Fostering Generalization in Single-view 3D Reconstruction by Learning a Hierarchy of Local and Global Shape Priors

1. Qualitative Examples

Here, we present more qualitative results to illustrate different behaviours of ONet[2] and our Hierarchical Prior Network (HPN). All shapes created by HPN are much closer to the ground truth shape. For each of the four training settings from Tbl. 1 we show one figure. Fig. 1 for *airplane*, Fig. 2 for *lamp*, Fig. 3 for *chair* and Fig. 4 for *multi-class*. Each figure contains twelve examples. For each sample we show the input image (first row), reconstructions for ONet (second row), reconstructions for Ours (third row) and the ground truth (fourth row) in two views. The figures are best viewed in PDF, such that one can zoom in.

2. Quantitative

In Tbl. 1, we present the full version of Tbl. 1 from the main paper. This contains the full set of evaluation classes (columns). Additionally, we provide the mean for all categories seen during training in the first column. Furthermore, we also report results for networks trained in the *single-class* airplane setting. Best scores are marked in **bold**. Our method is always best in the generalization settings (black, orange and green numbers) and best for one of the training classes (blue numbers). In Tbl. 2 we report the IoU values for completeness.

Figure 1. Qualitative results for networks trained on the *airplane* class.

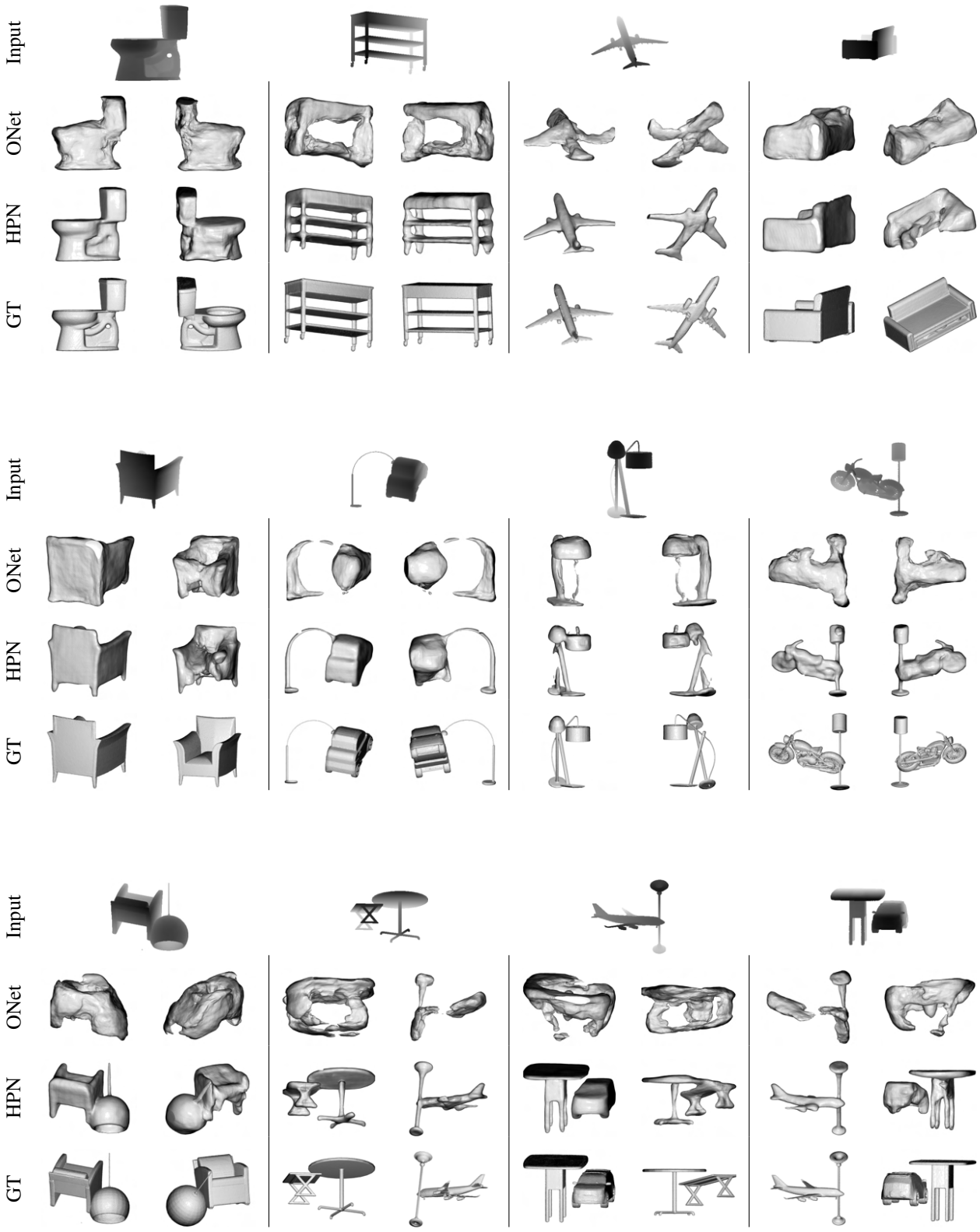

Figure 2. Qualitative results for networks trained on the *lamp* class.

