Supplementary Material for ShapeClipper: Scalable 3D Shape Learning from Single-View Images via Geometric and CLIP-based Consistency

Zixuan Huang[1] Varun Jampani[2] Anh Thai[1] Yuanzhen Li[2]
Stefan Stojanov[1] James M. Rehg[1]
[1]Georgia Institute of Technology, [2]Google Research

A. Generalization performance

Performance on unseen categories. We train on 6 categories (car, chair, diningtable, motorbike, train, tvmonitor) of Pascal3D+ and test on other unseen categories (see Fig. 1 (a)). We find our model can generalize to categories that are highly related to at least one training category (e.g. sofa - average CD 0.571), and does not generalize as well to categories less related to training categories (e.g. bottle - average CD 1.020).

We further quantify the relationship between generalization performance and semantic relevance. We analyze the correlation between reconstruction error and minimum CLIP distance from each test sample to training images. As in Fig. 2 (b), there exists a clear positive correlation (Pearson coefficient $\rho = 0.53$) between the two variables. This verifies our model often generalizes better to samples that are more semantically related to the seen categories.

Figure 1. (a) Reconstruction on unseen categories. (b) Inference of our Pix3D-trained model on CO3D chairs.

Direct inference on in-the-wild data. Our method can also reconstruct faithful shapes under reasonable domain gaps. We test our Pix3D model directly on CO3D chair images (without fine-tuning) and find most reconstructions are reasonable. See Fig. 1 (b) for examples.

B. Additional Analysis of the Model

Viewpoint robustness of CLIP embeddings. We quantitatively evaluate the viewpoint robustness of CLIP embeddings on Pix3D chairs. In our experiments, CLIP can find a significant number of neighbors with distinct viewpoints (see Fig. 2 (a)). Geodesic distance is the minimal angular difference between two rotations. The average geodesic distance from query images to CLIP neighbors is 64°, and 67% of the query images have at least one neighbor with distinct pose (at least 90°away).

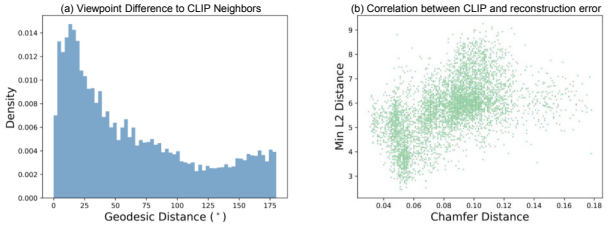

Figure 2. (a) Histogram of viewpoint distance from query images to 5 CLIP neighbors on Pix3D. (b) Correlation between reconstruction error and minimum CLIP distance on unseen categories.

Robustness to corrupted masks. To evaluate the performance under corrupted masks, we replace the Pix3D masks with masks corrupted by perlin noise and then train/test our models under different level of pixel corruption percentages. Under 0/5/10/20/30/50% corruption level, the chamfer distance is 0.612/0.626/0.632/0.651/0.689/0.763 respectively. This experiment shows that our model is not significantly affected by mild to moderate mask inaccuracy.

Performance with GT viewpoints. We further evaluate our model when GT viewpoints are given during training. An image can be explained by infinite combinations of shapes and viewpoints. When GT viewpoints are given, such entanglement is resolved and the learning will be much easier. Our model trained with GT viewpoints obtains average CD of 0.418 on Pix3D (vs. 0.612 w/o GT viewpoint).

Additional discussion on retrieval methods. Comparing reconstruction methods to retrieval methods has been one of the central topics in the area of single-view shape reconstruction [3]. Based on the finding about CLIP's re-

lationship to shape in our paper, it would be natural to consider the retrieval baseline using CLIP. While retrieving shapes with CLIP is an interesting direction, we would like to emphasize that it is not directly comparable to our proposed reconstruction method. Retrieval methods require a **large paired** image-3D shape database, similar to the non-scalable fully supervised 3D reconstruction setting. In contrast, our method only requires single-view 2D images during training, allowing it to learn to reconstruct objects from datasets like OpenImages for which there are no paired 3D shapes. Such datasets without any geometric annotations are our main application domain, and retrieval methods cannot be applied to these datasets.

