# Not Every Side Is Equal: Localization Uncertainty Estimation for Semi-Supervised 3D Object Detection

Chuxin Wang[1], Wenfei Yang[1], Tianzhu Zhang[1,2,*]
[1]University of Science and Technology of China, [2]Deep Space Exploration Lab
wcx0602@mail.ustc.edu.cn, yangwf@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

## A. Overview

In Scetion B, we first describe the computation process of category-specific scale factor used in the proposed pseudo-label selection strategy. We then provide training details for indoor and outdoor datasets in Section C. In Section D, we provide visual comparisons, overhead comparisons, and more detailed experimental comparisons with previous methods. Finally, we present additional ablation studies and visualizations to validate and analyze our method in Section E.

## B. Category-specific Scale Factor

In this section, we illustrate the process of calculating category-specific scale factor $\gamma_t(c)$. Since different categories of objects have different localization difficulties, we follow FlexMatch [5] and consider the learning progress of each category. Specifically, we define the learning progress $N_t(c)$ of a category as the number of pseudo-labeled categories used in the semi-supervised training process, as follows:

$$N_t(c) = \sum_{i=0}^{t} \text{Count}(\hat{y}_i^u, c), \tag{1}$$

where the $\text{Count}(\hat{y}_i^u, c)$ is the number of pseudo-labels for the category $c$ in the iteration $i$. By applying the normalization to $N_t(c)$, we obtain the relative learning progress $\sigma_t(c)$ of each category, the formula is as follows:

$$\sigma_t(c) = \frac{N_t(c)}{\max_c\{N_t(c)\}}, \tag{2}$$

Due to the instability in the early stage of model training, we introduce warm-up strategy in the above equation, as follows:

$$\beta_t(c) = \frac{N_t(c)}{\max\{\max_c\{N_t(c)\}, N - \sum_c N_t(c)\}}, \tag{3}$$

where $N$ is the hyper-parameter of the warm-up process and we set $N$ to be four times the quantity of unlabeled data. Finally, a convex function $\mathcal{M}(x) = \frac{x}{2-x}$ is applied to generate the category-specific scale factor $\gamma_t(c)$ as in Equation (4),

$$\gamma_t(c) = \mathcal{M}(\beta_t(c)). \tag{4}$$

## C. Training Details

In this section, we introduce the training details for indoor and outdoor datasets. **Training on Indoor Datasets.** During the pre-training stage, we use the Adam optimizer [1] with an initial learning rate of 0.008 and weight decay of 0.01. The batch size is set to 16, and we train the model for 360 epochs on a single NVIDIA GeForce RTX 3090 GPU, with the learning rate decayed by 0.1 at the 240th and 300th epochs. For the training stage, we form a batch by sampling 4 labeled samples and 8 unlabeled samples. The initial learning rate is set to 0.005, and we train the model for another 360 epochs with the same settings as the pre-training stage. The weight $\beta$ for the unsupervised loss is set to 2. **Training on Outdoor Dataset.** During the pre-training stage, we use the Adam optimizer with an initial learning rate 0.01 and weight decay 0.0025. We train the side-aware PV-RCNN [3] with a batch size of 32 for 80 epochs on 8 NVIDIA GeForce RTX 3090 GPUs. The learning rate is decayed by 0.1 at the 35th and 45th epochs. For the training stage, each batch consists of 8 labeled samples and 8 unlabeled samples. The other settings are the same as the pre-training stage, and the weight $\beta$ for the unsupervised loss is set to 1.

## D. Comparison with Different Methods

In this section, we first provide a comparison of the overhead between our method and other semi-supervised methods [6, 4]. We then present a more detailed comparison of performance metrics. Finally, we randomly select some test set samples for visualization and compare the visualization results of different methods.

### D.1. Overhead Comparison

In Table 1, we report the memory usage and time consumption for model training, including the pretraining stage and the training stage. All results are produced with the same experiment setting on a single GTX 3090 GPU. The runtime consumption is computed with a forward pass and

*Corresponding Author

Table 1. **Memory usage and runtime comparison of different methods.** Here we report the memory usage and time consumption for model training, including the pretraining stage and the training stage. The runtime consumption is calculated by performing a forward pass and a backward pass through the model.

