3D Guided Weakly Supervised Semantic Segmentation

Supplementary Material

1 Implementation Details of Competing Methods

In the Sec. 4.2 of the main paper, we compare our method with two state-of-the-art methods, i.e., SDI [1] and WSSL [2]. As code is not available for these methods we re-implement SDI [1] and WSSL [2] with Pytorch. We choose these methods as their performance is still competitive with SOTA and they have clear and explicit implementation details. In this section, we introduce how we implement these two competing methods in detail.

Following the SDI paper [1], we implement the $M \cap G+$ in their method, $M \cap G+$ denotes the method which achieves the best result in their paper. First we adopt MCG [3] to get object proposals from the 1822 2D images. Then we use GrabCut [4] to obtain segment proposals from our labeled bounding boxes. Finally, we mark as foreground where both MCG outputs and GrabCut outputs agree, and use the foreground areas as segment proposals for training.

For WSSL [2], we adopt the *Bbox-Seg*, which achieves their best performance of bounding box based results in their paper. In this method, we constrain the center area of the bounding box mask (20% of the pixels within the box) to be foreground, while constraining pixels outside the bounding box to be background. The cropped area within the bounding box is regarded as unknown area. Then we feed constrained masks into a CRF to get segment proposal masks.

We still use the DeepLabV3+ [5] as backbone, follow exactly the same training procedure we introduced in the main paper to train a semantic segmentation network supervised by the segment proposals above. The quantitative results have been shown in Table. 1 of the main paper, our proposed method outperforms all competing box based semantic segmentation methods. More qualitative results are shown in Figure 2.

2 Ablation: Source of Performances.

As introduced in the main paper, We propose a novel approach, where a small number of images are labeled with bounding boxes and these images have their corresponding 3D data. Our approach can extract segment proposals from bounding boxes on labeled images and creates new segment proposals on unlabeled images of the same object instance. Experimental results show that our method outperforms all competing methods. In this ablation study, we analyze the source of our good performances, and validate that the performances of our proposed method come from two sources: 1) Better segment proposal quality. 2) available segment proposals on more training data.

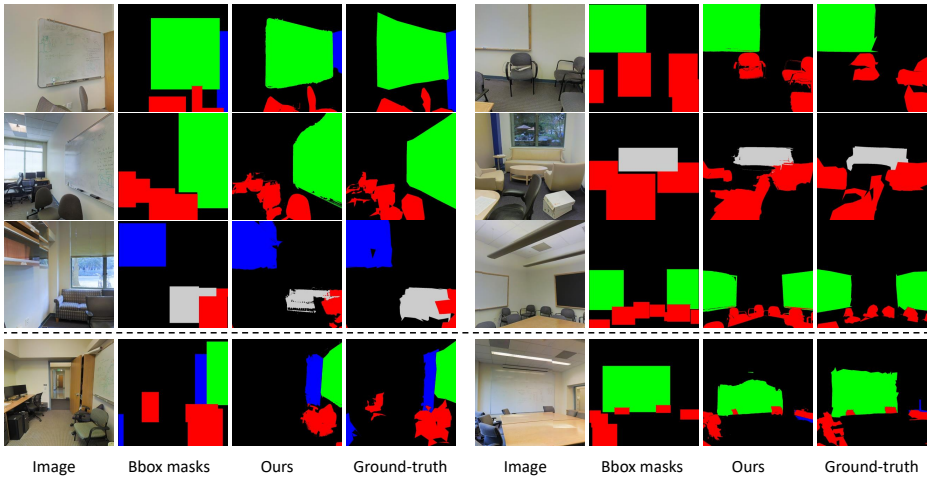

Fig. 1. Our segment proposals compared with labeled bounding box masks and ground-truth. The last row shows some failure cases.

Method	Training set amount	mIoU
Bounding box 1822	1822	61.78
WSSL [2] 1822	1822	69.06
SDI [1] 1822	1822	69.11
Ours 1822	1822	69.76
Ours 1822	4028	72.27

Table 1. Experiments of source of performances. The amount of images with hand annotations that were used in these settings are all 1822.

First, by labeling bounding boxes on 1822 images, we adopt our proposed framework to obtain pixel-wise segment proposals on these 1822 images. Then we regard these 1822 images as training set and train our semantic segmentation network. As shown in Table 1, we achieve 69.76 which already outperforms all competing bounding box based methods, demonstrating that our method can generate segment proposals(pseudo labels) with better quality. Moreover, we extend our method to generate pixel-wise segment proposals on more images by projecting the generated point clouds to more unlabeled images. With the training set that contains 4028 images, the performance is further improved to 72.27 which surpasses competing methods by a obvious margin.

