Incremental Human-Object Interaction Detection with Invariant Relation Representation Learning

Yana Wei[1,*] Zeen Chi[1,*] Chongyu Wang[1] Yu Wu[1] Shipeng Yan[1] Yongfei Liu[1] Xuming He[1,2]
[1]ShanghaiTech University [2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{weiyn1,chize,wangchy5,wuyu1,yansp,hexm}@shanghaitech.edu.cn liuyongfei314@gmail.com

Abstract

Inspired by human's ability to continually acquire knowledge throughout their growth, this paper focuses on incremental learning in human-object interaction (HOI) detection, aiming to develop agents capable of learning the relations between humans and objects in ever-changing environments. We not only address the common issue of catastrophic forgetting but also tackle the unique challenges of interaction drift and detecting zero-shot HOI combinations with training data arriving partially and sequentially. Due to the limitations of existing methods in simultaneously addressing these three challenges, we propose a novel incremental relation distillation framework (IRD). This framework first disentangles object and relation learning and then introduces Momentum Feature Distillation and Concept Feature Distillation to learn invariant relation features across different HOI combinations sharing the same relation, enhancing model robustness to changing data distributions and unseen HOI categories. Extensive experiments on HICO-DET and V-COCO datasets demonstrate that our method outperforms state-of-the-art baselines, not only in mitigating forgetting but also in its robustness against interaction drift and generalization on zero-shot HOIs.

1. Introduction

The task of human-object interaction (HOI) detection [6, 50, 78] aims to detect humans and objects in images and classify the interactions between them. While much progress has been made in HOI detection recently, existing approaches primarily focus on a closed-world setting with a fixed number of HOI classes. However, in open-world and dynamic environments, it is required to understand long-term human behavior with personalized or specific interactions that are hard to pre-define. For instance, home service robots should learn continually to adapt to new actions/habits of family members or guests. In certain scenarios, such as operating rooms and hospital wards, access to historical data is limited due to privacy concerns. Therefore, it is desirable to have the human-like ability to acquire new HOI concepts continually [37, 49, 71], which remains under-explored so far. In this work, we aim to tackle this problem by introducing an *incremental human-object interaction detection* (IHOID) setup, where the HOI model needs to learn to detect an increasingly larger set of interactions between humans and a fixed set of familiar objects. Such a problem setting reflects a usual daily living or working environment where novel objects often rarely appear. Additionally, due to the compositional nature of HOIs, the model should also generalize well to zero-shot object-relation combinations [6, 25, 27]. As illustrated in Fig. 1, the model learns the interactions `feed dog` earlier and incrementally learns new interactions like `ride horse` at a

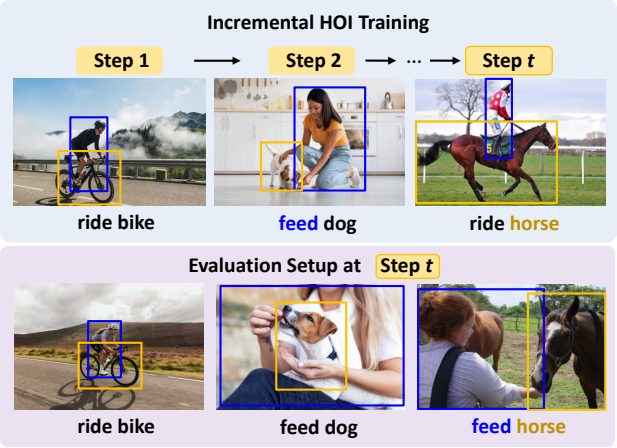


Figure 1. The training and evaluation of IHOID. The model is trained in multiple steps and learns different object-relation pairs in each step. During the evaluation, the model should not only detect HOIs learned in the previous and current steps such as `ride bike` and `feed dog`, but also overcome interaction drift of `ride` caused by subsequent `ride horse`. Additionally, the model is required to recognize the zero-shot HOI `feed horse` which is a combination of previously learned relations and objects.

*Both authors contributed equally to this work.

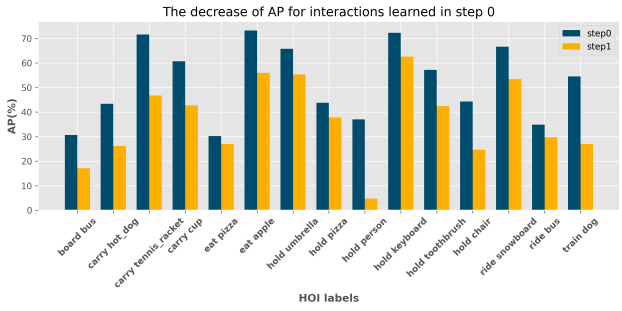


Figure 2. Demonstration of interaction drift. The statistics show the APs of HOI categories which are related to the same relation categories that occur across training step 0 and step 1. The APs of these categories suffer from obvious decreases.

later time step, so the model should naturally recognize the novel combinations `feed horse` during evaluation.

Along with catastrophic forgetting in class incremental learning (CIL) [37, 53], this incremental HOI task introduces two new challenges. Firstly, due to the compositional nature of the HOI classes, learning a new interaction class may interfere with learned interactions sharing the same relation class. This may result in a forgetting phenomenon at the component level of HOI pairs. In Fig. 1, for instance, the model initially learns the human-object interaction `ride bike`, and subsequently learns `ride horse`, the acquisition of the latter HOI could potentially lead to a forgetting of the former although both interactions fall within the same relational category. This phenomenon, which we refer to as *interaction drift*, arises principally due to the excessive dependence of interaction representation learning on object information. The detailed statistics are shown in Fig. 2. Secondly, our IHOID setup introduces a distinct challenge compared to the zero-shot HOI problem in common joint training scenarios [26, 27, 50]. Here, the objects and relations of zero-shot cases may appear in different time steps, and only a fraction of data at each step is visible to the model. As such, our incremental learning setting provides a limited context for the model to generalize to zero-shot scenarios. This can be observed in Fig. 1 with the test example `feed horse`.

To address the aforementioned challenges, we propose an Incremental Relation Distillation (IRD) framework for IHOID, which is also capable of generalizing to unseen HOI combinations. Our main idea is to disentangle the learning of HOIs into object and relation learning, and focus on continually learning robust and invariant relation representations, as shown in Fig. 3. Specifically, we introduce two new distillation strategies, namely Concept Feature Distillation (CFD) and Momentum Feature Distillation (MFD), to facilitate the learning of robust and invariant relation representations. Concretely, CFD captures the invariant aspects across different object-relation pairs that share the same relation, and MFD focuses on maintaining the discriminative feature of relations across incremental steps, alleviating forgetting caused by the introduction of new HOI categories. These two distillations are achieved by a dynamically updated concept-feature dictionary and a momentum teacher, which respectively store and provide reference relation features associated with each relation concept. Both modules are removed after training and only the HOI detection model is used to predict the HOI classes during inference.

We validate our approach by extensive comparison with prior incremental learning methods on two widely used HOI datasets: HICO-DET [6] and V-COCO [20]. The experimental results and ablation study show that our method outperforms other approaches in tackling forgetting and interaction drift and has better generalization on zero-shot HOIs.

