NOPE: Novel Object Pose Estimation from a Single Image Supplementary Material

Van Nguyen Nguyen[1], Thibault Groueix[2], Georgy Ponimatkin[1], Yinlin Hu[3], Renaud Marlet[1], Mathieu Salzmann[4], Vincent Lepetit[1]

[1]LIGM, Ecole des Ponts, [2]Adobe, [3]MagicLeap, [4]EPFL

1. Implementation details

Our implementation of U-Net is mainly based on denoising-diffusion-pytorch [4]. The encoder receives as input an image embedding of size $8 \times 32 \times 32$ and encodes it into a $1024 \times 4 \times 4$ representation. The decoder takes as input this representation and intermediate representations from the encoder through skip connections, and decodes them into an embedding of same size representing the novel view.

Both the encoder and decoder consist of four layers, each of which comprises a residual block, a cross-attention layer that facilitates the injection of pose embeddings into feature maps, and a convolution followed by group normalization and the SiLU activation function.

We show in Figure 1 an overview of the U-Net architecture we use to generate the novel views' embeddings.

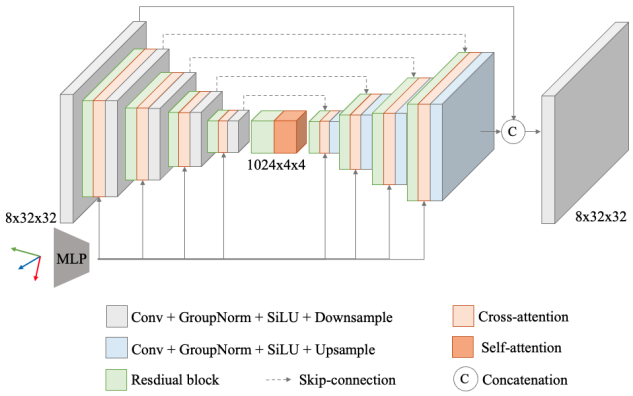

Figure 1. Architecture of our U-Net.

2. Additional results

We provide in Table 1 the 3D pose estimation results of the baselines [1–3,5] and our method on novel instances and novel categories of ShapeNet dataset, as detailed in Section 4 of the main paper, with standard deviation computed on 5 runs with 5 different reference images.

We also show in Figures 2 and 3 some additional results on **unseen categories** without and with occlusions.

