SUPPLEMENTARY MATERIAL: SEMI-SUPERVISED LEARNING OF MULTI-OBJECT 3D SCENE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

1 OVERVIEW

In this supplementary material, we present further evaluation results on various datasets. In particular, we present more qualitative results on the Clevr dataset (Johnson et al., 2017) (Fig. 1) as well as ShapeNet scenes (Chang et al., 2015) in Fig. 6, Fig. 7, and Fig. 8. Typical failure cases for ablations of our method are shown in Fig. 2. We further present more detailed results on our experiments with different object counts in Tab. 2, and Tab. 3, Tab. 4 and show qualitative results for these experiments in Fig. 4 and preliminary results on real-world images in Fig. 3. In Fig. 5 and Fig. 9, we present results for latent traversals for both Clevr and ShapeNet scenes. Renderings of novel views on ShapeNet scenes are shown in Fig. 10. We provide rotation error histograms in Fig. 11. Detailed explanations are provided in the captions of the corresponding Figures and Tables.

Moreover, we also provide more information about the network architecture, additional auxiliary loss functions and the applied parameter settings in Section 2. A more detailed listing of the reported metrics can be found in Section 3.

2 NETWORK ARCHITECTURE & PARAMETERS

Additional Loss Functions. For our experiments on the ShapeNet tabletop dataset, we use two additional loss functions:

- We favor poses which render the object visible in the image

$$L_p = \sum_i \max(-\min(x_i^p, w - x_i^p), 0) \ , \tag{1}$$

  where x_i^p is the pixel position of the object center and w is the image width.
- We penalize intersections between objects through

$$L_{int} = \sum_i \sum_{j<i} \frac{1}{K} \sum_{k=1}^{K} \max(-(\phi_i(\mathbf{x}_k) + \phi_j(\mathbf{x}_k)), 0) \ , \tag{2}$$

  where i, j are object indices, $\mathbf{x}_k$ are sample points distributed evenly between the object centers and $\phi_i(\mathbf{x}_k) := \mathbf{\Phi}\left(\mathbf{z}_{i,sh}, \mathbf{T}_c^o(\mathbf{z}_{i,ext})\mathbf{x}_k\right)$.

Network Parameters. We provide detailed information about our network architecture and parameter settings in Tab. 1.

3 EVALUATION METRICS

Instance Reconstruction. We evaluate the decomposition capability of our model by comparing the predicted object masks $\widehat{M}_{1:N}$ with the ground truth masks M_{gt}. For each object combination $(M_i, M_{gt,j})$, the IoU w.r.t. the occupied pixels is determined. We call object o_i to be a *true positive* if there is an object $o_{gt,j}$ for which $\text{IoU}(M_i, M_{gt,j}) \geq \tau$ for some threshold τ. All other predicted objects are considered as a *false positives*. Ground truth objects that were not associated with a

Table 1: **Network Parameters.** Wer report the parameter setting that was used for our experiments. *Notation:* *: same as Clevr, _: latent vector is concatenated to input of this layer (see (Park et al., 2019))

			Clevr	ShapeNet
network architecture	object encoder	Conv	$[32, 32, 64, 64]$	*
		FC	$[256, 64]$	*
	shape decoder	FC	$[64, 64, \underline{64}, 64]$	*
	color decoder	FC	$[64, 64, 64, 64]$	*
	obj. repr.	D_{sh}	8	16
		D_{tex}	7	15
training setup	# epochs		500	400
	batch size		8	*
	learning rate		0.0001	*
	loss functions, weights	λ_I	1.0	*
		λ_D	0.1	0.05
		λ_{gr}	0.01	*
		λ_{sh}	lin(0.025-0.0025; 500K)	lin(0.1-0.01; 500K)
		(λ_{inter})	-	(0.001)
		(λ_{view})	-	(0.005)
data	image size		(64, 64)	*
	dataset size		(9K/ 1K/ 2.5K)	(18K/ 2K/ 5K)
	# objects		2, 3, 4, 5	3
	position range		$[1.5, 1.5]^2$	*
	size range		$[0.625, 1.25]$	*cars:* $[1.0, 1.5]$, *chairs:* $[0.75, 1.25]$, *tabletop:* $[0.8, 1.5]$

prediction are stated to be *false negatives*. As objects might not be viewable in the image due to occlusion, we only consider masks with a minimum number of 25 occupied pixels. For an image pair $(\widehat{M}_{1:N}, M_{gt})$, we denote the total number of true positives as $TP_{0.5}(\widehat{M}_{1:N}, M_{gt})$, the number of false positives as $FP_{0.5}(\widehat{M}_{1:N}, M_{gt})$, and the number of false negatives as $TP_{0.5}(\widehat{M}_{1:N}, M_{gt})$.

From this, we compute our reported metrics as follows.