Our main contributions can be summarized as follows:

- We propose the incremental learning setting for human-object interaction detection (IHOID), which not only focuses on the catastrophic forgetting of HOI classes but also considers the model's robustness to interaction drift and generalization ability on zero-shot HOI combinations.
- To tackle the challenges introduced by IHOID, we propose an incremental relation distillation framework that disentangles the learning of objects and relations and focuses on learning robust and invariant relation representations.
- We conduct extensive experiments on two HOI datasets HICO-DET and V-COCO, demonstrating that our method outperforms the SOTA baselines under the aforementioned two new challenges along with catastrophic forgetting.

2. Related Works

2.1. HOI Detection

HOI detection [19, 20, 35, 36] is crucial for understanding deeper scenes and achieving better relation or action understanding in computer vision. Existing HOI detectors can be broadly categorized into two-stage and one-stage models. Traditionally, two-stage HOI detectors [6, 17, 18, 35, 60, 63, 64, 77, 78] often adopted a modular approach, where object instances are firstly detected and interaction classification is then learned through the grouped pairwise human-object proposals. On the contrary, one-stage models [10, 15, 31, 32, 38, 58, 67, 81, 82] directly predicted HOI triplets from the image features in parallel either in anchor-based [31, 67] or point-based [38] manners. Recent studies are based on HOI Transformers [32, 33, 58, 62, 69, 76, 78, 79, 82], which builds upon the object detection architecture DETR [4], or on Vision Transformer [12] as the feature extraction backbone [30, 51]. However, these models were primarily trained in a close-

world setting with a fixed number of HOI categories during training. In this work, we aim to address the challenges of HOI detection in the incrementally arriving data scenario.

2.2. Zero-shot HOI Detection

Zero-shot learning aims to predict categories that were not seen during training. Previous works have explored zero-shot HOI detection in three scenarios: unseen combination [3, 25, 44, 66], unseen object [3, 25, 27], and unseen relation [28]. Some approaches [3, 21, 25–28, 56] factorized the learning of objects and relations and then performed data augmentation by combining independent object and relation representations. Others [39, 72, 73, 80] proposed vision-language pre-training to improve the performance on zero-shot HOI cases. Unlike joint training where all objects and relations are seen at the same time, in IHOID, the step-by-step data exposure to the model presents unique zero-shot learning challenges. IHOID's sequential introduction of elements increases the risk of forgetting previous ones, making it harder for the model to generalize to unseen scenarios.

2.3. Class Incremental Learning

In the context of class incremental learning (CIL) [1, 14, 34, 37, 43, 52, 54, 68, 70, 71, 74], models are trained on a sequence of data batches and continually learn new classes. However, this process often leads to catastrophic forgetting [34]. Existing incremental learning methods can be broadly categorized into three types. First, dynamic architecture methods [46–48, 55, 70, 71] progressively expanded the model structure to accommodate new classes. Second, memory-based methods [2, 8, 9, 24, 52, 54, 65, 75] stored exemplars of previous steps and learned new data with memory replay in subsequent time steps. Third, regularization-based methods [1, 7, 11, 13, 29, 34, 37, 43, 57, 59, 74] imposed constraints on updating neural weights that are important for previous tasks to alleviate forgetting. Different from traditional CIL, where forgetting primarily occurs when introducing new categories, our IHOID framework is faced with an additional challenge known as *interaction drift*. When the same relation interacts with different objects in separate steps, the object-relation combinations introduced later can cause a decline in the performance of those that appeared in earlier steps. This unique challenge cannot be effectively addressed by existing methods designed for CIL.

3. Methods

In this work, we aim to tackle the *incremental human-object interaction detection* (IHOID) problem, where training data of different HOI categories sequentially arrive in multiple time steps. To overcome the challenges in incremental learning of those compositional classes, we propose an Incremental Relation Distillation (IRD) framework, as illustrated in Fig. 3. Our framework adopts a model architecture that disentangles the HOI detection into object detection and relation learning, and then specifically focuses on learning robust and invariant relation representations. To achieve this goal, we develop two novel distillation methods, achieved by a momentum teacher and a dynamically updated concept-feature dictionary.

In the following subsections, we first present the problem formulation in Sec. 3.1, followed by the presentation of model architecture in Sec. 3.2. In Sec. 3.3, we describe the proposed method that facilitates the learning of relation representations with distillations, and we conclude the training objective functions of this framework in Sec. 3.4.

3.1. Problem Formulation

In the IHOID setup, we aim at not only mitigating catastrophic forgetting of HOI classes but also maintaining the model's robustness to interaction drift and generalization ability on zero-shot HOI combinations. In our method formulation, the HOI detector undergoes continual training across T total steps. In each step $t \in \{1, \cdots, T\}$, the model is exposed to only a subset of annotations corresponding to specific HOI categories.

Formally, we define the training set as $\mathcal{D} = \{(I, y)\}$, where I represents images and y represents the corresponding HOI annotations. Within the annotations y, we denote $\mathcal{C} = \{C_i\}_{i=1}^{N_c}$, $\mathcal{O} = \{O_j\}_{j=1}^{N_o}$, and $\mathcal{R} = \{R_k\}_{k=1}^{N_r}$ as the sets of HOIs, objects, and relations, where N_c, N_o, and N_r are the number of HOI, object, and relation categories, respectively. Each HOI category C_i consists of an object category O_j and a relation category R_k.

To construct the IHOID task, we partition the dataset and HOI categories into T disjoint subsets $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_T$ and $\mathcal{C} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_T$, one for each training step. For each step t, we filter samples $\{(I, y)\} \subseteq \mathcal{D}_t$ so that y only contains HOI annotations belonging to $\mathcal{C}_t$. After step t is complete, training switches to the next step $t+1$, so the model observes a different image set $\mathcal{D}_{t+1}$ and HOI annotations $\mathcal{C}_{t+1}$. The detailed partition of HOI categories for each step is further explained in Section 4.1.

3.2. Model Architecture

We propose a model architecture that disentangles the learning of object and relation categories, allowing the model to decouple the learning of relation representations from object information.

As shown in Fig. 3, the model consists of two main components: an object branch and a relation branch. In the object branch, for an input image I, we use a pre-trained object detector with DETR architecture [4] to generate a global image feature $\mathbf{g}$ and a set of object detection results. Non-maximum suppression and thresholding are

