Black-Box Test-Time Shape REFINEment for Single View 3D Reconstruction Supplementary

1. Additional Notes on REFINE

1.1. Extended Figures and Tables

Several figures and tables are given in this supplementary to complement the main paper. Figure 1 illustrates that although OccNet [22], Pix2Mesh [29], and AtlasNet [11] produce very different failure cases and artifacts, REFINE improves the both the input image consistency and 3D accuracy of all methods. For detailed per-class results on ShapeNet, please refer to Table 1. For per-class results on RerenderedShapeNet, refer to Tables 2 and 3. Figures 10 and 11 provides performance measurements visualized as a heatmap for 3D-ODDS across the class/angle and domain/angle factors. Figures 21 and 22 plots reconstruction accuracies on a per-object basis, for all objects in 3D-ODDS.

Figures 12, 13, 14, and 15 show more REFINE examples on real-world images (Pix3D and 3D-ODDS). Figure 16 is on RerenderedShapeNet, while Figure 17 is on ShapeNet. All these figures use an OccNet to reconstruct the original mesh. For REFINEment examples using the AtlasNet, Pix2Mesh, and Pix2Vox reconstruction methods, please refer to Figures 18, 19, and 20 respectively.

1.2. Scope, Limitations, and Future Work

Test-time shape refinement explores whether or not reconstructions can be improved by the use of additional auxiliary test-time information. In the work of [23], this was performed by optimizing the parameters of a SVR network given a coarsely reconstructed mesh, object silhouette, and pose. We also follow this input setting, which allows us to focus on studying REFINE independently without confounding factors. Research in automatic image segmentation [1, 12, 24, 35] and pose estimation [16, 27, 31] is beyond the scope of this paper, and advancement in those tasks is left for future research. Additionally, we believe that the REFINE paradigm and 3D-ODDS dataset provide an excellent foundation for future improvements in test-time refinement and generalizable reconstruction. For example, it may be worthwhile to explore more complex architectures, high level learned priors, topological modifications, and generative/adversarial formulations. They may lead to

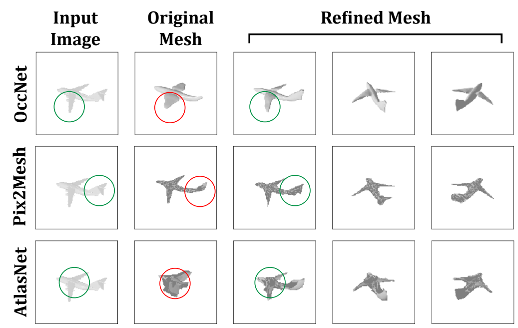

Figure 1. An airplane reconstructed by three different methods [11, 22, 29]. Since the methods differ greatly, they exhibit very different failure cases and artifacts. Nevertheless, REFINE improves all reconstructions.

more powerful refinements, but also significantly increased challenges in avoiding degenerate solutions.

1.3. Potential on Societal Impact

REFINE is a relatively lightweight instance-based, class-agnostic postprocessing step. It does not rely on any dataset to train on; its effectiveness is due its formulation, designed architecture, and proposed loss functions. Thus, we do not anticipate immediate negative environmental, fairness, or privacy concerns directly resulting from REFINE. However, it requires a black-box separate single view reconstruction network S which reconstructs the original meshes. In real world deployments we encourage understanding the design and training procedure of S, especially its potential biases and security/privacy concerns which may be problematic in some neural networks [28, 30].

1.4. REFINE Architecture

The feature map encoder is based on the first two convolutional layers of ResNet-18 [13]. The dimension of all 8 graph convolution layers used is 128, and each is followed by a ReLU nonlinearity. V_{dis} is predicted with a single fully connected layer, while V_{sConf} is predicted with fully connected layers of sizes 32, 16, and 1 (a ReLU follows each except for the last, which uses sigmoid). The feature map encoder is initialized using ImageNet [5] classification pretrained weights, while all other weights are randomly ini-

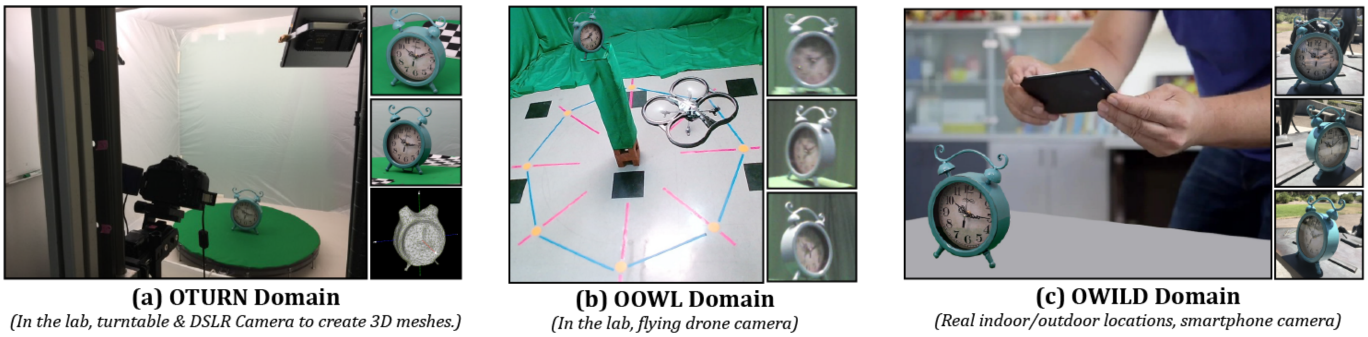

Figure 2. The data collection procedure differs for each domain of the 3D-ODDS dataset. OTURN uses a high resolution DSLR camera in a controlled turntable setup, which was used to generate 3D meshes using structure-from-motion. OOWL is captured using a drone camera, mid-flight. OWILD depicts objects in diverse indoor/outdoor locations, captured using smartphone cameras.

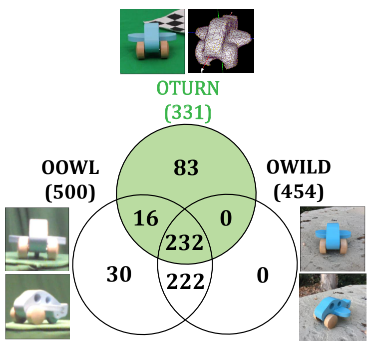

Figure 3. A venn diagram of objects that can be found in the 3D-ODDS dataset's three domains: OTURN, OOWL, and OWILD. 232 objects can be found simultaneously in all three domains.

tialized; no weights are frozen during optimization. We use the PyTorch3D differentiable renderer [18], which is implemented based on [19]. All hyperparameters were tuned with a small portion of RerenderedShapeNet, disjoint from the test set.

1.5. Loss Functions

The weights chosen for the loss functions are $\lambda_{Sil} = 10$, $\lambda_{Isym} = 80$, $\lambda_{Vsym} = 20$, $\lambda_{SymB} = 0.0005$ $\lambda_{Dis} = 100$, $\lambda_{Nc} = 10$, and $\lambda_{Lp} = 10$. We found that this configuration works well overall in practice; however, they are not overly sensitive and changing them by $\pm 25\%$ didn't change results significantly. Beyond this range, we observed that these weights operate intuitively as one would expect (as illustrated in Figure 9 of the main paper). In general, they are not difficult to tune and practitioners can modify them accordingly with their use case. For example, one might increase the weight of λ_{dis} if they are confident that in their use case, input reconstructions are already of relatively high quality. This would effectively apply a stronger prior towards minimizing the displacements' magnitudes. Alternatively, if only symmetric objects are considered λ_{SymB} can be increased.

Additionally for the symmetry losses, there are methods to predict planes of object symmetry [8, 34] but we found them to be unnecessary since most reconstruction methods output semantically aligned meshes for objects of the same class. In general, the objects are aligned so that $\mathcal{Z}$ is the vertical plane with $\vec{n} = [0, 0, 1]^\intercal$. We adopt this convention in all our experiments. For the image rendering based symmetry loss, we also tried to use differentiably rendered normal maps and depth maps instead of only silhouettes. However, we found that this increased the computational complexity, and resulted in nearly the same performance.

1.6. Time Efficiency

Ideally, test-time shape refinement postprocessing should support any mesh and be fast. REFINE intrinsically satisfies the first requisite, since it is black-box, class-agnostic, and allows variable number of vertices per mesh. Optimization from scratch converges in relatively few iterations, approximately 400 (i.e. 400 forward and backward passes). This requires about 90 seconds on a GTX 1080Ti GPU. Moreover, because instances are treated independently, the refinement is trivially parallelizable. Since 4 instances fit on a GPU, a two GPU server trivially achieves a per-instance refinement time of $90/(4*2) \approx 11$ seconds, which is effective in terms of the second requisite.

2. Dataset Additional Details

Three new datasets are proposed in this paper: 3D-ODDS, RerenderedShapeNet, and ShapeNetAsym. All datasets will be publicly released upon publication. More details about these datasets are provided as follows.

2.1. 3D-ODDS

The proposed 3D-ODDS dataset contains multiview objects in 3 domains, as illustrated in Figure 2. The first domain is the contribution of this paper, OTURN, which is taken in the lab using a turntable and DSLR camera. 331 objects were imaged with dense pose coverage; 3 elevation angles and 72 azumuth angles ($5°$ increments), for $331*3*72 = 71,496$ total images which are of high resolution with simple backgrounds. All the OTURN images for an example airplane object is provided in Figure 9. These images were then used to reconstruct a mesh for each object, using structure-from-motion software [21]. The sec-

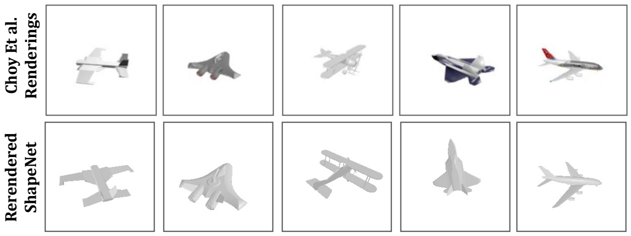

Figure 4. Comparisons between ShapeNet renderings from Choy Et al. [3] and RerenderedShapeNet. Both use the same 3D models, but a domain gap is intentionally created through viewpoint, lighting, and rasterization implementation differences in the rendering process.

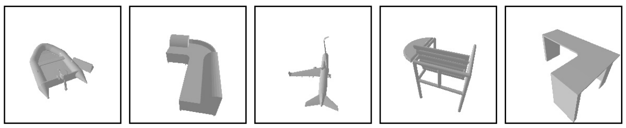

Figure 5. Example images from the ShapeNetAsym dataset.

ond domain is OOWL [14], with multiview ($45°$ azimuth increments) images of objects collected in the lab using a drone camera. These images have a green background and can be blurry, due to camera shake from flight. The third domain is OWILD [15], which contains multiview images of objects (also $45°$ azimuth increments) in diverse real world indoor/outdoor locations. These images are taken with a smartphone camera. A venn diagram of the objects found in these domains is given in Figure 3; 232 objects can be simultaneously found in OTURN, OOWL, and OWILD. More example objects in the 3D-ODDS dataset are shown in Figures 7 and 8.

