Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Master's Thesis

Out-of-distribution detection in 3D semantic segmentation

Lokesh Veeramacheneni

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr. Paul G Plöger
Dr. Matias Valdenegro
Prof. Dr. Sebastian Houben

April 2022

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

Date	Lokesh Veeramacheneni

Contents

List of Figures

List of Tables

1

Experiments and Results

This chapter discusses the experiments conducted for Out-Of-Distribution (OOD) detection on RandLA-Net using quantified uncertainty from Deep Ensembles and Flipout. In a detailed discussion about RandLA-Net, Deep Ensembles and Flipout can be found in Chapter [$]. In this chapter, we first discuss the training results of RandLA-Net using deep ensembles and flipout on Semantic3D as an In-Distribution (ID) dataset. Furthermore, we compare the Maximum Softmax Probability (MSP) as in [1] and entropy values for the proposed OOD benchmark datasets Semantic3D vs S3DIS and Semantic3D vs Semantic3D w/o colour. Finally, we visualize and evaluate the performance of OOD detection using the AUROC score.

1.1 Deep Ensembles-Semantic3D

In this experiment, we trained 20 models of RandLA-Net over the Semantic3D dataset using random initializations with the experimental setup described in Section [$]. The predictions from these 20 individual models are then averaged to compute the final predictions. The evaluation results of the Deep Ensembles are described in Table 1.1 using meanIoU, per-class IoU and accuracy. The predictions from the Deep Ensembles are depicted in Figure 1.1 and Figure 1.2.

		IoU per class								
Ensemble size	**meanIoU**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Accuracy**
1	68.19	94.55	81.19	84.67	29.43	81.37	18.85	64.74	90.74	88.78
5	69.51	94.73	81.92	84.42	28.05	**86.41**	28.50	61.03	91.03	90.04
10	69.97	95.25	83.73	86.63	30.36	84.13	18.60	**66.01**	92.61	89.94
15	70.32	95.27	83.54	**88.22**	**32.19**	84.82	26.17	61.67	90.75	**90.57**
20	**70.80**	**95.55**	**84.11**	86.65	29.60	85.41	**29.58**	62.47	**93.06**	90.56

Table 1.1: Illustration of performance of RandLA-Net on Semantic3D over number of ensembles. meanIOU and IOU per-class and overall accuracy are represented here. C1 to C8 are the classes of Semantic3D which are Manmadeterrain, Naturalterrain, Highvegetation, Lowvegetation, Buildings, Hardscapes, Scanningartifacts, and Cars.

[Talk about lower class scores for some classes] From the Table 1.1, we infer that the Deep Ensembles improve the model's overall performance in terms of meanIoU and Accuracy. With an ensemble size of 10, we observe a 2% increment in meanIoU. An increase in ensemble size also results in an improvement in

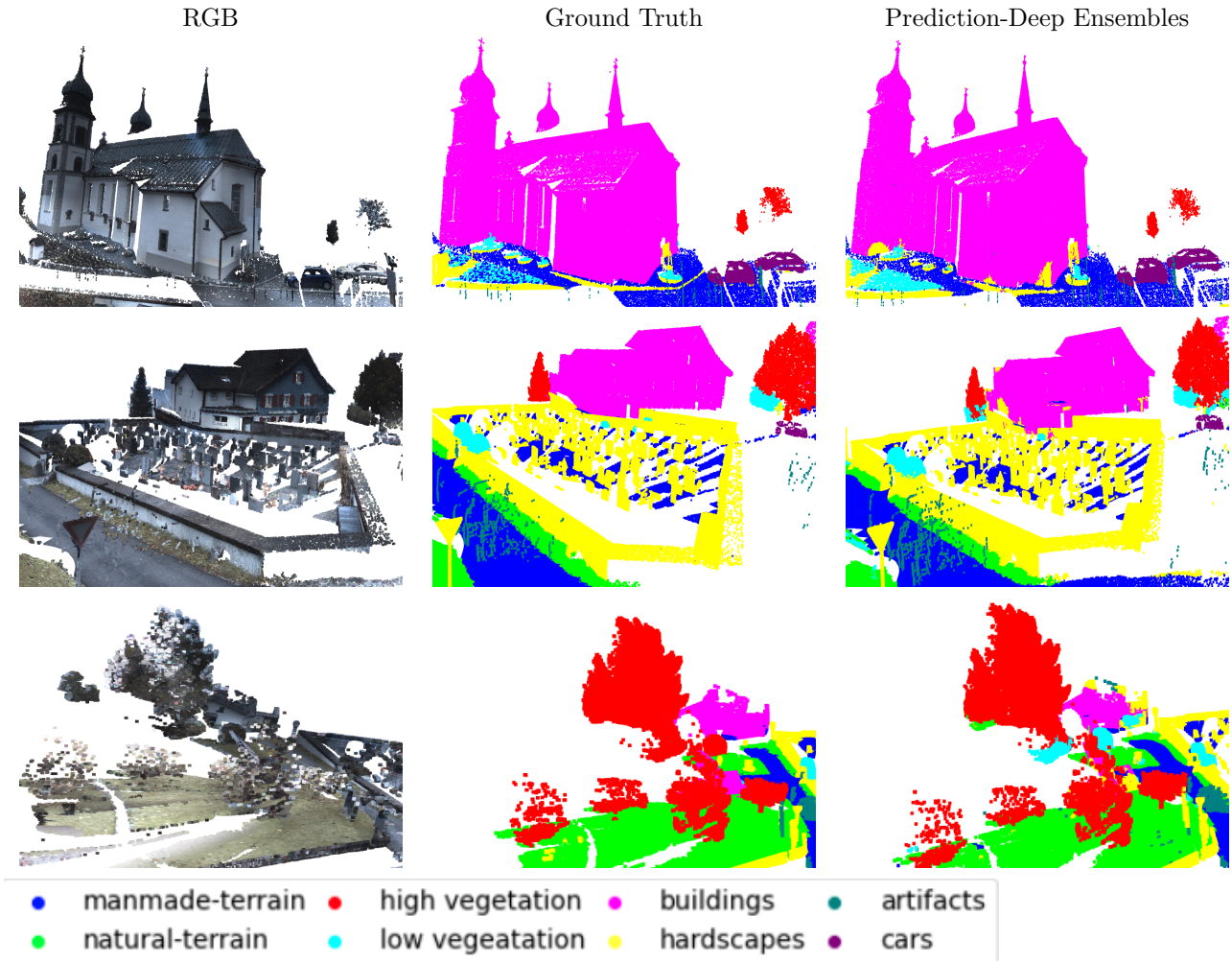


Figure 1.1: Output predictions of the RandLA-Net over the Semantic3D dataset (13 ensemble size) Legend spelling mistake.

per-class IoU performance. Figure 1.2 depicts the improvement in model classifications visually with the increase in ensemble size. From Figure 1.1, we also observe the misclassifications along the edges of the church, trees and ground. The possible explanation for these misclassifications is ambiguity in the feature vector of RandLA-Net. For example, a feature vector for the point along the lower edge of the church contains the part of ground points as the feature vector.

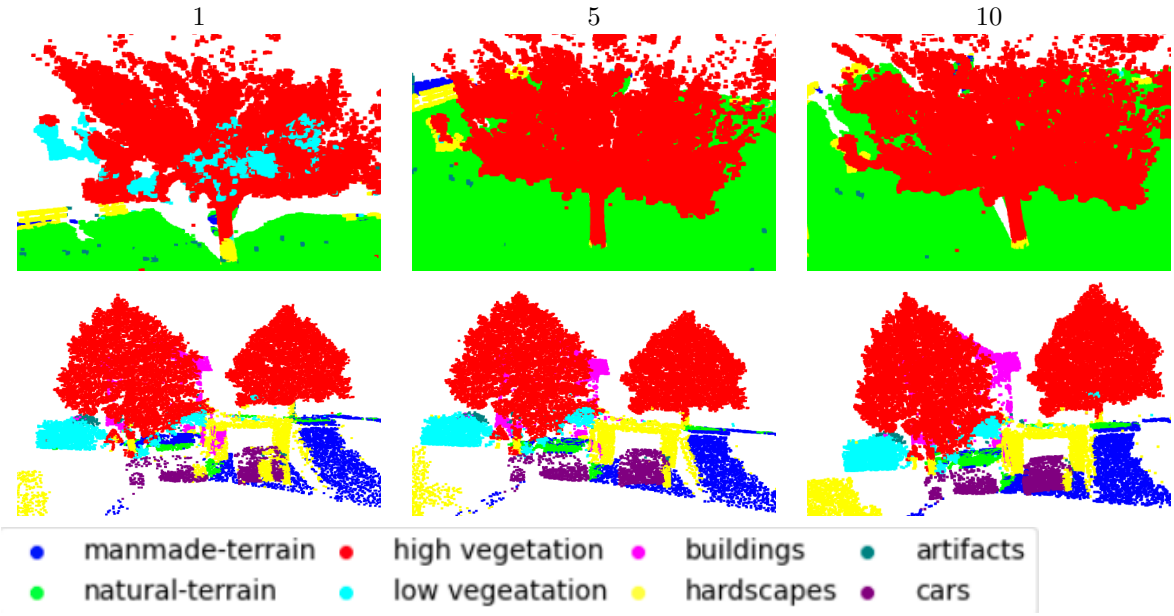


Figure 1.2: Deep ensembles performance on RandLA-Net over the Semantic3D dataset.

1.2 Flipout-Semantic3D

In this experiment, we trained a Flipout version of RandLA-Net as described in Section [$] over the Semantic3D dataset. Like Deep Ensembles, we performed 20 forward passes over the Flipout versioned RandLA-Net and averaged the predictions to obtain final predictions. Table 1.2 describes the performance of Flipout versioned RandLA-Net using meanIoU, per-class IoU and Accuracy. Figure 1.3 depicts the predictions of the Flipout versioned RandLA-Net visually.

