Supplement Material
Any-Shot GIN: Generalizing Implicit Networks for Reconstructing Novel Classes

Yongqin Xian[1,2] Julian Chibane[2,3] Bharat Lal Bhatnagar[2,3] Bernt Schiele[2]
Zeynep Akata[2,3,4] Gerard Pons-Moll[2,3]

[1] ETH Zurich [2] Max Planck Institute for Informatics
[3] University of Tübingen [4] Max Planck Institute for Intelligent System

In the following, we provide more implementation details (Sec. 1) and results. In Sec. 2, we ablate the number of renderings per training shape, showing that the performance first improves as the number of renderings increases but saturates after 25 renderings. In Sec. 3, we present the class-wise comparison on ShapeNet, demonstrating that our improvement is uniform over most of classes. Finally, in Sec. 4, we provide more qualitative comparison on synthetic renderings of ShapeNet, real images of Pix3D and single-view human reconstruction, empirically confirming that our method generalizes better than competitors.

1. Implementation details

In this section, we provide implementation details for reproducing reported results.

Depth estimation network. Our depth estimation network follows a U-Net architecture [6] with a ResNet-18 [2] being the backbone. The encoder is a ResNet-18 [2] which encodes the input image into 512 features maps of size 7×7. The decoder is the reverse of the encoder with transposed convolutional layers, producing an estimated depth map of size 224×224. The network has a single shared encoder and two different decoders for predicting front and back depth maps respectively. The original input image is with 256 spatial resolution and we resize it into 224. We optimize the berHu loss [4] for both front and back depth (equally weighted). We chose Adam [3] as the optimizer and use learning rate 0.0001 and batch size 128. We select the number of training epochs by inspecting the validation loss on seen classes. Note that novel classes are completely excluded during training and hyperparameter tuning.

Implicit shape completion network. Our shape encoder network takes as input a voxel grid (resolution 128^3) and compromises 11 3D convolutional layers. We consider 7 intermediate feature maps ($L = 7$) at 0, 1, 3, 5, 7, 9, 11 layers. When extracting point-aligned multi-scale features, we take in account the query point $p = (x, y, z)$ itself and its 6 surrounding points by adding a small displacement $\epsilon = 0.0722$ to its x, y, z coordinates, i.e., $\{(x+\epsilon, y, z), (x-\epsilon, y, z), \dots, (x, y, z-\epsilon])\}$. The implicit decoder takes as input the point features and consists of 4 fully connected layers with ReLU being the non-linearity. We chose Adam [3] as the optimizer and use learning rate 0.0001 and batch size 8. In every batch, we sample 50000 query points per shape (refer to the main paper query for point sampling). We select the number of training epochs by inspecting the validation loss on seen classes.

Dataset statistics. For seen classes, We follow the data splits provided by SDFNet [7]. There are 28008/3883/7481 watertight meshes for training/validation/test (some meshes are deleted because there are errors when converting the original mesh into the watertight one). SDFNet [7] use all shapes of novel classes for testing. In contrast, we generate a data split on novel classes for few-shot learning studies. Specifically, we randomly split all shapes into 10243/420/2020 training/validation/test.

Distance between a novel class and all seen classes. In Fig. 4 (right) of the main paper, we show that the reconstruction results are correlated with the distance between the given novel class and all seen classes. Formally, we define the distance as the following,

$$\frac{1}{|C_{y_n}|} \sum_{m_n \in C_{y_n}} \min_{m_s \in \mathcal{T}} \text{CD}(m_n, m_s), \tag{1}$$

where CD denotes the Chamfer distance, C_{y_n} is the set of meshes from novel class y_n, m_n is a mesh from a novel classes, $\mathcal{T}$ is the training set of all seen classes, and m_s is a mesh from the training set. Essentially, for every mesh from the novel class y_n, Eq. 1 computes its minimum distance to the training set followed by averaging those minimum distances.