Note that some imperfections are present in the mesh scans, as a natural consequence of real-world 3D data collection. This can be due to absence of texture or difficult material reflectance properties. To account for this, we manually annotated each mesh's quality; there are 101 excellent quality meshes, 198 high quality meshes, and 32 low quality meshes. High quality meshes are characterized by overall geometrical resemblance to the true shape; some small superficial noise artifacts may exist. In all experiments, we only considered objects found in OTURN, OOWL, and OWILD with excellent/high quality meshes; this subset consists of 212 objects.

2.2. RerenderedShapeNet and ShapeNetAsym

RerenderedShapeNet matches the ShapeNet [2] models in the test set given by [3]. However, a small domain gap is intentionally induced compared to the images from [3] through differences in the rendering process. This allows us to measure the robustness of SVR methods to domain gaps of various sizes between training on [3] and inference (on RerenderedShapeNet, ShapeNetAysm, Pix3D, or 3D-ODDS). In particular, [3] is rendered textured with Blender's Eevee engine [4] at distance 0.8, uses 2 sun light sources 180 degrees rotated from one another, with specular and diffuse shading disabled. Meanwhile, RerenderedShapeNet is rendered textureless with PyTorch3D's Hard Phong shading [18] at distance 1, uses a point light source at $(0, 5, -10)$, with ambient intensity 0.3, specular intensity 0.2, and diffuse intensity 0.3. Images in both have an elevation of $40°$ and randomly samples azimuths uniformly. An illustration of differences between RerenderedShapeNet and renderings from [3] is shown in Figure 4. In total, RerenderedShapeNet contains 8629 images and meshes.

We also introduced an asymmetric subset of RerenderedShapeNet called ShapeNetAsym containing 1259 images and meshes. The meshes are all asymmetrical, in the sense that each mesh has a symmetry loss $L_{Isym} < 0.01$ for $\lambda_{SymB} = 1$ and $\sigma_{j,k} = 1$. Some examples from ShapeNetAsym can be found in Figure 5.

3. Evaluation Details

3.1. Analysis of Variance Results

Analysis of Variance (ANOVA) [7] is a commonly used statistical model and hypothesis testing framework for splitting observed variability into systemic factors and random error. In particular, it can be used to model the influence of categorical independent variables (i.e. "factors") on a continuous dependent variable, to check if they are statistically significant. Due to its hierarchical structure, 3D-ODDS has 3 factors: class (14 levels, one for each class), domain (3 levels in OTURN, OOWL, OWILD) and pose (8 levels from $45°$ viewpoint azimuth increments). This suggests a 3-way ANOVA with blocked design, to account for object-based variability and dependencies (i.e. each object comprises a block). Our dependent variable in this case is F-Score after REFINEment of an OccNet. All factors, pairwise interaction effects, and triplet interaction effects were found to be statistically significant at the $\alpha = 0.05$ level. total variability was decomposed into 13% class, 2% pose, 1% domain, 19% object instance. Interaction effects between (class,domain), (class, angle), (domain, angle), and (class, domain, angle) were found to be 7.6% , 6.8% , 0.3%, and 2.5% respectively, for 17.2% in total attributed to interaction effects between the factors.

Note that ANOVA has several assumptions. The dependent variable should be additively influenced, and ideally errors should be independent, homoscedastic, and Gaussian (though ANOVA is considered relatively robust to some departures [10, 20], due to the central limit theorem). In light of this, we suggest viewing these ANOVA results as a simple summary heuristic useful for gaining further intuition and insight into 3D-ODDS, rather than dogma.

3.2. Metrics

We detail the metrics used in the main paper below. For Pix3D, we follow the practice of [23] and exclude images where the object is truncated resulting in 5325 test instances. The meshes used in this work have approximately

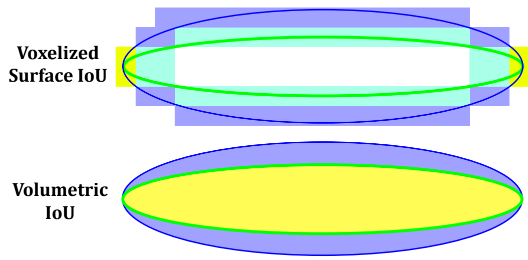

Figure 6. A 2D illustration of why a surface-based coarsely voxelized IoU metric (top) can be inaccurate, compared to the standard volumetric IoU (bottom). Highlighted in yellow are the intersections of the green and blue ellipse with major axis of length 1. Note that the surface-based voxelized IoU heavily underrepresents the intersection-based similarity of the two shapes compared to the volume based approach.

1500 vertices, although REFINE can handle much larger meshes (the only limitation being GPU memory).

The **Earth Mover Distance (EMD)** measures distance between point clouds S_1, S_2 sampled from two meshes, by solving the assignment optimization problem given by

$$d_{EMD}(S_1, S_2) = \min_{\phi:S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2, \quad (1)$$

where ϕ is an optimal bijection. Because exactly computing EMD is too expensive, we utilize the approximation given by [6]. Like [23], we sample 2048 points from the reconstructed mesh and target mesh, scaled so it fits in a sphere of radius 1. For this metric, lower is better.

Chamfer-l_2 Distance (CD-l_2) is a widely used metric in the 3D literature [6, 11, 22, 29]. It computes the average nearest neighbor distance between points sampled from two meshes. Given these two sampled point clouds S_1, S_2, their Chamfer-l_1 distance is

$$d_{CD-l_2}(S_1, S_2) = \sum_{p \in S_1} \min_{q \in S_2} \|p-q\|_2^2 + \sum_{q \in S_2} \min_{p \in S_1} \|p-q\|_2^2. \quad (2)$$

Just like EMD, we follow [23] and sample 2048 points from the reconstructed mesh and target mesh, scaled so it fits in a sphere of radius 1. For this metric, lower is better.

F-score is formulated as the harmonic mean between precision and recall at a distance threshold between two shapes. Precision involves the number of points on the reconstruction which lie a certain distance to ground truth; recall measures completeness by the number of points on the ground truth which lie within a certain distance to the reconstruction. Like [23], we set this distance threshold to be 0.05 and sample 10000 points after rescaling to a sphere of radius 1. For more details, please refer to [26]. For this metric, higher is better.

Volumetric IoU is a standard metric [26] computed by the volume of two meshes' union divided by the volume of their intersection. Like [22], we obtain an unbiased estimate by randomly sampling 100k points in the bounding volume and checking if points are inside the meshes (scaled to radius 1). A higher score is better. Non-watertight meshes were made watertight with ManifoldPlus [17]. Note that MeshSDF [23] reports scores for their non-standard version of the 3D IoU which only accounts for coarsely $30 \times 30 \times 30$ voxelized 3D surface, not internal volume. As this can be highly misleading (see Figure 6), we instead use the conventional definition of 3D IoU in all experiments.

3.3. Use of Existing Assets

The following code/dataset assets were used to conduct experiments for this paper.

- Occupancy Networks [22]. Used under the MIT License, copyright 2019 Lars Mescheder, Michael Oechsle, Michael Niemeyer, Andreas Geiger, Sebastian Nowozin. Commit 406f794. https://github.com/autonomousvision/occupancy_networks.
- Pix2Mesh [29]. Used under the Apache License 2.0. Commit 7c5a7a1. https://github.com/nywang16/Pixel2Mesh.
- Pix2Vox [32]. Used under the MIT License, copyright 2018 Haozhe Xie. Commit f1b8282. https://github.com/hzxie/Pix2Vox.
- AtlasNet [11]. Used under the MIT License, copyright 2019 ThibaultGROUEIX. Commit 22a0504. https://github.com/ThibaultGROUEIX/AtlasNet.
- Mesh R-CNN [9]. Used under a BSD License, copyright Facebook, Inc. and its affiliates. Commit d582649. https://github.com/facebookresearch/meshrcnn.
- OOWL [14]. Obtained explicit permission from authors to use in 3D-ODDS. http://www.svcl.ucsd.edu/projects/OOWL/CVPR2019_adversarial.html#dataset.
- OWILD (i.e. ObjectPI) [15]. Obtained explicit permission from authors to use in 3D-ODDS. http://www.svcl.ucsd.edu/projects/OOWL/CVPR2019_PIE.html#dataset.
- ShapeNet [2]. https://shapenet.org/.
- Pix3d [25]. http://pix3d.csail.mit.edu/.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3, 4

[3] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 3

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 4

[7] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992. 3

[8] Lin Gao, Ling-Xiao Zhang, Hsien-Yu Meng, Yi-Hui Ren, Yu-Kun Lai, and Leif Kobbelt. Prs-net: Planar reflective symmetry detection net for 3d models. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):3007–3018, 2020. 2

[9] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9785–9795, 2019. 4, 10

[10] Gene V Glass, Percy D Peckham, and James R Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, 42(3):237–288, 1972. 3

[11] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 1, 4, 10

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Chih-Hui Ho, Brandon Leung, Erik Sandstrom, Yen Chang, and Nuno Vasconcelos. Catastrophic child's play: Easy to perform, hard to defend adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9229–9237, 2019. 3, 4

[15] Chih-Hui Ho, Pedro Morgado, Amir Persekian, and Nuno Vasconcelos. Pies: Pose invariant embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12377–12386, 2019. 3, 4

[16] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3385–3394, 2019. 1

[17] Jingwei Huang, Yichao Zhou, and Leonidas Guibas. Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. *arXiv preprint arXiv:2005.11621*, 2020. 4

[18] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*, pages 1–1. 2020. 2, 3

[19] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[20] Lisa M Lix, Joanne C Keselman, and Harvey J Keselman. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test. *Review of educational research*, 66(4):579–619, 1996. 3

[21] Agisoft LLC. Agisoft metashape. 2

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 1, 4, 10

[23] Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoit Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Meshsdf: Differentiable iso-surface extraction. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22468–22478. Curran Associates, Inc., 2020. 1, 3, 4, 10

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[25] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 4

[26] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 4

[27] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018. 1

[28] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[29] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 1, 4, 10

[30] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[31] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

[32] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2690–2698, 2019. 4, 11

[33] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems*, pages 492–502, 2019. 10

[34] Yichao Zhou, Shichen Liu, and Yi Ma. NeRD: Neural 3d reflection symmetry detector. In *CVPR*, 2021. 2

