Supplementary Material for Environment-Agnostic Pose: Generating Environment-independent Object Representations for 6D Pose Estimation

In this supplementary document, we first provide more samples of self-made DiverseScenes Dataset in Figure 1, Figure 2, Figure 3 and Figure 4. We simulated different indoor and outdoor backgrounds with varying lighting to construct four distinct scenarios, which have a significant difference from the training data.

We also provide detailed results on LineMOD, LineMOD-Occluded (LM-O) and YCB-V in Table 1, Table 2 and Table 3. We also evaluate our method on YCB-V and LM-O and report results in BOP challenges average recall (AR) metrics and runtime in Table 4. Trained solely on synthetic data, our method still achieves the best results.

Next, we study the impact of different backbones for the image (condition) encoder with ResNet101 [2], Effecientnetb7 [12] and Swin-B [6]. The results are reported in Table 3 which shows that Swin-B performs better than other backbones. When replaceing the Swin-Transformer with CNN-based backbone ResNet, the accuracy on LM-O dropped by 2.5 % to 85.3%, yet it still outperforms Self6D++ (59.8%) and SO-Pose (62.3%).

At last, we provide more detailed visualization results including the environment-independent object representations under cluttered scenes and the same objects in different environments in Figure 5, Figure 6, Figure 7 and Figure 8 .

Table 1: Comparison with state-of-the-art methods on LineMOD dataset. The table reports results for the Average Recall (%) of ADD(-S). All results except ours are copied from SMOC-Net [13], TexPos [1] and RKHSPose [16]. R: annotated real RGB data. S: synthetic RGB data. R^{-}: unannotated real RGB data. D: depth data. The best pose method using the same kind of training data is underlined, and the overall best method is marked in bold.

Methods	Training data	Ape	Bench.	Cam	Can	Cat	Driller	Duck	Eggbox	Glue	Holep.	Iron	Lamp	Phone	Mean
DPOD [19]	R	53.3	95.2	90.0	94.1	60.4	97.4	66.0	99.6	93.8	64.9	99.8	88.1	71.4	82.6
PVNet [8]		43.6	**99.9**	86.9	95.5	79.3	96.4	52.6	99.2	95.7	81.9	98.9	**99.3**	92.4	86.3
CDPN [4]		64.4	97.8	91.7	95.5	83.8	96.2	66.8	99.7	99.6	**85.8**	97.9	97.9	90.8	89.9
Self6D [15]	S+R^{-}+D	38.9	75.2	36.9	65.6	57.9	67.0	19.6	99.0	94.1	16.2	77.9	68.2	50.1	58.9
Self6D++ [14]		75.4	94.9	97.0	99.5	86.6	98.9	68.3	99.0	96.1	41.9	99.4	98.9	94.3	88.5
RKHSPose [16]		90.3	99.7	99.1	99.8	96.4	99.3	86.5	99.8	99.8	80.7	99.6	98.8	97.2	95.9
DSC-PoseNet [18]	S+R^{-}	35.9	83.1	51.5	61.0	45.0	68.0	27.6	89.2	52.5	26.4	56.3	68.7	46.4	54.7
Self6D++ [14]		76.0	91.6	97.1	99.8	85.6	98.8	56.5	91.0	92.2	35.4	99.5	97.4	91.8	85.6
SMOC-Net [13]		85.6	96.7	**97.2**	**99.9**	95.0	**100.0**	76.0	98.3	**99.2**	45.6	**99.9**	98.9	**94.0**	91.3
TexPose [1]		80.9	99	94.8	99.7	92.6	97.4	83.4	94.9	93.4	79.3	99.8	98.3	78.9	91.7
AAE [11]	S	4.2	22.9	32.9	37.0	18.7	24.8	5.9	81.0	46.2	18.2	35.1	61.2	36.3	32.6
MHP [7]		11.9	66.2	22.4	59.8	26.9	44.6	8.3	55.7	54.6	15.5	60.8	-	34.4	38.8
DPODv2 [9]		35.1	59.4	15.5	48.8	28.1	59.3	25.6	51.2	34.6	17.7	84.7	45.0	20.9	40.5
EA6D		**95.8**	98.8	97.2	98.3	**97.4**	96.9	**96.5**	**100.0**	**99.2**	84.9	98.2	96.7	91.3	**96.3**

Table 2: Comparison with state-of-the-art methods on LineMOD-Occluded dataset. The table reports results for the Average Recall (%) of ADD(-S). All results except ours are copied from 6D-diff [17], SMOC-Net [13] and TexPos [1]. The best pose method using the same kind of training data is underlined, and the overall best method is marked in bold.