	Method	novel inst.	bottle*	bus	clock	dishwasher	guitar	mug	pistol	skateboard	train	washer	mean
Acc 30 ↑	ViewNet [1]	**77.5** ± 4.2	48.4 ± 2.4	36.2 ± 4.8	23.5 ± 3.1	16.4 ± 4.1	37.8 ± 8.1	31.3 ± 5.2	17.9 ± 4.1	33.9 ± 4.8	44.8 ± 4.1	25.1 ± 4.8	35.7 ± 4.4
	SSVE [2]	75.3 ± 3.2	61.5 ± 2.2	38.2 ± 3.7	41.8 ± 3.1	21.3 ± 2.9	46.8 ± 8.7	38.4 ± 4.6	36.8 ± 3.6	**62.3** ± 4.1	41.5 ± 3.5	50.8 ± 3.5	46.8 ± 4.6
	PIZZA [3]	72.3 ± 3.5	76.0 ± 2.8	38.6 ± 5.1	38.5 ± 2.7	32.6 ± 5.1	30.8 ± 8.6	35.6 ± 4.2	40.4 ± 2.3	58.3 ± 3.1	52.9 ± 5.1	**61.0** ± 4.6	48.8 ± 3.8
	3DiM [5]	77.3 ± 2.1	95.1 ± 1.9	43.5 ± 4.1	23.6 ± 3.2	24.5 ± 2.6	36.0 ± 5.3	32.0 ± 3.1	31.9 ± 2.5	50.3 ± 3.5	37.0 ± 3.2	56.1 ± 2.9	46.1 ± 3.6
	Ours (top 1)	75.5 ± 1.3	**96.0** ± 1.1	**53.6** ± 3.2	**48.0** ± 2.1	**48.0** ± 1.8	**49.0** ± 5.1	**44.6** ± 2.3	**69.0** ± 3.1	57.8 ± 2.8	**55.2** ± 2.3	60.6 ± 2.0	**59.8** ± 3.1
	Ours (top 3)	92.0 ± 1.5	97.4 ± 0.9	83.8 ± 2.5	73.4 ± 3.6	78.5 ± 2.2	66.8 ± 4.5	56.0 ± 2.0	83.8 ± 3.1	86.2 ± 2.2	86.0 ± 1.9	84.4 ± 2.0	80.8 ± 2.1
	Ours (top 5)	95.5 ± 0.8	97.8 ± 0.2	89.8 ± 2.0	80.4 ± 3.0	88.2 ± 2.0	74.6 ± 4.6	62.8 ± 2.1	88.4 ± 2.6	92.8 ± 2.0	95.4 ± 1.7	93.4 ± 2.1	87.1 ± 1.7
Median ↓	ViewNet [1]	6.6 ± 5.0	26.7 ± 3.5	35.8 ± 5.1	40.3 ± 4.6	96.3 ± 3.0	50.6 ± 6.9	51.6 ± 4.1	42.8 ± 4.1	37.4 ± 3.7	26.8 ± 3.1	44.3 ± 2.9	41.7 ± 4.9
	SSVE [2]	6.1 ± 3.1	23.8 ± 2.6	45.2 ± 3.0	41.9 ± 3.3	90.4 ± 2.3	47.6 ± 7.6	49.6 ± 4.1	24.0 ± 2.9	13.5 ± 3.8	24.9 ± 4.1	48.1 ± 3.8	37.7 ± 4.2
	PIZZA [3]	5.8 ± 2.6	25.5 ± 3.4	26.4 ± 4.1	43.2 ± 4.1	80.6 ± 2.7	40.2 ± 5.1	45.5 ± 4.5	23.4 ± 2.1	17.3 ± 3.1	20.3 ± 5.1	38.5 ± 5.1	33.3 ± 4.7
	3DiM [5]	**5.7** ± 3.1	**1.8** ± 2.4	19.8 ± 3.5	47.3 ± 3.7	98.8 ± 2.9	35.2 ± 5.8	35.7 ± 2.5	21.2 ± 2.8	**12.5** ± 2.9	**17.6** ± 2.5	19.2 ± 2.6	28.6 ± 3.0
	Ours (top 1)	8.1 ± 1.3	**1.8** ± 1.1	**18.4** ± 3.2	**39.9** ± 2.1	**77.6** ± 1.8	**31.6** ± 5.1	**35.5** ± 2.3	**13.4** ± 3.1	15.5 ± 2.8	18.3 ± 2.3	**8.5** ± 2.0	**24.4** ± 3.1
	Ours (top 3)	5.0 ± 1.0	1.3 ± 0.7	5.8 ± 3.6	9.1 ± 2.1	4.8 ± 2.2	16.0 ± 4.1	22.6 ± 3.1	8.1 ± 1.8	6.5 ± 1.6	6.7 ± 2.6	5.7 ± 2.6	8.3 ± 2.4
	Ours (top 5)	4.5 ± 0.8	1.2 ± 0.6	4.5 ± 3.0	7.1 ± 2.0	4.4 ± 2.1	11.6 ± 3.1	18.4 ± 2.6	6.1 ± 2.4	5.6 ± 1.7	4.9 ± 2.0	5.0 ± 2.8	6.6 ± 2.1

Table 1. **Quantitative results on ShapeNet dataset.** We treat "bottle" as a symmetric category, i.e., the error is only the difference of elevation angle. Since the quality of prediction may depend on the reference image, we report the average and the standard deviation of 5 runs with 5 different reference images.

Figure 2. **Visual results on unseen categories from ShapeNet, without occlusions**. The arrow indicates the pose with the highest probability as recovered by our method. We visually compare to PIZZA, which is the method with the second best performance. **We visualize the predicted poses by rendering the object from these poses, but the 3D model is only used for visualization purposes, not as input to our method. Similarly, we use the canonical pose of the 3D model to visualize this distribution, but not as input to our method.**

References

[1] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. ViewNet: Unsupervised Viewpoint Estimation from Conditional Generation. In *ICCV*, 2021. 1

[2] Octave Mariotti and Hakan Bilen. Semi-Supervised Viewpoint Estimation with Geometry-aware Conditional Generation. In *ECCV Workshop*, 2020. 1