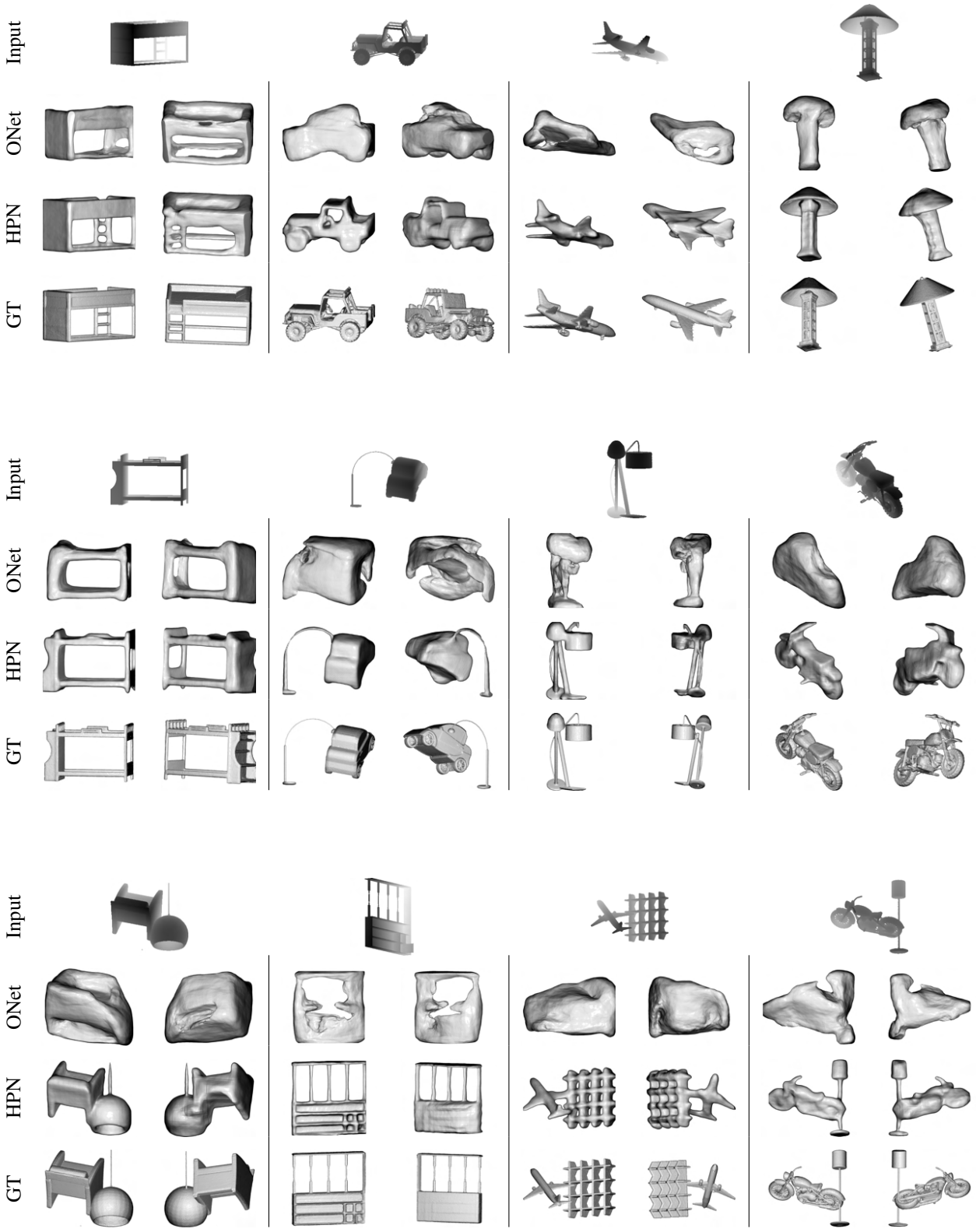

Figure 3. Qualitative results for networks trained on the *chair* class.