[35] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018. 1

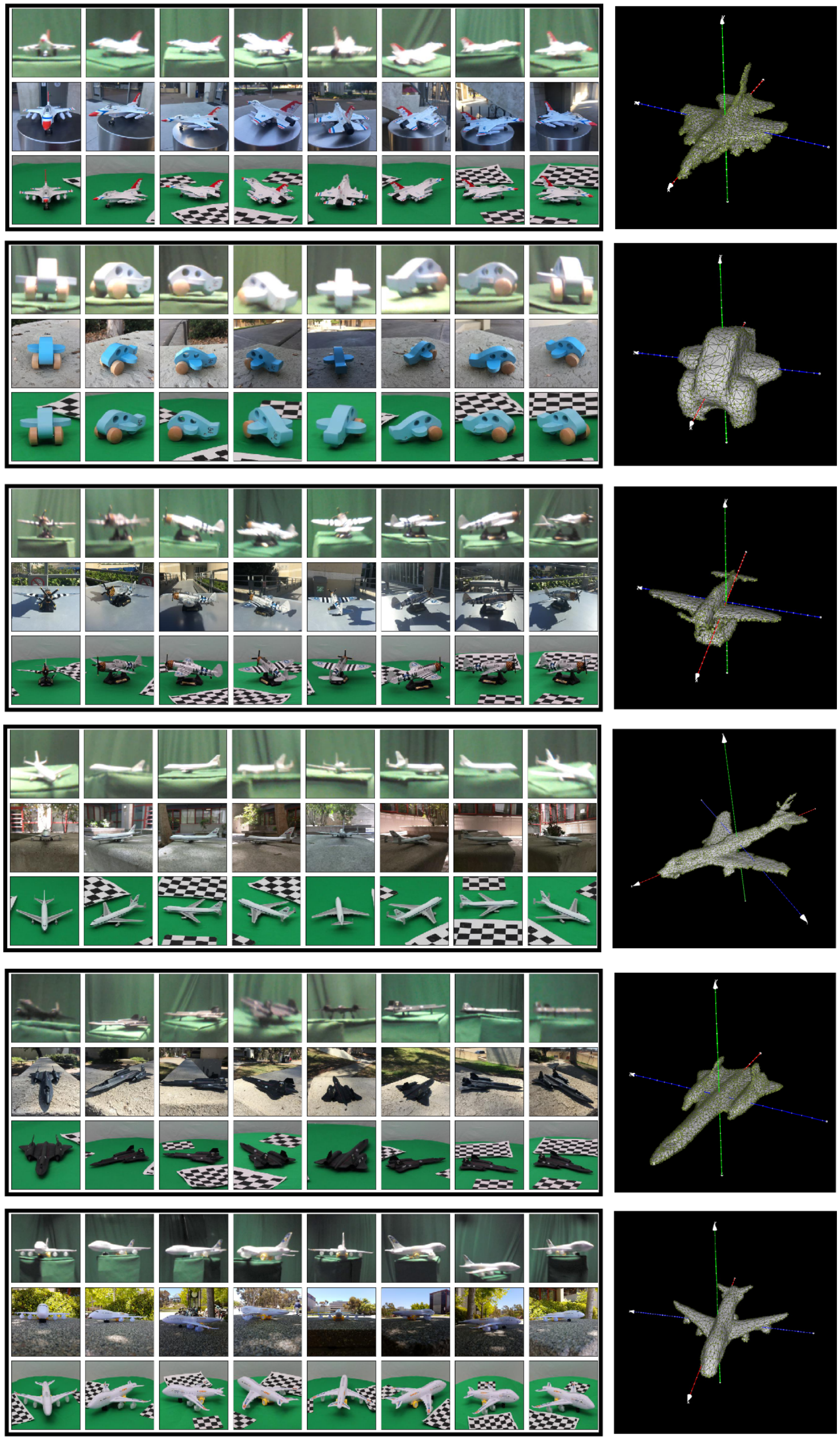

Figure 7. Additional example images and mesh for objects in the Airplane class of 3D-ODDS.

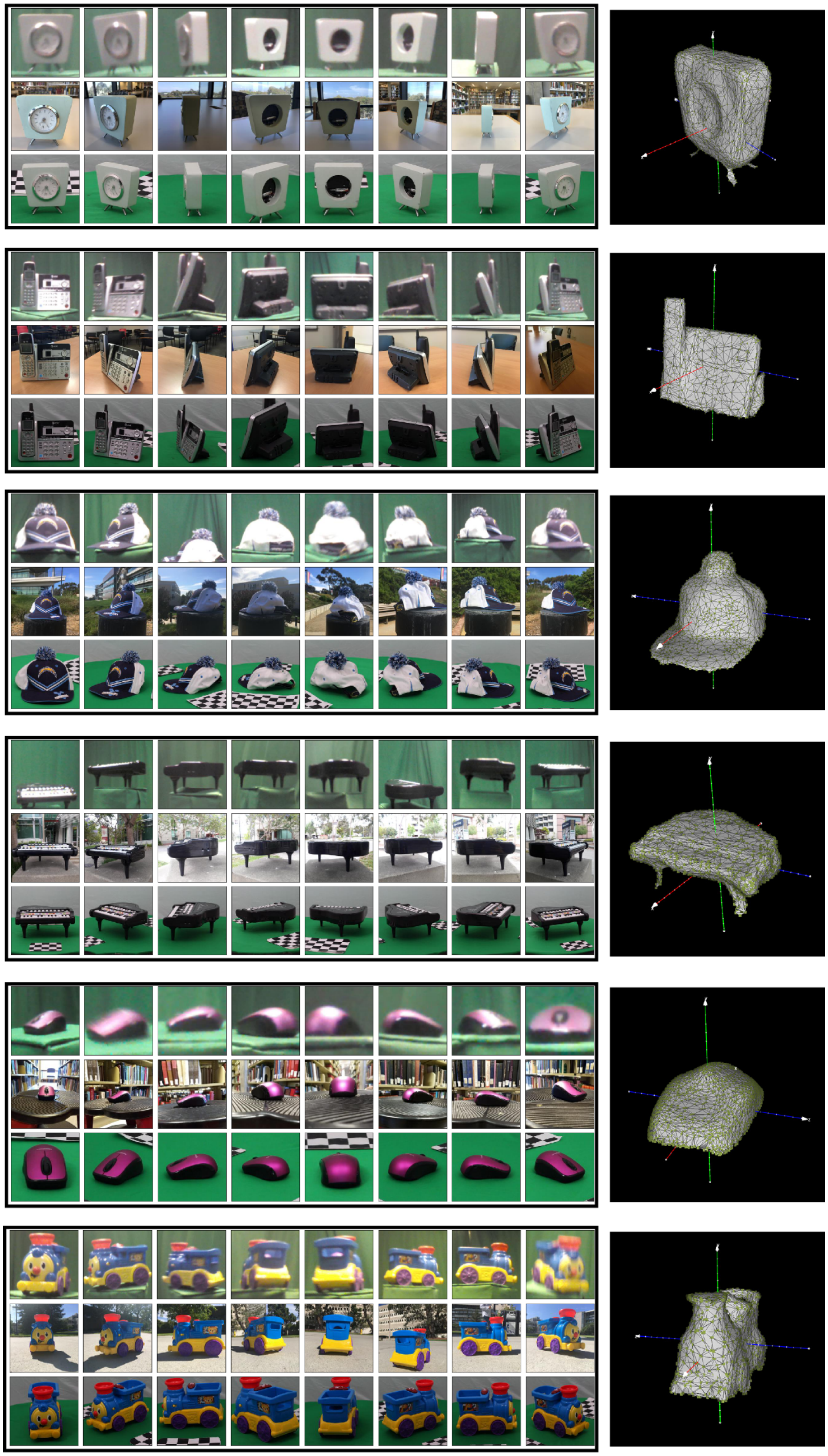

Figure 8. Additional example images and mesh for objects in 3D-ODDS.

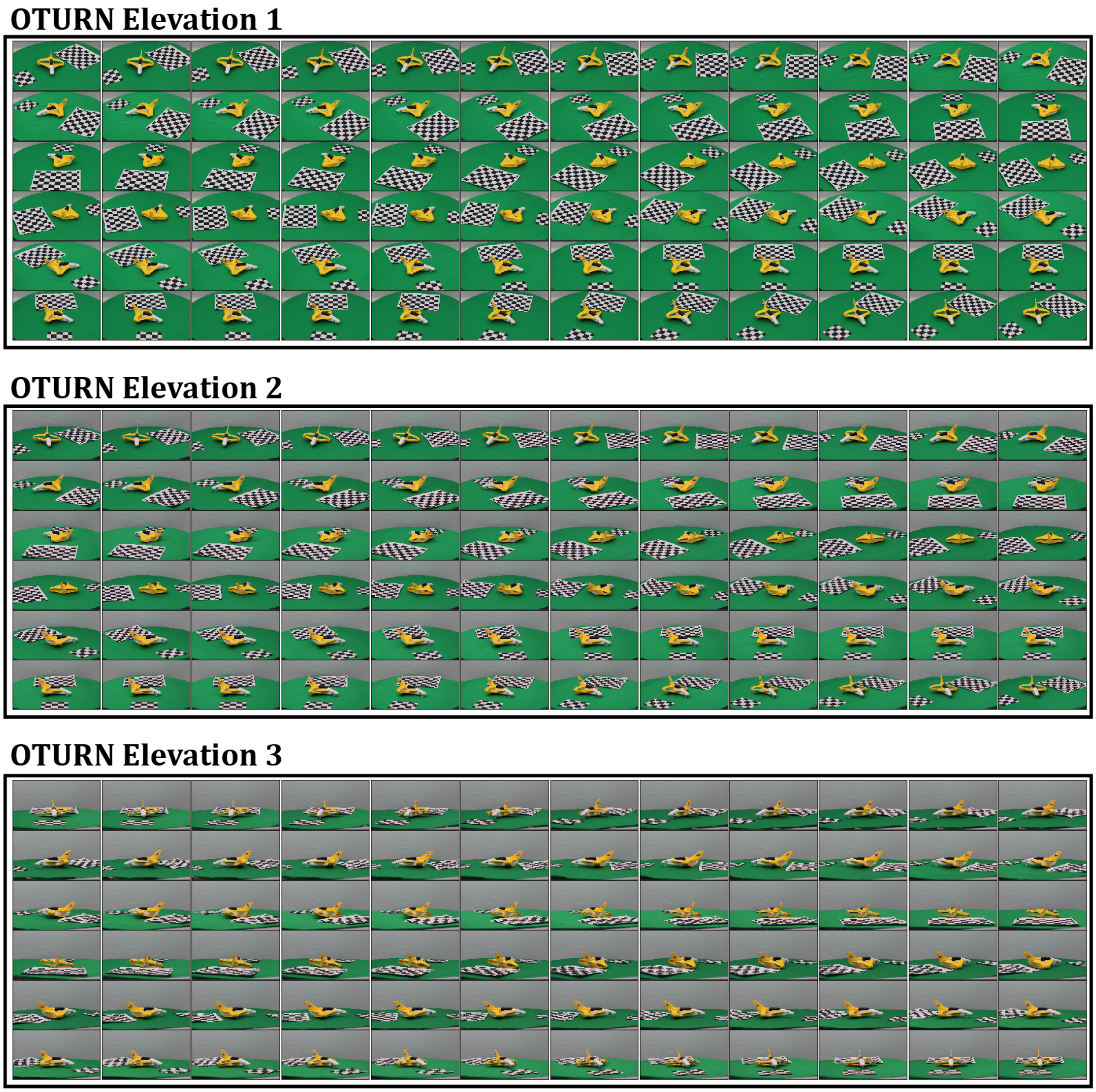

Figure 9. All 216 images for an airplane object in the OTURN domain of 3D-ODDS. There are 72 azimuth angles (increments of 5°) for 3 elevation angles.

