Detecting Human-Object Interaction via Fabricated Compositional Learning

Zhi Hou[1], Baosheng Yu[1], Yu Qiao[2,3], Xiaojiang Peng[4], Dacheng Tao[1]
[1] School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory
[4] Shenzhen Technology University
zhou9878@uni.sydney.edu.au, baosheng.yu@sydney.edu.au, yu.qiao@siat.ac.cn,
pengxiaojiang@sztu.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Human-Object Interaction (HOI) detection, inferring the relationships between human and objects from images/videos, is a fundamental task for high-level scene understanding. However, HOI detection usually suffers from the open long-tailed nature of interactions with objects, while human has extremely powerful compositional perception ability to cognize rare or unseen HOI samples. Inspired by this, we devise a novel HOI compositional learning framework, termed as Fabricated Compositional Learning (FCL), to address the problem of open long-tailed HOI detection. Specifically, we introduce an object fabricator to generate effective object representations, and then combine verbs and fabricated objects to compose new HOI samples. With the proposed object fabricator, we are able to generate large-scale HOI samples for rare and unseen categories to alleviate the open long-tailed issues in HOI detection. Extensive experiments on the most popular HOI detection dataset, HICO-DET, demonstrate the effectiveness of the proposed method for imbalanced HOI detection and significantly improve the state-of-the-art performance on rare and unseen HOI categories. Code is available at https://github.com/zhihou7/HOI-CL.

1. Introduction

Human-Object Interaction (HOI) detection, which aims to localize and infer relationships between human and objects in images/videos, $\langle human, verb, object\rangle$, is an essential step towards deeper scene and action understanding [7, 13]. In real-world scenarios, long-tailed distributions are common for the data perceived by human vision system, *e.g.*, actions/verbs and objects [39]. The combinatorial nature of HOI further highlights the issues of long-tailed distributions in HOI detection, while human can efficiently learn to recognize seen and even unseen HOIs from limited samples. An intuitive example of open long-tailed HOI detection is shown in Figure 1, in which one can easily recognize the unseen action "ride bear", nevertheless it never even happened. However, existing HOI detection approaches usually focus on either the head [13, 36, 57], the tail [62] or unseen categories [48, 43], leaving the problem of open long-tailed HOI detection poorly investigated.

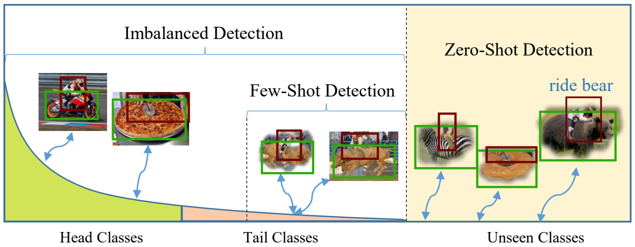

Figure 1. Open long-tailed HOI detection addresses the problem of imbalanced learning and zero-shot learning in a unified way. We propose to compose new HOIs for open long-tailed HOI detection. Specifically, the blurred HOIs, *e.g.*, "ride bear", are composite. See more examples in supplementary materials.

Open long-tailed HOI detection falls into the category of the long-tailed zero-shot learning problem, which is usually referred into several isolated problems, including long-tailed learning [26, 22], few-shot learning [10, 52], zero-shot learning [33]. To address the problem of imbalanced training data, existing methods mainly focus on three strategies: 1) re-sampling [9, 19, 27]; 2) re-weighted loss functions [8, 5, 21]; and 3) knowledge transfer [58, 39, 10, 33, 47, 11]. Specifically, re-sampling and re-weighted loss functions are usually designed for imbalance problem, while knowledge transfer is introduced to relieve all the long-tailed [58], few-shot [49], and zero-shot problem [61, 11]. Recently, two popular knowledge transfer methods have received increasing attention from the com-

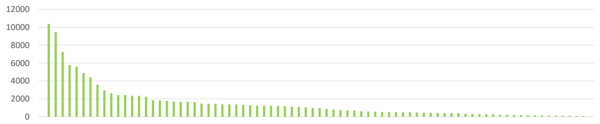

Figure 2. Illustration of distribution of the number of object box in HICO-DET dataset. The categories are sorted by the number of instances.

munity, data generation [58, 57, 61, 39, 10, 33, 47, 29] (transferring head/base classes to tail/unseen classes) and visual-semantic embedding [11, 42] (transferring from language knowledge). Along the first way, we address the problem of open long-tailed HOI detection from the perspective of HOI generation.

Unlike the samples in typical long-tailed zero-shot learning for visual recognition, each HOI sample is composed of a verb and an object, and different HOIs may share the same verb or object (*e.g.*, "ride bike" and "ride horse"). In cognitive science, human perceives concepts as the compositions of shareable components [4, 23] (*e.g.*, verb and object in HOI), which indicates that human can conceive a new concept through a composition of existing components. Inspired by this, several zero-and few-shot HOI detection approaches have been proposed to enforce the factored primitive (verb and object) representation of the same primitive class to be similar among different HOIs, such as factorized model [48, 3] and factor visual-language model [62, 43, 3]. However, regularizing factor representation, *i.e.* enforcing the same verb/object representation to be similar among different HOIs, is only sub-optimal for HOI detection. Recently, Hou *et al.* [24] present to compose novel HOI samples via combining decomposed verbs and objects between pair-wise images and within image. Nevertheless, it still remains a great challenge to compose massive HOI samples in each minibatch from images due to limited number of HOIs in each image, especially when the distribution of objects/verbs is also long-tailed. We demonstrate the distribution of the number of objects in Figure 2.

