Supplementary Material

We present additional evaluation results of our method, such as a visual comparison with [6, 41] on LDI prediction, ablation of the layout component, intermediate outputs, multi-layer performance, instance segmentation results, a video illustrating our performance in 3D photography, as well as dataset examples.

LDI visual comparison

Fig. 1 illustrates examples of visual comparison between our method, Dhamo *et al*. [6] and Tulsiani *et al*. [41]. We observe that the method from [41] has a rather local diminishing effect, *i.e*. around object borders. In contrast to our method, [6] simply separate the scene in a foreground and a background. For instance, one can see in the upper examples in Fig 1 that the originally occluded regions of foreground object are lost in the layered representation of [6], while in ours, one can see these parts on the second layer (oven behind furniture, sofa behind table). In the lower left example of Fig. 1, both methods perform comparably, given that the scene consists of one level of occlusion only.

Ablation of the layout component

Here, we motivate our design choices for the layout branch (Network B). For this experiment, we compare the layout predictions of our model, against the ground truth layouts. Table 1 shows that our added loss components improve the performance of the layout prediction, specially for color. In particular, the variant of our model that does not receive a depth prior, leads to considerably less accurate depth. This is an example of performance gain, due to decoupling of a hard task (*i.e*. layout depth prediction from visible color) to simpler tasks (*i.e*. standard depth prediction and RGBD inpainting).

Method	color		depth	
	MPE	RMSE	MPE	RMSE
Base, without input depth pred	21.42	42.94	0.662	1.091
Base with input depth pred	22.64	42.45	0.505	0.993
+ adversarial loss	20.93	41.47	0.495	0.953
+ perceptual loss	**19.40**	**39.89**	**0.482**	**0.919**

Table 1. **Ablation of the layout prediction (Network B) on the SunCG dataset.** Base refers to the model as introduced in the paper, where only the reconstruction loss is present $\mathcal{L}_r$. The errors are measured for color range $0 - 255$ and depth in meters.

Layout and object completion

In this paragraph we demonstrate an intermediate step of our method, which is object completion and layout prediction. Fig. 3 and 4 provide examples of the predicted mask probabilities, where opacity indicates confidence. From top to bottom, those are followed by the predictions of our network and ground truth. The two bottom rows visualize the predicted and ground truth layouts. Interestingly, the predictions tent to describe plausible object shape and texture, neglecting the color of front occluding objects. For Stanford 2D-3D, the collected ground truth contains holes, but the network learns from the available examples to regress continuous maps (specially layout).

Accuracy measurements for all layers

Here we report the error measures of all the predicted layers, for a more thorough insight on the performance of our method. Although not possible to compare against state-of-the-art approaches (restricted to two layers), we find it interesting to see the curve of accuracy as we move from layer to layer. For every layer l, whenever there is no novel content (zeros), we migrate the information from the previous layer $l-1$. Then the predicted maps are compared against the ground truth layers, *only in the areas where novel content appears*, *i.e*. ground truth dis-occlusion. This is in accordance with both the LDI representation, as well as the evaluation settings in previous works [6, 41]. Without this migration, the error values tent to be higher, as it leads to comparing the ground truth with zeros (missing information). Applied to view synthesis, these two settings lead to the same result, as a repetition of previous layers does not lead to novel content on dis-occlusion.

The results are shown in Fig. 2. The frequency plot (left) shows that almost every scene requires three or more LDI layers to be fully represented. As expected, the color and depth errors are the lowest in the first layer, where the level of uncertainty is lower. Further, the errors are roughly comparable in the middle range of layers. Interestingly, we observe a performance increase in the last layers. This is due to the increase of the contribution of the layout component in the composition of later layers. Regressing the box of the scene is an easier problem than completing objects behind occlusion, which makes the layout accuracy higher even behind occlusion. This further supports our choice to decouple the object completion from the layout prediction.

Instance segmentation

We show in Fig. 5 that our object completion inherently refines the input visible masks.

3D Photography video

We demonstrate a 3D Photography video, using our predictions. The frames are from the test set on SunCG and Stanford 2D-3D. We use inverse bilinear interpolation during the image-based rendering, to fill in the holes caused by pixel discretization of the target coordinates.

Datasets

We show in Fig. 6 and Fig. 7 an example from the automatically generated datasets.