| | Method | ScanNet | | SUNRGB-D | |
|---|---|---|---|---|---|
| | | Mem. (GB) | RunTime (s) | Mem. (GB) | RunTime (s) |
| Pretrain | VoteNet [2] | 16.573 | 0.322 | 16.527 | 0.301 |
| | SESS [6] | 16.573 | 0.322 | 16.527 | 0.301 |
| | 3DIoU [4] | 17.133 | 0.382 | 17.078 | 0.375 |
| | *Ours* | 18.253 | 0.391 | 18.211 | 0.381 |
| Train | SESS [6] | 12.281 | 0.403 | 12.264 | 0.391 |
| | 3DIoU [4] | 16.909 | 0.915 | 16.868 | 0.901 |
| | *Ours* | 17.789 | 0.951 | 17.767 | 0.919 |

a backward pass for the both pretraining and training stage. The memory usage is computed with batch size 16 for the pretraining stage and batch size 12 for the training stage (4 labeled samples and 8 unlabeled samples). During the pretraining stage, since we have an extra uncertainty estimation network and IoU prediction network, the memory usage and time consumption are both slightly increased. During the training stage, our method is similar in speed to the pseudo-label based method 3DIoUMatch [4], but is slower than the consistency based method SESS [6]. This is because both our method and 3DIoUmatch need extra time to generate pseudo-labels.

### D.2. Detailed Comparison of Metrics

We present per-category results using 50% labeled data on both ScanNet and SUNRGB-D datasets. Table 2 and Table 3 show the mAP@0.25 and mAP@0.5 for each category on the ScanNet 50% labeled data. Similarly, Table 4 and Table 5 show the mAP@0.25 and mAP@0.5 for each category on the SUNRGB-D 50% labeled data. These results indicate that the proposed method achieves the best results for most categories under both datasets. Our method shows superior performance in detecting objects belonging to certain challenging categories such as tables, pictures, windows, and bookshelves. For the outdoor datset, we divide all objects into three difficulty levels according to the height range, the occlusion level and the truncation of the bounding box. In the main paper, we report the results of the moderate difficulty level for all three categories. Table 6 shows the results of different levels on KITTI 1% labeled data. From the results, we can see that the proposed method achieves the best results for all classes and all difficulty levels.

### D.3. Visual Comparison

In Figure 1 and Figure 2, we present visual comparisons of the detection results obtained by different methods on the ScanNet 50% labeled data and the SUNRGB-D 50% labeled data, respectively. The results demonstrate that our method achieves superior performance with higher localization quality. This is attributed to the fact that we assign different weights to the sides with different localization qualities, which in turn helps to improve the localization ability by learning from sides with higher quality.

## E. More Experiments

In this section, we provide additional ablation experiments to further validate the impact of each module in the model on performance. Additionally, we visualize pseudo-labels to further illustrate the significance of our proposed method in the field of semi-supervised 3D object detection.

### E.1. More Ablation Studies

Table 7 illustrates the effect of the number of bins in side probability distributions on the model performance. Increasing the number of bins leads to a slight improvement in model performance, while decreasing the number of bins significantly reduces the model's performance. This indicates that the granularity of the distribution plays a critical role in uncertainty estimation and model performance. To balance computational efficiency and performance, we set the number of bins to 32.

We provide additional ablation experiments to investigate the effect of different distribution properties on the model performance. As shown in Table 8, besides using the distribution values, we introduce three statistical measures (top-k mean, variance, entropy) that reflect the flatness of the distribution. The Top-k mean is insensitive to relative shifts over the distribution, resulting in a robust representation that is independent of the object scale. Based on the experimental results, we ultimately use the distribution values, top-k mean, and variance as distribution property inputs into the uncertainty estimation module.

### E.2. Visualization of Pesudo-labels

To further demonstrate the significance of the proposed side-aware method for semi-supervised 3D object detection,

Table 2. **Pre-category mAP@0.25 on ScanNet 50% labeled data.** Results are reported as mean ± standard deviation across 3 runs with random data splits.

