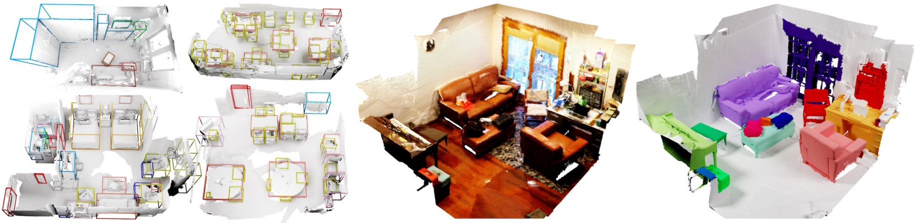
Box2Mask: Weakly Supervised 3D Semantic Instance Segmentation Using Bounding Boxes

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 Zürich, AI Center, Switzerland

virtualhumans.mpi-inf.mpg.de/box2mask



Training Data
3D Bounding Box Annotations
sparse, object-level

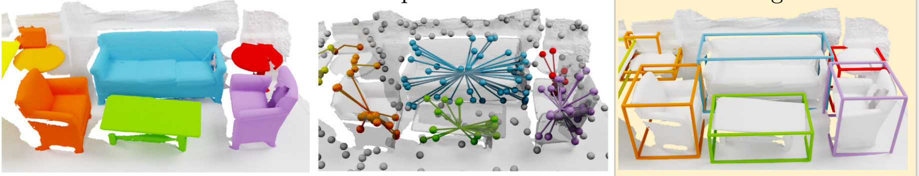
Input
3D Test Scene

Prediction
Instance Segmentation
dense, point-level

Fig. 1: Can we use 3D bounding box annotations alone to train dense 3D semantic instance segmentation models? We find that this is the case and propose a Deep Hough Voting based method that fully exploits bounding box annotations.

Abstract. Current 3D segmentation methods heavily rely on large-scale point-cloud datasets, which are notoriously laborious to annotate. Few attempts have been made to circumvent the need for dense per-point annotations. In this work, we look at weakly-supervised 3D semantic instance segmentation. The key idea is to leverage 3D bounding box labels which are easier and faster to annotate. Indeed, we show that it is possible to train dense segmentation models using only bounding box labels. At the core of our method, Box2Mask, lies a deep model, inspired by classical Hough voting, that directly votes for bounding box parameters, and a clustering method specifically tailored to bounding box votes. This goes beyond commonly used center votes, which would not fully exploit the bounding box annotations. On ScanNet test, our weakly supervised model attains leading performance among other weakly supervised approaches (+18 mAP_{50}). Remarkably, it also achieves 97% of the mAP_{50} score of current fully supervised models. To further illustrate the practicality of our work, we train Box2Mask on the recently released ARKitScenes dataset which is annotated with 3D bounding boxes only, and show, for the first time, compelling 3D instance segmentation masks.

Keywords: 3D Semantic Instance Segmentation, Weakly Supervised Learning, 3D Scene Understanding



	Dense Points	Sparse Points	Bounding Boxes
Predictions:	✓ Accurate	✗ Less Accurate	✓ Accurate
Labor:	✗ Costly	✓ Cheaper	✓ Cheaper
Availability:	✗ Few Datasets	✗ Single Dataset	✓ Many Datasets

Fig. 2: Annotation Types. Our key finding is that bounding box annotations serve as a surprisingly valuable annotation type for learning dense 3D instance masks. Prior work either requires per-point annotations *(left)* with instance ids and semantic classes for millions of points, or initial weak supervision methods [21] use sparse point annotations *(middle)*, where a subset of points is annotated with instance centers and semantic classes. We propose to use bounding box annotations *(right)*, where each object is annotated with its tight fitting box and a semantic label. We find boxes combine desirable properties: they allow for results on par with full supervision, reduce annotation effort to the object-level and are readily available in several large-scale 3D datasets [3,5,14,33].

1 Introduction

Semantic instance segmentation of 3D scenes is one of the fundamental challenges in computer vision and robotics. The goal is to predict a foreground-background mask and a semantic class (*e.g.*, 'chair', 'fireplace') for each object in a 3D scene (point cloud or mesh). Over the last years, the research community has contributed numerous methods [6,11,12,17,19,25,29,33,53]. This rapid development was made possible, not only by substantial advances in 3D deep learning backbones [8,16,39,40,43,45], but also by large-scale 3D datasets [2,9,36,44] crucial to train data-hungry deep models. While the acquisition of large datasets has become easier with commodity 3D scanners [3], per-point annotations (Fig. 2, left) – largely required by current methods – are still very labour-intensive. For example, labeling an average scene in ScanNet takes ∼22.3 minutes [9]. It is therefore highly desirable to alleviate the need for dense point labels. Only few works have addressed this challenge. Hou *et al.* [21] build on self-supervised pre-training [51] and propose contrastive learning techniques using sparse point annotations (Fig. 2, middle) which, however, depend on carefully selected points during the annotation process.

The key idea of this work is to use 3D bounding box annotations as weak supervision signal for dense 3D semantic instance segmentation (Fig. 2, right). Despite promising results on image understanding tasks [28], bounding boxes have so far been overlooked for dense 3D instance segmentation. We find that obtaining dense segmentation masks from object detection models (predicting one box per object) is non-trivial and leads to unsatisfying results (see Sec. 6.2). We present Box2Mask, the first method for dense instance segmentation trained solely on coarse bounding box annotations. The main result of this paper is that our weakly supervised method outperforms previous weakly supervised works [23,51] by a large margin, and is even competitive with fully-supervised state-of-the-art methods [6,33]. To achieve this goal, we face two key challenges. First, we lack dense per-point instance annotation. Second, there is no obvious representation for instance segmentation, because in contrast to semantic segmentation we cannot assign a single categorical label to points.

To address these challenges, we represent instances with the six parameters of an axis-aligned bounding box, fully exploiting the given labels. Since bounding boxes cover the full extent of the object, they are naturally a richer instance representation than the commonly used centers. Moreover, leveraging this representation leads to novel algorithms for voting, instance clustering, and a new training strategy to cope with weak labels. Specifically, we train a model where each point in the scene votes for the bounding box to which it belongs. We devise a new algorithm to cluster votes based on box volumetric overlap. Such overlap can be back-projected to the original scene points to obtain a probabilistic instance mask. Since we lack dense labels, we propose a training strategy where point to instance associations are approximated on the fly based on bounding boxes. An overview of Box2Mask is presented in Fig. 3. We evaluate our approach on three challenging indoor 3D datasets: ScanNet [9], S3DIS [2] and ARKitScenes [3].

In summary, our contributions are as follows:

- We propose a principled method leveraging bounding boxes both as a representation and to guide the training scheme. This comprises a novel method for voting, instance clustering and training with weak labels. Code, models and annotations are available at `virtualhumans.mpi-inf.mpg.de/box2mask`.
- We present the first dense 3D instance semantic segmentation method trained with only bounding boxes. It is competitive with the best fully supervised baselines (97% of HAIS [6] on ScanNet test, mAP_{50}) and it largely outperforms weakly supervised alternatives (+18 mAP_{50} compared to CSC [21]). On the largest scene dataset ARKitScenes annotated only with 3D bounding boxes, we obtain for the first time compelling 3D instance segmentation results.

2 Related Work

Densely Supervised 3D Instance Segmentation. The first deep models for 3D instance segmentation] (SGPN [47], 3D-BEVIS [11], ASIS [48]) estimated instances by grouping learned point features in an abstract embedding space. Extending this, MTML [29] proposes an additional learned directional embedding space. All these methods require non-learned, computationally expensive point-feature clustering. Similar to the popular MaskRCNN [18] for 2D instance segmentation, 3D-SIS [20] extracts bounding box proposals and extracts the per-voxel masks via a 3D-RoI layer. An interesting alternative is proposed in 3D-BoNet [53] which, from a single global scene descriptor, directly predicts all object bounding boxes which are then segmented into foreground and background. Both previous methods are sensitive to missed object detections since they cannot be recovered at later stages in the model. More recently, several works group points based on predicted semantics and object centers [6,12,17,25,33]. 3D-MPA [12] first combines points into sets of points and then groups these into objects based on features learned via a graph convolutional network. PointGroup [25] combines sets of points based on both the original point positions and learned center positions. OccuSeg [17] additionally predicts instance occupancy as a proxy for the physical size of an instance. Similar in spirit to [12], HAIS [6] groups points into sets of points, and performs additional set refinement steps as well as scoring as in [25]. SSTNet [33] generalizes the over-segmentation idea from [17] using a super-point semantic tree network which hierarchically merges segments into object instances. Different from the above methods, our Box2Mask does not estimate object centers but votes for object bounding boxes. By doing so, the voting mechanism is robust towards varying object sizes (see Fig. 4).