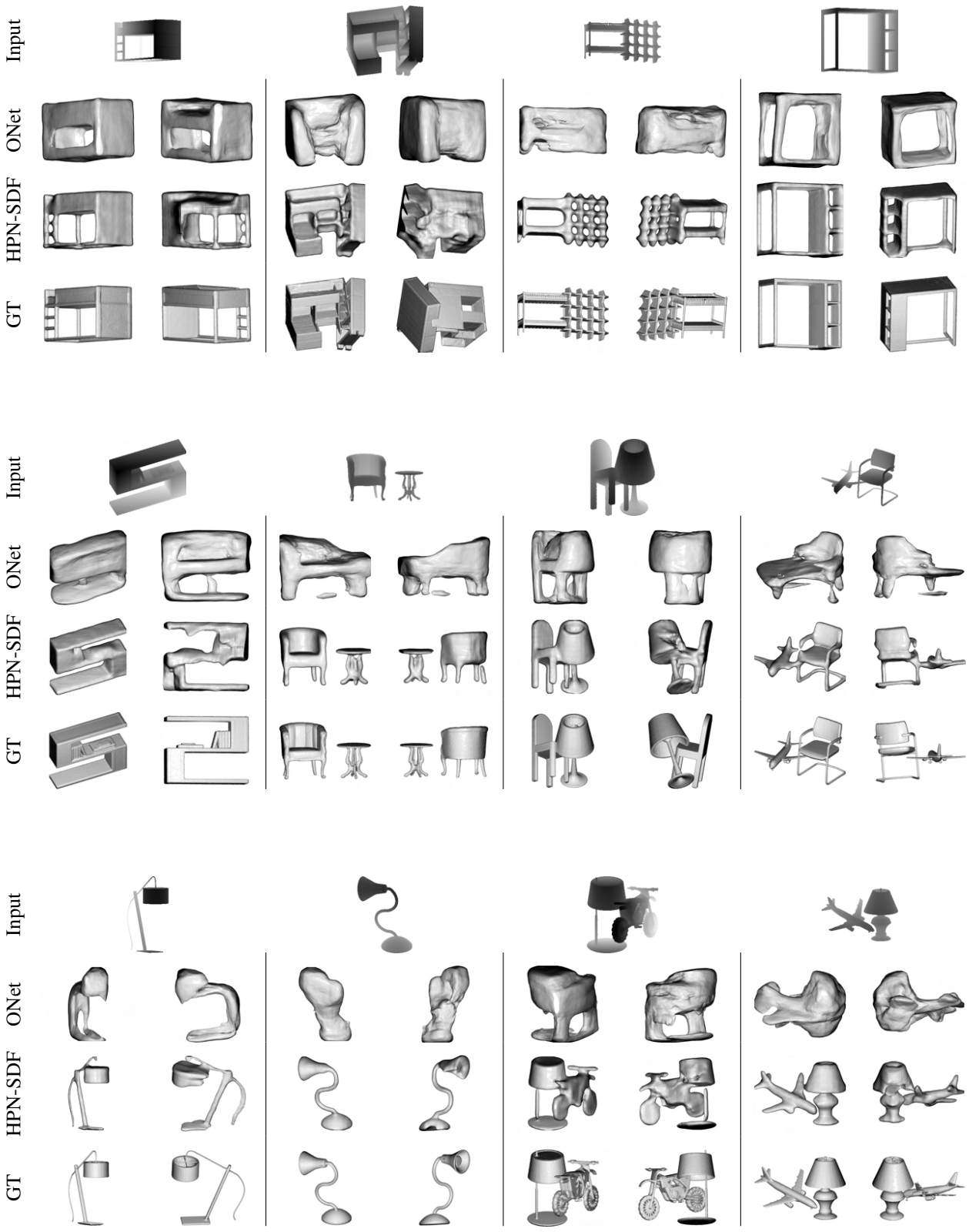

Figure 4. Qualitative results for networks trained in the *multi-class* setting on *airplanes*, *cars* and *chairs*.

