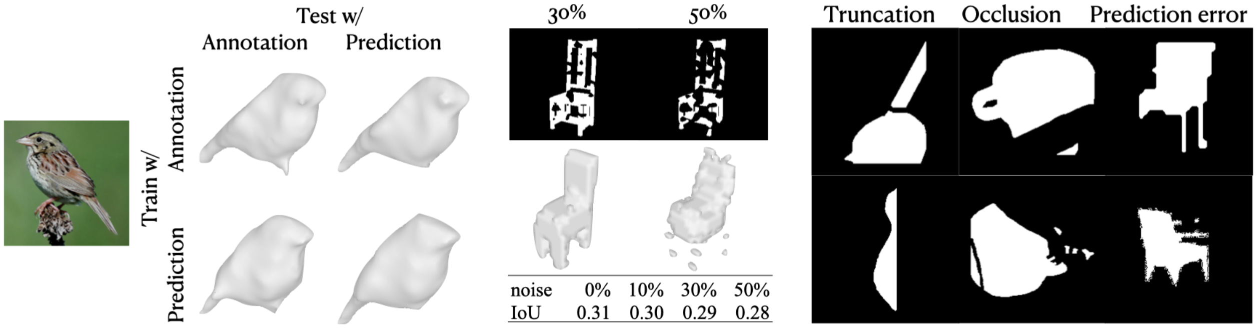

Figure 11. Left: qualitative results on CUB replacing annotation with prediction in training and/or test time. Middle: quantitative results on ShapeNet-chairs by adding random noise on masks during both training and inference. Right: masks used to train the models in the paper.

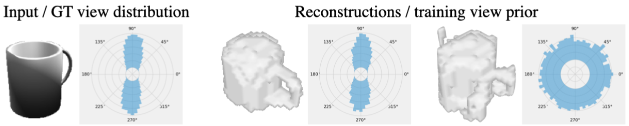
Input / GT view distribution
Reconstructions / training view prior

Figure 12. Results on training models with different viewpoint priors.

6. Ablation Study

Assumption of viewpoint distribution. We briefly analyze the effect of viewpoint prior. In figure 12 we visualize volumetric reconstruction training with different viewpoint prior on the mug category of ShapeNet. While our method is robust to some view distribution mismatch, the shapes display artifact (*e.g.* two handles) when the assumed prior is far from the ground-truth viewpoint distribution. It is because different viewpoint distribution may induce different 3D shapes as the adversarial loss matches its projections with the existing image collections. We notice similar artifacts when training on the real datasets (*e.g.* starfish and mugs on OpenImages), as camera pose biases exist by human photographers (*e.g.* front view of starfish or mugs with handles). While we assume azimuth from uniform distribution across all experiments and have achieved some promising results on various categories, we encourage more works to explore the direction of better viewpoint distribution prior.

Robustness against segmentation quality. Our model depends on the segmentation quality, as it is the only supervision. We ablate our model with noisy masks, both qualitatively and quantitatively. The model trained/tested with predictions from [19] (left) or with synthesized noise (mid) performs comparably to using GT, until considerably severe corruption. Our experiments in paper have already suggested that our model is robust to the noise as masks might be truncated, occluded, or corrupted due to prediction error (Fig 11 right). We also visualized the masks used in the main paper (Fig 11 right). Our experiments suggest that our model is robust to the noise as masks might be truncated, occluded, or corrupted due to prediction error.

7. Architecture Details

Neural Network Architecture. The encoder is comprised of 4 convolution blocks followed by two heads to output v and z. Each block consists of $Conv(3 \times 3) \rightarrow LeakyReLU$. The feature from the last block is fed to 2 fully-connected layers to get v and is fed to Average Pooling with another fully-connected layer to output z. v is in 2-dim to represent azimuth and elevation while the dimensionality of latent variable z is 128.

The decoder follows StyleGAN[18] to use the latent variable z as a "style" parameters to stylize a constant 256×4^3 feature. Given z, the constant is upsampled to the implicit 3D feature S_f by a sequence of style blocks. Then S_f is transformed to get the occupancy grid S_o by a $3 \times 3 \times 3$ Deconv layer with Sigmoid activation. Among all of our experiments, our decoder consists of 2 style blocks each of which are built with $Deconv \rightarrow AdaIN \rightarrow LeakyReLU$. The shape of S_f is 64×16^3 and the shape of S_o is 1×32^3.

Training Details. We optimize the losses with Adam [20] optimizer in learning rate 10^{-4}. The learning rate is scheduled to decay linearly after 10k iterations, following prior work [57]. We weight the losses such that they are around the same scale at the start of training. Specifically, we use $\lambda = 10$ for $\mathcal{L}_{pixel} + \mathcal{L}_{perc}$, 1 for $\mathcal{L}_{adv}$ and $\mathcal{L}_{content}$. The volumetric reconstruction network is optimized for 80k. Due to the diverse appearance and data noise on Quadrupeds, we additionally regularize the network by an L2 distance between the predicted voxels and the mean shape of all quadrupeds. The model can still capture the articulation for different instance.

8. F-score Calculation

In order to calculate F-score – the harmonic mean of recall and precision, the meshes are first converted to point cloud by uniformly sampling from surfaces. The recall is considered as the percentage of ground-truth points whose nearest neighbour in predicted point cloud is within a threshold while the precision is calculated as the other prediction-to-target way.