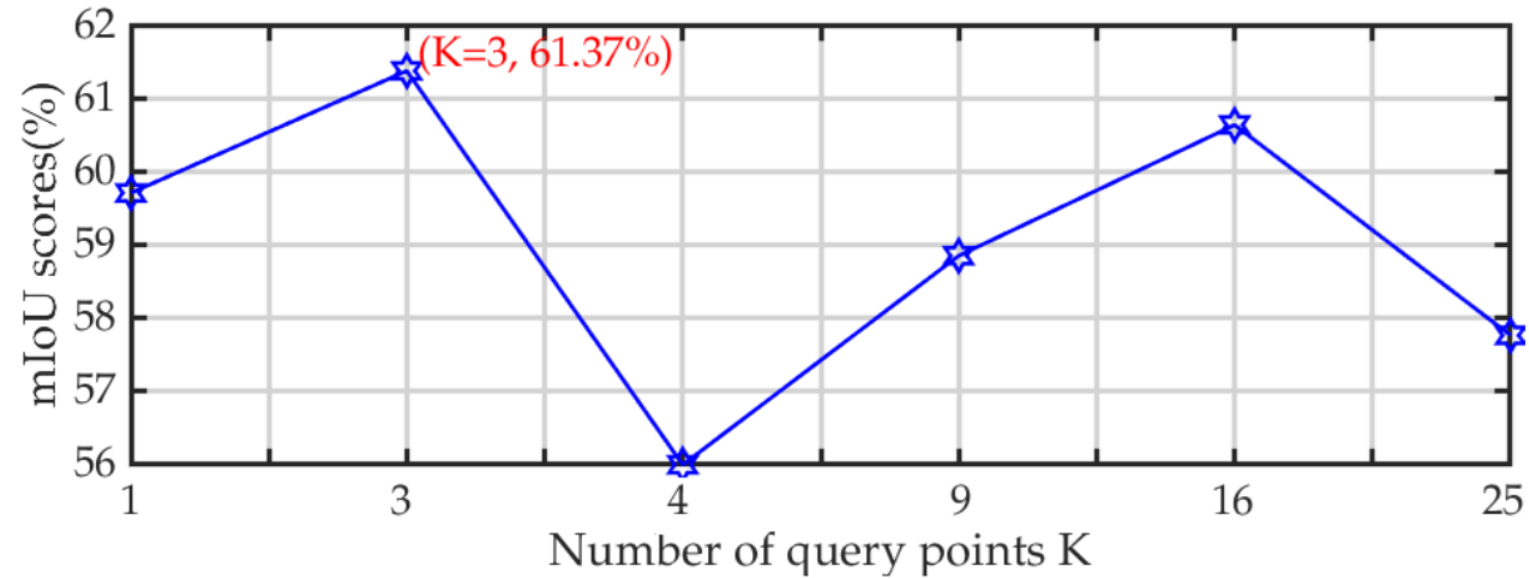
## ■ Variants of Semantic Queries

Model	1st	2nd	3rd	4st	OA(%)	mIoU(%)
A	✓				48.66	22.89
B				✓	75.54	46.02
C	✓	✓			70.76	38.18
D	✓	✓	✓		82.37	54.21
E	✓	✓	✓	✓	86.26	61.37

- Querying at the last layer can achieve much better results than in the first layer
- Querying at different encoding layers and combining them is likely to achieve better segmentation results



