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

Rosario Leonardi[1], Antonino Furnari[1,2], Francesco Ragusa[1,2], and Giovanni Maria Farinella[1,2]

[1] Department of Mathematics and Computer Science, University of Catania, Italy
[2] Next Vision s.r.l., 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/`.

**Keywords:** Synthetic Data · Egocentric HOI · Domain Adaptation

## 1 Introduction

Understanding how humans interact with the surrounding objects from egocentric images is a fundamental challenge in computer vision, with applications in diverse domains including collaborative robotics [4, 14], industrial behavior understanding [42, 47], human-computer interaction [32], and healthcare [1]. Previous works investigated the task of understanding human-object interactions from an egocentric perspective in different forms, including action recognition [8], object state-change detection [18], and hand-object interaction forecasting [29]. A line of work developed around the goal of identifying the actively manipulated object, the presence of hands, and the contact state between hands and objects [10, 24, 42, 49], which is generally referred to as *Hand-Object Interaction (HOI) detection*. Despite the progress in model design granted by the availability of egocentric benchmarks such as EPIC-KITCHENS [8, 9] and VISOR [10], performance in real application scenarios is closely tied to the availability of large amounts of annotated real-world and domain-specific data [42]. In addition, the need for spatial and interaction annotations makes acquiring and labeling such data an expensive and time-consuming process.

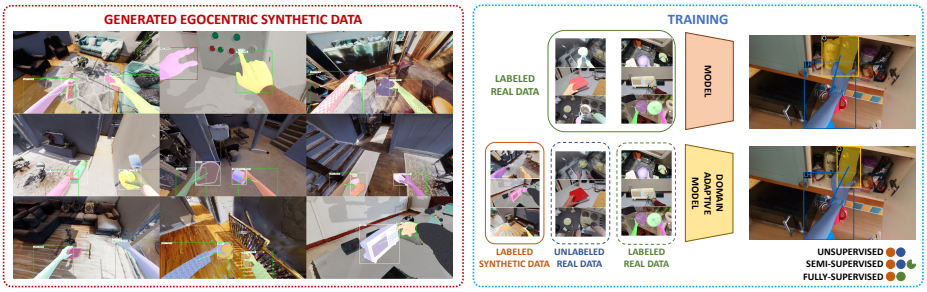


**Fig. 1:** We study the impact of synthetic data in egocentric hand-object interaction detection. We generate and automatically label large sets of synthetic data (left) and study a set of domain adaptation scenarios in which models are trained on both synthetic and real unlabeled data, plus different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

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 [22,46,57] and autonomous driving [13,15]. 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. As a result, 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) What scale of synthetic data is needed? 8) 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* [10], and *EgoHOS* [59], and a domain-specific dataset, *ENIGMA-51* [43]. 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 [17, 27, 30, 52] and hand-object interaction detection [10] in multiple settings. Our analysis provides several insights into the advantages of using properly generated synthetic data for egocentric hand-object interaction detection: A) Despite the progress in realistic data generation, a domain gap between synthetic and real data still exists, with models trained only on synthetic data lagging behind models trained on real data by large margins (a gap of $\sim 30\% - 40\%$ in AP), which we attribute to limits in photorealism, accuracy of grasping, and diversity of environments and objects; B) We show that performing domain adaptation allows to reduce the synth-real gap in the settings of *unsupervised domain adaptation*, where methods obtain large improvements of $\sim 20\% - 35\%$ AP when exposed to unlabeled real data, *semi-supervised domain adaptation*, where models achieve performance comparable to approaches trained only on real data by using only $\sim 10\% - 25\%$ of real data labels, and *fully-supervised domain adaptation*, where combining labeled real and synthetic data improves performance by $\sim 1\% - 4\%$ AP; C) While most of the improvements comes from 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) When available, in-domain synthetic data including objects and environments aligned to those of the target real domain, greatly improves performance in the unsupervised domain adaptation setting, with gains of up to $\sim +20\%$ AP, while advantages of in-domain synthetic data are limited if few real labeled data are available for semi-supervised adaptation.