| Method | cabinet | bed | chair | sofa | table | door | window | bkshf | picture |
|---|---|---|---|---|---|---|---|---|---|
| VoteNet [2] | 31.9±1.5 | 85.5±0.6 | 86.1±0.9 | 82.0±0.4 | 57.2±0.9 | 45.8±1.3 | 29.9±1.8 | 46.4±1.4 | 7.5±1.1 |
| SESS [6] | 41.0±1.6 | 86.4±1.2 | 88.2±1.1 | 88.7±0.8 | 59.8±1.1 | 49.5±1.5 | 35.7±1.9 | 52.8±0.5 | 10.6±0.6 |
| 3DIoU [4] | 44.2±0.5 | **87.3±0.7** | 88.4±0.4 | 91.0±0.3 | 59.1±1.0 | **51.8±1.3** | 37.1±0.5 | 51.9±0.9 | 11.4±0.8 |
| **Ours** | **46.3±1.1** | 86.8±0.5 | **89.1±0.3** | **91.3±0.8** | **64.4±0.9** | 50.9±1.3 | **39.6±1.8** | **55.5±1.1** | **15.8±1.7** |
| **Method** | **counter** | **desk** | **curtain** | **fridg** | **showr** | **toilet** | **sink** | **bathtub** | **ofurn** |
| VoteNet [2] | 68.1±0.9 | 67.3±1.2 | 44.1±1.0 | 46.2±1.6 | 63.4±0.7 | 96.5±1.2 | 34.8±1.5 | 89.4±0.6 | 29.5±0.9 |
| SESS [6] | 60.9±1.1 | 67.9±1.1 | 36.7±1.9 | 44.5±0.8 | **64.1±0.4** | 98.8±0.3 | 32.9±1.6 | 92.1±0.8 | 37.5±1.7 |
| 3DIoU [4] | 65.1±0.7 | 65.1±0.8 | 41.9±1.3 | 49.4±1.2 | 61.1±0.9 | 98.6±0.4 | **43.3±0.8** | 89.9±0.5 | 37.5±0.9 |
| **Ours** | **69.1±0.8** | **74.3±0.5** | **44.8±1.2** | **51.2±0.5** | 54.9±2.2 | **99.8±0.2** | 42.8±1.5 | **92.2±1.1** | **38.6±0.6** |

Table 3. **Pre-category mAP@0.50 on ScanNet 50% labeled data.** Results are reported as mean ± standard deviation across 3 runs with random data splits.

| Method | cabinet | bed | chair | sofa | table | door | window | bkshf | picture |
|---|---|---|---|---|---|---|---|---|---|
| VoteNet [2] | 8.7±1.2 | 70.9±0.9 | 68.8±1.0 | 69.6±0.8 | 44.5±1.1 | 16.4±1.6 | 8.7±1.8 | 36.5±0.5 | 1.2±1.1 |
| SESS [6] | 12.3±0.8 | 75.7±0.4 | 71.9±0.4 | 74.1±0.7 | 51.2±0.8 | 18.6±1.3 | 9.6±1.5 | 43.2±0.6 | 2.2±0.8 |
| 3DIoU [4] | 12.5±1.1 | **76.7±0.5** | 73.8±0.9 | **81.7±0.4** | 49.3±0.3 | 26.1±0.5 | **14.6±0.5** | 42.6±0.6 | 4.2±1.2 |
| **Ours** | **18.6±2.1** | 74.1±0.6 | **76.5±0.6** | 81.3±0.9 | **56.7±1.5** | **26.4±1.1** | 13.6±0.8 | **45.9±1.2** | **6.5±1.3** |
| **Method** | **counter** | **desk** | **curtain** | **fridg** | **showr** | **toilet** | **sink** | **bathtub** | **ofurn** |
| VoteNet [2] | 29 4±1.0 | 39.0±0.8 | 24.6±1.2 | 35.2±0.7 | 2.1±0.9 | 85.8±0.4 | 14.9±1.3 | 80.9±0.3 | 13.8±1.6 |
| SESS [6] | 19.7±1.2 | 38.6±0.6 | 25.0±0.9 | 33.4±0.7 | **3.7±0.8** | **89.7±0.6** | 15.3±1.7 | 89.6±0.5 | 19.7±1.2 |
| 3DIoU [4] | 31.1±0.8 | 40.4±0.5 | 29.1±1.1 | **33.8±0.9** | 2.3±1.1 | 84.8±0.4 | 24.9±1.1 | **89.7±0.8** | 20.8±1.5 |
| **Ours** | **42.3±0.7** | **48.9±0.9** | **29.2±1.0** | 33.3±1.7 | 2.1±1.8 | 85.8±0.5 | **25.5±1.1** | 82.9±0.6 | **26.1±1.2** |

Table 4. **Pre-category mAP@0.25 on SUNRGB-D 50% labeled data.** Results are reported as mean ± standard deviation across 3 runs with random data splits.