$$AP = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} AP_\tau, \qquad \text{with } \mathcal{T} = \{0.5, 0.55, ..., 0.95\} \tag{3}$$

$$AP_{0.5} = \frac{1}{\#\text{imgs}} \sum_{(\widehat{M}_{1:N}, M_{gt})} Prec_{0.5}(\widehat{M}_{1:N}, M_{gt}) \tag{4}$$

$$= \frac{1}{\#\text{imgs}} \sum_{(\widehat{M}_{1:N}, M_{gt})} \frac{TP_{0.5}(\widehat{M}_{1:N}, M_{gt})}{TP_{0.5}(\widehat{M}_{1:N}, M_{gt}) + FP_{0.5}(\widehat{M}_{1:N}, M_{gt})} \tag{5}$$

$$AR_{0.5} = \frac{1}{\#\text{imgs}} \sum_{(\widehat{M}_{1:N}, M_{gt})} Rec_{0.5}(\widehat{M}_{1:N}, M_{gt}) \tag{6}$$

$$= \frac{1}{\#\text{imgs}} \sum_{(\widehat{M}_{1:N}, M_{gt})} \frac{TP_{0.5}(\widehat{M}_{1:N}, M_{gt})}{TP_{0.5}(\widehat{M}_{1:N}, M_{gt}) + FN_{0.5}(\widehat{M}_{1:N}, M_{gt})} \tag{7}$$

$$F1_{0.5} = \frac{1}{\#\text{imgs}} \sum_{(\widehat{M}_{1:N}, M_{gt})} 2 \frac{Prec_{0.5}(\widehat{M}_{1:N}, M_{gt}) \cdot Rec_{0.5}(\widehat{M}_{1:N}, M_{gt})}{Prec_{0.5}(\widehat{M}_{1:N}, M_{gt}) + Rec_{0.5}(\widehat{M}_{1:N}, M_{gt})}. \tag{8}$$

Image Reconstruction. We use the following metrics for evaluating the reconstructed images.

$$MSE(\widehat{I}_{1:N}, I_{gt}) = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \left\| \widehat{I}_{1:N}(\mathbf{u}) - I_{gt}(\mathbf{u}) \right\|^2 \tag{9}$$

$$RMSE(\widehat{I}_{1:N}, I_{gt}) = \sqrt{MSE(\widehat{I}_{1:N}, I_{gt})} \tag{10}$$

$$PSNR(\widehat{I}_{1:N}, I_{gt}) = 10 \log_{10} \frac{L^2}{MSE(\widehat{I}_{1:N}, I_{gt})} \ , \tag{11}$$

where L is the dynamic range of allowable image pixel intensities (Wang & Bovik, 2009). We refer the reader to (Wang et al., 2004) for a detailed explanation of the SSIM metric.

We use the scikit-image implementation[1] to compute PSNR and SSIM scores.

Depth Reconstruction. Our depth reconstruction evaluation is based on (Eigen et al., 2014) and is evaluated with the following measures.

$$RMSE(\widehat{D}_{1:N}, D_{gt}) = \sqrt{\frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \left\| \widehat{D}_{1:N}(\mathbf{u}) - D_{gt}(\mathbf{u}) \right\|^2} \tag{12}$$

$$AbsRD(\widehat{D}_{1:N}, D_{gt}) = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \left| \widehat{D}_{1:N}(\mathbf{u}) - D_{gt}(\mathbf{u}) \right| / D_{gt}(\mathbf{u}) \tag{13}$$

$$SqRD(\widehat{D}_{1:N}, D_{gt}) = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \left\| \widehat{D}_{1:N}(\mathbf{u}) - D_{gt}(\mathbf{u}) \right\|^2 / D_{gt}(\mathbf{u}) \tag{14}$$

Pose Estimation. We evaluate the error on the predicted pose only for objects that were denoted as *true positive*, i.e. for which we found a valid ground truth object match. Since we are missing the association between object masks and object poses in our data, we compare each predicted object's position $\mathbf{p}_i$ to the closest ground truth object ($\mathbf{p}_{gt,j}$) according to its 3D position. Each ground truth object is assigned at most once in a greedy proceeding.

$$Err_{pos} = \frac{1}{|P|} \sum_{(\mathbf{p}_i, \mathbf{p}_{gt,j}) \in P} \sqrt{\|\mathbf{p}_i - \mathbf{p}_{gt,j}\|^2}, \qquad P = \{\text{found matches } (\mathbf{p}_i, \mathbf{p}_{gt,j})\} \tag{15}$$

$$Err_{rot} = \text{median}_{(\mathbf{p}_i, \mathbf{p}_{gt,j}) \in P} \left[\frac{360^\circ}{2\pi} \min \Big(|r_i - r_{gt,j}|, 2\pi - |r_i - r_{gt,j}| \Big) \right] \tag{16}$$

$$\text{with} \quad r_i = \text{arctan2}(z_{cos,i}, z_{sin,i})$$

[1] https://scikit-image.org/docs/dev/api/skimage.metrics.html

Table 2: **Absolute scores on the Clevr dataset (Johnson et al., 2017) for scenes with varied number of objects ($\#obj = o_{train}/o_{test}$).** Experiments are ordered w.r.t. o_{train}. We use the encoder that was trained on o_{train} objects and adapt the number of slots o_{test} to the number of objects in the test set. Models achieve slightly better results when evaluated on scenes with a lower number of objects. If tested on scenes with larger number of objects, our model is able to detect more object than it has seen during training as can be seen from the $AR_{0.5}$ and allObj score.

