Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?

Rosario Leonardi[1], Antonino Furnari[1,2], Francesco Ragusa[1,2], Giovanni Maria Farinella[1,2]
[1]FPV@IPLab - University of Catania, Italy
[2]Next Vision s.r.l. - Spinoff of the University of Catania, Italy

Abstract

In this study, we investigate the effectiveness of synthetic data in enhancing egocentric hand-object interaction detection. Via extensive experiments and comparative analyses on three egocentric datasets, VISOR, EgoHOS, and ENIGMA-51, our findings reveal how to exploit synthetic data for the HOI detection task when real labeled data are scarce or unavailable. Specifically, by leveraging only 10% *of real labeled data, we achieve improvements in Overall AP compared to baselines trained exclusively on real data of:* $+5.67\%$ *on EPIC-KITCHENS VISOR,* $+8.24\%$ *on EgoHOS, and* $+11.69\%$ *on ENIGMA-51. Our analysis is supported by a novel data generation pipeline and the newly introduced HOI-Synth benchmark which augments existing datasets with synthetic images of hand-object interactions automatically labeled with hand-object contact states, bounding boxes, and pixel-wise segmentation masks. Data, code, and data generation tools to support future research are released at:* https://fpv-iplab.github.io/HOI-Synth/.

1. Introduction

The use of synthetic data to reduce the dependence of prediction algorithms on labeled real data has been previously explored in different domains, including embodied AI [7] and autonomous driving [2]. However, the exploitation of synthetic data is currently under-explored in egocentric vision in general and hand-object interaction detection in particular, due to the challenges associated to generating accurate and photorealistic images of hand-object interactions, which requires the modeling of hands, objects and physical contact. While the use of synthetic data holds promise for reducing reliance on labeled real data in egocentric hand-object interaction detection, many questions still remain unanswered: *1) Is there a gap between real and synthetic data? 2) Where does it originate? 3) How can it be reduced? 4) Can synthetic data entirely replace real data? 5) Can synthetic data enable training in the presence of unlabeled real data? 6) Can synthetic data increase efficiency when few real data are labeled? 7) Is in-domain synthetic data, aligned to the target real domain in terms of objects and environment, beneficial?*

With the goal of advancing research in egocentric hand-object interaction detection and synthetic-to-real domain adaptation for egocentric vision, in this paper, we propose a systematic investigation to answer the questions above. To support our investigation, we propose a novel pipeline and develop a simulator able to generate synthetic images of realistic hand-object interactions in multiple environments, which are automatically labeled for the considered hand-object detection task (Figure 1-left). We generate three sets of synthetic data, paired with two popular domain-generic hand-object detection benchmarks, *EPIC-KITCHENS VISOR* [1], and *EgoHOS* [10], and a domain-specific dataset, *ENIGMA-51* [5]. We hence study three different domain adaptation tasks: *unsupervised domain adaptation*, where models are trained with synthetic data and unlabeled real data, *semi-supervised domain adaptation*, where models are trained with synthetic data, unlabeled real data, and few labeled real data, and *fully supervised domain adaptation*, where models are trained with labeled synthetic and real data (Figure 1-right). Collectively, the real and generated egocentric data define a new benchmark dataset, which we term *HOI-Synth*.

We leverage *HOI-Synth* to benchmark different approaches to domain adaptation for hand-object interaction detection based on previous literature on domain adaptation for object detection and hand-object interaction detection in multiple settings. Our analysis provides several insights into the advantages of using synthetic data for egocentric hand-object interaction detection: A) Despite advancements, there's still a gap between synthetic and real data, attributed to limitations in realism, grasping accuracy, and diversity of environments and objects; B) Domain adaptation reduces this gap: unsupervised domain adaptation yields improvements of $\sim 20-35\%$ AP; semi-supervised adaptation approaches achieve the performance of fully supervised methods on real data using only $\sim 10\%-25\%$ of the labels; fully-supervised adaptation sees a $\sim 1\%-4\%$ AP boost. C) While most of the improvements come from

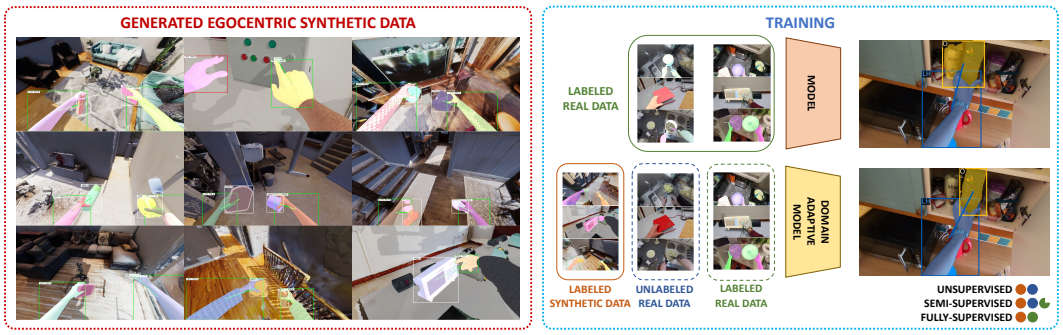

Figure 1. We study the impact of synthetic data in egocentric hand-object interaction detection. We generate and automatically label large sets of synthetic data (left) and study a set of domain adaptation scenarios in which models are trained on both synthetic and real unlabeled data, plus different amounts of labeled real data (right).