The contributions of this work are: 1) A systematic investigation of the egocentric hand-object interaction detection task assessing the effectiveness of properly generated synthetic data in three domain adaptation settings. Our investigation provides insights into the usefulness of synthetic data and will inform future model and experimental designs; 2) *HOI-Synth*, a novel benchmark for unsupervised, semi-supervised and fully-supervised domain adaptation which, for the first time, enables the study of synthetic-to-real domain adaptation for egocentric hand-object interaction detection. With the benchmark, we include several baseline results showcasing the potential of synthetic data in this domain and providing a basis for future comparisons and advances; 3) A novel data generation pipeline and a developed simulator, which will be able to support future investigations in the exploitation of synthetic data for egocentric vision. To enable future research on this topic, we publicly release the generated data, the simulator, and all the code required to reproduce the results.

## 2 Related Work

**Hand-Object Interaction Detection** The authors of [49] were among the first to frame hand-object interaction detection as the task of detecting hands, inferring contact states, and detecting manipulated objects, considered both egocentric and third-person vision images in which hands and objects are clearly

visible. This task formulation was later extended in [5] adding object and hands segmentation, secondary objects, and grasp type prediction. Similar investigations have been also performed considering purely egocentric vision scenarios. The authors of [31] proposed an architecture leveraging specific egocentric cues such as hand poses and object masks. The authors of [42] defined Egocentric Human-Object Interaction (EHOI) detection as the task of identifying manipulated objects and predicting interaction verbs for each of them. Hand-object interaction detection has also been considered in an industrial scenario in which manipulated object classes are known beforehand [24, 43]. Other investigations considered different tasks related to hand-object interaction understanding such as State Change Object Interaction Detection [18] or active object detection [16]. Previous investigations assumed different task formulations, which makes it hard to compare methods and assess progress. Recently, the authors of [10] provided a task formulation termed Hand-Object Segmentation (HOS) together with the EPIC-KITCHENS VISOR dataset, thus setting a standard benchmark for hand-object interaction detection in egocentric vision. HOS consists in estimating the contact relation between the hands and the objects and segmenting them given a single RGB frame. A similar formulation was proposed in [59] together with the EgoHOS dataset. In this paper, we adopt the HOS formulation and the baseline model of [10] to perform experiments on three datasets designed for hand-object interaction detection: VISOR [10], EgoHOS [59], and ENIGMA-51 [43].

**Simulators for Synthetic Visual Data Generation** Previous research introduced simulators to generate synthetic data that mimics the behavior of real-world agents, such as cars and robots. Some examples include CARLA [13], Gibson [25, 56], Habitat [33, 50], Omniverse [36], and Isaac Sim [37]. Additionally, with advancements in graphics, game engines have been exploited for synthetic data generation for tasks such as pedestrian detection and tracking in urban scenarios [12, 15], and safety monitoring in construction sites [41]. Other studies proposed simulators specifically crafted to represent human agents moving in the scene [38] and interacting with objects [23]. Specifically, [23] provides accurate modelling of the physics of the world and object manipulation actions, but it does not model hands, which are not visible in the scene. While these works advanced the understanding of synthetic data for computer vision tasks, no prior work studied the use of synthetic data for egocentric hand-object interaction detection, mainly due to the challenges associated with accurate modelling of environments, objects, and grasping. Based on recent advances in these fields [44, 55], we propose a novel data generation pipeline and develop a simulator allowing to obtain realistic and diverse images of hand-object interactions.

**Synthetic Data for Hand-Object Interaction Understanding** Few previous efforts used synthetic data to address tasks related to understanding egocentric hand-object interactions. The authors of [19] introduced the ObMan dataset designed for the joint reconstruction of hands and manipulated objects. The authors of [20] introduced a dataset of hand-object interactions designed for hand-object affordance understanding. Recently, [58] exploited diffusion models for generating hand-object interactions synthetic datasets. These works did not


**Fig. 2: The proposed data generation pipeline.** (a) An object-grasp pair is selected from DexGraspNet [55] 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 [44]. (c) Egocentric data of hand-object interactions is generated and automatically labeled. Labels include bounding boxes and segmentation masks of hands and interacted objects, contact-state, and hand-object relations.

tackle hand-object interaction detection and did not consider the generation of data with fine-grained labels required to address the task To enable the exploitation of synthetic data in this domain, we introduce a new benchmark comprising real images, paired with photorealistic labeled synthetic images annotated with labels useful for hand-object detection, such as 2D bounding boxes, semantic segmentation masks, depth maps, and hand-object relations.