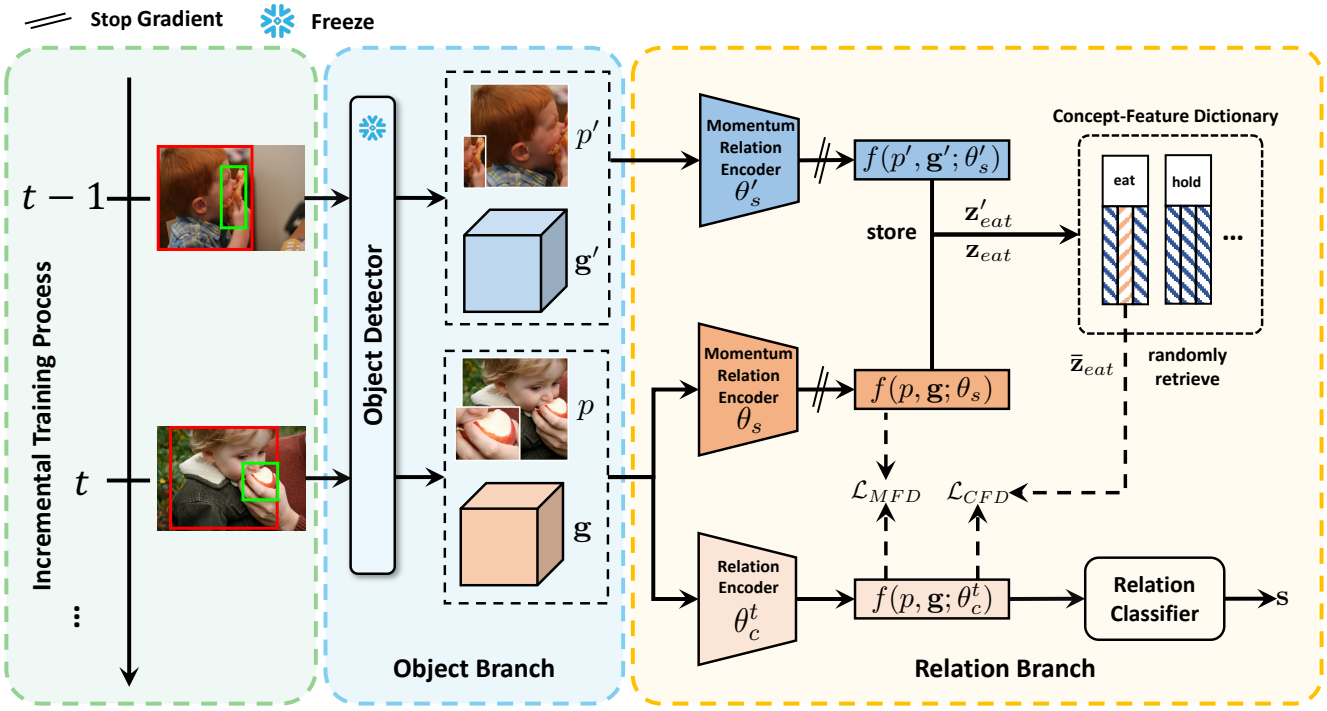


Figure 3. The pipeline of our relation representation learning framework. At each training step t, the object branch outputs the box pair information p and the global image feature $\mathbf{g}$. These are then fed into the relation branch, where a momentum teacher processes them to produce the reference relation feature $\mathbf{z} = f(p, \mathbf{g}; \theta_s)$, subsequently stored in the concept-feature dictionary. Concurrently, the current encoder takes the same input and yields $f(p, \mathbf{g}; \theta_c^t)$, facilitating the computation of distillation losses $\mathcal{L}_{MFD}$ and $\mathcal{L}_{CFD}$ with $\mathbf{z}$ and the invariant relation feature $\bar{\mathbf{z}}$ randomly retrieved from the dictionary, respectively.

subsequently applied, leaving a smaller result set $\{d_i\}_{i=1}^n$, where $d_i = (\mathbf{b}_i, s_i, c_i, \mathbf{x}_i)$ consists of the box coordinates $\mathbf{b}_i \in \mathbb{R}^4$, the confidence score $s_i \in [0, 1]$, the predicted object class $c_i \in \mathcal{O}$, and the object feature $\mathbf{x}_i$. The output boxes are paired as human-object candidates, forming the set $\mathcal{P} = \{p = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{b}_i, \mathbf{b}_j) \mid i \neq j, c_i = \text{human}\}$. In the relation branch, together with the global feature $\mathbf{g}$, p is taken as input to the relation encoder f parameterized by θ, producing the relation representation $f(p, \mathbf{g}; \theta)$ for the box pair p, and finally being fed to the relation classifier to predict the relation logits $\mathbf{s}$. To fully leverage the information from the pre-trained object detector, we integrate the object confidence scores into the final score computation of each human-object pair. The final score of p is formulated as:

$$\tilde{\mathbf{s}} = (s_i)^\lambda \cdot (s_j)^\lambda \cdot \sigma(\mathbf{s}) \tag{1}$$

where λ is a constant to suppress overconfident objects [78] and σ is the sigmoid function. The training loss $\mathcal{L}_{rel}$ for this architecture is the focal loss [42] on the relation classification, which deals with the imbalance between positive and negative examples.

For the relation encoder, we follow the interaction head of UPT [78] to design its structure, and we also follow [24] to introduce cosine normalization to the vanilla softmax function in the relation classifier so that the model can be less biased to new classes. In addition, given that the object categories $\mathcal{O}$ are known beforehand, we propose freezing the object detector, which is pre-trained on all object categories within the dataset. This strategy shifts our focus more toward enhancing the learning process of the relation branch.

3.3. Incremental Relation Distillation

To learn an HOI detector in the IHOID setting, we now introduce our incremental training strategy. Our method is motivated by a key observation: in each incremental step, the model encounters only a subset of object-relation combinations, which induce a strong dependency between the relation and object representations. Consequently, the change of object components in HOI combinations often leads to significant variations in relation representations. Such observations underscore the necessity of designing robust and invariant features for relations, thereby enhancing the model's ability to alleviate interaction drift and generalize to unseen HOI combinations.

To address this challenge, we propose two distillation strategies used in our framework, which enable us to learn robust and invariant relation representations during the incremental learning process. First, the Concept Feature Distillation (CFD) aims to help the model learn the invariant

features shared by different samples within the same relation category. Second, the Momentum Feature Distillation (MFD) focuses on maintaining the discriminative relation features across steps to mitigate catastrophic forgetting. These two distillations are achieved by a momentum teacher and a dynamically updated concept-feature dictionary. The two components and two distillations are introduced in the following paragraphs.

Momentum Teacher While previous knowledge distillation methods [13, 24, 37] in continual learning only used the model from the last step as the teacher, in our approach, we employ a more robust momentum teacher which was firstly used in unsupervised learning [5, 23]. The teacher model retains the knowledge from previous steps and mildly adapts to the data in the current step.

Specifically, in addition to maintaining a frequently changing current model θ_c^t at step t, we keep a model θ_s as the teacher, which remains detached from the training process. At each iteration, the current model θ_c^t adapts to the target distribution and simultaneously updates the model θ_s using exponential moving weighted average:

$$\theta_s = m\theta_s + \texttt{sg}[(1-m)\theta_c^t] \tag{2}$$

where $\texttt{sg}$ is the stop-gradient operation and m is the momentum value. By introducing this momentum teacher, we ensure a stable reference for the model's learning process, preserving old knowledge while allowing soft adaptation to new data distributions.

Concept-Feature Dictionary Inspired by RelViT [45], here we introduce the concept-feature dictionary and its store-retrieve strategy. The concept in our setup is the relation category, while this can be changed to other things in other incremental learning setups [16, 49], such as object, attribute, or HOI category and so on. For each concept, the dictionary stores a queue of invariant reference representations and is dynamically updated during the training process, serving as a storage and retrieval intermediary.

At training step t, denote the total number of learned concepts as N_t and the set of learned concepts as $\mathcal{R}_{1:t} = \{R_1, \cdots, R_{N_t}\}$. A concept-feature dictionary is denoted as $\{(R_1, Q_1), \cdots, (R_{N_t}, Q_{N_t})\}$, where each relation concept R_i is associated with a queue Q_i of capacity L. About the storage of features, for each image, we first sample candidate box pairs $\mathcal{P}_s$ from predicted candidates $\mathcal{P}$. Only a portion of relation representations, whose minimum box-pair IoU with the ground truth exceeding 0.5, is eligible to be stored in the dictionary. Since a given $p \in \mathcal{P}_s$ can have multiple relation labels, it thereby involves multiple concepts, so we denote the relation concept set associated with p as $\mathcal{R}_p \subseteq \mathcal{R}_{1:t}$. For any $(p, \mathcal{R}_p)$, we uniformly select a concept $R \in \mathcal{R}_p$ and randomly retrieve a relation feature $\bar{\mathbf{z}}$ from the queue Q corresponding to R, and the retrieved feature $\bar{\mathbf{z}}$ is used to compute CFD loss. Simultaneously, we input the box pair p into the teacher network θ_s to obtain a new relation feature $\mathbf{z} = f(p, \mathbf{g}; \theta_s)$, which is enqueued into Q. If Q is full, the oldest relation feature is dequeued.