synthetic sets in the order of $10,000$ images, methods still obtain benefits as the amount of synthetic data is increased up to $30,000$; D) In-domain synthetic data significantly enhances unsupervised domain adaptation ($\sim +20\%$ AP), while its advantage in semi and fully supervised adaptation seems limited.

2. The HOI-Synth Benchmark

HOI-Synth Data Generation Pipeline and Simulator Figure 2 shows a scheme of the proposed data generation pipeline, which is composed of three main steps. Our pipeline relies on state-of-the-art datasets and components to enable an accurate generation of egocentric images of hand-object interactions [6, 8, 9]. We first select a random hand-object grasp from the DexGraspNet dataset [9], which is fit to a randomly generated human model and integrated with the appropriate object mesh specified in the hand-object grasp [8] (Figure 2-a). We then select a random environment from the HM3D dataset [6] and place the human-object model in the environment (Figure 2-b). We finally place a virtual camera at human eye level to capture the scene from the first-person point of view. For each generated interaction, the simulator annotates the bounding boxes and the segmentation masks of the hands and interacted objects, the hand contact state, as well as the hand-object relations (see Figure 2-c). We developed the pipeline in the Unity3D framework and implemented a hand-object interaction simulator, which will support future research on synthetic data generation for egocentric vision.

Datasets The HOI-Synth benchmark extends three established datasets of egocentric images designed to study hand-object interaction detection, EPIC-KITCHENS VISOR [1], EgoHOS [10], and ENIGMA-51 [5], with automatically labeled synthetic data obtained through the proposed generation pipeline. Table 1 reports statistics of the training section of the HOI-Synth benchmark dataset, including the number of real and synthetic images, annotated hands, objects and HOIs. We use the official validation and test sets of the respective datasets for evaluation.

Table 1. Statistics of the training sets considered in our HOI-Synth benchmark.

Dataset	Images	Hands	Objects	HOI
VISOR [1]	32,857	52,906	42,785	42,787
Synthetic	30,259	60,098	45,219	45,353
EgoHOS [10]	8,107	15,015	11,393	13,659
Synthetic	8,107	16,101	12,170	12,129
ENIGMA-51 [5]	3,479	5,075	4,343	4,344
Synthetic-In-Domain	16,773	25,444	16,637	16,773
Synthetic-out-domain	20,321	40,135	27,499	27,370

3. Experimental Analysis and Results

We consider six different approaches to hand-object segmentation based on *VISOR HOS* [1]: (1) *Synthetic-Only*, (2) *Unsupervised Domain Adaptation (UDA)*, (3) *Real-Only*, (4) *Synthetic + Real* (5) *Semi-Supervised Domain Adaptation (SSDA)*, and (6) *Fully-Supervised Domain Adaptation (FSDA)*.

Evaluation measures Following [1], we evaluate performance using *COCO Mask AP* [4]. In particular, we adopted the Hand + Object (Overall) AP which assesses the correctness of the predicted hands and object bounding boxes of hands, the hand-state (contact vs. no contact) and the offset vector representing the relation between the hand and the active object. We also break down performance using Mask APs measures evaluating specific aspects of the predictions: Hand (H), Hand + Side (H+S), Hand + Contact (H+C), and Object (O).

Results on VISOR Table 2 shows the results on the validation set of EPIC-KITCHENS VISOR [1].

Figure 2. **The proposed data generation pipeline.** (a) An object-grasp pair is selected from DexGraspNet [9] and integrated with a randomly generated human model. (b) The human + object model is placed in an environment randomly selected from the Habitat-Matterport 3D dataset [6]. (c) Egocentric data of hand-object interactions is generated and automatically labeled.

Table 2. Results on the EPIC-KITCHENS VISOR validation set considering different real data settings available in training. Yellow rows indicate baseline models in each configuration, while green rows highlight models trained with synthetic and real data. In each group, the **best results** are in bold, while the best results among the models trained with synthetic and real data are underlined. Overall enhancements are shown in blue, indicating improvements of the models trained with synthetic and real data over the baseline.