**Domain Adaptation** Domain Adaptation (DA) techniques have gained significant attention in recent years [2, 7, 45, 60]. These methods rely on different strategies to reduce the domain gap between a source and a target domain, such as adversarial training [17, 53], transfer learning [17, 53] and pseudo-labeling [3, 11, 27, 30, 52]. While domain adaptation has been extensively studied in egocentric vision for different tasks, such as Action Recognition [34, 40], Person Re-identification [6], Video Retrieval [35], Ego-Exo Adaptation [26] and Object Detection [39], a study on DA for the HOI detection task is missing. Unlike standard object detection, HOI detection involves identifying hands and active objects and understanding the specific relations between them. Hand-object interactions can be significantly influenced by the variability in shapes and sizes of objects, the diverse poses of human hands as well as by different contexts, such as kitchens and industrial laboratories. We provide HOI-Synth, the first benchmark explicitly designed to support the study of domain adaptation for hand-object interaction detection.

## 3 The HOI-Synth Benchmark

In this section, we describe the HOI-Synth benchmark that we introduce to enable the study of synth-to-real domain adaptation for egocentric hand-object interaction detection. HOI-Synth is obtained by complementing three existing egocentric hand-object benchmarks with synthetic data, which is possible thanks to a novel data generation pipeline and a hand-object interaction simulator,

which we release together with the benchmark to support future investigations on the use of synthetic data in egocentric vision.

### 3.1 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 [44, 48, 54, 55]. We first select a random hand-object grasp from the DexGraspNet dataset [55], which is fit to a randomly generated human model and integrated with the appropriate object mesh specified in the hand-object grasp [54] (Figure 2-a). We then select a random environment from the HM3D dataset [44] 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 be publicly released to support future research on synthetic data generation for egocentric vision. The supplementary material reports details on the generation pipeline and visual examples of the generated data.

### 3.2 Datasets

The HOI-Synth benchmark extends three established datasets of egocentric images designed to study hand-object interaction detection, EPIC-KITCHENS VISOR [10], EgoHOS [59], and ENIGMA-51 [43], with automatically labeled synthetic data obtained through the proposed generation pipeline.

**EPIC-KITCHENS VISOR [10]** contains 36 hours of egocentric videos from EPIC-KITCHENS-100 [8], including 32,857 training images and pixel-wise annotations for 42,787 hand-object relations. We complement this dataset with 30,259 synthetic images including 45,353 HOIs.

**EgoHOS [59]** includes 8,107 egocentric training images of HOIs sparsely sampled from videos belonging to EGO4D [18], THU-READ [51], EPIC-KITCHENS [9], and other egocentric videos of people playing escape rooms. The dataset is labeled with pixel-wise annotations of 13,659 hand-object relations. We complement this dataset with 8,107 synthetic images including 12,129 HOIs.

**ENIGMA-51 [43]** is an egocentric dataset of subjects following instructions to repair electrical boards in an industrial laboratory. The dataset contains 3,479 training images with pixel-wise annotations of 13,659 hand-object interactions. The dataset also provides 3D models of the manipulated objects and the industrial laboratory. We complement this dataset with two sets of synthetic images: an in-domain set and an out-domain set. The in-domain set is generated using the 3D models of the environment and objects provided by the authors, thus obtaining synthetic images aligned to the real data. The out-domain set

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

| Dataset | Images | Hands | Objects | HOI |
|---|---|---|---|---|
| VISOR [10] | 32,857 | 52,906 | 42,785 | 42,787 |
| Synthetic | 30,259 | 60,098 | 45,219 | 45,353 |
| EgoHOS [59] | 8,107 | 15,015 | 11,393 | 13,659 |
| Synthetic | 8,107 | 16,101 | 12,170 | 12,129 |
| ENIGMA-51 [43] | 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 |

contains images of hand-object interactions in generic environments and with generic objects, akin to those generated to complement VISOR and EgoHOS.[3]

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.

## 4 Experimental Analysis and Results

We use *VISOR HOS* [10] as a baseline for our experiments. This method is based on the *PointRend* [21] instance segmentation network with the addition of three modules to detect the hand side, contact state ("contact" or "no contact"), and an offset vector that links the hand to the interacted object[3]. We consider five different approaches to hand-object segmentation based on *VISOR HOS*:

**Synthetic-Only** The *VISOR HOS* model is trained using only synthetic data and tested directly on real data. Experiments with this approach aim to assess whether synthetic data can entirely replace real data.