		Mean (seen)		Airplane		Car		Chair		Lamp		Bench		Cabinet		Display		Speaker		Rifle		Phone		Vessel		Sofa		Table		Mean (unseen)		Composition	
		F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓	F↑	CD↓
plane,car,chair	ONet [2]	44.4	3.8	34.7	4.1	57.7	**3.2**	40.8	4.1	18.8	9.3	31.7	5.2	46.9	4.8	19.5	9.0	38.4	6.0	13.3	9.0	19.8	8.0	26.5	6.7	43.2	4.7	35.2	5.3	29.3	6.8	18.3	8.7
	ONet-SDF [2]	37.2	4.5	29.0	4.9	46.7	3.9	35.9	4.6	19.9	8.5	28.4	5.6	41.7	5.2	23.0	8.0	37.6	5.8	14.7	8.5	23.2	7.2	26.7	6.5	38.3	5.1	33.0	5.6	28.6	6.6	19.3	8.0
	GenRe [4]	-	4.5*	-	-	-	-	-	-	-	6.0*	-	5.0*	-	7.6*	-	6.0*	-	7.7*	-	3.1*	-	5.4*	-	4.8*	-	5.9*	-	5.7*	-	5.7*	-	-
	LDIF$_{svim1d}$ [1]	**76.7**	**0.5**	**87.9**	**0.2**	**80.0**	**0.3**	**62.1**	**0.9**	20.8	9.4	**48.6**	**1.2**	26.2	**3.4**	15.8	**5.7**	22.9	5.2	32.2	**1.5**	20.6	**2.3**	**48.1**	**1.4**	52.7	**1.3**	33.0	**3.3**	32.1	**3.5**	16.4	10.9
	HPN (ours)	45.4	3.8	36.8	3.9	55.1	3.5	44.3	3.8	38.4	4.8	37.4	4.3	**54.0**	**4.3**	43.2	6.3	**49.7**	**4.8**	33.1	5.0	45.9	5.3	37.3	5.4	46.6	4.5	43.7	4.4	42.9	4.9	30.2	5.7
	HPN-SDF (ours)	54.9	3.2	52.8	2.9	58.3	3.5	53.6	3.3	**56.5**	**3.5**	36.9	4.4	50.7	5.0	**47.1**	6.0	49.4	5.0	**39.9**	4.3	**53.8**	4.9	40.1	5.7	**54.4**	3.9	**53.1**	3.7	**48.2**	4.6	**42.4**	**3.9**
plane	ONet [2]	48.1	2.8	48.1	2.8	27.7	5.7	12.7	13.0	7.6	22.6	24.1	7.2	13.9	11.3	12.4	13.0	11.9	13.4	19.6	7.1	8.3	16.6	27.1	7.0	22.4	7.6	19.9	11.6	17.3	11.3	16.5	10.0
	ONet-SDF [2]	30.5	4.7	30.5	4.7	23.0	6.7	11.4	12.5	8.0	21.5	19.4	8.1	14.1	11.1	11.0	13.1	12.2	13.1	16.1	8.0	9.0	14.5	22.4	7.6	20.3	8.1	19.0	11.2	15.5	11.3	13.8	10.4
	LDIF$_{svim1d}$ [1]	**86.1**	**0.2**	**86.1**	**0.2**	33.1	**1.8**	12.9	22.7	10.0	50.4	26.9	5.4	10.4	28.5	11.6	27.9	8.3	39.1	30.6	**1.8**	9.9	47.3	**41.1**	**3.4**	18.2	5.9	18.7	16.1	19.3	20.9	15.9	16.7
	HPN (ours)	46.1	3.1	46.1	3.1	**45.3**	4.7	**35.0**	**6.2**	30.7	18.4	**35.5**	**4.9**	39.8	7.2	39.3	7.5	35.4	7.9	35.4	4.7	36.0	8.5	40.7	5.4	**39.9**	**5.5**	**42.7**	5.5	38.0	7.2	**30.1**	**5.9**