	Instance Reconstruction					Image Reconstruction			Depth Reconstruction			Pose Est.
	mAP ↑	$AP_{0.5}$ ↑	$AR_{0.5}$ ↑	$F1_{0.5}$ ↑	allObj ↑	RMSE ↓	PSNR ↑	SSIM ↑	RMSE ↓	AbsRD ↓	SqRD ↓	Err_{pos}
# obj=2/2	0.782	0.977	0.963	0.967	0.928	0.039	28.389	0.941	0.432	0.012	0.040	0.138
# obj=2/3	0.606	0.877	0.842	0.854	0.622	0.060	24.827	0.884	0.671	0.027	0.085	0.214
# obj=2/4	0.406	0.698	0.629	0.655	0.186	0.083	21.972	0.81	0.906	0.049	0.149	0.293
# obj=2/5	0.294	0.583	0.474	0.516	0.031	0.095	20.714	0.769	0.921	0.056	0.151	0.312
# obj=3/2	0.756	0.974	0.969	0.97	0.942	0.041	28.011	0.937	0.452	0.013	0.044	0.14
# obj=3/3	0.712	0.949	0.942	0.943	0.85	0.049	26.466	0.914	0.554	0.019	0.061	0.155
# obj=3/4	0.613	0.883	0.853	0.863	0.512	0.06	24.669	0.88	0.665	0.028	0.083	0.179
# obj=3/5	0.478	0.775	0.71	0.735	0.212	0.072	23.093	0.841	0.69	0.033	0.086	0.201
# obj=4/2	0.720	0.969	0.959	0.961	0.923	0.044	27.39	0.929	0.484	0.015	0.051	0.146
# obj=4/3	0.708	0.953	0.943	0.945	0.852	0.05	26.252	0.911	0.564	0.020	0.064	0.153
# obj=4/4	0.688	0.941	0.919	0.926	0.746	0.054	25.632	0.899	0.584	0.022	0.064	0.151
# obj=4/5	0.575	0.869	0.81	0.832	0.397	0.063	24.258	0.869	0.600	0.026	0.067	0.165
# obj=5/2	0.606	0.919	0.913	0.914	0.845	0.053	25.959	0.908	0.582	0.021	0.075	0.174
# obj=5/3	0.628	0.914	0.908	0.908	0.778	0.057	25.181	0.892	0.657	0.026	0.091	0.168
# obj=5/4	0.640	0.916	0.899	0.903	0.691	0.058	24.950	0.885	0.649	0.027	0.082	0.161
# obj=5/5	0.604	0.895	0.861	0.872	0.539	0.061	24.568	0.876	0.593	0.025	0.067	0.149

Table 3: **Relative scores on the Clevr dataset (Johnson et al., 2017) for scenes with varied number of objects.** Experiments are ordered w.r.t. o_{train} and relative scores $(obj = n/m)$ / $(obj = n/n)$ are presented. Results are based on the same experiments as in Tab. 2.

	Instance Reconstruction					Image Reconstruction			Depth Reconstruction			Pose Est.
	mAP ↑	$AP_{0.5}$ ↑	$AR_{0.5}$ ↑	$F1_{0.5}$ ↑	allObj ↑	RMSE ↓	PSNR ↑	SSIM ↑	RMSE ↓	AbsRD ↓	SqRD ↓	Err_{pos}
# obj=2/3	0.775	0.898	0.874	0.883	0.670	1.538	0.875	0.939	1.553	2.250	2.125	1.551
# obj=2/4	0.519	0.714	0.653	0.677	0.200	2.128	0.774	0.861	2.097	4.083	3.725	2.123
# obj=2/5	0.376	0.597	0.492	0.534	0.033	2.436	0.730	0.817	2.132	4.667	3.775	2.261
# obj=3/2	1.062	1.026	1.029	1.029	1.108	0.837	1.058	1.025	0.816	0.684	0.721	0.903
# obj=3/4	0.861	0.930	0.906	0.915	0.602	1.224	0.932	0.963	1.200	1.474	1.361	1.155
# obj=3/5	0.671	0.817	0.754	0.779	0.249	1.469	0.873	0.920	1.245	1.737	1.410	1.297
# obj=4/2	1.047	1.030	1.044	1.038	1.237	0.815	1.069	1.033	0.829	0.682	0.797	0.967
# obj=4/3	1.029	1.013	1.026	1.021	1.142	0.926	1.024	1.013	0.966	0.909	1.000	1.013
# obj=4/5	0.836	0.923	0.881	0.898	0.532	1.167	0.946	0.967	1.027	1.182	1.047	1.093
# obj=5/2	1.003	1.027	1.060	1.048	1.568	0.869	1.057	1.037	0.981	0.840	1.119	1.168
# obj=5/3	1.040	1.021	1.055	1.041	1.443	0.934	1.025	1.018	1.108	1.040	1.358	1.128
# obj=5/4	1.059	1.023	1.044	1.036	1.282	0.951	1.016	1.010	1.094	1.080	1.224	1.081

Table 4: **Relative scores on the Clevr dataset (Johnson et al., 2017) for scenes with varied number of objects.** Experiments are ordered w.r.t. o_{test} and relative scores $(obj = n/m)$ / $(obj = m/m)$ are presented. Results are based on the same experiments as in Tab. 2.

	Instance Reconstruction					Image Reconstruction			Depth Reconstruction			Pose Est.
	mAP ↑	$AP_{0.5}$ ↑	$AR_{0.5}$ ↑	$F1_{0.5}$ ↑	allObj ↑	RMSE ↓	PSNR ↑	SSIM ↑	RMSE ↓	AbsRD ↓	SqRD ↓	Err_{pos}
# obj=3/2	0.967	0.997	1.006	1.003	1.015	1.051	0.987	0.996	1.046	1.083	1.100	1.014
# obj=4/2	0.921	0.992	0.996	0.994	0.995	1.128	0.965	0.987	1.120	1.250	1.275	1.058
# obj=5/2	0.775	0.941	0.948	0.945	0.911	1.359	0.914	0.965	1.347	1.750	1.875	1.261
# obj=2/3	0.851	0.924	0.894	0.906	0.732	1.224	0.938	0.967	1.211	1.421	1.393	1.380
# obj=4/3	0.994	1.004	1.001	1.002	1.002	1.020	0.992	0.997	1.018	1.052	1.049	0.987
# obj=5/3	0.882	0.963	0.964	0.963	0.915	1.163	0.951	0.976	1.186	1.368	1.492	1.084
# obj=2/4	0.590	0.742	0.684	0.707	0.249	1.537	0.857	0.901	1.551	2.227	2.328	1.940
# obj=3/4	0.891	0.938	0.928	0.932	0.686	1.111	0.962	0.979	1.139	1.272	1.297	1.185
# obj=5/4	0.930	0.973	0.978	0.975	0.926	1.074	0.973	0.984	1.111	1.227	1.281	1.066
# obj=2/5	0.487	0.651	0.550	0.592	0.058	1.557	0.843	0.878	1.553	2.240	2.254	2.094
# obj=3/5	0.791	0.866	0.825	0.843	0.393	1.180	0.940	0.960	1.164	1.320	1.284	1.349
# obj=4/5	0.952	0.971	0.941	0.954	0.737	1.033	0.987	0.992	1.012	1.040	1.000	1.107

