PatchNets: Patch-Based Generalizable Deep Implicit 3D Shape Representations
— Supplementary Material —

Edgar Tretschk[1] Ayush Tewari[1] Vladislav Golyanik[1]
Michael Zollhöfer[2] Carsten Stoll[2] Christian Theobalt[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] Facebook Reality Labs

In this supplemental material, we expand on some points from the main paper. We first perform an ablation study on the extrinsics losses in Sec. 1. In Sec. 2, we describe the error measures we employ. In Sec. 3, we compare error measures on the reduced and full test sets. Sec. 4 shows the randomly picked single shape that we use in one of the generalization experiments. Sec. 5 contains more experiments using object-level priors. Sec. 6 shows different number of patches and network/latent code sizes. Next, we measure the performance under synthetic noise in Sec. 7. We show preliminary results on a large scene in Sec. 8. Finally, in Sec. 9, we provide some remarks on the concurrent work DSIF [1].

1 Loss Ablation Study

We run an ablation study of each of the extrinsics losses. We also test whether guiding the rotation via initialization and a loss function helps. Table 1 contains the results. Due to our initialization, as described in Sec. 3.2, the extrinsics losses are not necessary in this setting. However, as shown in Sec. 5 in this supplementary material, they are necessary when the extrinsics are regressed instead of free. Initializing and encouraging the rotation towards normal alignment helps. We do not use $\mathcal{L}_{\text{recon}}$ on the mixture because that modification does not sufficiently constrain the patches to individually reconstruct the surface, as Fig. 1 shows.

2 Error Metrics

Similar to Genova *et al.* [1], we evaluate using IoU, Chamfer distance and F-score. We report the mean values across different test sets.

IoU: For a given watertight groundtruth mesh, we extract the reconstructed mesh using marching cubes at 128^3 resolution. We then sample $100k$ points uniformly in the bounding box of the GT and check for both the generated mesh and the GT whether each point is inside or outside. The final value is the fraction of intersection over union, multiplied by a factor of 100. Higher is better.

Chamfer Distance: Here, we sample 100k points on the surface of both the groundtruth and the reconstructed mesh. We use a kD-tree to compute the

Table 1. Ablation Study of PatchNet. We remove each of the extrinsics losses. We also impose the reconstruction loss on the mixture (using $g_i(\mathbf{x})$ from Eq. 11 instead of $f(\mathbf{x}, \mathbf{z}_{i,p}, \theta)$).

	IoU	Chamfer	F-score
no $\mathcal{L}_{\text{sur}}$	92.0	0.049	94.8
no $\mathcal{L}_{\text{cov}}$	90.7	0.051	93.6
no $\mathcal{L}_{\text{rot}}$	92.5	0.043	95.4
no $\mathcal{L}_{\text{scl}}$	91.2	0.031	94.3
no $\mathcal{L}_{\text{var}}$	91.6	0.045	94.4
random rotation initialization and no $\mathcal{L}_{\text{rot}}$	89.0	0.048	93.1
ours	91.6	0.045	94.5
ours with $\mathcal{L}_{\text{recon}}$ on mixture	94.0	0.026	96.8

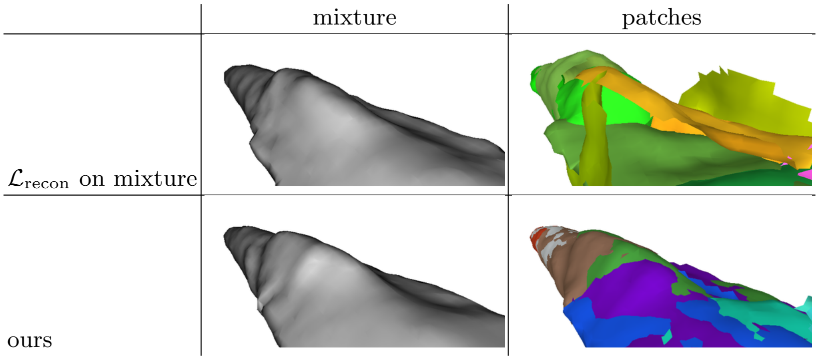


Fig. 1. Mixture Reconstruction Loss. Imposing the reconstruction loss on the mixture instead of directly on the patches leads to individual patches not matching the surface.

closest points from the reconstructed to the groundtruth mesh and vice-versa. We then square these distances ($L2$ Chamfer) and sum the averages of each direction. For better readability, we finally multiply by 100. Lower is better.

