Supplementary Material:
Peek-a-boo: Occlusion Reasoning in Indoor Scenes with Plane Representations

Ziyu Jiang[1,3*] Buyu Liu[1] Samuel Schulter[1] Zhangyang Wang[3] Manmohan Chandraker[1,2]
[1]NEC Laboratories America [2]UC San Diego [3]Texas A&M Universaity

This supplemental material contains the following details that we could not include in the main paper due to space restrictions.

- **(Sec. 1)** Detailed Description of Proposed Methods
- **(Sec. 2)** More Details for Semantic Merging Module
- **(Sec. 3)** Comparison of Different Warping Methods
- **(Sec. 4)** More Qualitative Results

1. Detailed Description of Proposed Methods

1.1. Semantic Merging Module

We depict the algorithm of Semantic Merging process for single input image in Algorithm. 1. Multiple images can be processed iteratively.

1.2. Plane Warping Module

We depict the algorithm for calculating plane warping loss of single input image in Algorithm. 2. The total loss of one batch can be averaged among losses of all images.

2. More Details for Semantic Merging Module

As mentioned in our main paper, we utilize semantic segmentation results to effectively merge the predictions from both layout and object branch. We introduce more details for the semantic segmentation model in the following.

Model: We utilize ENet [2] as our semantic segmentation network and obtain the per-pixel foreground and background outputs out of it for Semantic Merging module. Compare to other networks, ENet achieves good trade-offs between efficiency and accuracy. Note that better semantic segmentation networks can provide us higher performance and potentially lead to better merging results. We leave the exploration of SOTA semantic segmentation models to our future work.

Training setting: To train ENet, we follow the same training, validation and testing split with the proposed occlusion reasoning model. The input image is down-sampled to 240×320 for fast training and inference. We set the batch size to 64 and train our segmentation model for 10 epochs. The initial learning rate is set as 1e-3 and decreases by 10× every 3 epochs.

Quantitatively performance: We can achieve the mIOU of 65.97 in term of Layout/Object segmentation on the testing set.

3. Comparison of Different Warping Methods

To demonstrate the effectiveness of our proposed plane warping module from Sec. 3.2 of the main paper, we compare with the depth-based warping as suggested in PlaneRCNN [1]. Specifically, we apply the depth-based warping module by first generating a depth map with predicted complete masks. Then we map each pixel to 3D coordinates w.r.t. its depth prediction and obtain 3D points. We further transform these 3D points to nearby views w.r.t. camera poses. Finally, the average L2 distance between the transformed points and ground truth 3D points in nearby views are utilized as the loss. We refer to this baseline as Depth Warping and show the quantitative results in Tab. 1.

As shown in Tab. 1, the proposed Plane Warping Module can improve the AP value by at lease 0.8% and the APH by at least 0.2% compared to the Depth Warping baseline.

4. Additional Qualitative Results

Finally, we provide more qualitative examples in this section. We firstly show the 3D reconstruction results of our proposed method. As can be seen in Fig. 1, our method is able to predict the occluded regions quite well from the given single perspective RGB image. We also highlight the layout and object prediction results in 3D. Compared to the

Table 1: Comparison of the proposed Plane Warping Module with the Depth Warping Module of PlaneRCNN [1] in terms of AP and APH values. The proposed method can achieve an overall improvement.

	AP0.4	AP0.6	AP0.9	APH0.4	APH0.6	APH0.9
Depth Warping	0.325	0.374	0.397	0.095	0.109	0.116
Plane Warping	**0.334**	**0.382**	**0.405**	**0.097**	**0.111**	**0.118**

*Part of this work was conducted during a summer internship at NEC Laboratories America.