Note that initially, the concept-feature dictionary does not contain any key-value pairs. When the concept R belonging to $\mathcal{R}_p$ does not exist in the dictionary, we only create a new entry (R, Q) in the dictionary and add the feature $\mathbf{z}$ to Q without retrieval. The dictionary is updated in each training iteration, allowing the continuous expansion and refinement of the reference relation features.

Concept Feature Distillation To encourage the model to learn invariant characteristics of the same relation category across different samples, we introduce the Concept Feature Distillation (CFD). For each box pair p, the CFD loss is defined as

$$\mathcal{L}_{CFD} = \|f(p, \mathbf{g}; \theta_c^t) - \bar{\mathbf{z}}\|_2^2 \tag{3}$$

where $\bar{\mathbf{z}}$ is the invariant relation representation retrieved from the concept-feature dictionary.

Momentum Feature Distillation To simultaneously learn HOI classes with new relation categories while retaining the knowledge from previous steps, we develop Momentum Feature Distillation (MFD) to learn stable relation representations. For each box pair p, the MFD loss is computed as

$$\mathcal{L}_{MFD} = \left\|f(p, \mathbf{g}; \theta_s) - f(p, \mathbf{g}; \theta_c^t)\right\|_2^2 \tag{4}$$

where $f(p, \mathbf{g}; \theta_s)$ and $f(p, \mathbf{g}; \theta_c^t)$ are the relation representations obtained from the momentum teacher and the current model, respectively.

Concept Distribution Distillation In addition to the proposed two distillations, we employ a classic technique known as Concept Distribution Distillation (CDD) [37] to prevent the forgetting of the classifier. For each box pair p, with a maintained model θ_c^{t-1} from the last step, this distillation loss is defined as follows:

$$\mathcal{L}_{CDD} = -\sum_{i=1}^{N_{t-1}} \mathbf{q}_i^{t-1} \log \mathbf{q}_i^t \tag{5}$$

where $\mathbf{q}_i^t = \frac{e^{\mathbf{s}_i^t/T}}{\sum_{j=1}^{N_r^{t-1}} e^{\mathbf{s}_j^t/T}}$, $\mathbf{q}_i^{t-1} = \frac{e^{\mathbf{s}_i^{t-1}/T}}{\sum_{j=1}^{N_r^{t-1}} e^{\mathbf{s}_j^{t-1}/T}}$, N_r^{t-1} is the number of learned relation categories until the end of step $t-1$, $\mathbf{s}_i^t$ is the i^{th} element in the logits $\mathbf{s}^t$ given by the current step model θ_c^t, $\mathbf{s}_i^{t-1}$ is the i^{th} element in the logits $\mathbf{s}^{t-1}$ given by the last step model θ_c^{t-1}, and T is the temperature set as $T = 1$ by default.

Additionally, we allow the model to use a memory [53] with limited size to store and replay exemplar samples from previous steps.

3.4. Training Objectives

In the training stage, the total loss $\mathcal{L}_{total}$ is the weighted sum of four components calculated over all box-pair candidates: the standard relation classification loss $\mathcal{L}_{rel}$ illustrated in Sec. 3.2, CDD loss, CFD loss, and MFD loss. $\mathcal{L}_{total}$ is thereby formulated as

$$\mathcal{L}_{total} = \sum_{p \in \mathcal{P}_s} (\mathcal{L}_{rel} + \alpha_0 \mathcal{L}_{CDD} + \alpha_1 \mathcal{L}_{MFD} + \alpha_2 \mathcal{L}_{CFD}) \tag{6}$$

where $\alpha_0, \alpha_1, \alpha_2$ are tunable hyperparameters used to balance the contribution of each loss term.

4. Experiments

We conduct a series of experiments to verify the effectiveness of our method. In this section, we first introduce the experiment setup in Sec. 4.1. Then we show our experimental results in Sec. 4.2, followed by the ablation study in Sec. 4.3.

4.1. Experiment Setup

Baselines The baselines we compare with include three types of state-of-the-art (SOTA) methods, allowing us to showcase the advantages of our approach in this task comprehensively. Firstly, we investigate whether the challenges of this new problem can be effectively addressed by several SOTA class incremental learning (CIL) methods, namely LwF [37], ER [53], PODNet [13], ESMER [54], and PCR [40], in comparison to our approach. Additionally, we examine the adaptability of General-Inc [70], a proposed method for solving the general incremental learning problem, to our task. Lastly, for a comprehensive and fair comparison, we apply VCL [25] and SCL [28] to our model architecture, along with PODNet-flat, as baselines for zero-shot HOI detection.

Datasets To investigate the IHOID setting, we conduct experiments on two widely used HOI datasets HICO-DET [6] and V-COCO [20]. We perform preprocessing on them, including removing the `no interaction` category in HICO-DET and excluding four body motion categories and the `point instr` category in V-COCO following [78]. Specifically, any HOI and its corresponding bounding box annotations related to these relation categories are removed, and images lacking annotations after the removal are also discarded. The detailed statistics of two datasets before and after preprocessing are shown in Tab. 1.

Training Set Partition When partitioning the training set for each step, we follow the problem formulation guidelines in Sec. 3.1. Object-relation pairs that do not appear during training are considered as unseen HOI combinations, constituting our zero-shot test samples. Specifically, each new HOI class that emerges in training step t is characterized by the introduction of either a new object or a new relation category not present in previous steps. Formally, for $C_i = (O_j, R_k)$ in $\mathcal{C}_t$, either $O_j \notin \mathcal{O}_{1:t-1}$ or $R_k \notin \mathcal{R}_{1:t-1}$ holds true. We segment HICO-DET into 5-step and 10-step training subsets, whereas V-COCO is only split into 5-step subsets, as a 10-step division results in too small subsets to effectively train. Detailed information on the statistics of the partitions is shown in the Suppl.

Evaluation Metrics In the IHOID setup, we adopt the mean Average Precision (mAP) as the primary evaluation metric for both datasets, aligning with the standard test setting of HICO-DET. The matching criterion for a detected human-object pair hinges on the intersection over union (IoU) between the predicted and ground truth bounding boxes for both human and object. A pair is deemed a match if the IoU surpasses 0.5. Among these matched pairs, the one with the highest score is labeled as a true positive, while others are regarded as false positives. Any pairs lacking a corresponding ground truth match are also classified as false positives.

To evaluate the model's forgetting, we test the mAP of not only all learned HOI categories but also all old HOI categories at the end of each time step, which is denoted as *Old* in Tab. 2. We also evaluate three category sets on HICO-DET: all HOI categories (*Full*), HOI categories with less than 10 training instances (*Rare*), and the remaining ones (*Non-rare*). To assess the model's robustness against Interaction Drift (RID), we test them on a subset of HOI categories featuring relation classes present in both current and previous steps. Additionally, the generalization performance on zero-shot HOIs is demonstrated by testing models on unseen HOI combinations until the end of each time step.