	HPN-SDF (ours)	39.0	3.7	39.0	3.7	43.7	5.1	33.2	**6.2**	**36.0**	**10.8**	31.7	5.3	**46.4**	**6.1**	**43.5**	**6.3**	**42.0**	**6.8**	**36.9**	4.8	**47.7**	**5.9**	40.1	5.5	39.7	5.7	41.4	**5.3**	**40.2**	**6.1**	27.5	6.3
chair	ONet [2]	36.2	4.6	15.9	9.1	29.6	5.8	36.2	4.6	16.6	10.3	26.6	6.0	37.6	5.7	18.7	9.6	34.2	6.5	8.6	12.1	17.7	9.1	19.1	8.8	35.3	5.4	31.6	5.9	24.3	7.9	16.5	9.3
	ONet-SDF [2]	37.9	4.4	19.2	7.6	31.5	5.6	37.9	4.4	21.5	8.3	29.2	5.6	39.2	5.5	21.5	8.4	37.4	6.0	10.9	10.7	21.6	7.5	20.8	8.2	37.3	5.0	34.0	5.5	27.0	7.0	18.4	8.8
	LDIF$_{svim1d}$ [1]	**59.2**	**1.0**	24.5	6.7	31.9	**1.8**	**59.2**	**1.0**	17.8	10.6	**42.5**	**1.4**	28.5	**3.7**	15.7	6.2	21.6	5.6	18.7	3.5	23.7	**3.4**	32.3	**2.2**	**44.4**	**1.4**	31.4	**3.9**	27.7	**4.2**	14.9	13.0
	HPN (ours)	43.0	3.9	37.2	4.4	**48.0**	4.6	43.0	3.9	40.2	4.6	37.3	4.1	51.3	4.7	44.9	6.0	48.6	**4.8**	40.6	3.8	38.6	5.5	41.9	5.3	**44.4**	4.6	**44.2**	4.3	43.1	4.7	31.2	5.3
	HPN-SDF (ours)	41.2	4.2	**40.9**	**4.0**	47.6	5.0	41.2	4.2	**43.6**	**4.3**	38.0	4.1	**51.4**	5.2	**46.6**	**5.7**	**48.8**	5.0	**46.7**	**3.3**	**47.5**	4.7	**44.7**	5.3	43.8	5.0	**44.2**	4.5	**45.3**	4.7	**31.7**	**5.2**
lamp	ONet [2]	42.0	4.7	20.8	7.7	26.8	6.4	20.4	8.1	42.0	4.7	23.7	7.1	35.1	5.5	24.9	7.0	37.8	5.6	21.0	6.7	30.8	6.8	27.2	6.4	24.2	7.2	29.1	7.1	26.8	6.8	18.1	8.5
	ONet-SDF [2]	31.6	5.5	17.2	8.1	24.0	6.9	18.3	8.3	31.6	5.5	18.5	8.3	34.5	5.6	23.5	7.3	34.3	6.0	17.4	7.3	28.3	6.9	21.9	7.6	21.5	7.9	23.4	7.9	23.6	7.3	16.1	8.9
	LDIF$_{svim1d}$ [1]	48.1	**2.5**	18.1	5.4	22.4	**2.4**	12.4	12.2	48.1	**2.5**	14.6	7.1	23.4	**3.4**	14.1	6.8	21.6	5.1	**48.6**	**1.3**	15.5	**3.6**	34.1	**2.0**	11.8	7.4	17.1	10.5	21.1	5.6	12.5	14.0
	HPN (ours)	**50.3**	3.6	**43.0**	**4.2**	**46.5**	5.1	**42.4**	**4.7**	**50.3**	3.6	**41.0**	**4.3**	**54.5**	4.5	48.6	5.4	**53.2**	**4.6**	43.8	4.0	54.0	5.4	**45.7**	5.0	**45.2**	**5.0**	**47.1**	**4.7**	**47.1**	**4.7**	**35.8**	**5.0**
	HPN-SDF (ours)	48.4	3.6	41.6	**4.2**	44.5	5.1	41.1	4.8	48.4	3.6	38.3	4.5	53.2	4.4	**49.8**	**5.2**	51.5	**4.6**	43.0	4.0	**56.7**	5.4	44.4	5.1	44.7	**5.0**	44.8	4.8	46.1	4.8	33.9	5.2