**Unsupervised Domain Adaptation (UDA)** It replicates the *VISOR HOS* architecture within the Adaptive Teacher (AT) unsupervised domain adaptation framework proposed in [27]. While AT was originally designed to tackle cross-domain object detection, we adapted it to perform HOI detection by adding modules to estimate hand side, contact state, offset vector, and segmentation masks. The model is hence trained using labeled synthetic data and unlabeled real data following an unsupervised domain adaptation scheme. This approach aims to assess whether current domain adaptation techniques in conjunction with high-quality labeled synthetic data allow avoiding labeling real data.

**Real-Only** It consists in training the *VISOR HOS* model on labeled real data only. We experiment with different amounts of labeled real data to assess how much performance depends on the scale of training data when synthetic images are not available. This method provides baseline performance and corresponds to the standard fully supervised setup.

[3] See the supplementary material for examples of in-domain and out-domain generated images and for additional details about architectures and training setups.

**Synthetic + Real** It consists in pre-training the *VISOR HOS* model on labeled synthetic data and fine-tuning it on labeled real data. Also in this case, we experiment with different amounts of labeled real data. These experiments aim to assess the potential of labeled synthetic data to reduce the amount of real data required for training, without any explicit synth-to-real adaptation.
**Semi-Supervised Domain Adaptation (SSDA)** It consists in training the Adaptive Teacher model with labeled synthetic data, unlabeled real data, and a set of labeled real data. We experiment with different proportions of labeled real data (i.e., 10%, 25%, 50%). To allow the Adaptive Teacher model to work in a semi-supervised regime, we merge the set of labeled synthetic data with the labeled real data. These experiments aim to assess whether synthetic data can improve results when only some real data are labeled.
**Fully-Supervised Domain Adaptation (FSDA)** We train the AT model merging labeled synthetic and all real data, while also performing domain adaptation. This approach aims to assess whether synthetic data can improve state-of-the-art results, even in the presence of large quantities of real labeled data.
**Evaluation measures** Following [10], we evaluate performance using *COCO Mask AP* [28]. 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).

### 4.1 Results on VISOR

Table 2 shows the results on the validation set of EPIC-KITCHENS VISOR [10].[4] When no real labeled data are considered (Table 2-a), training the model only on synthetic data leads to poor performance, with an overall AP of 9.88%, which is not comparable to results achieved in fully supervised settings (45.33% when all real data are considered, as shown by *Real-Only* in Table 2-c). This highlights that, despite the photorealism of state-of-the-art data generation pipelines, there is still a consistent gap between synthetic and real data. Adopting the Unsupervised Domain Adaptation (UDA) settings significantly improves model performance across all the evaluation criteria compared to the *Synthetic-Only* approach. We observe an absolute improvement of +23.45% for the Overall Mask AP. Improvements are noticeable also across the different breakdown Mask APs: +51.75%, +41.09%, +24.83% for hand-dependent APs, and +7.12% for Object AP. This confirms the usefulness of synthetic data when Unsupervised Domain Adaptation approaches are used to mitigate the synthetic-real domain gap. When different percentages of real labeled data (i.e., 10%, 25% and 50%) are considered in the semi-supervised setting (Table 2-b), models trained on synthetic and real

[4] Note that, in our implementation, the results of the HOS model differ from those reported in [10] because, for fair comparisons, we adopted a batch size of 4, the largest batch size achievable with domain adaptation models in our configuration.

**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 |

data either via pre-training (*Synthetic+Real*) or semi-supervised domain adaptation (*SSDA*) achieve consistently higher performance with respect to baselines trained only with real data. In particular, the *SSDA* approach achieves improvements of +5.67%, +7.65%, and +8.32% considering the Overall Mask AP metric when 10%, 25%, and 50% of labeled real data are considered respectively. Significant improvements are observed when considering Object Mask AP (O) (+3.59%, +4.30% and +5.57%), and H+C Mask AP (+0.27%, +1.89% and +5.74%), highlighting that synthetic data enhances the detection of active objects and improves the prediction of the hand contact state. Results are comparable with respect to the H and H+S measures, with improvements in the $[-0.99, +1.6]$ range, due to the fact that there is less room for improvement in these measures (all numbers in the $80\% - 90\%$ range) and that real-only tends to overfit to these sub-tasks, while reaching suboptimal overall results. When