Implementation Details Regarding the object detector, we fine-tune the DETR model [4] on the HICO-DET and V-COCO datasets before incremental training, subsequently freezing its weights. All experiments on each dataset use the same DETR weight for a fair comparison. Specifically for HICO-DET, we utilize publicly available DETR models that are pre-trained on MS COCO [41]. However, for V-COCO, to avoid overlap with its test set present in the COCO val2017 subset, we train DETR models from scratch on MS COCO, excluding images from the V-COCO test set. Across all setups, ResNet-50 [22] serves as our backbone for image feature extraction. As for the input to the relation branch, the post-processing of detection results obtained from DETR follows the same approach as UPT [78]. The length of each queue in the concept-feature dictionary L is 10, the object scores exponential parameter λ is 1 during training and 2.8 during evaluation, and the momentum value m is set as 0.999 [23].

During training, we utilize the AdamW optimizer with a total of 15 epochs for each training step. The learning rate

Table 1. Statistics of HICO-DET and V-COCO (in format *before/after* preprocessing).

Datasets	# training images	# test images	# object categories	# action categories	# interaction categories
HICO-DET	37,633/33,601	9,546/8,528	80/80	117/116	600/520
V-COCO	5,400/3,923	4,946/3,501	80/80	29/24	287/259

Table 2. Experiment results of our model compared with other incremental learning methods on HICO-DET.

Methods	$T = 5$					$T = 10$				
	Old	Full	Rare	Non-rare	RID	Old	Full	Rare	Non-rare	RID
Joint (Upper Bound)	-	39.20	30.34	41.70	-	-	40.06	30.26	42.78	-
Finetune	19.50	22.71	17.36	24.22	22.63	19.22	20.79	17.38	21.74	19.49
LwF [37]	21.83	24.57	18.54	26.27	27.08	20.30	21.66	15.83	23.28	24.20
ER [53]	23.85	25.96	16.07	28.76	27.19	22.78	23.94	19.06	25.29	24.87
PODNet-flat [13]	29.06	30.19	22.50	32.36	29.82	28.28	29.03	20.34	31.45	28.17
General-Inc [70]	26.88	28.03	22.10	29.71	28.72	28.06	28.46	21.21	30.47	29.29
ESMER [54]	24.65	24.05	15.78	26.40	28.99	26.04	25.64	20.41	27.10	31.75
PCR [40]	26.27	27.37	21.94	28.91	28.62	25.01	25.91	19.54	27.68	25.27
IRD (Ours)	**30.96**	**31.47**	**24.90**	**33.32**	**33.64**	**30.69**	**30.65**	**21.30**	**33.25**	**36.18**

Table 3. Experiment results of our model compared with other incremental learning methods on V-COCO under the 5-step setting.

Methods	Old	Full	RID
Joint (Upper Bound)	-	46.54	-
Finetune	27.81	33.65	24.35
LwF [37]	30.60	36.01	28.83
ER [53]	32.38	37.40	27.41
PODNet-flat [13]	34.32	38.48	29.06
General-Inc [70]	32.70	37.15	29.39
ESMER [54]	30.42	33.23	27.94
PCR [40]	28.45	33.62	23.59
IRD (Ours)	**36.54**	**40.91**	**30.44**

is initially set at 10^{-4} and decreases by a factor of 10 after the 10th epoch. The coefficients of the loss terms are set as $\alpha_0 = 2.5, \alpha_1 = 0.05, \alpha_2 = 0.05$. The training is conducted on 4 GPUs, with a batch size of 8 per GPU.

4.2. Results

Here we summarize the experimental results for the IHOID setting on both the HICO-DET and V-COCO datasets. The tables show the results after the last training step. Except for LwF, all baselines and our method use a memory size of 100 exemplar images to ensure a fair comparison. Tab. 2 and Tab. 3 show that our IRD consistently outperforms the baselines in alleviating forgetting and interaction drift on both datasets, and Tab. 4 shows the superiority of IRD on zero-shot HOI combinations.

Catastrophic Forgetting Our model effectively mitigates forgetting of old HOI classes, achieving mAP of 30.96% and 30.69% on old classes and shows better robustness on full categories with mAP of 31.47% and 33.22% on HICO-DET with 5 steps and 10 steps, respectively. On V-COCO, our method also achieves SOTA performance with over 2% mAP improvement compared to baselines.

Robustness to Interaction Drift Our model demonstrates advantages in addressing the challenges of interaction drift as reflected in the metrics of Robustness to Interaction Drift (RID). On HICO-DET, our model surpasses the best baseline by 3.82% and 4.43% under 5-step and 10-step setups, respectively, and achieves over 30% mAP on V-COCO. This is partly attributed to the better knowledge retention of old concepts by our model. Additionally, our model learns invariant relation representations for samples with the same relation but different HOI classes, enabling generalization to new object-relation pairs.

Zero-shot HOI Detection In zero-shot HOI detection shown by Tab. 4, our model not only achieves SOTA performances with around 2% mAP improvement compared with baselines but also surpasses the models trained in the joint training scenario. This is because the joint training model, unlike our CFD loss, only uses focal loss for learning and does not consider maintaining the consistency of representations among samples within the same relation class. The VCL and SCL methods show limited improvement, partly due to the noise introduced by the unconstrained combination of object and relation features. Moreover, such data augmentation can only generate limited new combinations within a single training step and cannot handle zero-shot

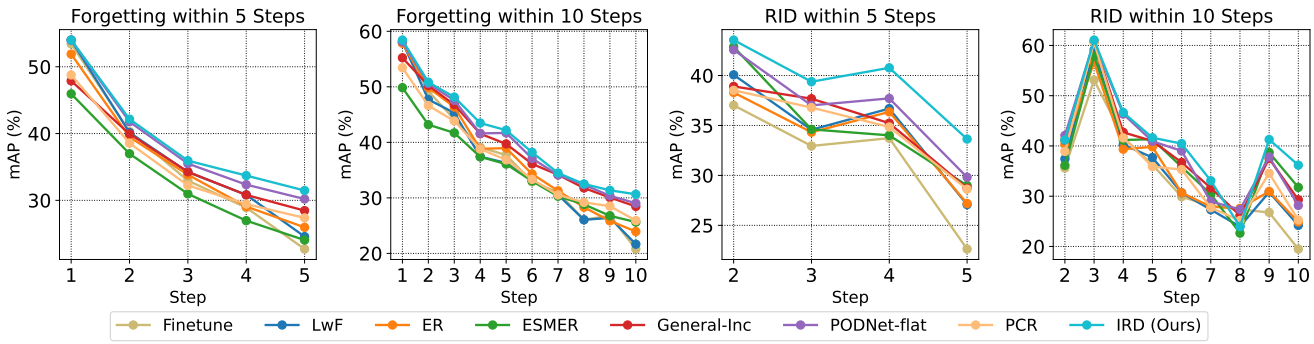


Figure 4. Performances w.r.t. steps on HICO-DET benchmark for catastrophic forgetting and robustness to interaction drift (RID).

Table 4. Zero-shot evaluation on both HICO-DET and VCOCO datasets.

Methods	HICO-DET		V-COCO
	$T = 5$	$T = 10$	$T = 5$
Joint	20.46	20.44	30.52
Finetune	12.67	11.58	23.01
LwF [37]	13.99	12.94	26.48
ER [53]	15.16	13.13	27.89
PODNet-flat [13]	18.33	18.15	30.05
General-Inc [70]	19.21	18.65	27.39
ESMER [54]	16.37	19.04	27.45
PCR [40]	18.67	18.31	25.68
PODNet-flat+VCL [13, 25]	19.88	19.16	30.70
PODNet-flat+SCL [13, 28]	20.99	19.37	31.46
IRD (Ours)	**22.70**	**21.76**	**32.13**

combinations consisting of object and relation classes that appear in different steps.