	0°	45°	90°	135°	180°	225°	270°	315°	All
Airplane	38.4 → 45.5 (18.1 → 15.0)	38.5 → 46.6 (17.9 → 14.1)	41.2 → 53.8 (18.7 → 15.6)	37.5 → 46.7 (17.3 → 13.3)	37.6 → 48.0 (20.6 → 19.6)	39.7 → 48.5 (20.3 → 12.6)	38.1 → 50.0 (20.6 → 15.9)	38.2 → 47.3 (18.9 → 14.1)	38.7 → 48.3 (18.9 → 15.2)
Boat	25.0 → 33.9 (14.6 → 12.9)	41.9 → 46.1 (18.6 → 15.0)	41.5 → 49.1 (19.9 → 15.3)	36.1 → 45.0 (17.0 → 16.0)	22.4 → 30.2 (11.6 → 12.4)	36.7 → 43.4 (21.3 → 17.1)	38.7 → 48.6 (22.3 → 15.7)	36.5 → 43.5 (18.9 → 12.4)	34.9 → 42.5 (19.4 → 15.9)
Bottle	59.0 → 71.6 (18.3 → 8.8)	58.1 → 70.1 (25.6 → 17.1)	60.7 → 73.2 (14.4 → 13.9)	56.6 → 74.4 (23.0 → 14.3)	59.8 → 75.7 (19.4 → 14.5)	60.0 → 74.6 (20.5 → 12.5)	56.0 → 71.6 (21.3 → 16.2)	59.0 → 70.1 (16.0 → 13.9)	58.7 → 72.7 (19.6 → 13.8)
Bowl	40.5 → 46.4 (14.0 → 13.9)	40.0 → 46.5 (13.7 → 12.2)	40.8 → 48.0 (13.1 → 12.2)	39.0 → 43.3 (13.5 → 12.7)	36.8 → 42.0 (13.0 → 13.5)	38.6 → 43.7 (14.2 → 13.8)	38.8 → 46.0 (11.6 → 12.2)	40.4 → 46.1 (13.2 → 13.1)	39.4 → 45.2 (13.2 → 13.0)
Can	52.8 → 49.8 (24.5 → 24.8)	48.6 → 51.7 (19.7 → 24.1)	51.8 → 53.9 (24.2 → 25.6)	47.6 → 47.9 (22.9 → 25.5)	46.3 → 47.6 (22.2 → 24.8)	49.8 → 52.8 (22.8 → 26.5)	49.0 → 50.6 (24.8 → 24.6)	50.5 → 50.3 (24.5 → 27.1)	49.6 → 50.6 (23.1 → 25.2)
Car	30.0 → 33.2 (11.5 → 11.9)	53.5 → 54.5 (22.3 → 17.2)	54.7 → 62.6 (21.6 → 16.8)	51.3 → 52.8 (20.8 → 17.1)	26.9 → 33.6 (13.0 → 14.9)	47.0 → 50.3 (20.5 → 17.5)	51.6 → 61.5 (23.6 → 20.1)	51.7 → 52.3 (19.9 → 16.0)	45.8 → 50.1 (22.0 → 19.5)
Clock	35.0 → 40.1 (12.9 → 13.1)	33.0 → 36.8 (11.5 → 11.8)	33.2 → 36.1 (14.4 → 13.3)	32.7 → 36.0 (12.7 → 13.3)	34.2 → 37.1 (13.9 → 13.7)	35.2 → 37.7 (12.8 → 10.6)	36.7 → 39.5 (17.5 → 14.9)	33.7 → 38.1 (10.9 → 12.3)	34.2 → 37.7 (13.4 → 12.9)
Mouse	25.3 → 29.8 (13.3 → 11.3)	45.9 → 50.7 (16.4 → 12.8)	44.8 → 54.3 (18.0 → 12.3)	36.5 → 40.1 (16.4 → 13.5)	26.9 → 31.3 (15.1 → 15.3)	42.7 → 48.0 (19.3 → 13.8)	39.9 → 47.1 (16.2 → 13.7)	40.5 → 45.9 (16.8 → 12.5)	37.8 → 43.3 (17.9 → 15.5)
Hat	31.7 → 37.2 (11.5 → 9.5)	31.3 → 33.6 (11.4 → 8.6)	30.8 → 35.4 (11.0 → 10.1)	30.1 → 35.5 (12.7 → 9.8)	30.9 → 37.7 (11.1 → 9.1)	29.7 → 37.4 (10.8 → 11.1)	33.7 → 38.5 (11.5 → 10.8)	33.2 → 35.2 (11.9 → 8.3)	31.4 → 36.3 (11.5 → 9.7)
Keyboard	38.5 → 55.4 (26.2 → 21.1)	35.6 → 44.7 (21.6 → 21.1)	26.6 → 33.0 (17.7 → 17.0)	34.2 → 43.9 (24.8 → 22.3)	30.7 → 51.8 (23.7 → 26.4)	33.9 → 43.2 (22.3 → 18.7)	33.9 → 37.9 (21.0 → 16.6)	36.9 → 45.8 (27.1 → 18.1)	33.7 → 44.3 (23.2 → 21.2)
Piano	41.5 → 44.8 (16.3 → 12.3)	34.6 → 38.9 (16.9 → 13.3)	36.2 → 40.1 (11.7 → 11.5)	36.6 → 41.8 (14.2 → 10.4)	40.6 → 46.2 (15.9 → 14.1)	37.3 → 42.5 (16.8 → 13.0)	36.9 → 39.4 (10.0 → 10.8)	37.6 → 41.8 (11.8 → 11.8)	37.7 → 41.9 (14.4 → 12.3)
Remote	19.7 → 30.3 (14.7 → 21.9)	22.7 → 35.4 (18.6 → 25.2)	32.9 → 38.6 (23.7 → 27.9)	22.9 → 34.1 (18.6 → 24.2)	18.5 → 30.7 (11.1 → 19.1)	24.5 → 37.5 (14.8 → 25.0)	29.0 → 35.2 (20.4 → 26.5)	26.4 → 36.0 (18.8 → 25.6)	24.6 → 34.7 (18.3 → 24.4)
Telephone	28.5 → 36.3 (13.5 → 13.9)	28.1 → 36.2 (12.3 → 17.0)	31.6 → 37.9 (14.1 → 18.1)	28.4 → 34.0 (14.2 → 13.7)	26.4 → 34.5 (11.9 → 14.4)	27.8 → 36.7 (12.0 → 15.6)	29.7 → 35.0 (14.6 → 17.9)	29.5 → 35.5 (12.2 → 13.2)	28.7 → 35.8 (13.1 → 15.5)
Train	25.7 → 28.7 (12.8 → 12.9)	46.9 → 49.0 (23.8 → 21.2)	48.5 → 56.8 (24.4 → 20.8)	36.9 → 43.7 (20.4 → 18.5)	22.4 → 28.0 (12.2 → 12.6)	38.3 → 45.3 (20.7 → 20.3)	41.6 → 51.1 (22.3 → 20.9)	44.5 → 47.5 (23.5 → 19.5)	38.1 → 43.8 (22.2 → 20.9)
All	33.8 → 39.9 (18.4 → 17.7)	39.3 → 44.7 (20.0 → 18.6)	40.4 → 47.2 (20.3 → 19.9)	36.9 → 43.0 (19.3 → 18.1)	31.4 → 39.1 (17.7 → 19.0)	37.7 → 44.5 (19.5 → 18.2)	38.9 → 45.8 (20.0 → 19.4)	39.2 → 44.4 (19.6 → 17.6)	37.2 → 43.6 (19.6 → 18.7)

Figure 10. F-score performance and standard deviation (in parenthesis) on the 3D-ODDS dataset across the class and angle factors, before → after REFINEment. Colors correspond to accuracy after REFINEment, normalized across the table. Red indicates lower accuracy, green indicates higher. Margins correspond to Figure 11 in the main paper.

	OOWL	OTURN	OWILD	All
0°	32.8 → 39.1 (17.0 → 16.7)	33.8 → 39.1 (19.6 → 19.4)	34.9 → 41.6 (18.6 → 16.9)	33.8 → 39.9 (18.4 → 17.7)
45°	38.9 → 44.2 (20.7 → 19.4)	41.2 → 45.8 (19.5 → 18.6)	37.8 → 44.2 (19.7 → 17.7)	39.3 → 44.7 (20.0 → 18.6)
90°	40.5 → 47.5 (20.5 → 19.8)	42.1 → 47.5 (19.2 → 20.1)	38.7 → 46.6 (21.1 → 19.9)	40.4 → 47.2 (20.3 → 19.9)
135°	34.5 → 40.8 (17.7 → 17.6)	39.4 → 44.4 (19.1 → 19.0)	36.8 → 43.6 (20.7 → 17.5)	36.9 → 43.0 (19.3 → 18.1)
180°	30.9 → 37.7 (17.8 → 18.5)	32.9 → 39.6 (19.0 → 19.4)	30.4 → 39.9 (16.1 → 19.0)	31.4 → 39.1 (17.7 → 19.0)
225°	37.6 → 43.9 (19.5 → 19.2)	40.3 → 46.1 (19.2 → 18.5)	35.3 → 43.4 (19.5 → 16.8)	37.7 → 44.5 (19.5 → 18.2)
270°	38.9 → 46.5 (20.5 → 20.5)	39.6 → 44.4 (19.0 → 18.5)	38.3 → 46.3 (20.6 → 19.2)	38.9 → 45.8 (20.0 → 19.4)
315°	37.1 → 42.4 (19.1 → 17.3)	41.5 → 45.8 (18.9 → 18.2)	39.1 → 44.9 (20.7 → 17.1)	39.2 → 44.4 (19.6 → 17.6)
All	36.4 → 42.8 (19.4 → 18.9)	38.8 → 44.1 (19.4 → 19.1)	36.4 → 43.8 (19.8 → 18.1)	37.2 → 43.6 (19.6 → 18.7)

Figure 11. F-score performance and standard deviation (in parenthesis) on the 3D-ODDS dataset across the domain and angle factors, before → after REFINEment. Colors correspond to accuracy after REFINEment, normalized across the table. Red indicates lower accuracy, green indicates higher. Margins correspond to Figure 11 in the main paper.