## ■ Varying Number of Queried Neighbours



## ■ Extension to Region-wise Annotated Data

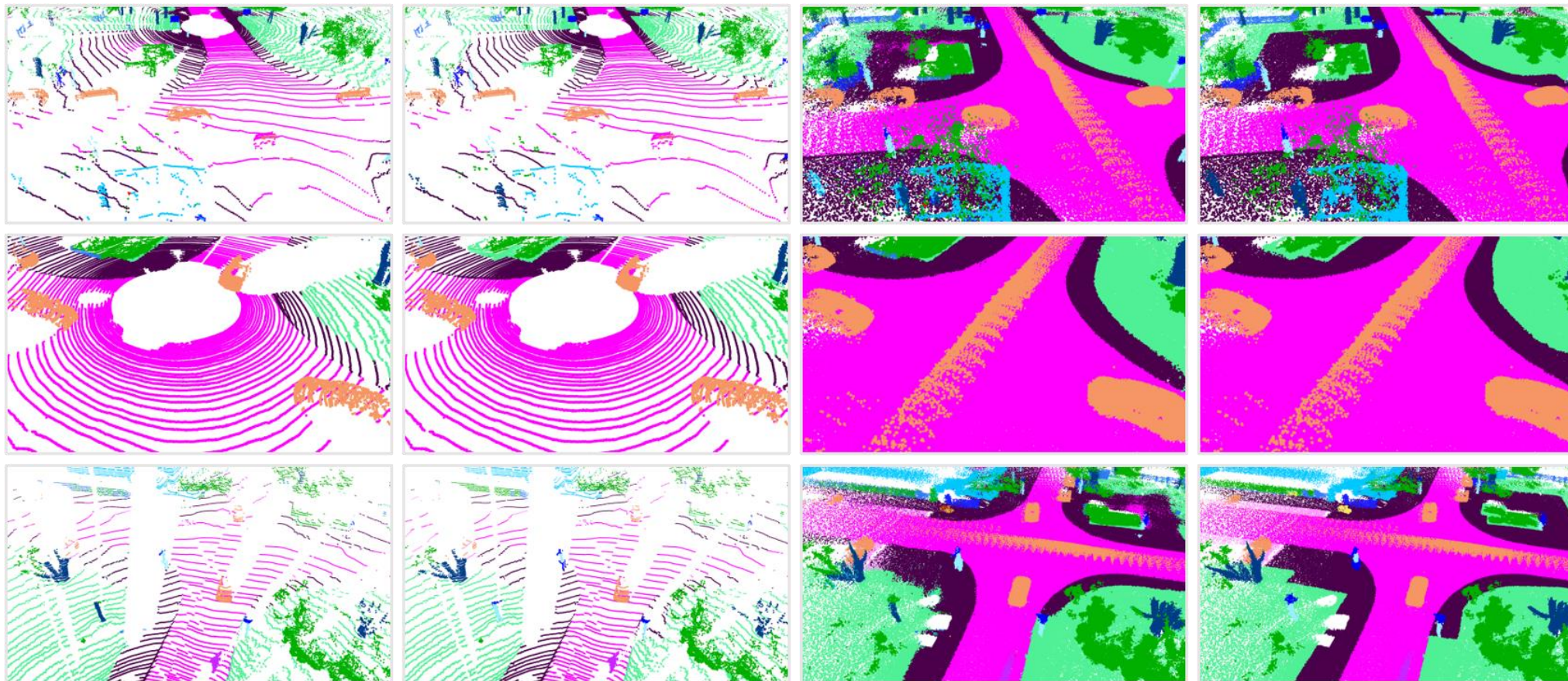
	SPVCNN [63]	MinkowskiUnet [11]	SQN (Ours)
Random	49.61	46.15	<b>60.19</b>
Softmax Confidence [74]	51.05	45.45	<b>57.24</b>
Softmax Margin [74]	50.80	44.33	<b>57.94</b>
Softmax Entropy [74]	50.35	49.99	<b>57.98</b>
MC Dropout [18]	50.39	49.94	<b>58.30</b>
ReDAL [85]	50.89	47.88	<b>54.24</b>

### ■ SQN with different backbones

Methods	mIoU(%)	Params(M)	<i>road</i>	<i>sidewalk</i>	<i>parking</i>	<i>other-ground</i>	<i>building</i>	<i>car</i>	<i>truck</i>	<i>bicycle</i>	<i>motorcycle</i>	<i>other-vehicle</i>	<i>vegetation</i>	<i>trunk</i>	<i>terrain</i>	<i>person</i>	<i>bicyclist</i>	<i>motorcyclist</i>	<i>fence</i>	<i>pole</i>	<i>traffic-sign</i>
MinkUNet 0.1%	55.5	21.9	92.5	79.0	43.0	0.9	88.7	95.0	64.5	0.9	47.4	46.4	87.3	63.4	73.7	45.3	70.3	0.3	53.4	59.9	43.2
<b>SQN (MinkUNet) 0.1%</b>	<b>55.8</b>	<b>8.8</b>	<b>91.5</b>	<b>78.0</b>	<b>41.1</b>	<b>0.9</b>	<b>88.5</b>	<b>94.9</b>	<b>66.8</b>	<b>5.7</b>	<b>43.2</b>	<b>43.6</b>	<b>88.2</b>	<b>64.0</b>	<b>75.5</b>	<b>49.5</b>	<b>66.3</b>	<b>0.0</b>	<b>55.9</b>	<b>61.1</b>	<b>45.2</b>
MinkUNet 0.01%	43.2	21.9	89.3	74.8	32.1	0.0	87.6	92.4	25.8	0.0	24.8	20.1	87.1	56.4	73.2	9.6	15.9	0.0	55.1	48.2	28.3
<b>SQN (MinkUNet) 0.01%</b>	<b>50.0</b>	<b>8.8</b>	<b>89.7</b>	<b>75.6</b>	<b>31.9</b>	<b>0.2</b>	<b>87.6</b>	<b>93.5</b>	<b>47.2</b>	<b>0.2</b>	<b>35.6</b>	<b>31.6</b>	<b>88.2</b>	<b>58.0</b>	<b>76.0</b>	<b>33.8</b>	<b>59.1</b>	<b>0.0</b>	<b>52.9</b>	<b>52.1</b>	<b>36.4</b>