Also, we include curves of performance w.r.t. steps on the HICO-DET benchmark with three splits, which are shown in Fig. 4 and Fig. 5. Fig. 4 illustrates the mAP of all learned HOI classes and the model's RID until the last time step. It demonstrates that our method maintains a consistent advantage throughout the learning process. Fig. 5 showcases the model's performance on zero-shot HOI combinations. The advantage of our model over multiple training steps, especially on 10 steps, can be attributed to the proposed Concept Feature Distillation (CFD) and Momentum Feature Distillation (MFD), which enables the model to learn robust and invariant relation representations.

4.3. Ablation Study

To assess the necessity and effectiveness of our proposed two distillations in the IRD framework, ablative experiments are conducted on the HICO-DET dataset, starting with the naive model with $\mathcal{L}_{rel}$ and $\mathcal{L}_{CDD}$. The results are summarized in Tab. 5. The CFD (Concept Feature Distillation) component significantly improves the performance of

Table 5. Ablation study on HICO-DET under the 5-step setting.

MFD	CFD	Memory	Old	Full	Rare	Non-Rare	UC
-	-	-	21.83	24.57	18.54	26.27	13.99
-	✓	-	28.83	29.17	22.92	30.93	21.37
✓	-	-	27.61	28.91	23.00	30.58	17.86
✓	✓	-	29.90	30.55	23.78	32.46	21.27
✓	✓	✓	**30.96**	**31.47**	**24.90**	**33.32**	**22.70**

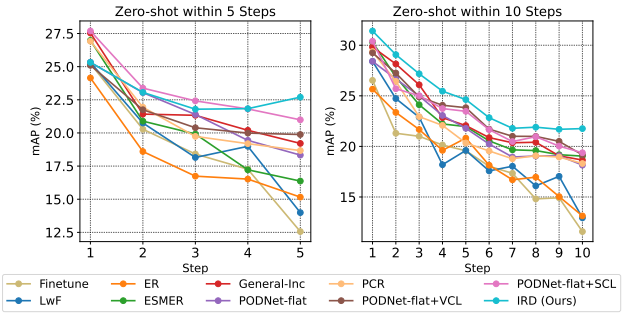


Figure 5. Zero-shot performances w.r.t. steps on HICO-DET.

unseen combinations (UC) and that of previously learned HOIs. It enhances the model's stability and generalization capability by maintaining invariant relation representations for samples with the same relation class but different HOI classes across different steps. The MFD (Momentum Feature Distillation) component aims to ensure learning robust relation representations, effectively mitigating the issue of forgetting. The Memory component refers to the conventional memory replay operation. Its inclusion in the framework demonstrates that replaying old exemplars still provides some benefits to the overall performance.

5. Conclusion

In summary, we introduce the incremental learning setting for human-object interaction detection (IHOID), which is accompanied by three challenges including forgetting previously learned HOI categories, the interaction drift on the relation classes that appear across multiple steps, and the

difficulty in generalizing to zero-shot HOI combinations. Our proposed incremental relation distillation framework offers a novel approach by first disentangling the learning of objects and relations and then emphasizing the acquisition of robust and invariant relation representations through carefully designed distillations. These distillation losses are supported by a momentum teacher and a dynamically updated concept-feature dictionary. Through extensive experiments on the HICO-DET and V-COCO datasets, we have demonstrated the effectiveness of our method to tackle all three challenges.

6. Limitation

Although the effectiveness of our work has been comprehensively evaluated through extensive experiments, the datasets with significantly small training subsets in each step merit further discussion. The larger variation and uncertainty brought by this real-world scenario are valuable in further exploration of incremental HOI detection.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 3

[2] 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. 3

[3] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10460–10469, 2020. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3, 6

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 1, 2, 6, 14

[7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018. 3

[8] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 3

[9] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 3

[10] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 2

[11] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2534–2543, 2021. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2020. 3, 5, 6, 7, 8, 13

[14] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022. 3

[15] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1291–1299, 2021. 2

[16] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022. 5

[17] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[18] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 696–712. Springer, 2020. 2

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2

[20] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 6, 14

[21] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5, 6

[24] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 3, 4, 5

[25] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer, 2020. 1, 3, 6, 8, 13

[26] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 2

[27] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 1, 2, 3

[28] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 461–478. Springer, 2022. 3, 6, 8, 13

[29] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2021. 3

[30] Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Lin Liang, Jenq-Neng Hwang, and Zicheng Liu. The overlooked classifier in human-object interaction recognition. *arXiv preprint arXiv:2203.05676*, 2022. 2

[31] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 498–514. Springer, 2020. 2

[32] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*. IEEE, 2021. 2

[33] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2925–2934, 2023. 2

[34] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[35] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2

[36] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33:5011–5022, 2020. 2

[37] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1, 2, 3, 5, 6, 7, 8, 13

[38] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2

[39] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 3

[40] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24246–24255, 2023. 6, 7, 8, 13

[41] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[43] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2262–2268. IEEE, 2018. 3

[44] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 3

[45] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar.

Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022. 5

[46] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 3

[47] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–82, 2018.

[48] Marc Masana, Tinne Tuytelaars, and Joost van de Weijer. Ternary feature masks: continual learning without any forgetting. *arXiv preprint arXiv:2001.08714*, 4(5):6, 2020. 3

[49] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1114–1124, 2021. 1, 5

[50] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 1, 2

[51] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 2

[52] 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. 3, 13

[53] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018. 2, 5, 6, 7, 8, 13

[54] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Error sensitivity modulation based experience replay: Mitigating abrupt representation drift in continual learning. *arXiv preprint arXiv:2302.11344*, 2023. 3, 6, 7, 8, 13

[55] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018. 3

[56] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018. 3

[57] Yujun Shi, Li Yuan, Yunpeng Chen, and Jiashi Feng. Continual learning via bit-level information preserving. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 16674–16683, 2021. 3

[58] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2

[59] Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9634–9643, 2021. 3

[60] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13617–13626, 2020. 2

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 15

[62] Bo Wan and Tinne Tuytelaars. Exploiting clip for zero-shot hoi detection requires knowledge distillation at multiple levels, 2023. 2

[63] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 2

[64] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 248–264. Springer, 2020. 2

[65] Liyuan Wang, Kuo Yang, Chongxuan Li, Lanqing Hong, Zhenguo Li, and Jun Zhu. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5383–5392, 2021. 3

[66] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 3

[67] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 2

[68] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 3

[69] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15275–15284, 2023. 2

[70] Jiangwei Xie, Shipeng Yan, and Xuming He. General incremental learning with domain-aware categorical representa-

tions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14351–14360, 2022. 3, 6, 7, 8, 13

[71] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021. 1, 3

[72] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[73] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training, 2023. 3

[74] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 3

[75] Mengyao Zhai, Lei Chen, and Greg Mori. Hyper-lifelonggan: Scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2255, 2021. 3

[76] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 2

[77] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2

[78] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 1, 2, 4, 6

[79] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10411–10421, 2023. 2

[80] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19392–19402, 2023. 3

[81] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 2

[82] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021. 2

The supplementary material is structured as follows: Sec. A provides additional details on our proposed IRD method and the application of baseline methods in the IHOID setup. Sec. B elaborates on the experimental setup, including dataset partitioning and evaluation metrics. Additional results and analyses on HICO-DET, such as further sensitive studies and t-SNE visualizations of relation features, are presented in Sec. C. Finally, Sec. D discusses per-step performance on V-COCO.