| Method | bed | table | sofa | chair | toilet | desk | dresser | nights | bkshf | bathtub |
|---|---|---|---|---|---|---|---|---|---|---|
| VoteNet [2] | 82.2±0.6 | 47.0±0.9 | 61.1±0.7 | 76.7±0.4 | 85.8±0.4 | 16.6±2.7 | 28.3±0.5 | 54.7±0.2 | 23.5±1.7 | 74.4±0.3 |
| SESS [6] | 83.5±0.3 | 48.8±1.1 | 63.0±0.5 | 77.7±0.6 | 86.7±0.2 | 20.3±1.6 | 30.3±0.7 | 56.1±0.5 | 29.0±0.9 | **79.8±1.2** |
| 3DIoU [4] | 83.9±0.8 | 48.5±0.4 | 65.2±0.9 | 77.3±0.2 | 87.6±0.7 | 25.8±1.1 | 29.8±1.1 | 56.8±0.4 | 29.4±1.4 | 78.9±0.6 |
| **Ours** | **85.5±0.5** | **54.1±0.8** | **67.4±0.6** | **78.9±0.8** | **90.6±0.4** | **27.3±0.7** | **31.4±0.9** | **62.3±0.6** | **32.3±0.2** | 71.2±0.4 |

Table 5. **Pre-category mAP@0.50 on SUNRGB-D 50% labeled data.** Results are reported as mean ± standard deviation across 3 runs with random data splits.

| Method | bed | table | sofa | chair | toilet | desk | dresser | nights | bkshf | bathtub |
|---|---|---|---|---|---|---|---|---|---|---|
| VoteNet [2] | 46.1±0.4 | 19.5±0.7 | 46.2±0.5 | 57.8±0.4 | 52.6±0.3 | 3.1±1.1 | 14.2±0.6 | 30.8±0.3 | 2.1±0.6 | 47.3±0.5 |
| SESS [6] | 41.7±0.9 | 20.2±0.5 | 48.4±1.3 | 58.3±1.2 | 57.2±0.5 | 4.7±0.6 | 15.6±0.2 | 38.7±0.6 | 3.3±0.5 | **50.6±0.7** |
| 3DIoU [4] | 53.0±0.7 | 22.9±0.6 | 50.1±0.7 | 59.5±0.5 | 59.0±0.4 | 7.1±1.3 | 17.5±0.7 | 36.8±0.4 | 5.4±0.8 | 45.3±0.2 |
| **Ours** | **55.3±0.3** | **26.2±1.1** | **51.3±0.6** | **60.0±0.8** | **62.7±1.2** | **8.4±0.9** | **20.9±0.5** | **42.3±0.2** | **8.1±1.2** | 42.9±0.4 |

we visualize the ground-truth and pseudo-labels generated by the pre-trained model in Figure 3 and Figure 4. In addition, we also provide the uncertainty estimation value of each side for the pesudo-labels with the red circle. Due to the incompleteness and irregular shape of the objects, there are many detection results with poor localization qual-

Table 6. **Results on on KITTI 1% labeled data.** TThe results for all difficulty levels are evaluated by the mAP with 40 recall positions.

| Model | Car | | | Pedestrian | | | Cyclist | | |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| PVRCNN [3] | 87.4±0.0 | 73.1±0.2 | 66.9±0.8 | 30.2±14.9 | 21.4±11.1 | 19.2±9.5 | 47.1±10.1 | 28.0±6.0 | 26.1±5.8 |
| 3DIoU [4] | 88.0±1.9 | 75.2±1.8 | 69.8±1.0 | 36.8±18.7 | 32.9±16.1 | 27.4±12.0 | 48.7±17.3 | 29.4±10.8 | 27.5±10.1 |
| **Ours** | **89.3±0.7** | **76.3±1.0** | **70.7±0.5** | **38.3±17.3** | **33.1±13.6** | **30.1±11.3** | **55.8±7.6** | **33.6±5.2** | **30.6±4.7** |

Table 7. **Effect of the number of bins of different side probability distributions on model performance.** $N$ denotes the number of bins.

| Number of bins | ScanNet 20% | | ScanNet 50% | |
|---|---|---|---|---|
| | $mAP_{25}$ | $mAP_{50}$ | $mAP_{25}$ | $mAP_{50}$ |
| N = 16 | 53.76 | 36.83 | 60.84 | 42.26 |
| N = 24 | 54.08 | 37.11 | 61.34 | 42.87 |
| N = 32 | 54.51 | 37.29 | 61.52 | 43.13 |
| N = 64 | **54.81** | 37.59 | **61.89** | **43.51** |
| N = 96 | 54.71 | **37.66** | 61.39 | 43.22 |

Table 8. **Effect of different distribution properties on uncertainty estimation of each sied.** "All Values" refers to using all distribution values as input. "Top-k Mean" involves selecting the top-k values and computing their mean value as the property. "Variance" and "Entropy" correspond to calculating the distribution variance and entropy as properties, respectively.

