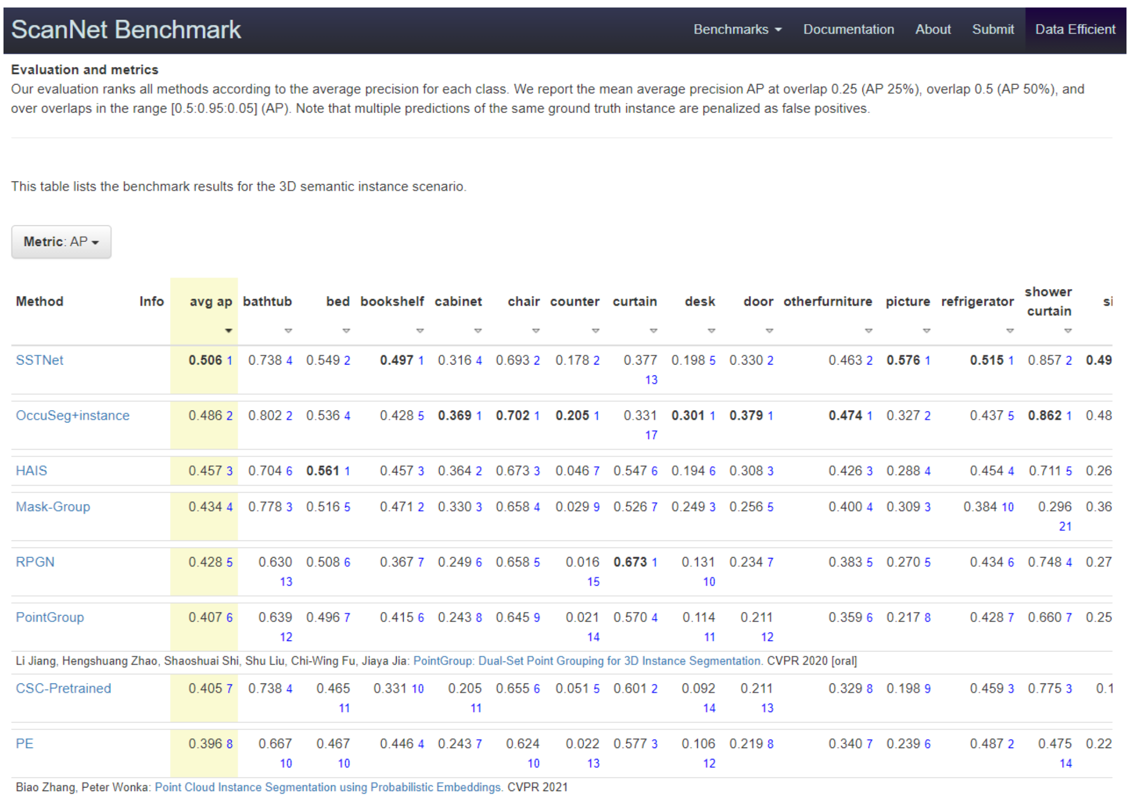
Instance Segmentation in 3D Scenes using Semantic Superpoint Tree Networks — Supplementary Material

Zhihao Liang[1,2], Zhihao Li[3], Songcen Xu[3], Mingkui Tan[1] and Kui Jia[1,4,5]*
[1]South China University of Technology, [2]DexForce Technology Co., Ltd.
[3]Noah's Ark Lab, Huawei Technologies, [4]Pazhou Laboratory, [5] Peng Cheng Laboratory
eezhihaoliang@mail.scut.edu.cn, {kuijia, mingkuitan}@scut.edu.cn,
{zhihao.li, xusongcen}@huawei.com



ScanNet Benchmark | Benchmarks | Documentation | About | Submit | Data Efficient

Evaluation and metrics
Our evaluation ranks all methods according to the average precision for each class. We report the mean average precision AP at overlap 0.25 (AP 25%), overlap 0.5 (AP 50%), and over overlaps in the range [0.5:0.95:0.05] (AP). Note that multiple predictions of the same ground truth instance are penalized as false positives.

This table lists the benchmark results for the 3D semantic instance scenario.

