X-PDNet: Accurate Joint Plane Instance Segmentation and Monocular Depth Estimation with Cross-Task Distillation and Boundary Correction

Cao Dinh Duc[1,3]
duccd@hanyang.ac.kr
Jongwoo Lim[2]
jongwoo.lim@snu.ac.kr

[1] Dept of Computer Science
Hanyang University
Seoul, South Korea

[2] Dept of Mechanical Engineering
Seoul National University
Seoul, South Korea

[3] AI Research Lab, MAXST Co., Ltd
Seoul, South Korea

Abstract

Segmentation of planar regions from a single RGB image is a particularly important task in the perception of complex scenes. To utilize both visual and geometric properties in images, recent approaches often formulate the problem as a joint estimation of planar instances and dense depth through feature fusion mechanisms and geometric constraint losses. Despite promising results, these methods do not consider cross-task feature distillation and perform poorly in boundary regions. To overcome these limitations, we propose X-PDNet, a framework for the multitask learning of plane instance segmentation and depth estimation with improvements in the following two aspects. Firstly, we construct the cross-task distillation design which promotes early information sharing between dual-tasks for specific task improvements. Secondly, we highlight the current limitations of using the ground truth boundary to develop boundary regression loss, and propose a novel method that exploits depth information to support precise boundary region segmentation. Finally, we manually annotate more than 3000 images from Stanford 2D-3D-Semantics dataset and make available for evaluation of plane instance segmentation. Through the experiments, our proposed methods prove the advantages, outperforming the baseline with large improvement margins in the quantitative results on the ScanNet and the Stanford 2D-3D-S dataset, demonstrating the effectiveness of our proposals. The code is available at: https://github.com/caodinhduc/X-PDNet-official.

1 Introduction

Piecewise planar regions frequently appear in man-made environments, especially in indoor scenes (wall, floor, furniture, etc.). The detection and segmentation of such piecewise planar surfaces in images has attracted much attention due to its wide range of applications.

In the indoor environment, planar instance segmentation offers an essential representation for scene understanding [17], augmented reality (AR) applications, robot navigation, and visual SLAM [15]. In the outdoor scenes, ground and wall plane cues benefit 6-DoF object pose estimation, building reconstruction [10], and drivable surface detection in autonomous driving. Recently, with the advancement of deep neural networks, the piecewise estimation of planar region can be reformulated to the plane instance segmentation task. Starting with PlaneNet [11] and PlaneRecover [24], which make breakthroughs in using convolutional neural networks (CNNs) to segment planar or non-planar region instances. Next, PlaneR-CNN [12] inherits Mask R-CNN [7] to segment plane instances with their plane parameters and segmentation masks. PlaneSegNet [21] builds upon Yolact++ [1], which was presented as the first real-time single-stage plane instance segmentation method in this field. PlaneRec-Net [22] forms a multi-task learning framework by jointly training a single-stage plane instance segmentation network with depth estimation from a single RGB image. Unlike other existing approaches [11, 12, 24, 25], PlaneRecNet concentrates on enforcing cross-task consistency by introducing multiple loss functions (geometric constraints) that cooperatively enhance the accuracy of plane instance segmentation and depth estimation. Despite achieving solid quantitative results on both tasks besides computational efficiency, PlaneRecNet still has several limitations that need to be improved. 1) Since the instance segmentation mask candidates are fused to hidden depth features through multiplication and concatenation computations, this design inherently limits the adaptive feature distillation capability between cross-tasks and further limits the performance of the plane instance segmentation while over-focusing on depth estimation. 2) Current single-stage plane instance segmentation methods do not explicitly utilize the boundary information of the ground truth masks, which results in imprecise predicted masks. Furthermore, because the ground truth plane masks are generated by RANSAC-based methods, it produces incorrect and coarse boundary ground truth instances. Hence, predicted masks optimized by traditional boundary regression loss not to be tightly aligned to the true boundaries. To address these issues, we propose X-PDNet (X indicates a cross design), a framework for joint plane instance segmentation and depth estimation, which is based upon PlaneRecNet [22] with several major improvements. We introduce the cross-task distillation design, where distillation modules are dual-integrated between the aggregated depth feature layer and the feature mask layer of SOLO V2 [19] network. In addition, we propose **Depth Guided Boundary Preserving Loss**, which alleviates the effect of incorrect ground truth masks by evaluating the gradient difference between the boundary ground truth and its neighbors at the pixel level. Our main contributions can be summarized as follows:

- Developing from PlaneRecNet [22], we design X-PDNet, a multi-task learning framework for joint plane instance segmentation and depth estimation, which allows the respective task decoder to adaptively distill the cross-supplementary information for the specific task optimization.
- We introduce a novel Depth Guided Boundary Preserving Loss, which combats noisy ground truth to produce more accurate segmentation at boundary regions.
- We contribute manual annotations of over 3000 images from the Stanford 2D-3D-Semantics dataset as a reliable evaluation set for plane instance and boundary segmentation.
- Extensive experiments on the ScanNet and the 2D-3D-S datasets demonstrate the effectiveness of our method in both plane instance segmentation and depth estimation tasks by a large margin improvements over previous methods with no additional computational cost.

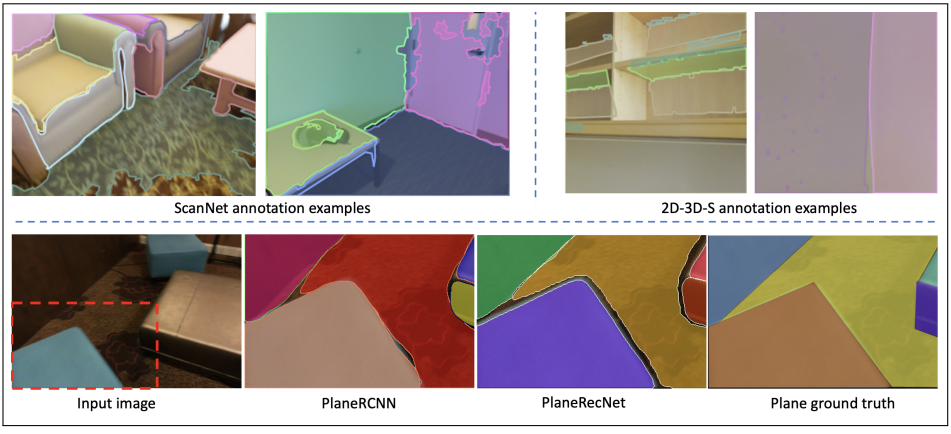

Figure 1: The first row shows examples of incorrect instance ground truth from different datasets. The second row visualises the segmentation results produced by existing methods on the ScanNet [5] dataset with poor quality predictions at the boundary regions of planes.

2 Related Work

Plane Instance Segmentation: PlaneNet [11] is the first attempt employing a deep neural network to reconstruct piecewise planar regions from a single RGB image. It shares an encoder and provides three prediction branches: plane parameter estimation, plane segmentation, and non-planar depth map estimation. Later, in PlaneRecover [24], Yang and Zhou indicate the obstacles to obtaining the ground truth of the plane annotation dataset. Then they present a novel plane structure-induced loss to train the plane segmentation and plane parameter estimation for outdoor scenes through an unsupervised learning approach. In spite of generating promising results, both PlaneNet and PlaneRecover require a fixed number of predicted planar regions, which severely restricts the generalization capabilities of the application to different scenarios. PlaneAE [25] trains a CNN to map each pixel to an embedding space where pixels from the same plane instance have similar embeddings. It then groups embedding vectors into piecewise plane instances using its mean shift clustering algorithm. PlaneRCNN [12] proposes an effective plane segmentation branch built upon Mask R-CNN [7] and jointly refines the segmentation mask with their novel warping loss function. The method shows high localization ability and generalization across different domains but fails to achieve real-time execution. PlaneSegNet [21] introduces a fast single-stage instance segmentation method for high-resolution piece-wise planar regions, the approach adapts strongly at large-scale planar regions but misses depth estimation. Differently, PlaneRecNet [22] designs a multi-task network for jointly studying plane instance segmentation and depth estimation and boosting the cross-task consistency by exploiting geometric constraints.

