EgoObjects: A Large-Scale Egocentric Dataset for Fine-Grained Object Understanding

Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu
Hichem El-Mohri, Sean Chang Culatana, Roshan Sumbaly, Zhicheng Yan
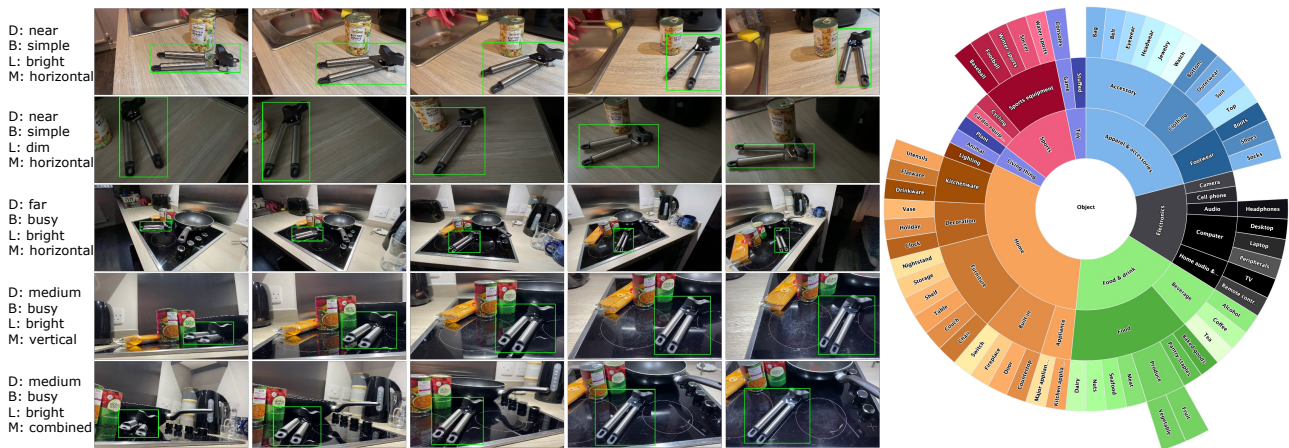Meta AI

{chenchenz, fanyix, zyan3}@meta.com

https://github.com/facebookresearch/EgoObjects

Figure 1: **EgoObjects dataset**. **Left**: *It contains videos of objects captured from the first-person viewpoint under 10 diverse conditions (only 5 are shown for clarity). Multiple objects in each video are annotated with instance ID and category label. In each row, we visualize the annotations of one instance track ("can opener") in one video captured under one set of condition variable choices. For clarity, we use shorthand notations: D – Distance, B – Background, L – Lighting, M – Camera Motion. Also annotations on other objects are not shown.* **Right:** *A visualization of a subset of non-leaf nodes in our hierarchical object taxonomy, covering diverse object categories. Leaf nodes and other non-leaf nodes are omitted for clarity.*

Abstract

Object understanding in egocentric visual data is arguably a fundamental research topic in egocentric vision. However, existing object datasets are either non-egocentric or have limitations in object categories, visual content, and annotation granularities. In this work, we introduce EgoObjects, a large-scale egocentric dataset for fine-grained object understanding. Its Pilot version contains over 9K videos collected by 250 participants from 50+ countries using 4 wearable devices, and over 650K object annotations from 368 object categories. Unlike prior datasets containing only object category labels, EgoObjects also annotates each object with an instance-level identifier, and includes over 14K unique object instances. EgoObjects was designed to capture the same object under diverse background complexities, surrounding objects, distance, lighting and camera motion. In parallel to the data collection, we conducted data annotation by developing a multi-stage federated annotation process to accommodate the growing nature of the dataset. To bootstrap the research on EgoObjects, we present a suite of 4 benchmark tasks around the egocentric object understanding, including a novel instance level- and the classical category level object detection. Moreover, we also introduce 2 novel continual learning object detection tasks. The dataset and API are available at https://github.com/facebookresearch/EgoObjects.

1. Introduction

Object understanding tasks, such as classification and detection, are arguably fundamental research topics in computer vision. Enormous amount of advances achieved so

far have been accelerated by the availability of large-scale datasets, such as ImageNet [14], COCO [28], LVIS [18], Open Images [24] and Objectron [1]. Those datasets often contain images captured from a third-person or exocentric viewpoint and curated from given sources (e.g. Flicker). Albeit the large volume, they often only capture individual object instances in a single image or video, and do not capture the same object under diverse settings, which are important for fine-grained object understanding task, such as instance-level object detection. In contrast, object understanding tasks in egocentric vision processes visual data containing objects captured from a first-person or egocentric viewpoint. The approaches to those tasks have wide applications in augmented reality and robotics, such as robustly anchoring virtual content at a real world object under various conditions (e.g. background, lighting, distance), and are often required to perform well from the egocentric viewpoint and distinguish objects at both category- (e.g. mug vs kettle) and instance level (e.g. my mug vs your mug) under various conditions. Therefore, there are clear gaps in adopting existing exocentric datasets for egocentric object understanding.

