Supplementary for Category Level Object Pose Estimation via Neural Analysis-by-Synthesis

Xu Chen[1,3⋆], Zijian Dong[1⋆], Jie Song[1], Andreas Geiger[2,4], and Otmar Hilliges[1]

[1] ETH Zürich
[2] University of Tübingen
[3] Max Planck ETH Center for Learning Systems
[4] Max Planck Institute for Intelligent Systems, Tübingen

1 Implementation Details

The deep neural network architectures for the encoder, the 3D decoder and the 2D decoder are illustrated in Fig. 1. We set the output resolution of the image generation network to 64×64 pixels. For each category, we train the network for 10 epochs using Adam with a learning rate of $3e^{-2}$, which takes 8 hours on a GTX 1080Ti GPU. At inference time, we also use Adam and optimize for 50 iterations which we empirically found sufficient for our algorithm to converge. We initialize with 32 random seeds.

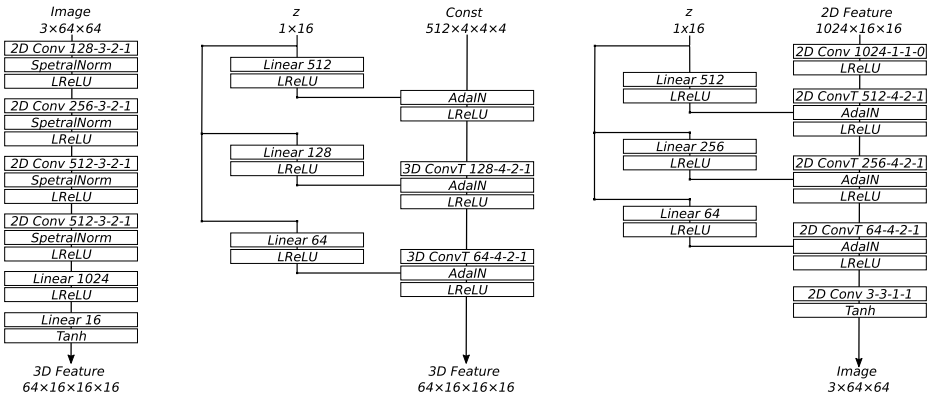

Fig. 1: **Network Architectures.** The parameters for the convolution and transposed convolution layers are specified in the following order: number of filters, filter size, stride and padding size.

⋆ Equal contribution.

2 Qualitative Results

Additional qualitative results on the iterative pose and appearance update procedure as well as pose-aware image synthesis are provided in Fig. 2 and Fig. 3.

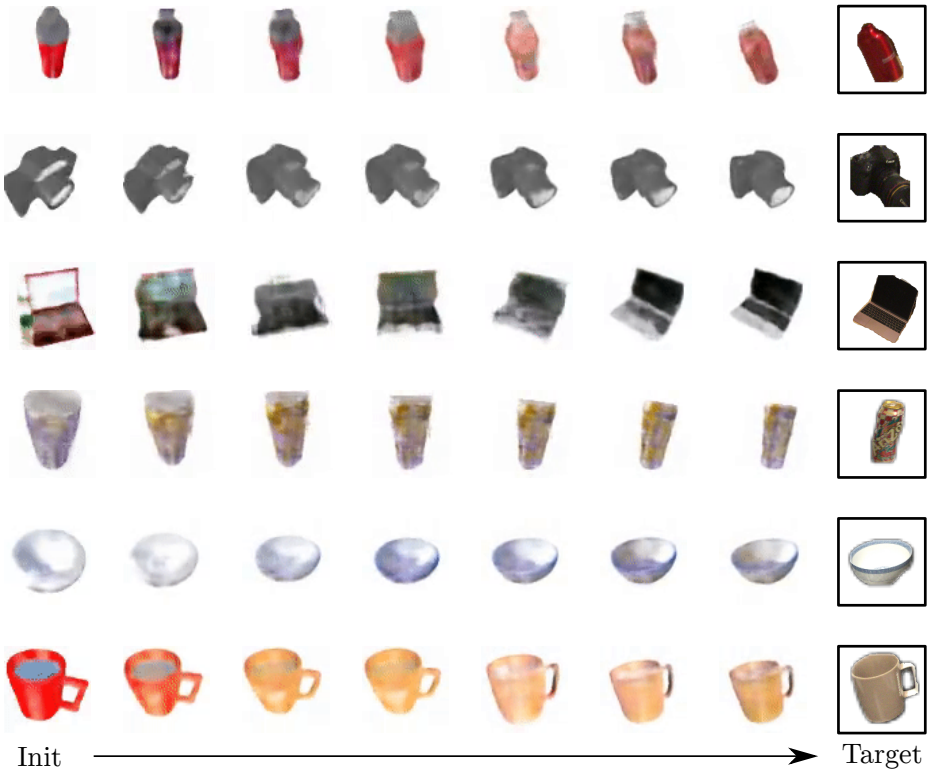

Fig. 2: **Qualitative Results for Pose and Appearance Fitting.** Starting from the initialization (left), our method iteratively updates the pose and appearance parameters to generate an image that aligns with the target (right).