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	09.88	28.41	24.89	08.64	01.23
	UDA	**33.33**	**80.16**	**65.98**	**33.47**	**8.35**
Absolute Improvement		**+23.45**	+51.75	+41.09	+24.83	+7.12

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
10% (3,286 images)	Real-Only	38.55	87.45	**83.27**	51.98	19.47
	Synthetic+Real	37.62	86.39	82.85	**52.25**	**23.03**
	SSDA	**44.22**	**89.05**	80.77	46.83	20.41
Absolute Improvement		**+5.67**	+1.60	−0.42	+0.27	+3.56
25% (8,215 images)	Real-Only	37.90	90.14	**85.66**	53.99	17.85
	Synthetic+Real	38.19	89.98	84.67	**55.88**	18.49
	SSDA	**45.55**	**90.37**	84.42	52.59	**22.15**
Absolute Improvement		**+7.65**	+0.23	−0.99	+1.89	+4.30
50% (16,429 images)	Real-Only	38.15	91.16	**86.05**	52.28	17.92
	Synthetic+Real	43.52	**91.34**	85.85	54.09	19.06
	SSDA	**46.47**	90.94	85.73	**58.02**	**23.49**
Absolute Improvement		**+8.32**	+0.18	−0.20	+5.74	+5.57

c) Fully-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
100% (32,857 images)	Real-Only	45.33	**92.25**	88.54	**59.24**	24.23
	Synthetic+Real	44.52	91.45	**88.94**	56.55	**27.77**
	FSDA	**46.48**	91.83	87.65	57.63	24.03
Absolute Improvement		**+1.15**	−0.42	+0.40	−1.61	+3.54

Results on EgoHOS Table 3 reports the results on the test set of EgoHOS [10].

Results on ENIGMA-51 Table 4 reports the results on the test set of ENIGMA-51 [5]. In this case, we also compare performance when in-domain and out-domain synthetic data are used.

The full discussion of the results is available in the complete version of the paper [3] at the following link: https://arxiv.org/abs/2312.02672.

Table 3. Results on the EgoHOS [10] test set.

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	07.16	18.25	15.93	05.33	01.24
	UDA	**28.16**	**70.30**	**59.21**	**20.84**	**09.65**
Absolute Improvement		**+21.00**	+52.05	+43.28	+15.51	+8.41

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
10% (857 images)	Real-Only	28.44	76.28	68.92	35.84	16.59
	Synthetic+Real	28.74	77.15	71.64	39.25	17.33
	SSDA	**36.68**	**83.25**	**73.72**	**47.20**	**22.40**
Absolute Improvement		**+8.24**	+6.97	+4.80	+11.36	+5.81
25% (2,026 images)	Real-Only	33.73	78.94	70.62	41.67	21.83
	Synthetic+Real	33.78	79.60	71.61	46.11	19.87
	SSDA	**37.16**	**83.79**	**74.28**	**49.00**	**23.82**
Absolute Improvement		**+3.43**	+4.85	+3.66	+7.33	+1.99
50% (4,379 images)	Real-Only	36.30	81.82	73.63	47.27	25.73
	Synthetic+Real	34.30	82.54	74.03	47.92	23.47
	SSDA	**39.85**	**85.17**	**76.80**	**52.58**	**26.90**
Absolute Improvement		**+3.55**	+3.97	+3.17	+5.31	+1.17

c) Fully-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
100% (8,758 images)	Real-Only	36.16	84.39	76.24	51.81	26.46
	Synthetic+Real	34.68	84.56	71.56	49.72	23.16
	FSDA	**39.61**	**85.58**	**76.80**	**51.99**	**27.05**
Absolute Improvement		**+3.45**	+1.19	+0.56	+0.18	+0.59

4. Discussion and Conclusion

With the proposed analysis we aimed to address several questions.

Is there a gap between synthetic and real data? Where does it originate? How can it be reduced? Despite progress in realistic data generation, a gap remains between synthetic and real data. Our analysis offers insights into the extent of such gap, which is in the order of $30\% - 40\%$ depending on the dataset. In the context of VISOR, the estimated gap (35.45%) is narrowed by unsupervised domain adaptation to 12.00% and further shrunk to 1.11% adopting semi-supervised domain adaptation strategies. Similar considerations can be made for the other datasets. We suggest this gap is caused by the photo-realism of generated synthetic data, the diversity of context-aware characteristics (as shown by

Table 4. Results on the ENIGMA-51 [5] test set.