Cross-Task Distillation Mechanism: Related to our work are methods that facilitate feature sharing or distillation across tasks, inspired by the idea that each task could benefit from complementary information from the others. PAD-Net [23] uses an attention mechanism to distill information across multimodal features. MTI-Net [18] extends PAD-Net with a multi-scale solution to better distill multimodal information. [28] proposes to learn a single-task affinity matrix, then which is then combined to diffuse and refine the task-specific features. [2] in-

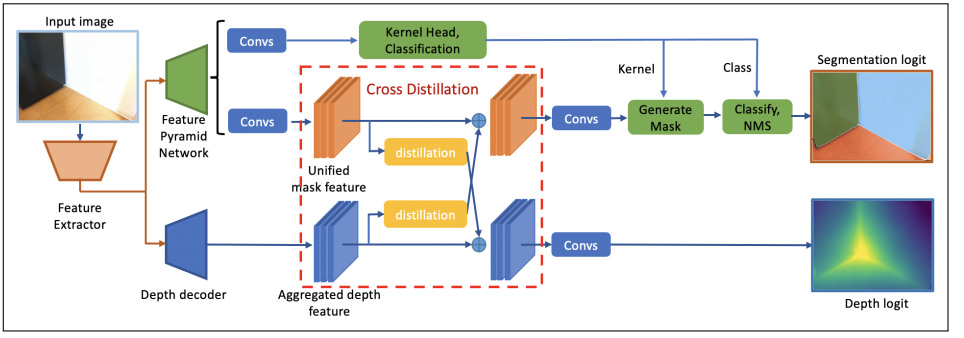

Figure 2: The architecture of **X-PDNet**: The network consists of a shared backbone and two parallel branches for plane instance segmentation and monocular depth estimation. A couple of Cross Distillation Modules are integrated between the unified mask feature of the segmentation branch and aggregated feature of the depth decoder to facilitate early cross-task feature distillation. Detail of distillation module is described in the Fig. 3.

troduces ATRC to refine each task prediction by capturing cross-task contexts dependent on four relational context types.

Boundary Preserving Loss for Segmentation: Obtaining sharp boundaries is important for the high-quality instance segmentation task. Existing methods are solid in terms of plane localization but do not pay attention to the exploitation of boundary information. As a result, these models produce planes with coarse and imprecise contours that are typically illustrated by the overlaps or the gaps between two adjacent planes as shown in Fig.1. Observing in the segmentation field, enhancing segmentation accuracy in boundary regions has been studied in some existing methods [4, 9, 16, 20, 26, 29] but mostly developed for detection-then-segmentation methods. BMask R-CNN [4] achieves a better result by combining the representation of object boundaries to guide mask prediction. Gate-SCNN [16] jointly supervised segmentation and boundary map prediction. [20] introduces a boundary-preserving reweighting mechanism that forces the model to focus on boundary-relevant areas. BSOLO [27] designs a Hungarian algorithm based border loss to calculate the cost of matching between borders. While these methods show that they can lead to higher quality predicted masks, they still suffer from several limitations, including the high computational cost due to the additional branch for edge detection, the lack of ideal edge ground truth, and the unstable or low quality of predicted edges.

3 Method

3.1 X-PDNet Overview

Our proposed X-PDNet is built upon the PlaneRecNet [22] with several major improvements to address the aforementioned problems. As described in Fig. 2, given a single color image as an input, our network consists of two branches with a shared backbone to predict a piece-wise planar segmentation S_{pred} and a depth estimation D_{pred} in parallel. A couple of distillation modules are dual-integrated between the aggregated feature of the depth decoder and the mask feature of the segmentation branch to distill cross-task complementary signals.

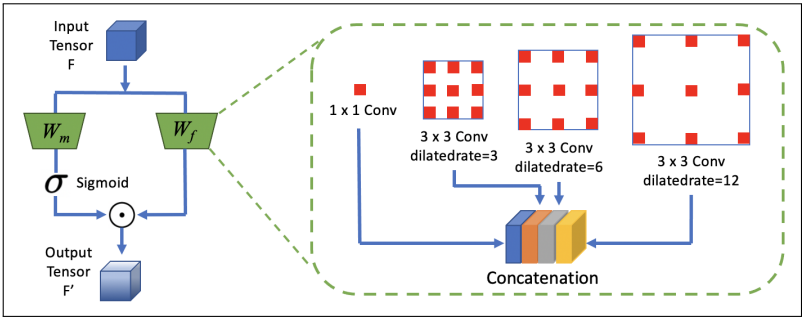

Figure 3: Illustration of the cross-task distillation module, which involves a feature branch (convolution layers with different dilated rates (green box)) and mask branch (a convolution layer with a sigmoid function to construct an attention mask), and the output feature is combined by the element-wise multiplication.