Methods	Training data	Ape	Can	Cat	Driller	Duck	Eggbox	Glue	Holep.	Mean
ZebraPose [10]	R+S	57.9	95.0	60.6	94.8	64.5	70.9	88.7	83.0	76.9
CheckerPose [5]	R+S	58.3	95.7	62.3	93.7	69.9	70.0	86.4	83.8	77.5
6D-diff [17]	R+S	60.6	97.9	63.2	**96.6**	67.2	73.5	92.0	**85.5**	79.6
Self6D [15]	S+R^{-}+D	13.7	43.2	18.7	32.5	14.4	**57.8**	54.3	22.0	32.1
Self6D++ [14]	S+R^{-}+D	59.4	**96.5**	**60.8**	92.0	30.6	51.1	88.6	38.3	64.7
RKHSPose [16]	S+R^{-}+D	62.7	93.7	58.2	92.7	58.7	48.3	88.7	46.7	68.7
DSC-PoseNet [18]	S+R^{-}	13.9	15.1	19.4	40.5	6.9	38.9	24.0	16.3	21.9
Self6D++ [14]	S+R^{-}	57.7	95.0	52.6	90.5	26.7	45.0	87.1	23.5	59.8
SMOC-Net [13]	S+R^{-}	60.0	94.5	59.1	93.0	37.2	48.3	89.3	25.0	63.3
TexPose [1]	S+R^{-}	60.5	93.4	56.1	92.5	55.5	46.0	82.8	46.5	66.7
EA6D	S	**92.3**	94.2	**70.5**	95.4	**86.6**	**85.6**	**95.6**	82.4	**87.8**

Table 3: Comparison with RGB-based 6D object pose estimation methods on YCB-V dataset. (*) denotes symmetric objects.

Method	Self6D++ [14]		RKHSPose [16]		ZebraPose [10]		CheckerPose [5]		6D-diff [17]		EA6D	
Metric	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)
002_master_chef_can	88.8	8.4	88.7	13.7	93.7	75.4	87.5	67.7	94.3	77.3	96.4	78.4
003_cracker_box	94.2	84.9	94.7	86.2	93.0	87.8	93.2	86.7	93.7	88.1	95.8	86.6
004_sugar_box	95.8	88.0	96.2	91.3	95.1	90.9	95.9	91.7	96.3	91.8	98.2	88.8
005_tomato_soup_can	90.8	79.4	92.2	83.2	94.4	90.1	94.0	89.9	95.4	91.3	97.1	87.6
006_mustard bottle	98.6	92.7	99.5	92.7	96.0	92.6	95.7	90.9	96.6	92.9	96.5	88.2
007_tuna_fish_can	97.5	89.7	98.2	92.3	96.9	92.6	97.5	90.9	96.9	93.8	97.3	85.6
008_pudding_box	98.4	93.9	98.3	94.3	97.2	95.3	94.9	91.5	97.6	95.6	94.8	84.7
009_gelatin_box	94.0	83.9	95.2	84.2	96.8	94.8	96.1	93.4	97.3	95.3	95.1	86.6
010_potted_meat_can	89.3	75.7	92.7	76.3	91.7	83.6	86.4	80.4	92.5	84.5	98.5	87.2
011_banana	98.5	91.8	98.4	93.7	92.6	84.6	95.7	90.1	94.7	89.4	96.4	95.2
019_pitcher_base	98.9	92.1	99.1	94.3	96.4	93.4	95.8	91.9	96.7	93.9	98.7	91.4
021_bleach_cleanser	93.5	84.5	94.2	86.0	89.5	80.8	90.6	83.2	90.3	82.8	92.3	88.3
024_bowl*	89.1	89.1	92.3	92.5	37.1	37.1	82.5	82.5	41.8	42.5	80.4	78.5
025_mug	94.1	81.4	95.2	83.2	96.1	90.8	96.9	92.7	96.7	91.7	98.4	92.1
035_power_drill	95.2	84.2	95.5	86.3	95.0	89.7	94.7	88.8	95.6	91.4	87.3	92.3
036_wood_block*	78.3	78.3	81.2	81.2	84.5	84.5	68.3	68.3	87.8	88.1	89.6	93.5
037_scissors	69.2	45.2	71.3	62.3	92.5	84.5	91.7	81.6	93.1	86.5	94.2	95.3
040_large_marker	87.5	74.6	89.2	75.6	80.4	69.5	83.3	72.3	85.6	72.5	75.8	88.4
051_large_clamp*	79.2	79.2	83.4	83.4	85.6	85.6	90.0	90.0	89.3	88.6	88.4	89.2
052_extra_large_clamp*	87.3	87.3	90.2	90.2	92.5	92.5	91.6	91.6	92.7	92.7	90.5	87.4
061_foam_brick*	95.5	95.5	95.5	95.5	95.3	95.3	94.1	94.1	96.5	95.7	94.2	89.7
mean	91.1	80.0	92.4	82.8	90.1	85.3	91.3	86.4	91.5	87.0	93.1	88.3