The long-tailed distribution of objects/verbs makes it difficult to compose new HOIs from each mini-batch, significantly degrading the performance of compositional learning-based methods for rare and zero-shot HOI detection [24]. Inspired by recent success of visual object representation generation [61, 20, 57], we thus apply fabricated object representation, instead of fabricated verb representation, to compose more balanced HOIs. We referred to the proposed compositional learning framework with fabricated object representation as Fabricated Compositional Learning or FCL. Specifically, we first extract verb representations from input images, and then design a simple yet efficient object fabricator to generate object representation. Next, the generated visual object features are further combined with the verb features to compose new HOI samples. With the proposed object fabricator, we are able to generate balanced objects for each verb within the mini-batch of training data as well as compose massive balanced HOI training samples.

The main contributions of this paper can be summarized as follows: 1) proposing to compose HOI samples for Open Long-Tailed HOI detection; 2) designing an object fabricator to generate objects for HOI composition; 3) significantly outperforming recent state-of-the-art methods on HICO-DET dataset among rare and unseen categories.

2. Related Works

HOI Detection. HOI detection is essential for deeper scene and action understanding [7]. Recent HOI detection approaches usually focus on representation learning [13, 66, 51, 55, 53], zero/few-shot generalization [48, 62, 43, 3, 24], and One-Stage HOI detection [36, 56]. Specifically, existing methods improve HOI representation learning by exploring the relationships among different features [44, 66, 51], including pose information [35, 53, 34], context [13, 55], and human parts [66]; Generalization methods for HOI detection mainly include visual-language model [43, 62], factorized model [48, 18, 51, 3], and HOI composition [24]. Recently, Liao *et al.* [36] and Wang *et al.* [56] propose to detect the interaction point for HOI by heatmap-based localization [41]. Wang *et al.* [54] try to detect HOI with novel objects by leveraging human visual clues to localize interacting objects. However, existing HOI approaches usually fail to investigate the imbalance issue and zero-shot detection. Inspired by the factorized model [48], we propose to compose visual verb and fabricated objects to address the open long-tailed issue in HOI detection. Furthermore, according to whether detect the objects with a separated detector or not, existing HOI detection approaches can be divided into two categories: 1) one-stage [48, 36, 56, 14] and two-stage [13, 35, 66, 51, 55, 62, 3, 51]. Two-stage methods usually achieve better performance and our method falls into this category.

Compositional Learning. Irving Biederman illustrates that human representations of concepts are decomposable [4]. Meanwhile, Lake *et al.* [32] argue compositionality is one of the key blocks in a human-like learning system. Tokmakov *et al.* [50] apply the compositional deep representation into few-shot learning. External knowledge graph and graph convolutional networks in [28] are used to compose verb-object pairs for HOI recognition. Recently, Hou *et al.* [24] propose a novel visual compositional learning framework to compose HOIs from image-pairs for HOI detection, failing to address the open and long-tailed issues. Therefore, we further compose verb and fake object representations for HOI detection.

Generalized Zero/Few-Shot Learning. Different from typical zero/few-shot learning [10, 33, 52], generalized