Weak Supervision for 3D Segmentation. Many efforts have been made to reduce the labeling cost of dense annotation for 2D images, with models learning from only weak annotations such as bounding boxes [10,22,28], scribbles [34,46,54], points [4] and image-level labels [1,26,41,59].

In 3D, mostly semantic segmentation on point clouds was addressed with weak labels, where only sparsely labeled points are given [7,21,24,35,52,52,55,56]. SSPC-Net [7] and Liu *et al.* [35] assign sparse labels to super-points, construct graphs over super-points and learn to propagate semantic labels between nodes in the graph from labeled super-points to the unlabeled super-points. In this fashion, and with only 10% of labeled points, Xu *et al.* [52] achieve a performance close to fully supervised semantic segmentation methods. Another line of work explores scene-level annotations or subcloud-level annotations for semantic segmentation [42,49,58]. Here, only a list of semantic classes contained in a scene (or part of it) are assumed, without precise localization. However, since scene and sub-scene labels are the coarsest annotations assumed, results typically lack details.

Only recently, the first work on weak supervision for 3D semantic instance segmentation was proposed, assuming a sparse number of points is annotated with their instance centers [21]. To reduce data hungriness, unsupervised pre-training on 3D point clouds [15,32,51] is used. PointContrast [51] improves supervised down-stream tasks significantly via contrastive pre-training on point clouds. CSC [21] additionally incorporates point-level correspondences and spatial contexts in a scene. CSC achieves encouraging initial results, but still leaves a significant gap to fully supervised methods.

3 Hough Voting for Bounding Boxes

Our model votes for instances represented as bounding boxes (Fig. 3). This is unlike prior work, which represents instances as centers. Our experiments show that the proposed box representation has several advantages over centers.

Encoding. Input are scene points $\mathcal{S} \in \mathbb{R}^{N \times F}$, where N is the number of scene points and F the number of per-point input features ($F = 9$ in our experiments, for position, color and estimated surface normal). We use the popular sparse convolutional U-Nets [8,16] as backbone, to obtain per scene point features.

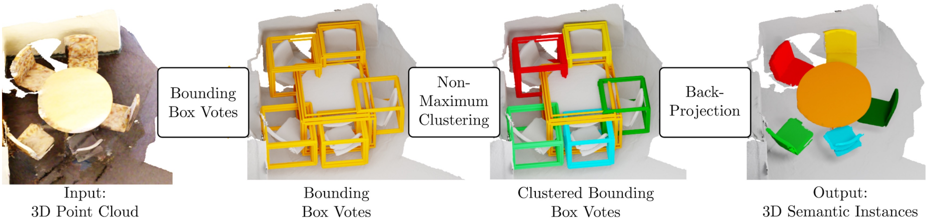


Fig. 3: Box2Mask Overview. Input to Box2Mask is a colored 3D point cloud of a scene. **Bounding Box Voting:** For each point in the input scene, our model predicts the points instance, parameterized as 6-DoF bounding box. The key contribution is the training procedure with only coarse bounding box labels (requiring no per-point labels) by associating points to bounding box labels. **Non-Maximum Clustering:** Votes are clustered using our Non-Maximum Clustering (NMC) that is specifically tailored to the bounding box representation. **Back-Projection:** A point is associated with the cluster of the box it predicted. Doing this for each point yields the final instance masks.

Those require discretization of point positions into regular grids, but allow for high resolution (we use 2 cm). In case multiple points fall into the same grid location, instead of averaging features and associated ground truth annotations leading to a blurry combination, we pick the features of the nearest neighboring scene point. We retain this mapping to allow reversing the discretization later.

Decoding. Based on the point features, we predict a bounding box, a score, and a semantic label per scene point. We branch out decoding with separate (3 layers) MLP networks. We predict axis-aligned bounding boxes, parameterized using their centers, and sizes (width, height, depth), with one MLP for center and one for size. Similarly, another MLP predicts a scalar score, estimating the intersection-over-union of the predicted bounding box with the ground truth, and a fourth branch predicts the semantic label ("chair", "table", ...).

4 Clustering and Back-projection

Next, we want to turn our model predictions into instance masks. Since, scene points vote for the bounding box of their instance, all points voting for the same box define an instance and points voting for a different box define a different instance. However, due to noisy predictions, box votes from points of the same instance will not be perfectly aligned which demands a clustering strategy.

In contrast to clustering based on centers with Euclidean distance metrics (as commonly employed in 3D instance segmentation), we make full use of our bounding box votes, by defining a novel 3D clustering method based on volumetric similarity (Fig. 4 A, B). Specifically, we define our clustering similarity-metric over two bounding box votes, $\mathbf{b}_a$ and $\mathbf{b}_b$, as the intersection-over-union (IoU):

$$\text{vote-space similarity: } \mathrm{IoU}(\mathbf{b}_a, \mathbf{b}_b) = \frac{\text{area of overlap}}{\text{area of union}}. \qquad (1)$$

Voting and clustering of bounding boxes has two key benefits: First, IoU allows to separate two instances when no box overlap is present, which requires care-

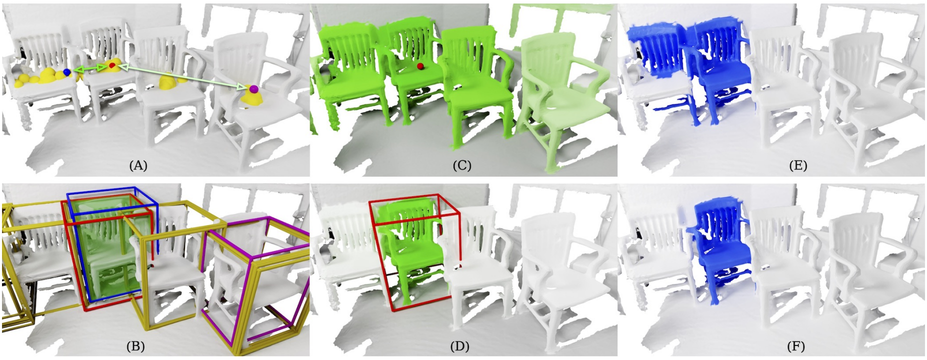


Fig. 4: Center Clustering (top row) *v.s.* Our Box Clustering (bottom row). **(A)** Scene points predict their instance centers (yellow). For clustering, the Euclidean metric between votes is used (arrows between highlighted votes). **(B)** In contrast, we propose to use IoU (Eq. 1) on bounding box votes. Intersection of blue and red votes visualized in green. Since boxes define the extend of objects, this metric can discriminate distinctively when overlap is not present, as with the violet box vote. This is key for obtaining sharp similarity decay of the scene points **(D)**, instead of smooth decay with distance **(C)**. The latter is sensitive to errors in binarization thresholds (handcrafted and dataset specific) when converting to instance masks **(E)**. In contrast, our method naturally encodes this threshold via box sizes, *i.e.*, converting from **(D)** to **(F)** is robust.

ful handcrafted thresholds for center voting (see Fig. 4). Second, while center clustering will fail when two instances have the same center (*e.g.*, an apple in a bowl), boxes additionally distinguish instance size.