Table 4: The average processing time and AR metrics in BOP challenges [3] on YCB-V and LM-O using different training data.

Training data	Real+Synthetic		Synthetic		
Method	GPose [3]	HccePose [3]	GDRNPP [3]	ZebraPose [10]	EA6D
AR (YCB-V)	0.809	0.839	0.713	0.830	0.835
AR (LM-O)	0.699	0.768	0.713	0.729	0.744
Runtime	0.26s	0.10s	0.277	0.25s	0.23s

Table 5: The results of different backbones on LM-O dataset.

Backbone	ResNet [2]	EffecientNet [12]	Swin-B [6]
ADD(-S)	85.3	86.4	87.8

References

[1] Hanzhi Chen, Fabian Manhardt, Nassir Navab, and Benjamin Busam. Texpose: Neural texture learning for self-supervised 6d object pose estimation. In *CVPR*, 2023. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[3] Tomáš Hodaň, Martin Sundermeyer, Yann Labbé, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiří Matas. BOP challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 2

[4] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *ICCV*, 2019. 1

[5] Ruyi Lian and Haibin Ling. Checkerpose: Progressive dense keypoint localization for object pose estimation with graph neural network. In *ICCV*, 2023. 2

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 2

[7] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *ICCV*, 2019. 1

[8] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 1

[9] Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. Dpodv2: Dense correspondence-based 6 dof pose estimation. *TPAMI*, 2021. 1

[10] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *CVPR*, 2022. 2

[11] M. Sundermeyer, Z. Marton, M. Durner, M. Brucker, and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *ECCV*, 2018. 1

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 2019. 1, 2

[13] Tao Tan and Qiulei Dong. Smoc-net: Leveraging camera pose for self-supervised monocular object pose estimation. In *CVPR*, 2023. 1, 2

[14] Gu Wang, Fabian Manhardt, Xingyu Liu, Xiangyang Ji, and Federico Tombari. Occlusion-aware self-supervised monocular 6d object pose estimation. *TPAMI*, 2021. 1, 2

[15] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *ECCV*, 2020. 1, 2

[16] Yangzheng Wu and Michael Greenspan. Pseudo-keypoint rkhs learning for self-supervised 6dof pose estimation. In *ECCV*, 2024. 1, 2

[17] Li Xu, Haoxuan Qu, Yujun Cai, and Jun Liu. 6d-diff: A keypoint diffusion framework for 6d object pose estimation. In *CVPR*, 2024. 2

[18] Zongxin Yang, Xin Yu, and Yi Yang. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In *CVPR*, 2021. 1, 2

[19] S. Zakharov, I. Shugurov, and S. Ilic. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019. 1

Figure 1: Scene 1 of DiverseScenes Dataset.

Figure 2: Scene 2 of DiverseScenes Dataset.

Figure 3: Scene 3 of DiverseScenes Dataset.

Figure 4: Scene 4 of DiverseScenes Dataset.

