Supplementary Material: Robust Category-Level 3D Pose Estimation from Diffusion-Enhanced Synthetic Data

A. P3D-Diffusion

A.1. Dataset Statistics

We generate P3D-Diffusion using the CAD models provided in the PASCAL3D+ [2] and OOD-CV [4] datasets. Table 1 shows the number of CAD models for each category. We add a random jitter of 5% to the length, width and height of the CAD models to generate more shape variety. For all categories, we randomly sample the azimuth pose from a uniform distribution in range [0, 360]. The range for elevation defined in PASCAL3D+ is [-90, 90], yet we add some constraints in our P3D-Diffusion empirically (e.g. one hardly looks at the car or bus from the bottom). We sample the elevation for each category from a gaussian distribution with a cutoff. Table 2 shows the parameters of the gaussian distribution for each category. During the training, we also add a random in-plane rotation in range [-5, 5].

We first sample 7000 images for each category and randomize the texture of the objects by sampling textures from the describable texture database. And we use the Canny edge detector [1] to produce the 2D edge maps. Then we use a pre-trained style transfer generative model [3] which takes the edge maps as input to generate realistic images. The background images are sampled from a collection of 100 images that we collected from the internet by searching for the keywords "wallpaper". It is worth noting that we do not use specific backgrounds for different categories, which helps the model learn disentangled semantic features and improve robustness in O.O.D. scenarios. See more examples in Section A.2.

A.2. Sample Images

P3D-Diffusion dataset contains 84000 images from 12 categories in PASCAL3D+. Figure 1 shows the sample images from each category. The background for different categories are from a same image set.

B. Qualitative Examples

We provide some qualitative examples of our model on the PASCAL3D+ and OOD-CV datasets in Figure 2. Results show that our CC3D model is able to estimate the pose correctly for a variety of objects in challenging scenarios, such as unusual background (d & e), complex textures (f) and object shapes (c) and camera views (b).

C. Quantitative resutls

C.1. Category-Specific Results on PASCAL3D+

Table 3 shows the pose estimation results on PASCAL3D+ for all categories respectively, comparing our models with fully supervised models trained on the full annotated PASCAL3D+ dataset. For most of the categories, our P3D-Diffusion+CC3D+50% outperforms all the fully supervised models by a significant margin. It should be noted that the number of images of category "car" is much larger than other categories, and the pose estimation accuracy of "car" is almost saturated in fully supervised methods ($98.2\%@\frac{\pi}{6}$ and $94.7\%@\frac{\pi}{18}$). Our P3D-Diffusion+CC3D+50% achieves a larger improvement in non-car categories.

References

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 1

[2] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 1

[3] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[4] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv
num	8	6	6	8	6	10	10	6	5	6	4	4

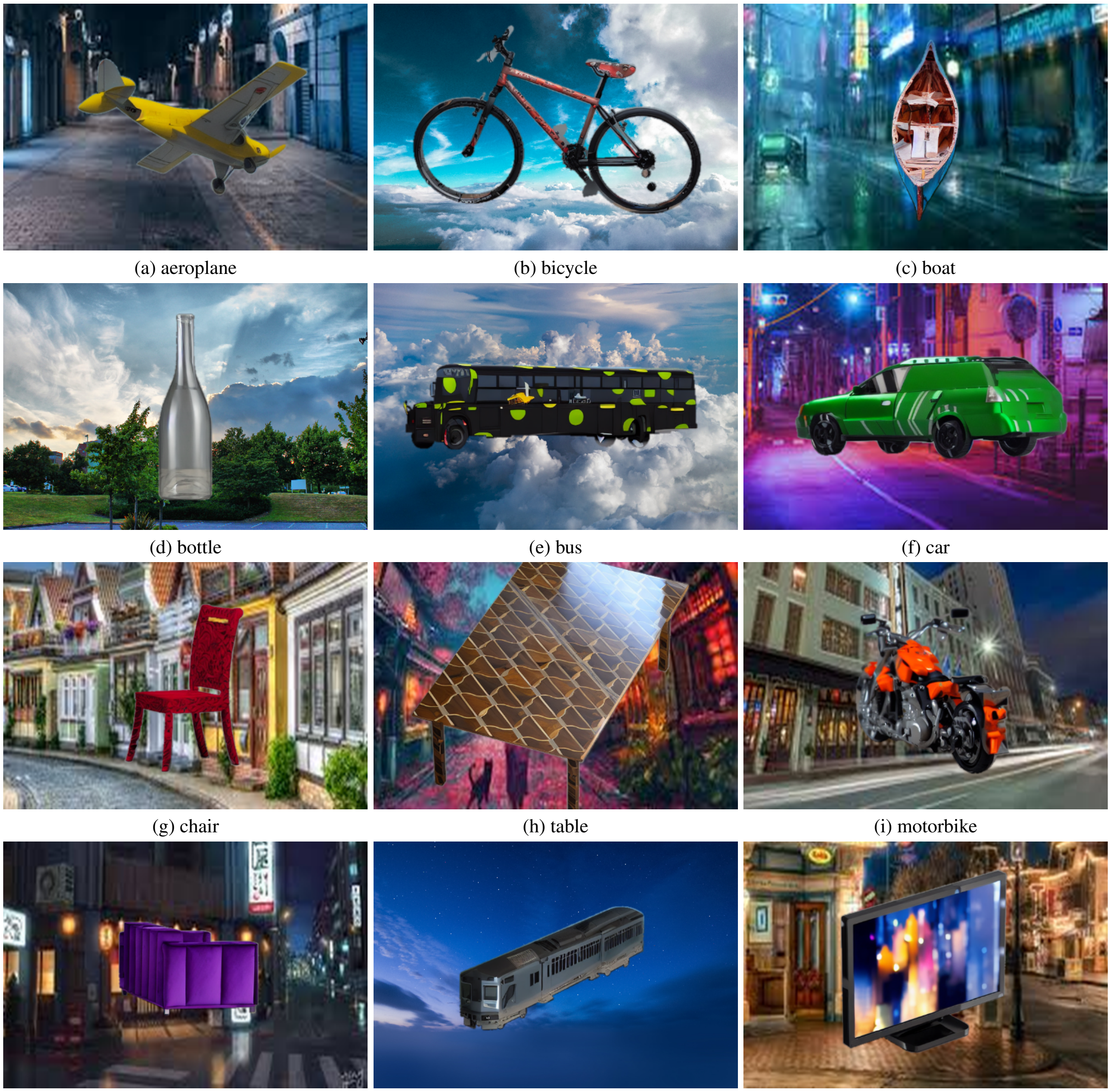
Table 1. The number of CAD models used in generating P3D-Diffusion.