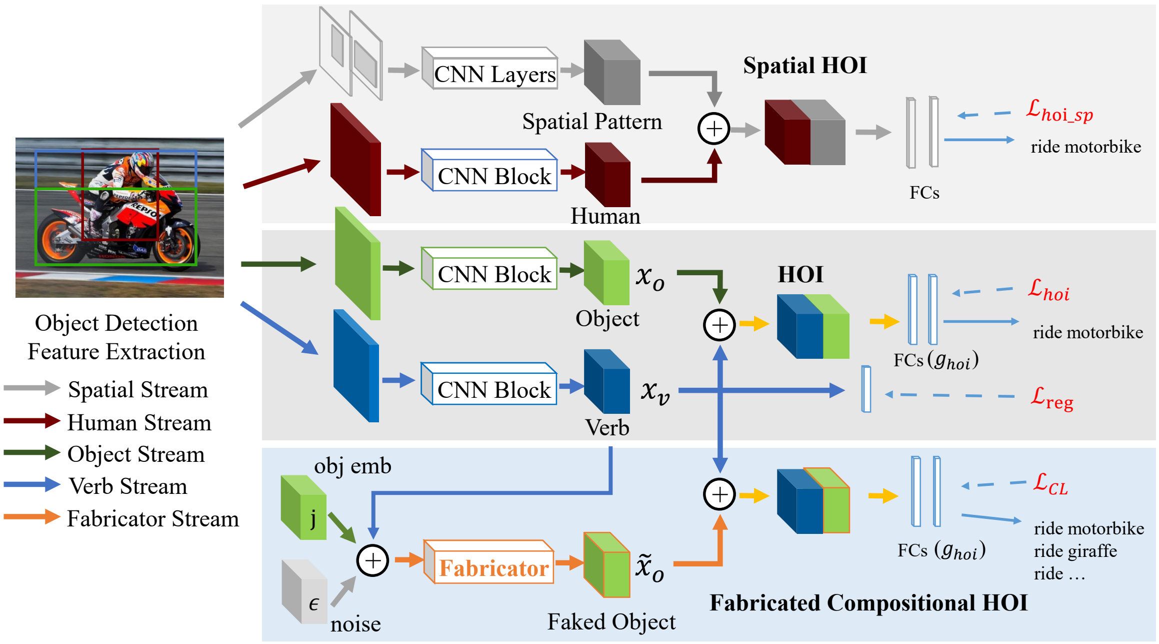

Figure 3. An overview of the proposed multi-branch fabricated compositional learning framework for HOI detection. We first detect human and object with Faster-RCNN [45] from the image. Next, with ROI-Pooling and residual CNN blocks, we extract human features, verb features and object features. Meanwhile, an object identity embedding, verb feature and noise are input into Fabricator to generate fake object feature. Then, these features are fed into the following branches: individual spatial HOI branch, HOI branch and fabricated compositional HOI branch. Finally, HOI representations from HOI branch and fabricated branch are optimized by a shared FC-Classifier, while HOI representations from spatial branch are classified by an individual FC-Classifier. In fabricated compositional HOI branch, verb features are combined with fabricated objects to construct fabricated HOIs.

zero/few-shot learning [60] is a more realistic variant, since the performance is evaluated on both seen and unseen classes [47, 6]. The distribution of HOIs is naturally long-tailed [7], *i.e.*, most classes have a few training examples. Moreover, the open long-tailed HOI detection aims to handle the long-tailed, low-shot and zero-shot issue in a unified way. The long-tailed data distribution [26, 22, 25] is one of challenging problem in visual recognition. Currently, re-sampling [16, 27], specific loss [37, 8, 65, 5, 21], knowledge transfer [58, 39], and data generation [57, 31, 61, 2] are major strategies for imbalanced learning [26, 22, 25]. To make full use of the composition characteristic of HOI, we aim to compose HOI samples by visual feature generation to relieve the open long-tailed issue in HOI detection. Recent feature generation methods [31, 61] mainly depend on Variational Autoencoder [30] and Generative Adversarial Network [15], which usually suffer from the problem of model collapse [46]. Wang *et al*. [57] present a new method for low-shot learning that directly learns to hallucinate examples that are useful for classification. Similar to [57], we compose HOI samples with an object fabricator in an end-to-end optimization without using the adversarial loss.

3. Method

In this section, we first describe the multi-branch compositional learning framework for HOI detection. We then introduce the proposed fabricated compositional learning for open long-tailed HOI detection.

3.1. Multi-branch HOI Detection

HOI detection aims to find the interactions between human and different objects in a given image/video. Existing HOI detection methods [13, 35, 3] usually contain two separated stages: 1) human and object detection; and 2) interaction detection. Specifically, we first use a common object detector, *e.g.*, Faster R-CNN [45], to localize the positions and extract the features for both human and objects. According to the union of human and object bounding boxes, we then extract the verb feature from the feature map of backbone networks via the ROI-Pooling operation. Similar to [13, 18, 35], an additional stream for spatial pattern, *i.e.*, spatial stream, is defined as the concatenation of human and object masks, *i.e.*, the value in the human/object bounding box region is 1 and 0 elsewhere. As a result, we obtain several input streams from the first stage, *i.e.*, human stream, object stream, verb stream, and spatial stream.

The input streams from the first stage then are used to construct different branches in the second stage: 1) **the spatial HOI branch**, which concatenates the spatial and the human streams to construct spatial HOI feature for HOI recognition; 2) **the HOI branch**, which concatenates the verb and the object streams; and 3) **the fabricated compositional branch**, which is based on a new stream, the fabricator stream, to generate fake object features for composing new HOIs. Specifically, the fabricated compositional branch generates novel HOIs by combining visual verb features and generated object features. The main multi-branch HOI detection framework is shown in Figure 3, and we leave the details of the fabricated compositional branch in next section.

3.2. Fabricated Compositional Learning

The motivation of compositional learning is to decompose a model/concept into several sub-models/concepts, in which each sub-model/concept focuses on a specific task, and then all responses are coordinated and aggregated to make the final prediction [4]. Recent compositional learning method for HOI detection considers each HOI as the combination of a verb and an object to compose new HOIs from objects and verbs within the mini-batch of training samples [28, 24]. However, existing compositional learning methods fail to address the problem of long-tailed distribution on objects.

To address the open long-tailed issue, we propose to generate balanced objects for each decoupled visual verb as follows. Formally, we denote $\mathbf{l}_v$ as the label of a verb x_v, $\mathbf{l}_o$ as the label of an object x_o and $\mathbf{y}$ as the HOI label of $\langle x_v, x_o \rangle$. Given another verb representation $\hat{x}_v$ (sharing the same label $\mathbf{l}_v$ with x_v), and another object representation $\hat{x}_o$ (sharing the same label $\mathbf{l}_o$ with x_o), regardless of the sources of the verb and object representations, an effective composition of verb and object should be

$$g_{hoi}(\hat{x}_v, \hat{x}_o) \approx g_{hoi}(x_v, x_o), \tag{1}$$

where g_{hoi} indicates the HOI classification network. By doing this, we can compose new verb-object pair $\langle \hat{x}_v, \hat{x}_o \rangle$, which have similar semantic type $\mathbf{y}$ to the real pair $\langle x_v, x_o \rangle$, to relieve the scarcity of rare and unseen HOI categories. To generate effective verb-object pair $\langle \hat{x}_v, \hat{x}_o \rangle$, we regularize the verb representation $\hat{x}_v$ and object representation $\hat{x}_o$ such that same verbs/objects have similar feature representation.