Clustering Volumetrically. First, all bounding box votes are sorted in descending order according to the predicted scores. Then, the highest-scoring box is picked and serves as the representative, $\mathbf{b}_r$, of the first cluster. All boxes, $\mathbf{b}$, that are sufficiently similar to the representative, $\text{IoU}(\mathbf{b}, \mathbf{b}_r) > \tau$, are assigned to this cluster. Higher values of $\tau \in (0, 1)$ will result in numerous smaller clusters and lower values will result in fewer larger clusters. The next step is to take the next highest scoring box that has not yet been clustered: it will serve as next representative. This process is repeated until all boxes are assigned to a representative or are chosen as representative themselves. We call this clustering *non-maximum clustering* (NMC). A pseudo-code description is given in the appendix.

Back-projection to Instance Masks. Ultimately, we are interested in clusters in the original point cloud. Therefore, we *back-project* each clustered bounding box to the point that voted for it. All points that voted for boxes within the same cluster form an instance mask. For semantic instance segmentation, each instance mask should be accompanied with a semantic class and a score. Since our model predicts semantics for each scene point, we obtain instance labels by performing a majority-vote per mask. For the score, we rely on the predicted IoU score of the point that voted for the instance's cluster representative $\mathbf{b}_r$.

5 Training with Weak Bounding Box Labels

Fully supervised 3D instance segmentation methods rely on densely annotated point clouds and learn to predict at each scene point $\mathbf{p}$ some ground truth value, $gt(\mathbf{p})$ (*e.g.* instance center). However, this strategy cannot be applied when a scene is annotated only with a set of bounding boxes. More specifically, boxes do not define instance ground truth on a point-level, such that $gt(\mathbf{p})$ is unclear. We address this issue by finding a strategy to approximate point-to-bounding box associations. More formally, let $\mathcal{B} = \{\mathbf{b}_1, ..., \mathbf{b}_B\}$, $\mathbf{b}_i = [\text{center}, \text{size}, \text{label}] \in \mathbb{R}^6 \times \mathbb{N}$ be the set of annotated boxes in a scene, we define a function $a : \mathcal{P} \rightarrow \mathcal{B}$ which maps a 3D scene point $\mathbf{p} \in \mathcal{P}$ to a ground truth bounding box $a(\mathbf{p}) \in \mathcal{B}$. Once such a function is found, the model can be trained in a similar fashion as fully supervised models, replacing the exact point-to-point ground truth $gt(\mathbf{p})$ with our approximate point-to-box ground truth $gt\big(a(\mathbf{p})\big)$.

How Should the Mapping Function a Be Defined? Since ideally, an object bounding box contains all the points of its instance, the possible box associations of a point are reduced to only those boxes containing it. In turn, if a point is contained in no bounding box, it can only be part of the background (*e.g.*, wall, ceiling, floor). This simple observation has an important effect: with high certainty, we can learn to segment (or discard) non-instance points, a crucial part of instance segmentation. We can specify our approximate associations further for points contained in only a single box: all those points will actually belong to the instance of the box, up to points from non-annotated background points. If a point, however, is located in multiple bounding boxes, we cannot get exact point-to-box association. These observations can be formulated into our initial approximate association function:

$$a(\mathbf{p}) = \begin{cases} \text{background}, & \text{if } \mathbf{p} \text{ is not contained in any } \mathbf{b} \in \mathcal{B} \\ \mathbf{b}, & \text{if } \mathbf{p} \text{ is only contained in a single } \mathbf{b} \in \mathcal{B} \\ \text{undecided} & \text{else} \end{cases} \tag{2}$$

and updates the co-domain of a to $\mathcal{B} \cup \{\text{background}, \text{undecided}\}$. These associations are already surprisingly useful for supervising on *decided* points only (i.e. none or single box points): in experiments, this initial strategy achieves 87% of current fully-supervised methods with dense per-point labels.

A key remaining question however is: can we increase prediction quality by integrating approximate associations for points in multiple boxes? In our analysis (Tab. 5), we find that choosing the smallest of multiple available boxes improves over other strategies. This makes sense since smaller objects are often fully contained in bounding boxes of larger objects (a pillow on a sofa, a sink in a cabinet). Using this strategy, and only relying on bounding box annotations, we achieve 97% of the performance of fully-supervised methods trained with dense per-point labels.

Losses. Let $\mathcal{P}$ be the set of scene points and $\mathcal{B}$ the set of annotated bounding boxes. Using our association function, a, we define our losses, only given the box annotations. We define our instance losses only for points associated with the scene foreground $\mathcal{F} := \{\mathbf{p} \in \mathcal{P} | a(\mathbf{p}) \in \mathcal{B}\}$, excluding "background" and "undecided" points. Our instance prediction losses are defined as:

$$\begin{aligned}
\mathcal{L}_{\text{offset}} &:= \frac{1}{|\mathcal{F}|} \sum_{\mathbf{p} \in \mathcal{F}} \| \, o\big(\mathbf{p}, a(\mathbf{p})\big) - \widehat{o}(\mathbf{p}) \, \|_1, \\
\mathcal{L}_{\text{size}} &:= \frac{1}{|\mathcal{F}|} \sum_{\mathbf{p} \in \mathcal{F}} \| \, s\big(a(\mathbf{p})\big) - \widehat{s}(\mathbf{p}) \, \|_1, \\
\mathcal{L}_{\text{score}} &:= \frac{1}{|\mathcal{F}|} \sum_{\mathbf{p} \in \mathcal{F}} \text{CE}\Big(iou\big(a(\mathbf{p})\big), \widehat{iou}(\mathbf{p}) \Big),
\end{aligned} \tag{3}$$

where o is the offset from $\mathbf{p}$ to the center of its associated bounding box, $a(\mathbf{p})$; s is the size (width, height, depth) and iou is the IoU of the predicted bounding box with the associated box, $a(\mathbf{p})$. We denote the predicted values with a hat and the cross-entropy loss with CE. Similarly, the dense semantic segmentation is learned from only bounding boxes, relying on their semantic label. In contrast to above instance losses, the semantic loss, $\mathcal{L}_{\text{sem}}$, includes the points associated with the background $\mathcal{D} := \{\mathbf{p} \in \mathcal{P} | a(\mathbf{p}) \neq \text{undecided}\}$:

$$\mathcal{L}_{\text{sem}} := \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} \text{CE}\Big(sem\big(a(\mathbf{p})\big), \widehat{sem}(\mathbf{p}) \Big). \tag{4}$$

where sem defines the ground truth, including a generic semantic class ”background” for all points associated with it:

$$sem(\mathbf{p}) := \begin{cases} \text{background\_class}, & \text{if } \mathbf{p} \text{ in no box} \\ \text{label}\big(a(\mathbf{p})\big), & \text{else} \end{cases} \tag{5}$$

Importantly, this allows us at inference time, to predict and filter background points, not defining any instances. Our network prediction consists of a forward pass, fully implemented with convolutions and trained end-to-end with the combined, multi-task loss defined as $\mathcal{L} := \mathcal{L}_{\text{offset}} + \mathcal{L}_{\text{size}} + \mathcal{L}_{\text{score}} + \mathcal{L}_{\text{sem}}$.

5.1 Implementation and Training Details

We train our network end-to-end and from scratch with the Adam optimizer, using an initial learning rate of 0.001, a batch size of 8 entire scenes, and train for 500 epochs on a single NVIDIA Quadro RTX 8000. For data augmentation, scenes are randomly rotated around height, flipped, and scaled in Uniform[0.9, 1.1]. Our backbone is a 6-layer sparse-convolutional encoder-decoder including skip connections based on [6]. The MLP heads are implemented using 3 layers with 96 hidden units. Similar to other current segmentation methods [17, 33, 37], we perform point over-segmentation [13, 27] on ScanNet and ARKitScenes, and similarly employ [30,31] on S3DIS. This reduces the number of votes by averaging over segments before clustering, alleviating the computational load. Empirically, we set the NMC clustering threshold $\tau = 0.3$. More details are in the appendix.

6 Experiments

6.1 Comparing with State-of-the-art Methods.

Datasets. S3DIS [2] consists of 272 scans of 6 large-scale indoor areas collected from three different buildings. Scans are represented as point clouds and points are annotated with instance- and semantic-labels out of 13 object classes. To obtain bounding box annotations from masks, we use the standard approach of [12, 19–21, 38, 50, 53, 57], *i.e.*, we obtain axis-aligned bounding boxes from the instance point annotations. We report scores on Area-5 and 6-fold cross-validation.