**Table 3:** Results of different semi-supervised adaptation approaches trained with synthetic data and 25% EPIC-KITCHENS VISOR labeled training data.

| Method | Overall | H | H+S | H+C | O |
|---|---|---|---|---|---|
| Synthetic + Real | 38.19 | 89.98 | **84.67** | 55.88 | 18.49 |
| MT [52] | 43.69 | 88.78 | 84.40 | **60.94** | 21.89 |
| MT+GRL [17] | 43.97 | 88.64 | 84.27 | 58.21 | 21.82 |
| UT [30] | 44.32 | **90.60** | 84.49 | 52.55 | 22.11 |
| AT [27] | **45.55** | 90.37 | 84.42 | 52.59 | **22.15** |

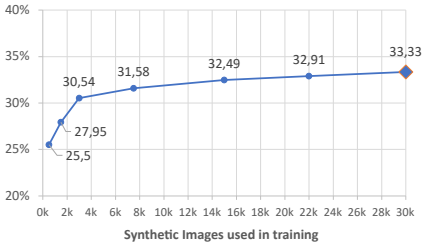


**Fig. 3:** UDA Overall AP on VISOR for different amounts of synthetic data.

we consider a Fully-supervised configuration (Table 2-c), *FSDA* improves the results over *Real-Only* by a +1.15% in Overall Mask AP and by a +3.54% in Object Mask AP, with comparable performance on H, H+S, and H+C, with improvements in the range $[-1.61, +0.4]$. It is worth noting that with as little as 25% labeled real training data, which corresponds to only 8,215 images, *SSDA* achieves an Overall Mask AP value of 45.55%, i.e., a +7.65% with respect to *Real-Only* and even a +0.22% with respect to the fully supervised baseline trained with 100% of all labeled real data (32,857 images).

**Benchmark of Domain Adaptation Approaches** In Table 3, we compare the performance of different choices of semi-supervised domain adaptation approaches on EPIC-KITCHENS VISOR when 25% of real labeled data are considered. We choose this setup as a challenging benchmark for semi-supervised domain adaptation when the amount of labeled real data is limited. We considered the following methods: *Mean Teacher [52]* (MT), *Mean Teacher + Adversarial Loss [17]* (MT+GRL), *Unbiased Teacher [30]* (UT), *Adaptive Teacher [27]* (AT). In all cases, we adapted the methods to perform HOS by including appropriate layers for the prediction of hand side, contact state, offset vector, and segmentation masks. Methods are compared to the *Synthetic + Real* baseline. *MT* and *MT+GRL* obtained the worst and second-worst results in the *Overall AP* measure (43.69% and 43.97%), but achieved the best and second best results according to the *AP Hand+Contact* measure (60.94% and 58.21%). *UT* obtains the second best result in the *Overall AP* (44.32%), but also best result in the *AP Hand* metric (90.60%). It's worth noting that *UT* outperforms the second-best results of *AT* by a thin margin (+0.23%) considering the *Hand AP*, while *AT* excels in other breakdown metrics. Finally, *AT* outperforms competitors in terms of the *Overall AP*, surpassing UT by +1.23%. Additionally, when considering the *AP Object* measure, *AT* achieves the best result of 22.15%. Results confirm the superior performance of the selected *AT* method according to the *Overall AP* performance measure with respect to the other domain adaptation strategies, despite not always achieving best results in breakdown metrics. This also suggests space for improvement in the proposed benchmark.

**Scale of Synthesized Data** Given that synthesized images can be easily generated at low cost and in large quantities using the proposed tool, we endeavoured

**Table 4:** Results on the EgoHOS [59] 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 |

to determine the scale of synthesized data required to maximize or plateau model performance. To address this issue, we trained our UDA approach on VISOR using different amounts of labelled synthetic data and 25% of real labeled data. Results reported in Figure 3 show how the model benefits from integrating additional quantities of synthetic data, approaching a plateau between 22k and 30k training images.