On the other hand, several egocentric datasets containing object annotations have been built. A family of such datasets are focused on capturing human activities and hand-object interactions. Ego4D [17] contains a large number of egocentric videos of human activities. However, according to the PACO-Ego4D [39] which mines the objects from Ego4D, there are only 75 object categories with at least 20 samples, and each object instance often only appears in one video. Epic-Kitchens-100 [11] contains over 700 videos depicting human activities in the kitchen, but only annotates objects within the kitchen. HOI4D [30] is collected for category-level human-object interaction, and only contains 800 different object instances from 16 categories. There are several other datasets that are more object-centric, including TREK-150 [15], FPHA [16] and CO3D [41], but only contain objects from a limited set of categories ($<$50). Objects there are often captured in a single setup or few setups with limited variations in surrounding objects, background, distances and camera motions. Moreover, semantic granularity of the object annotations are often limited at category-level, and object instances from the same category are not distinguished, which impedes the development of instance-level object understanding approaches. Therefore, there are still significant gaps with existing egocentric datasets in the dataset scale, visual content variations around individual objects, object semantic diversity, and instance-level object annotation.

To address these gaps, we introduce *EgoObjects*, a new large-scale egocentric video dataset for fine-grained object understanding (Figure 1). Unlike prior egocentric datasets which are limited to a small dataset scale, a specific do-

main or a small number of object categories, EgoObjects includes a large number of videos containing objects from hundreds of object categories commonly seen in the households and offices worldwide. For video capture, 4 wearable devices with various field-of-view are used, including Vuzix Blade smart glasses[1], Aria glasses[2], Ray-Ban Stories smart glasses[3] and mobile phones with ultra-wide lens[4], which provide representative media formats of egocentric visual data. Each main object is captured in multiple videos with different choices of nearby secondary objects, background complexity, lighting, viewing distance and camera motion. We annotate both the main and secondary objects in the sampled frames with bounding boxes, category level semantic labels and instance-level object identifiers (ID). In current Pilot version release, it contains over $9,200$ videos of over 30 hours collected by 250 participants from 50+ countries and regions, and 654K object annotations with 368 object categories and 14K unique object instance IDs from 3.8K hours of annotator efforts. To our best knowledge, EgoObjects is the largest egocentric video dataset of objects in terms of object categories, videos with object annotations, and object instances captured in multiple conditions. Comparisons between EgoObjects and other datasets can be seen in Table 1.

To bootstrap the research on EgoObjects, we introduce 4 benchmark tasks spanning over both non-continual learning and continual learning settings. For non-continual learning setting, we include a novel instance-level object detection task, largely under-explored previously due to the lack of a dataset with object ID annotations, as well as conventional category-level object detection task. For continual learning setting, we present novel object detection tasks at instance- and category level. Evaluations of different approaches to all tasks are also presented to establish the baseline benchmarks. In particular, for instance-level object detection task, a novel target-aware instance detection approach is proposed and validated to outperform a baseline target-agnostic object detection method.

To summarize, we make the following contributions.

- We created a large-scale egocentric dataset for object understanding, which features videos captured by various wearable devices at worldwide locations, objects from a diverse set of categories commonly seen in indoor environments, and videos of the same object instance captured under diverse conditions.

- We proposed a multi-stage federated annotation process for the continuously growing dataset to accompany the

---

[1] https://www.vuzix.com

[2] https://about.meta.com/realitylabs/projectaria

[3] https://www.meta.com/glasses

[4] Participants are asked to hold mobile phone close to their eyes to simulate egocentric viewpoints

	Exocentric			Egocentric			
	Objectron	CO3D	BOP	Epic-K*.	HOI4D	Ego4D**	EgoObjects
#category	9	50	89	300*	16	75	368+
#participant	int'l.	-	-	45	9	859 int'l.	250 int'l.
#image	4M	1.5M	330K	20M	2.4M	23.9K	114K+
#instance	17K	19K	89	-	800	17K	14K+
#bbox	-	-	-	38M	-	50K	654K+
inst ID	✗	✗	✓	✗	✗	✗	✓
device	M	M	PC,K	G	K,I	G,V,Z,W,PP	R,A,V,M

Table 1: ***Comparing EgoObjects with other datasets.*** *For EgoObjects, we report statistics of the current Pilot version, which is estimated to account for 10% of the full dataset (thus the "+" notation). *Epic-Kitchen-100 [11] only contain object categories in the kitchen. **Ego4D statistics are reported by the PACO-Ego4D [39], which annotates objects in the Ego4D [17]. Abbreviation for devices: M=Mobile, K=Kinect, A=Aria, G=GoPro, PC=Primesense Carmine, I=Intel RealSense, V=Vuzix Blade, R=Ray-Ban Stories, PP=Pupil, Z=Zetronix zShades, W=Weeview.*

parallel data collection at scale. Rich annotations at video level (e.g. location, background description) and object-level (e.g. bounding box, object instance ID, category level semantic label) are collected from 3.8K hours of human annotator efforts.

- We introduced 4 benchmark tasks on EgoObjects, including the novel instance-level and the conventional category-level object detection tasks as well as their continual learning variants. We evaluated multiple approaches on all tasks, and also proposed a novel target-aware approach for instance-level object detection task.