Figure 5: Visulization results on the LINEMOD dataset. Top: input images and the visulizations of poses, green bounding boxes and blue bounding boxes represent GT poses and estimated poses, respectively. Middle: Visualization of generated pure object representation. Bottom: The image after decoding the pure object representation.

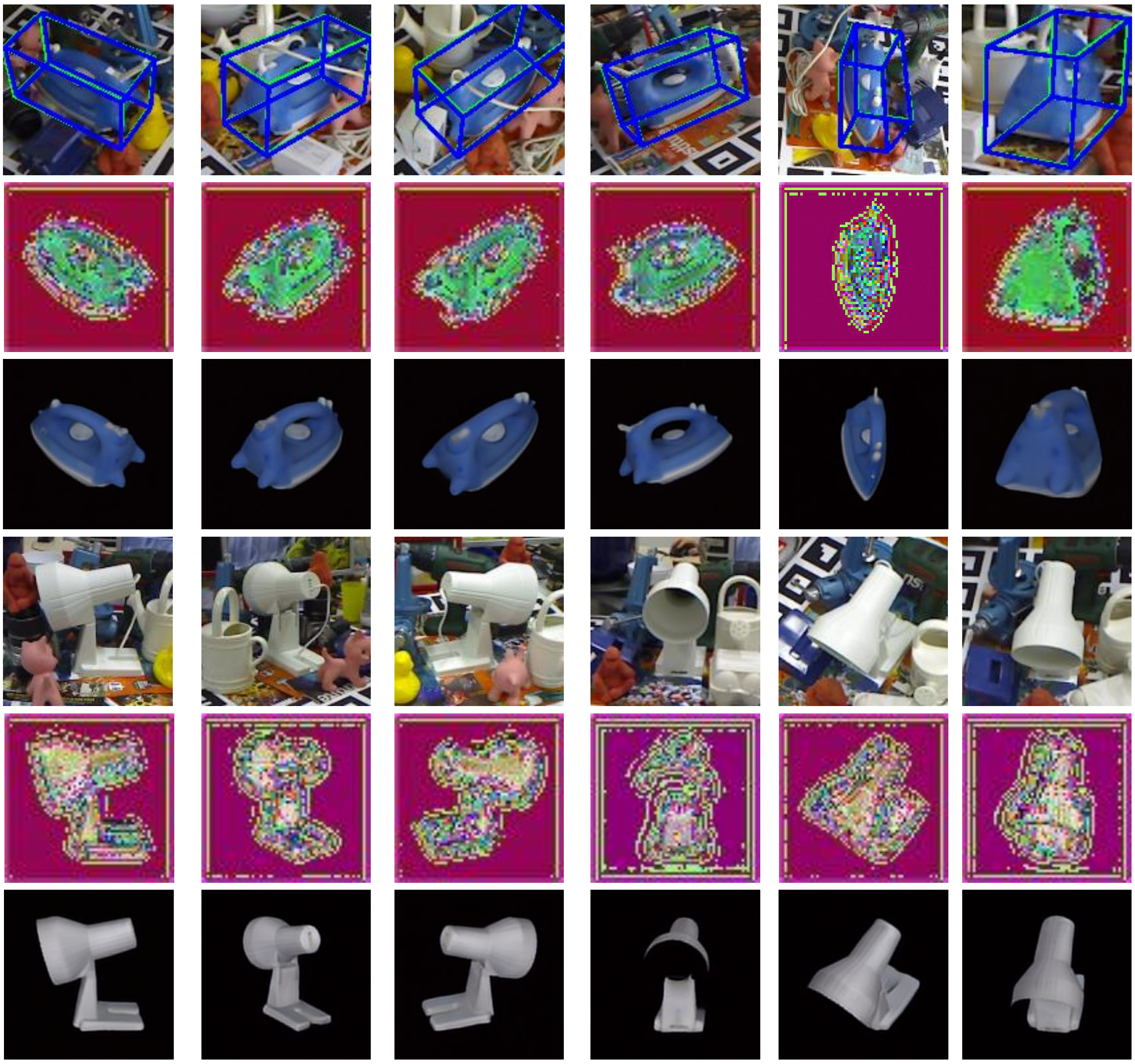

Figure 6: Visulization results on the LINEMOD dataset. Top: input images and the visulizations of poses, green bounding boxes and blue bounding boxes represent GT poses and estimated poses, respectively. Middle: Visualization of generated pure object representation. Bottom: The image after decoding the pure object representation.