3 Further Qualitative Results

Figure 1 shows original images, the bounding boxes, our segment proposals and ground truth. Figure 2 shows our network predictions compared with competing

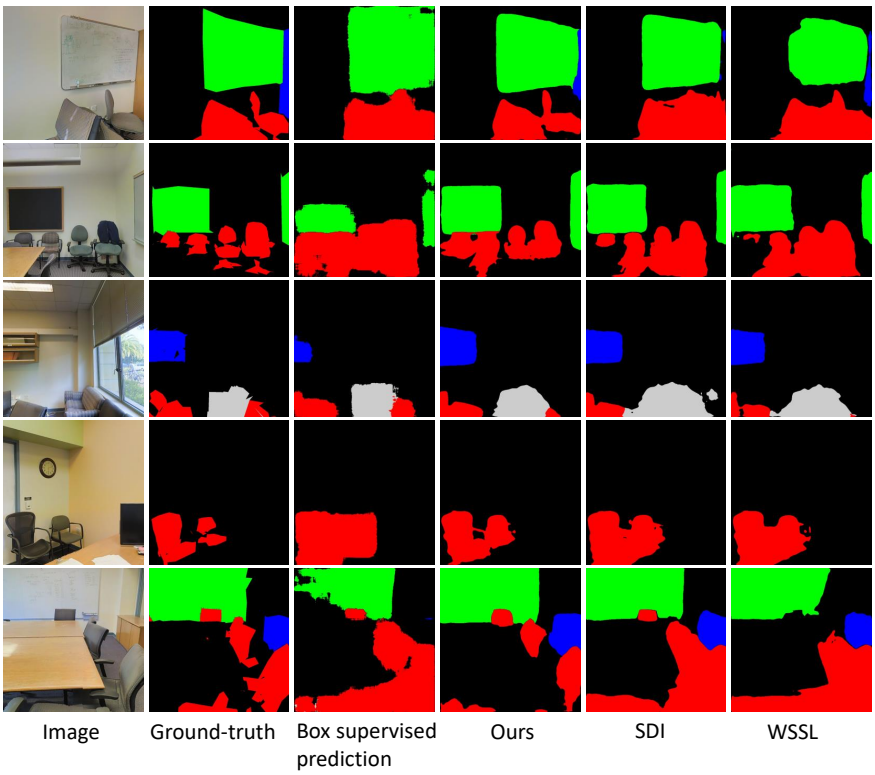

Fig. 2. Our network predictions compared with competing methods, i.e., box supervised predictions, SDI [1], and WSSL [2]. Fully connected CRF [6] is not used for results refinement.

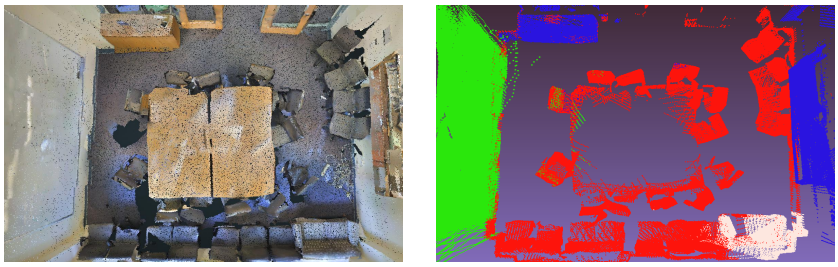

Fig. 3. Our reconstructed point cloud compared with the raw point cloud provided by the 2D-3D-S [7]. We use different colours to display points of different classes. Green: board, red: chair, blue: bookcase, white: sofa.

methods, It can be seen that our method gives substantially better fine-grained segmentation in many cases.

The last row of Figure 1 shows some failure cases. In the probability inference step, we calculate the objectness score of every point based on detection frequency across 2D images from different camera viewpoints. However, for some points that are on objects, and are detected as such, we do not have enough 2D images to reliably differentiate them from background due to the unbalanced distribution of camera viewpoints. Consequently after back-projection, these areas with low objectness probability will be ignored in the refinement step and which results in the object being incomplete or missing. This can be addressed by ensuring that sufficient images are available of objects during segment proposal generation.