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv
μ	0	10	5	10	5	5	20	20	5	10	0	10
σ	20	10	15	20	5	10	10	10	10	10	10	10
Max	90	50	70	80	30	50	60	50	80	50	30	60
Min	-90	-50	-20	-80	-20	-20	-30	-20	-40	-20	-20	-40

Table 2. Elevation distribution parameters. For each category, we sample the elevation from a gaussian distribution $N(\mu, \sigma^2)$ with a cutoff.

		aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	Mean
$ACC_{\frac{\pi}{6}} \uparrow$	NeMo fully supervised	82.2	78.4	68.1	88.0	91.7	**98.2**	87.0	76.9	85.0	95.0	83.0	82.2	86.1
	Res50 fully supervised	83.0	79.6	**73.1**	87.9	96.8	95.5	91.1	82.0	80.7	97.0	**94.9**	83.3	88.1
	StarMap fully supervised	85.5	84.4	65.0	**93.0**	**98.0**	97.8	94.4	82.7	85.3	**97.5**	93.8	**89.4**	89.4
	P3D-Diffusion+Res50	47.9	68.8	23.2	73.5	41.1	62.1	76.5	52.6	66.2	69.5	30.7	43.4	53.5
	P3D-Diffusion+NeMo	75.6	75.0	38.7	77.1	68.7	91.8	83.0	55.3	76.8	66.5	74.9	47.4	71.8
	P3D-Diffusion+CC3D	76.6	77.6	45.7	81.1	84.7	92.6	86.7	61.7	79.6	75.6	81.1	53.0	76.3
	P3D-Diffusion+CC3D+10%	87.4	83.5	62.1	85.2	90.9	96.1	94.8	79.8	82.7	95.1	89.3	86.8	86.7
	P3D-Diffusion+CC3D+50%	**90.9**	**88.8**	72.0	88.4	96.0	97.9	**95.6**	**86.4**	**87.9**	97.2	92.5	88.2	**90.7**
$ACC_{\frac{\pi}{18}} \uparrow$	NeMo fully supervised	49.7	29.5	37.7	49.3	89.3	**94.7**	49.5	52.9	29.0	58.5	70.1	42.4	61.0
	Res50 fully supervised	31.3	25.7	23.9	35.9	67.2	63.5	37.0	40.2	18.9	62.5	51.2	24.9	44.6
	StarMap fully supervised	49.8	34.2	25.4	**56.8**	90.3	81.9	67.1	57.5	27.7	70.3	69.7	40.0	59.5
	P3D-Diffusion+Res50	6.9	14.1	4.3	21.1	14.4	16.7	22.8	13.7	12.9	22.9	6.4	6.0	13.2
	P3D-Diffusion+NeMo	42.8	28.3	13.0	25.5	54.3	70.9	38.8	28.1	28.5	9.9	44.4	14.4	39.5
	P3D-Diffusion+CC3D	43.4	27.7	18.3	11.6	71.4	73.9	37.6	33.8	27.5	7.8	50.9	14.1	41.4
	P3D-Diffusion+CC3D+10%	58.2	37.9	36.5	45.9	84.5	89.8	63.5	60.9	36.0	55.7	71.9	44.2	62.4
	P3D-Diffusion+CC3D+50%	**65.0**	**41.8**	**44.8**	54.2	**91.7**	93.9	**70.6**	**70.8**	**40.4**	**63.9**	**82.0**	**52.8**	**71.4**
$MedErr \downarrow$	NeMo fully supervised	10.1	16.3	14.9	10.2	3.2	**3.2**	10.1	9.3	14.1	8.6	5.4	12.2	8.8
	Res50 fully supervised	13.3	15.9	15.6	12.1	8.9	8.8	11.5	11.4	16.6	8.7	9.9	15.8	11.7
	StarMap fully supervised	10.0	14.0	19.7	**8.8**	3.2	4.2	6.9	8.5	14.5	**6.8**	6.7	12.1	9.0
	P3D-Diffusion+Res50	32.6	21.9	68.6	16.2	8.4	21.7	15.4	26.4	22.1	16.5	85.4	34.4	26.7
	P3D-Diffusion+NeMo	12.2	16.4	46.1	15.2	8.5	7.0	12.1	23.9	14.8	23.3	11.3	31.8	17.6
	P3D-Diffusion+CC3D	11.7	15.9	37.1	15.8	6.8	6.7	12.6	17.1	15.4	22.2	9.8	28.2	15.5
	P3D-Diffusion+CC3D+10%	8.2	13.0	16.7	10.9	3.3	4.0	8.0	7.5	12.8	9.0	5.6	11.4	8.4
	P3D-Diffusion+CC3D+50%	**6.8**	**11.8**	**11.9**	9.2	**2.8**	**3.4**	**6.8**	**6.1**	**12.2**	7.7	**4.6**	**9.5**	**6.9**