2. Related Work

Egocentric object understanding datasets. Given the growing needs of egocentric object understanding in augmented reality and robotics, several egocentric datasets focused on objects have been built. TEgO [25] contains egocentric images of only 19 distinct objects for training object recognizers. TREK-150 [15] consists of 150 annotated videos for tracking objects from 34 categories merely. Despite the availability of object annotations, other larger egocentric datasets are more focused on human activities and hand-object interactions. For example, Epic-Kitchens-100 [11] captures 700 videos of nearly 100 human activities involving 300 object categories in the kitchen, but is limited to the kitchen scenario. The ADL [37] dataset features people performing everyday activities in kitchens, which has object boxes, object track ID, action labels. However, it only has 42 object categories and the track ID is not used for analysis. The MECCANO [38] is a multimodal dataset of egocentric videos to study humans behavior understanding in industrial-like settings with object, depth, and gaze annotations, supporting a suite of 5 tasks. However, the diversity of participants and locations is limited. FPHA [16] captures

45 different daily hand-object action categories involving only 26 different objects. HOI4D [30] contains videos of human-object interaction with only 800 different object instances from 16 categories. Albeit the large number of human activity videos, the recent Ego4D [17] only contains object annotations from around 75 object categories with at least 20 samples. Object instances often only appear in a single video, and only 870 instances have more than 5 occurrences. Meanwhile, synthetic egocentric datasets are built to scale up the data collection. xR-EgoPose [47] is a large-scale synthetic dataset containing realistic renderings of people in various poses and serves as a benchmark of 3D human pose estimation. It is focused on ego-body and simulates fisheye lens where the surrounding environment, including objects, are largely distorted. EHOI [26] is also a synthetic dataset, consisting of 20K images and 124K object instances from 19 categories with interactions with human hands. Its fidelity is low compared with real data, and it has limited complexities in lighting, background and viewpoints. To summarize, existing egocentric datasets have limitations in the number of object categories, the variations in the setting of capturing the same object, the granularity of object semantic labeling where instance-level object ID is not available and photorealism in synthetic datasets.

Instance-level object detection and datasets. Being able to localize and recognize different object instances is critical to applications in augmented reality and robotics, such as detecting a specific toy or a custom industrial part. However, such task has been severely less explored due to the lack of object ID annotations at scale in existing datasets. In the cases of a growing number of object instances to detect, which is arguably a realistic setup, instance-level detection approaches are often required to adapt with little-to-no fine-tuning time. Mercier et al [34] proposed a template-based detector that uses example viewpoints of the target object to detect it in query images without extra training, and evaluated it on a small exocentric dataset of 20 object instances only. Hu et al [22] proposed a template-based detection approach, which incorporated a multi-level correlation model and a similarity-refine module, for handling the category-agnostic instance. On the dataset side, T-less [20] is an object dataset with 6D pose annotation for only 30 industry-relevant object instances. In [43], a small dataset of 10K RGBD images of 24 object instances were created for object detection and pose estimation. BOP dataset [21] combines 8 public datasets, and consists of 89 object instances with 3D groundtruth and over 330K RGBD images from different viewpoints. All those datasets aforementioned are not egocentric, and only contains a small number of object instances. In contrast, EgoObjects contains over 14K unique object instances captured under diverse settings. We also propose a target-agnosic baseline approach and a novel target-aware approach, and evaluate them on EgoObjects.
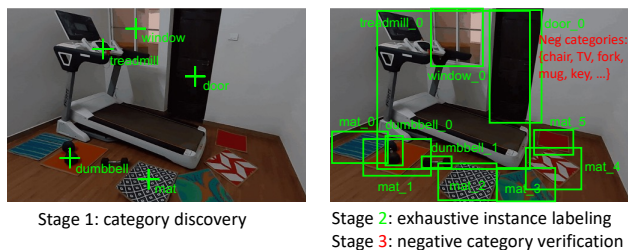
(a) *Number of instances per category.* (b) *Number of annotations per category.* (c) *Distribution of object centers.*

(d) *Relative bounding box sizes.* (e) *Geo distribution of participants.* (f) *Distribution of video metadata.*

Figure 2: ***Dataset statistics. (a) Left****: the number of instances per category in the log scale.* **Right***: the word cloud highlights the head categories, including box, bottle.* **(b) Left** *the number of annotations per category in the log scale.* **Right***: the word cloud is similar to (a), but a few new head categories emerge, including mug, boot. **(c)** Spatial distribution of the main objects center coordinates, confirming the diverse locations of main objects. **(d)** Relative bounding box sizes compared between EgoObjects, LVIS, and COCO. EgoObjects has more objects of medium and large sizes in the egocentric view. **(e)** Diverse distribution of participants' geographic locations in 50+ countries from 5 continents. **(f)** Distribution of video metadata including lighting (left) and background (right). Most objects are collected indoor, where lighting is more likely either artificial or low light. The background is uniformly distributed across rooms within the household.*

Continual learning. Conventional object understanding approaches build static models incapable of adapting their predictive behaviors over time. In contrast, continual learning models can learn from an infinite stream of data and grow their predicative capabilities while reducing catastrophic forgetting of previous knowledge [33, 5, 3, 9, 35, 2, 40, 8, 48]. Broadly speaking, they can be categorized into 3 classes [12] with increasing complexities and practicalities. In *Task Incremental Learning*, individual tasks with respective training data arrive sequentially, and the model is often built with separate heads for individual tasks. At inference time, a task ID is required for each sample. In the *Class Incremental Learning*, no task ID is provided at any time, and the model often has only one head. In the most general *Data Incremental Learning* [13], more assumptions on the stationary data distribution and the paradigm of sequentially growing tasks and classes are removed. In this work, we use EgoObjects to set up 2 new continual learning tasks, which covers both Class- and Data Incremental Learning paradigms. Moreover, existing approaches are often assessed on small object classification datasets, such as Core50 [32] and OpenLORIS-Object [45]. To our best knowledge, EgoObjects is the first dataset to support the benchmarking of *continual learning of object detection* at both instance and category level.