| All Values | Top-k Mean | Variance | Entropy | ScanNet 20% | | ScanNet 50% | |
|---|---|---|---|---|---|---|---|
| | | | | $mAP_{25}$ | $mAP_{50}$ | $mAP_{25}$ | $mAP_{50}$ |
| ✗ | ✗ | ✗ | ✗ | 52.56 | 35.15 | 60.26 | 41.82 |
| ✓ | ✗ | ✗ | ✗ | 52.98 | 35.15 | 60.26 | 41.82 |
| ✓ | k=4 | ✗ | ✗ | 53.84 | 36.51 | 60.93 | 42.85 |
| ✓ | k=8 | ✗ | ✗ | 54.28 | 36.88 | 61.21 | 43.01 |
| ✓ | k=12 | ✗ | ✗ | 53.45 | 36.26 | 60.49 | 42.74 |
| ✓ | k=8 | ✓ | ✗ | **54.51** | 37.29 | **61.52** | **43.13** |
| ✓ | k=8 | ✓ | ✓ | 54.31 | **37.38** | 61.45 | 43.07 |

ity in the pseudo-labels. Our method addresses this issue by predicting the uncertainty of each side and focusing on the sides with higher localization quality while ignoring the sides with poor localization quality during model training.

## References


[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[2] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3

[3] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 1, 4

[4] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021. 1, 2, 3, 4

[5] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[6] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 1, 2, 3