ScanNet [9] is a richly annotated dataset of 3D reconstructed indoor scenes represented as meshes. Similar to S3DIS, each scene is annotated with semantic- and instance-segmentations of 18 object categories. It consists of 1201 training scenes, 312 validation scenes, and 100 hidden test scenes. Bounding box annotations are obtained the same way as for S3DIS.

ARKitScenes [3] is the largest of these datasets with 4499 training scenes and 550 validation scenes. The scenes are represented as reconstructed meshes and are recorded in real-world homes. The dataset is annotated with oriented object bounding boxes across 17 semantic classes. Importantly, per-point labels are not available. Nevertheless, our approach is able to leverage the bounding box annotations as weak supervision signal, which is an *immense practical advantage* over existing 3D instance segmentation methods which require dense per-point annotations and can therefore not be trained on this dataset.

Methods in Comparison. We compare to both fully-supervised and weakly-supervised SOTA prior methods. Fully-supervision methods are the majority: we compare top-down segmentation methods 3D-BoNet [53], 3D-SIS [20] and bottom-up methods MTML [29], PointGroup [25], 3D-MPA [12], OccuSeg [17], HAIS [6] and SSTNet [33]. See Sec. 2 for more details.

Weakly-supervised methods are much less, and only recently explored. PointContrast [51] and CSC [21] both make use of unsupervised pre-training via contrastive-learning. Compared to PointContrast, CSC follows a more sophisticated approach by taking spatial scene context into account. For 3D instance segmentation, the pre-trained models are supervised with a limited number of sparsely annotated points (20, 50, 100 or 200 Points), for which the ground truth object centers and semantic classes are known during training.

Results on S3DIS and ScanNet are summarized in Tab. 1. Our approach improves upon prior (point-based) weakly-supervised methods [21, 51] by more than **10 mAP**. While sparse point labels and bounding box labels might not be directly comparable, it is noteworthy that this improvement is achieved without pre-training as used by [21, 51]. Compared to fully-supervised approaches, our weakly-supervised method achieves 92% and 94% of the performance of leading methods on ScanNet (SSTNet, val, mAP$_{50}$) and S3DIS (HAIS, A5, mPrec) respectively. This is extremely encouraging, as it indicates that densely labeled points might not be entirely necessary. Qualitative ScanNet results are shown in Fig. 6. Our method predicts clear masks in heavily cluttered environments

			ScanNet Validation		ScanNet Hidden Test		S3DIS Area 5		S3DIS 6-fold CV	
	Method	Supervision	mAP @50%	mAP @25%	mAP @50%	mAP @25%	mPrec	mRec	mPrec	mRec
Weakly Supervised	CSC [21]	20 Points	26.3	–	28.9	49.6				
	PointContrast [51]	20 Points	–	–	25.9	47.4				
	CSC [21]	50 Points	32.6	–	41.4	62.0				
	PointContrast [51]	50 Points	–	–	40.0	60.3				
	CSC [21]	100 Points	39.9	–	46.0	65.4				
	PointContrast [51]	100 Points	–	–	45.6	63.7				
	CSC [21]	200 Points	48.9	–	49.4	70.2				
	PointContrast [51]	200 Points	44.5	–	47.1	66.2				
	Box2Mask (Ours)	Boxes	**59.7**	**71.8**	**67.7**	**80.3**	66.7	65.5	72.2	70.5
	Relative to SOTA		92.8%	95.0%	96.9%	100%	93.8%	99.8%	98.2%	96.0%
Fully Supervised	3D-SIS [20]	All Points	18.7	35.7	38.2	55.8				
	3D-BoNet [53]	All Points	–	–	48.8	68.7	–	–	65.6	47.6
	MTML [29]	All Points	40.2	55.4	54.9	73.1				
	PointGroup [25]	All Points	56.9	71.3	63.6	77.8	61.9	62.1	69.6	69.2
	3D-MPA [12]	All Points	59.1	72.4	61.1	73.7	46.7	**65.6**	66.7	64.1
	OccuSeg [17]	All Points	60.7	71.9	63.4	73.9	–	–	72.8	60.3
	HAIS [6]	All Points	64.1	**75.6**	**69.9**	**80.3**	**71.1**	65.0	73.2	69.4
	SSTNet [33]	All Points	**64.3**	74.0	69.8	78.9	65.5	64.2	**73.5**	**73.4**

Table 1: State-of-the-art 3D Semantic Instance Segmentation. We show fully-supervised methods (dense point annotations) and weakly-supervised methods (sparse points and bounding boxes) on ScanNet [9] and S3DIS [2]. [51] is as reported in [21].

and accurately segments even very large objects like tables. The difference between weak- and full-supervision is marginal, however, bounding boxes need only be annotated on object-level in contrast to per-point annotations. Additional qualitative results and analysis, including S3DIS, are in the appendix.

Quantitative results on ARKitScenes are shown in Tab. 2, visual results in Fig. 5. As per-point instance labels are not available, we cannot report segmentation scores. Instead, as a proxy, we compare to recent object detection methods [38, 50, 57] by fitting oriented bounding boxes to our predicted masks. This indirectly measures mask quality. However, high detection scores are only obtained if the predicted point masks are accurate. Therefore, the correctness of position and size of the masks are measured. Our approach achieves leading performance among all methods (**+4 mAP**) suggesting good quality masks.

Limited Annotations 3D Semantic Instance Benchmark. On this benchmark, the ground truth labels are given for only a limited number of annotated points per scene. We compare to the baseline methods introduced in [21]. These methods perform instance segmentation by predicting centers, which means that they rely on annotated centers (see Fig. 2, middle). Instead, our approach relies on bounding box annotations. We believe that bounding boxes are more realistic and easier to annotate than 3D object centers, which are usually located somewhere in empty space and can be hard for an annotator to precisely locate. Results are shown in Tab. 3. Our approach consistently outperforms prior work with a large margin, even without relying on any pre-training.

	Cabinet	Refrig.	Shelf	Stove	Bed	Sink	Washer	Toilet	Bathtub
VoteNet [38]	37.1	62.7	12.4	0.3	85.0	31.1	45.3	75.5	93.3
H3DNet [57]	40.2	59.4	10.0	1.6	**88.2**	40.1	49.0	83.8	93.0
MLCVNet [50]	45.1	**70.0**	16.9	2.4	88.0	**40.2**	51.5	85.9	**94.1**
Box2Mask(ours)	**45.9**	62.6	**28.0**	**5.2**	87.1	30.6	**53.8**	**89.4**	92.9
	Oven	Dishw.	Fireplace	Stool	Chair	Table	Monitor	Sofa	**mAP**
VoteNet [38]	18.3	2.9	22.1	3.0	20.1	31.0	0.6	68.3	35.8
H3DNet [57]	24.1	**3.9**	19.5	8.8	25.2	32.2	1.5	70.4	38.3
MLCVNet [50]	24.2	3.0	38.5	8.0	31.5	36.6	4.1	71.9	41.9
Box2Mask (ours)	**28.1**	3.8	**59.9**	**20.8**	**35.2**	**60.3**	**7.3**	**82.8**	**46.7**

Table 2: Whole-scene 3D Object Detection Scores on ARKitScenes [3]. The ground truth includes only oriented bounding box annotations, no point-level instance masks. Therefore, we cannot directly compute instance segmentation scores. Instead, as a proxy, we compare to recent object detection methods by fitting an oriented bounding box containing our predicted masks. We report the average precision on the validation set with an IoU threshold of 50% as in [38]. All other scores are as reported in [3].