		IoU per class								
#Passes	**MeanIoU**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Accuracy**
1	69.95	94.24	80.09	86.16	22.48	88.70	39.41	57.42	91.12	90.71
5	69.83	94.38	80.21	84.10	23.32	87.80	39.68	57.75	91.43	90.43
10	69.84	94.38	80.16	83.90	23.46	87.73	39.75	57.83	91.47	90.40
15	69.86	94.38	80.17	83.80	23.48	87.73	39.82	57.96	91.57	90.40
20	69.87	94.38	80.18	83.80	23.57	87.72	39.84	57.92	91.57	90.40

Table 1.2: Illustration of performance of Flipout versioned RandLA-Net on Semantic3D. meanIOU and IOU per-class and overall accuracy are represented here. C1 to C8 are the classes of Semantic3D which are Manmadeterrain, Naturalterrain, Highvegetation, Lowvegetation, Buildings, Hardscapes, Scanningartifacts, and Cars.

From Table 1.2, we infer that the Flipout versioned RandLA-Net has a similar performance to the original RandLA-Net model proposed in [2] and also Deep Ensembles with ensemble size one . We also observe a significant improvement in the Hardscapes class represented as C6 in Table 1.2. There is a decrement in performance of classes Lowvegetation represented as C4 and Scanningartifacts represented

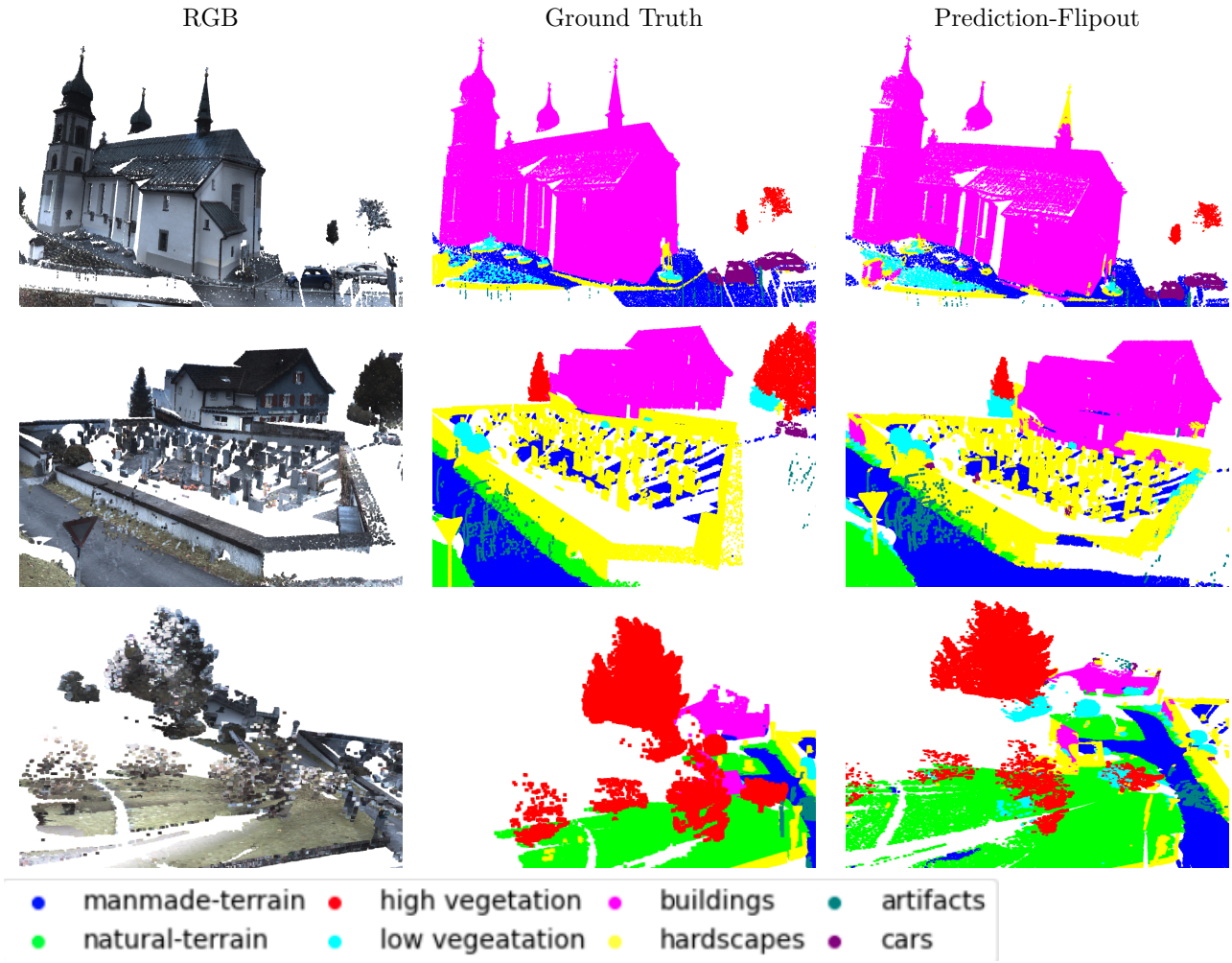


Figure 1.3: Output predictions of the RandLA-Net over the Semantic3D dataset (10 forward passes) Legend spelling mistake.

as C7 in Table 1.2, keeping the overall meanIoU same.

1.3 OOD benchmark - Semantic3D vs S3DIS

In the previous section, we studied the performance of the Deep Ensembles and Flipout over the Semantic3D (In-Distribution) dataset. In this section, we study the predictions of the RandLA-Net model on the S3DIS (Out-Of-Distribution) dataset using Deep Ensembles and Flipout. We also compare the distribution of Maximum Softmax Probability (MSP) and entropy scores for Semantic3D and S3DIS datasets.

Figure 1.4 depicts the predictions of the RandLA-Net model and Flipout versioned RandLA-Net. We observe that most objects, such as ceilings and bookshelves, are labelled as buildings when using Deep Ensembles. We also observe that most point clouds are labelled as hardscapes class when using Flipout versioned RandLA-Net. The classifications on S3DIS datasets are in triangles because of the property of the scanner. As discussed in Section [$], data collected from the Matterport scanner is represented as triangular mesh, and then a point cloud is extracted from this mesh.

1.3.1 Maximum Softmax Probability (MSP)

In this experiment, we study the probability values of the ID dataset (Semantic3D), and OOD dataset (S3DIS) computed using Deep Ensembles and Flipout methods of RandLA-Net. We compute the average of the maximum softmax probability of all the points in the dataset, and this averaged value is called here the mean probability value. Figure 1.5a and Figure 1.5b depicts the mean probability values across the ID (green) and OOD (red) datasets and their variance represented as error bars. Figure 1.5a represents the change in mean probability value to ensemble size. Figure 1.5b represents the change in mean probability value to the number of passes in Flipout. Here, we represent the mean probability values across the odd number of ensembles size and the odd number of passes in case of flipout.

From Figure 1.5a, we can infer that as the increase in ensemble size, the mean probability of the ID (Semantic3D) dataset remains stable. The variance is reduced until the ensemble size of 9 and then stabilizes. In the case of the OOD (S3DIS) dataset, we observe a decrement in mean probability value and then remain the same after an ensemble size of 3 with a larger variance. With the increase in ensemble size, we also observe that the overlap in the variance of ID and OOD is getting lower. This smaller overlap in higher ensemble size should result in higher OOD detection performance. In the case of Flipout, as in Figure 1.5b the mean probability and variance remain mostly the same for the ID dataset. With the OOD dataset, we observed a reduction in the mean probability value in the case of multiple passes. The variance from the Flipout is higher than the Deep Ensembles for the ID dataset. This phenomenon is to be expected because the Deep Ensembles combine predictions from various radomly initializated models and in the case of Flipout same model is used for multiple forward passes.

Figure 1.6 and Figure 1.7 represent the ground truth, prediction and probability map of the ID (Semantic3D) dataset using Deep Ensembles and Flipout respectively. Similarly Figure 1.8 and Figure 1.9 depict the prediction and probability map for OOD (S3DIS) dataset using Deep Ensembles and Flipout respectively. On visual inspection of Figure 1.6 and Figure 1.7, we observed that the probability scores are low for points which are misclassified. The points which lie on the edge of the structures are low

