PASR: Pose-Aware 3D Shape Retrieval from Occluded Single Views

Supplementary Material

A. Additional Implementation Details

2D Encoder. We adopt the `vit_base_patch16` variant of DINOv3 [**?**] as our 2D encoder. We use the output of the final transformer layer as the 2D feature representation. The input query images are resized to a resolution of $3 \times 592 \times 592$, and the resulting feature maps have a spatial resolution of 37×37 with 768 channels, *i.e.*, the encoder output has shape $768 \times 37 \times 37$.

Renderer. We employ the point-based renderer from PyTorch3D, using `PointRasterizer` followed by `NormWeightedCompositor` to project point-level features into the 2D image plane. The rendered feature maps have a spatial resolution of 37×37, matching the resolution of the 2D encoder output. The point radius is set to 0.04, and the number of points per pixel is set to 16. The focal length and principal point are adjusted according to the output resolution, and normalized device coordinates (NDC) are disabled. The background color is set to black.

Data Preprocessing. For 3D shapes, we first convert the meshes into point clouds using PyTorch3D and then normalize the point clouds. Specifically, we uniformly sample 4096 points on the mesh surface, with sampling probability proportional to the face area, to obtain a batch of point clouds from a batch of meshes. For the CMIC [**?**] and SC-IBSR [**?**] baselines, we apply standard data augmentation, including horizontal flipping, random cropping, and color transfer. For PASR, OpenShape [**?**], and Uni3D [**?**], we do not use any data augmentation. Figure 1 shows the original occluded images. All input images are centered and resized to a resolution suitable for downstream 3D pose estimation, ensuring consistent scale and alignment across the dataset.

B. Additional Results

In this section, we provide additional results on Pix3D [**?**] and Pascal3D [**?**]. Table 1 presents all pose metrics for our method and comparison baselines, Swin-T and ResNet-50. On Pix3D, our method outperforms the baselines under L0, L1, and L2 occlusion. Under severe occlusion (L3), the pose estimation accuracy is inherently bounded by the retrieval quality. Since our method retrieves the 3D shape first, incorrect shape retrieval in extreme occlusion scenarios leads to higher pose errors, resulting in performance lower than the baselines in this specific setting. On Pascal3D, our method achieves the best performance or highly competitive scores across all metrics and occlusion levels compared to Swin-T and ResNet-50.

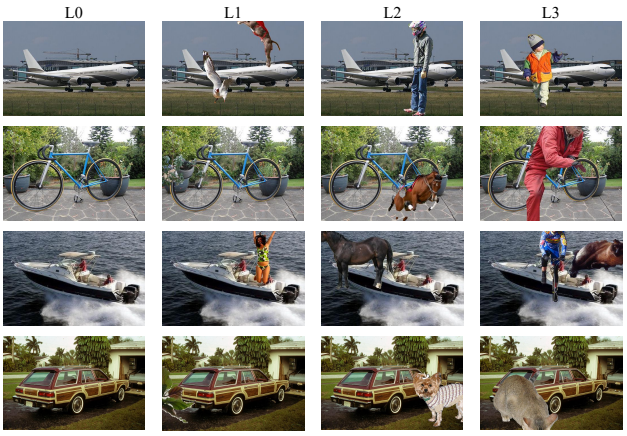

Figure 1. Examples of the query images across different occlusion levels (L0-L3).

As shown in Table 2, on Pascal3D, our method outperforms the other compared models overall and in most categories. Notably, for some categories with a large number of samples, our method is minimally affected by occlusion. For instance, in the aeroplane category, our method only saw a decrease of 8.52% from L0 to L3, whereas other competitive models, OpenShape and CMIC decreased by 39.5% and 22.5%, respectively. In the car category, our method only saw a 5.14% drop from L0 to L3 as well, whereas CMIC decreased by 26.6%, and SC-IBSR decreased by 34.14%. We observe a counter-intuitive performance gain, such as Uni3D in the aeroplane category from L0 to L1. Upon qualitative inspection, we hypothesize that this is because the L0 images in Pascal3D often contain significant background clutter. The synthetic occlusion introduced in L1 may inadvertently suppress these distracting background features, forcing the attention of the model towards the discriminative geometric parts of the aeroplane (e.g., wings/fuselage), to which the ViT backbone is robust.

C. Additional Qualitative Results

Figure 2 shows some additional 3D shape retrieval results on Pascal3D (L3). PASR not only retrieves the best-matching 3D shapes, but also estimates the pose of each shape accurately.

Method	$Acc_{\pi/6}$ ↑				$Acc_{\pi/18}$ ↑				MedErr ↓			
	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3
∇Pix3D												
Swin-T	72.0	66.0	60.8	**50.6**	39.4	30.6	25.0	19.3	12.5	16.1	19.4	**29.0**
ResNet-50	80.4	73.4	61.8	43.9	48.1	40.4	30.9	17.7	10.3	12.5	17.3	41.4
Ours	**89.1**	**82.8**	**66.7**	38.6	**70.0**	**58.0**	**41.3**	**19.9**	**6.5**	**8.4**	**13.5**	50.3
∇Pascal3D												
Swin-T	63.6	59.8	56.4	51.3	29.6	26.2	22.8	17.7	17.6	20.7	23.6	28.8
ResNet-50	71.8	67.2	63.5	**56.0**	37.6	32.6	28.6	20.7	13.7	15.6	18.6	24.6
Ours	**80.2**	**76.2**	**68.2**	55.2	**40.3**	**36.7**	**34.3**	**25.7**	**12.1**	**13.3**	**15.4**	**23.9**

Table 1. Performance comparison on Pix3D and Pascal3D under different rotation granularities ($\pi/6$, $\pi/18$) and median error (MedErr).