Metric	Method	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp	Speakers	Rifle	Sofa	Table	Telephone	Watercraft	Mean
EMD ↓	AtlasNet [11]	6.3	7.9	9.5	8.3	7.8	8.8	9.8	10.2	6.6	8.2	7.8	9.9	7.1	8.0
	Mesh R-CNN [9]	4.5	3.7	4.3	3.8	4.0	4.6	5.7	5.1	3.8	4.0	3.9	4.7	4.1	4.2
	Pix2Mesh [29]	3.8	2.9	3.6	3.1	3.4	3.3	4.8	3.8	3.2	3.1	3.3	2.8	3.2	3.4
	DISN [33]	2.2	2.3	3.2	2.4	2.8	2.5	3.9	3.1	1.9	2.3	2.9	1.9	2.3	2.6
	MeshSDF [23]	3.3→2.5	2.5→2.1	3.2→3.0	2.2→2.0	2.8→**2.4**	3.0→2.4	4.2→**3.2**	3.5→2.9	2.6→**1.9**	2.7→2.4	3.1→2.7	1.9→1.7	2.9→**2.3**	3.0→2.5
	REFINEd OccNet [22]	3.0→**2.4**	2.4→**2.0**	3.1→**1.9**	2.3→**1.9**	2.8→2.6	2.4→**2.3**	5.4→3.4	4.8→**2.7**	2.5→2.4	2.8→**1.7**	3.4→**2.3**	1.3→**1.2**	2.9→2.4	2.9→**2.3**
CD-l_2 ↓	AtlasNet [11]	10.6	15.0	30.7	10.0	11.6	17.3	17.0	22.0	6.4	11.9	12.3	12.2	10.7	13.0
	Mesh R-CNN [9]	13.3	8.3	10.5	7.2	9.8	10.9	16.4	14.8	6.9	8.7	10.0	6.9	10.4	10.3
	Pix2Mesh [29]	12.4	5.5	8.2	5.6	6.9	8.2	12.3	11.2	6.0	6.8	7.9	4.7	7.9	8.0
	DISN [33]	6.3	6.6	11.3	5.3	9.6	8.6	23.6	14.5	4.4	6.0	12.5	5.2	7.8	9.7
	MeshSDF [23]	10.6→**6.3**	9.5→5.4	8.8→7.8	4.2→**3.5**	8.2→**5.9**	12.4→**7.3**	25.9→14.9	20.4→12.1	8.9→**3.4**	11.5→7.8	14.6→10.7	6.2→3.9	17.1→10.0	12.0→7.8
	REFINEd OccNet [22]	7.5→6.5	8.5→**5.3**	7.4→**5.2**	5.3→4.9	13.1→8.1	18.7→11.7	30.2→**13.1**	18.5→**10.5**	5.9→3.9	10.0→**7.1**	11.7→**8.8**	7.6→**3.5**	11.9→**9.1**	12.2→**7.5**
F-Score ↑	AtlasNet [11]	91	86	74	94	91	84	81	80	96	91	91	90	90	89
	Mesh R-CNN [9]	87	91	90	95	90	89	83	85	93	92	90	95	91	90
	Pix2Mesh [29]	88	95	94	97	94	92	89	89	95	96	93	97	94	93
	DISN [33]	94	94	89	96	90	92	78	85	96	96	87	96	93	91
	MeshSDF [23]	92→**96**	95→**97**	92→94	98→**98**	94→**97**	91→95	85→91	86→91	96→**98**	94→**96**	91→94	95→**98**	93→**95**	91→95
	REFINEd OccNet [22]	94→**96**	95→**97**	94→**97**	93→95	90→94	89→**96**	81→**92**	86→**92**	95→95	92→95	92→**95**	96→96	90→94	91→**96**
Vol. IoU ↑	AtlasNet [11]	39	34	21	22	26	36	21	23	45	28	23	43	28	30
	Pix2Mesh [29]	42	32	66	55	40	49	32	60	40	61	40	66	40	48
	DISN [33]	58	53	52	74	**54**	**56**	35	55	**59**	66	48	**73**	**56**	57
	REFINEd OccNet [22]	57→**59**	49→**55**	73→**73**	73→**74**	50→51	47→49	37→**43**	65→**65**	47→49	68→**69**	51→**52**	72→72	53→54	57→**59**

Table 1. Extended, per-class results for reconstruction accuracy with no domain shift. Corresponds to Table 3 in the main paper.

	REFINEd OccNet [22]				REFINEd Pix2Mesh [29]				REFINEd AtlasNet [11]			
	EMD ↓	CD-l_2 ↓	F-Score ↑	Vol. IoU ↑	EMD ↓	CD-l_2 ↓	F-Score ↑	Vol. IoU ↑	EMD ↓	CD-l_2 ↓	F-Score ↑	Vol. IoU ↑
Airplane	3.5 → **2.2**	20.6 → **11.4**	86 → **91**	38 → **40**	3.7 → **2.3**	22.3 → **11.0**	65 → **88**	12 → **22**	5.3 → **3.8**	41.9 → **18.2**	60 → **82**	5 → **13**
Bench	2.9 → **2.2**	28.6 → **17.0**	84 → **86**	20 → **20**	3.6 → **2.6**	28.0 → **19.9**	65 → **76**	9 → **11**	4.9 → **4.6**	50.0 → **37.7**	58 → **68**	5 → **8**
Cabinet	3.4 → **2.7**	17.0 → **14.8**	83 → **85**	45 → **46**	3.6 → **3.0**	20.2 → **16.4**	74 → **78**	37 → **39**	4.3 → **4.1**	30.7 → **19.9**	59 → **75**	14 → **17**
Car	2.9 → **2.5**	19.9 → **12.9**	86 → **87**	30 → **31**	2.7 → **2.3**	10.8 → **7.8**	85 → **90**	24 → **27**	7.6 → **4.8**	98.8 → **27.0**	44 → **72**	6 → **12**
Chair	6.5 → **5.4**	48.5 → **39.4**	72 → **76**	29 → **32**	6.3 → **4.5**	35.4 → **25.2**	60 → **73**	17 → **22**	6.8 → **5.0**	49.5 → **27.3**	53 → **71**	8 → **13**
Display	3.5 → **2.7**	30.8 → **18.1**	76 → **83**	31 → **37**	4.2 → **3.0**	28.0 → **17.4**	72 → **81**	25 → **32**	4.9 → **4.5**	43.1 → **30.0**	61 → **71**	10 → **14**
Lamp	8.9 → **6.3**	90.5 → **59.1**	68 → **73**	22 → **23**	9.2 → **7.0**	71.6 → **40.6**	50 → **66**	11 → **14**	10.2 → **7.5**	102.4 → **51.1**	44 → **62**	5 → **10**
Speakers	4.4 → **3.6**	29.8 → **22.3**	73 → **76**	43 → **44**	4.3 → **3.8**	31.4 → **25.5**	65 → **70**	36 → **38**	5.4 → **4.7**	46.6 → **27.7**	55 → **69**	13 → **17**
Rifle	6.5 → **3.9**	37.7 → **14.6**	86 → **91**	30 → **30**	3.5 → **3.4**	18.1 → **10.1**	76 → **91**	12 → **21**	6.3 → **4.5**	61.4 → **28.6**	70 → **84**	7 → **14**
Sofa	3.0 → **2.7**	23.8 → **17.9**	83 → **85**	48 → **49**	4.3 → **3.2**	24.8 → **21.8**	71 → **79**	34 → **40**	5.3 → **4.7**	48.0 → **31.1**	63 → **73**	15 → **19**
Table	4.5 → **3.9**	40.6 → **34.3**	72 → **77**	17 → **20**	9.3 → **6.2**	159.3 → **81.8**	30 → **44**	6 → **8**	8.9 → **7.4**	129.6 → **82.7**	36 → **47**	4 → **8**
Telephone	2.3 → **2.0**	10.9 → **8.0**	90 → **92**	48 → **50**	2.2 → **1.8**	10.9 → **8.2**	89 → **92**	40 → **44**	3.4 → **3.3**	33.6 → **20.8**	66 → **79**	11 → **16**
Watercraft	4.3 → **2.9**	42.4 → **23.5**	80 → **86**	32 → **36**	5.0 → **2.7**	32.7 → **14.0**	71 → **86**	16 → **27**	7.1 → **4.3**	76.8 → **25.5**	55 → **79**	6 → **15**
Mean	4.3 → **3.3** **(-1.0)**	34.0 → **22.5** **(-11.5)**	80 → **84** **(+4)**	33 → **35** **(+2)**	4.8 → **3.5** **(-1.3)**	38.0 → **23.1** **(-14.9)**	67 → **78** **(+11)**	22 → **27** **(+5)**	6.2 → **4.9** **(-1.3)**	62.5 → **32.9** **(-29.6)**	56 → **72** **(+16)**	8 → **13** **(+5)**

Table 2. REFINEment in the presence of mild domain shift, namely RerenderedShapeNet reconstructions by ShapeNet trained networks OccNet, Pix2Mesh, and AtlasNet. REFINE achieves gains under all networks, classes, and metrics. Corresponds to the first 3 rows of Table 4 in the main paper.

	REFINEd Pix2Vox [32]			
	EMD ↓	CD-l_2 ↓	F-Score ↑	Vol. IoU ↑
Airplane	4.5 → **2.3**	19.7 → **7.3**	71 → **93**	19 → **38**
Bench	2.9 → **2.5**	25.3 → **16.5**	72 → **80**	12 → **16**
Cabinet	2.8 → **2.8**	17.0 → **15.5**	79 → **80**	43 → **43**
Car	3.2 → **2.5**	26.3 → **14.6**	80 → **85**	29 → **32**
Chair	5.3 → **3.5**	30.2 → **18.4**	64 → **79**	23 → **32**
Display	3.9 → **3.2**	33.1 → **20.4**	71 → **80**	28 → **34**
Lamp	9.6 → **6.1**	78.0 → **44.6**	53 → **65**	18 → **23**
Speakers	3.5 → **3.5**	27.5 → **22.0**	72 → **75**	42 → **44**
Rifle	4.8 → **3.0**	23.2 → **12.2**	83 → **92**	25 → **35**
Sofa	4.1 → **3.2**	32.4 → **20.2**	72 → **82**	43 → **50**
Table	6.8 → **5.4**	121.3 → **62.5**	35 → **51**	8 → **11**
Telephone	2.1 → **2.1**	20.3 → **14.6**	79 → **85**	34 → **38**
Watercraft	5.1 → **2.7**	30.3 → **15.2**	75 → **87**	26 → **42**
Mean	4.5 → **3.3** **(-1.2)**	37.3 → **21.8** **(-15.5)**	70 → **80** **(+10)**	27 → **34** **(+7)**

Table 3. REFINEment in the presence of mild domain shift, namely RerenderedShapeNet reconstructions by a ShapeNet trained Pix2Vox Network. REFINE achieves gains under all classes and metrics. Corresponds to the last row of Table 4 in the main paper.

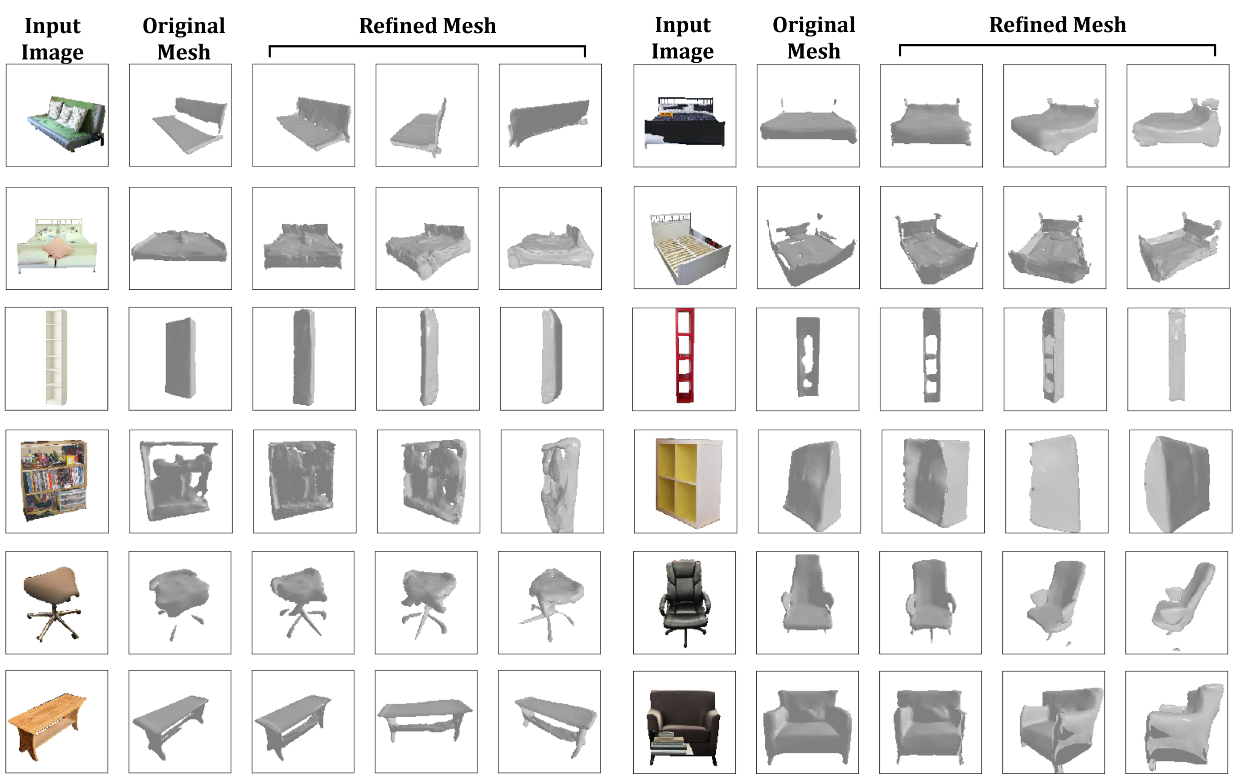

Figure 12. Occupancy Network mesh REFINEments for Pix3D images in the bed, bookcase, and chair classes.