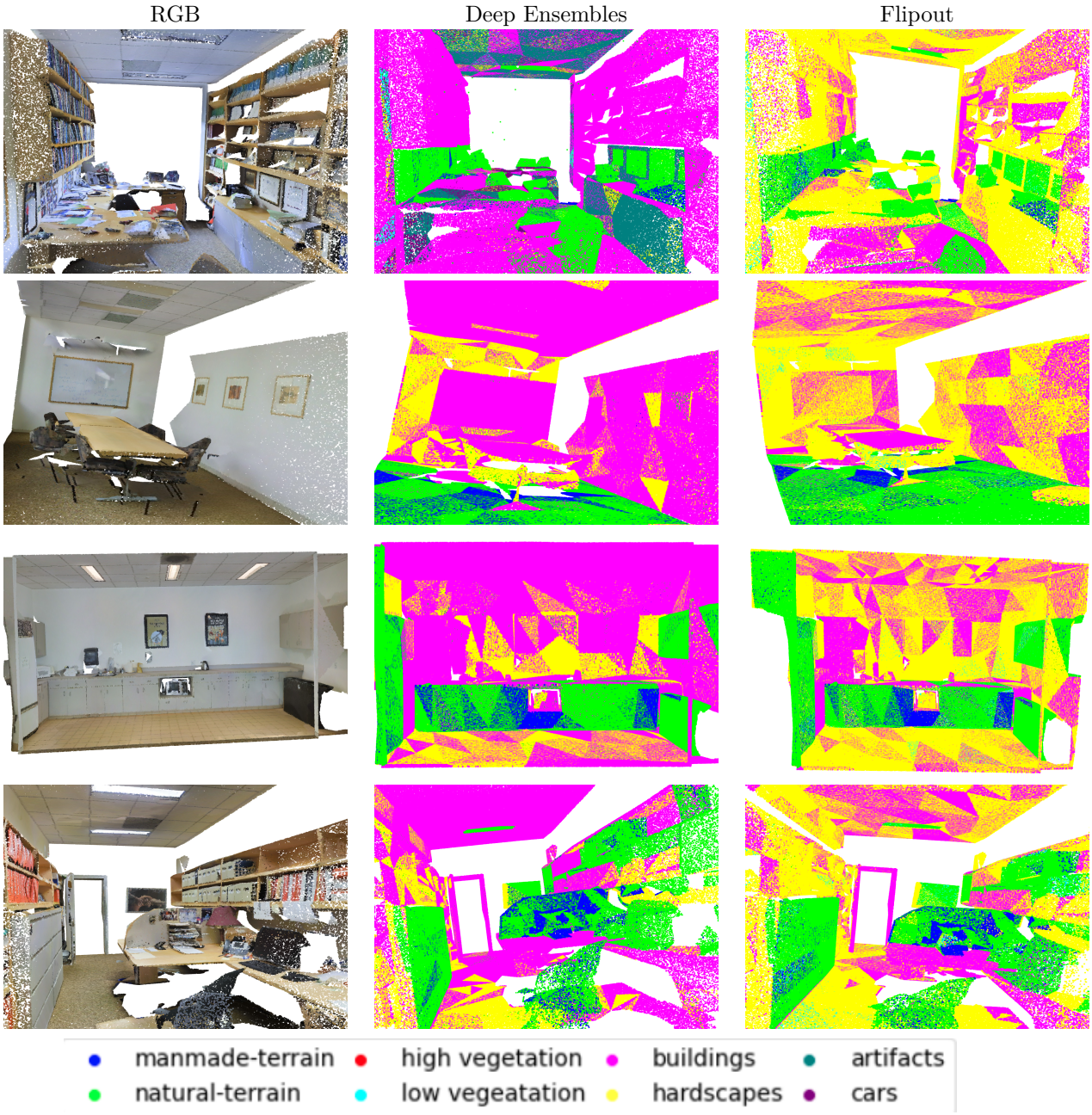
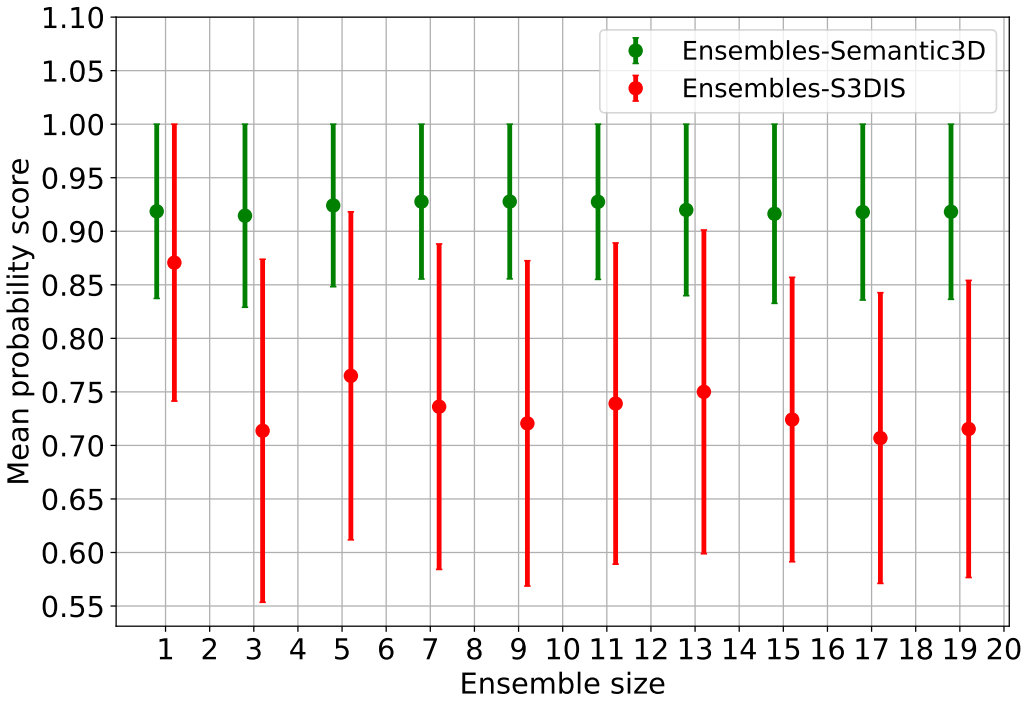
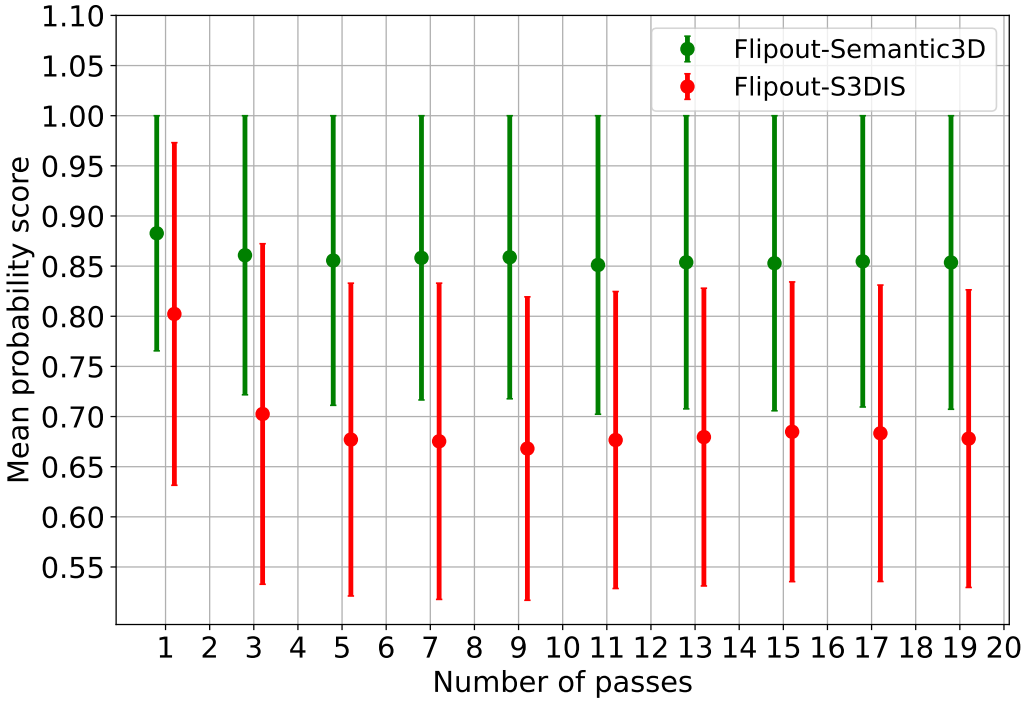


Figure 1.4: Output predictions of the RandLA-Net over the S3DIS dataset.

scored. This effect is profound near the edges of church, and also edges of walls. In case of OOD dataset presented in Figure 1.8 and Figure 1.9, the overall probability scores are low, as the whole point cloud has greener shade than the ID dataset probability map represented in yellow shade.



(a) MSP deep ensembles



(b) MSP flipout

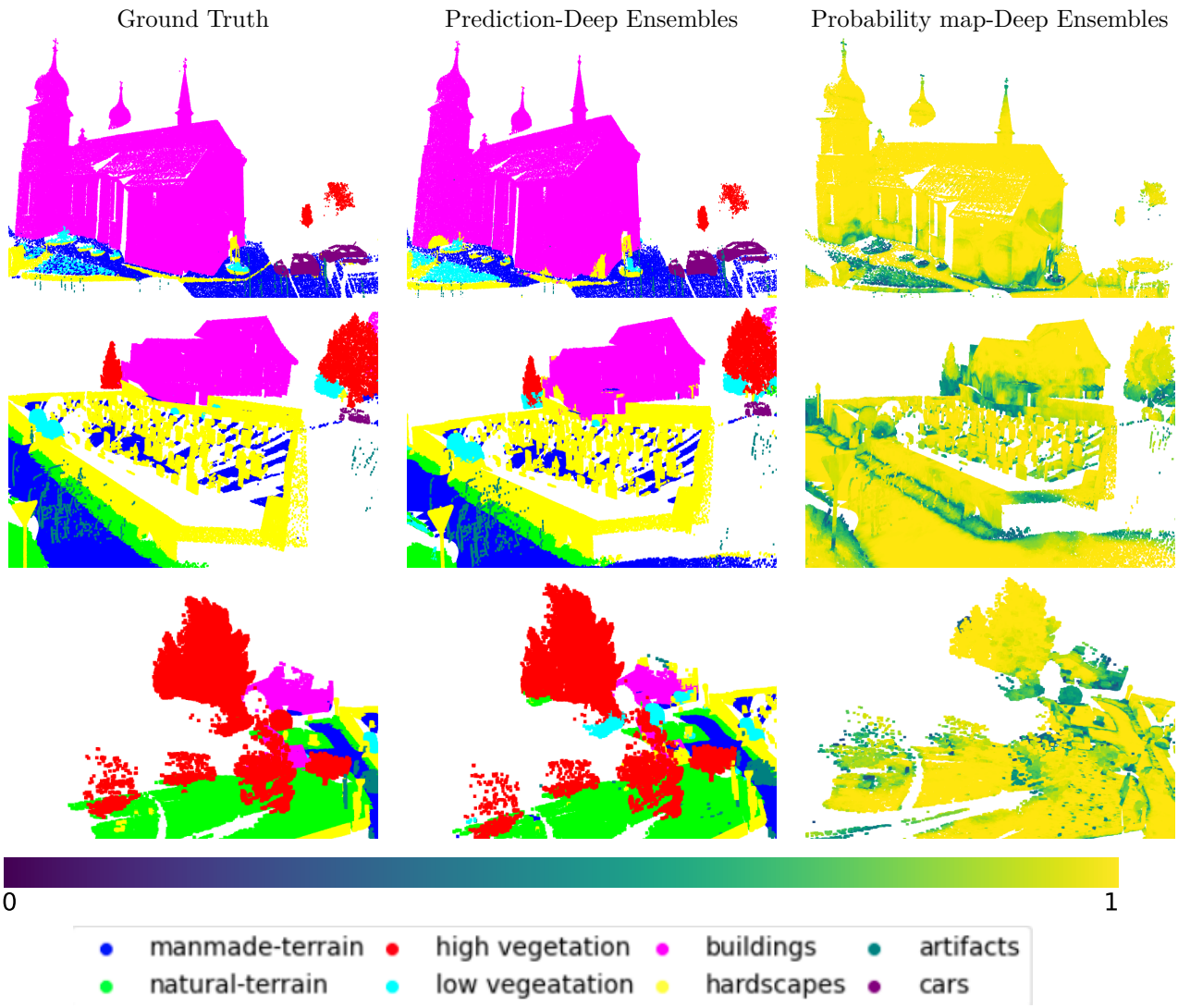


Figure 1.6: Perpoint probability visualization of the semantic3D dataset.