F-score: For each shape, we threshold the point-wise distances computed before at 0.01 (all meshes are normalized to a unit cube). We then compute the fraction of distances below the threshold, separately for each direction. Finally, we take the harmonic mean of both these values and multiply the result by 100. Higher is better.

In cases where a network does not produce any surface, we set the value of IoU to 0, the Chamfer distance to 100, and the F-score to 0.

3 Reduced Test Set

Our reduced test set on ShapeNet consists of 50 randomly chosen test shapes per category. Table 2 shows how well the error measures on this reduced test set approximate the error measures on the full test set.

Table 2. Reduced Test Set vs. Full Test Set. The computed metrics on the reduced test set of ShapeNet are a good approximation of the computed metrics on the full test set. This is an extended version of Table 1 from the main paper.

	IoU						Chamfer						F-score					
Category	DeepSDF		Baseline		Ours		DeepSDF		Baseline		Ours		DeepSDF		Baseline		Ours	
	full	red.	full	red.	full	red.	full	red.	full	red.	full	red.	full	red.	full	red.	full	red.
airplane	84.9	84.0	65.3	64.2	91.1	90.7	0.012	0.023	0.077	0.084	0.004	0.006	93.0	92.3	72.9	71.6	97.8	97.5
bench	78.3	77.1	68.0	65.7	85.4	83.7	0.021	0.015	0.065	0.043	0.006	0.006	91.2	90.4	80.6	80.1	95.7	94.9
cabinet	92.2	89.1	88.8	84.8	92.9	91.6	0.033	0.027	0.055	0.047	0.110	0.119	91.6	90.3	86.4	84.3	91.2	91.8
car	87.9	88.4	83.6	84.3	91.7	92.6	0.049	0.057	0.070	0.074	0.049	0.050	82.2	82.1	74.5	74.4	87.7	87.8
chair	81.8	80.1	72.9	70.3	90.0	88.6	0.042	0.041	0.110	0.118	0.018	0.013	86.6	86.0	75.5	74.8	94.3	93.5
display	91.6	92.9	86.5	89.1	95.2	95.5	0.030	0.010	0.061	0.034	0.039	0.049	93.7	95.1	87.0	89.8	97.0	97.3
lamp	74.9	72.3	63.0	63.4	89.6	88.0	0.566	2.121	0.438	0.257	0.055	0.063	82.5	79.9	69.4	70.1	94.9	94.0
rifle	79.0	78.0	68.5	66.0	93.3	93.1	0.013	0.012	0.039	0.046	0.002	0.001	90.9	90.7	82.3	80.4	99.3	99.3
sofa	92.5	92.2	85.4	84.5	95.0	95.1	0.054	0.075	0.226	0.236	0.014	0.012	92.1	91.3	84.2	83.0	95.3	95.3
speaker	91.9	90.5	86.7	84.9	92.7	90.8	0.050	0.060	0.094	0.121	0.243	0.242	87.6	84.7	79.4	75.7	88.5	85.1
table	84.2	83.4	71.9	69.5	89.4	90.3	0.074	0.043	0.156	0.169	0.018	0.017	91.1	91.5	79.2	79.1	95.0	96.1
telephone	96.2	96.0	95.0	94.1	98.1	98.0	0.008	0.010	0.016	0.016	0.003	0.004	97.7	97.3	96.2	94.7	99.4	99.3
watercraft	85.2	84.9	79.1	78.5	93.2	93.1	0.026	0.019	0.041	0.031	0.009	0.006	87.8	88.2	90.2	80.6	96.4	96.6
mean	86.2	85.3	78.1	76.9	92.1	91.6	0.075	0.193	0.111	0.098	0.044	0.045	89.9	89.2	80.6	79.9	94.8	94.5

4 Generalization Experiments – Single Shape

Fig. 2 shows the randomly picked single shape on which we trained PatchNet in Sec. 4.2 of the main paper.

5 Object-Level Priors

5.1 Surface Reconstruction

We report surface reconstruction errors using object-level priors (see Sec. 4.3 from the main paper). Note that the experiments in Sec. 4.3 of the main paper

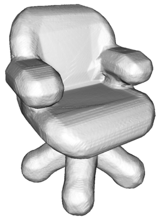


Fig. 2. Single Shape. In one of the generalization experiments in Sec. 4.2 of the main paper, we train PatchNet on this randomly chosen groundtruth shape.

use the *most* competitive setting of the global-patch baseline (*i.e.*, pretrained on all categories and then refined) and the *least* competitve setting of PatchNet (*i.e.*, pretrained on one category and not refined). This demonstrates how well our proposed PatchNet generalizes. For consistency, for the DeepSDF-based baseline, we choose the same setting as for the global-patch baseline. Note that that setting is virtually on par with the most competitive DeepSDF setting (*i.e.*, pretrained on one category and then refined).