3.2 Cross Distillation Design

Attention or distillation mechanisms [2, 13, 18, 23, 28] have been commonly used to facilitate cross-task optimization in multi-task learning for a long time, this builds on the intuition that each decoder could learn from the complementary signal of another branch. Moreover, since the cross-task feature is not always beneficial for the primary task, the distillation module can act as a filter to select only useful information from the other tasks. Considering the baseline method [22], which is based on SOLO V2 [19] for the plane instance segmentation branch, where the mask candidates are fused into the depth branch through the Plane Prior Attention module, we argue that fusing plane-predicted masks into depth aggregated features imposes the model optimize for the depth estimation task but may affect the segmentation accuracy due to the depth backpropagation gradient through the plane instance mask candidates. To facilitate the early cross-task information distillation, in our work, we introduce a lightweight but efficient cross-distillation design to guide the message passing between the aggregated feature maps generated by the depth decoder and feature mask of the segmentation branch as illustration in Fig. 2. We leverage the idea presented in PAD-NET [23] with a reformulation to help the model adapt robustly with multiple scale plane instances as reported in [21]. Given the context that we want to pass the message from the secondary task to facilitate the primary task. As visualization in Fig. 3, firstly, an attention map (the output of sigmoid function) A is generated from the secondary task feature F as follows:

$$A \leftarrow \sigma(W_m \otimes F), \tag{1}$$

Where W_m is the 2-D convolution parameters and σ is a sigmoid function to normalize the attention map. Then the message passed from the secondary task F is controlled by the attention map A as described:

$$F' \leftarrow A \odot (W_f \otimes F). \tag{2}$$

In the equation 2, inspired by the ASPP design [3], we extract the cross-task feature F by a set of convolution layers (W_f) with different dilation rates ([1, 3, 6, 12] in our experiments) to enlarge the spatial scale of cross-task contexts, then stack the outputs together as depiction in the Fig. 3, $\odot$ and $\otimes$ denote the element-wise multiplication and convolution operation. Finally, the passed message F' is merged into the primary task for specific task optimization

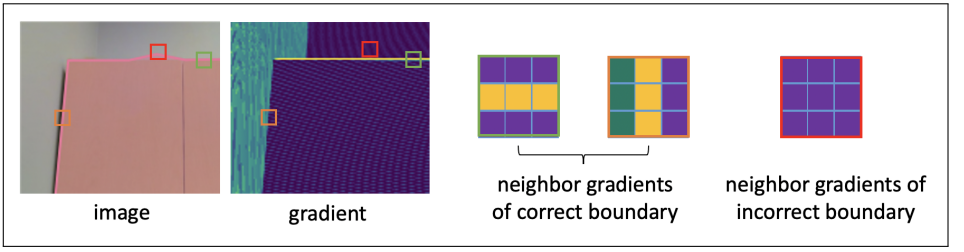

Figure 4: Visualization of gradient based analysis. With a boundary ground truth point (center of each window), we consider its depth gradient and that of its neighbors within a window, yellow elements indicate boundary points.

as shown in Fig. 2. Through the experiments in Tab. 3, we demonstrate the effectiveness of our design as well as the importance of receptive field expansion in the cross-task feature distillation by significantly improving from not only the segmentation but also the depth estimation.

3.3 Depth Guided Boundary Preserving Loss (DGBPL)

The traditional approach (**Vanilla**) to the problem of poor performance at boundary regions is boundary regression loss, which obtains the ground truth and predicted boundaries by the edge detectors (Sobel, Laplacian), then uses Mean Square Error to teach the predicted masks to align with the true boundaries. However, since the quality of the GT masks is poor (Fig. 1 first row), it forces the segmentation method trained with the vanilla method produce segmentation masks that is far from ideal as example in Fig. 6. To alleviate this limitation, we evaluate the confidence of the ground truth boundaries at pixel level by measuring the difference between its depth gradient with that of its neighbors. We observe that the depth gradient fluctuates slightly over the plane area but changes suddenly at the occluded areas or junctions between adjacent planes (Fig. 4). To exploit this constraint, we first construct the gradient mask G_{gt} from the ground truth depth D_{gt} using Sobel-Filter:

$$G_{gt} = abs(G_x) + abs(G_y) \quad \text{with} \quad G_x = Sobel_x(D_{gt}),\ G_y = Sobel_y(D_{gt}). \tag{3}$$