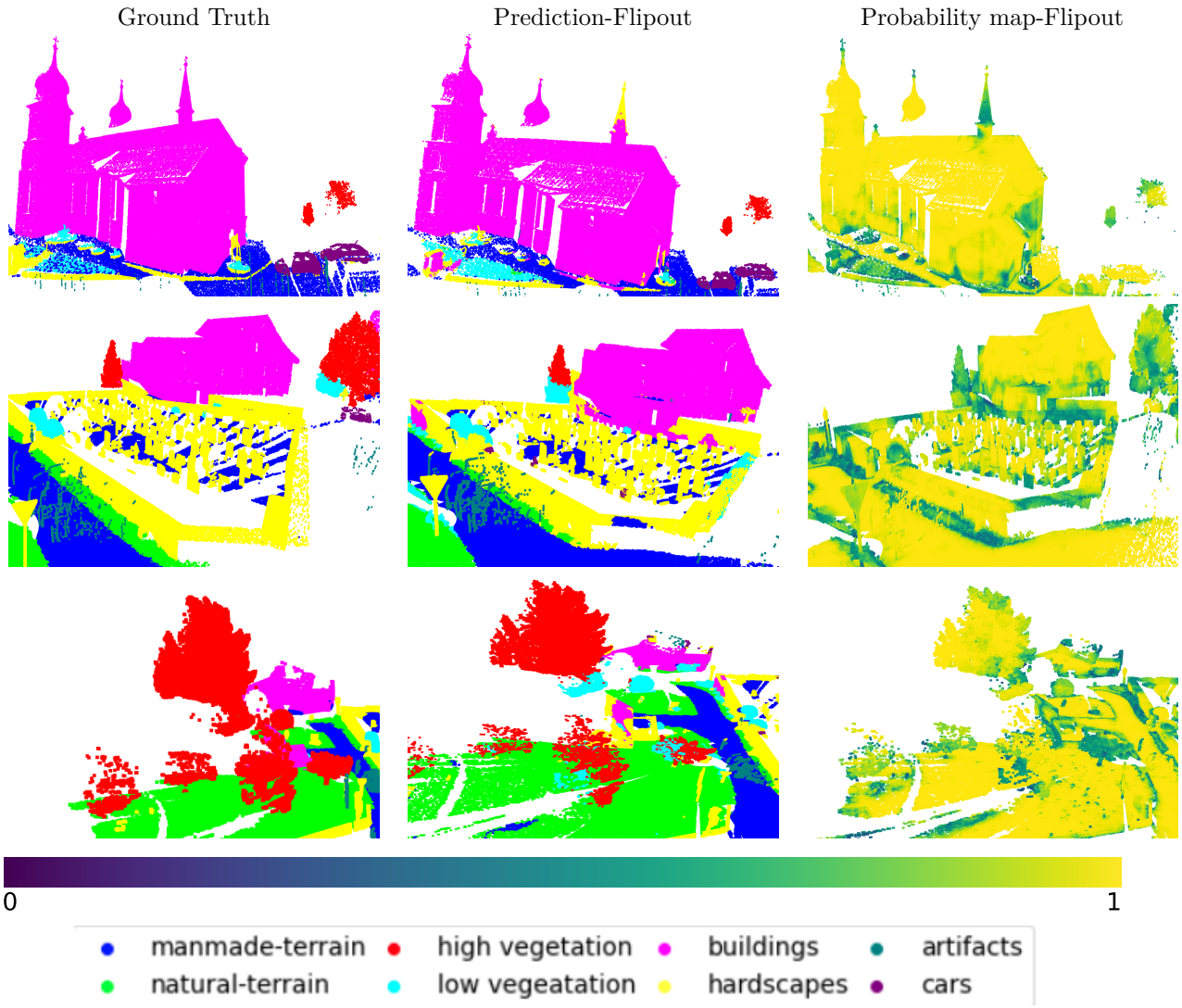


Figure 1.7: Perpoint probability visualization of the semantic3D dataset.

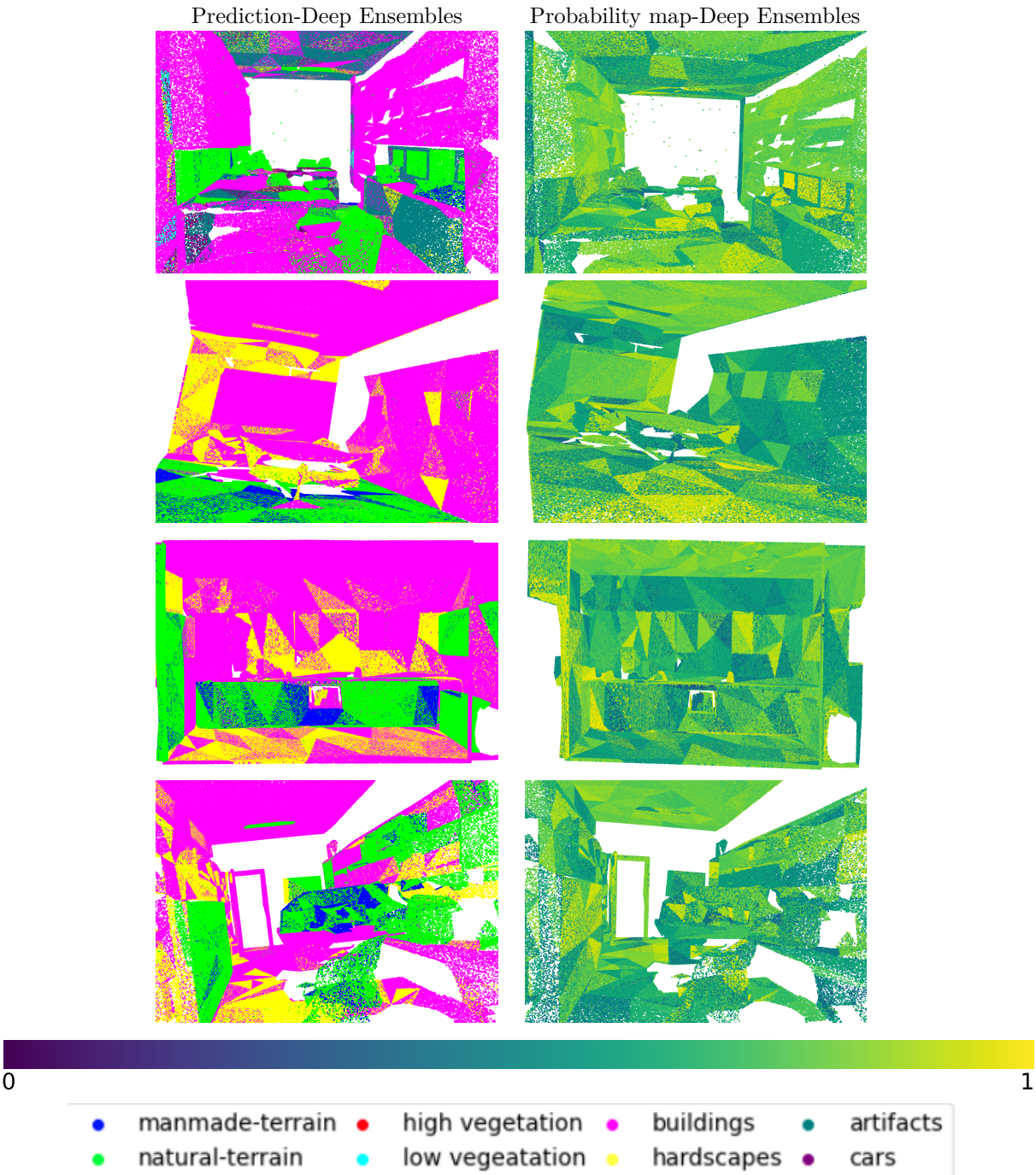


Figure 1.8: Perpoint probability visualization of the S3DIS dataset.