Similar to previous approaches, such as factor visual-language joint embedding [62, 43] and factorized model [48, 18], when $\hat{x}_v$ is similar to x_v and $\hat{x}_o$ is similar to x_o, we then have that Equation (1) can be generalized to HOI detection via the compositional branch. We refer to the proposed compositional learning framework with fabricated object representation as Fabricated Compositional Learning or FCL. We train the proposed method with composited HOI samples $\langle \hat{x}_v, \hat{x}_o \rangle$ in an end-to-end manner, and the overall loss function are defined as follows:

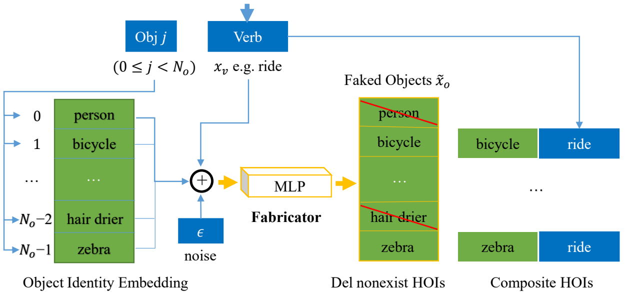

Figure 4. For a given visual verb feature and each j_{th} $(0 \le j < N_o)$, we firstly select the j_{th} object identity embedding. Then, we concatenate verb feature, object embedding and Gaussian noise to input to fabricator for generating a fake object feature. We can fabricate N_o objects for a verb feature. We finally remove nonexisting HOIs as described in Section 3.2.2.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{hoi} + \lambda_2 \mathcal{L}_{CL} + \lambda_3 \mathcal{L}_{reg} + \mathcal{L}_{hoi\_sp}, \tag{2}$$

where $\mathcal{L}_{reg}$ aims to regularize verb and object features, $\mathcal{L}_{CL}$ indicates a typical compositional learning loss function for the classification network g_{hoi} with composite HOI samples $\langle \hat{x}_v, \hat{x}_o \rangle$ as the input, $\mathcal{L}_{hoi\_sp}$ is the loss for Spatial HOI Branch. $\lambda_1, \lambda_2, \lambda_3$ are the hyper-parameters to balance different loss functions. Specifically, object feature extracted from a pre-trained object detector backbone network (*i.e.* Faster-RCNN [45]) are usually discriminative. Thus, we only regularize verb representation.

3.2.1 Object Generation

The HOI is composed of a verb and an object, in which the verb is usually a very abstract notation compared to the object, making it difficult to directly generate verb features. Recent visual feature generation methods have demonstrated the effectiveness of feature generation for visual object recognition [57, 61]. Therefore, we devise an object fabricator to generate object feature representations for composing novel HOI samples.

The overall framework of object generation is shown in Figure 4. Specifically, we maintain a pool of object identity embeddings, *i.e.*, v_{id}. We provide three kinds of embeddings in supplementary material. In each HOI, the pose of the object is usually influenced by the human who is interacting the object [64], and the person who is interacting with the object is firmly related to verb feature representation. Thus, for each extracted verb and the j_{th} object $(0 \le j < N_o$ and N_o is the number of all different objects), we concatenate the j_{th} object identity embedding v^j_{id}, the verb feature x_v and a noise vector $\epsilon \sim \mathcal{N}(0, 1)$, as the input of the object fabricator, *i.e.*,

$$\hat{x}_o = f_{obj}(\{v^j_{id}, x_v, \epsilon\}), \tag{3}$$

where $\hat{x}_o$ is the fake object feature and f indicates the object fabricator network. Here, the noise ϵ is used to increase the diversity of generated objects. We then combine the fake object feature $\hat{x}_o$ and the verb x_v to compose a new HOI sample $\langle x_v, \hat{x}_o \rangle$. Specifically, during training, both real HOIs and composite HOIs share the same HOI classification network g_{hoi}.