As presented in equation. 3, we formulate this mask as a combination of absolute gradients following the x and y directions. Next, for each pair ground truth mask $y_m^{gt} \in (H/4 \times W/4)$ and predicted mask $y_m^{pr} \in (H/4 \times W/4)$ generated by the plane instance segmentation branch, we follow the traditional edge detection method (Laplacian operation) to obtain the ground truth boundary (y_b^{gt}) and the predicted boundary (y_b^{pr}), respectively. For each boundary point in (y_b^{gt}), we consider the local gradient variation by obtaining the corresponding gradient values within the window (3x3 with the target point at the center in our experiment), then measure the standard deviation of these points. As visualization in Fig. 4, we expect the standard deviation (std) computed from the correct GT boundary (green and orange boxes) will be higher than that computed from the incorrect GT boundary (red box). We then normalize these std values to estimate the weights (W) before using them to reweight the boundary regression loss (MSE($y_b^{gt} * W$, $y_b^{pr} * W$)) at the pixel level to guide the this loss to focus on the correct contour while reducing the impact of noise ground truth boundaries. DGBPL mitigates the impact of an imperfect plane GT mask on normal boundary regression

loss, as demonstrated by better results on a manually annotated dataset while maintaining the same training set. See section. 5.2 for a detailed evaluation.

4 Experiments Setup

4.1 Datasets and metrics

To measure the performance of our proposals, we conduct experiments on two public datasets: ScanNet with annotation provided by [12] and 2D-3D-S with annotation provided by [21]. For the 2D-3D-S dataset, we additionally test the networks on our manually annotated evaluation set to figure out clearly the effectiveness of the proposed loss function. For the quantitative metrics in plane instance segmentation, we use Average Precision for both masks (AP_m) and bounding boxes (AP_b) at different NMS thresholds (overall, 50, and 75). In terms of depth estimation evaluation, the metrics include Absolute Relative Error (rel), Log 10 error (log_{10}), linear Root Mean Square Error (RMS), and accuracy under the thresholds ($\delta_1, \delta_2, \delta_3$).

4.2 Implementation details

Similar to PlaneRecNet [22], our proposed X-PDNet is implemented using the Pytorch [14] framework. It adopts ResNet101 [6] with deformable convolution [30] as the backbone network. We use Adam optimizer [8] and a batch size of 8 images for model training. For a fair comparison, we keep the loss functions, loss weights, and training strategies from the baseline [22], To be more specific, losses include:

$$L = L_{Focal} + L_{Dice} + L_{RMSE} + L_{constraints} + DGBPL \tag{4}$$

Where focal and dice losses are for the segmentation task, RMSE is for the optimization of the depth estimation, and geometric constraint losses. Our model is trained for 10 epochs on ScanNet and 15 epochs on 2D-3D-S with the plane annotation given by [12] and [21], respectively. In both datasets, training data is augmented with random photometric distortion, horizontal and vertical flipping, and Gaussian noise. All training sessions are conducted on an NVIDIA RTX A5000 GPU device.

5 Experiments

5.1 Comparison with existing methods

Methods	Dataset	Segmentation Metrics						Depth Metrics					
		AP_m	AP_m^{50}	AP_m^{75}	AP_b	AP_b^{50}	AP_b^{75}	$rel \downarrow$	$log_{10} \downarrow$	$RMS \downarrow$	$\delta 1$	$\delta 2$	$\delta 3$
PlaneAE [23]	ScanNet	5.92	14.72	4.00	7.86	17.83	6.25	0.111	0.049	0.409	0.864	0.967	0.991
PlaneRCNN [12]	ScanNet	14.23	28.23	12.88	17.51	33.00	16.00	0.124	0.050	0.265	0.865	0.972	0.994
PlaneRecNet [22]	ScanNet	16.61	31.59	15.56	21.05	36.45	20.29	0.076	0.032	0.180	0.950	0.992	0.998
X-PDNet	ScanNet	**17.62**	**33.05**	**16.60**	**22.23**	**37.53**	**21.91**	**0.069**	**0.029**	**0.175**	**0.955**	**0.993**	**0.999**
PlaneRecNet [22]	2D-3D-S	24.10	38.99	24.39	27.13	41.14	27.23	0.062	0.027	**0.294**	**0.966**	0.990	**0.996**
X-PDNet	2D-3D-S	**25.20**	**39.63**	**25.79**	**28.62**	**41.80**	**29.15**	**0.061**	**0.026**	**0.294**	**0.966**	**0.991**	**0.996**

Table 1: Evaluation of plane instance segmentation and depth estimation on **ScanNet** and **2D-3D-S** datasets. **X-PDNet** outperforms existing methods in most metrics.