Algorithm 1: Semantic Merging for single input image

Input: The input image I; Trained semantic segmentation model $\mathcal{F}$; Overlap threshold θ;
Input: Layout plane predictions $\{P_{L_i}\}_{i=1}^{n}$, which is composed by complete mask, visible mask, normal and offset $\{M_{L_i}^C, M_{L_i}^V, \vec{n}_{L_i}, o_{L_i}\}_{i=1}^{n}$.
Input: Object plane predictions $\{P_{O_j}\}_{j=1}^{n}$, which is composed by complete mask, visible mask, normal and offset $\{M_{O_j}^C, M_{O_j}^V, \vec{n}_{O_j}, o_{O_j}\}_{j=1}^{n}$.
Initialize an empty drop set D ;
Predict the layout segmentation map S_L and object segmentation map S_O with $\{S_L, S_O\} = \mathcal{F}(I)$;
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** m **do**
    Calculate the Intersection of Union iou between $M_{L_i}^V$ and $M_{O_j}^V$;
    if $P_{L_i} \notin D$ ***and*** $P_{O_j} \notin D$ ***and*** $iou > \theta$ **then**
      Calculate the confidence score of P_{L_i} with $c_i^L = \frac{\text{Area}(S_L \cap M_{L_i}^V)}{\text{Area}(M_{L_i}^V)}$;
      Calculate the confidence score of P_{O_j} with $c_j^O = \frac{\text{Area}(S_O \cap M_{O_j}^V)}{\text{Area}(M_{O_j}^V)}$;
      if $c_i^L > c_j^O$ **then**
        Add P_{O_j} to the drop set D;
      else
        Add P_{L_i} to the drop set D;
      end
    end
  end
end
The merged planes set is $\{P_M\} = \{P_L\} \cup \{P_O\} \cap \overline{D}$

Algorithm 2: Plane warping loss calculation for single input images

Input: Intrinsic parameters I; Extrinsic parameters of the original viewpoint and a neighbor[a] viewpoint $\{E_o, E_n\}$; intersection of union (IoU) calculation function IoU; Matching IoU threshold η_{iou}; Matching depth threshold η_{depth}; Plane warping function $\mathcal{W}$; Merged plane predictions $\{P_i\}_{i=1}^{m}$, which is composed by complete mask, visible mask, normal and offset $\{M_{P_i}^C, M_{P_i}^V, \vec{n}_{P_i}, o_{P_i}\}_{i=1}^{m}$; Ground truth planes of another point of view $\{G_j\}_{j=1}^{n} = \{M_{G_j}^C, M_{G_j}^V, \vec{n}_{G_j}, o_{G_j}\}_{j=1}^{n}$; Cross Entropy loss $\mathcal{L}_{CE}$;
Initialize an empty pair set S;
Warp prediction planes $\{P_{W_i}\}_{i=1}^{m} = \mathcal{W}(\{P_i\}_{i=1}^{m}; I, E_o, E_n)$;
for $i \leftarrow 1$ **to** m **do**
  Find the plane $M_{G_j}^C$ with highest IoU$(M_{P_{W_i}}^C, M_{G_j}^C)$ among all $\{G_j\}_{j=1}^{n}$ with $D_n(P_{W_i}, G_j) > \eta_{\text{depth}}$[b];
  if $IoU(M_{P_{W_i}}^C, M_{G_j}^C) > \eta_{iou}$ **then**
    Add the pair (P_{W_i}, G_j) to pair set S.
  end
end
The loss $\mathcal{L} = \frac{1}{\text{length}(S)} \mathcal{L}_{CE} \sum_{k \in S} (M_{P_{W_k}}^C, M_{G_k}^C)$;

[a] We follow the Warping Loss Module in [1] for picking neighbor viewpoint.
[b] $D_n(P_{W_i}, G_j) = \left\| o_{P_{W_i}} \cdot \vec{n}_{P_{W_i}} - o_{G_j} \cdot \vec{n}_{G_j} \right\|^2$

baseline method, PlaneRCNN-OR, we further show that the proposed method outperforms it significantly (see Fig. 2 for more details).