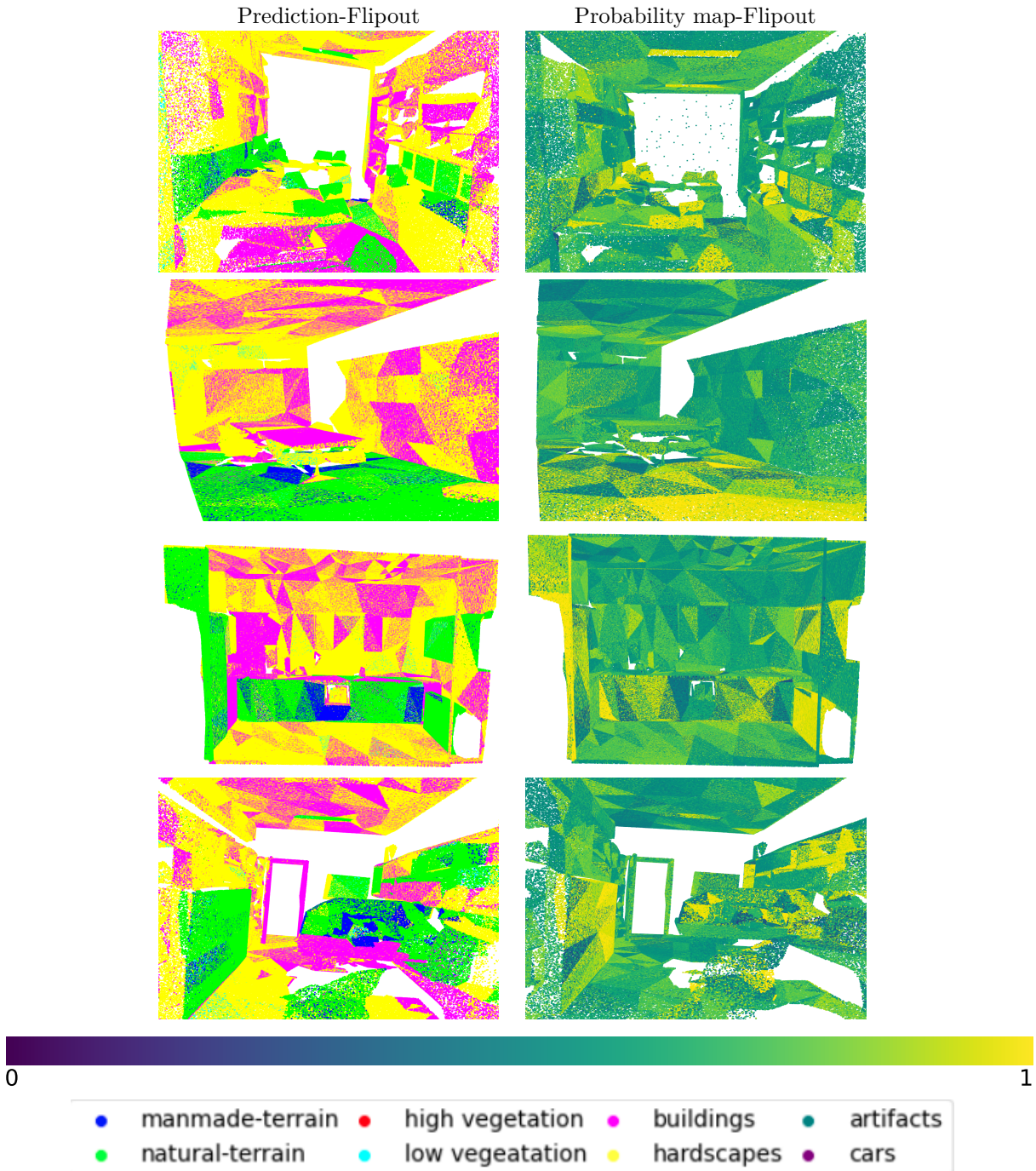


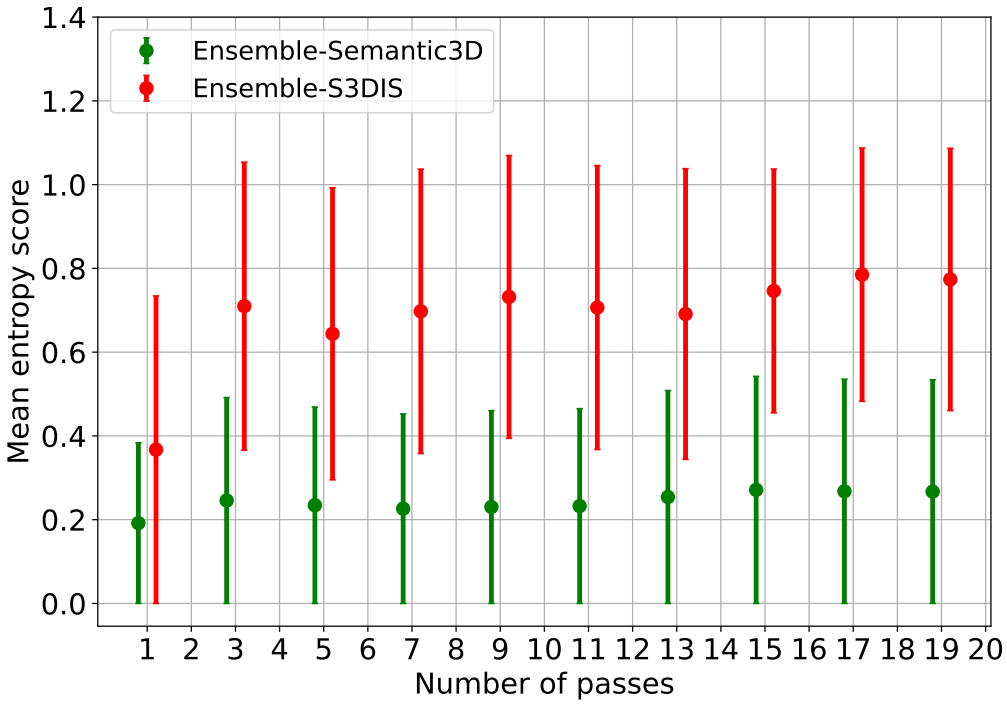
Figure 1.9: Perpoint probability visualization of the S3DIS dataset flipout.

1.3.2 Entropy

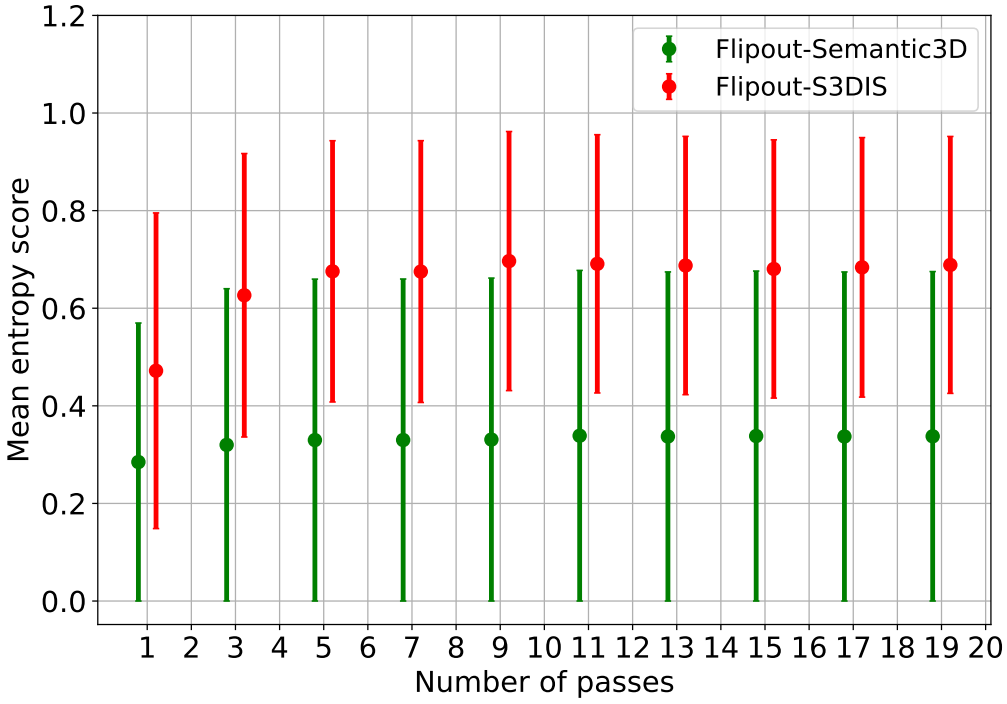
Similar to the experiment in Section [$], in this experiment, we study the distribution entropy values for ID (Semantic3D) and OOD (S3DIS) datasets. Entropy scores are a sum log function of softmax probabilities and detailed discussion in Section [$]. Figure 1.10a and Figure 1.10b depict the mean entropy score and variance plotted as error bars for Deep Ensembles and Flipout methods. Figure 1.11 and Figure 1.12 represents the entropy map for the ID dataset generated using Deep Ensembles and Flipout. Similarly, Figure 1.13 and Figure 1.14 represents the entropy map for the OOD dataset.

From Figure 1.10a and Figure 1.10b, we observe the entropy for the ID dataset is lower because the softmax values are not highly distributed across all the classes. Whereas in the case OOD dataset, the softmax probabilities are highly distributed across all the classes, so we observe a higher entropy score. Similar to the probability score, we also observe a decrement in the variance of entropy score until the ensemble size of 13 and then stabilizes out. Also, we infer that there is a lesser overlap between the error bars of OOD and ID datasets in the case of Deep Ensembles than the Flipout. So in both cases of MSP and entropy, we expect the Deep Ensembles to detect OOD objects better than the Flipout.

On careful observation of Figure 1.11 and Figure 1.12, we infer that overall entropy map is more bluish in shade representing the lower entropy scores. A higher entropy (yellow) is observed at the points of misclassifications and along the edges (similar to the case of probability map in Figures 1.6, 1.7). An example of misclassified points having higher entropy is observed in Figure 1.12, where the misclassified top of the church is greener on the entropy map compared to the rest of the church building. Visually observation of entropy maps for the OOD dataset in Figure 1.13 and Figure 1.14 reveals the colours to be at the higher end of the entropy scale resulting in a more yellowish shade in the case of both Flipout and Deep Ensembles.



(a) Entropy deep ensembles, (mistake in x axis)