A. Our IRD and Baselines

A.1. Details on Memory Design in IRD

As mentioned in Sec. 3.3, our proposed IRD incorporates a limited-size memory [52, 53] to further mitigate forgetting. We allocate the memory equally among the data from different steps, dividing it based on the number of previous steps. For instance, with a total memory size of S and $t-1$ previous steps, we allocate $S/(t-1)$ memory slots for each step's training subset. Upon the arrival of a new step t, a portion of data from each existing step in the memory is randomly removed to accommodate data from step t. Consequently, the allocation for each step is then adjusted to S/t. This strategy ensures a balanced and diverse set of HOI categories within the memory.

A.2. Modified Baselines in IHOID

In this section, we detail the adaptation of the baselines from Sec. 4.2 to our framework for use in the IHOID setting. This includes LwF [37], ER [53], ESMER [54], General-Inc [70], PODNet [13], PCR [40], PODNet+VCL [13, 25], and PODNet+SCL [13, 28]. Given the differences in data and model structures between the CIL and IHOID tasks, we have retained the original implementations of LwF and ER, while the other baselines were modified to fit our specific context.

ESMER While preserving ESMER's fundamental weighting strategy, we have adapted its episodic memory management. In our approach, an image is selected for memory storage if it meets two criteria: first, its predicted box pairs must have an Intersection over Union (IoU) greater than 0.5 with the ground truth pairs, and second, the focal loss values for these pairs should classify them as low-loss samples, in line with the guidelines set out in [54].

General-Inc In the IHOID setting, the problem of interaction drift is similar to the continuous domain shift for data belonging to the same category in the general incremental learning setting [70]. Drawing from General-Inc's strategy, we adopt the concept of maintaining and dynamically expanding multiple prototypes per category. Specifically, for each new data related to a relation class R that involves n object categories, we create n additional prototypes for class R.

PODNet We mainly adopt the feature distillation idea from PODNet. Since the whole module before the relation classifier is Transformer-based rather than CNN-based as designed in PODNet, we discard the spatial distillation loss and only retain the distillation of the final embedding, which is denoted as *POD-flat* in the original paper [13]. Specifically, we take the box pair information into the models from both step $t-1$ and step t, and calculate $\mathcal{L}_{POD-flat}$ using the output relation representations. This modified baseline method is referred to as **PODNet-flat** in our paper.

PCR We integrate Proxy-based Contrastive Replay (PCR) into our framework, as outlined in [40], due to its compatible memory replay approach and contrastive-based loss, which align well with our IHOID setting. We utilize both original and augmented samples as inputs to the model, employing the proxy-based classifier during training. For inference, we follow the same process mentioned in the paper.

VCL VCL was originally designed for zero-shot HOI detection in the joint training setting. We adapt the idea of recombining object features and relation features from different images for data augmentation in VCL. In our IHOID setting, we recombine human box features and object box features from different images. This modified VCL method serves as a plugin in our framework. As a result, we introduce a baseline method for zero-shot HOI detection in the incremental learning setup, denoted as **PODNet-flat+VCL**, which combines the PODNet-flat approach with the modified VCL technique.

SCL SCL tackles the same problem setting as VCL. Building upon the ideas of VCL, SCL further introduces the concept confidence matrix which is essentially the cross-product space of objects and relations. This enables many more combinations than VCL so that zero-shot HOIs can be detected more effectively during inference. In each step of our incremental setting, we separately maintain the confidence matrix and dynamically update the confidence scores during training. We add the *concept discovery loss* term corresponding to SCL to the baseline with VCL, giving **PODNet-flat+SCL**, which combines the PODNet-flat approach with SCL.

B. Experiment Setup

B.1. Statistics on Training Set Partition

In this section, detailed statistics of dataset partitioning under the IHOID setting are presented in Tables 6, 7, and 8.

Table 6. Statistics of the HICO-DET dataset partitioned into 10 steps.

	step 1	step 2	step 3	step 4	step 5	step 6	step 7	step 8	step 9	step 10
HOI	17	17	17	17	17	17	17	17	17	17
Relation	13	13	13	15	17	13	15	15	15	15
Object	15	14	16	14	16	16	15	15	15	14
Training Images	3497	1837	1411	2538	2590	1496	1668	1949	2941	1203
Drift Interaction	-	5	5	9	18	13	12	7	19	15
Unseen Combination	37	69	141	185	245	287	319	332	340	342

Table 7. Statistics of the HICO-DET dataset in the 5-step setup.

	step 1	step 2	step 3	step 4	step 5
HOI	40	40	40	40	35
Relation	26	28	32	33	29
Object	30	32	29	27	24
Training images	5745	6178	2580	4348	3804
Drift Interaction	-	16	26	34	30
Unseen Combination	89	211	294	325	325

Table 8. Statistics of the V-COCO dataset partitioned into 5 steps.

	step 1	step 2	step 3	step 4	step 5
HOI	20	20	20	20	16
Relation	10	8	7	7	10
Object	17	19	19	15	10
Training images	1088	743	1055	1021	1756
Drift Interaction	-	10	29	45	48
Unseen Combination	33	75	118	138	138

Specifically, the first four rows of each table indicate the quantities of HOI categories, relation categories, object categories, and training images, respectively. The fifth row, labeled **Drift Interaction**, represents all HOIs affected by the interaction drift issue discussed in Sec. 1. The final row, **Unseen Combination**, quantifies the zero-shot HOI combinations.

HICO-DET Tables 6 and 7 detail the division of the training subset into 10 and 5 steps, respectively, for the HICO-DET dataset [6], as described in Sec. 4.1. Notably, at the end of both steps 4 and 5, the model encounters an identical number of zero-shot combinations. This is because the new relations and objects introduced in step 5 do not form any additional valid unseen combinations, leaving the count of zero-shot HOI combinations unchanged.

V-COCO For the V-COCO [20] dataset, we follow the data partitioning described in Sec. 3.1 and Sec. 4.1, and the specific statistics of subsets are presented in Tab. 8.

B.2. Evaluation Metrics

As mentioned in Sec. 4.1, we mainly evaluate our method using three metrics: robustness to catastrophic forgetting, robustness against interaction drift (**RID**), and performance on zero-shot HOI categories (**UC**), which are tested on different HOI category subsets. Here, we provide a detailed explanation of how these metrics are conducted after training at each time step t.

Catastrophic Forgetting First, for the robustness to catastrophic forgetting, we measure the mAP on all the HOI categories $\mathcal{C}_{1:t}$ that have been learned up to step t.

Robustness against Interaction Drift Second, for RID, we evaluate the model's mAP on a subset $\mathcal{C}_t^{id}$ of previously learned classes that encounter interaction drift. Specifically, $\mathcal{C}_t^{id}$ consists of HOI categories $C_i = (O_j, R_k)$ where $C_i \in \mathcal{C}_{1:t-1}$, $R_k \in \mathcal{R}_{1:t-1}$, and $R_k \in \mathcal{R}_t$ at the same time. In other words, for each old class C_i, the corresponding relation category has appeared in both the previous steps and the current step.