2. Ablating number of renderings

In all of our previous experiments, we follow SDFNet [7] to render 25 images from random view points for each

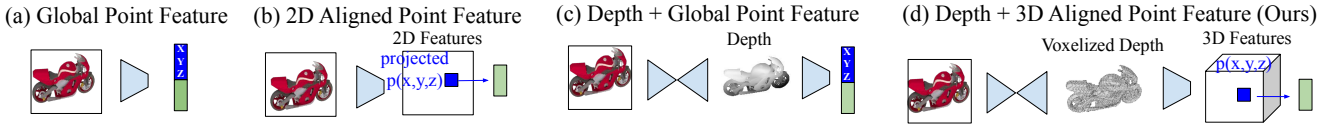

Figure 1. Our method differs fundamentally with existing class of methods. (a) ONet [5] directly compress the input image into a 1D latent representation, thus losing spatial structure. (b) DISN [8] learns input aligned features that allows it to retain more details but it does not leverage depth. (c) SDFNet [7] leverages intermediate depth but still compress the image into 1D vector. (d) Our GIN leverages intermediate depth, and unprojects it to 3D enabling 3D reasoning, which is key for details and generalisation.

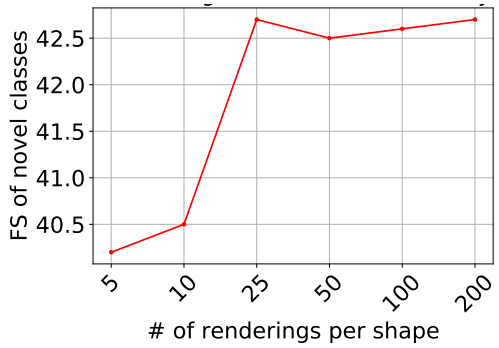

Figure 2. Number of renderings per shape vs F-score (FS) on novel classes. We conduct this experiment on the 5-shot learning setting where we have access to 5 training shapes per novel class. It shows that increasing the number of renderings does not lead to a performance boost after 25 renderings, which explains the choice of 25 renderings used in other experiments.

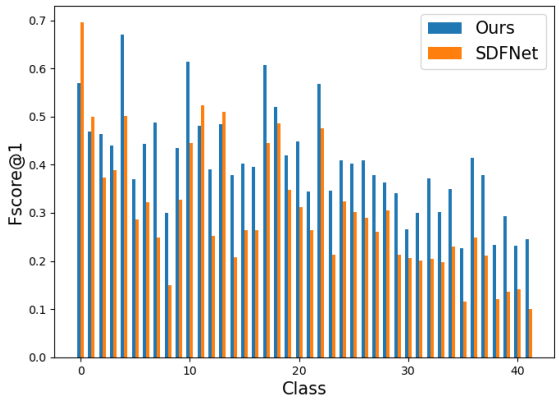

Figure 3. Class-wise comparison with SDFNet on novel classes of ShapeNet where novel classes are ordered by descending similarities (defined in Eq. 1) to seen classes. We observe that our method significantly outperforms the best baseline (SDFNet) in 38 out of 42 novel classes, indicating that the improvement is uniform over most classes, particularly the ones dissimilar to seen classes.

training shape. In this experiment, we study the impact of the number of renderings per shape. Specifically, we render each shape from R uniformly sampled viewpoints ($\theta_{azimuth} \in [0°, 360°)$, $\theta_{elevation} \in [-50°, 50°)$). As this experiment is expensive to run, we conduct it under the few-shot learning setting (specifically 5-shot) where we have access to 5 training shapes per novel class. We then render $R = \{5, 10, 25, 50, 100, 200\}$ images for each training shape of novel classes. In Fig. 3 (left), we observe that increasing the number of renderings does not lead to a performance boost after 25 renderings, which explains the choice of 25 renderings used in other experiments.

3. Class-wise comparison

We show class-wise comparison with the best baseline SDFNet [7] in Fig. 3 (right). We order the novel classes in the x-axis by their descending similarities to seen classes. We observe that our method significantly outperforms SDFNet in 38 out of 42 novel classes, indicating that the improvement is uniform over most of classes, particularly the ones dissimilar to seen classes (right side of the x-axis).

