Supplementary Material
Box2Mask: Weakly Supervised 3D Semantic Instance Segmentation Using Bounding Boxes

Julian Chibane[1,2], Francis Engelmann[3], Tuan Anh Tran[2], and Gerard Pons-Moll[1,2]

[1] University of Tübingen, Germany
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3] ETH Zurich AI Center, Switzerland

A Baseline: Object Detector followed by Segmentation

In the main paper, we address how per-point instance masks can be learned from bounding box annotations only. To show that this is a non-trivial task, and that our proposed method generalizes beyond the weak supervision signal, we present an additional baseline experiment. This baseline is an object detector predicting bounding boxes and is trained on the given ground truth bounding box annotations. Then, the instance masks are obtained by segmenting the points inside each predicted bounding box into foreground and background. The baseline implementation closely follows the implementation of our main model: using a sparse convolutional network [2] we obtain deep learned features for each point in the input point cloud. The learned point-features then vote for object bounding box proposals. These steps are identical to the first part our main model shown in Fig. 2 of the main paper. We then perform non-maximum-suppression (NMS) to obtain object detection bounding boxes from the proposals. The final instance masks are obtained from the predicted bounding boxes, which are segmented into foreground and background based on the number of bounding boxes each point is contained in. This is the same mechanism as described in the main paper to obtain per-point supervision signals (Sec. 5, Eq. 2 in the main paper). By doing so, it is guaranteed that the baseline is directly comparable with the proposed weakly-supervised approach. Visual results, including the object detections are shown in Fig. 1. Scores are shown in Tab. 1. Our proposed approach largely outperforms this baseline (+11.8 mAP_{50}). In particular, this experiment shows that learning instance masks from bounding box annotations alone is non-trivial, and that our trained model is able to generalize beyond the weak training signal obtained from the bounding box annotations.

B Per-Category Results

In this section, we show per-category results on the ScanNet validation and test splits, and on S3DIS 6-fold cross validation, as summarised in Tab. 2, 3, 4 and 5. On ScanNet validation and S3DIS, we show also per-category scores for the fully-supervised model trained with per-point instance labels.

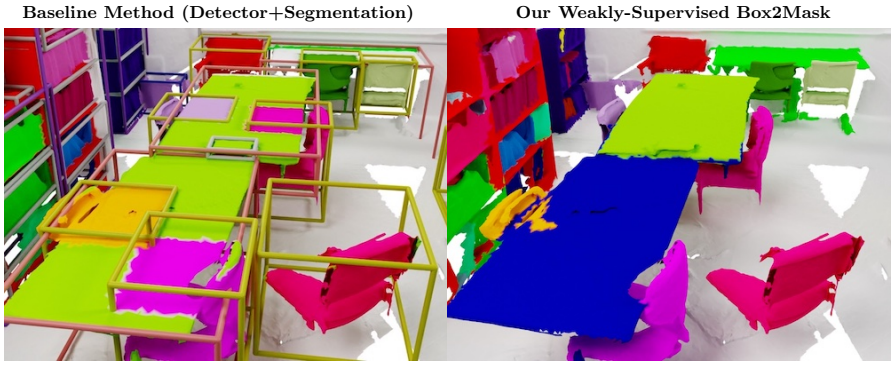

Fig. 1: Qualitative comparison of the baseline *(left)* and our approach *(right)*. For the baseline, the outputs of the object detector and the subsequent foreground background segmentations are shown. The baseline fails whenever two object bounding boxes are intersecting (table top). While our Box2Mask is supervised with comparable labels during training, it learns to generalize beyond these weak labels and infers the correct instance masks for objects with intersecting bounding boxes (see chairs and table).

	mAP	mAP_{50}	mAP_{25}
Baseline (ours)	26.5	47.9	64.8
Box2Mask (ours)	**39.1** (+12.6)	**59.7** (+11.8)	**71.8** (+7.0)