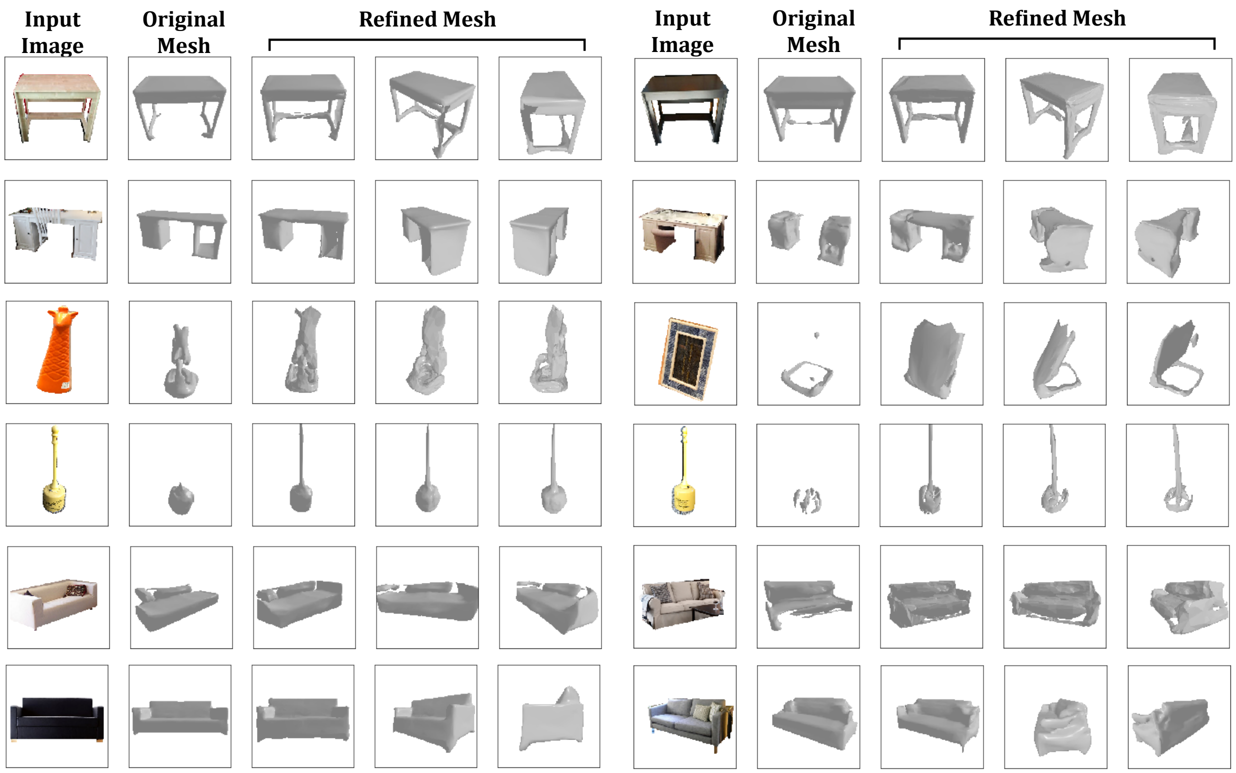

Figure 13. Occupancy Network mesh REFINEments for Pix3D images in the desk, misc, and sofa classes.

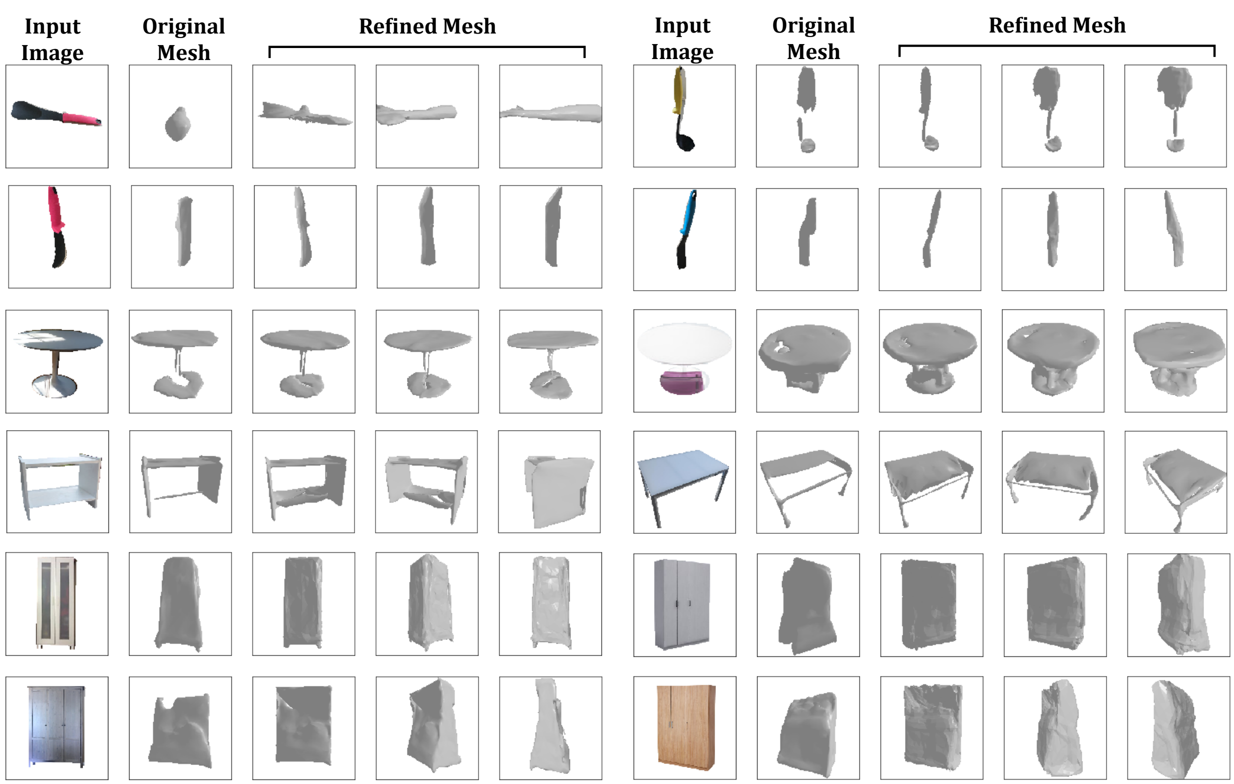

Figure 14. Occupancy Network mesh REFINEments for Pix3D images in the tool, table, and wardrobe classes.

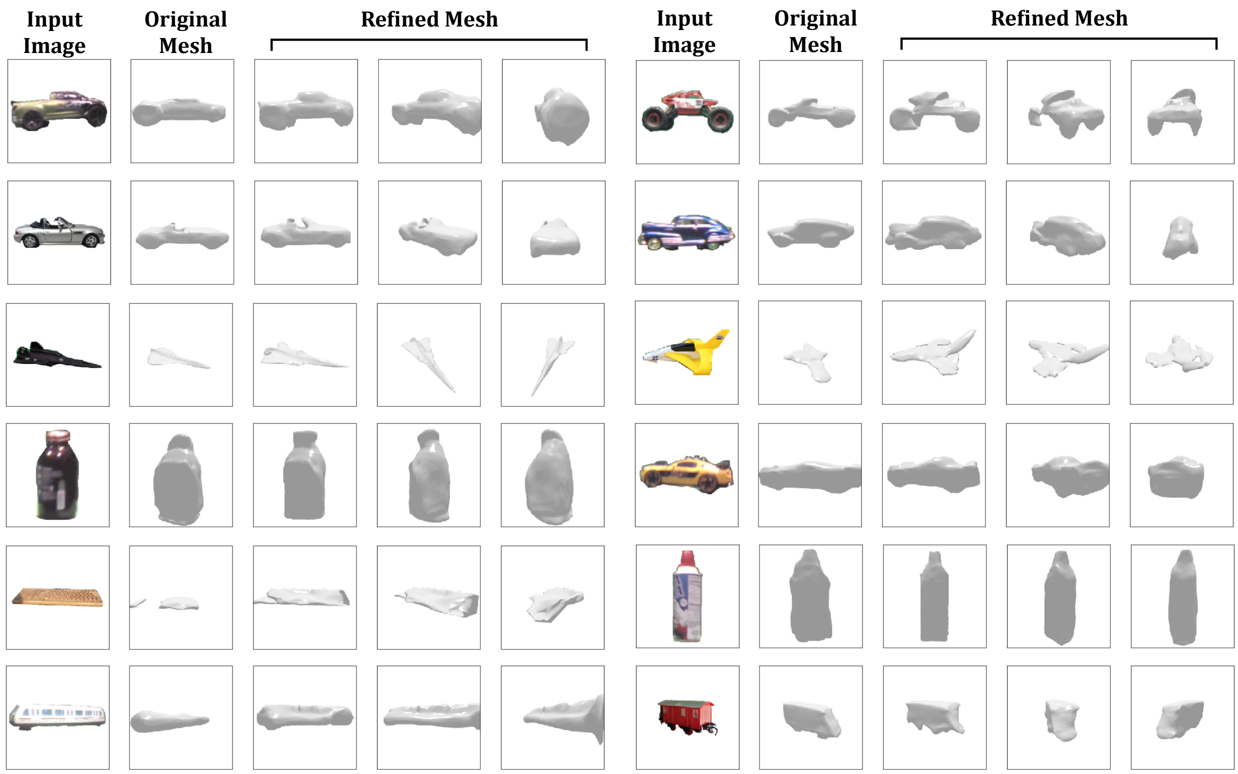

Figure 15. Occupancy Network mesh REFINEments for example 3D-ODDS images.