Metric: AP

Method	Info	avg ap	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	otherfurniture	picture	refrigerator	shower curtain	si
SSTNet		**0.506** 1	0.738 4	0.549 2	**0.497** 1	0.316 4	0.693 2	0.178 2	0.377 13	0.198 5	0.330 2	0.463 2	**0.576** 1	**0.515** 1	0.857 2	**0.49**
OccuSeg+instance		0.486 2	0.802 2	0.536 4	0.428 5	**0.369** 1	**0.702** 1	**0.205** 1	0.331 17	**0.301** 1	**0.379** 1	**0.474** 1	0.327 2	0.437 5	**0.862** 1	0.48
HAIS		0.457 3	0.704 6	**0.561** 1	0.457 3	0.364 2	0.673 3	0.046 7	0.547 6	0.194 6	0.308 3	0.426 3	0.288 4	0.454 4	0.711 5	0.26
Mask-Group		0.434 4	0.778 3	0.516 5	0.471 2	0.330 3	0.658 4	0.029 9	0.526 7	0.249 3	0.256 5	0.400 4	0.309 3	0.384 10	0.296 21	0.36
RPGN		0.428 5	0.630 13	0.508 6	0.367 7	0.249 6	0.658 5	0.016 15	**0.673** 1	0.131 10	0.234 7	0.383 5	0.270 5	0.434 6	0.748 4	0.27
PointGroup		0.407 6	0.639 12	0.496 7	0.415 6	0.243 8	0.645 9	0.021 14	0.570 4	0.114 11	0.211 12	0.359 6	0.217 8	0.428 7	0.660 7	0.25
Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, Jiaya Jia: PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. CVPR 2020 [oral]																
CSC-Pretrained		0.405 7	0.738 4	0.465 11	0.331 10	0.205 11	0.655 6	0.051 5	0.601 2	0.092 14	0.211 13	0.329 8	0.198 9	0.459 3	0.775 3	0.1
PE		0.396 8	0.667 10	0.467 10	0.446 4	0.243 7	0.624 10	0.022 13	0.577 3	0.106 12	0.219 8	0.340 7	0.239 6	0.487 2	0.475 14	0.22
Biao Zhang, Peter Wonka: Point Cloud Instance Segmentation using Probabilistic Embeddings. CVPR 2021																

Figure 1: The snapshot from ScanNet (V2) benchmark on March 18th 2021. Our SSTNet ranks top on the mAP leaderboard.

1. Network Specifics

In this section, we present architectural specifics of our proposed Semantic Superpoint Tree Network (SSTNet).

1.1. Backbone and Learning Branches

Fig.2-(a) illustrates the architecture of our backbone, where we employ a U-Net [4] style network with a depth of 5. Fig.2-(b) illustrates the branch specifics of semantic scoring and offset prediction.

1.2. The Classifier for Tree Traversal and Splitting

Fig. 3 presents the multi-layer perceptron (MLP) of the binary classifier ϕ that is used for generation of object proposals, where we also show how the classifier is used when traversing the tree.

1.3. CliqueNet

An illustration on how a tree branch can be converted as a graph clique and the thus constructed CliqueNet.

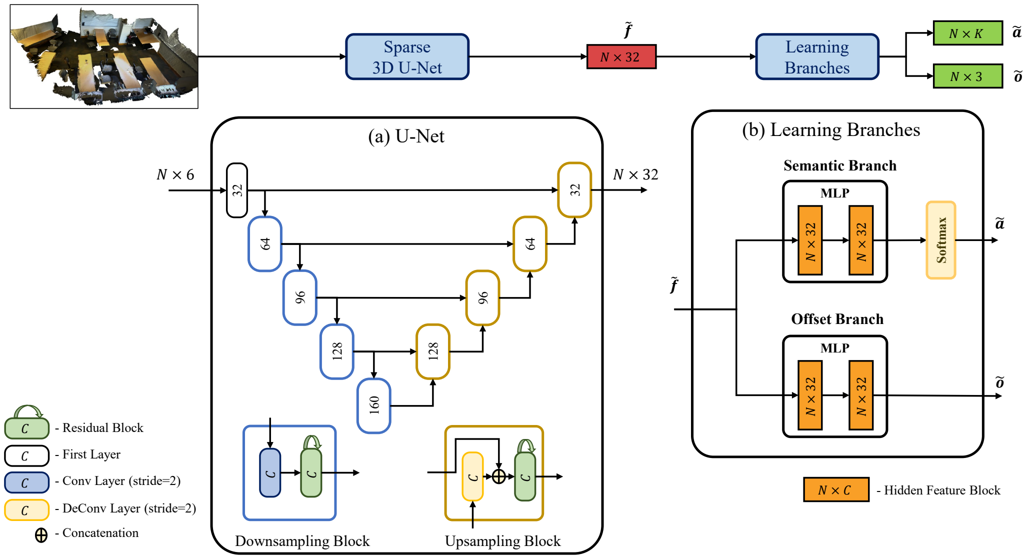


Figure 2: Module specifics of the backbone, semantic scoring, and offset prediction used in SSTNet. N is the number of input points, and numbers in each block denote those of output channels.