References

[1] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *The IEEE Conference on Computer*

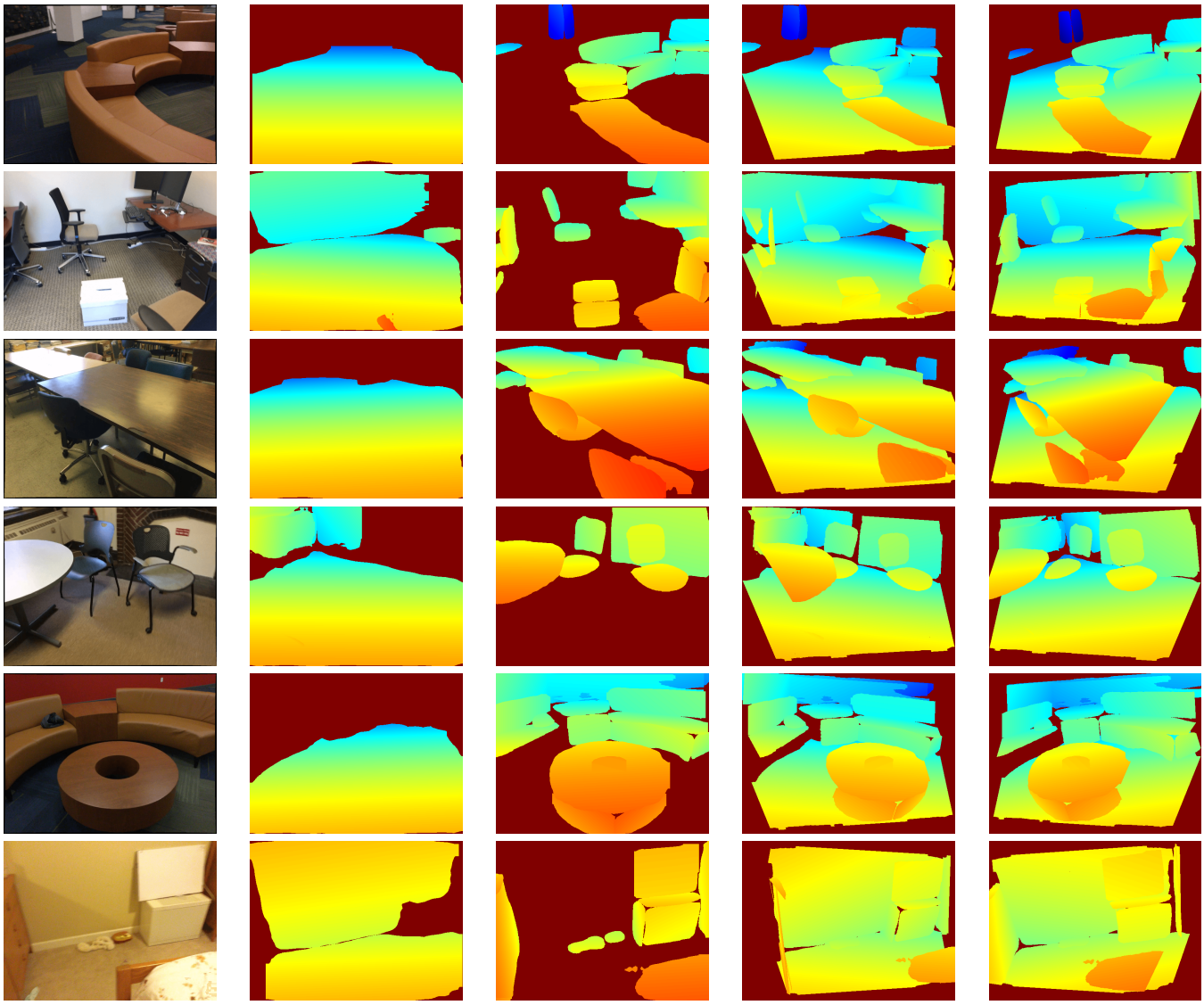

Figure 1: Visualization of proposed method, The first column is the input image. The second column and third column show the depth maps of layout and object 3D reconstruction from predicted complete planes. The fourth column and the last column are two depth maps from novel views. As can be viewed in our examples, the proposed method is able to predict complete planes that can be used for 3D reconstruction with aware of occlusion area.