3.2.2 Efficient HOI Composition

To compose new HOIs from verb and object representations, we need to remove some infeasible composite HOIs (*e.g.*, "ride vase") as illustrated in Figure 4. To avoid frequently checking the pair (x_v, x_o), we use an efficient HOI composition similar to [24]. Specifically, the HOI label space is decoupled into verb and object spaces, *i.e.*, the co-occurrence matrices $\mathbf{A}_v \in R^{N_v \times C}$ and $\mathbf{A}_o \in R^{N_o \times C}$, where N_v, N_o, and C indicate the number of verbs, objects and HOI categories, respectively. Given an one-hot HOI label vector $\mathbf{y} \in R^C$, we then have the verb label vectors,

$$\mathbf{l}_v = \mathbf{y}\mathbf{A}_v^{\mathsf{T}}, \tag{4}$$

where $\mathbf{l}_v \in R^{N_v}$ can be a multi-hot vector with multiple verbs, *e.g.*, $\langle \{hold, read\}, book \rangle$). Similarly, combining the verb $\mathbf{l}_v$ with all N_o objects, we have the matrix $\hat{\mathbf{l}}_o \in R^{N_o \times N_o}$ as labels of all N_o fake objects. Let $\hat{\mathbf{l}}_v \in R^{N_o \times N_v}$ denote the verb labels corresponding to fake object features $\hat{\mathbf{l}}_o$, the new interaction label can then be evaluated as follows,

$$\hat{\mathbf{y}} = (\hat{\mathbf{l}}_o \mathbf{A}_o) \ \& \ (\hat{\mathbf{l}}_v \mathbf{A}_v), \tag{5}$$

where $\&$ indicates the logical operation "**and**". Finally, the logical operation automatically filters out the infeasible HOIs since the labels of those infeasible HOIs are all-zero vectors in the label space.

3.3. Optimization

Training. The verb feature contains the pose information of the object, making it difficult to jointly train the network with an object fabricator from scratch. Therefore, we introduce a step-wise training strategy for the long-tailed HOI detection. Firstly, we pre-train the network by $\mathcal{L}_{hoi}$, $\mathcal{L}_{hoi\_sp}$ and $\mathcal{L}_{reg}$ without the fabricator branch. Then, we fix the pre-trained model and train the randomly initialized object fabricator via the loss function for the fabricator branch $\mathcal{L}_{CL}$. Lastly, we jointly fine-tune all branches by $\mathcal{L}$ in an end-to-end manner. To avoid the bias to seen data in the first step, we optimize the network in one step for zero-shot HOI detection (See analysis in Section 4.4).

Inference. The fabricated branch is only used in the training stage, *i.e.*, we remove it during the inference stage. Similar to previous multi-branch methods [13, 35, 24], for each human-object bounding box pair (b_h, b_o), the final HOI prediction $S^c_{h,o}$ for each category $c \in 1, ..., C$, can be evaluated as follows,

$$S^c_{h,o} = s_h \cdot s_o \cdot S^c_{sp} \cdot S^c_{hoi}, \tag{6}$$

where s_h and s_o indicate the object detection scores for the human and object, respectively. S^c_{sp} and S^c_{hoi} are the scores from the Spatial branch and the HOI branch, respectively.

4. Experiments

In this section, we first introduce datasets and metrics, and then provide the details of the implementation of our method. Next, we present our experimental results compared with state-of-the-art approaches. Finally, we conduct ablation studies to validate the components in our work.

4.1. Datasets and Metrics

We adopt the largest HOI datasets HICO-DET [7], which contains 47,776 images including 38,118 images for training and 9,658 images for testing. All 600 HOI categories are constructed from 80 object categories and 117 verb categories. HICO-DET provides more than 150k annotated human-object pairs. In addition, V-COCO is another small HOI dataset with 29 categories [17]. Considering that V-COCO mainly focuses to verb recognition and do not contain a severe long-tailed issue, we mainly evaluate the proposed method on HICO-DET. We also illustrate the result on visual relation detection [40, 63], which requires to detect the triplet (subject, predicate, object) in supplementary materials. We follow the evaluation settings in [7], *i.e.* a HOI prediction is a true positive if 1) both the human and object bounding boxes have IoUs larger than 0.5 with the reference ground truth bounding boxes; and 2) the HOI prediction is accurate.

4.2. Implementation Details

Similar to [3, 24], our HOI detection model contains two separated stages: 1) we finetune the Faster R-CNN detector pre-trained on COCO [38] using HICO-DET to detect the human and objects [1]; 2) we use the proposed FCL model for HOI classification. Specifically, all branches are two-layer MLP sigmoid classifiers with 2048-d input and 1024-d hidden units. Fabricator is a two-layer MLP. The $\mathcal{L}_{reg}$ is a sigmoid classifier for verb representation. $\mathcal{L}_{CL}$, $\mathcal{L}_{hoi}$ and $\mathcal{L}_{hoi\_sp}$ are binary cross entropy losses. $\mathbf{A}_v$ and $\mathbf{A}_o$ are set according to HOI dataset, and we can also set them by prior knowledge to detect more types of unseen HOIs. Besides, to prevent the fabricated HOIs from dominating the model optimization process, we randomly sample fabricated HOIs in each mini-batch to keep that the number of

[1] We use the Faster R-CNN detector implemented in detectron2 [59].

fabricated HOIs is not more than three times the number of non-fabricated HOIs. We train our network for one million iterations by SGD optimizer on the HICO-DET dataset with an initial learning rate of 0.01, a weight decay of 0.0005, and a momentum of 0.9. We set λ_1 as 2.0, λ_2 as 0.5 and λ_3 as 0.3, while we set 1 for the coefficient of $\mathcal{L}_{hoi\_sp}$. The hyper-parameters are ablated in supplementary materials. We jointly fine-tune the model with the object fabricator for 500k iterations, and decay the initial learning rate 0.01 with a cosine annealing schedule. All our experiments on HICO-DET are conducted using TensorFlow [1] on a single Nvidia GeForce RTX 2080Ti GPU. We evaluate V-COCO based on PMFNet [53] with two GPUs. We do not use auxiliary verb loss since there are only two kinds of objects on V-COCO. We set λ_1 as 1 and λ_2 as 0.25 on V-COCO.