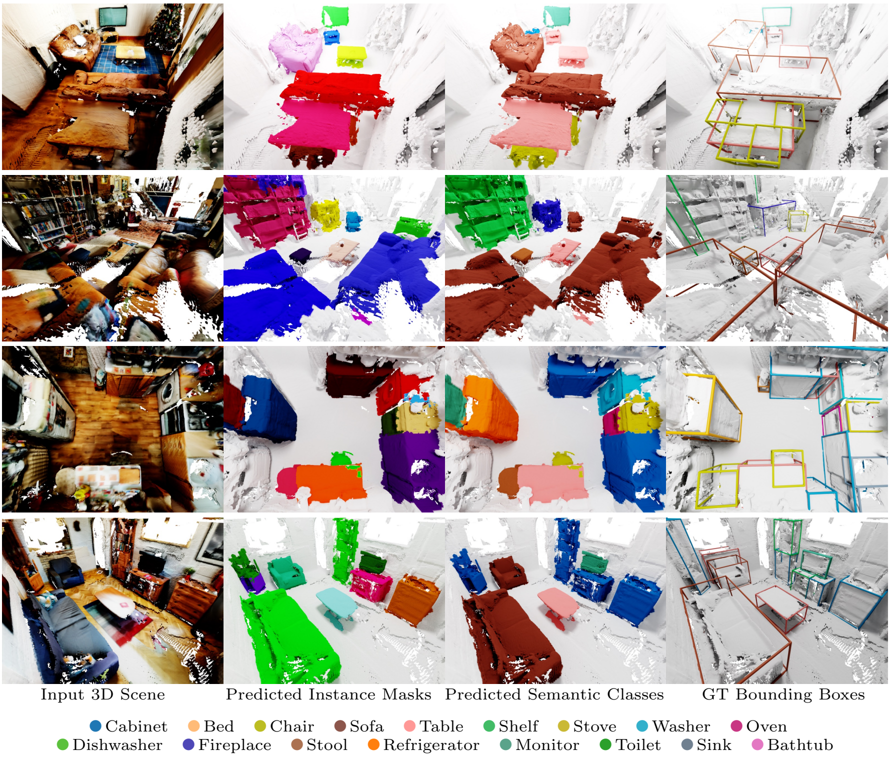


Fig. 5: Qualitative Instance Segmentation Results on ARKitScenes [3]. Individual instance masks are colored randomly. Semantic classes are colored as indicated. Ground truth boxes are shown for reference only and are not used during inference.

mAP_{50}	200 points	100 Points	50 Points	20 Points
CSC trained from scratch [21]	46.4	41.8	31.1	20.0
PointContrast* [51]	47.1	45.6	40.0	25.9
CSC* [21]	49.4	46.0	41.4	28.9
Ours	**59.2** (+9.8)	**56.5** (+10.5)	**49.8** (+8.4)	**46.5** (+17.6)

Table 3: ScanNet Data Efficient Benchmark Test. Instance segmentation on limited annotations (LA). Scores as in [21]. Star (*) indicates usage of pre-training.

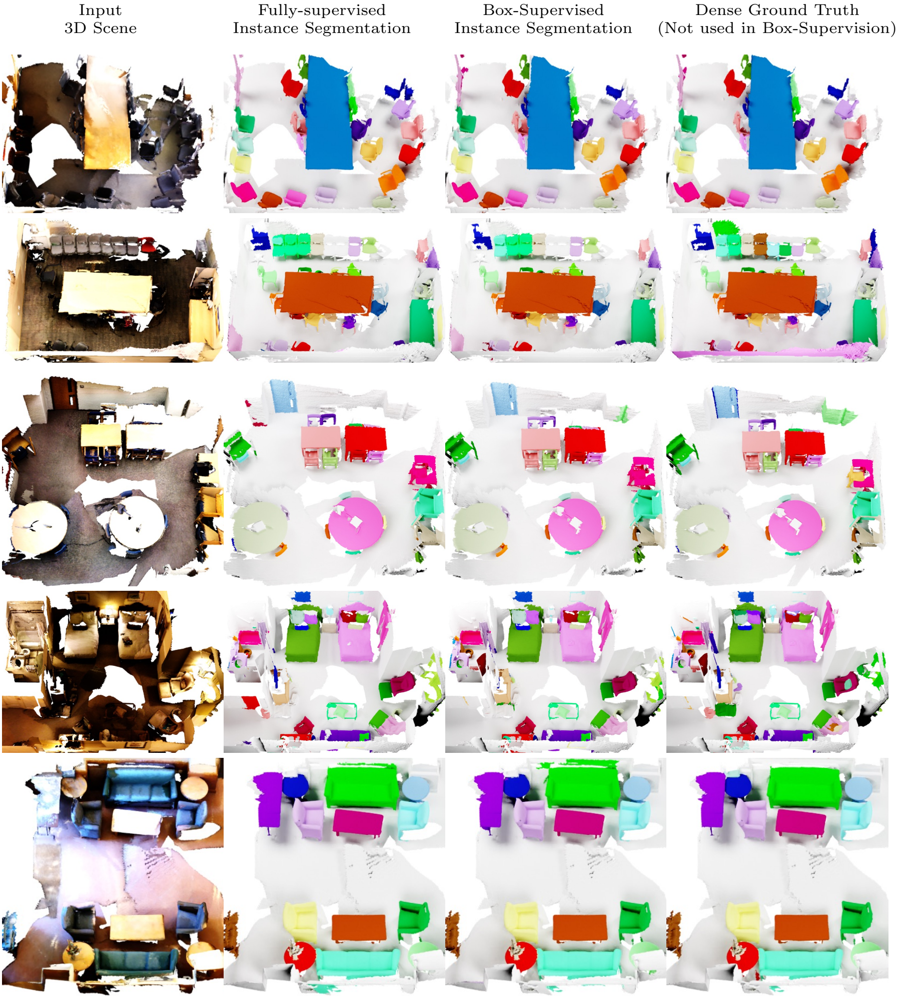


Fig. 6: Qualitative Instance Segmentation Results (validation) on ScanNet [9]. Results trained only on bounding boxes well resemble the fully supervised model, and both are close to densely annotated ground truth. Instance masks have the same random color as the corresponding ground truth mask.

6.2 Analysis

	mAP_{50}	mAP_{25}
Centers + SC	51.4	63.8
Centers + SC (per Sem.)	52.0	67.7
Boxes + SC (per Sem.)	53.1	67.9
Boxes + NMC (per Sem.)	55.1	68.3
Boxes + SC	53.5	65.2
Boxes + NMC	**59.7**	**71.8**

Table 4: Non-maximum clustering (NMC) and spatial clustering (SC) on center- and bounding box-votes.

Boxes or Centers? An important baseline to the proposed *bounding-box* representation is the popular *center* representation [6,21,25,33,41]. We also analyse different techniques for clustering the voting space and compare to the proposed *non-maximum clustering* (NMC). Spatial clustering (SC), such as breadth-first search as in [25] or DBScan as in [12], groups votes based on their pairwise Euclidean distance. Further, it is common practice to cluster votes separately *per semantic class*, which ensures that points of disagreeing semantics are in different instances. Tab. 4 shows that clustering conditioned on the semantic class is beneficial only for centers. This indicates that box votes already encode sufficient semantics (via the size) increasing robustness to wrongly predicted semantics. More importantly, the proposed bounding boxes consistently outperform centers, suggesting that object size is important for vote clustering. The largest improvement is observed by NMC over SC. While SC treats all dimensions in the voting space equally, NMS is tailored to bounding boxes, using the actual geometric meaning of each feature dimension in the voting space.

Supervision	mAP_{50}	mAP_{25}
(1) Decided Only	56.0	70.8
(2) Decided + Closest Box	58.7	71.7
(3) Decided + Smallest Box	**59.7**	**71.8**

Table 5: Analysis of association strategies

Weak Supervision Analysis. We introduced *undecided points* as points inside multiple ground truth bounding boxes. *Decided* points are either supervised as background (if they are in no box) or with the single box they are in (*c.f.* Eq.2). For all others, the undecided points, we compare multiple heuristics, as summarized in Tab. 5. The simplest baseline (1) does not supervise undecided points at all, which results in 56 mAP_{50}. This is already 87% of the performance of the fully-supervised state-of-the-art SSTNet [33] (64.3 mAP_{50}). We then compare two additional heuristics: points that are in multiple ground truth bounding boxes are supervised with the closest bounding box in terms of distance to the center (2), and the smallest bounding box in terms of volume (3). The additional supervision improves scores by +3.7 mAP while the smallest box performs a bit better than the closest. Importantly, using these associations, our weakly supervised model obtains 97% of the performance of a comparable fully supervised model which shows that coarse bounding box annotations are surprisingly strong supervision signal compared to dense per-point annotations.