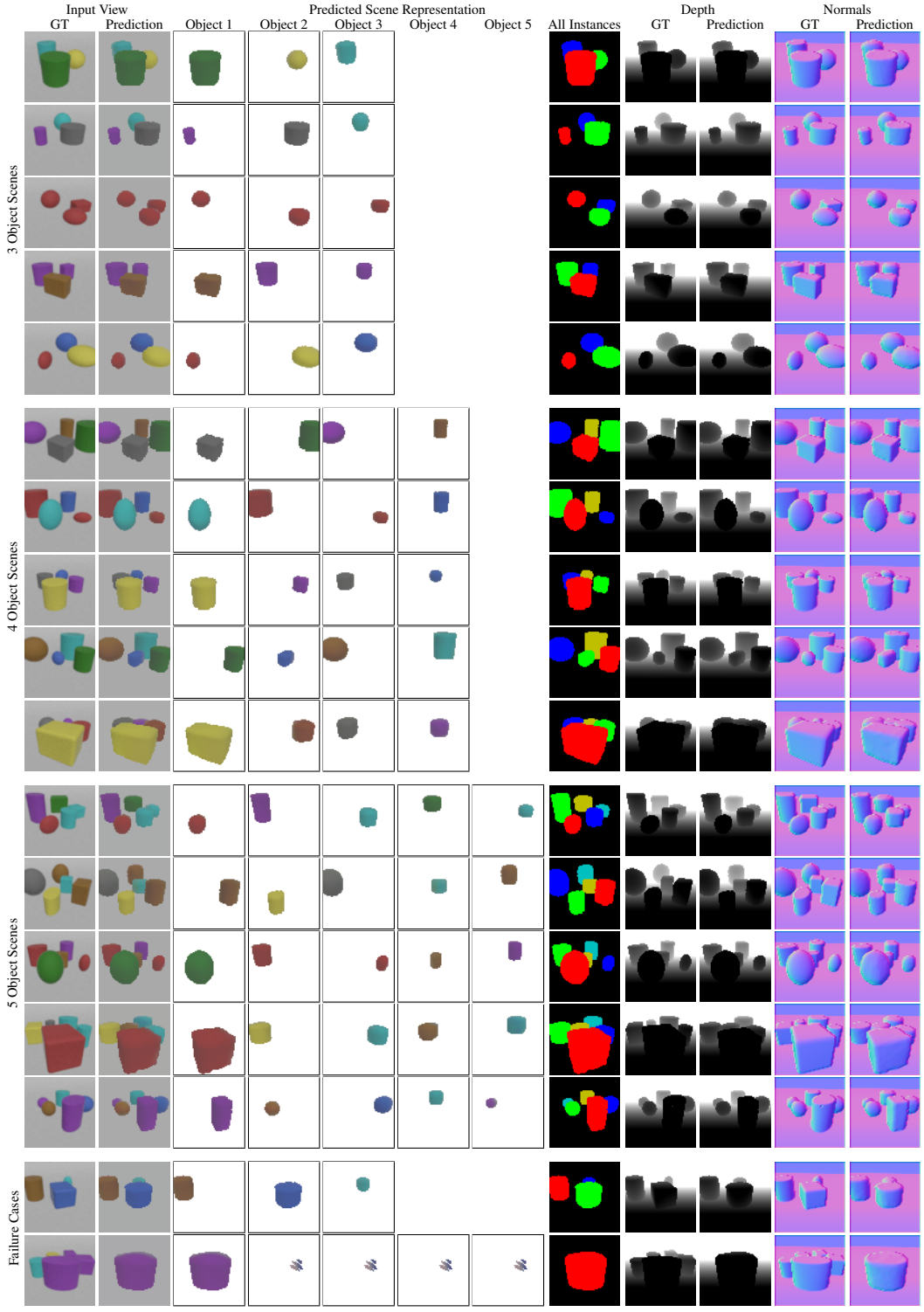

Figure 1: **Qualitative results on the Clevr dataset (Johnson et al., 2017) with three, four, and five objects.** Our model is able to decompose the scene into the individual objects. It recognizes basic color appearance and geometric properties like basic shape type and deformations (best seen in normal map). It is able to infer complete objects although some of them might be partly occluded by others in the input image. In the last two rows we also show failure cases: We found that our model sometimes misinterprets cubes as cylinders which is presumably due to the similarity of their shape and appearance at the image resolution. In few cases, it only detects a low number of objects, predominantly the most significant ones.