4 Example Images of Reconstructed Point Clouds

We display example images of reconstructed point clouds in this section. In the proposed method, when we reconstruct point clouds of objects, the class of every point is determined by the class of the projecting bounding box, so the reconstructed point clouds are semantically classified (as shown in Figure 3). In this dataset, our 2D images are categorized by room, points in different rooms are isolated by walls. So we reconstruct independent point cloud for every room respectively, which achieves memory efficiency and boosts computational speed. These point clouds of different rooms are in the same world coordinates and can be combined together (as shown in Figure 4).

We can observe some noise in the point cloud, the reason is that the bounding boxes contain not only objects but also background noise, so background regions are also projected into 3D space as noise. Thus, we adopt probabilistic inference to emphasize the correct points and weakens the irrelevant points (background noise). Refer to the top right image of Figure 3 of the main paper to see the results after the probabilistic inference.

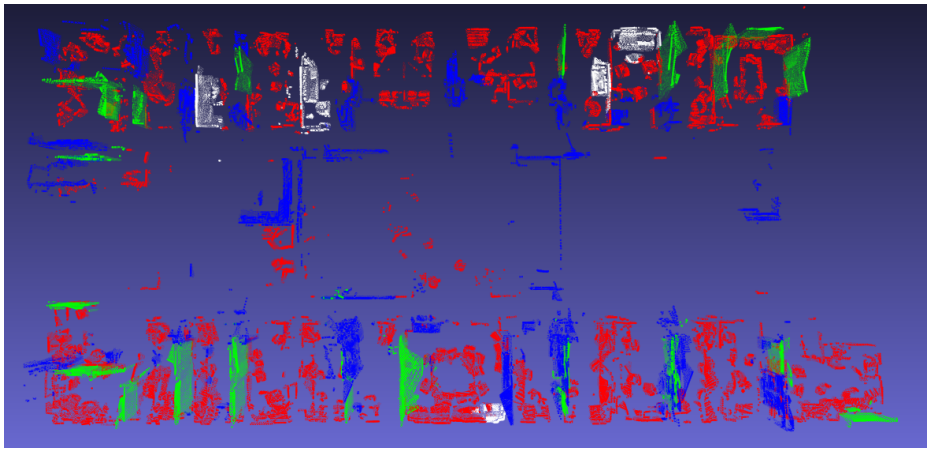

Fig. 4. Combination of point clouds of all rooms. All points are semantically labeled. Some points may locate outside the rooms since the doors of the rooms are not closed.

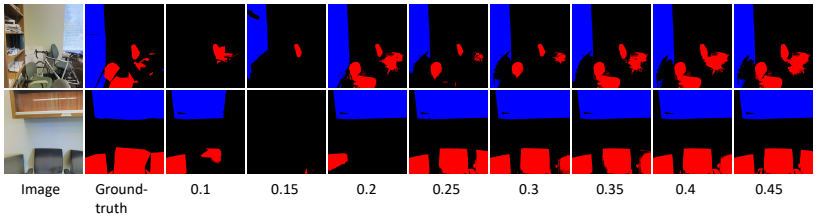

Fig. 5. Examples of the segment proposals generated with different ratios of available annotated data. The numbers under the figures represent different percentages of available annotated data. More annotated data leads to more reliable and accurate segment proposals.

5 Example Images of Segment Proposals of Different Ratios of Available Annotated Data

In our method, we select different ratios of available annotated data and generate segment proposals for training with our proposed pipeline. We evaluate network performances when different percentages of available annotated data is provided. More annotation data leads to manifest performance improvements which validates effectiveness of our method. We visualize segment proposals generated with different ratios of annotated data. As shown in Figure 5, more annotated data provides more information from different camera viewpoints, and leads to more reliable and accurate proposals.