Settings Both our network and the baselines consist of a four-layer ObjectNet and the standard final eight FC layers. We pretrain the final eight FC layers either on the reduced training set of all categories or on all shapes from the *Cabinets* category training set. We then either keep those pretrained weights fixed while training ObjectNet or we allow them to be refined. While at training time, each phase lasts 1000 epochs, we reduce this to 800 epochs at test time.

Results Table 3 contains the quantitative results. The baselines do not generalize well if they are kept fixed. Refinement improves error measures.

Table 3. Surface Reconstruction with ObjectNet. We pretrain the final eight layers either on one category (*one*) or on all categories (*all*). We then either keep those layers fixed (*fix.*) or refine them (*ref.*).

		baseline				DeepSDF-based				ours			
		one		all		one		all		one		all	
		fix.	ref.	fix.	ref.	fix.	ref.	fix.	ref.	fix.	ref.	fix.	ref.
airplanes	IoU	35.9	70.9	60.2	73.3	47.0	75.6	69.9	74.1	67.5	68.5	71.9	74.2
	Chamfer	0.710	0.146	0.218	0.147	0.546	0.049	0.127	0.050	0.203	0.182	0.179	0.170
	F-score	37.5	76.0	63.6	78.3	49.1	82.5	76.4	81.6	71.7	74.1	77.9	79.7
sofas	IoU	76.1	81.8	76.3	84.3	76.4	79.7	82.4	76.6	85.3	86.2	84.9	86.0
	Chamfer	0.416	0.159	0.398	0.171	0.467	0.178	0.282	0.406	0.118	0.139	0.236	0.082
	F-score	69.0	75.2	71.8	77.9	70.1	72.3	77.5	71.8	79.0	80.7	79.5	79.9

5.2 Ablation Study

We evaluate the extrinsics losses in the context of surface reconstruction with object-level priors. We use the version of our method from the main paper: pretrained on the *Cabinets* category and without refinement. We perform the ablation study on the *Sofas* category.

The quantitative results are in Table 4. The network failed to reconstruct without $\mathcal{L}_{\text{cov}}$.

Table 4. Ablation Study with Object-level Priors. We remove each of the extrinsics losses.

	IoU	Chamfer	F-score
no $\mathcal{L}_{\text{sur}}$	87.6	0.076	82.6
no $\mathcal{L}_{\text{scl}}$	75.5	0.154	54.2
no $\mathcal{L}_{\text{var}}$	71.8	0.269	47.3
ours	85.3	0.118	79.0
ours with $\mathcal{L}_{\text{recon}}$ on mixture	84.9	0.116	78.1

5.3 Partial Point Cloud Completion

We report additional depth-map completion results using the same settings for our method that we use for the baselines in the main paper (pretrained on all categories and refined). Note that in the main paper, we report the shape-completion results of the most disadvantageous version of our method (according to Table 3). Table 5 contains the quantitative results. In all cases, our method after local refinement yields the best results.

Table 5. Partial Point Cloud Completion from Depth Maps. We complete depth maps from a fixed camera viewpoint and from per-scene random viewpoints.

	sofas fixed		sofas random		airplanes fixed		airplanes random	
	acc.	F-score	acc.	F-score	acc.	F-score	acc.	F-score
baseline	0.094	43.0	0.092	42.7	0.069	58.1	0.066	58.7
DeepSDF-based baseline	0.106	33.6	0.101	39.5	0.066	56.9	0.065	55.5
ours (main paper)	0.091	48.1	0.077	49.2	0.058	60.5	0.056	59.4
ours+refined (main paper)	**0.052**	53.6	**0.053**	52.4	**0.041**	67.7	**0.043**	65.8
ours (baseline-matched)	0.088	47.5	0.074	50.0	0.052	64.8	0.050	64.3
ours+refined (baseline-matched)	0.061	**54.7**	0.056	**53.5**	0.045	**70.3**	0.044	**69.9**

6 Number of Patches and Network/Latent Code Sizes

Fig. 3 shows the mean error metrics on the reduced ShapeNet test set when training on the reduced ShapeNet training set. We try out different sizes. Size refers to both the dimensions of the patch latent vector and the hidden dimensions of PatchNet, as in Sec. 4.2. The gap between size 128 and 512 is much smaller than between 32 and 128. Furthermore, using 100 patches instead of 30 yields only marginal gains at best.