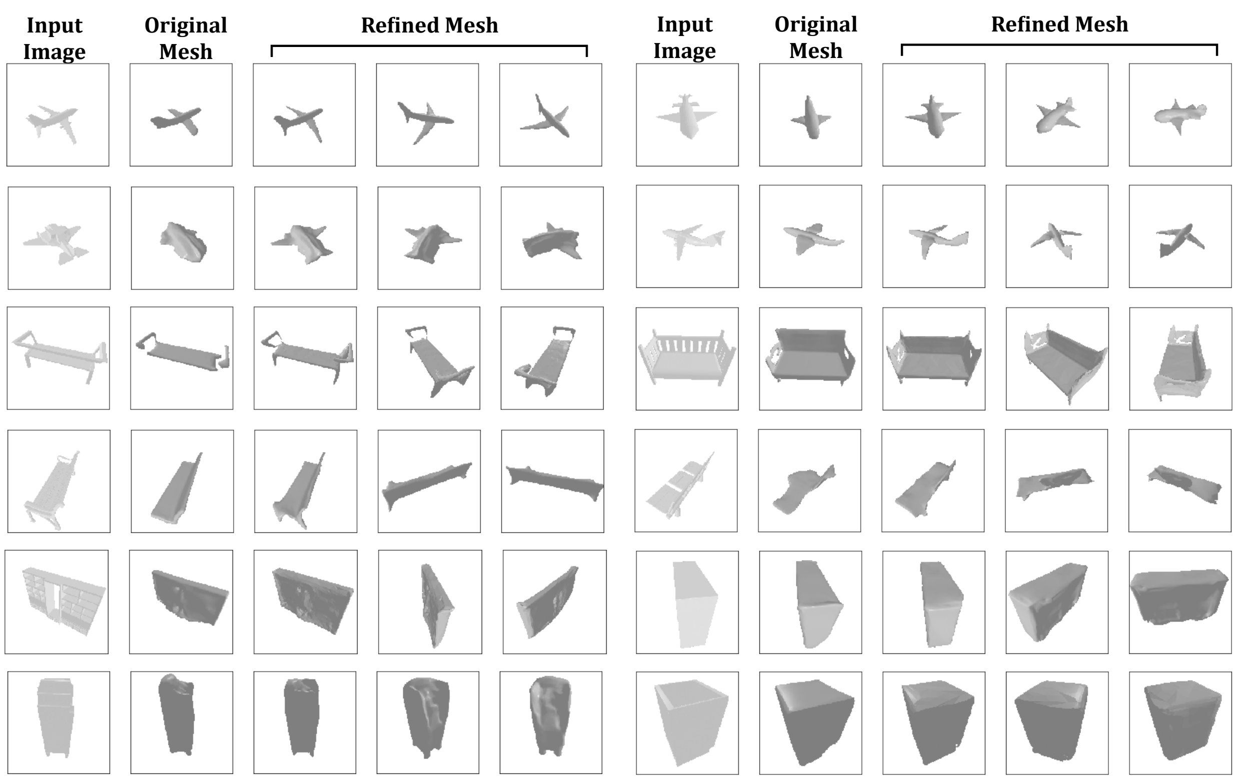

Figure 16. Occupancy Network mesh REFINEments for RerenderedShapeNet images in the airplane, bench, and cabinet classes.

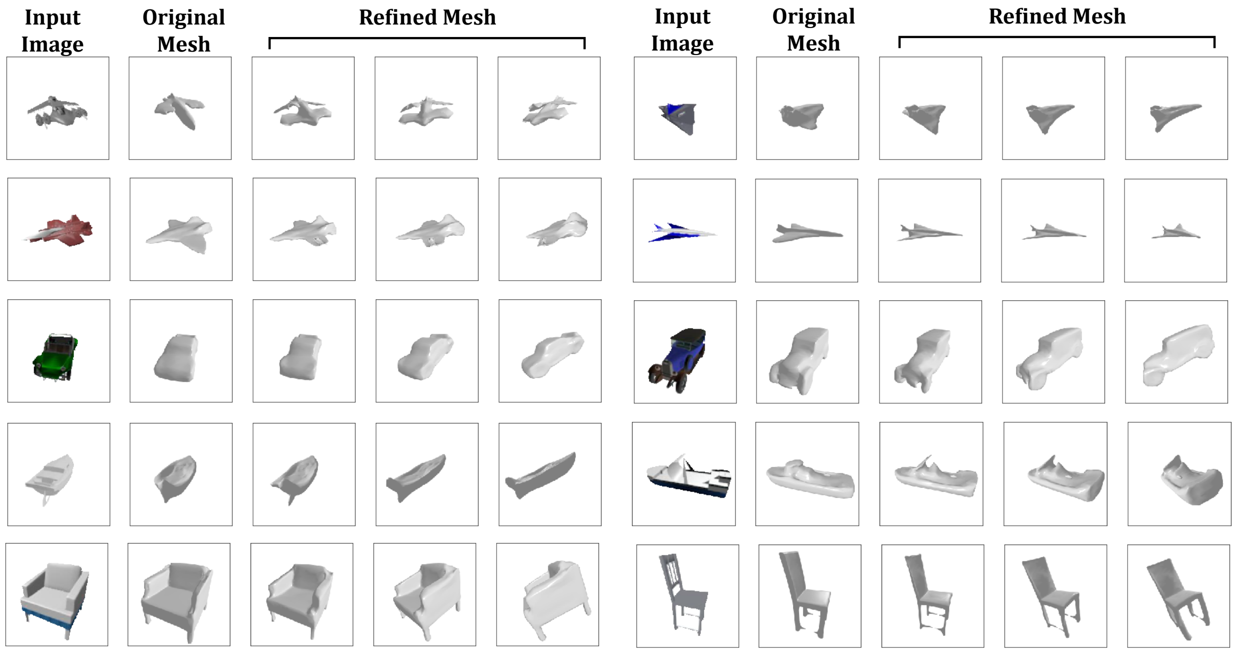

Figure 17. Occupancy Network mesh REFINEments for several ShapeNet images.

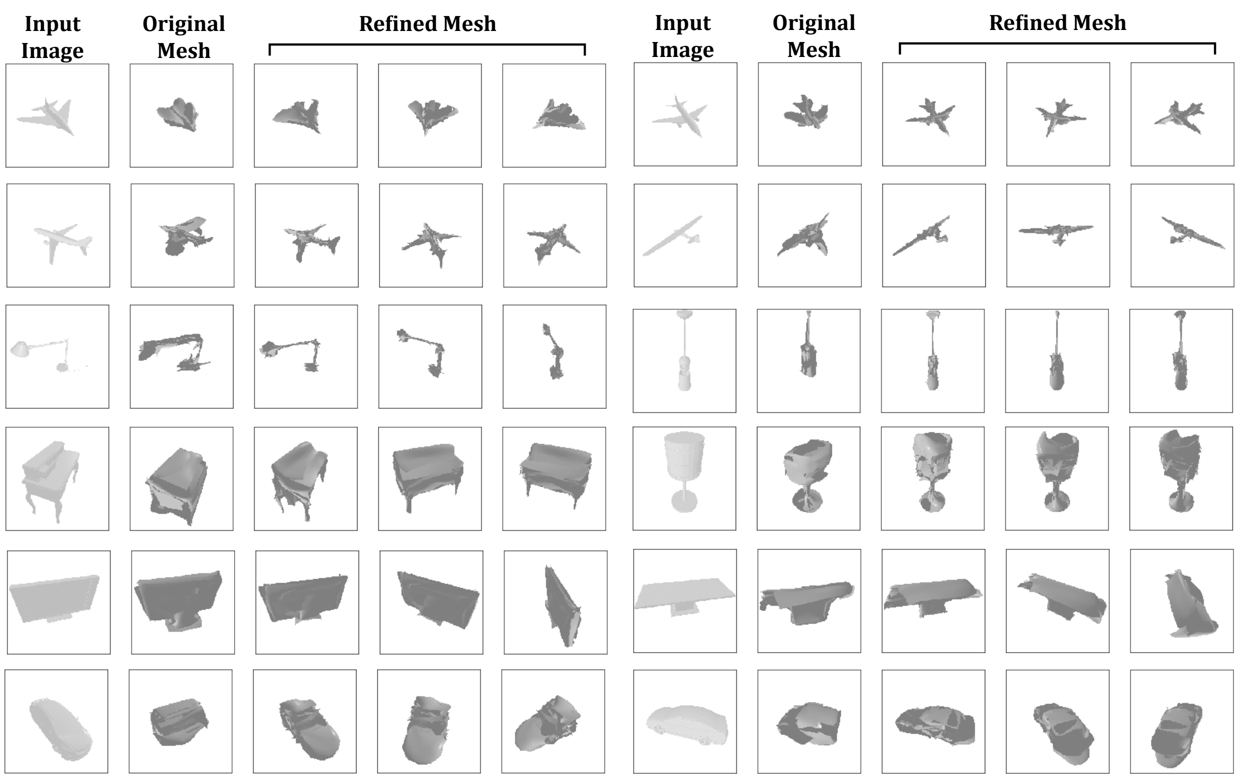

Figure 18. AtlasNet mesh REFINEments for several RerenderedShapeNet images.

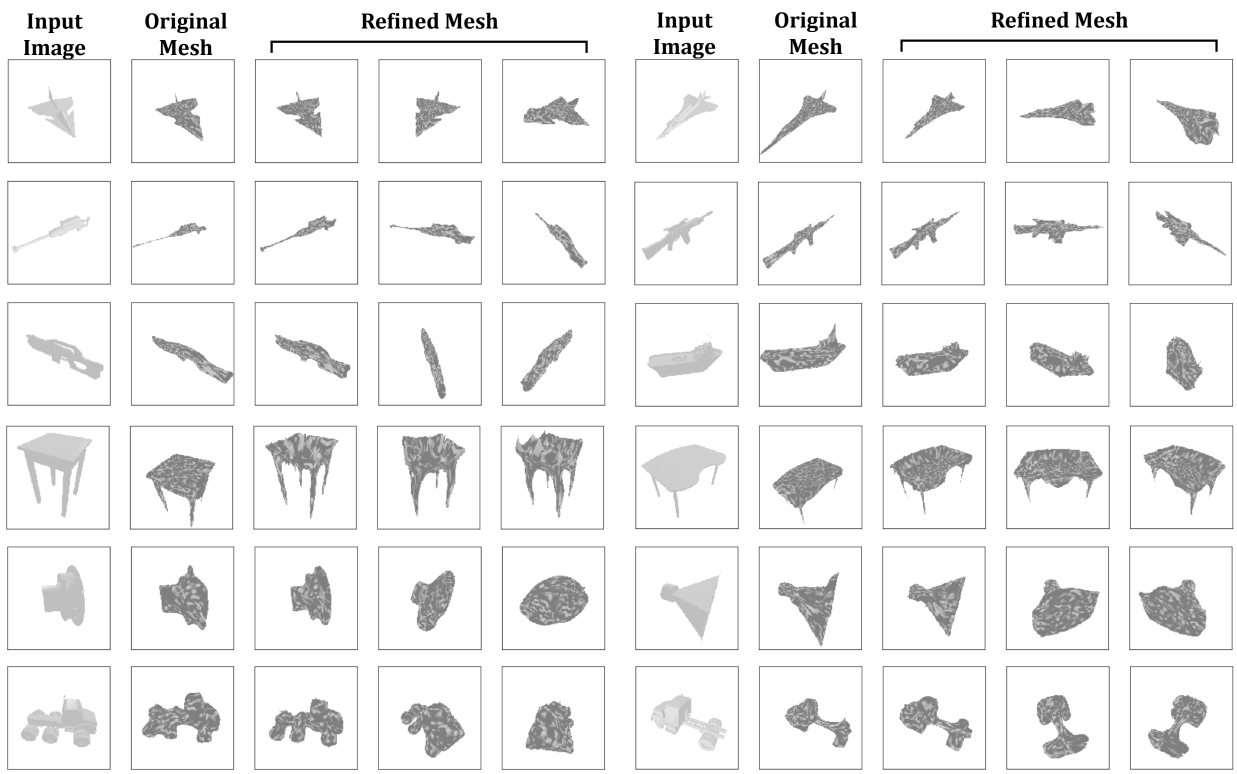

Figure 19. Pix2Mesh mesh REFINEments for several RerenderedShapeNet images.

