Object-Centric Data Synthesis for Category-level Object Detection (Student Abstract)

Vikhyat Agarwal[1]*, Jiayi Cora Guo[2]*, Declan Hoban[3]*, Sissi Zhang[4]*, Nicholas Moran[5], Peter Cho[5], Srilakshmi Pattabiraman[5], Shantanu Joshi[2]

[1]University of Richmond, Richmond, Virginia
[2]UCLA, Los Angeles, California
[3]UC Berkeley, Berkeley, California
[4]UT Austin, Austin, Texas
[5]Analog Devices, Inc.
vikhyat.agarwal@richmond.edu

Abstract

Deep learning approaches to object detection have achieved reliable detection of specific object classes in images. However, extending a model's detection capability to new object classes requires large amounts of annotated training data, which is costly and time-consuming to acquire, especially for long-tailed classes with insufficient representation in existing datasets. We compare four distinct methods of generating synthetic data to finetune object detection models on novel object categories, particularly when limited data is available in an object-centric format (multi-view images/3D models). Our approaches are based on simple image processing techniques, 3D rendering, and image generation models, each varying in complexity and realism. We assess how our methods, which use object-centric data to synthesize realistic, cluttered images with varying contextual coherence, enable models to achieve category-level generalization in real-world data. We demonstrate significant performance boosts within this data-constrained experimental setting.

Code — https://github.com/RIPS25-Analog/OC-Synthesis

Introduction

Object detection is a computer vision task with widespread applications, and deep learning models have achieved fast and robust detection of different object classes in images. A significant obstacle to the practical adoption of such techniques is the lack of high-quality labeled training data, especially when dealing with rare object classes that are not well-represented in existing image datasets. Models trained on too little data tend to overfit and struggle to generalize to novel complex scenes (Antonelli et al. 2022).

Synthetic data generation is a popular model-agnostic method to improve performance in limited-data scenarios. Algorithms for data generation automatically provide annotations for generated images, but they differ in the levels of image realism and diversity they achieve. Previous approaches to data synthesis for object detection have employed simple image composition tools, 3D rendering, or generative image models (Westerski and Fong 2024).

Existing works often fail to address key considerations of practical applications, due to **1**, insufficient modeling of occlusion or cluttered environments; and **2**, dependence on generative models, which cannot reliably synthesize long-tail object categories that are unrepresented in image datasets (e.g. specialized industrial parts). Moreover, previous works have reported mixed findings regarding the effectiveness of incorporating extra visual context in synthetic images (Dvornik, Mairal, and Schmid 2018; Ghiasi et al. 2021). In this work we develop data synthesis methods of varying levels of contextual coherence that address limitations **1** and **2**, and we compare these methods in the particular data-limited setting where we just have access to object-centric data (complete, isolated views of objects, e.g. 3D models and multi-view images) instead of expensively annotated data of the objects placed in various environments.

Data Synthesis Methods

We now describe the four data synthesis methods (Figure 1).

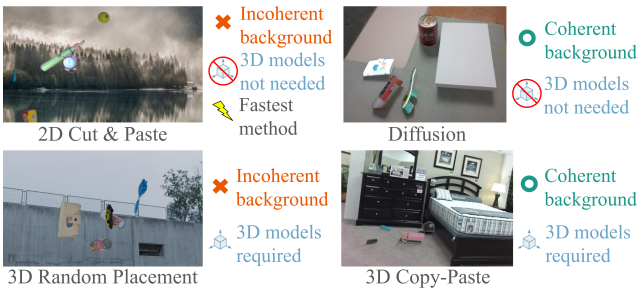

Figure 1: Qualitative comparison of our proposed data generation methods

Cut-Paste Adapting from (Dwibedi, Misra, and Hebert 2017), we simply extract object cutouts from multi-view images (or from annotated images) using a segmentation

*These authors contributed equally.

model, and then 'paste' them onto random background scenes with augmentations like rotation, scaling, occlusion, and mask blending. Our adaptation improves the object scaling and inter-object occlusion checking.

Diffusion Copy-Paste (Diffusion CP) This method randomly arranges object cutouts into a layout (similar to Cut-Paste) and uses a diffusion model to synthesize images conditioned on the object edge map and a reference background image. This creates a contextually coherent background for the cutouts, while allowing the same precise control over object position and orientation as Cut-Paste.

3D Random Placement (3D RP) This method places 3D models of objects into 360° HDRI background images, which encode environment lighting. We render multiple images from these scenes while randomly varying object placements & orientations, camera position & angle, camera exposure, and additional light sources.

3D Copy-Paste (3D CP) Adapted from (Ge et al. 2023), this method makes 3D RP more coherent by placing objects on feasible horizontal planes (identified within RGBD background images using clustering algorithms) in a collision-free manner. Our adaptation improves the collision detection and enables multi-object and multi-surface placement.

Experiments

As real-world evaluation data for our methods, we select a subset of the annotated PACE video dataset (You et al. 2024), focusing on object classes not commonly present in existing datasets. We split different instances of each class into training, validation, and test, so the model has to generalize to new instances of the same class. We generate synthetic images with the four methods using either object cutouts (extracted from annotated frames) or 3D models (provided by PACE) from the train set, and use three different data schemes to finetune a pretrained object detection model (YOLO11 (Jocher and Qiu 2024)). In the simplest **synthetic-only** scheme, we finetune on synthetic data only; in the **mixed** scheme, we finetune on a mix of synthetic data and a fraction of real-world train data (PACE video frames); and in the **sequential** scheme, we first finetune on synthetic data and then on a fraction of real train data. The sequential and mixed schemes model the practical scenario where synthetic data is used to augment limited amounts of real annotated data. Our results are summarized in Table 1.