### 4.2 Results on EgoHOS

Table 4 reports the results on the test set of EgoHOS [59]. Also in this case, using only synthetic data (Table 4-a) does not allow to achieve satisfactory performance, highlighting the existence of a domain gap between real and synthetic data. Indeed, synthetic-only achieves an Overall AP of 7.16%, a $\sim$ 20% drop with respect to a fully supervised baseline trained on 100% labeled real data (Table 4-c). *UDA* significantly improves over *Synthetic-Only*, achieving +21.00% on the Overall Mask AP, and improvements of +52.05%, +43.28%, +15.51%, and +8.41% across the breakdown metrics (H, H+S, H+C and O). In the semi-supervised settings (Table 4-b), both *Synthetic+Real* and *SSDA* improve over *Real-Only*, with major improvements obtained by SSDA, which obtains the best

**Table 5:** Results on the ENIGMA-51 [43] 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 |

*Overall APs*, surpassing *Real-Only* by +8.24%, +3.43% and +3.55% when 10%, 25% and 50% amounts of real labeled data are considered. In the fully supervised settings (Table 4-c), *FSDA* consistently achieves best results across all measures, surpassing *Real-Only* by +3.45% in *Overall AP*. Notably, *SSDA* trained with only 25% real labeled data, corresponding to 2,026 images, obtains an improvement of +0.52% with respect to *Real-Only* trained with 100% real labeled data (8,758 images) according to the *Overall AP* measure. These results confirm the effectiveness of synthetic data in reducing the need for real labeled data, and as a way to improve performance over standard fully supervised methods.

### 4.3 Results on ENIGMA-51

Table 5 reports the results on the test set of ENIGMA-51 [43]. In this case, we also compare performance when in-domain and out-domain synthetic data are used. In the unsupervised settings (Table 5-a), using only generic synthetic data leads to poor performances, confirming the domain gap between synthetic and real images also in this case. Using in-domain synthetic real data greatly reduces the gap, with the Overall AP passing from 0.21% to 12.85%. The UDA approach improves results both when paired with generic and in-domain data. In the last

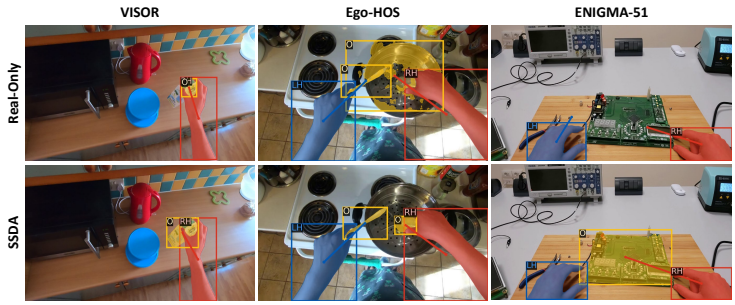


**Fig. 4:** Qualitative examples of *Real-Only* and *SSDA* on the three datasets. *SSDA* achieves better object segmentation and detection performance across the datasets.

case, UDA achieves an improvement of +21.93% in Overall AP, with improvements also in the breakdown APs: +22.78% (H), +35.77% (H+S), +12.90% (H+C) and +21.05% (O). The choice of the synthetic data source (in-domain vs. out-domain) is crucial, impacting the performance of models in this unsupervised setting. The UDA approach, trained with in-domain synthetic data, outperforms the same model trained with out-domain synthetic data by 27.91% (*Overall AP*). This result highlights that in-domain information contained in the generated images helps the detection of hand-object interactions when a specific domain is considered. In the semi-supervised setting (Table 5-b), *SSDA* systematically outperforms the baseline obtaining gains across *Overall Mask AP* of *+11.69%*, *+7.65%* and *+5.63%* when *10%*, *25%* and *50%* real labeled data are considered. Gains are also observed across all measures in the fully supervised setting (Table 5-c), e.g., with a +0.57% in overall AP, and a +1.95% in Object AP. Interestingly, the choice of in-domain vs. out-domain synthetic data in the semi- and fully-supervised settings is not as crucial as in the case of unsupervised domain adaptation, with both data sources achieving overall similar performance across the different AP measures, suggesting that even small quantities of real labeled can bridge the gap between out-domain synthetic and real data, hence making the generation of in-domain data less critical.