Table 3. Pose estimation results on PASCAL3D+ for all categories respectively. Results reported in Accuracy (percentage, higher better) and Median Error (degree, lower better).



(a) aeroplane

(b) bicycle

(c) boat

(d) bottle

(e) bus

(f) car

(g) chair

(h) table

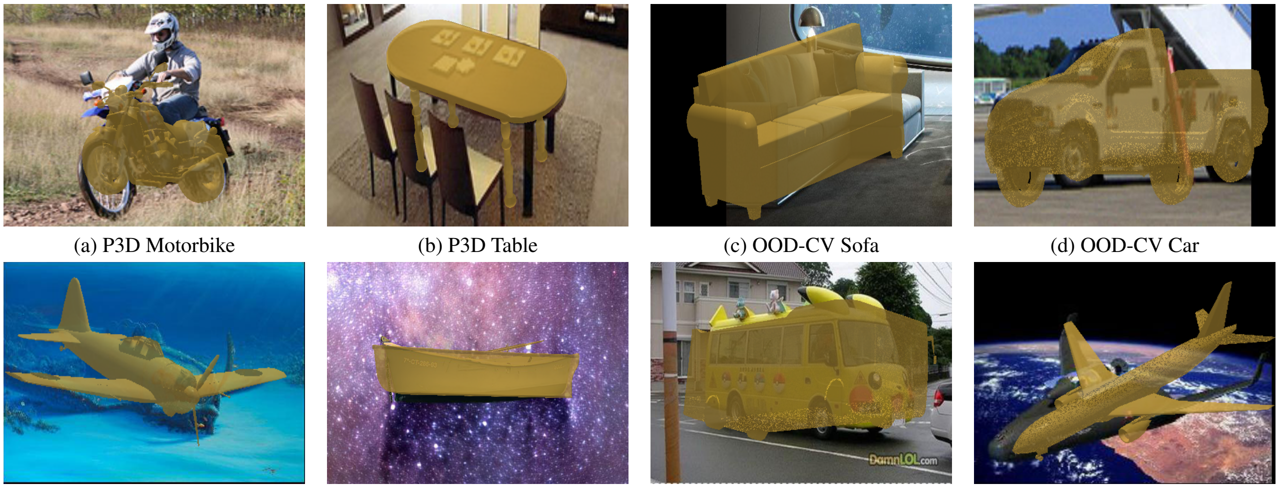
(i) motorbike

(j) sofa

(k) train

(l) tvmonitor

Figure 1. Samples of images in P3D-Diffusion. Figure (d), (f), and (h) show the realistic specular reflection produced by the graphics-guided style transfer module. Figure (e), (f), and (g) show the diverse textures from the 3D-consistent prior noise. Figure (a), (i), and (j) demonstrate that the simple prompt engineering can encourage objects with varying colors. Finally, all images are created with OOD-aware generation so our model could effectively break the spurious correlation between task-related semantic information and domain-specific background features.



(a) P3D Motorbike (b) P3D Table (c) OOD-CV Sofa (d) OOD-CV Car

(e) OOD-CV Plane (f) OOD-CV Boat (g) OOD-CV Bus (h) OOD-CV Aeroplane

Figure 2. Qualitative results of our proposed model on the PASCAL3D+ and OOD-CV datasets. We illustrate the predicted 3D pose using the CAD models from the respective datasets. Note that in our approach object are represent as cuboid without detailed shape. Our P3D-Diffusion+CC3D model is able to estimate the pose correctly for a variety of objects in challenging scenarios, such as unusual background (d & e), complex textures (f) and object shapes (c) and camera views (b).