Algorithm 1 Segment Proposal Generation for 3DWSS

Input: collection of images $Y = \{Y_1, ..., Y_i, ..., Y_N\}$;
camera parameters $M_i = (R_i, \tilde{C}_i, f_i)$ and depth map Y_{depth};
A subset of images labeled with bounding boxes $Y_b \subset Y$.
Point Cloud Reconstruction from Bounding Boxes:
for each $Y_i \in Y_b$ **do**
  Get collection of pixels inside the bounding boxes of Y_i: X^i
  for each pixel $x_j^i \in X^i$ **do**
    Project pixel into 3D space: $P_j^i = K_i^{-1}[R_i \mid -R_i\tilde{C}_i]^{-1} * x_j^i * d_j^i$
    where $K_i^{-1} = \begin{bmatrix} \frac{1}{f^i} & & -\frac{p_x^i}{f^i} \\ & \frac{1}{f^i} & -\frac{p_y^i}{f^i} \\ & & 1 \end{bmatrix}$
    Include P_j^i into point cloud P
  end for
end for
Point Cloud Probabilistic Inference:
for each $P_j \in P$ **do**
  Use O_j to represent objectness score of P_j
  for each $Y_i \in Y_b$ **do**
    Get bounding boxes B_i of image Y_i
    if back projected P_j is inside B_i **then**
      $O_j = O_j + 1$
    end if
  end for
end for
Normalize objectness score: $p(O_j) = \frac{O_j}{max(O)}$
Segment Proposal Generation by Point Cloud Back-projection:
for each $Y_i \in Y$ **do**
  for each $P_j \in P$ **do**
    Project point back to 2D images: $x_j^i = K_i[R_i \mid -R_i\tilde{C}_i]P_j$
    where $K_i = \begin{bmatrix} f_i & & p_x^i \\ & f_i & p_y^i \\ & & 1 \end{bmatrix}$
    depth information at x_j^i is d_j^i, distance between P_j and camera is z_j
    if $z_j < d_j^i$ **then**
      x_j^i position of image Y_i is O_j(objectness score).
    end if
  end for
end for
Feed every image Y_i into refinement step to get segment proposals.
Feed segment proposals into network as supervision signal.

6 Error Propagation in Recursive Training

Question might be raised that what would happen if network predictions of the first iteration are sub-optimal. First, refer to [8], suitable iterative technique can refine both label quality and model accuracy. While weakly supervised semantic segmentation methods [1, 2, 9] propose different iterative training methods to refine results gradually.

Second, in our proposed method, the segment proposals become more reliable and accurate when more annotation data is provided, since we have sufficient object information from different views. Thus, our method ensures accurate segment proposals for training. In addition, we adopt bounding box masking to remove irrelevant background regions, which also help improve stability of recursive training. Refer to Table 4 of our main paper, the proposed method achieves reasonably good result in the first iteration and can gradually refine the performance. In future work we would like to achieve our method in a end-to-end manner and explore more about recursive training in the weakly supervised semantic segmentation task.

7 Discussion

Extensive experiments and qualitative results have shown that the proposed method is effective, which helps achieve competitive results with less annotation cost. To provide a comprehensive analysis of our method, we provide the pseudo code of our proposed segment proposal generation pipeline above. Moreover, we discuss some further questions in this section.

First, in our work the supervision information comes from bounding boxes on 1822 images. We compare our method with two state-of-the-art methods trained with 1822 labeled images. Question might be raised that what will the results be if we have bounding box labels on all 4028 training images. We expect obvious performance improvement when more annotated data is provided. However, one of the advantages of our method is, guided by 3D information, we only need to label a small portion of the images and can transductively infer segment proposals on all images, so annotation cost is saved. Refer to Figure 7 of the main paper, When trained using only label 25% of training images, we achieve 70.03 mIoU which already outperforms competing methods (trained on 45%).

In the proposed method, we choose images of area 1 of the 2D-3D-S dataset [7] . In the area 1, there are over 10000 images and the amount we use for training and validation is over 5000, which can already effectively validate our proposed method with four classes. If we put more effort to obtain more annotation data in the future, our method can be easily scaled to more areas and scenes.

In addition, in our early version of method, we tried to use a pre-trained generic object detection network to generate bounding box labels instead of hand annotation. However, currently available trained model performs poorly on our dataset or does not have same classes. Hand labeled bounding boxes are still required to fine-tune the model on the 2D-3D-S dataset [7]. It will be

interesting to explore noisy bounding boxes that predicted by network in weakly supervised semantic segmentation, we leave this problem as our future work.

References

1. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2017) 876–885
2. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proc. IEEE Int. Conf. Comp. Vis. (2015) 1742–1750
3. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2016) 128–140
4. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG). Volume 23. (2004) 309–314
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. Eur. Conf. Comp. Vis. (2018)
6. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Proc. Adv. Neural Inf. Process. Syst. (2011) 109–117
7. Armeni, I., Sax, A., Zamir, A.R., Savarese, S.: Joint 2D-3D-Semantic Data for Indoor Scene Understanding. ArXiv e-prints (2017)
8. Haase-Schütz, C., Stal, R., Hertlein, H., Sick, B.: Trust your model: Iterative label improvement and robust training by confidence based filtering and dataset partitioning. arXiv preprint arXiv:2002.02705 (2020)
9. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proc. IEEE Int. Conf. Comp. Vis. (2015) 1635–1643