Category-level object detection. Early CNN-based approaches include both two-stage methods, which tackles

object proposal generation and object recognition separately [42, 6], and single-stage methods which remove the explicit object proposal generation for simplicity and efficiency [29, 4, 46]. Recent transformer-based methods introduce attention building blocks into both the backbone and the detection head to significantly improve the detection performance [7, 50, 10]. However, those approaches are often only evaluated on exocentric datasets, such as COCO [28] and LVIS [18], while their performance on egocentric datasets are largely unknown. EgoObjects contains nearly 400 object categories, and we assess both CNN and transformer models on it.

3. EgoObjects Dataset

3.1. An Overview

In current Pilot version, EgoObjects contains over 9K videos collected by 250 participants. A total of 114K frames are sampled and annotated.

Object instances captured under diverse conditions. A total of 14.4K unique object instances from 368 categories are annotated. Among them, there are 1.3K main object instances from 206 categories and 13.1K secondary object instances (*i.e.*, objects accompanying the main object) from 353 categories. On average, each image is annotated with 5.6 instances from 4.8 categories, and each object instance appears in 44.8 images, which ensures diverse viewing di-

Stage 1: category discovery

Stage 2: exhaustive instance labeling
Stage 3: negative category verification

Figure 3: *EgoObjects multi-stage annotation*. *See text for details at each stage.*

rections for the object. To further break down, for the main object, each instance appears in 95.9 images, whereas each secondary instance *i.e.* 39.8 images on average. See distributions of unique object instances and object annotations in Figure 2a and 2b, respectively. Both figures indicate the long-tailed nature of the dataset, making the benchmark more challenging and closer to the real-world distributions.

Diverse object spatial distribution. We encourage participants to avoid center bias during the capture of main objects with the moving camera. In Figure 2c, we confirm both the center coordinates are widely spread in the image.

Object scale distribution in egocentric view. Figure 2d compares EgoObjects with other datasets on the relative size distribution of object bounding boxes. The relative size is defined as the square root of the box-area-over-image-area ratio. Compared to COCO and LVIS, EgoObjects has more medium and large-sized objects, and suits the applications of egocentric object understanding where the users are more likely to interact with closer objects.

Metadata statistics. We further accumulate the per-video statistics across several metadata tags. As shown in Figure 2e, our data is geographically diverse, covering 50+ countries from five continents. Finally, as presented in Figure 2f, our data also has a diverse distribution covering various video capture conditions for lighting and backgrounds.

3.2. Data Collection

We work with third-party vendors to recruit participants for capturing videos of common indoor objects at worldwide locations. They use various glasses such as Vuzix Blade, Aria Glasses, and Ray-Ban Stories. They also use the ultra-wide lens on mobile phones and hold the phone close to the eyes for simulating the egocentric viewpoint. Participants are asked to capture videos of objects from a predefined list of 400 object categories, referred as main object categories. Each main object should be unique in the individual location and captured under various conditions.

We define 4 variables of capture conditions, including background complexity, camera motion, object distance and lighting. The background complexity can be either "simple" or "busy". The simple background has at least 3 sur-

rounding objects besides the main object, whereas the busy background has at least 5 other objects. In either background, we ask participants to capture the main object in natural settings (vs intentional setting). We also instruct the participants to move the camera around and capture different views of the main object, while avoiding the bias that the main object always stays in the center of the view. We define three 3 of camera motion: 1) "horizontal": move the camera from left to right or right to left. 2) "vertical": move the camera upwards or downwards and 3) "combined": rotate the camera both horizontally and vertically. The object distance also has three levels, i.e. "near", "medium", and "far". We define the object scale and the frame scale as the longer object dimension and the shorter frame edge, respectively. Near distance refer to those that the object-scale/frame-scale ratio is larger than 30%, whereas the medium distance has the ratio fall in between 20% and 30%. All remaining images are considered as having far object distances to the camera. For lighting conditions, there are two levels bright and dim. Lighting is considered as bright when a light meter reads above 250 lux and dim otherwise.

Given these 4 variables, participants are instructed to collect 10 videos of each main object according to 10 predefined configurations (see details in supplement), and each video lasts at least 10 seconds. Finally, videos are further tagged with rich metadata including the associated participant ID, main object category, location, background description and capture time.

3.3. Federated Annotation of the Growing Dataset

EgoObjects data collection was planned to operate at large scale and lasted for 14 months. To reduce the overall dataset creation time, we conducted data annotation in parallel to the data collection, which continuously grew the dataset and introduced more complexities to the data annotation. Inspired by LVIS [18], we adopt the idea of federated annotation to achieve a balance between annotation cost and annotation exhaustiveness, and further propose a 3-stage annotation pipeline tailored to the continuously growing nature of our dataset. Figure 3 illustrates our annotation pipeline, which is used to annotate video frames evenly sampled at 1 FPS.