Effect of Noisy Box Labels. Since training bounding boxes on ScanNet are obtained from point masks, they are perfectly aligned to the points – an accuracy a human annotation might not achieve. This motivates an experiment on the robustness towards more incorrect labels. We trained separate models on annotation with missing labels (levels 0 to 10%) and inaccurate placement

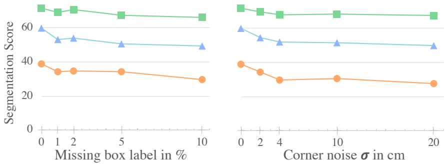


Fig. 7: Reduced Annotation Quality. Semantic instance segmentation scores on ScanNet val. trained with missing box labels *(left)* and noisy box labels *(right)*.

Fully Supervised	ScanNet		S3DIS	
	mAP_{50}	mAP_{25}	mPrec	mRec
PointGroup [25]	56.9	71.3	69.6	69.2
3D-MPA [12]	59.1	72.4	66.7	64.1
OccuSeg [17]	–	–	72.8	60.3
HAIS [6]	64.1	**75.6**	73.2	69.4
SSTNet [33]	64.3	74.0	73.5	**73.4**
Box2Mask (Ours)	**64.7**	74.5	**75.4**	69.3

Table 6: Fully Supervised Setting. Our model achieves competitive instance segmentation scores on ScanNet validation and S3DIS 6-fold cross validation.

(0 to 20 cm error in box corners). See Fig. 7 for the results. We observe good robustness, with only around 4 mAP differences.

Fully Supervised Setting. Our model can also be adapted to the fully supervised setting, where dense per-point labels are available. The association function a returns the corresponding ground truth point label. Our model compares favorably to recent state-of-the-art approaches, as summarized in Tab. 6.

Is a Detection Model Enough? As a simple baseline, instead of using the box annotations for directly training instance segmentation, we train a detection model that predicts one box per object. We obtain an instance mask via post-processing with the best performing box-to-point association strategy (Tab. 5), which was also used for weak supervision of our model. Our proposed approach largely outperforms this baseline quantitatively (+11.8 mAP_{50} on ScanNet) as well as qualitatively, see appendix (Sec. A) for details. This suggests that our model generalizes beyond the weak point associations, to complete object priors.

7 Conclusion

In this work, we show that 3D bounding box annotations serve surprisingly well as weak supervision for training dense instance segmentation models. Prior works either use dense supervision on all points (which is costly to label), or weak supervision from only a few annotated points (which performs less well). Bounding boxes provide an attractive alternative: the annotation effort is drastically reduced compared to dense point labeling, and they perform notably better than prior sparse labels and are even close to fully-supervised methods. We demonstrate the effectiveness of our instance segmentation approach on several benchmarks, and in particular on the recent, largest scene dataset, ARKitScenes. Although annotated with 3D bounding boxes only, we obtain for the first time compelling 3D instance segmentation results. This unlocks a large body of 3D detection datasets to be viable for learning instance segmentation.

Acknowledgements. We thank Alexey Nekrasov and Jonas Schult for their feedback. This work is funded by the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1, Project number 390727645. J. Chibane is a fellow of the Meta Research PhD

Fellowship Program - area: AR/VR Human Understanding. F. Engelmann is a post-doctoral research fellow at the ETH Zürich AI Center.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
2. 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, 9, 10, 20, 22
3. Baruch, G., Chen, Z., Dehghan, A., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A.: ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data. In: Neural Information Processing Systems (NIPS) (2021) 2, 3, 9, 11
4. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision (ECCV) (2016) 4
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2
6. Chen, S., Fang, J., Zhang, Q., Liu, W., Wang, X.: Hierarchical Aggregation for 3D Instance Segmentation. In: International Conference on Computer Vision (ICCV) (2021) 2, 3, 4, 8, 9, 10, 13, 14
7. Cheng, M., Hui, L., Xie, J., Yang, J.: Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In: Conference on Artificial Intelligence (AAAI) (2021) 4
8. Choy, C., Gwak, J., Savarese, S.: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2, 4, 19
9. 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, 3, 9, 10, 12, 20, 22, 23, 24
10. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: International Conference on Computer Vision (ICCV) (2015) 4
11. Elich, C., Engelmann, F., Kontogianni, T., Leibe, B.: 3D Bird's-Eye-View Instance Segmentation. In: German Conference on Pattern Recognition (GCPR) (2019) 2, 3
12. Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., Nießner, M.: 3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 3, 4, 9, 10, 13, 14
13. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. In: International Journal of Computer Vision (2018) 8
14. Gählert, N., Jourdan, N., Cordts, M., Franke, U., Denzler, J.: Cityscapes 3D: Dataset and Benchmark for 9 DoF Vehicle Detection. arXiv preprint arXiv:2006.07864 (2020) 2
15. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (ICLR) (2018) 4

16. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2, 4
17. Han, L., Zheng, T., Xu, L., Fang, L.: OccuSeg: Occupancy-aware 3D Instance Segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 3, 4, 8, 9, 10, 14
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
19. He, T., Shen, C., van den Hengel, A.: DyCo3d: Robust Instance Segmentation of 3D Point Clouds through Dynamic Convolution. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 9
20. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3, 9, 10
21. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring Data-efficient 3D Scene Understanding with Contrastive Scene Contexts. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 3, 4, 9, 10, 12, 13
22. Hsu, C.C., Hsu, K.J., Tsai, C.C., Lin, Y.Y., Chuang, Y.Y.: Weakly supervised instance segmentation using the bounding box tightness prior. Advances in Neural Information Processing Systems (2019) 4
23. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional Projection Network for Cross Dimensional Scene Understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2
24. Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C.W., Jia, J.: Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
25. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: PointGroup: Dual-set Point Grouping for 3D Instance Segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 3, 4, 9, 10, 13, 14
26. Joon Oh, S., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4
27. Karpathy, A., Miller, S., Fei-Fei, L.: Object discovery in 3D scenes via shape analysis. In: Robotics and Automation (ICRA) (2013) 8
28. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2, 4
29. Lahoud, J., Ghanem, B., Pollefeys, M., Oswald, M.R.: 3D Instance Segmentation via Multi-task Metric Learning. In: International Conference on Computer Vision (ICCV) (2019) 2, 3, 9, 10
30. Landrieu, L., Boussaha, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 8
31. Landrieu, L., Boussaha, M.: Point cloud over-segmentation with graph-structured deep metric learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 8
32. Li, J., Chen, B.M., Lee, G.H.: So-net: Self-organizing network for point cloud analysis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4

33. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance Segmentation in 3D Scenes using Semantic Superpoint Tree Networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 3, 4, 8, 9, 10, 13, 14
34. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4
35. Liu, Z., Qi, X., Fu, C.W.: One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
36. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-level 3D Object Understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2
37. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In: 3DV (2021) 8
38. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep Hough Voting for 3D Object Detection in Point Clouds. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 9, 10, 11
39. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2
40. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In: Neural Information Processing Systems (NIPS) (2017) 2
41. Qi, X., Liu, Z., Shi, J., Zhao, H., Jia, J.: Augmented feedback in semantic segmentation under image level supervision. In: European Conference on Computer Vision (ECCV). pp. 90–105 (2016) 4, 13
42. Ren, Z., Misra, I., Schwing, A.G., Girdhar, R.: 3d spatial recognition without spatially labeled 3d. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
43. Schult*, J., Engelmann*, F., Kontogianni, T., Leibe, B.: DualConvMesh-Net: Joint Geodesic and Euclidean Convolutions on 3D Meshes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2
44. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2
45. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and Deformable Convolution for Point Clouds. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2
46. Wang, B., Qi, G., Tang, S., Zhang, T., Wei, Y., Li, L., Zhang, Y.: Boundary perception guidance: A scribble-supervised semantic segmentation approach. In: IJCAI International joint conference on artificial intelligence (2019) 4
47. Wang, W., Yu, R., Huang, Q., Neumann, U.: SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
48. Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively Segmenting Instances and Semantics in Point Clouds. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
49. Wei, J., Lin, G., Yap, K.H., Hung, T.Y., Xie, L.: Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4