4.3. Comparison to Recent State-of-the-Arts

Our method aims to relieve open long-tailed HOI detection. However current approaches usually focus on full categories, rare categories and unseen categories separately. In order to compare with state-of-the-art methods, we evaluate our method on long-tailed detection and generalized zero-shot detection separately. The HOI detection result is evaluated with mean average precision (mAP) (%).

4.3.1 Effectiveness for Zero-Shot HOI Detection

There are different settings [3] for zero-shot HOI detection: 1) unseen composition; and 2) unseen object. Specifically, for the unseen composition setting, it indicates that the training data contains all factors (*i.e.*, verbs and objects) but misses the verb-object pairs; for the unseen object setting, it requires to detect unseen HOIs, in which the object do not appear in the training data. For unseen composition HOI detection, similar to [24], we select two groups of 120 unseen HOIs from tail preferentially (rare first) and from head preferentially (non-rare first) separately, which roughly compares the lowest and highest performances. As a result, we report our result in the following settings: Unseen (120 HOIs), Seen (480 HOIs), Full (600 HOIs) in the "Default" mode on HICO-DET dataset. For a better comparison, we implement the factorized model [48] under our framework for unseen composition zero-shot HOI detection. For unseen object HOI detection, we use the same HOI categories for unseen data as [3] (*i.e.* randomly selecting 12 objects from the 80 objects and picking all HOIs containing there objects as unseen HOIs). Then, we report our results in the setting: Unseen (100 HOIs), Seen (500 HOIs), Full (600 HOIs). To compare with the contemporary work [24], we use the same object detection result released by [24]. Here, our baseline method is the model without object fabricator, *i.e.*, the compositional branch.

Table 1. Comparison of zero-shot detection results of our proposed method. UC indicates unseen composition zero-shot HOI detection. UO indicates unseen object zero-shot HOI detection. For better illustration, we choose the mean UC result of [3].

Method	Type	Unseen	Seen	Full
Shen *et al.* [48]	UC	5.62	-	6.26
FG [3]	UC	11.31	12.74	12.45
VCL [24] (rare first)	UC	10.06	24.28	21.43
Baseline (rare first)	UC	8.94	24.18	21.13
Factorized (rare first)	UC	7.35	22.19	19.22
FCL (rare first)	UC	**13.16**	24.23	**22.01**
VCL [24] (non-rare first)	UC	16.22	18.52	18.06
Baseline (non-rare first)	UC	13.47	19.22	18.07
Factorized (non-rare first)	UC	15.72	16.95	16.71
FCL (non-rare first)	UC	**18.66**	**19.55**	**19.37**
FG [3]	UO	11.22	14.36	13.84
Baseline	UO	12.86	20.77	19.45
FCL	UO	**15.54**	20.74	**19.87**

Table 2. Comparison to the state-of-the-art approaches on HICO-DET dataset [7]. FCL ^{DRG} is FCL with object detector provided by [12]. FCL + VCL means we fuse the result provided in [24] with FCL. VCL^{DRG} uses the released model of VCL.

Method	Default			Known Object		
	Full	Rare	NonRare	Full	Rare	NonRare
FG [3]	21.96	16.43	23.62	-	-	-
IP-Net [56]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [36]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [24]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [12]	24.53	19.47	26.04	27.98	23.11	29.43
Baseline	23.35	17.08	25.22	25.44	18.78	27.43
FCL	**24.68**	**20.03**	**26.07**	**26.80**	**21.61**	**28.35**
FCL + VCL	**25.27**	**20.57**	**26.67**	**27.71**	**22.34**	**28.93**
VCL [24] ^{DRG}	28.33	20.69	30.62	30.59	22.40	33.04
Baseline^{DRG}	28.12	21.07	30.23	30.13	22.30	32.47
FCL ^{DRG}	**29.12**	**23.67**	**30.75**	**31.31**	**25.62**	**33.02**
(FCL + VCL) ^{DRG}	**30.11**	**24.46**	**31.80**	**32.17**	**26.00**	**34.02**
VCL [24] ^{GT}	43.09	32.56	46.24	-	-	-
FCL^{GT}	**44.26**	**35.46**	**46.88**	-	-	-
(FCL + VCL)^{GT}	**45.25**	**36.27**	**47.94**	-	-	-

Unseen composition. Table 1 shows that FCL achieves large improvement on Unseen category by **4.22%** and **5.19%** than baseline, and by **3.10% and 2.44%** compared to previous works [3, 24] on the two selection strategies respectively. Meanwhile, the two selection strategies witness a consistent improvement with FCL on nearly all categories, which indicates that composing novel HOI samples contributes to overcome the scarcity of HOI samples. In rare first selection, FCL has a similr result to baseline and VCL [24] on Seen category. But step-wise optimization can improve the result on Seen category and Full category (See Table 6). In addition, the factorized model has a very poor performance in the head classes compared to our baseline. Noticeably, factorized model achieves better performance on Unseen category than baseline in non-rare first selection while has worse result on Unseen category in rare first selection. FCL witnesses a consistent improvement in differ-

Table 3. Illustration of proposed modules under step-wise optimization. FCL means proposed Fabricated Compositional Learning. V indicates the verb regularization loss.