Stage 1: category discovery. The annotators are instructed to identify object categories from a predefined vocabulary $\mathcal{V}$ of 600+ categories commonly seen in indoor egocentric view. Annotators are asked to find at least 5 categories per each image if possible, including the main object category and other salient secondary objects.

Stage 2: exhaustive instance labeling. For each image, 3 annotators exhaustively annotate *all* object instances of the discovered categories with bounding box and category label c. To enable instance-level object understanding, we further enhance the bounding box annotation with a unique

object instance ID[5] that is consistent across the dataset. To reconcile the work from 3 annotators, we compare the annotations from one annotator to all other annotators to get an averaged IoU based consensus score for each annotator. Then, we select the annotator with the highest consensus score as the source of truth for final annotations.

Stage 3: negative category verification. By the design of federated annotation [18], not all object categories are handled in each image. However, for evaluation purpose for each image we would need to collect a set of negative categories, defined as categories that do not appear in the image. To operationalize this, we randomly sample several categories from the vocabulary $\mathcal{V}$ as the candidates of the negative categories, and ask annotators to verify. We remove a candidate category from the negative set if any annotator flags any corresponding object instance in the image. Finally, we get a set of negative categories per image.

4. Benchmark Tasks on EgoObjects

We introduce 4 benchmark tasks on EgoObjects, starting with a novel instance-level object detection task, which has been under-explored due to the lack of object ID annotations on individual objects captured in various conditions in existing datasets. We further present 2 novel continual learning object detection tasks, which are newly enabled by EgoObjects. Finally, we assess the performance of classical category-level object detection models on EgoObjects.

4.1. Instance-Level Detection

In the applications of egocentric object understanding in AR and robotics, we are often confronted with the situation where the model is presented with few examples of object instances unseen during training and needs to detect those novel objects on-the-fly. Inspired by this, we introduce the instance-level detection below, and present two models, including a novel target aware- and a baseline target agnostic instance detector.

4.1.1 Task Specification

At training time, the model has access to instance-level annotations of objects captured under diverse conditions. At inference time, the user can use the model to register a novel target instance T, regardless of whether its category is seen during training, by providing one or more 2D bounding boxes on reference images containing the target instance. After that, on a query image I, the detector needs to predict the bounding box of T with T's ID, or no box if T is absent. To simulate the realistic setup where model back-propagation is difficult for deployed models, we *disallow*

---

[5]We exclude objects from categories that have indistinguishable appearances between instances, such as those related to animals and food.
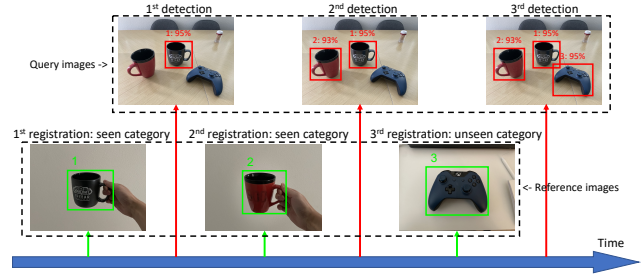

Figure 4: ***Instance detection at inference time.*** *Continuously registering more targets leads to more detected objects, while previously registered targets are not forgotten. The targets can be from either seen (target 1 and 2) or unseen (target 3) categories.*

model fine-tuning on the target object instance annotations. The model is required to allow the user to continuously register more model targets, and all registered targets should be considered during detection. Figure 4 contains an example where the user registers 3 targets sequentially and the model gradually detects more target objects in the same image.

Dataset split. We divide the dataset into 4 splits: train/target/val/test. The train split contains 9.6k instances with a total of 450k annotations from 79k images. The target, val, test splits share the remaining 4.1k instances which do not appear in the train images, and their categories can also be unseen during training. In the target split, there is a single reference image and one annotation for each instance. The val and test splits have 5.7K and 29.5K images with 3.8K and 4.1K instances, respectively.

Evaluation protocols. Under various IoU thresholds, we report Average Precision (AP) metrics, which are averaged across instances. Furthermore, we break down the metrics into two buckets for object instances from categories seen and unseen during training to assess the model capability of generalizing to instances from novel categories.

4.1.2 Target-aware Instance Detector

We propose an instance-level object detector aware of target objects during object localization, and refer to it as *Target-Aware Instance Detector* (TA-IDet). It supports 2 modes, namely target registration and target detection (Figure 5).

Target registration. To register a new target object, we feed the reference image into a ResNet based FPN backbone [27], generate target features at different pyramid levels by using ROIAlign operator [19] according to the target bounding box annotation, and average them over pyramid levels. Target features of two different resolutions are obtained from ROIAlign. T^{loc} feature of resolution 1×1 is used to efficiently localize the bounding box. T^{cls} feature has higher resolution $S \times S$ ($S = 5$ by default), and will be

used to compute a confidence score of classifying the localized target object. If several reference images per target are provided, the target features are averaged.