Figure 2 shows performance in the real-only and synthetic-only settings while varying the training set size. As expected, training on a few real images outperforms training on a few synthetic images, but as we scale data generation, we see training on synthetic data only can outperform training on all 10k real images. Unlike the results in Table 1, we do not optimize hyperparameters here, to emulate the lack of a real-data validation set in the synthetic-only scheme.

To test whether performance gains hold across different model architectures, we run our sequential scheme experiments with the RT-DETR (Zhao et al. 2024) model and observe significant mAP boosts across all methods (Figure 3).

	0%	2.5%	5%	10%
Cut-Paste	44.6	55.6 / 34.8	56.1 / 46.2	56.3 / 51.1
Diff. CP	41.2	60.5 / 52.1	60.2 / 55.3	58.4 / 55.0
3D RP	**45.2**	**62.8** / 51.7	**64.0** / 52.9	**63.2** / 56.9
3D CP	31.1	58.2 / 59.8	59.6 / 53.5	54.7 / 59.7
Real only	-	52.5	57.3	57.7

Table 1: YOLO11 finetuning results (test set mAP@50) with *Sequential* / *Mixed* training schemes and varying proportions of real data. The 0% column is *Synthetic-only*.

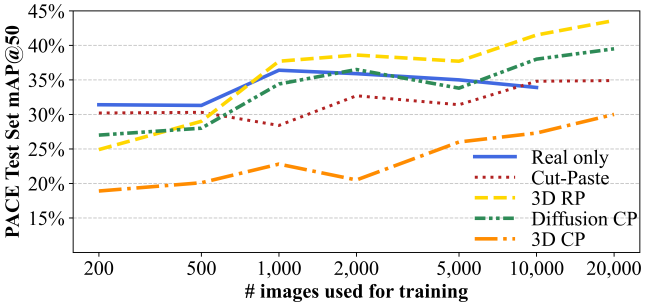

Figure 2: Head-to-Head comparison of different methods, training the model only on the given number of images of each type (e.g., only 200 Cut-Paste images).

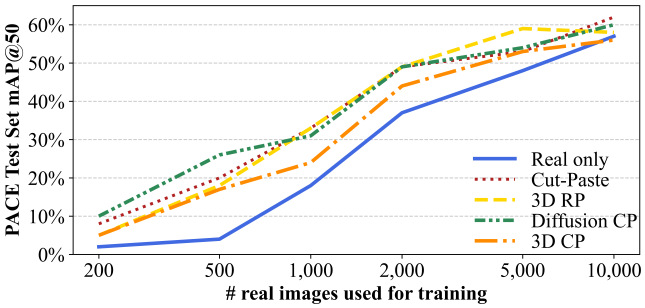

Figure 3: Sequential training of RT-DETR on different amounts of real data, compared to only training on real data.

Conclusion

We find that synthetic data yields significant performance improvements for category-level object detection, particularly by finetuning pretrained models sequentially. Comparison across four methods reveals mixed results for when modeling visual context and realism can lead to improved performance. 3D Random Placement and Diffusion Copy-Paste produce more realistic datasets than Cut-Paste by incorporating geometric, lighting, & perspective information, and contextual background information respectively, and both outperform the Cut-Paste baseline. 3D Copy-Paste combines aspects of both but fails to surpass Cut-Paste. We note that adding certain *realistic* constraints (e.g. placing objects on surfaces) can unintentionally *reduce diversity* along other aspects, such as viewpoint coverage, which may counteract the advantages of modeling the real world more faithfully. We hope our work invites further empirical study of the under-addressed yet practical object-centric setting.

Acknowledgments

This work was supported by the Institute for Pure and Applied Mathematics (IPAM) at UCLA, in conjunction with the industry sponsor Analog Devices, Inc. (ADI), through the 2025 Research in Industrial Projects for Students (RIPS) program.

References

Antonelli, S.; Avola, D.; Cinque, L.; Crisostomi, D.; Foresti, G. L.; Galasso, F.; Marini, M. R.; Mecca, A.; and Pannone, D. 2022. Few-Shot Object Detection: A Survey. *ACM Comput. Surv.*, 54(11s).

Dvornik, N.; Mairal, J.; and Schmid, C. 2018. Modeling Visual Context Is Key to Augmenting Object Detection Datasets. In *2018 ECCV Proceedings, Part XII*, 375–391.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection . In *2017 ICCV Proceedings*, 1310–1319.

Ge, Y.; Yu, H.-X.; Zhao, C.; Guo, Y.; Huang, X.; Ren, L.; Itti, L.; and Wu, J. 2023. 3D Copy-Paste: Physically Plausible Object Insertion for Monocular 3D Detection. In *37th NeurIPS Proceedings*, 17057–17071.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In *2021 CVPR Proceedings*, 2918–2928.

Jocher, G.; and Qiu, J. 2024. Ultralytics YOLO11. https://github.com/ultralytics/ultralytics.

Westerski, A.; and Fong, W. T. 2024. Synthetic Data for Object Detection with Neural Networks: State-of-the-Art Survey of Domain Randomisation Techniques. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(1): 1–20.

You, Y.; Xiong, K.; Yang, Z.; Huang, Z.; Zhou, J.; Shi, R.; Fang, Z.; Harley, A. W.; Guibas, L.; and Lu, C. 2024. PACE: A Large-Scale Dataset with Pose Annotations in Cluttered Environments. In *2024 ECCV Proceedings, Part LI*, 473–489.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. DETRs Beat YOLOs on Real-time Object Detection. In *2024 CVPR Proceedings*, 16965–16974.