a) Unsupervised Setting

% Real Labeled Data	Approach	In-domain	Overall	H	H+S	H+C	O
0%	Synthetic-Only		00.21	01.07	00.11	00.03	00.99
	Synthetic-Only	✓	12.85	56.05	35.14	15.24	4.79
	UDA		6.87	42.81	14.52	7.97	3.29
	UDA	✓	**34.78**	**78.83**	**70.91**	**28.14**	**25.84**
Absolute Improvement			**+21.93**	+22.78	+35.77	+12.90	+21.05

b) Semi-supervised Setting

% Real Labeled Data	Approach	In-domain	Overall	H	H+S	H+C	O
10% (347 images)	Real-Only	✓	45.39	81.25	76.22	37.96	39.53
	SSDA		**57.08**	**85.40**	**78.62**	**43.56**	**46.97**
	SSDA	✓	56.69	84.58	78.42	41.17	46.50
Absolute Improvement			**+11.69**	+4.15	+2.40	+5.60	+7.44
25% (870 images)	Real-Only	✓	51.83	82.95	78.70	43.52	45.25
	SSDA		58.17	**84.99**	**80.41**	**46.31**	49.34
	SSDA	✓	**59.48**	84.85	80.30	44.24	**49.37**
Absolute Improvement			**+7.65**	+2.04	+1.71	+2.79	+4.12
50% (1,739 images)	Real-Only	✓	57.62	84.65	80.43	47.41	48.79
	SSDA		**63.25**	**85.67**	82.00	**52.20**	**52.56**
	SSDA	✓	61.93	85.12	**82.01**	48.96	51.94
Absolute Improvement			**+5.63**	+1.02	+1.58	+4.79	+3.77

c) Fully-supervised Setting

% Real Labeled Data	Approach	In-domain	Overall	H	H+S	H+C	O
100% (3,479 images)	Real-Only	✓	63.84	85.01	81.05	52.32	51.35
	FSDA		**64.41**	**85.94**	**82.91**	**54.13**	52.50
	FSDA	✓	64.20	85.37	82.45	51.60	**53.30**
Absolute Improvement			**+0.57**	+0.93	+1.86	+1.81	+1.95

results with in/out-domain synthetic data) and hand-object interactions.

Can synthetic data entirely replace real data? Our study suggests that synthetic data cannot yet entirely replace real data for egocentric hand-object interaction detection, with synthetic-only baselines achieving poor results in all scenarios.

Can synthetic data enable training in the presence of unlabeled real data? While synthetic data cannot entirely replace real data, we show that it greatly improves models' performance in the presence of unlabeled real data. Indeed, significant gains are obtained by UDA across all scenarios, when compared to a synthetic-only baseline, while the gap with respect to fully supervised baselines is narrowed.

Can synthetic data increase efficiency when few real data are labeled? When different amounts of real labeled data are exploited together with synthetic data, SSDA and FSDA models obtain improvements in *Overall AP* over baselines trained on real data only in the considered benchmark. Notably, the performance gap diminishes as the quantity of real data increases: from +23.45% (0% of real data) to +1.15% (100% of real data) in VISOR, from +21.00% (0% of real data) to +3.45% (100% of real data) in EgoHOS and from +21.93% (0% of real data) to +2.33% (100% of real data) for ENIGMA-51. These results highlight the effectiveness of using synthetic data when real labeled data are scarce.

What scale of synthetic data is needed Our findings reveal that models benefit from large quantities of synthetic data. For instance, in the context of VISOR, a plateau is reached when 22K-30K synthetic images are included for training.

Is in-domain synthetic data beneficial? Our analysis shows that in-domain data is highly beneficial in unsupervised settings, where it helps narrow down the domain gap. For instance, in the ENIGMA-51 dataset, using in-domain synthetic data only allows to obtain an overall AP of 12.85, about $+10\%$ with respect to out-domain data. With UDA, performance jumps to 34.78%, a major increase. With few real labeled data, choice of in-domain data is less crucial, with models achieving comparable performance, regardless of the training data source.

References

[1] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS*, pages 13745–13758, 2022. 1, 2

[2] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *ICCV*, 2021. 1

[3] Rosario Leonardi, Antonino Furnari, Francesco Ragusa, and Giovanni Maria Farinella. Are synthetic data useful for egocentric hand-object interaction detection? In *European Conference on Computer Vision*, 2024. 3

[4] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2

[5] Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, and Giovanni Maria Farinella. Enigma-51: Towards a fine-grained understanding of human-object interactions in industrial scenarios. *arXiv preprint, 2309.14809*, 2023. 1, 2, 3, 4

[6] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *NeurIPS*, 2021. 2, 3

[7] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, pages 9339–9347, 2019. 1

[8] Unity. Synthetichumans package, 2022. https://github.com/Unity-Technologies/com.unity.cv.synthetichumans. 2

[9] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *CVPR*, pages 11359–11366, 2023. 2, 3

[10] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *ECCV*, pages 127–145, 2022. 1, 2, 3