FCL	V	Full	Rare	NonRare	Unseen
-	-	18.12	15.99	20.65	12.41
✓	-	19.08	17.47	20.95	14.90
-	✓	18.32	16.73	20.82	12.23
✓	✓	**19.61**	**18.69**	**21.13**	**15.86**

ent evaluation settings. In the remaining data, unseen HOIs of rare first zero-shot have more rare verbs (less than 10 instances) than that of non-rare first zero-shot.

Unseen object. We further evaluate FCL in novel object zero-shot HOI detection, which requires to detect HOIs that is interacting with novel objects. Table 1 shows FCL effectively improves the baseline by 2.68% on Unseen Category, although there are no real objects of unseen HOIs in training set. This illustrates the ability of FCL for detecting unseen HOIs with novel objects. Here, the same as [3], we also use a generic detector to enable unseen object detection.

4.3.2 Effectiveness for Long-Tailed HOI Detection

We compare FCL with recent state-of-the-art HOI detection approaches [56, 36, 3, 24, 12] using fine-tuned object detector on HICO-DET to validate its effectiveness on long-tailed HOI detection. For fair comparison, we use the same fine-tuned object detector provided by [24]. For evaluation, we follow the settings in [7]: Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) in "Default" and "Known Object" on HICO-DET.

In Table 2, we find that the proposed method achieves new state-of-the-art performance, **24.68%** and **26.80%** mAP on "Default" and "Known Object". Meanwhile, we achieve a significant performance improvement of **2.82%** over the contemporary best rare performance model [24] under the same object detector, which indicates the effectiveness of the proposed compositional learning for the long-tailed HOI detection. Furthermore, with the same object detection result to [12], our results surprisingly increase to **29.12%** on "Default" mode. Here, we merely change the detection result provided in [24] to that provided in [12] during inference. Particularly, we find our method is complementary to compose HOIs between images [24]. By simply fusing the result provided by [24] with FCL, we can further largely improve the results under different object detectors.

4.3.3 Effectiveness on V-COCO

We also evaluate FCL on V-COCO. Although the data on V-COCO is balanced, FCL still improves the baseline (reproduced PMFNet [53]) in Table 5.

Table 4. Ablation study of fabricator under step-wise optimization. FCL within image means we compose HOIs within image. + verb fabricator is we fabricate verb and object features.

Method	Full	Rare	NonRare	Unseen
FCL	**19.61**	**18.69**	21.13	**15.86**
FCL w/o noise	19.45	17.69	21.22	15.74
FCL w/o verb	19.20	18.02	21.04	14.71
FCL + verb fabricator	19.47	16.93	21.43	15.89

Table 5. Illustration of Fabricated Compositional Learning on V-COCO based on PMFNet [53]

Method	AP_{role}
PMFNet [53]	52.0
Baseline	51.85
FCL	**52.35**

4.4. Ablation Studies

For a robust validation of the proposed method in rare categories and unseen categories simultaneously, we select 24 rare categories and 96 non-rare categories for zero-shot learning (remained 30,662 training instances). This result is roughly between non-rare first selection and rare first selection in Table 1. See supplementary material for unseen type details and ablation study of long-tailed HOI detection based on Table 2. We conduct ablation study on FCL, verb regularization loss, verb fabricator, step-wise optimization and the effect of object detector.

Fabricated Compositional Learning. In Table 3, we find that the proposed compositional method with fabricator can steadily improve the performance and it is orthogonal to verb feature regularization (verb regularization loss).

Verb Feature Regularization. We use a simple auxiliary verb loss to regularize verb features. Although verb regularization loss can slightly improve the rare and unseen category performance (See row 1 and row 3 in Table 3), FCL further achieves better performance. This indicates that regularizing factor features is suboptimal compared to the proposed method. Semantic verb regularization like [62] has a similar result (See supplementary materials).

Verb and Noise for Fabricator. Table 4 demonstrates that performance drops without verb representation or noise. This shows verb representations can provide useful information for generating objects and noise efficiently improves the performance by increasing feature diversity. We meanwhile find the fabricator still effectively improves the baseline without verb or noise by comparing Table 3 and Table 4, which indicates the efficiency of FCL.

Verb Fabricator. The result of fabricating verb features (from verb identity embedding, object features and noise) is even worse as in Table 4. This verifies that it is difficult to directly generate useful verb or HOI samples due to the complexity and abstraction. Supplementary materials provide more visualized analysis of verb and object feature.

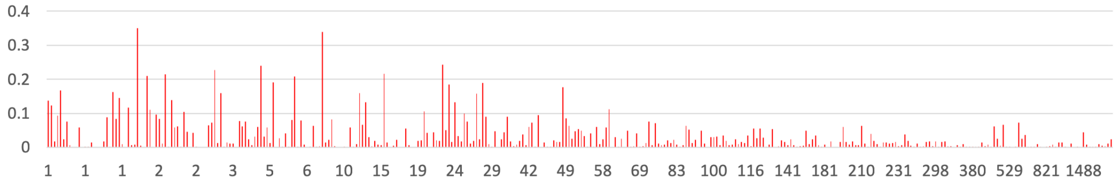

Figure 5. Illustration of the improvement in those improved categories between FCL and baseline on HICO-DET dataset under default setting. The graph is sorted by the frequency of category samples and the horizontal axis is the number of training samples for each category. The result is reported in mAP (%). The details of category name are provided in supplementary materials.