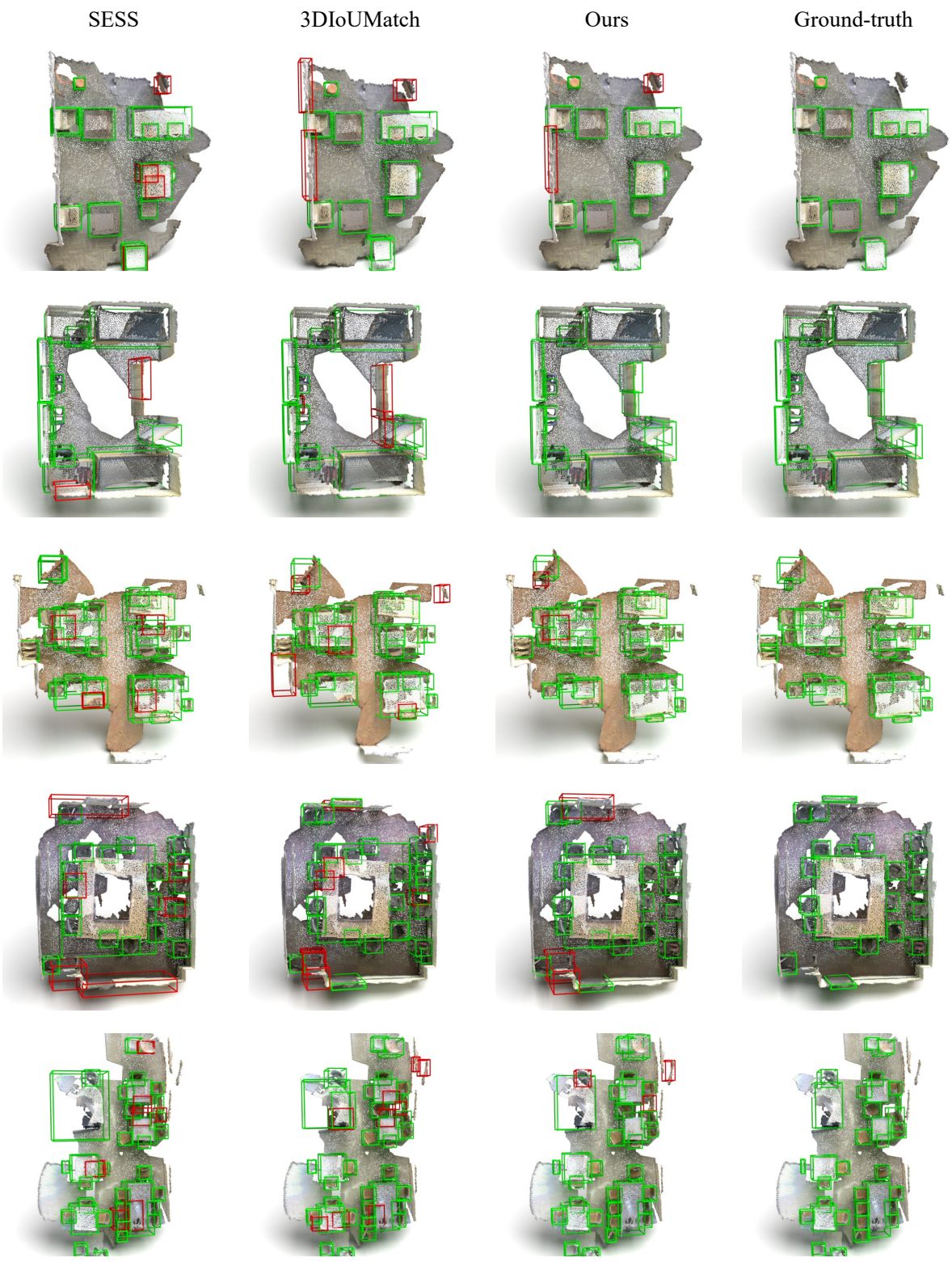


Figure 1. **Visual comparisons of the detection results on ScanNet 50% labeled data.** Here green bounding boxes have $IoU \geq 0.25$ and red bounding boxes have $IoU < 0.25$.

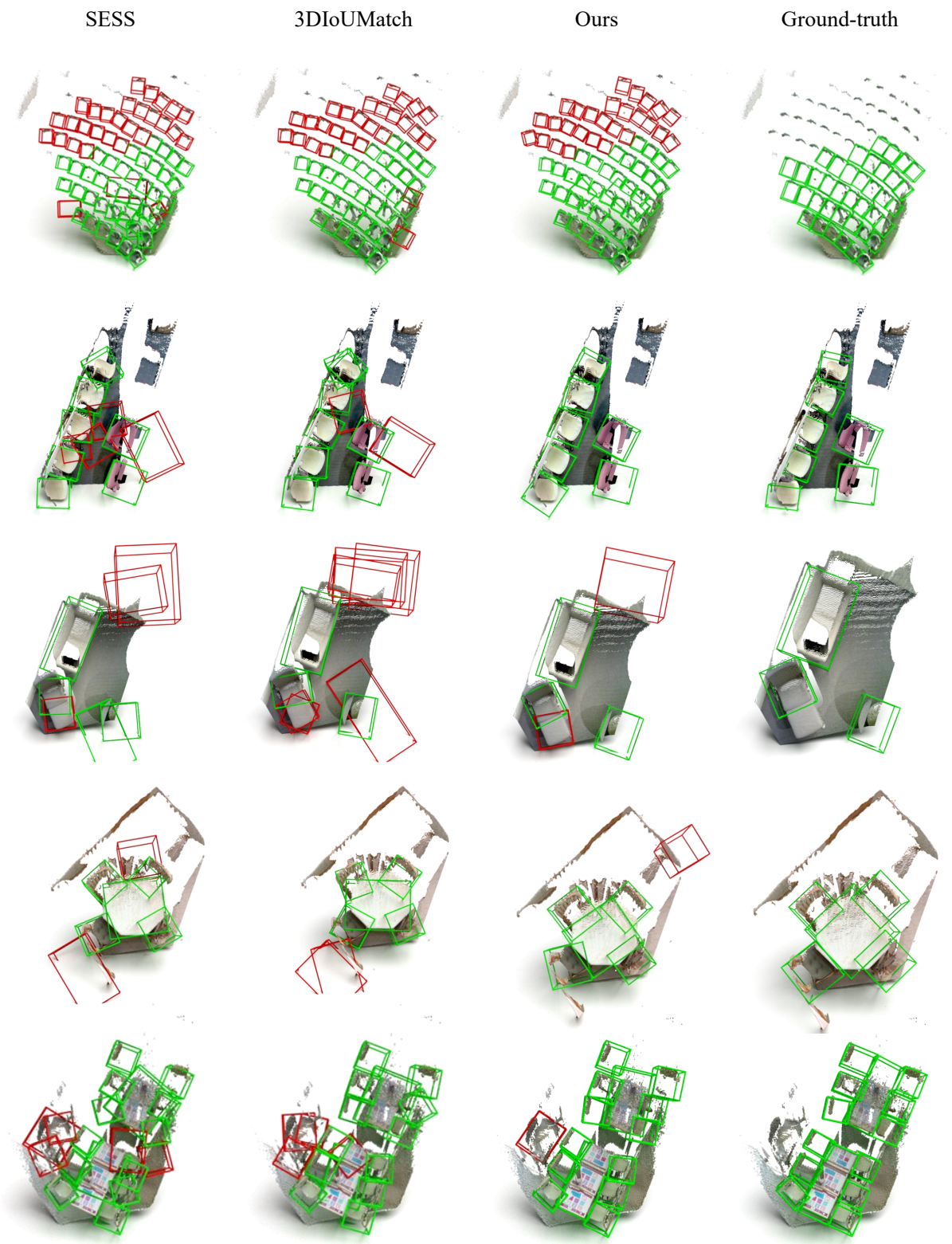


Figure 2. **Visual comparisons of the detection results on SUNRGB-D 50% labeled data.** Here green bounding boxes have $IoU \geq 0.25$ and red bounding boxes have $IoU < 0.25$.