C. Additional Implementation Details.

Implicit representation and rendering. The surface representation and texture rendering follow [4, 6]. We use an implicit SDF field and convert it to densities for volumetric rendering. The conversion from SDF to densities is done via the Cumulative Distribution Function (CDF) of the Laplace distribution:

$$\sigma(s) = \begin{cases} \frac{1}{\beta} \cdot \frac{1}{2} \exp(\frac{s}{\beta}) & \text{if } s \leq 0 \\ \frac{1}{\beta} \cdot (1 - \frac{1}{2} \exp(-\frac{s}{\beta})) & \text{if } s > 0 \end{cases}, \tag{1}$$

where s is the SDF. Because our focus is shape, and learning radiance is challenging without viewpoint annotation, we represent texture as an RGB field without any view-dependency. The surface normal rendering follows MonoSDF [6], where the local normal vectors are estimated by the gradient of the SDF field and aggregated via the standard volume rendering.

Uniform viewpoint prior. We use a uniform prior to regularize the viewpoint learning, which helps to prevent the rotation estimation from degeneration. Specifically, for each minibatch during training, we estimate the empirical distribution of the predicted azimuth. We then minimize the Earth-Mover Distance (EMD) between the empirical azimuth distribution and a uniform prior within $[0°, 360°]$.

D. Additional Results on OpenImages

We show more reconstruction results of our model trained on all 53 OpenImages [2] categories used in [5]. We further compare with SS3D [1] qualitatively, where our model demonstrates state-of-the-art reconstruction performance. Note the training is category-specific and the performance may be further improved by training a joint model via distillation [1], which is parallel to our research direction here.

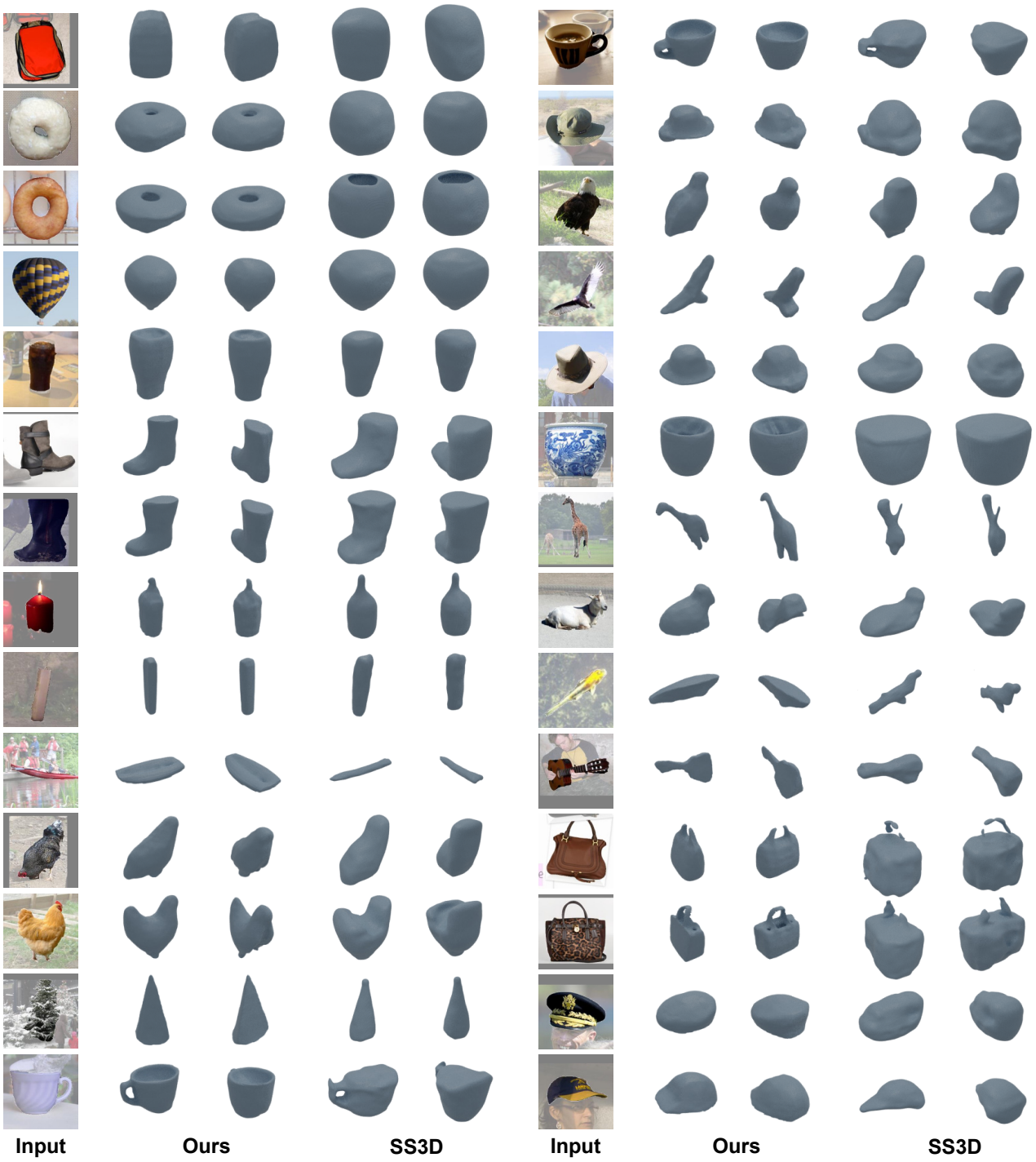

Figure 3. Additional qualitative results and comparison on full OpenImages.