4. Qualitative results

In this section, we provide more qualitative comparison on the test set of both novel and seen classes. All methods i.e., Ours, ONet [5] and DISN [8] are trained on 13 seen classes on ShapeNet. For each method, we visualize the reconstructed meshes (256^3 resolution) given a single RGB image as the input. In addition, we also show more qualitative results on real images of Pix3D dataset.

Reconstructing unseen shapes of seen classes. We first compare with ONet [5] and DISN [8] on seen classes, which is the standard setting for single-image 3D reconstruction. As shown in Fig. 4, our results are obviously more consistent with the input image than competitors. For example, our reconstructed sofa (the first row) does not have handles, resembling the input image. In contrast, the results of DISN and ONet incorrectly hallucinate the handles, which seem to be retrieved from training shapes. Our advantage becomes more clear for challenging cases e.g., lamp in the third row. These results empirically demonstrate that our method is able to outperform SOTA under the standard single-image 3D reconstruction setting.

Reconstructing unseen shapes of novel classes. In Fig.5, our results again attain good consistency with the input images of novel classes, while DISN and ONet lack structural details in the input. Moreover, we observe that results of

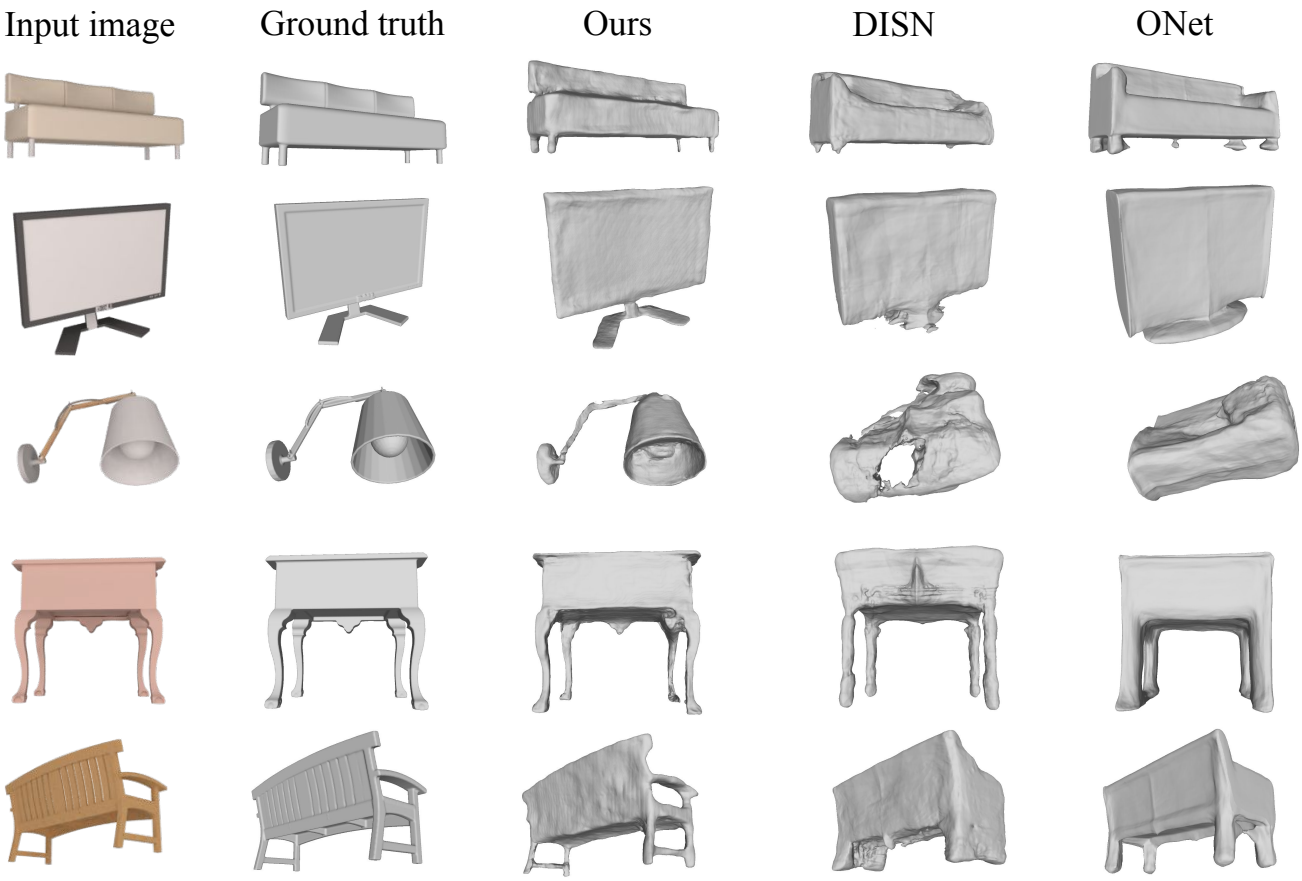

Figure 4. Qualitative results of seen classes on ShapeNet. We visualize the reconstructed meshes (256^3 resolution) given a single RGB image as the input. All the methods i.e., Ours, DISN [8], and ONet [5], are trained on the 13 seen classes on ShapeNet. Our results are obviously more consistent with the input image than competitors. For example, our reconstructed sofa (the first row) does not have handles, resembling the input image. In contrast, the results of DISN and ONet incorrectly hallucinate the handles, which seem to be retrieved from training shapes. Our advantage becomes more clear for challenging cases e.g., lamp in the third row.