Table 1: Comparison of our approach to the baseline (object detector followed by segmentation) on ScanNet validation set, trained with bounding box supervision only. The results indicate that obtaining instance masks from bounding boxes is non-trivial and that our training technique efficiently leverages weak bounding box annotations to predict dense and accurate instance masks. This is further visualized in Fig. 1.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg.**
Ours (mAP)	28.3	46.3	62.5	71.1	30.9	26.0	27.2	43.3	32.2	10.3	12.5	29.8	47.2	70.1	88.2	36.3	74.1	42.4	43.3
Ours (mAP@50%)	50.9	84.7	81.6	85.2	57.8	56.2	48.8	77.1	44.8	27.7	48.2	55.8	79.0	100	99.7	66.6	100	64.0	67.7
Ours (mAP@25%)	70.7	96.2	88.7	90.2	75.3	71.5	63.7	87.4	46.9	68.6	96.1	59.8	70.0	100	99.7	91.2	100	69.4	80.3

Table 2: Instance Segmentation on ScanNetV2 [3] Test Set. Trained only on *bounding boxes* on training and validation splits, no per-point annotations used.

	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**avg.**
Ours (mAP)	27.6	40.2	74.0	52.5	33.2	25.9	24.2	25.9	27.8	8.4	16.9	34.5	32.9	42.9	80.5	42.9	70.0	43.6	39.1
Ours (mAP@50%)	48.0	72.0	91.8	77.5	62.9	48.6	43.3	49.9	40.9	27.9	44.3	51.8	43.4	56.8	96.9	72.7	87.1	59.6	59.7
Ours (mAP@25%)	59.5	83.8	94.5	87.0	75.5	59.8	61.4	68.2	45.6	58.5	78.6	65.1	46.9	77.4	96.9	79.5	87.1	67.1	71.8

Table 3: Instance Segmentation on ScanNetV2 [3] Validation Set. Trained only on *bounding boxes* on the training split, no per-point annotations used during training.

	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookshelf	board	clutter	**avg.**
Ours (mPrec)	97.1	99.6	77.1	43.4	65.9	82.9	76.5	65.9	88.3	80.7	65.3	73.4	64.5	75.4
Ours (mRec)	68.3	95.6	64.1	63.2	66.6	83.9	88.4	55.5	69.7	68.6	50.6	69.1	58.0	69.4

Table 4: Instance Segmentation on S3DIS [1] 6-fold cross validation. Models are trained *fully supervised* with per-point semantic instance annotations.

	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookshelf	board	clutter	**avg.**
Ours (mPrec)	96.8	99.2	76.4	46.9	54.1	68.1	72.9	59.9	87.6	76.8	67.5	70.4	62.7	72.3
Ours (mRec)	68.1	95.3	64.0	67.5	63.8	77.0	90.7	60.0	70.4	68.9	53.4	79.9	57.7	70.5

Table 5: Instance Segmentation on S3DIS [1] 6-fold cross validation. Models are trained with only *bounding box supervision*, no per-point annotations used to train.

C Non-Maximum-Clustering (NMC) Algorithm

In Sec. 4 of the main paper, we introduced a clustering algorithm tailored specifically towards bounding box votes. The pseudo-code is below. Further, we analyse the effect of the threshold parameter τ, which can be between 0 (all boxes in single cluster) and 1 (each box is a separate cluster). In Fig. 2, we report mask prediction scores on ScanNet validation, and find that $\tau \approx 0.3$ performs best.

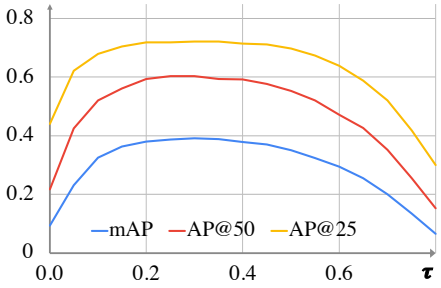

Fig. 2: Effect of parameter τ.

Algorithm 1: Non-Maximum-Clustering (NMC)

input : $P = (B, score)$ // Set of bounding box votes and corresponding scores
output: Clustered bounding box votes.
$P_{candidates} \leftarrow$ P.sort (score) // Sort bounding box votes based on score
Results $\leftarrow \emptyset$
while $P_{candidates} \neq \emptyset$ **do**
  $P_r \leftarrow P.pop()$ // Pop the highest scoring proposal
  $cluster \leftarrow \{p' \mid IoU(p_r.B, p'.B) > \tau \;\&\; p' \in P\}$ // Clustering with IoUs
  Results $\leftarrow$ Results $\cup \{cluster\}$ // Update the list of predictions
  $P_{candidates} \leftarrow P_{candidates} \setminus cluster$ // Remove the clustered votes from the // list of representative candidates
end
return *Results*