Fig. 4 shows the per-category error metric on the reduced ShapeNet test set when training on the reduced ShapeNet training set. We conduct this experiment with different numbers of patches. Apart from the outlier categories *cabinet*, *car*, and *speaker*, we observe that the error metrics behave very similar across categories. They improve strongly when going from 3 to 10 and from 10 to 30 patches and they improve at most slightly when going from 30 to 100 patches.

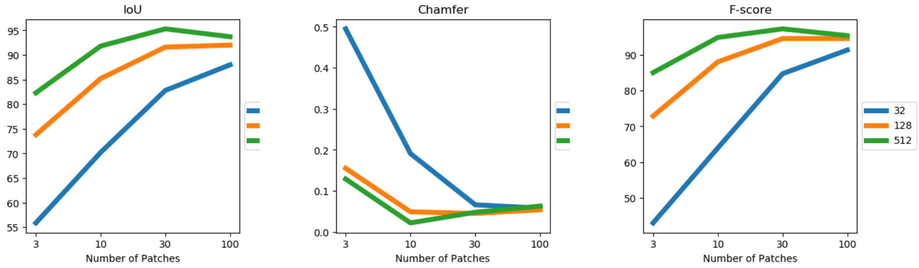


Fig. 3. Mean error metrics on the reduced ShapeNet test set for different numbers of patches and network/latent code sizes.

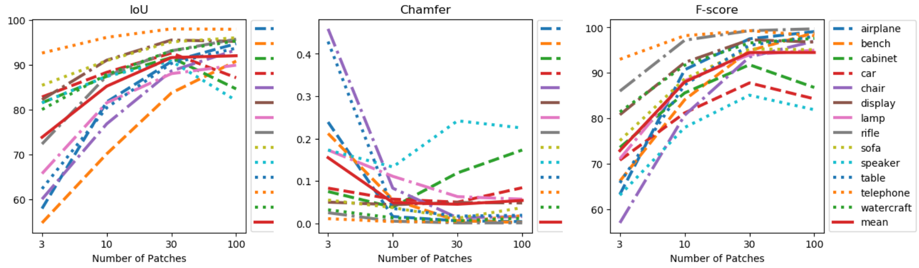


Fig. 4. Per-category error metrics on the reduced ShapeNet test set for different numbers of patches.

7 Synthetic Noise

We investigate the robustness of PatchNet by adding Gaussian noise to the groundtruth SDF values of the reduced test set. We use the PatchNet trained with default settings, which also means that it has only seen unperturbed SDF data during training. The Gaussian noise has zero mean and different standard deviations σ. For reference, the mesh fits tightly into the unit sphere, as mentioned in Sec. 3.2. The results are in Table 6.

Table 6. Synthetic Noise at Test Time.

	IoU	Chamfer	F-score
$\sigma = 0.1$	81.2	0.037	85.3
$\sigma = 0.01$	90.3	0.045	94.3
$\sigma = 0.001$	91.5	0.047	94.4
$\sigma = 0$ (ours)	91.6	0.045	94.5

8 Preliminary Results on ICL-NUIM

Once trained, a PatchNet can be used for any number of patches at test time. Here, we present some preliminary results on the large living room from ICL-NUIM [3].

Since the scene is already watertight, we skip the depth fusion step of the preprocessing method. We reduce the standard deviation used to generate SDF samples by a factor of 100 to account for scaling differences. Overall, we sample 50 million SDF samples.