(b) Entropy flipout

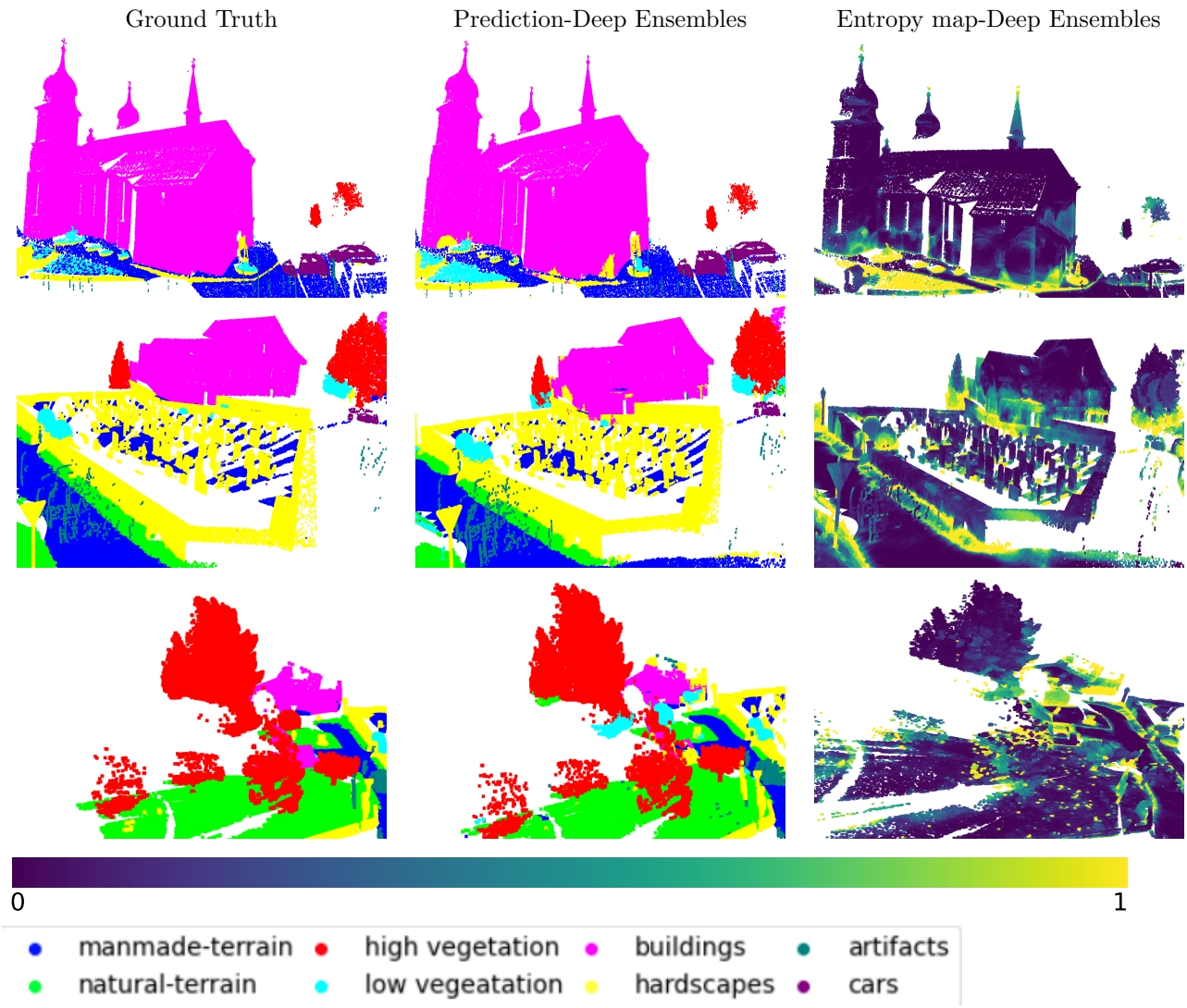


Figure 1.11: Perpoint entropy visualization of the semantic3D dataset-Ensembles. (Chnage the scale)

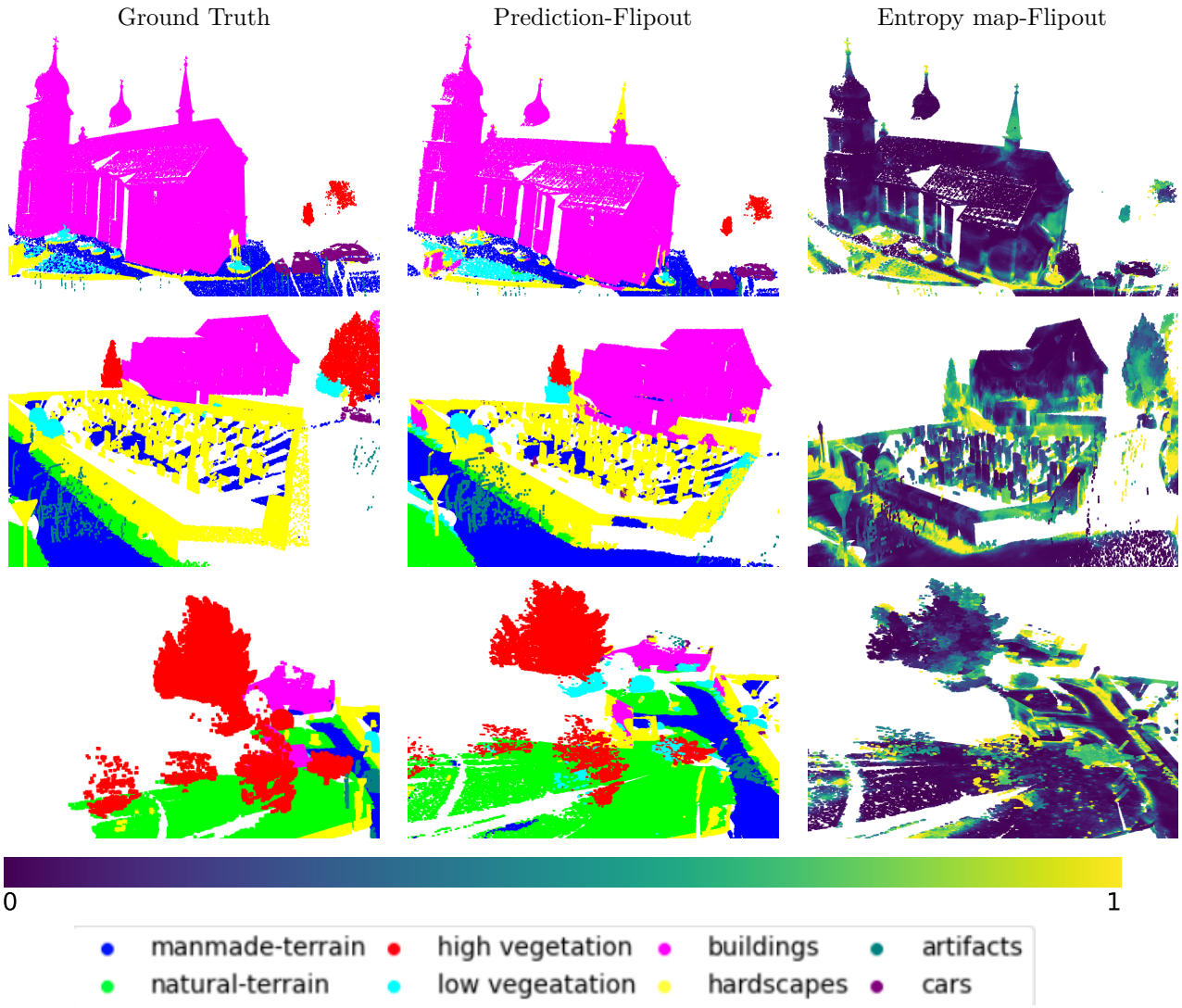


Figure 1.12: Perpoint entropy visualization of the semantic3D dataset.

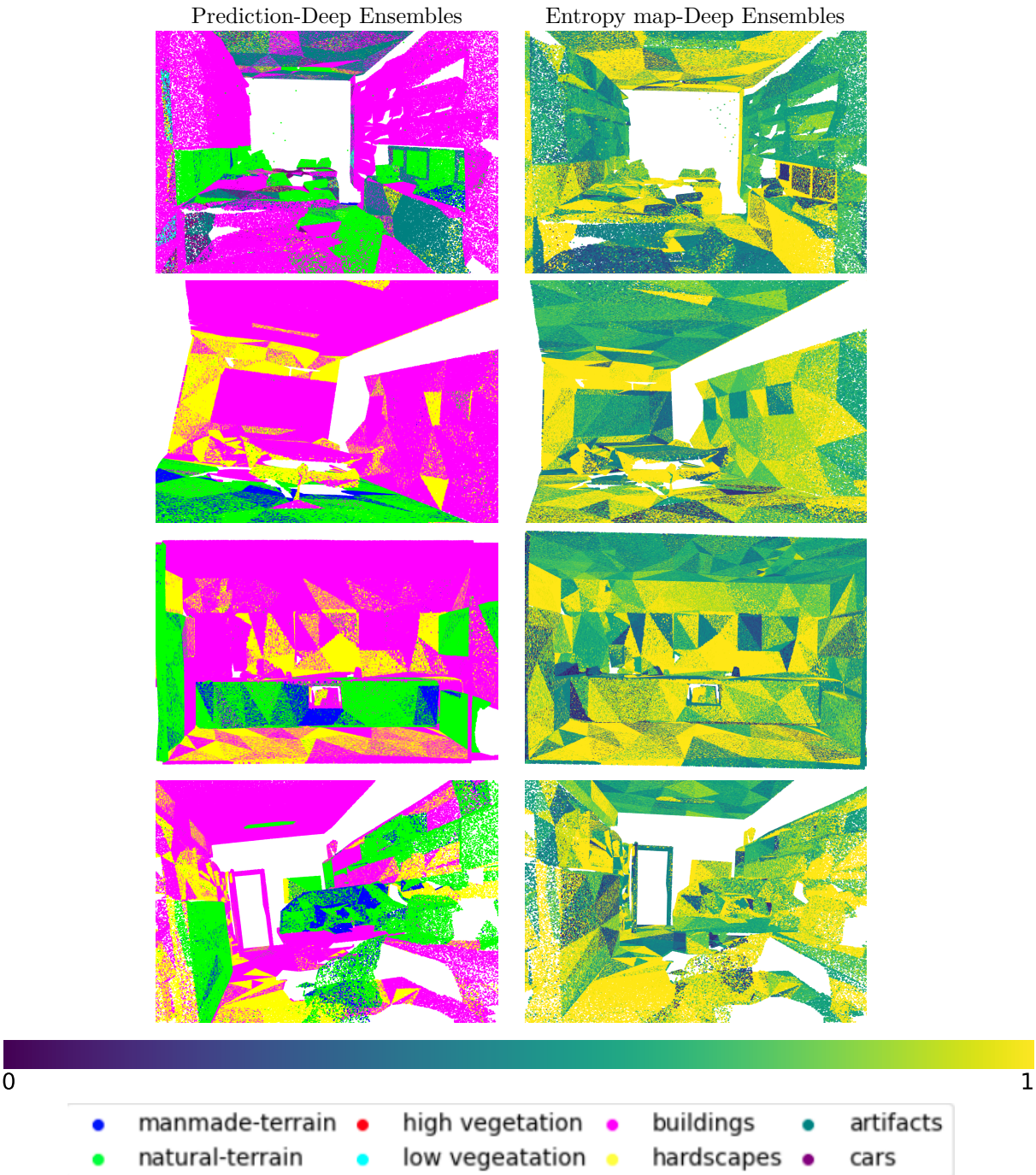


Figure 1.13: Perpoint entropy visualization of the S3DIS dataset.