Table 1. Comparison of the hierarchical prior network (HPN) to the state of the art in terms of generalization. The top part of the table shows training in the *multi-class* setting, the lower part shows training on a single class. We report two metrics: F-score (F, shown in %) and Chamfer distance (CD, multiplied by 100 for better readability). * denotes results taken from the original paper. Results on categories seen during training are marked in blue. *Mean (unseen)* shows the average of per-class scores over unseen categories. *Composition* shows results on the composition of two objects per image. On compositions, HPN is more than twice as accurate as the state of the art and generally better on unseen classes, while LDIF is better on seen classes. Best viewed in color.

		Mean (seen)	Airplane	Car	Chair	Lamp	Bench	Cabinet	Display	Speaker	Rifle	Phone	Vessel	Sofa	Table	Mean (unseen)	Composition
plane,car,chair	ONet [2]	77.5	77.2	**83.6**	71.7	51.6	60.3	73.3	50.8	68.5	62.4	58.5	65.7	74.4	57.9	62.3	46.1
	ONet-SDF [2]	75.1	74.5	81.2	69.8	55.5	60.3	73.0	53.8	69.6	65.0	62.8	66.2	73.6	56.6	63.6	49.4
	LDIF$_{svim1d}$ [1]	72.6	74.6	82.3	60.8	18.6	32.1	49.5	12.8	44.0	28.9	19.7	50.2	66.5	27.4	35.0	16.2
	HPN (ours)	**79.9**	**80.6**	82.9	**76.1**	**72.1**	**70.6**	**77.7**	63.9	**75.6**	80.0	73.5	**72.2**	**77.5**	**67.8**	**73.1**	64.2
	HPN-SDF (ours)	76.5	78.9	78.1	72.5	65.3	70.0	75.2	**64.9**	74.8	**82.6**	**74.7**	71.2	72.5	66.5	71.8	70.1
plane	ONet [2]	80.0	80.0	66.1	28.3	20.1	36.4	50.7	36.6	42.7	62.6	33.4	59.3	53.9	23.4	42.8	35.0
	ONet-SDF [2]	75.4	75.4	67.3	34.7	24.2	43.8	51.3	37.7	44.3	65.3	37.4	61.6	58.8	30.1	46.4	40.9
	LDIF$_{svim1d}$ [1]	73.4	73.4	51.6	12.6	6.4	11.6	18.3	8.4	15.4	23.5	6.7	42.7	21.4	11.6	19.2	14.3
	HPN (ours)	**82.9**	**82.9**	71.7	55.9	42.6	61.2	62.0	54.4	57.3	77.8	52.5	67.6	67.2	51.8	60.2	59.2
	HPN-SDF (ours)	81.4	81.4	**71.8**	**58.8**	**53.1**	**63.7**	**64.8**	**60.3**	**61.8**	**80.4**	**64.6**	**69.6**	**68.6**	**58.8**	**64.7**	**59.7**
chair	ONet [2]	69.0	46.0	68.8	69.0	49.2	56.9	69.9	45.9	66.8	46.6	54.2	54.8	71.5	54.8	57.1	42.5
	ONet-SDF [2]	70.5	49.6	70.5	70.5	51.4	58.6	70.1	48.0	67.8	44.4	54.0	55.9	72.9	55.0	58.2	45.5
	LDIF$_{svim1d}$ [1]	59.5	15.3	49.2	59.5	13.3	28.0	46.7	13.7	39.6	9.1	13.3	28.2	63.7	24.7	28.7	12.9
	HPN (ours)	**76.3**	72.3	**76.2**	**76.3**	72.6	**72.3**	**75.8**	65.1	**75.6**	82.5	73.2	**71.1**	**77.2**	**68.8**	**73.6**	66.0
	HPN-SDF (ours)	74.4	**74.3**	74.6	74.4	**72.7**	72.2	73.6	**66.7**	74.4	**84.0**	**75.8**	70.5	75.0	67.6	73.4	**67.2**
lamp	ONet [2]	69.8	44.2	64.0	40.6	69.8	39.9	69.0	53.6	67.8	64.4	66.2	58.5	54.8	43.9	55.6	38.0
	ONet-SDF [2]	70.5	48.8	66.1	44.3	70.5	44.4	70.8	54.8	69.4	67.8	67.1	60.8	57.9	44.9	58.1	43.2
	LDIF$_{svim1d}$ [1]	41.8	9.9	42.6	9.6	41.8	7.3	45.6	25.3	48.8	37.0	35.0	38.0	10.3	13.2	26.9	10.3
	HPN (ours)	78.0	68.3	72.1	**65.0**	78.0	64.4	74.8	65.1	74.3	80.5	**72.1**	68.8	68.4	**62.4**	69.7	**64.1**
	HPN-SDF (ours)	**78.2**	**69.6**	**72.2**	64.9	**78.2**	**65.7**	**75.7**	**65.5**	**75.0**	**81.4**	70.5	**68.9**	**69.3**	**62.4**	**70.1**	63.5