- Train in partial point clouds, test in complete point clouds



Predictions

Ground Truth

Predictions

Ground Truth

04

# Semantic Query Network

Demo



*Performance on Public Benchmarks*

■ Number of annotated points in practice

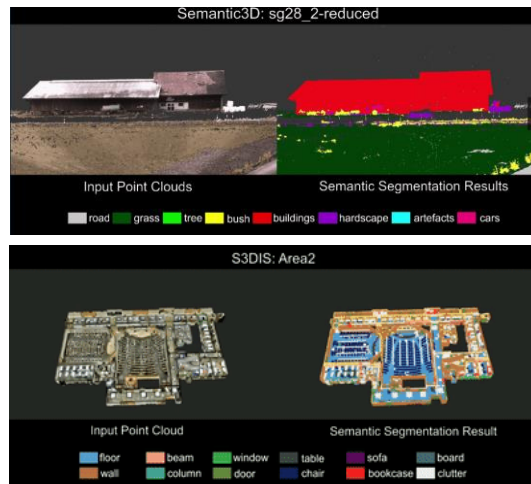
	Grid size	Raw pts	Grid sampled pts	Anno. pts (0.1%)	
S3DIS [2]	0.04	273M	18.6M	18,600	
Semantic3D [18]	0.06	4000M	78.1M	78,100	→ 0.002%
ScanNet [67]	0.04	242M	60.2M	60,200	
SemanticKITTI [3]	0.06	5299M	3401M	3.4M	
DALES [67]	0.32	505M	211M	211,000	
SensatUrban [23]	0.2	2847M	221M	221,000	
Toronto3D [57]	0.04	78.3M	24.3M	24,300	



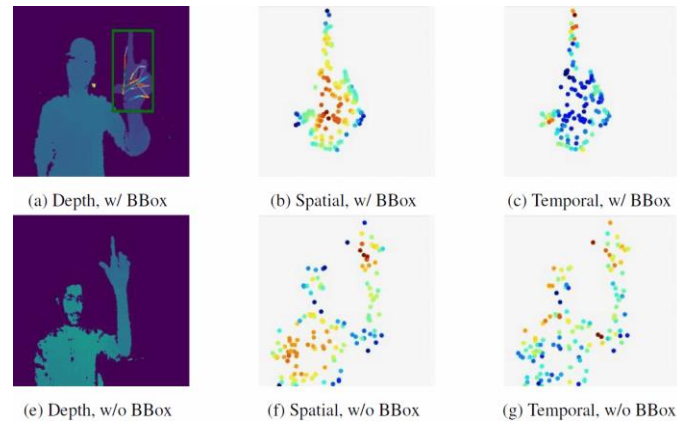
## *Sparse Point Annotation Pipeline*

- ✓ Save up to 98% annotation cost for large-scale 3D point clouds

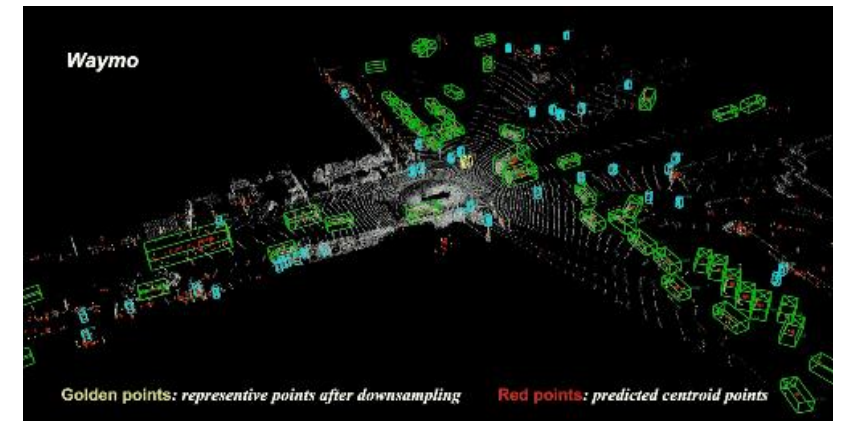
- Dynamic point cloud processing; (*Kinet, CVPR2022*)
- Efficient semantic segmentation of large-scale point clouds; (*RandLA-Net, CVPR 2020*)
- Efficient 3D object detection; (*IA-SSD, CVPR 2022*)
- Geometry/Attribute compression of 3D scenes; (*3DAC, CVPR 2022*)
- Generalized 3D point clouds registration; (*SpinNet, CVPR 2021*)



[RandLA-Net](#)



[KiNet](#)



[IA-SSD](#)

05

# Conclusion

Future directions

- Learning Unified 3D Representation





# Conclusion

## Future Directions

### ■ Future directions



Figure from “Immerse View for Google Maps”



Figure from Matthew et al. “Block-NeRF: Scalable Large Scene Neural View Synthesis”



UNIVERSITY OF  
**OXFORD**

Department of  
**COMPUTER SCIENCE**

**MANY THANKS !**

Follow us if you are interested in our work:

Homepage: <https://qingyonghu.github.io/>

GitHub: <https://github.com/QingyongHu>