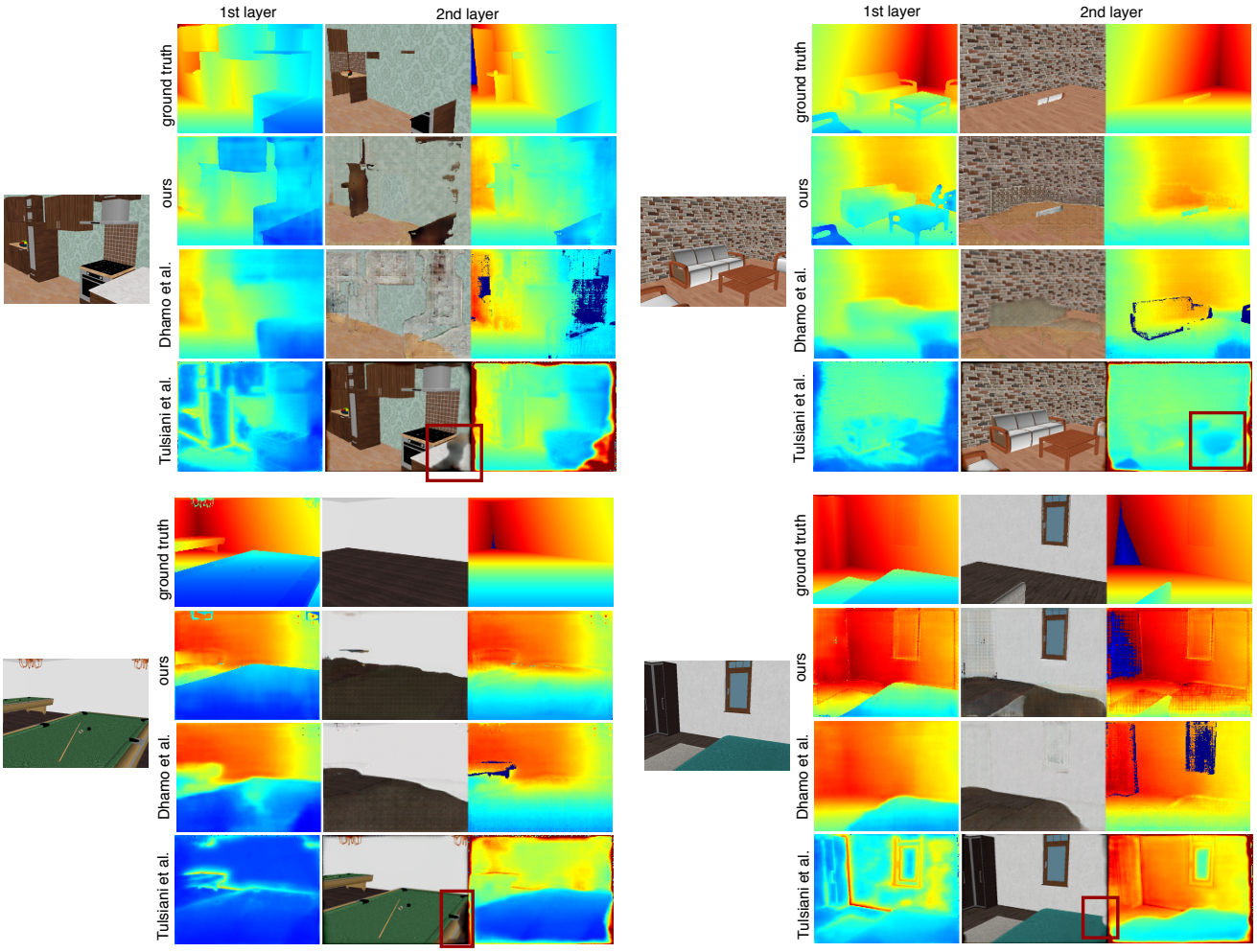

Figure 1. **LDI prediction results on SunCG.** For each example, *Left:* The input color image. *Right:* From top to bottom - ground truth, two-layer predictions of the proposed method, Dhamo *et al*. [6] and Tulsiani *et al*. [41] for the first two layers.