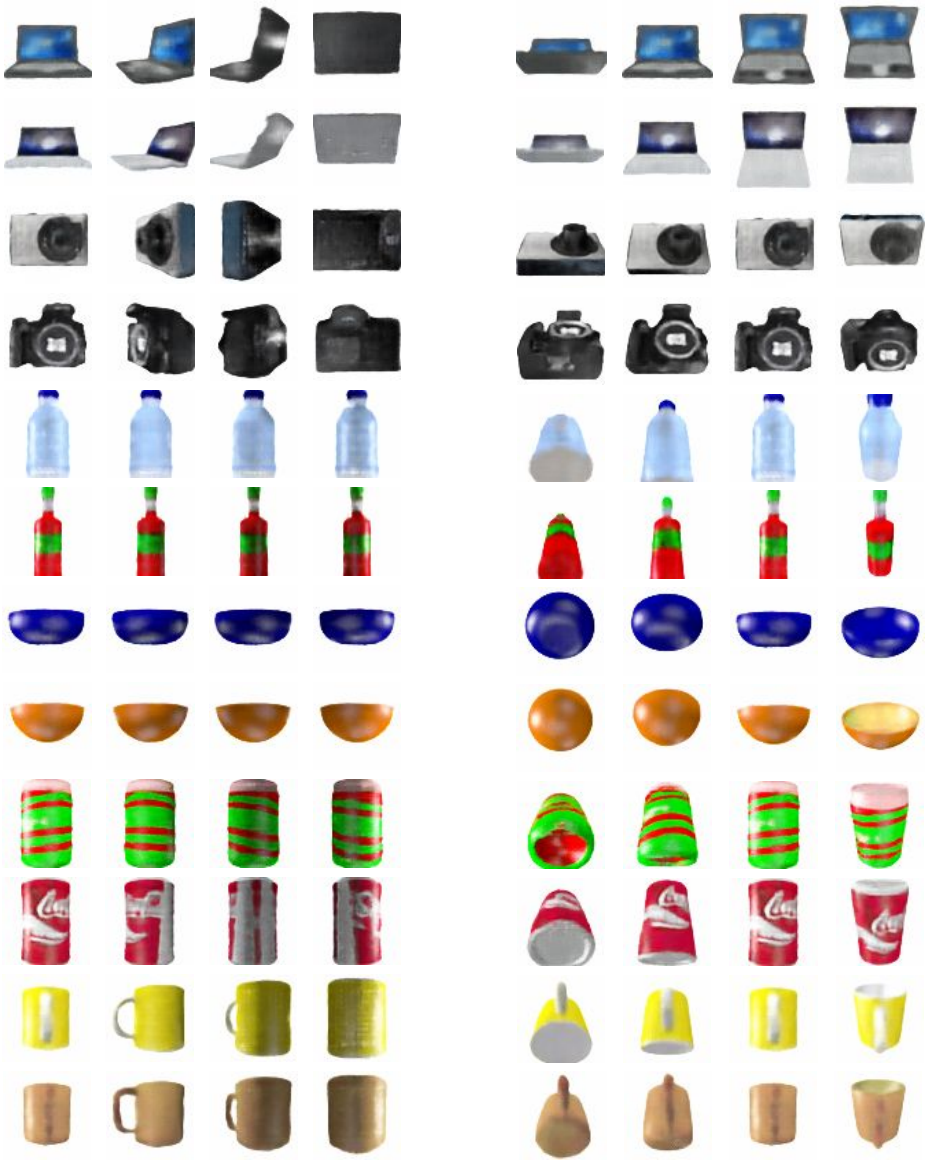

Fig. 3: **Qualitative Results for Pose-Aware Image Synthesis.** We generate images for different azimuth *(column 1-4)* and elevation *(column 5-8)* angles for all six categories: laptop *(row 1-2)*, camera *(row 3-4)*, bottle *(row 5-6)*, bowl *(row 7-8)*, can *(row 9-10)* and mug *(row 11-12)*.