D Additional Qualitative Results

In Fig. 3, we show exemplary qualitative results of our method on the S3DIS dataset [1]. We show the 3D input scene, our predicted instance masks learned from weak bounding box annotations and the groundtruth instance masks as well as the groundtruth bounding box annotations for comparison. In Fig. 4 and Fig. 5, we show additional close-up qualitative results on the ScanNet dataset [3]. Besides results of our weakly-supervised model, we also show results of the same model fully-supervised with dense per-point labels. Notably, the predicted instance segmentation masks of the two models hardly differ, indicating that bounding box annotations are appropriate to train dense segmentation models.

Fig. 3: Qualitative Instance Segmentation Results on S3DIS [1] Individual instance masks are colored randomly and match the ground truth instance mask colors. During training, only bounding box annotations are used (last column), per-point instance masks (third column) are not used, and are shown here only for judging the quality of the predicted instance masks (second column).

Fig. 4: Qualitative Instance Segmentation Results on ScanNet [3] Individual instance masks are colored randomly and match the ground truth instance mask colors. Left: results from full per-point supervision. Right: weak bounding-box supervision.

Fig. 5: Qualitative Instance Segmentation Results on ScanNet [3] Individual instance masks are colored randomly and match the ground truth instance mask colors. Left: results from full per-point supervision. Right: weak bounding-box supervision.

E Bounding Box labels *v.s.* Full Point Labels

In this section, we analyse the question: "Is our initial point-to-box association strategy (Eq. 2, main paper) enough to obtain the good performance of our model?"

It is indeed correct, that this simple strategy can give good results (87% of current fully-supervised state-of-the-art models). It would, however, be wrong to assume that the differences between point and box labels are insignificant. To clearly investigate this aspect, we quantitatively compare the quality of the bounding box labels to the full point labels. Our bounding box labels achieve **70.4 mAP** (measured on ScanNet scenes) when evaluated against the full per-point labels (which naturally define 100 mAP). This is a performance gap of 30%. The reason for this difference are the "undecided" points that fall into multiple bounding boxes, Fig. 6. They are generally between two neighboring instances and make up **13.5%** of all points. It is specifically these points, that are crucial for learning accurate and sharp masks of adjacent instances.

Then how is it possible that our method still achieves close to fully-supervised scores? The reason is twofold: **1)** We observe generalization beyond the weak bounding box labels which enable precise masks on full instances (Fig. 1). During training, the model sees a large variety of scenes where the correctly supervised regions of objects outweigh the noisy ones. This likely allows our model to build specific priors of full instance masks such that the model learns to generalize beyond the weaker box labels.
2) Our novel algorithm for voting and clustering based on bounding boxes can fully leverage the weak supervision. This is shown in Tab. 4 (main paper) where our proposed bounding box approach largely outperforms prior center-based approaches (+8 mAP). This is the main factor enabling almost fully-supervised performance.

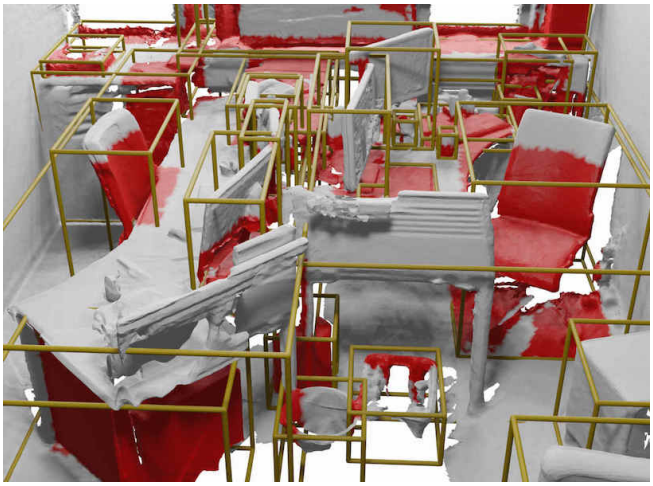

Fig. 6: ●: Undecided Points

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D Semantic Parsing of Large-Scale Indoor Spaces. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2, 3, 4
2. Choy, C., Gwak, J., Savarese, S.: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1
3. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2, 4, 5, 6