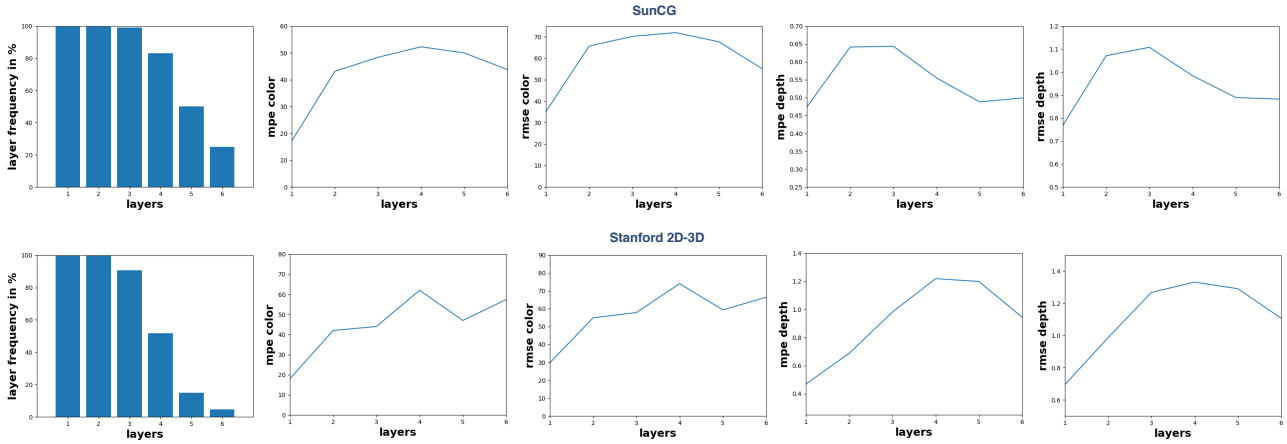

Figure 2. **Multi-layer evaluation for SunCG (top) and Stanford 2D-3D (bottom).** *Left:* The layer frequency, *i.e*. for layer l the frequency of images that have an l^{th} layer. *Center:* Color MPE and RMSE errors. *Right:* Depth MPE and RMSE errors.