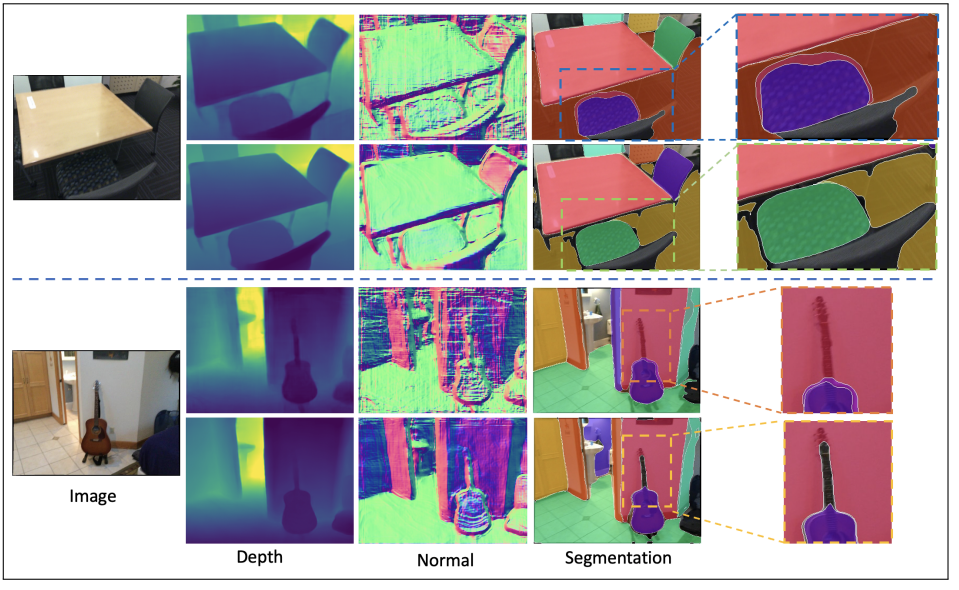

Figure 5: Qualitative comparison between X-PDNet and the baseline (PlaneRecNet) on images from the ScanNet dataset. It contains two examples, for each one, the first row is the output of **PlaneRecNet**, while the second row is generated by **X-PDNet**, (normal is recovered from the predicted depth). The obvious difference can be seen in the rectangle boxes. In the chair area of the first example, with cross-task feature distillation, X-PDNet is able to distinguish the chair surface from the floor, even though the RGB feature in this area is quite similar. In terms of depth, ours is better with the perception of visual information from the segmentation branch, resulting in a smooth normal vector converted from depth prediction in each plane area. The same improvement is observed in the second example (guitar surface).

This section is to illustrate the effectiveness of Cross Distillation Design, proved by remarkable improvements over the existing approaches on evaluation datasets. We first evaluate our proposed model on the ScanNet dataset which is the most popular dataset in plane instance segmentation with annotation generated by [12]. We utilize the same data setup with [12] and [22], which contains 100,000 training and 5,000 test images. Next, we further conduct the experiments on the 2D-3D-S dataset, which includes 60,000 training images and 5,000 test images. As the quantitative results are shown in Tab. 1, for the ScanNet dataset, X-PDNet outperforms the existing methods by a large margin in both task plane instance segmentation and depth estimation. Furthermore, for the 2D-3D-S dataset, there is still a large improvement in the segmentation performance of X-PDNet compared to the baseline, While the performance in terms of depth metrics increases slightly. Fig. 5 shows the qualitative improvements of **X-PDNet** compared to the baseline (**PlaneRecNet**). With Cross Distillation Design, the segmentation estimator has better geometric understanding to predict more accurate plane masks, especially in occluded areas or areas where RGB information is ambiguous. Meanwhile, perception of visual information allows the depth logit to be smoother in planar areas, resulting in noise reduction.