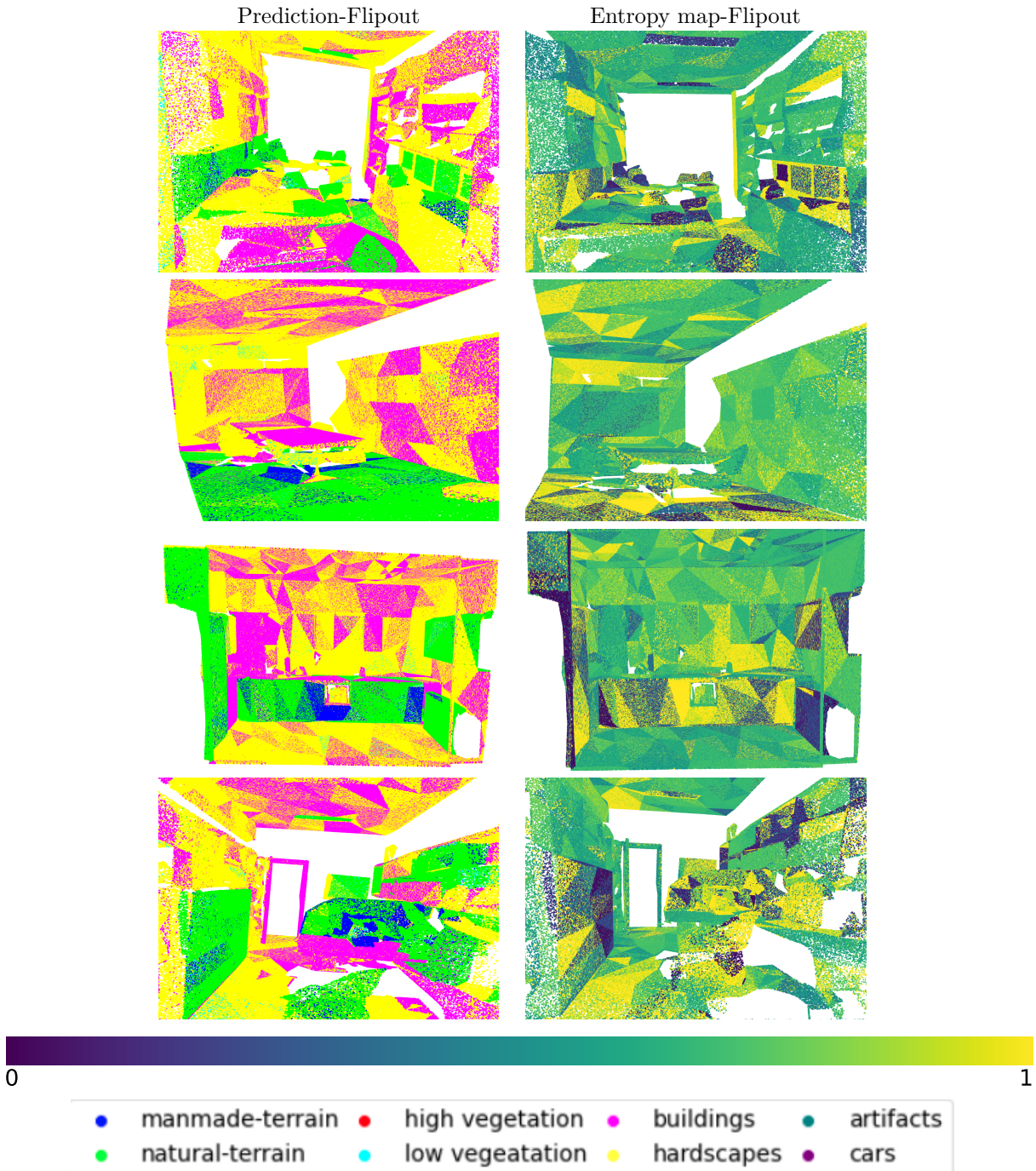


Figure 1.14: Perpoint entropy visualization of the S3DIS dataset flipout.

1.4 OOD detection evaluation - Semantic3D vs S3DIS

Ensemble size/ #passes	Method	AUROC	
		MSP	Entropy
1	Dropout	0.53311	0.53041
	Flipout	**0.69988**	**0.69368**
	Deep Ensembles	0.62020	0.62529
5	Dropout	0.58439	0.57821
	Flipout	0.77885	0.76934
	Deep Ensembles	**0.84013**	**0.83665**
10	Dropout	0.60168	0.59925
	Flipout	0.78728	0.78327
	Deep Ensembles	**0.87929**	**0.87541**
15	Dropout	0.59773	0.59557
	Flipout	0.7667	0.76741
	Deep Ensembles	**0.88486**	**0.88246**
20	Dropout	0.59766	0.59661
	Flipout	0.77331	0.77237
	Deep Ensembles	**0.89338**	**0.89052**

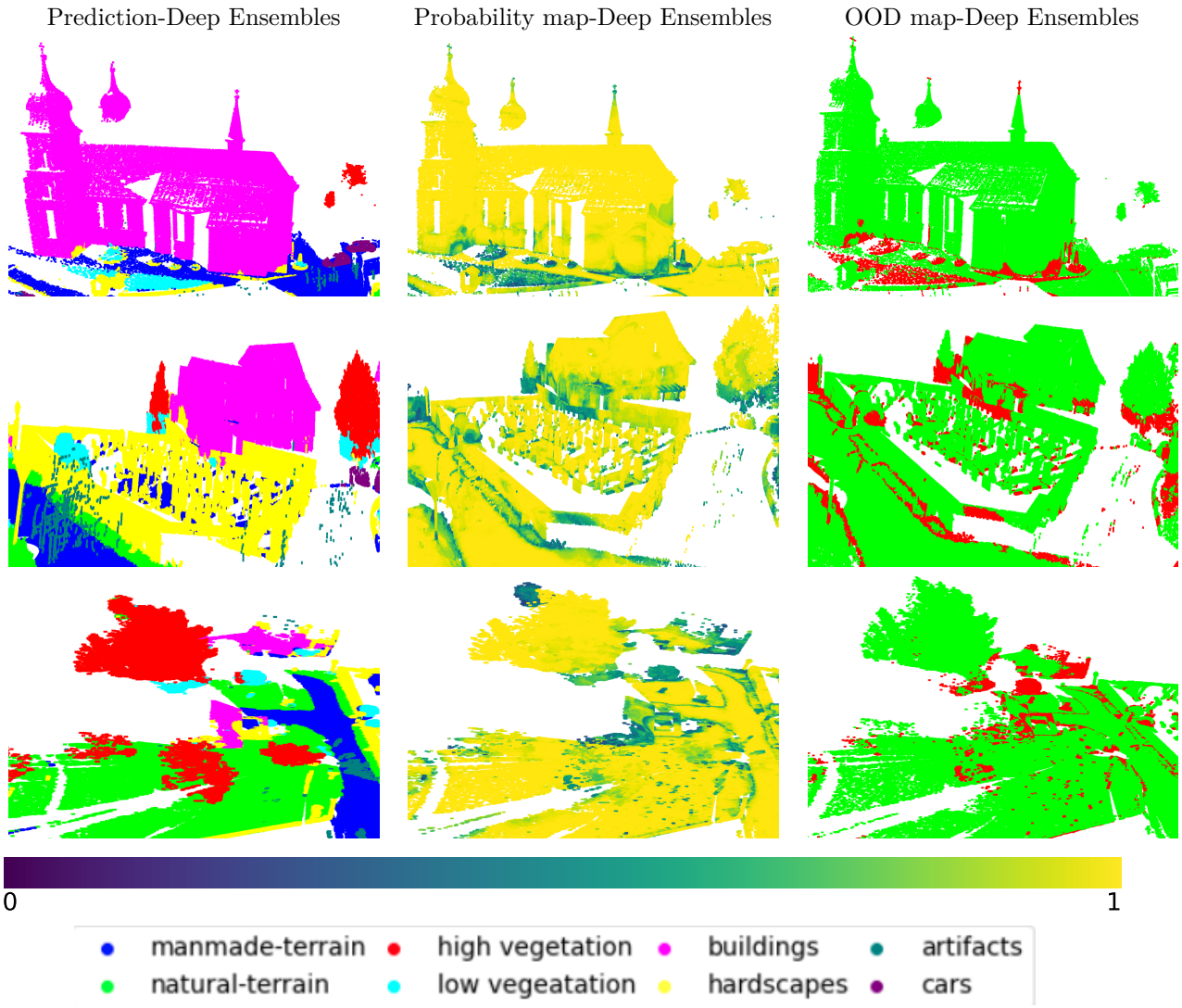


Figure 1.15: OOD point visualization of the semantic3D dataset Deep Ensembles-size 10.