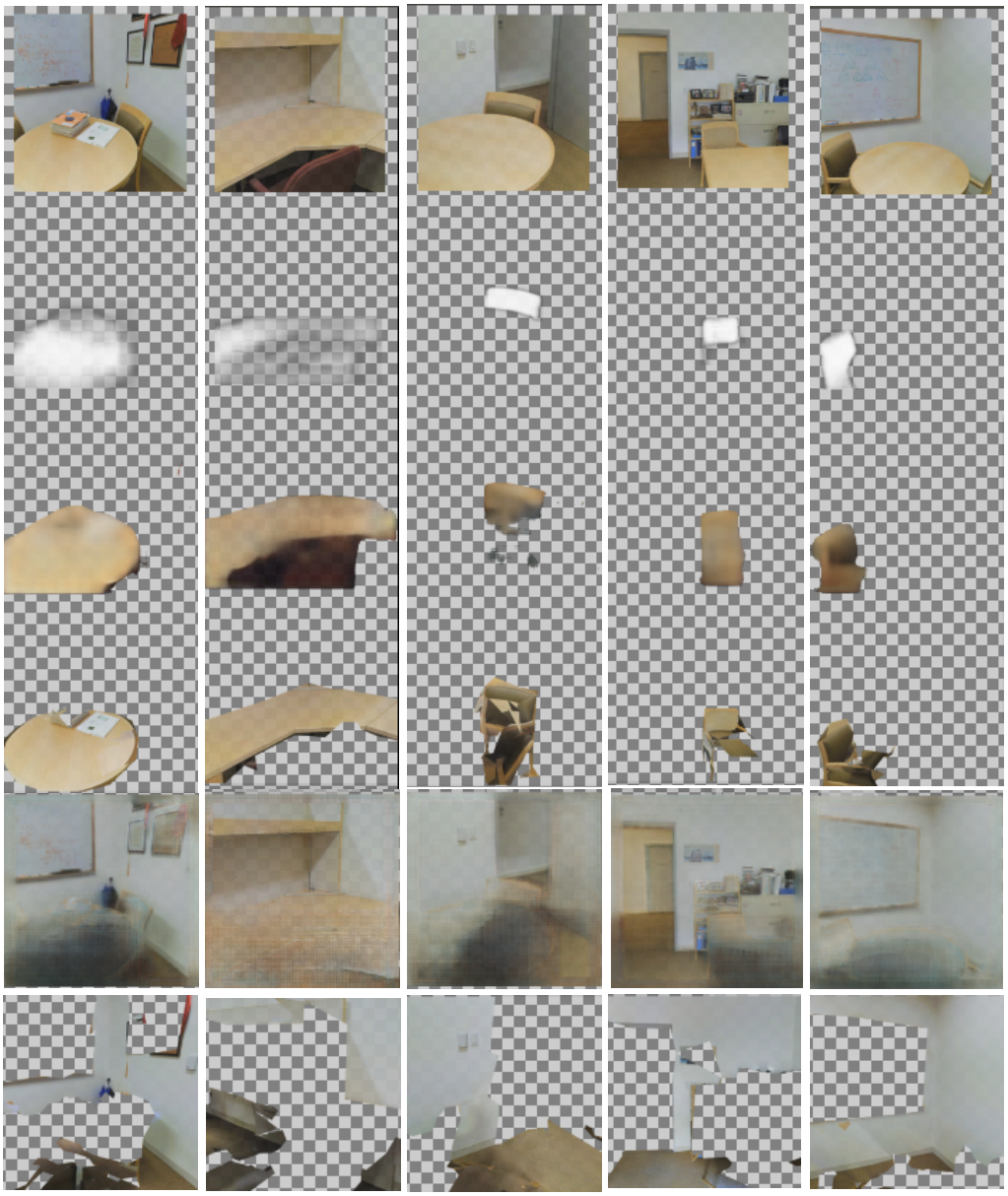

Figure 3. **RGBA object completion and layout prediction results on Stanford 2D-3D.** Input image, instance examples (top to bottom: mask, prediction, ground truth) as well as layout prediction.

Figure 4. **RGBA object completion and layout prediction results on SunCG.** Input image, instance examples (top to bottom: mask, prediction, ground truth) as well as layout prediction.

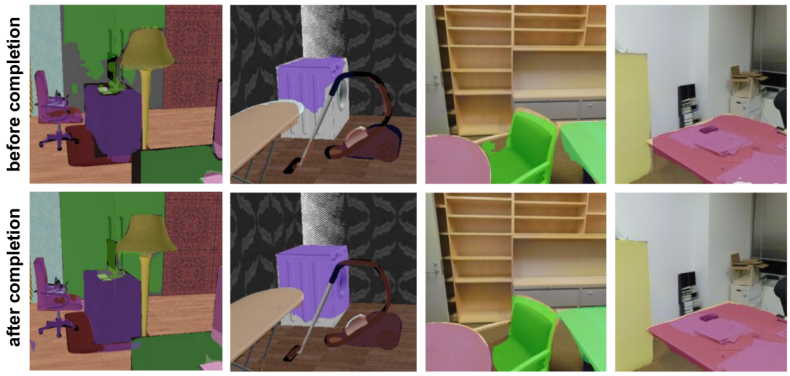

Figure 5. **Visualization of the visible masks.** *Left:* SunCG, *Right:* Stanford 2D-3D. *Top:* Instance masks as predicted from Mask R-CNN, *i.e.* input to our object completion network. *Bottom:* Instance masks using the visible parts of our predicted object extent. We observe that the object completion task inherently refines the visible masks, and aligns them better with the texture borders.