DISN and ONet degrade significantly on novel classes compared their results to seen classes. For example, our method recovered the feet and pedals of the piano (the first row), while DISN and ONet generated cuboids without any details. Furthermore, our method reconstructed the shape of the bed (the fourth row), while results of ONet and DISN look more like a sofa, which is a training class. These results empirically indicate that our method generalizes better to diverse unseen shapes of novel classes.

Qualitative results on real images of Pix3D. In Fig.6, our methods once again achieve the best results on real images. For example, our method reconstruct handle and feet of the chair (the second row), while SDFNet, ONet and MarrNet generate meshes that are inconsistent with the input image. Moreover, our method reconstructed the shape of the desk (the fifth row), while the second best method i.e., SDFNet, produces a table surface with holes. Also, ONet consistently fails on real images as it always outputs a chair or sofa shape which is a training class. These results empirically demonstrate that our method generalizes well to real images despite being trained only on synthetic renderings.

Qualitative results on single-view human reconstruction. We explore the generalization limit of our method by evaluating it on articulated human shapes from Human [1] dataset. In this experiment, the input is a partial 3D point cloud (around 5000 points on the visible side) obtained from a depth camera. The task is to reconstruct human shapes from the input point cloud using our shape completion model trained solely on ShapeNet. This is extremely challenging because human shapes are significantly dissimilar to the object shapes of ShapeNet. As shown in Fig. 7, our method is able to recover the visible details and generate reasonable surfaces at the backside region (occluded). These results once again demonstrate the strong generalization capability of our approach.