Zero-shot HOI Detection Finally, for zero-shot HOI detection, we evaluate the model on a set of HOI categories $\mathcal{C}_t^{uc}$ that the model has not seen before, but they are reasonable combinations of object and relation categories based on the objects and relations the model has encountered up to the current step. Specifically, $\mathcal{C}_t^{uc}$ consists of HOI categories $C_i = (O_j, R_k)$ where $O_j \in \mathcal{O}_{1:t}$, $R_k \in \mathcal{R}_{1:t}$, and $C_i \notin \mathcal{C}_{1:t}$.

C. Experiment Results on HICO-DET

C.1. More Analysis

Memory Size We explore the impact of memory size on our proposed method. Tab. 9 presents the results of different memory sizes used in our method. It can be observed that as the memory size increases, the model's performance generally improves. Furthermore, even without using memory, our method still outperforms the baselines, indicating the superiority of our proposed method in handling the IHOID setting.

Length of Queue In Tab. 10, we show the sensitive study on the length L of each queue in our concept-feature dictionary. We observe our method works better when $L = 10$.

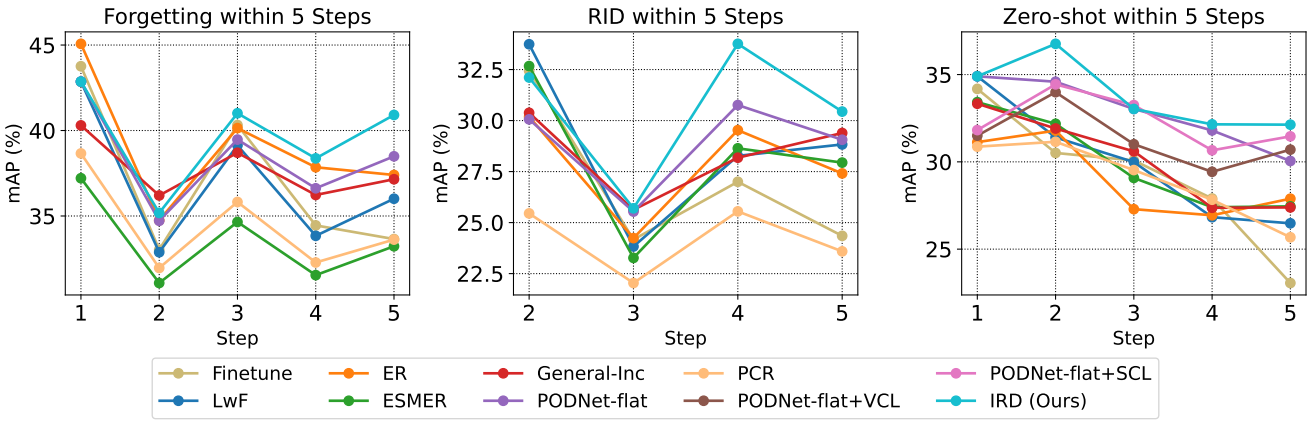
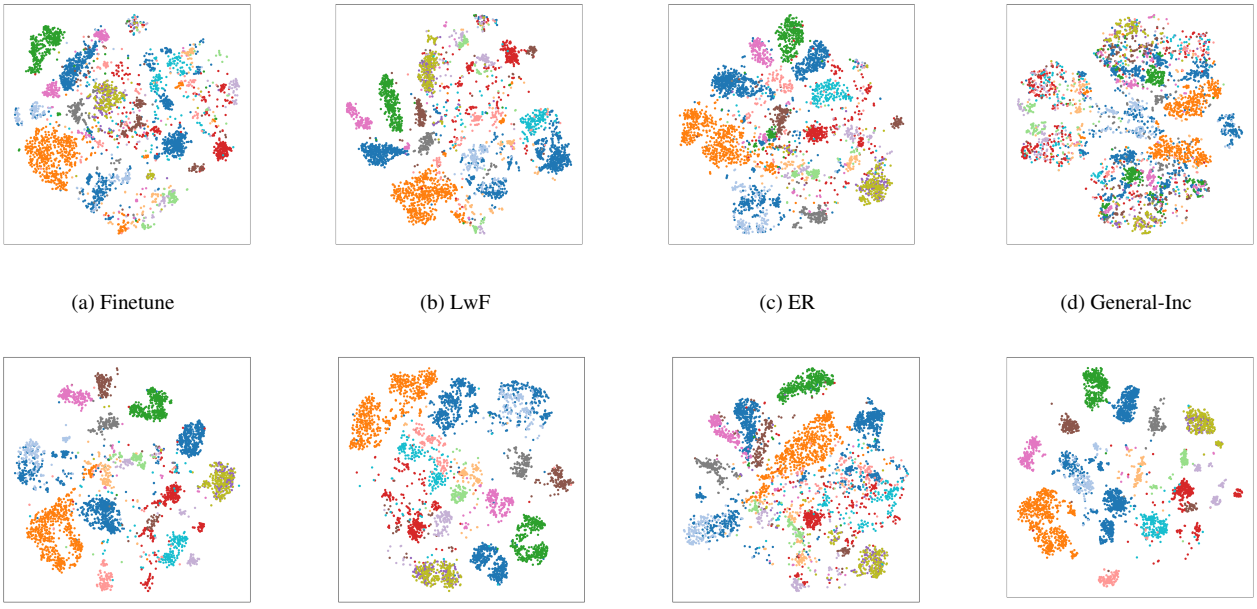


Figure 6. Performances w.r.t. steps on V-COCO benchmark for three evaluation indices under the 5-step setting.



(a) Finetune (b) LwF (c) ER (d) General-Inc

(e) PODNet-flat (f) ESMER (g) PCR (h) IRD (Ours)

Figure 7. t-SNE visualization on relation features after t steps.

The maximum performance difference is only 0.81% when using different values for L, which indicates our method is robust to this hyperparameter.

Table 9. Analysis of the memory size.

size	Old	Full	Rare	Non-Rare	UC	RID
0	29.90	30.55	23.78	32.46	21.27	31.11
20	30.04	30.64	23.74	32.60	22.56	33.06
50	30.69	31.13	23.00	**33.40**	22.11	**34.46**
100	**30.96**	**31.47**	**24.90**	33.32	**22.70**	33.64

C.2. t-SNE Visualization

We utilized the t-SNE visualization technique [61] to demonstrate the robustness and invariance of relation fea-

Table 10. Sensitive study on the length L of each queue in the concept-feature dictionary.

L	Old	Full	Rare	Non-Rare	UC	RID
5	30.80	31.46	24.09	**33.55**	22.31	33.09
10	**30.96**	**31.47**	**24.90**	33.32	**22.70**	33.64
20	30.63	31.16	23.61	33.29	22.39	**34.11**

tures learned by our method. Fig. 7 shows the t-SNE visualization of relation features from the test set at the final step, where features of the same relation category are indicated by identical colors. Our method enables a more compact distribution of features for each relation, suggesting that despite varying HOI classes, the relation features remain consistent across combinations with different objects. This pattern underscores our method's effectiveness in learning relation features that are invariant to the specific objects involved.

D. Experiment Results on V-COCO

In this section, we present the model's performance across different steps on the V-COCO dataset within the 5-step framework. Fig. 6 displays the model's resilience against catastrophic forgetting, its ability to handle interaction drift, and its effectiveness in identifying zero-shot HOI combinations. Each subplot in the figure highlights our model's performance on the V-COCO benchmark, showcasing its robustness and adaptability in various scenarios.