Method	Level	Aero	Bike	Boat	Bottle	Bus	Car	Chair	Table	MBike	Train	Sofa	TV	Overall
Uni3D	L0	64.66	27.81	94.36	55.43	59.93	58.08	61.98	27.21	48.20	20.43	20.75	71.71	54.01
	L1	77.75	24.06	91.09	51.90	52.81	47.65	59.32	24.96	51.48	16.67	13.52	65.79	50.25
	L2	49.48	24.37	88.36	49.46	52.06	46.48	51.71	22.70	51.48	20.43	22.96	61.18	46.73
	L3	63.20	19.69	84.55	44.84	52.81	45.08	45.25	22.88	50.82	11.29	27.04	54.28	45.84
OpenShape	L0	82.12	32.19	82.91	39.67	37.08	45.67	53.61	46.97	71.15	75.27	65.09	52.96	55.80
	L1	49.90	33.12	75.64	55.71	31.46	42.07	48.67	37.44	53.77	91.40	60.69	47.70	49.80
	L2	46.36	28.12	74.91	49.73	38.95	39.94	44.49	28.08	65.25	92.47	**70.13**	37.83	47.99
	L3	42.62	24.69	65.45	46.20	17.60	37.81	42.59	26.52	57.38	86.56	**59.43**	29.61	42.56
CMIC	L0	78.80	56.90	89.10	68.20	81.30	**72.70**	67.70	**73.10**	82.30	96.20	**71.10**	77.00	75.44
	L1	74.20	51.20	84.50	65.50	77.50	65.60	62.40	**69.50**	77.00	93.00	**67.90**	75.70	70.68
	L2	66.50	52.50	81.10	59.00	71.50	58.40	52.90	**67.60**	75.70	91.40	64.50	72.00	65.87
	L3	56.30	**42.20**	67.30	47.80	61.80	46.10	45.60	**62.20**	61.00	83.30	56.60	52.00	54.76
SC-IBSR	L0	67.57	42.50	83.27	54.08	72.28	62.85	58.94	63.95	73.11	92.47	66.67	73.36	66.42
	L1	54.05	35.94	72.00	49.73	68.16	53.96	48.29	60.31	66.56	87.63	62.26	65.46	58.65
	L2	47.19	29.69	65.64	45.92	61.42	41.78	32.70	53.73	61.64	80.11	55.66	61.18	50.58
	L3	33.26	18.44	46.36	39.13	44.19	28.71	25.86	47.66	55.74	73.12	38.68	40.13	38.12
Ours	L0	**89.19**	**57.50**	**96.74**	**74.73**	**86.14**	70.41	**75.67**	59.00	**86.56**	**97.85**	62.58	**84.87**	**76.43**
	L1	**89.19**	**54.69**	**96.01**	**73.37**	**84.64**	**67.40**	**69.96**	47.92	**84.59**	**97.85**	57.86	**82.24**	**73.21**
	L2	**87.87**	**52.81**	**95.39**	**70.16**	**82.86**	**66.28**	**67.89**	45.63	**83.61**	**96.77**	56.74	**79.47**	**71.49**
	L3	**80.67**	36.25	**92.57**	**63.04**	**71.54**	**65.27**	**52.09**	30.62	**72.46**	**96.24**	37.42	**60.53**	**63.05**

Table 2. Per-category 3D shape retrieval accuracy (%) across occlusion levels L0–L3 on Pascal3D.

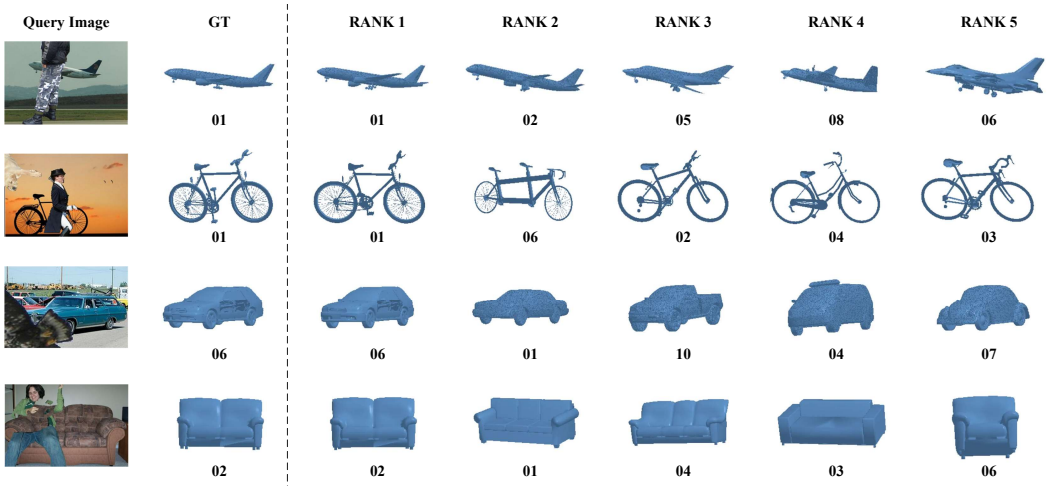

Figure 2. Additional 3D shape retrieval visualizations on Pascal3D (L3). Ground Truth (GT) is shown for reference.