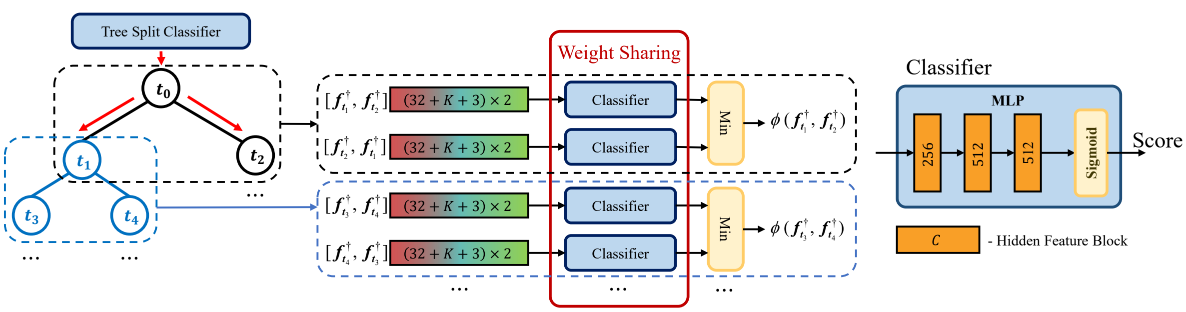


Figure 3: An illustration on the node-splitting classifier ϕ and how it is used when traversing the semantic superpoint tree.

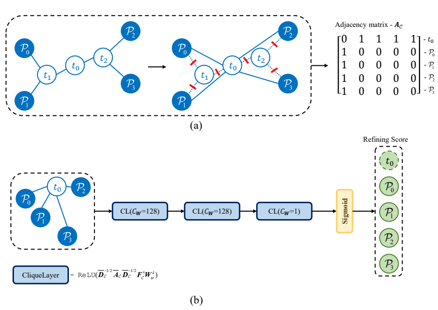


Figure 4: An illustration on how a tree branch can be converted as a graph clique and the thus constructed CliqueNet. $C_{\boldsymbol{w}}$ denotes the number of output channels in each CliqueLayer (CL).

Fig. 4 illustrates how a tree branch can be converted as a graph clique and also the construction of CliqueNet ψ. Given an input feature $\boldsymbol{F}_{\mathcal{C}}^{\dagger}$, the i^{th} layer of the CliqueNet (i.e., the i^{th} CliqueLayer) performs the following computation

$$\texttt{ReLU}(\bar{\boldsymbol{D}}_{\mathcal{C}}^{-1/2}\bar{\boldsymbol{A}}_{\mathcal{C}}\bar{\boldsymbol{D}}_{\mathcal{C}}^{-1/2}\boldsymbol{F}_{\mathcal{C}}^{\dagger}\boldsymbol{W}_{\psi}^{i}), \tag{1}$$

where the adjacency matrix $\boldsymbol{A}_{\mathcal{C}}$ is shown in Fig. 4-(a), $\bar{\boldsymbol{A}}_{\mathcal{C}} = \boldsymbol{A}_{\mathcal{C}} + \boldsymbol{I}$, and $\bar{\boldsymbol{D}}_{\mathcal{C}}$ is the diagonal degree matrix of $\bar{\boldsymbol{A}}_{\mathcal{C}}$. CliqueNet specifics are given in Fig. 4-(b).

2. Traing of the Proposal Evaluation Module

We follow [3] and use a ScoreNet (denoted as ω) to evalaute the proposals refined by CliqueNet. For such a proposal $\mathcal{B}_t^-$, we get the corresponding point-wise features $\widetilde{\boldsymbol{F}}_{\mathcal{B}_t^-} = [\tilde{\boldsymbol{f}}_1, \dots, \tilde{\boldsymbol{f}}_{N_t^-}] \in \mathbb{R}^{n \times N_t^-}$, and use the following