Table 2. Comparison of the hierarchical prior network (HPN) to the state of the art in terms of generalization. The top part of the table shows training in the *multi-class* setting, the lower part shows training on a single class. This table reports the intersection over union (IoU) values in %.

3. Local retrieval

It was shown that single-view reconstruction with shape retrieval is competitive with network approaches [3]. The principle of recombination, enabled by the local parts, is also compatible with retrieval. Instead of a local reconstruction network, could we also use local retrieval for reconstruction?

Fig. 5 shows a study for patches of size $N = 64$. For each patch in the test image (first row), we retrieved the nearest neighbor patch by absolute L_1 distance from the multi-class training set. The resulting approximated test image (second row) shows that the silhouette of the nearest neighbors agrees well with the target image, especially for the car and the table. We cropped and assembled the corresponding 3D parts from the ground truth mesh to obtain the reconstruction (third row). The result is roughly right. However, compared to the Local@64 network (forth row), the reconstructed shape is not smooth and shows some strange artifacts. Another advantage of networks is the fast inference time with ~ 4 seconds per shape versus 3 hours for a naive nearest neighbor search over all parts of the training set.

That said, the non-smoothness and the runtime could both be mitigated with a more sophisticated retrieval approach. This shows: the key concept to enable generalization in single-view reconstruction across object categories is not a particular choice of network but the recombination and aggregation of local parts. The local retrieval counterpart to our network implementation is a viable alternative, even though the network version is probably more elegant.

Figure 5. Comparison to a retrieval baseline which locally retrieves the nearest neighbors for depth patches of size $N = 64$. **First row:** Test image. **Second row:** Image assembled from nearest neighbor patches from the training set. **Third row:** Reconstruction from the nearest neighbors (opposite viewpoint). **Fourth row:** Reconstruction from our Local@64 network.

References

[1] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, 2020. 6, 7

[2] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 1, 6, 7

[3] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 8

[4] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In *NeurIPS*, 2018. 6