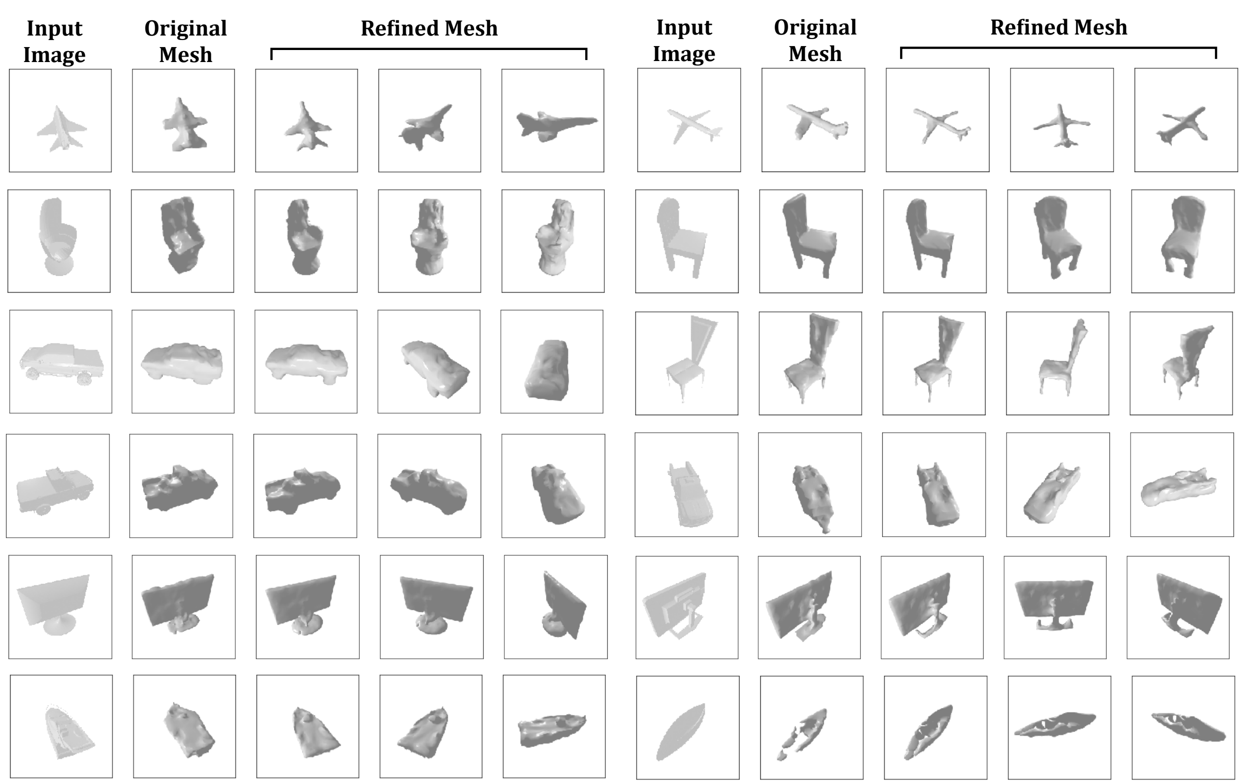

Figure 20. Pix2Vox mesh REFINEments for several RerenderedShapeNet images.

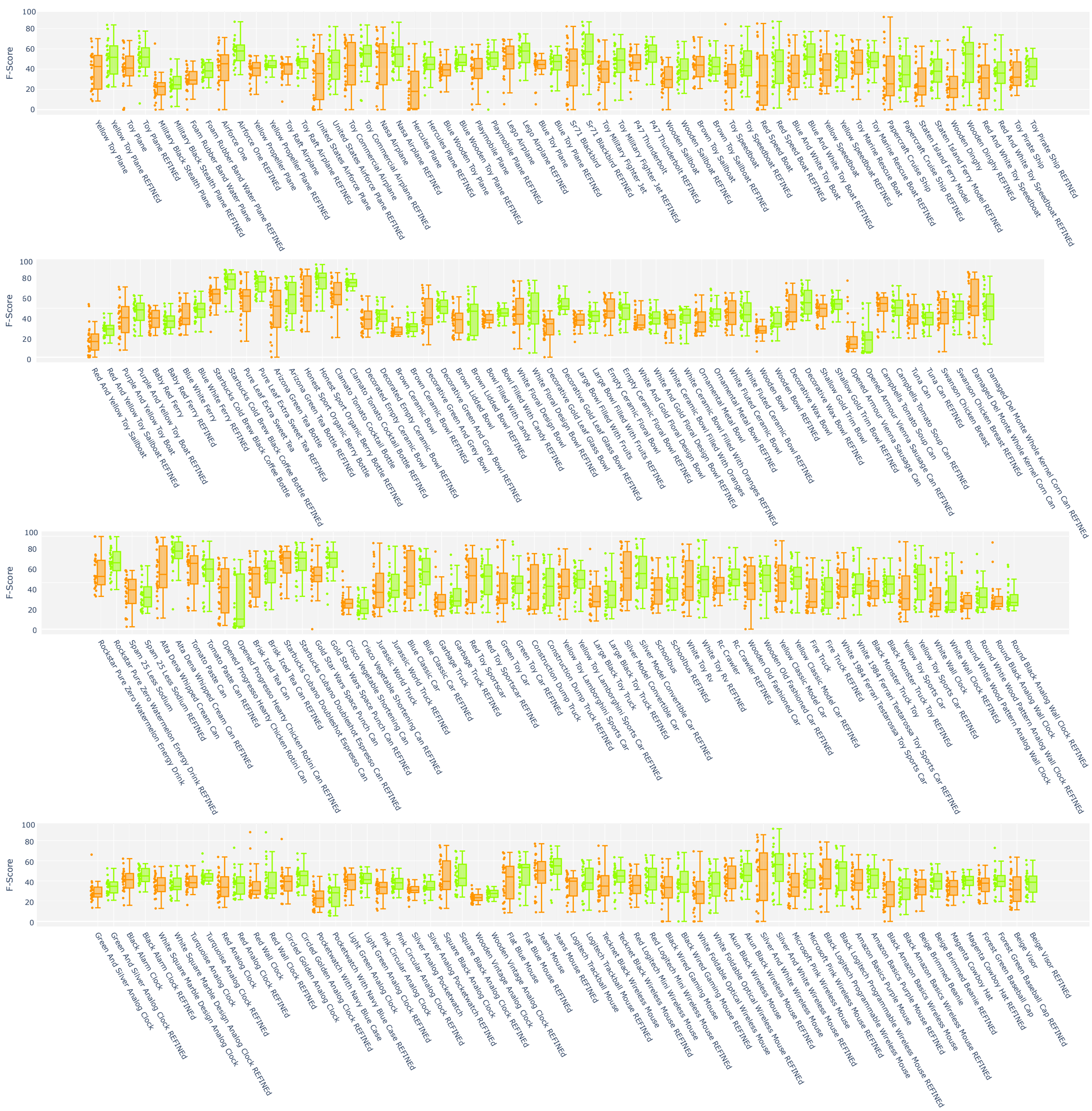

Figure 21. 3D-ODDS objects have 24 images (3 domains, 8 viewpoints). Reconstruction accuracies plotted before (after) REFINE as orange (green). Generally, REFINE improves performance invariance. Extended version of Figure 10 in the main paper.

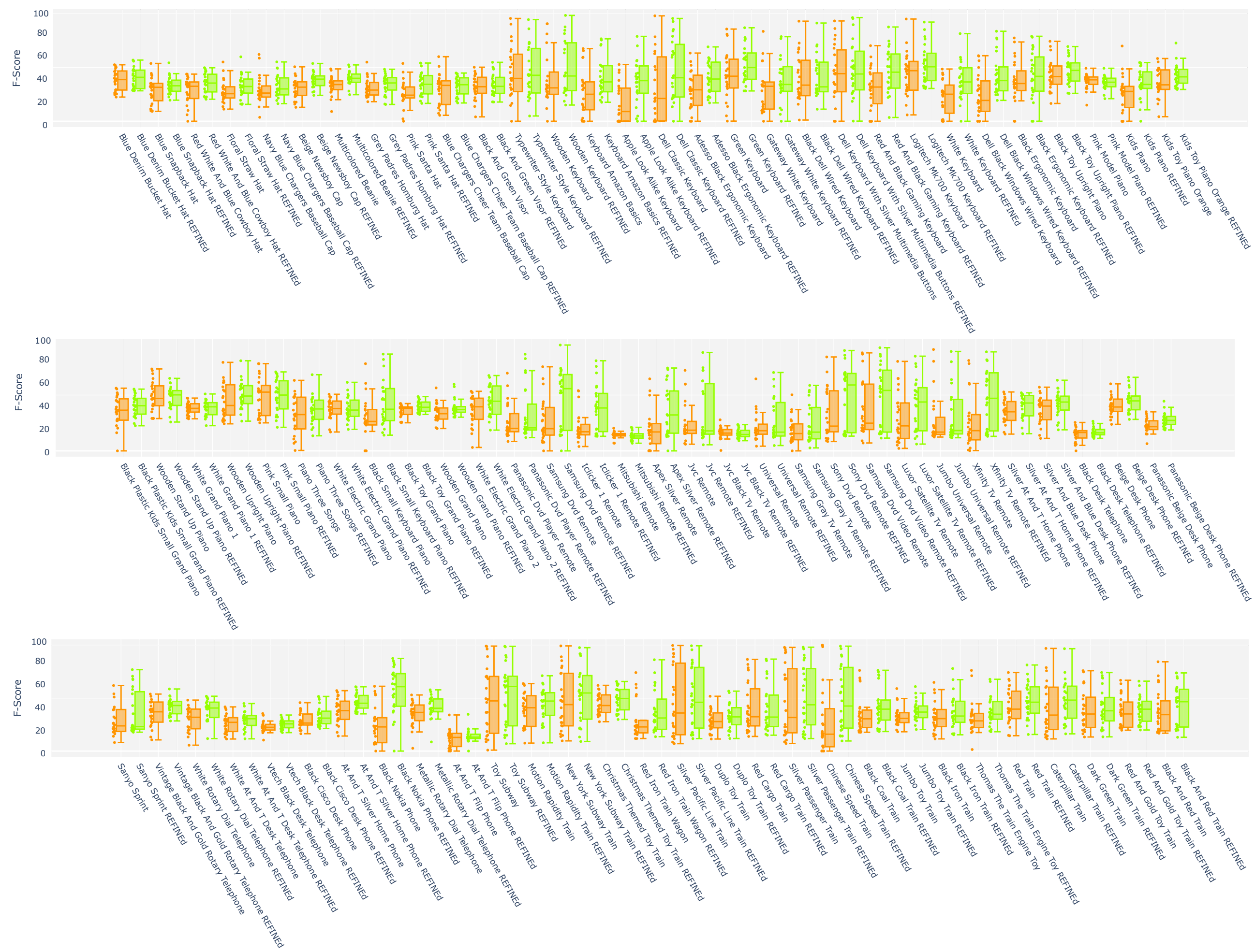

Figure 22. 3D-ODDS objects have 24 images (3 domains, 8 viewpoints). Reconstruction accuracies plotted before (after) REFINE as orange (green). Generally, REFINE improves performance invariance. Extended version of Figure 10 in the main paper.