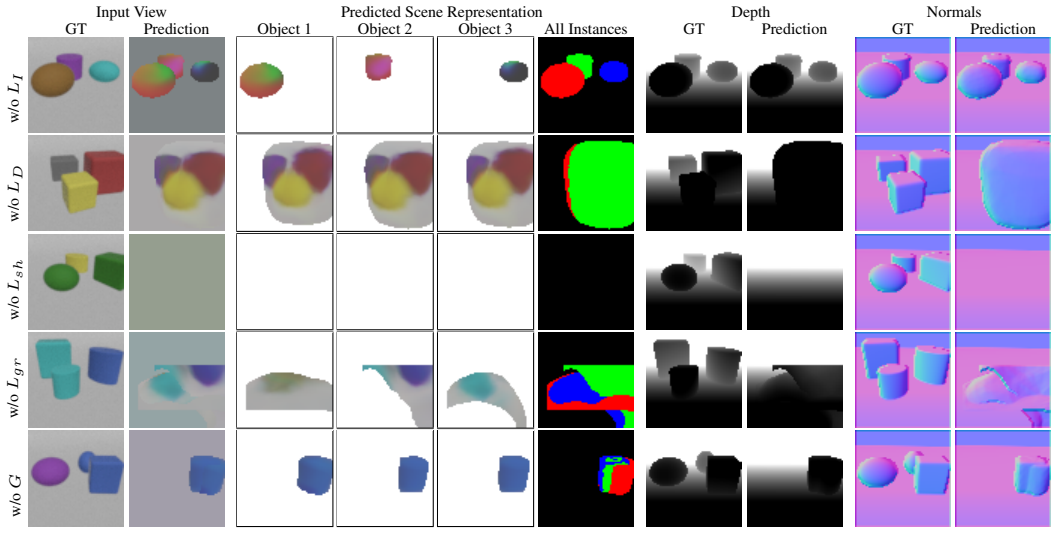

Figure 2: **Qualitative results for ablation study.** When training variants of our model without specific components, we observe typical failure cases. While a model trained without L_I is able to decompose the scene in the input image, it is obviously not able to recover the correct objects' appearance. L_D is crucial for learning the decomposition as otherwise the model can adapt the texture to obtain similar RGB reconstructions. If the shape is not regularized (L_{sh}) to match the pre-trained shape latent space, the model was not able to predict any reasonable object at all. L_{gr} helps to prevent the objects from being merged into the ground as well as to make sure that objects have a closed surface towards the ground. Without the Gaussian blur at the beginning of the training, the model often fails to detect the different objects but focuses on a single one instead.

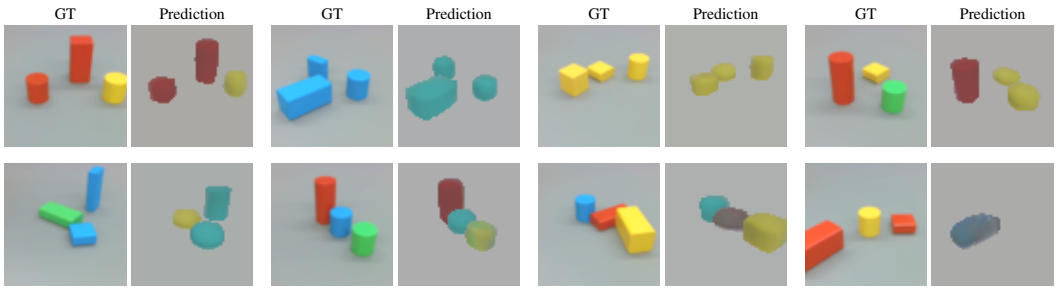

Figure 3: **Demo on real images.** We show preliminary results on real images by our model that was trained on the synthetic Clevr dataset. In some images our model can capture the coarse scene layout and shape properties of the objects. However, challenges arise due to domain, lighting, camera intrinsics and view point changes indicating interesting directions for future research.