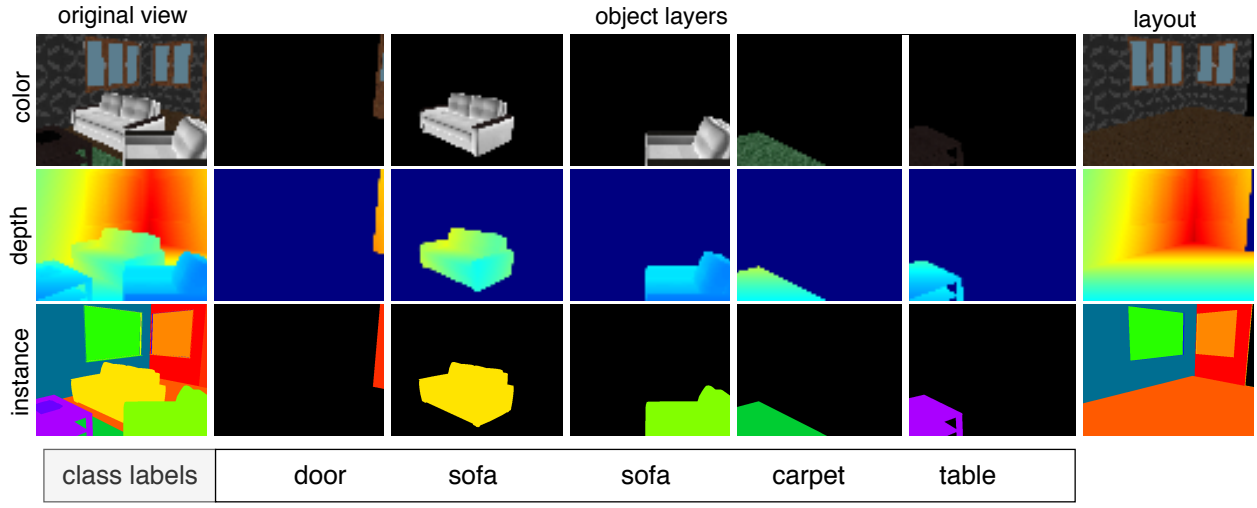

Figure 6. **Illustration of the SunCG dataset.** For every view, we provide the RGBA, depth, instance segmentation and class categories. This applies for the full-image content, object-wise layers as well as the layout. Even though the layout components are merged into a single layer, we keep track of the individual instances, as this can be exploited in future work.

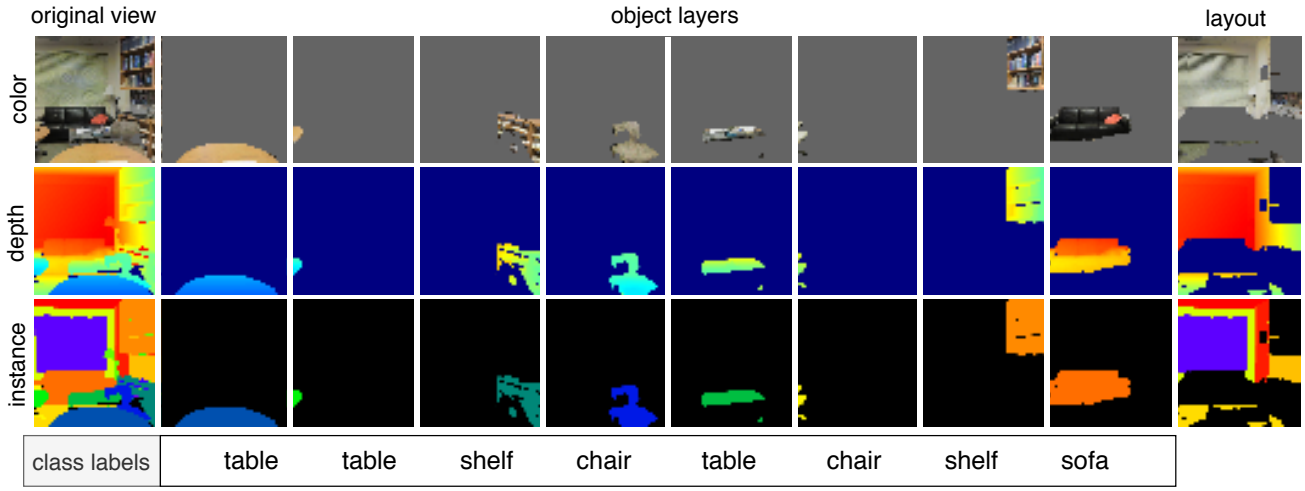

Figure 7. **Illustration of the Stanford 2D-3D dataset.** For every view, we provide the RGBA, depth, instance segmentation and class categories. This applies for the full-image content, object-wise layers as well as the layout. Even though the layout components are merged into a single layer, we keep track of the individual instances, as this can be exploited in future work.

Training details

We train Network A, B and C separately, using the Adam Optimizer with a learning rate of $1 \cdot 10^{-4}$ for Network A, $2 \cdot 10^{-3}$ for Network B and $1 \cdot 10^{-3}$ for Network C. We used a batch size of 4 (resolution 384×512) for SunCG and 8 (resolution 256×256) for Stanford 2D-3D.

Failure cases

The performance of the proposed method depends on the quality of predicted masks. For instance, if there are repetitions in the detection for a certain object, our algorithm produces two layers. Additionally, objects that are not detected might be lost from the layered representation, especially affecting scenes that contain a considerable amount of objects.