Target detection. At detection time, TA-IDet use the same FPN backbone to extract query image feature map F of size $C \times H \times W$ where C denotes feature channels, and $\{H, W\}$ feature map size. A feature modulation block will transform F according to target localization feature T^{loc} of size $C \times 1 \times 1$, which attenuates the features in regions where the target object is less likely to appear. The detection head takes as input the modulated query feature map, and processes it using a *Score* module, which consists of 4 convolutional layers with ReLU activations, to gradually reduce the channels from 256 to 1. The resulting score map is normalized by a *Softmax* operation, and the target object center (C_y, C_x) is predicted as the weighted sum of spatial coordinates according to the normalized score map.

$$
\begin{aligned}
F^{mod} &= (T^{loc} \circledast F) \odot F \\
P &= \text{Softmax}(\text{Score}(F^{mod}).\text{reshape}(-1)) \\
Y^g &= \text{ls}(0, H - 1, \text{steps} = H).\text{view}(H, 1).\text{repeat}(1, W) \\
X^g &= \text{ls}(0, W - 1, \text{steps} = W).\text{view}(1, W).\text{repeat}(H, 1) \\
C_y &= \text{sum}(P \odot Y^g.\text{reshape}(-1)) \\
C_x &= \text{sum}(P \odot X^g.\text{reshape}(-1))
\end{aligned}
\tag{1}
$$

where $\circledast$ denotes convolution, $\odot$ element-wise multiplication and ls torch.linspace. To refine object center and predict target object size, we sample a target feature at (C_y, C_x) in F via bilinear interpolation, and employ a 3-layer MLP with hidden dimension 256 to predict the spatial offset $(\delta C_y, \delta C_x)$ and target object size (S_y, S_x) with a ReLU activation. After predicting target object box, we use ROIAlign to sample a spatial feature of the resolution $S \times S$ in F, and compute its dot product with T^{cls} using a sigmoid activation function as the box confidence score.

Model training. During training, we sample three images for each instance: one reference image containing the instance, one positive image containing the instance captured in a different setting, and one negative image that does not contain the instance. In positive image, we consider both bounding box localization loss and classification loss. For localization loss, we use a linear combination of L_1 loss and generalized IoU loss [44]. For classification loss, we use the binary cross entropy loss between the predicted box confidence score and groundtruth box label, which is positive when IoU is above IoU^{pos}, negative when IoU is below IoU^{neg} and ignored otherwise. By default, $(IoU^{pos}, IoU^{neg}) = (0.7, 0.3)$. In negative image, only the classification loss is used and groundtruth label is negative. See more studies in the supplement.
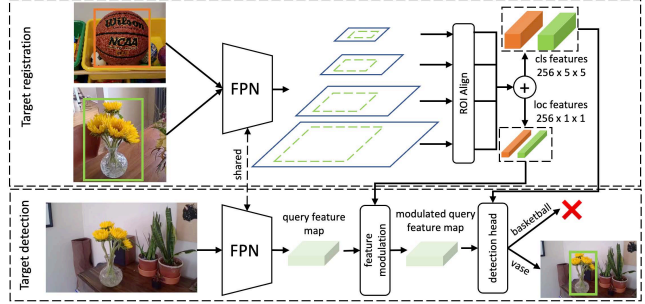

Figure 5: *Architecture of target-aware instance detector TA-IDet. Top: in target registration, localization and classification feature for each target are generated. Bottom: during target detection, the model predicts 1 bounding box per target and computes a confidence score to decide whether the prediction should be rejected via thresholding.*

4.1.3 Baseline Target-agnostic Instance Detector

We also consider a simple baseline approach *RPN+SFNet* which consists of a Region Proposal Network (RPN) [42] for object localization and a SFNet model [49], commonly used in metric learning, for object classification. We briefly review its target registration, detection and model training below, and include more details in the supplement.

Target registration. We crop the target object from reference images and feed it through the SFNet model to obtain the target object feature, which is then added to an index of target object features.

Target detection. For a given query image, the RPN generates a large number of object proposals *agnostic* to the target objects in the index. Each object proposal is cropped from the query image and fed into the SFNet model to extract the feature. These object features are then matched against all the added target features in the index. We pick the target object in the index with the highest matching score. The final confidence score of an object proposal against the top target object is the product of RPN object proposal confidence score and its matching score with the target object.

Model training. The RPN is trained on the train split using all bounding box annotations. The SFNet model is trained with SphereFace2 [49] loss function using all instance-level annotations, which encourages small distances between features of multiple views of the same instance, and large distances for features of different instances.

4.1.4 Results

The benchmark results of both models are presented in Table 2. For both approaches *TA-IDet* and *RPN+SFNet*, we build models with ResNet-50 and ResNet-101 backbones. There are several intriguing observations. First, *TA-IDet* substantially outperforms the *RPN+SFNet* on all metrics.

backbone	method	val				test			
		AP	AP50	$AP50_{se}$	$AP50_{un}$	AP	AP50	$AP50_{se}$	$AP50_{un}$
R50	RPN+SFNet	17.8	29.0	29.1	19.8	15.7	25.4	25.5	16.8
	TA-IDet	18.7	35.0	35.0	21.7	18.5	35.2	35.2	24.8
R101	RPN+SFNet	19.3	32.0	32.0	22.3	17.0	27.7	27.8	20.0
	TA-IDet	22.6	37.9	38.0	28.5	21.9	37.9	38.0	26.4