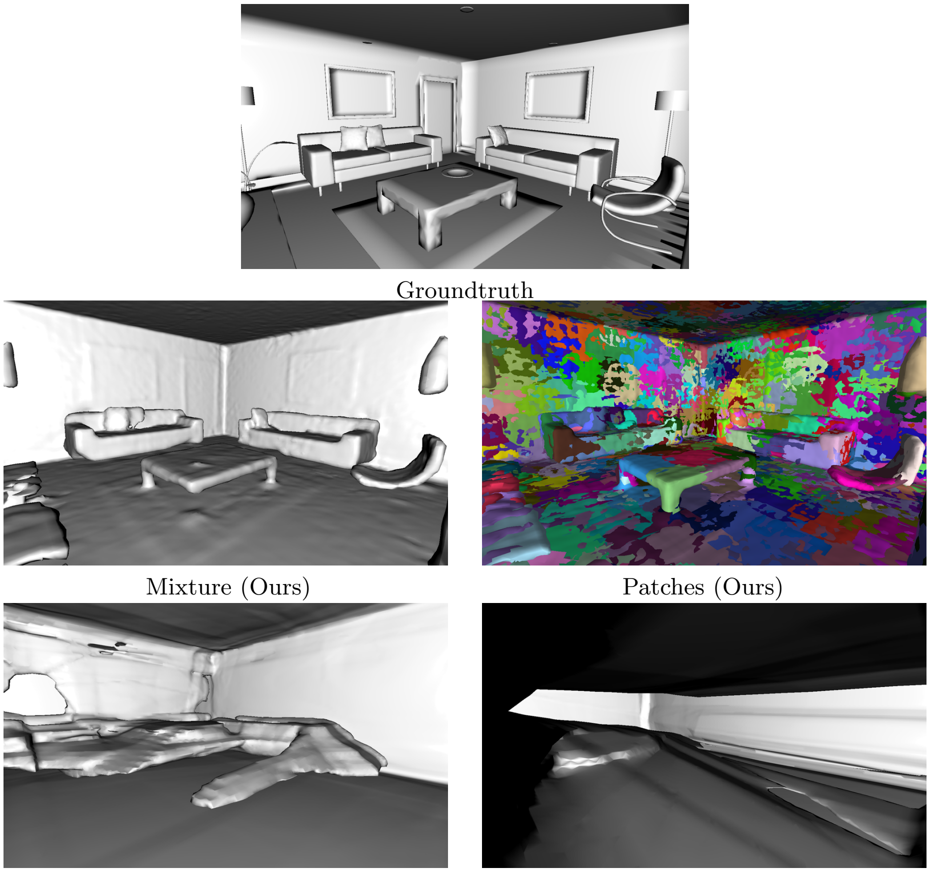
For PatchNet, we use 800 patches. We keep the extrinsics fixed at their initial values since we found that to improve the reconstruction. We optimize for 10k iterations, halving the learning rate every 2k iterations. During optimization, 25k SDF samples are used per iteration. The baselines are trained with the same modified settings.

The results are in Fig. 5. Note that due to our extrinsics initialization (Sec. 3.2) and $\mathcal{L}_{\text{var}}$, all patches have similar sizes, which leads to a wasteful distribution.

9 Remarks on the Concurrent Work DSIF [1]

For completeness, we provide some remarks on the unpublished, but concurrent related work *Deep Structured Implicit Functions (DSIF)* by Genova *et al.* [1][1]. In our terminology, they use a network from prior work (SIF [2]) to regress patch extrinsics from depth maps of 20 fixed viewpoints. They then use a point-set encoder to regress patch latent codes from backprojected depth maps according to the regressed extrinsics. Finally, they propose a modified version of OccupancyNetworks [4] to regress point-wise occupancy probabilities.

[1] After submission of this work, DSIF was published at CVPR 2020 and renamed to *Local Deep Implicit Functions for 3D Shape.*



Groundtruth

Mixture (Ours)

Patches (Ours)

DeepSDF

Our Global Baseline

Fig. 5. Preliminary Results on ICL-NUIM.

As Table 2 in the main paper shows, our proposed method outperforms theirs almost everywhere despite being trained on only $\sim 4\%$ of the training data. Since they impose their reconstruction loss on the final mixture, we do the same for a comparison in Table 1 in this supplementary material. Using 32 patches and $N_z = 128$, their method obtains an F-score below 95 (on the full test set), while our method reaches 96.8 (on the reduced test set; which is very representative of the full test set, see Sec. 3).

Furthermore, they regress the patch extrinsics with a network taken from prior work [2], while we show that it is possible to directly and effectively initialize them. Because DSIF regresses extrinsics, it can have issues predicting extrinsics for shapes very different from the training data, while we *by construction* do not have such issues. It also turns out that the isotropic Gaussian weights we use in our proposed method are sufficient to outperform their method, which uses more complicated anisotropic Gaussians. Finally, for their encoder to work, the input geometry needs to be represented in some way (which is a non-trivial decision that might impact performance), while we avoid this issue by auto-decoding.

References

1. Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3d shape. In: Computer Vision and Pattern Recognition (CVPR) (2020)
2. Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W.T., Funkhouser, T.: Learning shape templates with structured implicit functions. In: International Conference on Computer Vision (ICCV) (2019)
3. Handa, A., Whelan, T., McDonald, J., Davison, A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: International Conference on Robotics and Automation (ICRA) (2014)
4. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3D reconstruction in function space. In: Computer Vision and Pattern Recognition (CVPR) (2019)