Supplementary Material: 3D Copy-Paste: Physically Plausible Object Insertion for Monocular 3D Detection

Yunhao Ge[◇†], **Hong-Xing Yu**[◇], **Cheng Zhao**[§], **Yuliang Guo**[§], **Xinyu Huang**[§], **Liu Ren**[§], **Laurent Itti**[†], **Jiajun Wu**[◇]
[◇]Stanford University [†]University of Southern California
[§]Bosch Research North America, Bosch Center for Artificial Intelligence (BCAI)
{yunhaoge, koven, jiajunwu}@cs.stanford.edu {yunhaoge, itti}@usc.edu
{Cheng.Zhao, Yuliang.Guo2, Xinyu.Huang, Liu.Ren}@us.bosch.com

A Experiments on more Monocular 3D Object Detection methods

In our main paper, we utilize ImVoxelNet [Rukhovich et al., 2022] for monocular 3D object detection. To show the robustness of our 3D Copy-Paste across different downstream detection methods. We conducted additional experiments with another monocular 3D object detection model: Implicit3DUnderstanding (Im3D [Zhang et al., 2021]). The Im3D model predicts object 3D shapes, bounding boxes, and scene layout within a unified pipeline. Training this model necessitates not only the SUN RGB-D dataset but also the Pix3D dataset [Sun et al., 2018], which supplies 3D mesh supervision. The Im3D training process consists of two stages. In stage one, individual modules - the Layout Estimation Network, Object Detection Network, Local Implicit Embedding Network, and Scene Graph Convolutional Network - are pretrained separately. In stage two, all these modules undergo joint training. We incorporate our 3D Copy-Paste method only during this second stage of joint training, and it's exclusively applied to the 10 SUN RGB-D categories we used in the main paper. We implemented our experiment following the official Im3D guidelines[1].

Table 1 displays the Im3D results for monocular 3D object detection on the SUN RGB-D dataset, adhering to the same ten categories outlined in main paper. Im3D without insertion, attained a mean average precision (mAP) detection performance of 42.13%. After applying our 3D Copy-Paste method—which encompasses physically plausible insertion of position, pose, size, and light—the monocular 3D object detection mAP performance increased to 43.34. These results further substantiate the robustness and effectiveness of our proposed method.

Table 1: Im3D [Zhang et al., 2021] 3D monocular object detection performance on the SUN RGB-D dataset (same 10 categories as the main paper).

Setting	Insertion Position, Pose, Size	Insertion Illumination	mAP
Im3D	N/A	N/A	42.13
Im3D + 3D Copy-Paste	Plausible position, size, pose	Plausible dynamic light	**43.34**

B More experiment details

We run the same experiments multiple times with different random seeds. Table 2 shows the main paper Table **??** results with error range.

We also show our results with mAP@0.15 on SUN RGB-D dataset (Table 3), our method shows consistent improvements.

[1]https://github.com/chengzhag/Implicit3DUnderstanding

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 2: ImVoxelNet 3D monocular object detection performance on the SUN RGB-D dataset with different object insertion methods (with error range).

Setting	Insertion Position, Pose, Size	Insertion Illumination	mAP@0.25
ImVoxelNet	N/A	N/A	40.96 $\pm$ 0.4
ImVoxelNet + random insert	Random	Camera point light	37.02$\pm$ 0.4
ImVoxelNet + 3D Copy-Paste (w/o light)	Plausible position, size, pose	Camera point light	41.80$\pm$ 0.3
ImVoxelNet + 3D Copy-Paste	Plausible position, size, pose	Plausible dynamic light	**43.79** $\pm$ 0.4

Table 3: ImVoxelNet 3D monocular object detection performance on the SUN RGB-D dataset with mAP@0.15.

Setting	Insertion Position, Pose, Size	Insertion Illumination	mAP@0.15
ImVoxelNet	N/A	N/A	48.45
ImVoxelNet + 3D Copy-Paste	Plausible position, size, pose	Plausible dynamic light	**51.16**

C Discussion on Limitations and Broader Impact

Limitations. Our method, while effective, does have certain limitations. A key constraint is its reliance on the availability of external 3D objects, particularly for uncommon categories where sufficient 3D assets may not be readily available. This limitation could potentially impact the performance of downstream tasks. Moreover, the quality of inserted objects can also affect the results. Possible strategies to address this limitation could include leveraging techniques like Neural Radiance Fields (NeRF) to construct higher-quality 3D assets for different categories.

Broader Impact. Our proposed 3D Copy-Paste method demonstrate that physically plausible 3D object insertion can serve as an effective generative data augmentation technique, leading to state-of-the-art performance in discriminative downstream tasks like monocular 3D object detection. The implications of this work are profound for both the computer graphics and computer vision communities. From a graphics perspective, our method demonstrates that more accurate 3D property estimation, reconstruction, and inverse rendering techniques can generate more plausible 3D assets and better scene understanding. These assets not only look visually compelling but can also effectively contribute to downstream computer vision tasks. From a computer vision perspective, it encourages us to utilize synthetic data more effectively to tackle challenges in downstream fields, including computer vision and robotics.

References

D. Rukhovich, A. Vorontsova, and A. Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022.

X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018.

C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, and S. Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021.