### 4.4 Qualitative results

Figure 4 reports qualitative examples comparing *SSDA* with respect to *Real-Only* when 25% real labeled data are considered. In the VISOR example (first column), *SSDA* (second row) obtains better object segmentation than *Real-Only* (first row). This behavior can also be observed in the EgoHOS example (second column), where *SSDA* better detects and segments the objects involved in the interaction. In the ENIGMA-51 example (third column), *SSDA* detects and segments the interacted object, that was not detected by the *Real-Only* approach.[5]

[5] Additional qualitative examples are reported in the supplementary material.

## 5 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 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. For instance, in the VISOR dataset, UDA obtained a +23.45% improvement with respect to real-only in Overall AP, obtaining a score of 33.33%, about 10% smaller than the fully supervised baseline trained on real data.
*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.

We hope that our analysis will inform future application and model developments and that the release of the HOI-Synth benchmark and data generation pipeline will support future research in this field.

## Acknowledgments

This research has been supported by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006. This research has been partially supported by the project EXTRA-EYE - PRIN 2022 - CUP E53D23008280006 - Finanziato dall'Unione Europea - Next Generation EU.

## References

1. Besari, A.R.A., Saputra, A.A., Chin, W.H., Kubota, N., et al.: Hand–object interaction recognition based on visual attention using multiscopic cyber-physical-social system. International Journal of Advances in Intelligent Informatics **9**(2) (2023)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR. pp. 3722–3731 (2017)
3. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: CVPR. pp. 11457–11466 (2019)
4. Carfì, A., Patten, T., Kuang, Y., Hammoud, A., Alameh, M., Maiettini, E., Weinberg, A.I., Faria, D., Mastrogiovanni, F., Alenyà, G., et al.: Hand-object interaction: From human demonstrations to robot manipulation. Frontiers in Robotics and AI **8**, 714023 (2021)
5. Cheng, T., Shan, D., Hassen, A.S., Higgins, R.E.L., Fouhey, D.: Towards a richer 2d understanding of hands at scale. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
6. Choudhary, A., Mishra, D., Karmakar, A.: Domain adaptive egocentric person re-identification. In: Computer Vision and Image Processing (CVIP). pp. 81–92 (2021)
7. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey (2017), `https://arxiv.org/abs/1702.05374`
8. Damen, D., Doughty, H., Farinella, G.M., , Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV pp. 1–23 (2021)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV. pp. 720–736 (2018)
10. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. In: NeurIPS. pp. 13745–13758 (2022)
11. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: CVPR. pp. 4091–4101 (2021)
12. Di Benedetto, M., Carrara, F., Meloni, E., Amato, G., Falchi, F., Gennaro, C.: Learning accurate personal protective equipment detection from virtual worlds. Multimedia Tools and Applications **80**, 23241–23253 (2021)
13. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16 (2017)

14. Edsinger, A., Kemp, C.C.: Human-robot interaction for cooperative manipulation: Handing objects to one another. In: RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication. pp. 1167–1172. IEEE (2007)
15. Fabbri, M., Brasó, G., Maugeri, G., Ošep, A., Gasparini, R., Cetintas, O., Calderara, S., Leal-Taixé, L., Cucchiara, R.: Motsynth: How can synthetic data help pedestrian detection and tracking? In: ICCV (2021)
16. Fu, Q., Liu, X., Kitani, K.M.: Sequential voting with relational box fields for active object detection. In: CVPR. pp. 2374–2383 (2022)
17. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z.Q., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., González, C., Hillis, J.M., Huang, X., Huang, Y., Jia, W., Khoo, W.Y.H., Kolár, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K.K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbeláez, P., Crandall, D.J., Damen, D., Farinella, G.M., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R.A., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2021)
19. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
20. Jian, J., Liu, X., Li, M., Hu, R., Liu, J.: Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In: ICCV. pp. 14713–14724 (October 2023)
21. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR. pp. 9799–9808 (2020)
22. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: Ai2-thor: An interactive 3d environment for visual ai (2017), https://arxiv.org/abs/1712.05474
23. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv (2017)
24. Leonardi, R., Ragusa, F., Furnari, A., Farinella, G.M.: Egocentric human-object interaction detection exploiting synthetic data. In: International Conference on Image Analysis and Processing. pp. 237–248. Springer (2022)
25. Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K.E., Gokmen, C., Dharan, G., Jain, T., Kurenkov, A., Liu, K., Gweon, H., Wu, J., Fei-Fei, L., Savarese, S.: igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In: Faust, A., Hsu, D., Neumann, G. (eds.) Proceedings of the 5th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 164, pp. 455–465. PMLR (08–11 Nov 2022), https://proceedings.mlr.press/v164/li22b.html

26. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: CVPR. pp. 6943–6953 (2021)
27. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: CVPR. pp. 7581–7590 (2022)
28. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
29. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: CVPR. pp. 3282–3292 (2022)
30. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: ICLR (2021)
31. Lu, Y., Mayol-Cuevas, W.W.: Egocentric hand-object interaction detection and application (2021), https://arxiv.org/abs/2109.14734
32. Lv, Z., Poiesi, F., Dong, Q., Lloret, J., Song, H.: Deep learning for intelligent human–computer interaction. Applied Sciences **12**(22), 11457 (2022)
33. Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
34. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR. pp. 122–132 (2020)
35. Munro, J., Wray, M., Larlus, D., Csurka, G., Damen, D.: Domain adaptation in multi-view embedding for cross-modal video retrieval. ArXiv **abs/2110.12812** (2021), https://api.semanticscholar.org/CorpusID:239768993
36. NVIDIA: Nvidia omniverse (2020), https://www.nvidia.com/en-us/omniverse/synthetic-data/
37. NVIDIA: Nvidia isaac sim (2021), https://developer.nvidia.com/isaac-sim
38. Orlando, S., Furnari, A., Farinella, G.M.: Egocentric visitor localization and artwork detection in cultural sites using synthetic data. Pattern Recognition Letters - Special Issue on Pattern Recognition and Artificial Intelligence Techniques for Cultural Heritage (2020), https://iplab.dmi.unict.it/SimulatedEgocentricNavigations/
39. Pasqualino, G., Furnari, A., Signorello, G., Farinella, G.M.: An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites. Image and Vision Computing **107**, 104098 (2021)
40. Plizzari, C., Perrett, T., Caputo, B., Damen, D.: What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In: ICCV2023 (2023)
41. Quattrocchi, C., Mauro, D.D., Furnari, A., Lopes, A., Moltisanti, M., Farinella, G.M.: Put your ppe on: A tool for synthetic data generation and related benchmark in construction site scenarios. In: International Conference on Computer Vision Theory and Applications. pp. 656–663 (2023)
42. Ragusa, F., Furnari, A., Livatino, S., Farinella, G.M.: The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In: Winter Conference on Applications of Computer Vision. pp. 1569–1578 (2021)
43. Ragusa, F., Leonardi, R., Mazzamuto, M., Bonanno, C., Scavo, R., Furnari, A., Farinella, G.M.: Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4549–4559 (2024)

44. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J.M., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In: NeurIPS (2021)
45. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
46. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV. pp. 9339–9347 (2019)
47. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: CVPR. pp. 21096–21106 (2022)
48. shadowrobot: Shadowhand (2005), `https://www.shadowrobot.com/dexterous-hand-series/`
49. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR. pp. 9869–9878 (2020)
50. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. Advances in Neural Information Processing Systems **34**, 251–266 (2021)
51. Tang, Y., Tian, Y., Lu, J., Feng, J., Zhou, J.: Action recognition in rgb-d egocentric videos. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3410–3414. IEEE (2017)
52. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017)
53. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. pp. 7167–7176 (2017)
54. Unity: Synthetichumans package (unity computer vision) (2022), `https://github.com/Unity-Technologies/com.unity.cv.synthetichumans`
55. Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., Wang, H.: Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In: CVPR. pp. 11359–11366 (2023)
56. Xia, F., R. Zamir, A., He, Z.Y., Sax, A., Malik, J., Savarese, S.: Gibson env: real-world perception for embodied agents. In: CVPR (2018)
57. Xia, F., Shen, W.B., Li, C., Kasimbeg, P., Tchapmi, M.E., Toshev, A., Martín-Martín, R., Savarese, S.: Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. IEEE Robotics and Automation Letters **5**(2), 713–720 (2020)
58. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: CVPR. pp. 22479–22489 (2023)
59. Zhang, L., Zhou, S., Stent, S., Shi, J.: Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In: ECCV. pp. 127–145 (2022)
60. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)