Figure 7: Visulization results of generated environment-independent object representation under occlusion on the LM-O dataset.

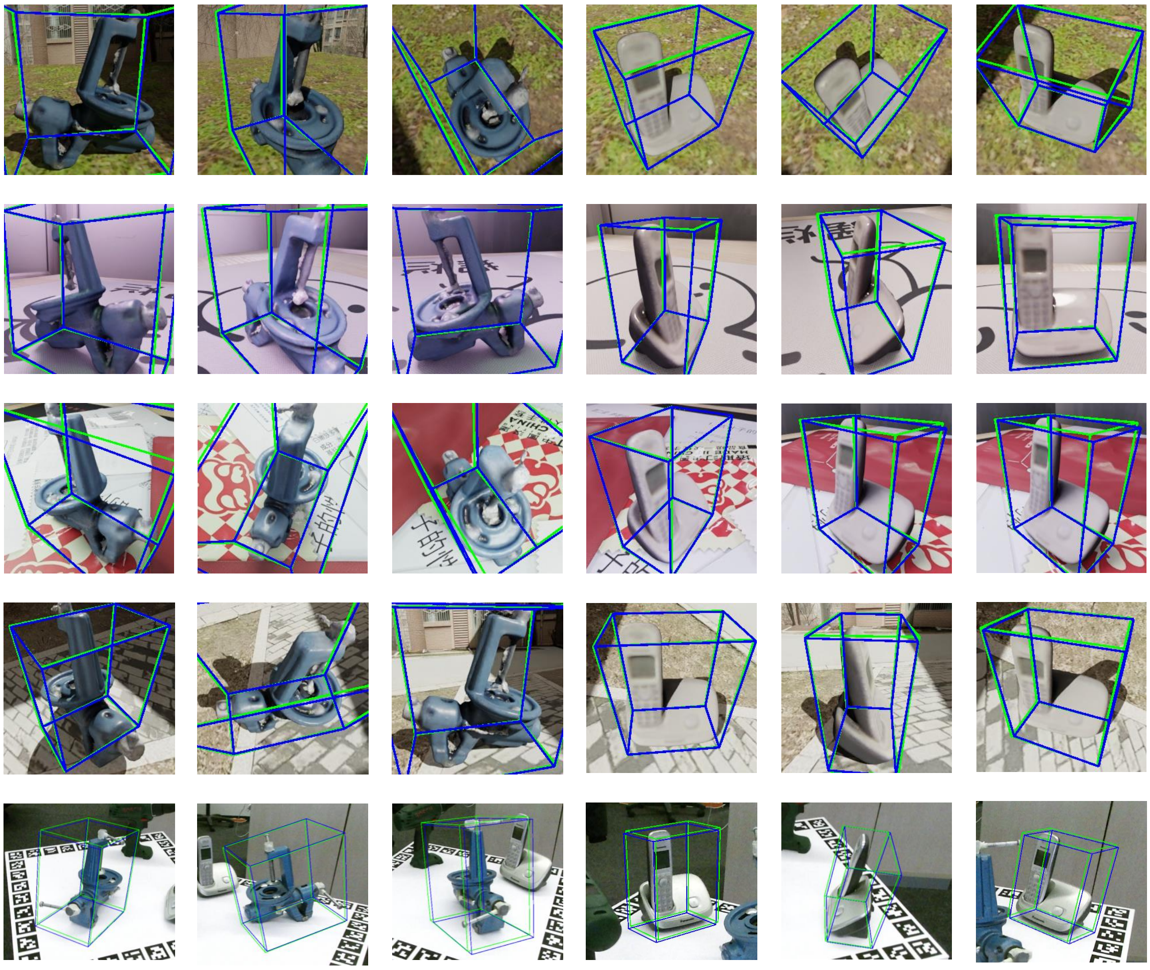

Figure 8: Visulization results of same objects on DiverseScenes Dataset and HomebrewedDB dataset. Green bounding boxes and blue bounding boxes represent GT poses and estimated poses, respectively.