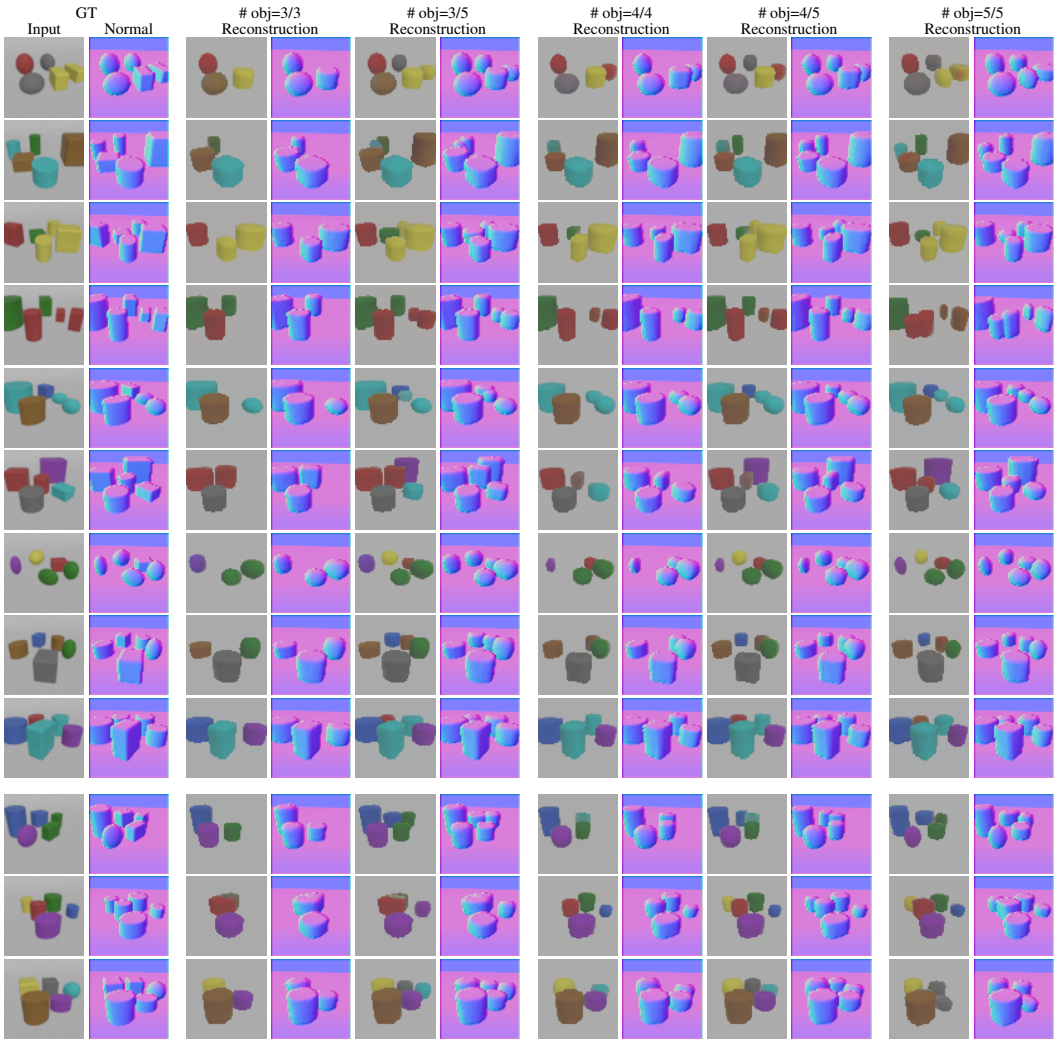

Figure 4: **Qualitative results on the Clevr dataset (Johnson et al., 2017) with varied number of objects.** As we use a shared encoder for detecting the objects in a recurrent architecture, it is possible to evaluate our model on a different number o_{test} of objects than it was trained on (o_{train}). We show reconstruction results for varying numbers $\#obj = o_{train}/o_{test}$. Remarkably, our models that were trained only on either three or four objects are able to recognize larger number of objects.

Figure 5: **Latent traversal on the Clevr dataset (Johnson et al., 2017).** We linearly adapt the first object's shape (top) or texture (middle) latent to match each of the other objects' respective representation. Moreover, we move the first object within the scene (bottom). As we reason about objects in 3D, we are able to recognize intersections between objects and exclude invalid scenes (missing images in last row). By doing so, we are able to generate new plausible scenes. Object shapes are best seen in normal maps.