5.2 Evaluation of Depth Guided Boundary Preserving Loss

Because the annotation of plane instances at boundary regions of existing datasets is underqualified to verify the contribution of Depth Guided Boundary Preserving Loss. We provide a manually annotated label on the 2D-3D-S dataset. It is separate from the training and test set provided by [21]. To measure how the depth gradient is beneficial in guiding the boundary preserving loss. We compare the segmentation performance of X-PDNet adding **DGBPL**, with and without vanilla boundary regression loss. In addition to segmentation metrics, we use the Boundary IoU $((y_b^{gt} \cap y_b^{pr})/(y_b^{gt} \cup y_b^{pr}))$ measurement to claim that our proposed technique is beneficial for boundary region prediction. The details of the comparison given in Tab. 2 show that **DGBPL** outperforms the vanilla regression loss in the manually annotated, while being competitive in the original evaluation set. When guided by the depth gradient, **X-PDNet** produces the correct predictions in boundary regions, as shown by the reduced overlaps and narrow apertures between two adjacent planes as examples in Fig. 6. Examples of the original planar ground truth and after manual correction can be found in the supplementary document.

Methods	Eval set	Boundary IoU	Segmentation Metrics					
			AP_m	AP_m^{50}	AP_m^{75}	AP_b	AP_b^{50}	AP_b^{75}
X-PDNet	Provided by [21]	-	25.20	39.63	25.79	28.62	41.80	29.15
X-PDNet+Vanilla	Provided by [21]	-	**26.49**	41.61	**27.09**	**30.23**	44.18	**30.7**
X-PDNet+DGBPL	Provided by [21]	-	25.86	**41.79**	26.34	29.94	**45.55**	29.98
X-PDNet	Manually annotated	13.36	24.09	36.84	25.08	25.80	37.08	26.72
X-PDNet+Vanilla	Manually annotated	14.82	25.27	38.24	26.59	27.08	38.93	**27.77**
X-PDNet+DGBPL	Manually annotated	**16.68**	**26.12**	**39.47**	**26.68**	**28.18**	**40.86**	27.46

Table 2: Evaluation of segmentation results on **2D-3D-S** annotation provided by [21] and human labelling evaluation datasets.

6 Ablation Study

Since our distillation module is based on the attention-guided message passing mechanism introduced in PAD-Net [23]. In this section, we analyze how our proposed modification affects the performance of joint instance segmentation and depth estimation. Specifically, we train the baseline **PlaneRecNet** (Plane Prior Attention), with no attention, cross design with attention module presented in PAD-Net [23], and our (**X-PDNet**). The comparison shown in Tab. 3 demonstrates the effectiveness of the cross-task distillation module (Fig. 3) through the quantitative improvements in both depth and segmentation metrics. Refer the supplementary material for detail architecture of each design.

Attention/ distillation	Segmentation Metrics						Depth Metrics					
	AP_m	AP_m^{50}	AP_m^{75}	AP_b	AP_b^{50}	AP_b^{75}	$rel \downarrow$	$log_{10} \downarrow$	$RMS \downarrow$	$\delta 1$	$\delta 2$	$\delta 3$
No Attention or distillation	16.05	30.38	14.99	20.82	35.77	19.86	0.078	0.033	0.183	0.950	0.992	0.997
Plane Prior Attention [23]	16.61	31.59	15.56	21.05	36.45	20.29	0.076	0.032	0.180	0.950	0.992	0.998
PAD-Net [23]	17.41	32.54	16.48	22.11	37.24	**21.96**	0.071	0.031	0.176	**0.955**	0.992	0.998
Ours	**17.62**	**33.05**	**16.60**	**22.23**	**37.53**	21.91	**0.069**	**0.029**	**0.175**	**0.955**	**0.993**	**0.999**

Table 3: Ablation study of the performance of the network with different selection of attention or distillation designs on **ScanNet** dataset. **Ours** performs better in both tasks.