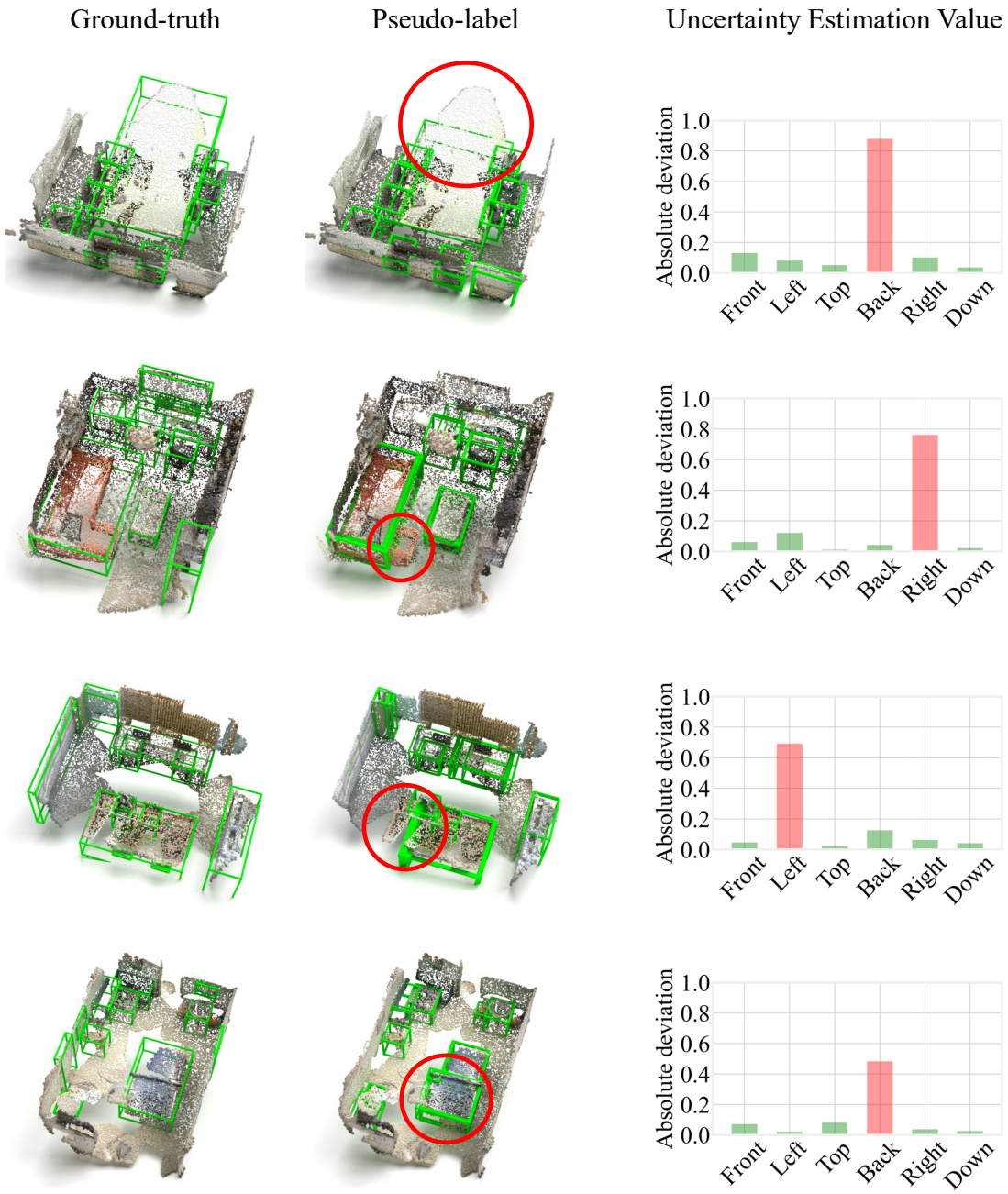


Figure 3. **Pesudo-labels and ground-truth visualization on the ScanNet dataset.** We pre-train the model with 50% labeled data and then generate pseudo-labels by threshold on the classification score and IoU score. Sides with poor localization quality are marked with red circles, the results indicate that there are many pseudo bounding boxes with poor localization quality on several sides.