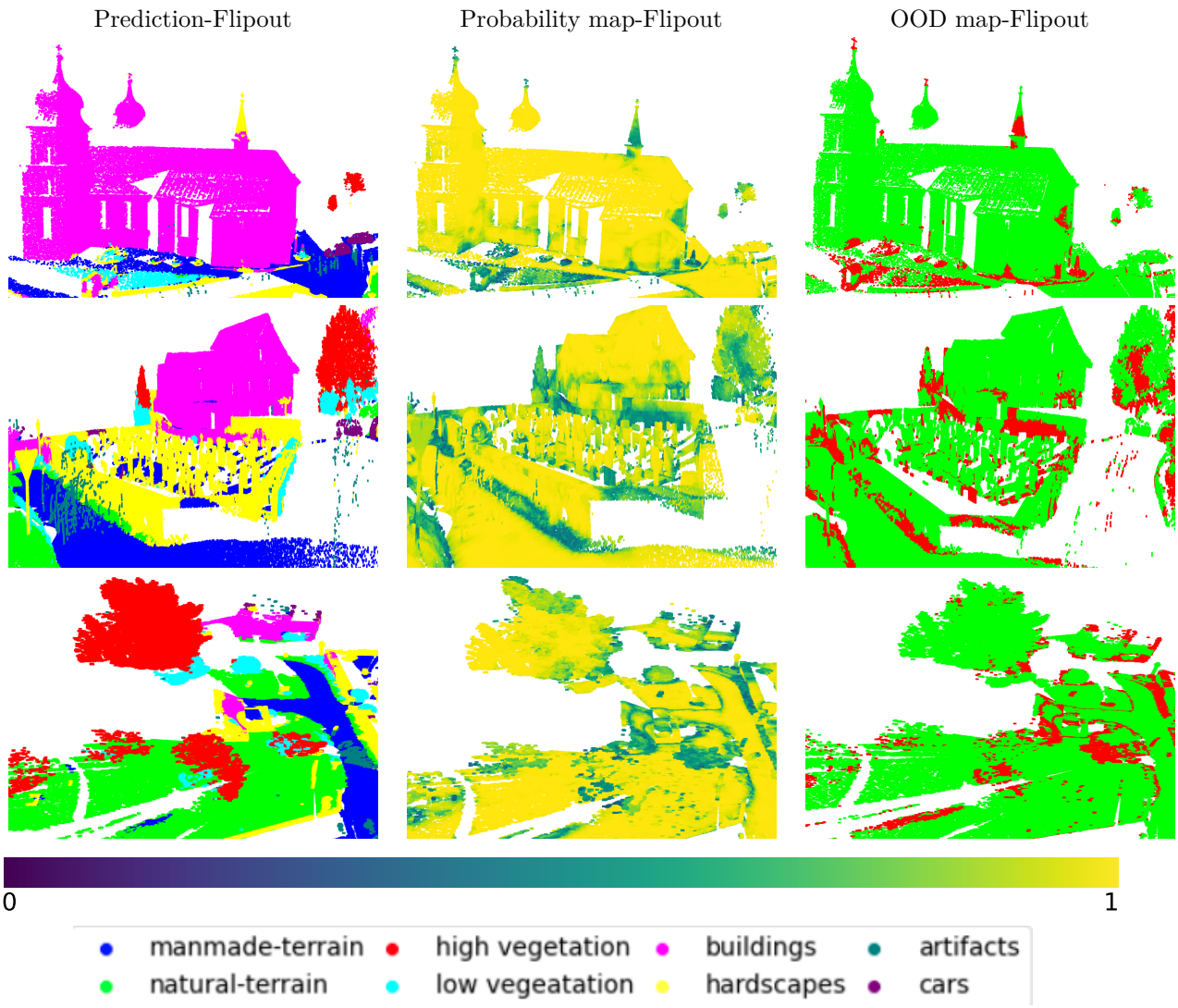


Figure 1.16: OOD point visualization of the semantic3D dataset Deep Ensembles-size 10.

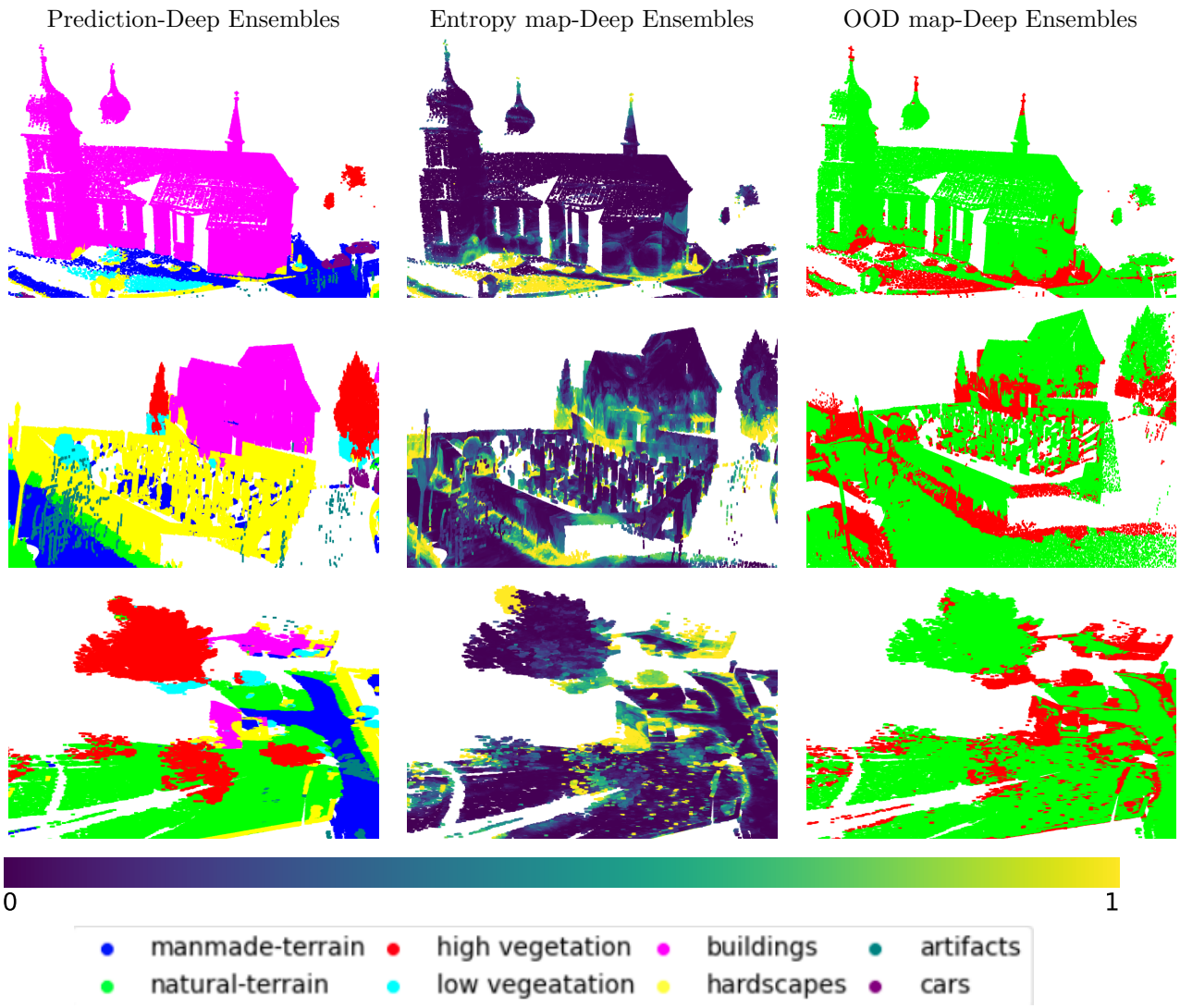


Figure 1.17: OOD point visualization of the semantic3D dataset Deep Ensembles-size 10.

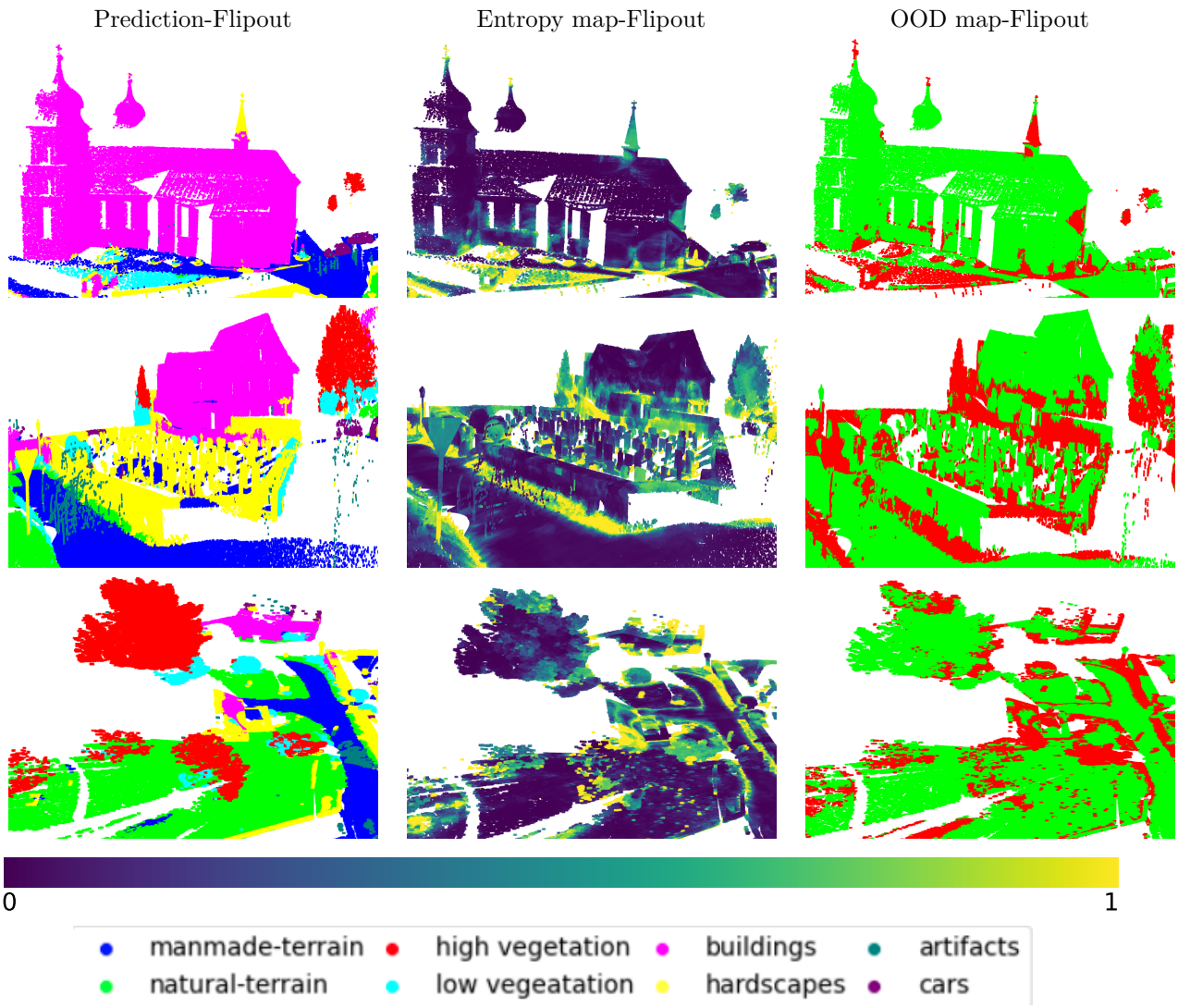


Figure 1.18: OOD point visualization of the semantic3D dataset Deep Ensembles-size 10.

1.5 OOD Benchmark - Semantic3D vs Semantic3D without color

1.5.1 Deep ensembles

1.5.2 Flipout

1.5.3 Maximum Softmax probability (MSP)

1.5.4 Entropy

1.6 OOD detection evaluation - Semantic3D vs Semantic3D without color

References

[1] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[2] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.