[3] Van Nguyen Nguyen, Yuming Du, Yang Xiao, Michael Ra-

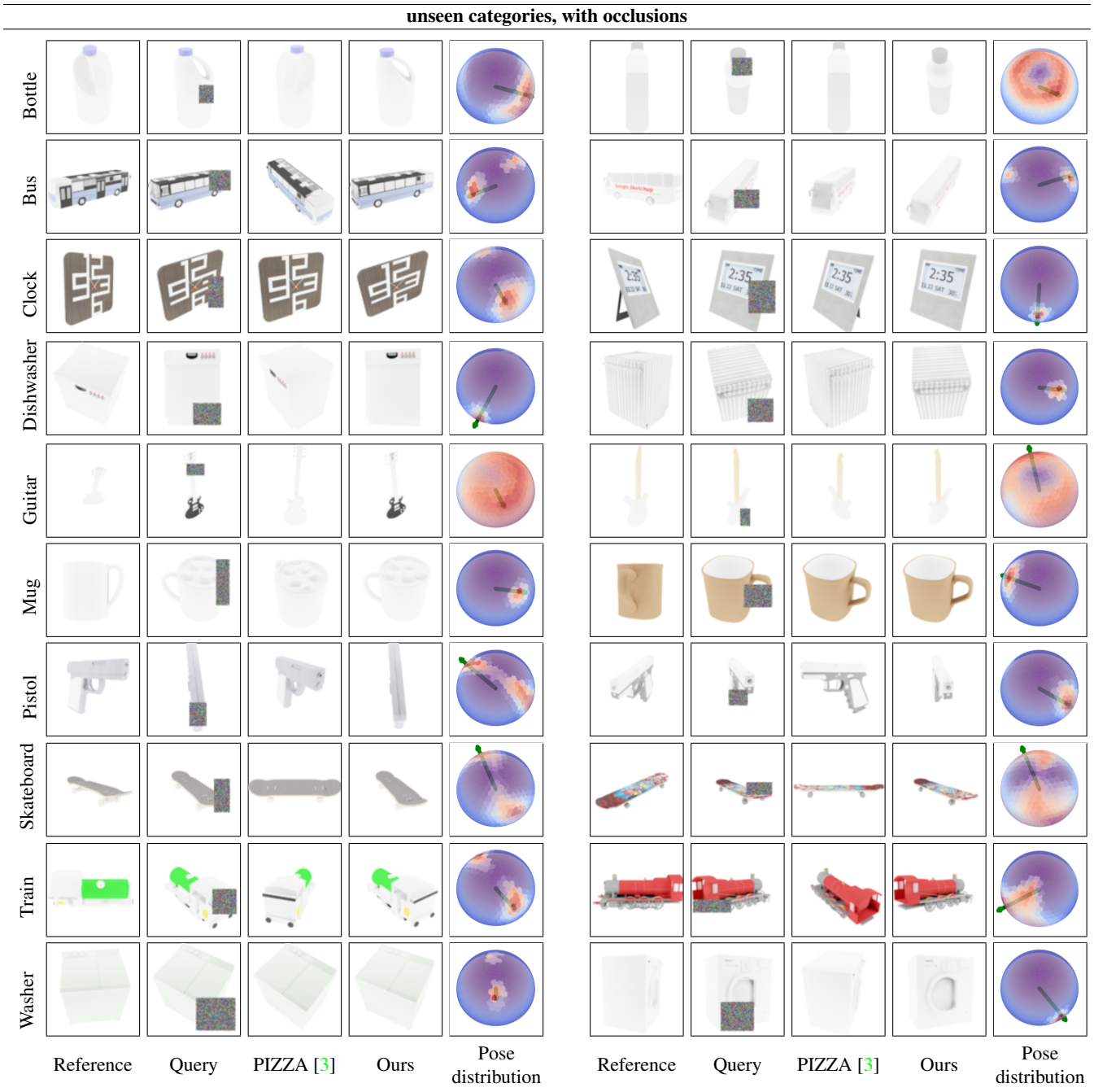

Figure 3. **Visual results on unseen categories from ShapeNet, with occlusions**. The arrow indicates the pose with the highest probability as recovered by our method. We visually compare with PIZZA, which is the method with the second best performance. **We visualize the predicted poses by rendering the object from these poses, but the 3D model is only used for visualization purposes, not as input to our method. Similarly, we use the canonical pose of the 3D model to visualize this distribution, but not as input to our method.**

mamonjisoa, and Vincent Lepetit. PIZZA: A Powerful Image-only Zero-Shot Zero-CAD Approach to 6 DoF Tracking. In *3DV*, 2022. 1, 2, 3

[4] Phil Wang. Implementation of denoising diffusion probabilistic model in pytorch. https://github.com/lucidrains/denoising-diffusion-pytorch, 2022. 1

[5] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel View Synthesis with Diffusion Models. In *ICLR*, 2023. 1