50. Xie, Q., Lai, Y.K., Wu, J., Wang, Z., Zhang, Y., Xu, K., Wang, J.: MLCVNet: Multi-level Context VoteNet for 3D Object Detection. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 9, 10, 11
51. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: European Conference on Computer Vision (ECCV) (2020) 2, 4, 9, 10, 12
52. Xu, X., Lee, G.H.: Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
53. Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., Trigoni, N.: Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. In: Neural Information Processing Systems (NIPS) (2019) 2, 3, 9, 10
54. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
55. Zhang, Y., Li, Z., Xie, Y., Qu, Y., Li, C., Mei, T.: Weakly supervised semantic segmentation for large-scale point cloud. In: Conference on Artificial Intelligence (AAAI) (2021) 4
56. Zhang, Y., Qu, Y., Xie, Y., Li, Z., Zheng, S., Li, C.: Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In: International Conference on Computer Vision (ICCV) (2021) 4
57. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3DNet: 3D Object Detection using Hybrid Geometric Primitives. In: European Conference on Computer Vision (ECCV) (2020) 9, 10, 11
58. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4
59. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4

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 [8] 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 of 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. 8. Scores are shown in Tab. 7. 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.

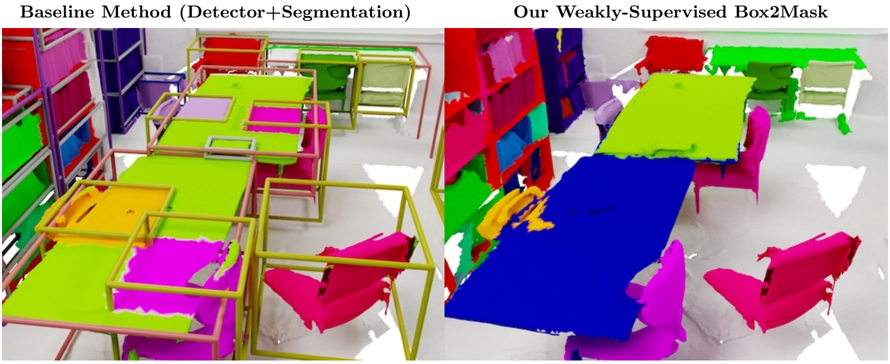

Baseline Method (Detector+Segmentation)
Our Weakly-Supervised Box2Mask

Fig. 8: 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 (tabletop). 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 7: 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. 8.

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. 8, 9, 10 and 11. On ScanNet validation and S3DIS, we show also per-category scores for the fully-supervised model trained with per-point instance labels.

	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 8: Instance Segmentation on ScanNetV2 [9] 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 9: Instance Segmentation on ScanNetV2 [9] 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 10: Instance Segmentation on S3DIS [2] 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 11: Instance Segmentation on S3DIS [2] 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. 9, we report mask prediction scores on ScanNet validation, and find that $\tau \approx 0.3$ performs best.

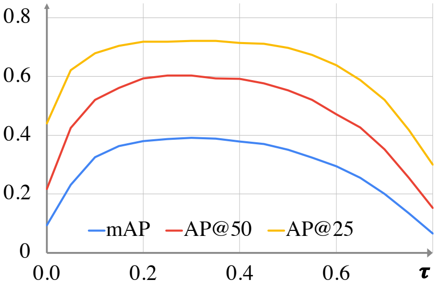


Fig. 9: Effect of parameter τ.

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. 10, we show exemplary qualitative results of our method on the S3DIS dataset [2]. We show the 3D input scene, our predicted instance masks learned from weak bounding box annotations and the ground truth instance masks as well as the ground truth bounding box annotations for comparison. In Fig. 11 and Fig. 12, we show additional close-up qualitative results on the ScanNet dataset [9]. 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.

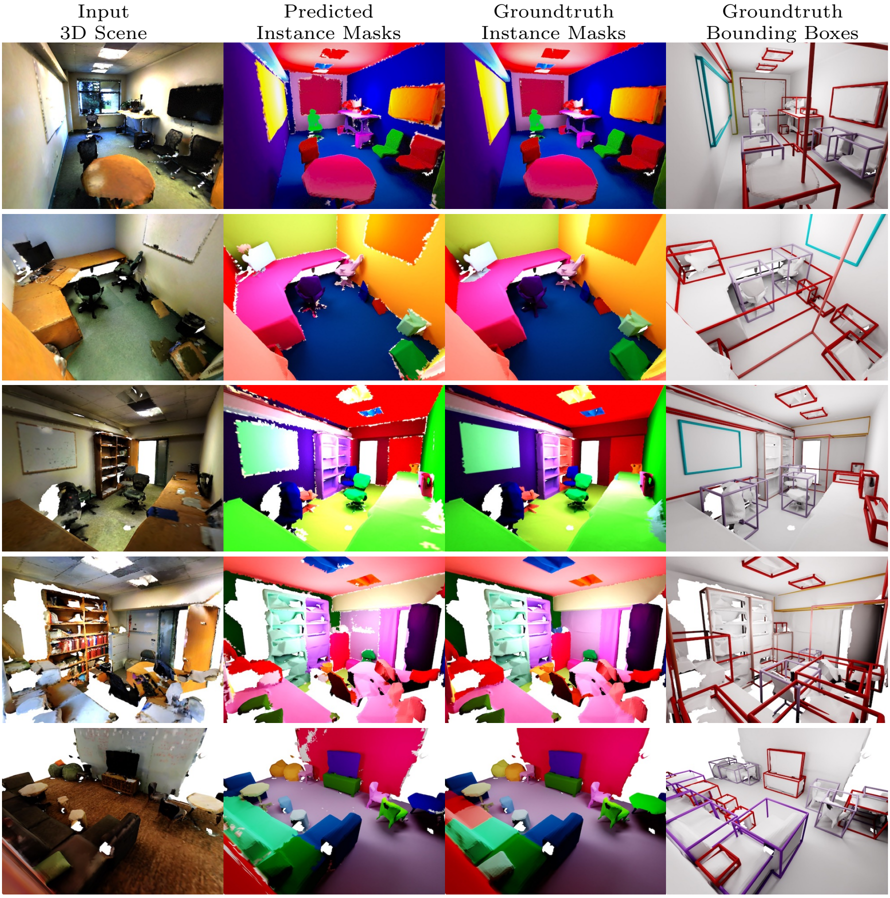


Fig. 10: Qualitative Instance Segmentation Results on S3DIS [2] Individual instance masks are colored randomly and match the ground truth instance mask colors. During training, only bounding box annotations are used (last column), per-point instance masks (third column) are not used, and are shown here only for judging the quality of the predicted instance masks (second column).