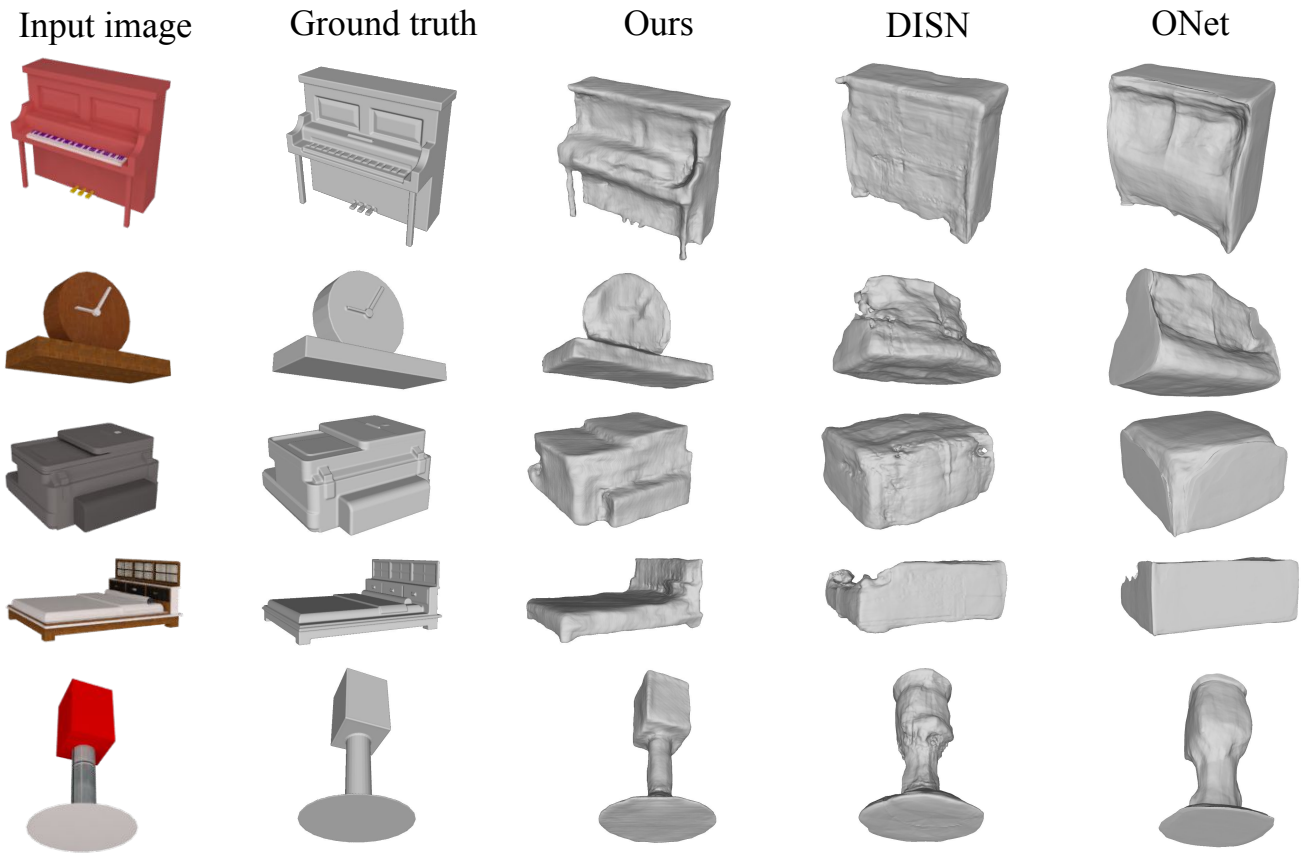

Figure 5. Qualitative results of **novel classes** on ShapeNet. We visualize the reconstructed meshes (256^3 resolution) given a single RGB image as the input. All the methods i.e., Ours, DISN [8], and ONet [5], are trained on the 13 seen classes on ShapeNet. Our results attain good consistency with the input images, while DISN and ONet lack structural details in the input. For example, our method recovered the feet and pedals of the piano (the first row), while DISN and ONet generated cuboids without any details. Moreover, our method reconstructed the shape of the bed (the fourth row), while results of ONet and DISN look more like a sofa, which is a training class.

References

[1] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *CVPR*, 2020. 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 1

[5] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2, 3, 4

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015. 1

[7] Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M Rehg. 3d reconstruction of novel object shapes from single images. In *2021 International Conference on 3D Vision (3DV)*, 2021. 1, 2