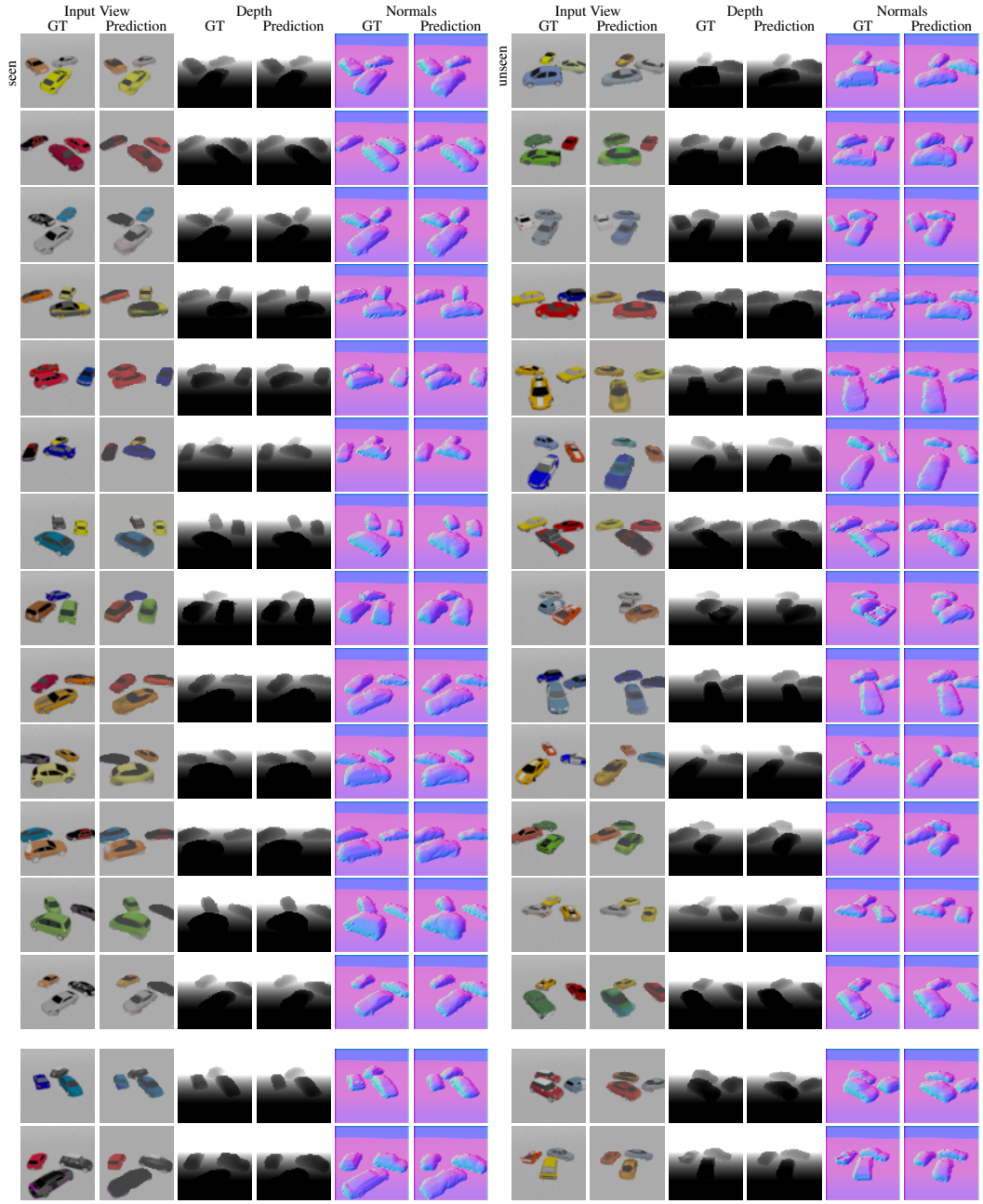

Figure 6: **Qualitative results on ShapeNet datasets (Chang et al., 2015) with car models.** Our model generates reasonable reconstruction for scenes with both seen and unseen object instances. For the latter case, it describes objects with similar shapes and textures is has seen in training. Typical failure cases are related to a pseudo-180-degree symmetry of the cars that is not distinguished by the model but handled by adapting the texture. In the lower two rows, all cars face in the wrong direction. This is in most cases not obvious from the reconstructed images only.

Figure 7: **Qualitative results on ShapeNet datasets (Chang et al., 2015) with chair models.** For the chair models it is more important to predict the correct rotation to infer a well matching shape than for other models in our datasets. The model still got easily trapped in local minima of 90-degree rotation steps where it would rather adapt shape and texture reconstruction instead of the estimated rotation. Due to the low resolution as well as the discrete sampling by the renderer, our model is prone to miss fine structural elements like armrests or thin legs.

Figure 8: **Qualitative results on tabletop scenes with ShapeNet (Chang et al., 2015) models.** For our mixed dataset, our model needs to predict object shapes from three different categories (mugs, bottles, cans) as well as respective typical size ranges. We found that our model is able to distinguish between the objects based on their typical characteristics. Unseen objects in the second test set are typically replaced by known objects from the training set which are similar in appearance. Handles of cups as well as thin, long bottlenecks are often neglected by the model. Especially for small objects, the model sometimes misses to reconstruct an object in the scene. The last row shows reconstructions from a failed training run in which only one object can be found.

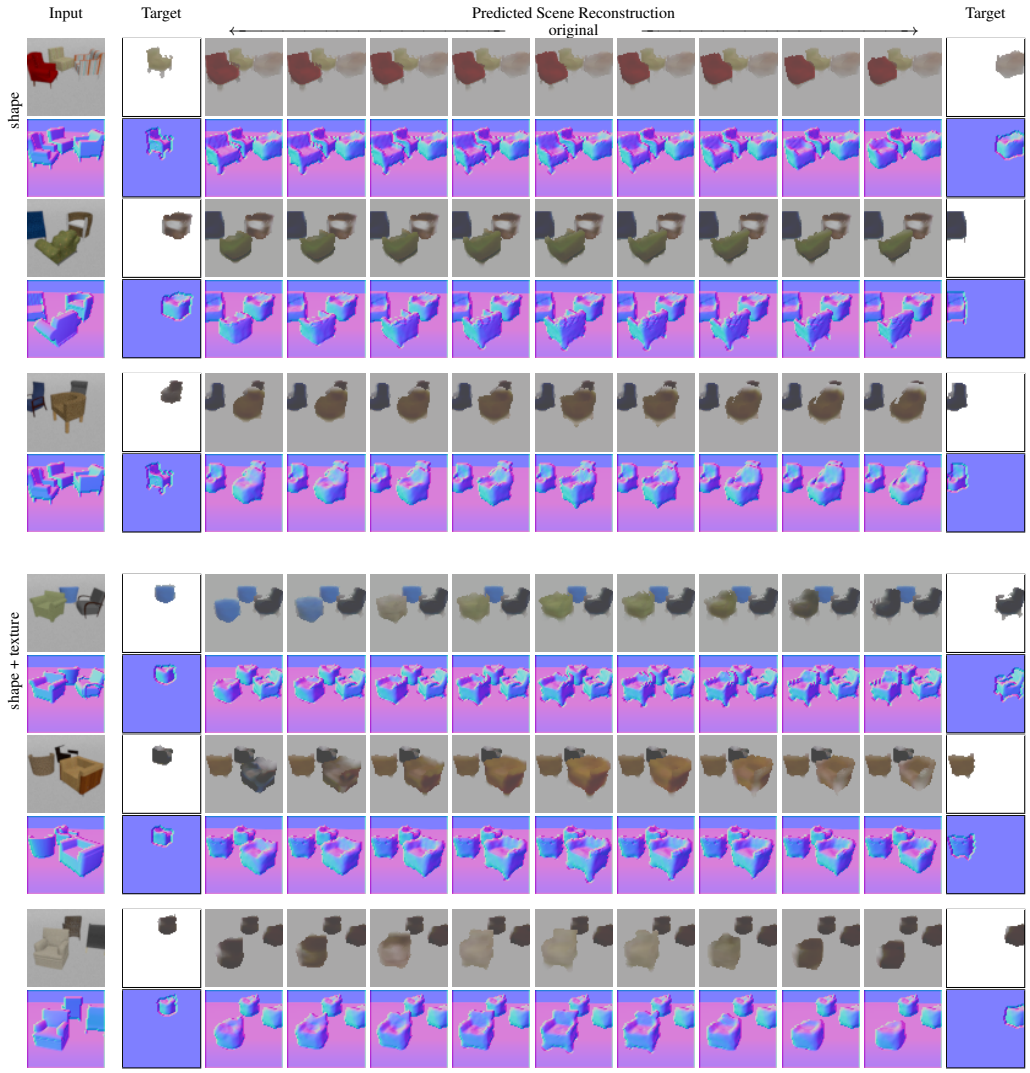

Figure 9: **Latent traversal on ShapeNet datasets (Chang et al., 2015) with chair models.** We linearly adapt the first object's latent to match each of the other objects' respective representation in either shape alone (top rows) or shape and texture (bottom rows). By this, we are able to generate new plausible scenes. Object shapes are best seen in the normal maps.