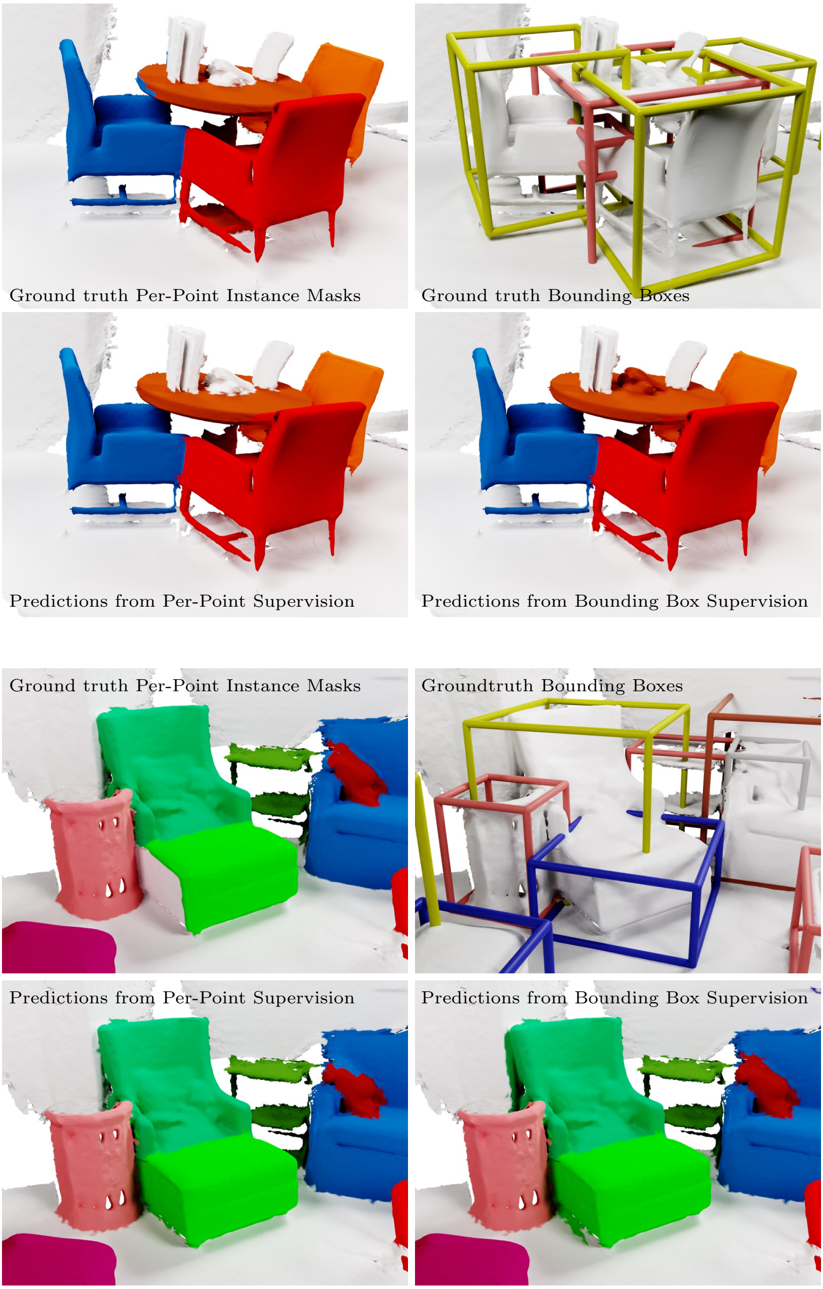

Ground truth Per-Point Instance Masks
Ground truth Bounding Boxes
Predictions from Per-Point Supervision
Predictions from Bounding Box Supervision
Ground truth Per-Point Instance Masks
Groundtruth Bounding Boxes
Predictions from Per-Point Supervision
Predictions from Bounding Box Supervision

Fig. 11: Qualitative Instance Segmentation Results on ScanNet [9] 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.

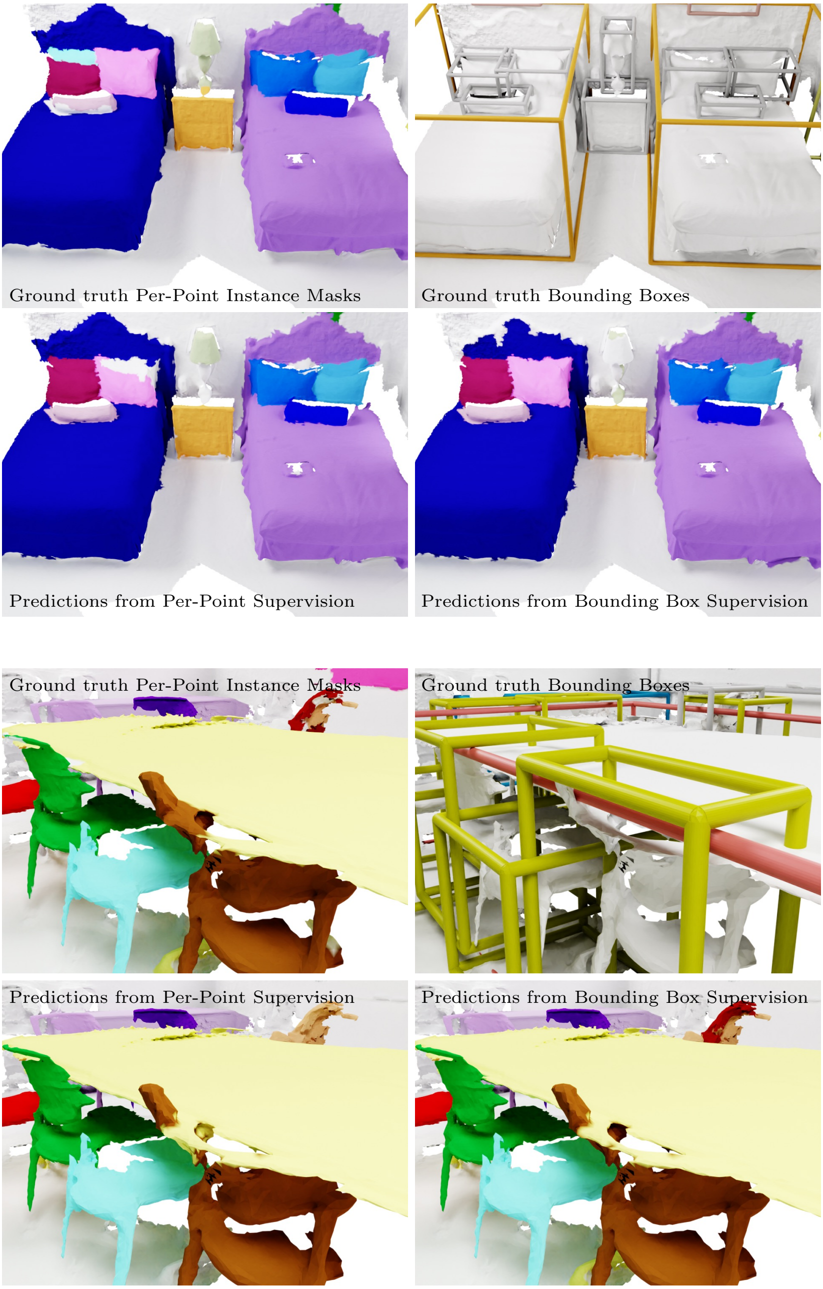

Ground truth Per-Point Instance Masks
Ground truth Bounding Boxes
Predictions from Per-Point Supervision
Predictions from Bounding Box Supervision
Ground truth Per-Point Instance Masks
Ground truth Bounding Boxes
Predictions from Per-Point Supervision
Predictions from Bounding Box Supervision

Fig. 12: Qualitative Instance Segmentation Results on ScanNet [9] 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. 13. 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. 8). 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.

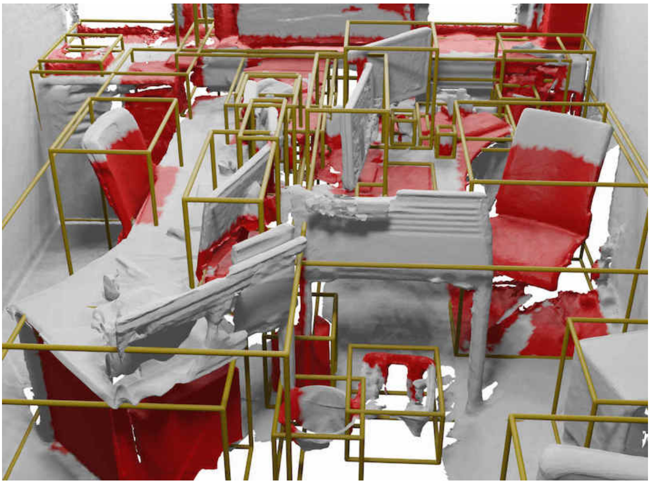


Fig. 13: ●: Undecided Points