Table 2: ***Instance-level detection benchmarking results on EgoObjects.*** *The proposed* TA-IDet *model significantly outperforms the baseline* RPN + SFNet *approach.* $AP50_{se}$ *and* $AP50_{un}$ *are computed for instances with categories seen and unseen during training. On the more challenging test split with more targets object instances, TA-IDet can maintain the performance whereas* RPN + SFNet *baseline has a significant performance drop. R50 and R101 denote ResNet-50/101 backbones.*

rank	CL Instance Detection						CL Category Detection					
	E_0	E_1	E_2	E_3	E_4	EAP	E_0	E_1	E_2	E_3	E_4	EAP
1st	23.3	39.5	54.6	70.2	85.6	54.7	30.6	47.2	58.1	67.5	76.2	55.9
2nd	15.1	30.4	45.5	60.8	75.4	45.4	28.4	44.7	57.6	67.9	78.2	55.4
3rd	14.7	29.1	42.3	55.4	66.9	41.7	19.5	34.5	43.9	52.7	61.5	42.4

Table 3: ***CL detection benchmarks on EgoObjects.*** *We report detection accuracy (mAP) after each experience and final Experience Average Precision (EAP), which is the averaged mAP over experiences.*

For example, the gains in AP50 are large ($+6\%$ on val and $+10\%$ on test split). We attribute this to the design that *TA-IDet* localizes the target object by using query image feature maps modulated by the target feature, and does not rely on the target-agnostic RPN to generate object proposals.

Second, the best *TA-IDet* model with ResNet-101 backbone only achieves less than 23% AP on both val and test split which have around 4K novel instances each, indicating the unique challenges in the large-scale instance-level object detection, such as large changes in viewing direction, lighting, background and distance as well as less distinguishable appearance between instances from the same category. See more examples in the supplement.

Third, there are significant gaps between AP_{se} and AP_{un}, reflecting the challenges in generalizing the models to detect instances from categories unseen during training.

4.2. Continual Learning

Existing continual learning (CL) approaches often tackle object classification problem while continual learning object detection task is not well explored due to the lack of a large-scale dataset that is annotated with instance- and category-level labels, and contains individual objects in multiple images captured under diverse conditions. We introduce 2 novel CL tasks, namely *CL instance detection* and *CL category detection*, on a subset of EgoObjects which contains 100K images with 250K box annotations for 1.1K main object instances from 277 categories. There are 3.4K and 3.1K instances in the train- and test set, respectively.

CL Instance Detection. In this task, we simulate the setting when a system continuously encounters new batches of instance detection training data, where each batch indexed at i is called an experience E_i. In E_i, each image only carries the annotation of its main object with instance ID being the class. The system can only access data in the latest experience, which means no access to previous experiences apart from the use of a limited replay memory. Previous experiences share no common main object instances with later experiences, which makes it a *Class-Incremental Learning*

setup. We evenly split 1110 main instances into 5 experiences, hence 222 instances per experience. For evaluation, the system is benchmarked after each experience on a fixed testing set with all the main instances, making it a 1110-way detection problem. The evaluation metric for each experience is mean average precision (mAP). The final performance is the averaged metric across all the experiences.

CL Category Detection. In this task, the goal is to predict the object category labels instead of instance IDs. We create 5 experiences by applying the class-incremental ordering on the 277 categories of the main object instances. Additionally, we also include the annotations of secondary objects, which makes it a *Data-Incremental Learning* setup, i.e. previous experiences share no common images or annotations with later ones. This differentiates our task from other CL object detection tasks focusing on annotation incrementality, where the same images are repeatedly encountered in successive experiences but with a different set of annotations (usually class-incremental). We believe our task provides a more realistic setting. The evaluation metric for each experience is also mAP.

Results. We benchmark the methods from the top submissions of the 3rd CLVision workshop challenge [36] and report the results on above CL tasks in Table 3. In general, these submissions build upon classic 1-stage/2-stage detectors and adopt techniques to mitigate catastrophic forgetting of early experiences when trained on later experiences, such as sampling data from previous experiences cached in a replay buffer, and distilling models from early experiences to the model for the current experience. However, these winning methods still have limitations. They treat the instance detection as a close-set problem same as the category detection, which cannot scale up to flexibly accommodate more instances. Additionally, there is no change to the detector architecture to better tailor to the CL tasks. See more discussions in the supplement.

4.3. Category-Level Detection

EgoObjects also supports the classical category-level object detection task given the nearly 400 object categories in the current Pilot version.

Evaluation protocols. We use the same dataset splits as the instance-level detection task. In total, there are

method	pretrain	AP	AP75	AP50	$AP50_l$	$AP50_m$	$AP50_s$	$AP50_{bright}$	$AP50_{dim}$	$AP50_{simple}$	$AP50_{busy}$	AP	AP50
							val					test	
FasterRCNN-R50	IN1k	20.5	21.8	32.9	41.4	31.2	18.1	35.6	32.3	42.3	32.7	20.1	32.6
FasterRCNN-R101	IN1k	22.5	23.6	37.9	47.0	35.8	20.8	40.7	37.5	49.2	37.1	21.6	36.6
FCOS-R50	IN1k	23.1	25.0	31.4	40.2	28.1	16.9	33.2	31.7	40.0	31.4	22.4	30.6
FCOS-R101	IN1k	24.2	26.4	32.1	40.5	28.9	17.8	34.2	31.9	41.3	31.9	23.2	31.0
DeformDETR-SwinT	IN22k	29.1	31.6	37.7	47.0	34.3	22.5	39.8	37.9	48.3	37.3	28.6	37.1
DeformDETR-SwinS	IN22k	32.0	34.4	41.0	52.1	37.5	21.9	42.8	41.7	52.8	39.8	31.1	40.3
DeformDETR-SwinB	IN22k	32.2	34.8	41.7	51.0	39.3	23.8	43.8	41.7	53.0	40.2	31.1	40.4
DeformDETR-SwinL	IN22k	33.5	36.3	42.6	53.7	39.5	25.6	44.7	42.9	53.2	42.2	32.6	41.8