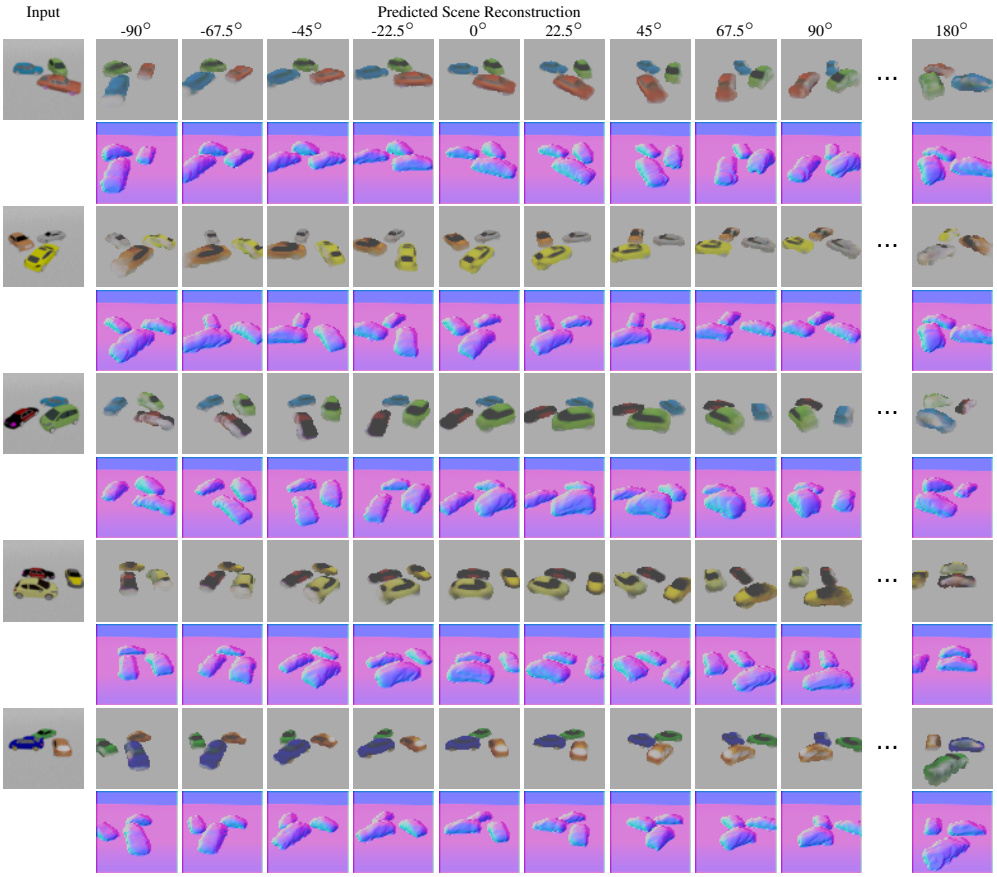

Figure 10: **Rendering of novel views.** As our model reasons about the underlying 3D structure of a given image, it is able to render novel views of a scene. This is possible although our model was trained exclusivity from single images. The reconstructed normal maps show that the model learned to reason about the depicted objects in 3D space. It can be observed that our model renders the reverse side of the car objects less accurate than the visible parts. This might be due to limited range in rotation that the model infers due to pseudo-symmetry.

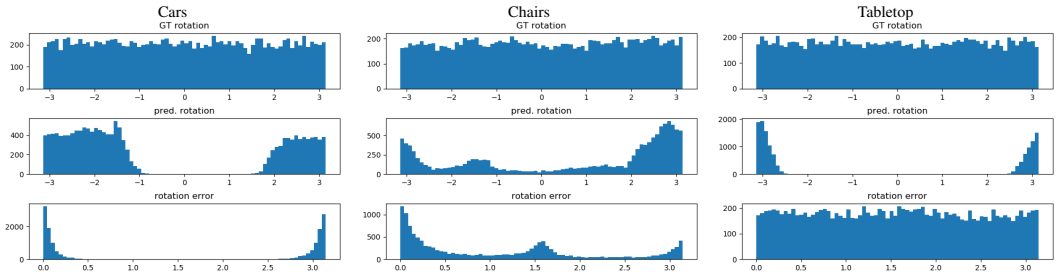

Figure 11: **Rotation Prediction on ShapeNet dataset (Chang et al., 2015).** From top to bottom: GT and predicted rotation angles for each dataset and resulting rotation angles. While values for GT rotation are naturally uniformly distributed over the entire range of $[-\pi, \pi]$ for all scenes, we found that predicted rotation estimates can be spread over a smaller sub-range. Peaks in the histogram for cars ($\sim \pi$) and chairs ($\sim \frac{\pi}{2}, \sim \pi$) indicate that the model got stuck in local minima where it predicts a rotation up to a pseudo-symmetry. In contrast, it predicts rotation almost uniformly for the tabletop scenes due to the rotational symmetry of the shapes and the capability of adapting the texture.

REFERENCES

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2366–2374, Cambridge, MA, USA, 2014. MIT Press.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.