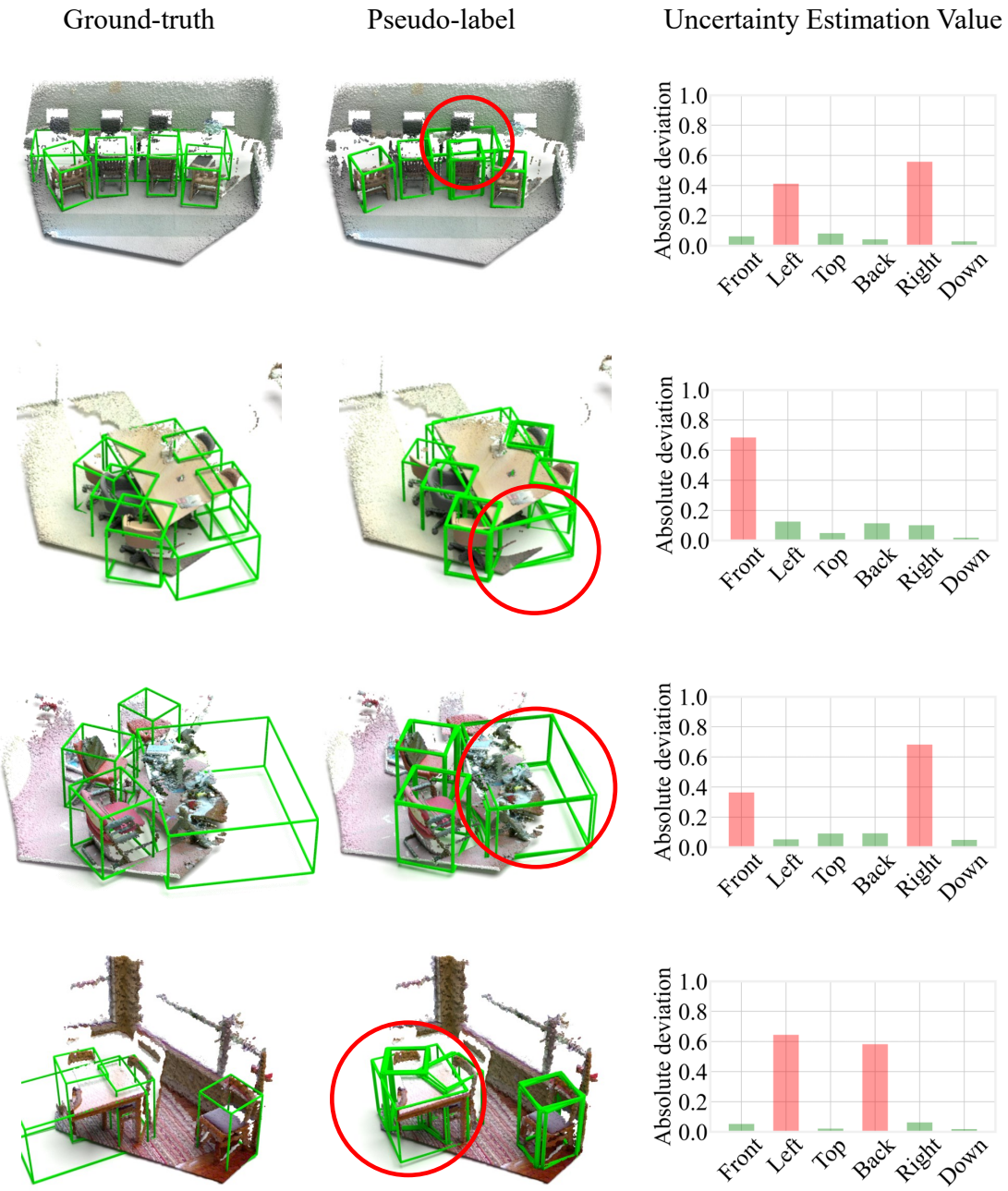


Figure 4. **Pesudo-labels and ground-truth visualization on the SUNRGB-D dataset.** We pre-train the model with 50% labeled data and then generate pseudo-labels by threshold on the classification score and IoU score. Sides with poor localization quality are marked with red circles, the results indicate that there are many pseudo bounding boxes with poor localization quality on several sides.