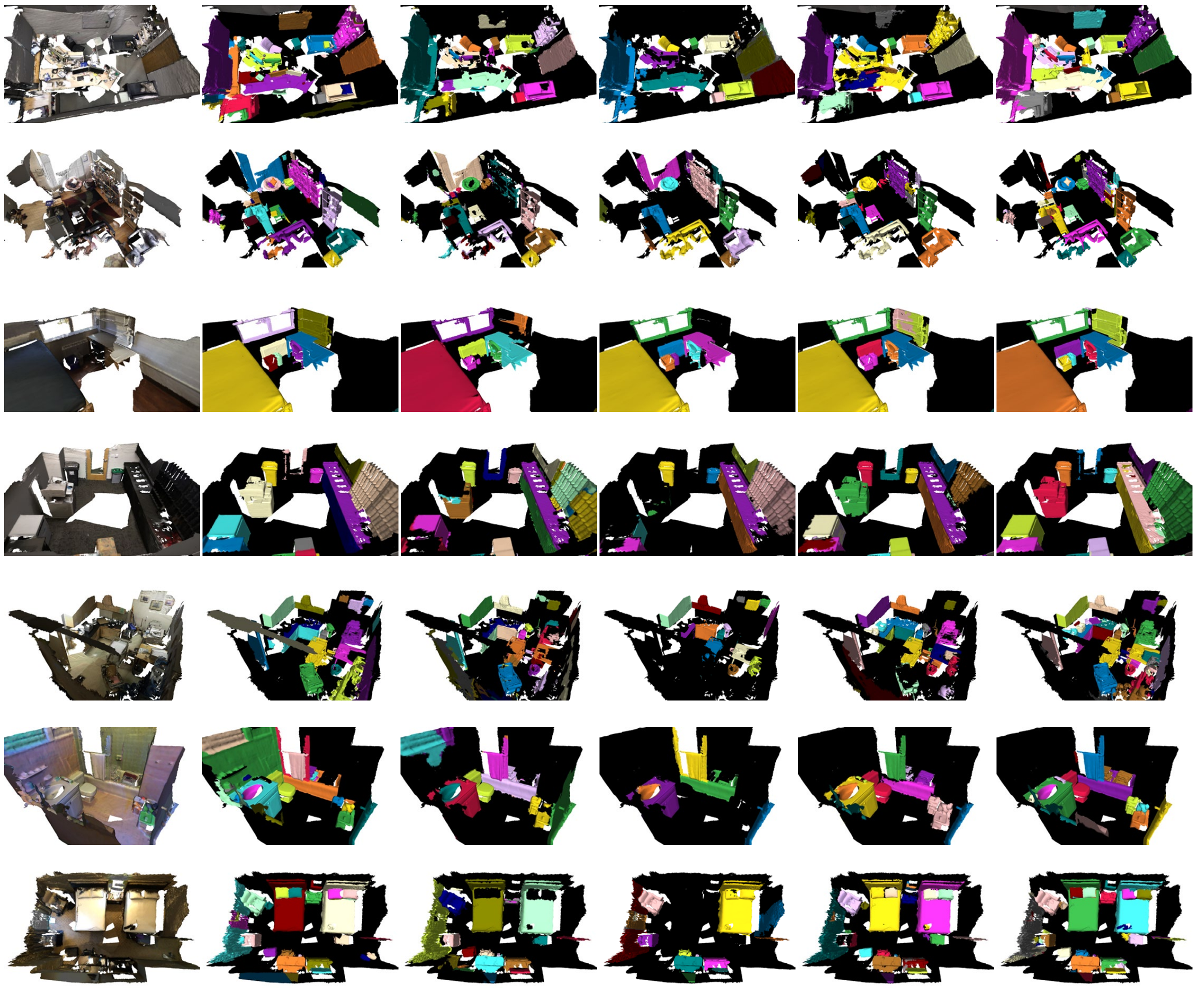


Figure 5: Visualization of the semantic and instance segmentation results on the validation set of ScanNet v2 (top) and S3DIS (bottom).

loss to train the ScoreNet

$$L_{\text{evaluation}} = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \texttt{BCE}(\omega(\widetilde{\boldsymbol{F}}_{\mathcal{B}_t^-}), v_t^*), \tag{2}$$

where the value v_t^* of supervision used in binary cross-entropy loss (BCE) is determined by the Intersection over Union (IoU) between the proposal $\mathcal{B}_t^-$ and its best matched ground-truth instance; we denote the IoU value as $\text{IoU}_{\mathcal{B}_t^-}$. Given $\text{IoU}_{\mathcal{B}_t^-}$, v_t^* is determined as

$$v_t^* = \begin{cases} 0 & \text{if} \quad \text{IoU}_{\mathcal{B}_t^-} < \theta_l \\ 1 & \text{if} \quad \text{IoU}_{\mathcal{B}_t^-} > \theta_h \\ \frac{1}{\theta_h - \theta_l}(\text{IoU}_{\mathcal{B}_t^-} - \theta_l) & \quad \text{otherwise} \end{cases} \tag{3}$$

where we set the hyperparameters $\theta_l = 0.25$ and $\theta_h = 0.75$.

3. Results Comparison Visualization

In this section, we show more comprehensive comparisons with 3D-MPA[2], SSEN[5] and PointGroup[3] on ScanNet(V2)[1]. As shown in Fig. 5, our results can better maintain the boundaries and the integrity of the segmentation results.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[2] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9031–9040, 2020. 3

[3] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[5] Dongsu Zhang, Junha Chun, Sang Cha, and Young Min Kim. Spatial semantic embedding network: Fast 3d instance segmentation with deep metric learning. In *arXiv preprint arXiv:2007.03169*, 2020. 3