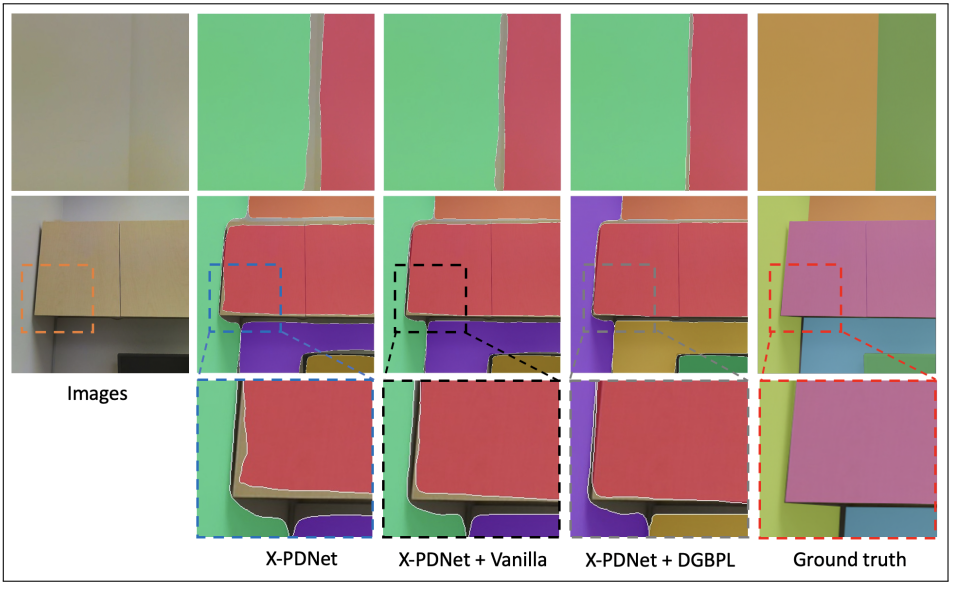

Figure 6: Effect of Depth Guided Boundary Preserving Loss (**DGBPL**) on the segmentation results on 2D-3D-S examples compared to traditional regression boundary loss. With (**DGBPL**), X-PDNet performs impressively at boundary related regions. Focus on rectangle boxes for clear difference.

7 Conclusion

In this paper, we present two techniques to achieve precise joint learning of plane instance segmentation and depth estimation. We formulate a cross-task distillation design and explicitly exploit the depth information support for accurate segmentation at boundary related regions. Through extensive experiments, we demonstrate the effectiveness of our proposals by a considerable improvement in both tasks compared to the baselines.

8 Acknowledgments

This work was partially supported by the Technology Innovation Program (No. 20018110, "Development of a wireless tele-operable relief robot for detecting searching and responding in narrow space") funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1A2C2010245).

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++ better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1108–1121, 2022.

[2] David Bruggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15849–15858, 2021.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[4] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *ECCV*, 2020.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[9] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2020.

[10] Minglei Li, Liangliang Nan, Neil G. Smith, and Peter Wonka. Reconstructing building mass models from uav images. *Comput. Graph.*, 54:84–93, 2016.

[11] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018.

[12] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4445–4454, 2019.

[13] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. *ArXiv*, abs/2206.08927, 2022.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019.

[15] Jason R. Rambach, Paul Lesur, Alain Pagani, and Didier Stricker. Slamcraft: Dense planar rgb monocular slam. *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6, 2019.

[16] Towaki Takikawa, David Acuna, V. Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5228–5237, 2019.

[17] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. *2011 International Conference on Computer Vision*, pages 121–128, 2011.

[18] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020.

[19] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2020.

[20] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16805–16814, 2022.

[21] Yaxu Xie, Jason R. Rambach, Fangwen Shu, and Didier Stricker. Planesegnet: Fast and robust plane estimation using a single-stage instance segmentation cnn. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13574–13580, 2021.

[22] Yaxu Xie, Fangwen Shu, Jason R. Rambach, Alain Pagani, and Didier Stricker. Planerecnet: Multi-task learning with cross-task consistency for piece-wise plane detection and reconstruction from a single rgb image. In *BMVC*, 2021.

[23] Dan Xu, Wanli Ouyang, Xiaogang Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[24] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *ECCV*, 2018.

[25] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1029–1037, 2019.

[26] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6857–6865, 2021.

[27] Yuxuan Zhang and Wei Yang. Bsolo: Boundary-aware one-stage instance segmentation solo. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2594–2598, 2022.

[28] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, N. Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4101–4110, 2019.

[29] Chen Zhu, Xuanye Zhang, Yanran Li, Liangdong Qiu, K. Han, and Xiaoguang Han. Sharpcontour: A contour-based boundary refinement approach for efficient and accurate instance segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4382–4391, 2022.

[30] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9300–9308, 2019.