[8] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. 2019. 2, 3, 4

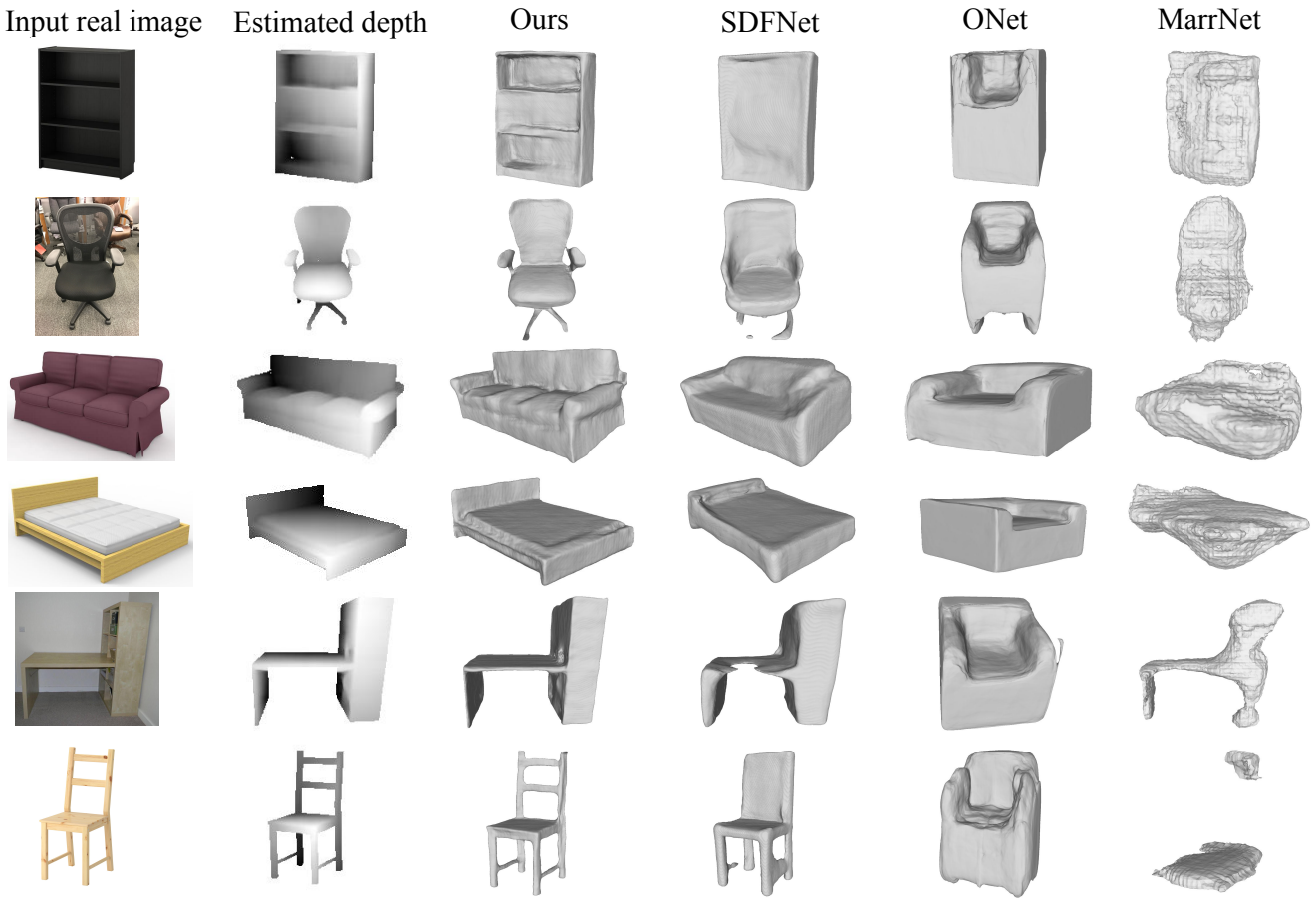

Figure 6. Qualitative results on **real images** of Pix3D. We visualize our estimated depth and reconstructed meshes (256^3 resolution except MarrNet is with 128^3) given a single real color image as the input. All the methods are trained on synthetic renderings of ShapeNet, then evaluated on real images, which is particularly challenging due to the substantial domain gap between synthetic renderings and real images. Our qualitative results are significantly better than other methods on real images, implying the strong generalization performance of our method.

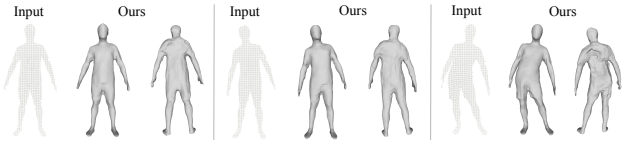

Figure 7. Single-view shape completion of articulated human bodies from partial point clouds (note that the back is completely occluded). Our model is trained on the 13 ShapeNet classes e.g., cars, chairs, and airplanes, and *has never seen any human shapes*. For three point clouds, we show our reconstructions from two different viewpoints.