Table 4: ***Category detection benchmark on EgoObjects.*** *We benchmark three mainstream detectors in this table: FasterRCNN [42], FCOS [46], and Deformable-DETR [50]. All detectors are trained on* `train` *split and tested on* `val` *split.* $\{AP50_l, AP50_m, AP50_s\}$ *are computed for different object sizes,* $\{AP50_{bright}, AP50_{dim}\}$ *for different lighting conditions, and* $\{AP50_{simple}, AP50_{busy}\}$ *for different background scenes.*

447K/31K/164K object annotations from 368 categories in the `train/val/test` split, respectively. Due to its federated annotation process, we only penalize false positive predictions on an image if the predicted class is in the list of negative categories for that image.

Benchmarking models. We consider 3 types of performant object detectors. The first one is FasterRCNN [42], which is a two-stage detector. Next, we include the representative single-stage detector FCOS [46] which skips the explicit proposal generation to accelerate the model inference. Finally, we also consider the recent transformer-based detectors (*i.e.*, DETR [7]). Specifically, we adopt the Deformable-DETR [50] due to its stable and fast training. For both FasterRCNN and FCOS, we use the ResNet50/101 backbone pretrained on ImageNet-1K, whereas for Deformable-DETR, we use the Swin-Transformers [31] backbone pretrained on ImageNet-22K.

Results. The results are presented in Table 4. Notably, single stage FCOS models outperform two-stage FasterRCNN detectors particularly for high IOU threshold (*e.g.* AP75), while DeformDETR-Swin models significantly outperform both types of CNN detectors at the cost of large model size and significantly more compute. However, even for the largest DeformDETR-SwinL model, its AP metrics on EgoObjects are still 10% lower than its AP on LVIS 43.7% reported in Table 3 of prior work [23]. We hypothesize due to the egocentric view and its data capture setting, EgoObjects contains larger variations in background, viewpoints, lighting and object distances, which together render it more difficult even for category-level detection. We also implemented the metrics of different buckets on experimental conditions (e.g. object scale, lighting, background complexity) in our evaluation API. We observe model's performance is lower under the more challenging conditions (small scale, dim lighting, busy background).

5. Conclusions

We present EgoObjects, a large-scale egocentric dataset containing tens of thousands of videos, and more than half million object annotations. By design, it captures the same

object under diverse conditions while annotating it with both category label and consistent object IDs in multiple images. To stimulate the egocentric object understanding research on it, we introduce 4 tasks and also provide the benchmarking results of various models, including a novel target-aware instance-level detector which largely outperforms an off-the-shelf baseline based on RPN and SFNet.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021.

[2] Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless sequential learning. *arXiv preprint arXiv:1806.05421*, 2018.

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8227, 2021.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[8] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[9] Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

[10] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *ICCV*, 2021.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[12] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[13] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259, 2021.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *International Journal of Computer Vision (IJCV)*, 2022.

[16] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017.

[21] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[22] Zhongxu Hu, Runjia Tan, Yanxin Zhou, Junyang Woon, and Chen Lv. Template-based category-agnostic instance detection for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(4):12451–12458, 2022.

[23] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[25] Kyungjun Lee and Hernisa Kacorri. Hands holding clues for object recognition in teachable machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[26] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Egocentric human-object interaction detection exploiting synthetic data. In *International Conference on Image Analysis and Processing*, pages 237–248. Springer, 2022.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[30] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, June 2022.

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

[32] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017.

[33] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[34] Jean-Philippe Mercier, Mathieu Garon, Philippe Giguere, and Jean-Francois Lalonde. Deep template-based object instance detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1507–1516, 2021.

[35] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[36] Lorenzo Pellegrini, Chenchen Zhu, Fanyi Xiao, Zhicheng Yan, Antonio Carta, Matthias De Lange, Vincenzo Lomonaco, Roshan Sumbaly, Pau Rodriguez, and David Vazquez. 3rd continual learning workshop challenge on egocentric category and instance level object understanding. *arXiv preprint arXiv:2212.06833*, 2022.

[37] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012.

[38] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[39] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. *arXiv preprint arXiv:2301.01795*, 2023.

[40] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[41] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[43] Colin Rennie, Rahul Shome, Kostas E Bekris, and Alberto F De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016.

[44] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Qi She, Fan Feng, Xinyue Hao, Qihan Yang, Chuanlin Lan, Vincenzo Lomonaco, Xuesong Shi, Zhengwei Wang, Yao Guo, Yimin Zhang, et al. Openloris-object: A dataset and benchmark towards lifelong object recognition. *arXiv preprint arXiv:1911.06487*, 2019.

[46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.

[47] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019.

[48] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[49] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. Sphereface2: Binary classification is all you need for deep face recognition. *arXiv preprint arXiv:2108.01513*, 2021.

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.