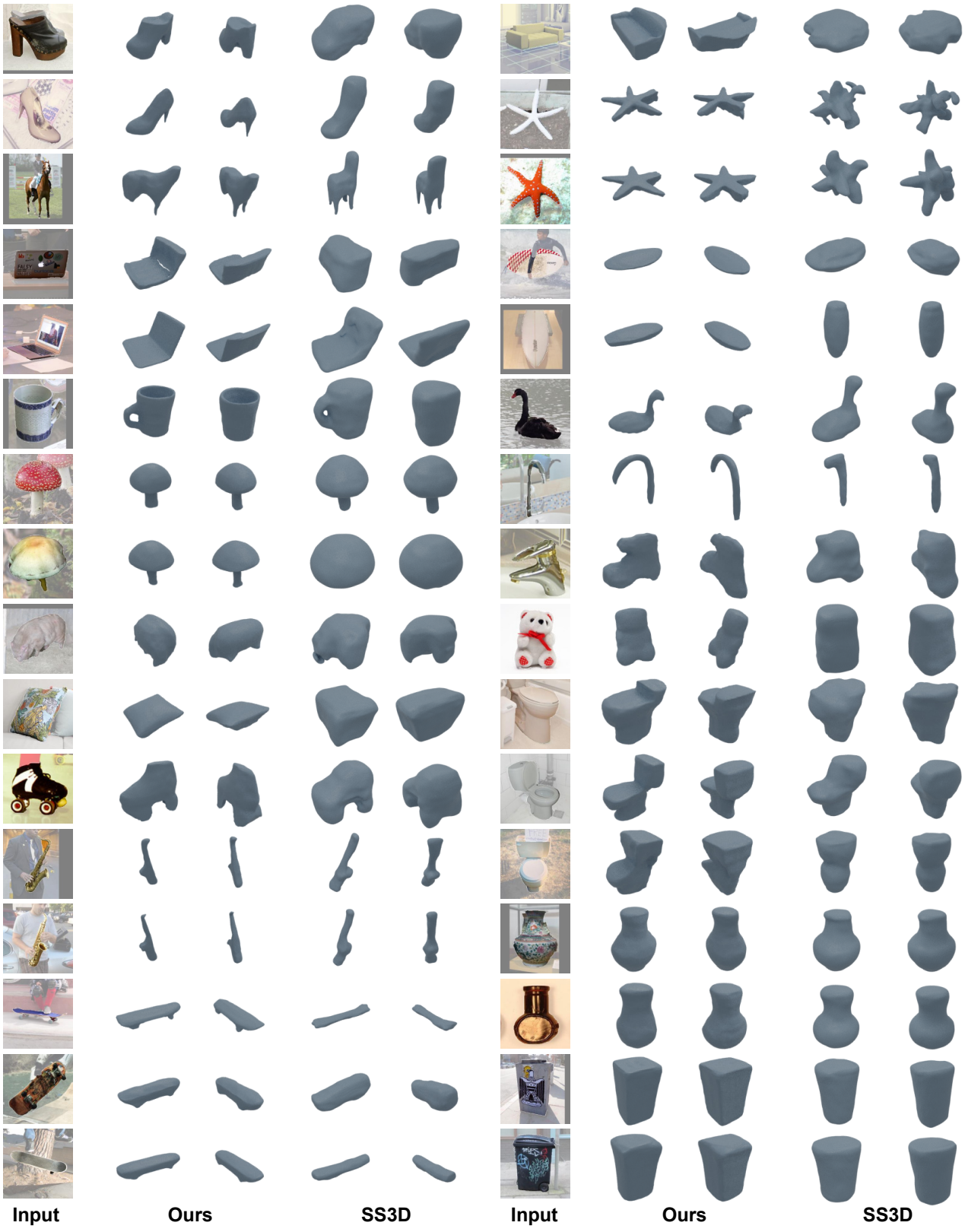

Figure 4. Additional qualitative results and comparison on full OpenImages.

References

[1] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2022. 2

[2] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 2

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 1

[4] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[5] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021. 2

[6] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2