Vision and Pattern Recognition (CVPR), June 2019. 1, 2, 4

[2] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 1

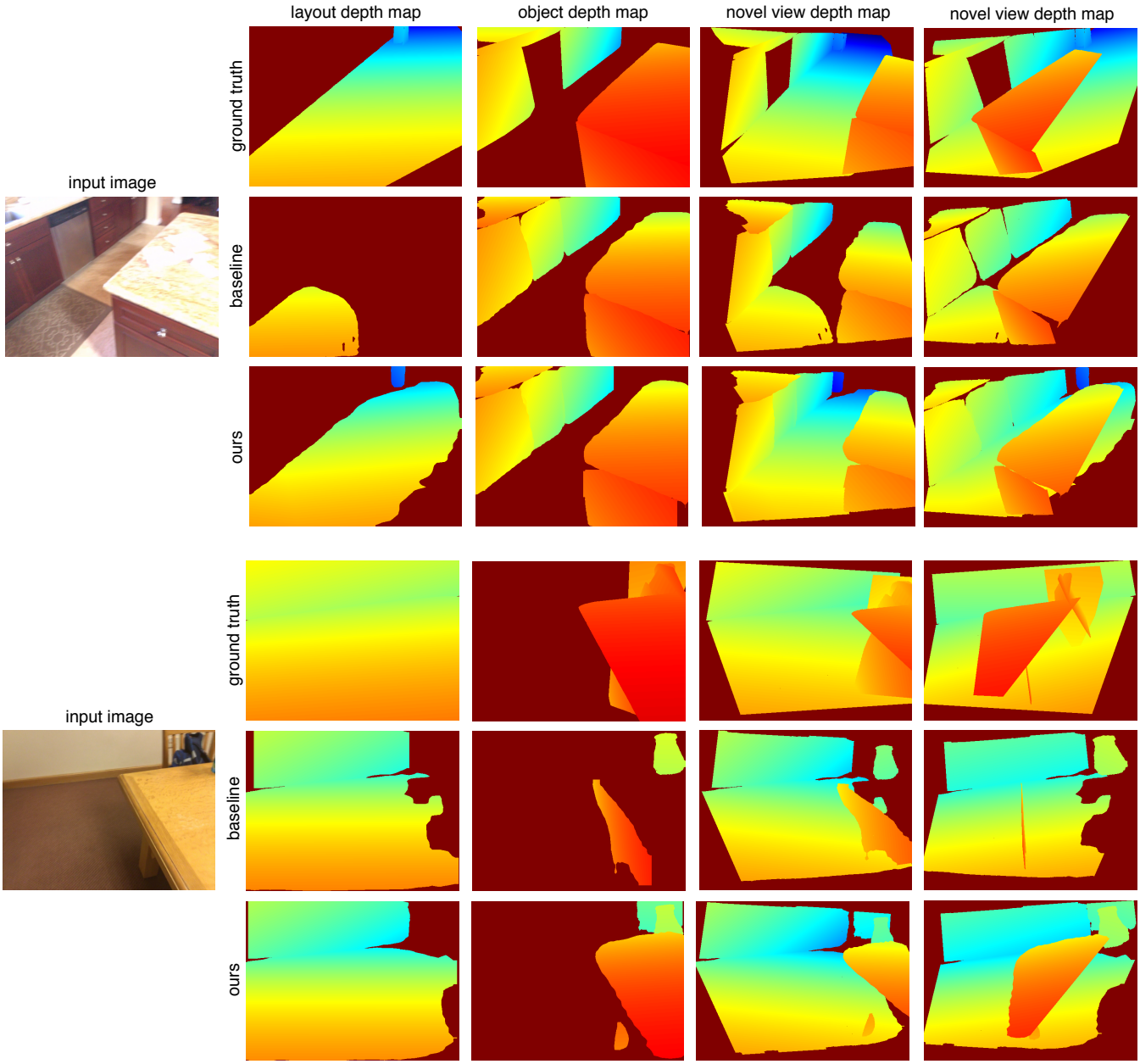

Figure 2: Qualitatively comparison of our method with PlaneRCNN-OR[1]. The first column is the input. The second column and third column are layout and object depth maps respectively. The fourth column and last column are two depth maps from novel perspective. For each sample, The top row, middle row and last row correspond to ground truth, baseline and the proposed method respectively.