Table 6. Comparison between step-wise optimization and one step optimization. ZS is the setting in our ablation study.

Method	Full	Rare	NonRare	Unseen
one step (long-tailed)	24.03	18.42	25.70	-
step-wise (long-tailed)	**24.68**	**20.03**	**26.07**	-
one step (ZS)	**19.69**	18.22	20.82	**17.64**
step-wise (ZS)	19.61	**18.69**	**21.13**	15.86
one step (rare first ZS)	22.01	15.55	24.56	**13.16**
step-wise (rare first ZS)	**22.45**	**17.19**	**25.34**	12.12
one step (non-rare ZS)	**19.37**	15.39	20.56	**18.66**
step-wise (non-rare ZS)	19.11	**17.12**	**21.02**	15.97

Table 7. Illustration of the effect of fine-tuned detectors on FCL. The COCO detector is trained on COCO dataset provided in [59]. We fine-tune the ResNet-101 Faster R-CNN detector based on Detectron2 [59]. Here, the baseline is our model without fabricator. The last column is object detection result on HICO-DET test.

Method	Detector	Full	Rare	NonRare	Object mAP
Baseline	COCO	21.24	17.44	22.37	20.82
FCL	COCO	**21.80**	**18.73**	**22.71**	20.82
Baseline	HICO-DET	23.94	17.48	25.87	30.79
FCL	HICO-DET	**24.68**	**20.03**	**26.07**	30.79
Baseline	GT	43.63	34.23	46.43	100.00
FCL	GT	**44.26**	**35.46**	**46.88**	100.00

Step-wise Optimization. Table 6 illustrates that step-wise training has better performance in rare and non-rare categories while has worse performance in unseen categories. We think it might be because the model with the step-wise training has the bias to seen categories in the first step since there are no training data for unseen categories.

Object Detector. The quality of detected objects has important effect on two-stage HOI Detection methods [24]. Table 7 shows that the improvement of FCL over baseline is higher with the fine-tuned detector on HOI data. COCO detector without finetuning on HICO-DET contains a large number of false positive and false negative boxes on HICO-DET due to domain shift, which is in fact less useful to evaluate the effectiveness of modeling human interactions for HOI detection. If the detected boxes during inference are false, the features extracted from the false boxes are also unreal and have large shift to the fabricated objects during training. This causes that fabricated objects are less useful for inferring HOIs during inference. Besides, GT boxes provide a strong object label prior for verb recognition.

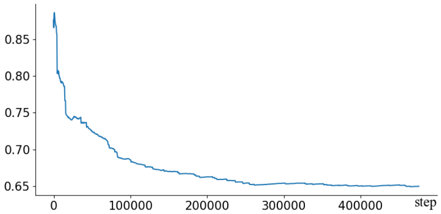

Figure 6. The changing trend of cosine similarity between fabricated object features and real object features during optimization in long-tailed HOI detection in step-wise training.

5. Qualitative Analysis

Illustration of improvement among categories. In Figure 5, we find that *the rarer the category is, the more the proposed method can improve*. The result illustrates the benefit of FCL for long-tailed issue in HOI Detection.

Visualized Analysis between fabricated and real object features. Figure 6 presents that cosine similarity between fabricated and real object features gradually goes down to stability in step-wise training. This demonstrates the end-to-end optimization with shared HOI classifier helps fabricate efficient and similar objects during optimization process. *More analysis of generated object representations by t-SNE is provided in Supplementary Materials.*

6. Conclusion

In this paper, we introduce a Fabricated Compostional Learning approach to compose samples for open long-tailed HOI Detection. Specifically, we design an object fabricator to fabricate object features, and then stitch the fake object features and real verb features to compose HOI samples. Meanwhile, we utilize an auxiliary verb regularization loss to regularize the verb feature for improving Human-Object Interaction generalization. Extensive experiments illustrate the efficiency of FCL on the largest HOI detection benchmarks, particularly for low-shot and zero-shot detection.

Acknowledgements This work was supported in part by Australian Research Council Projects FL-170100117, DP-180103424, IH-180100002, and IC-190100031.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th symposium on operating systems design and implementation (OSDI)*, pages 265–283, 2016. 6

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *CVPR*, pages 6548–6557, 2019. 3

[3] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 2, 3, 5, 6, 7

[4] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 2, 4

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1565–1576, 2019. 1, 3

[6] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016. 3

[7] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389. IEEE, 2018. 1, 2, 3, 5, 6, 7

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 1, 3

[9] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8, 2003. 1

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 28(4):594–611, 2006. 1, 2

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 1, 2

[12] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 6, 7

[13] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 1, 2, 3, 5

[14] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 3

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 3

[17] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[18] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, pages 9677–9685, 2019. 2, 3, 4

[19] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 1

[20] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pages 3018–3027, 2017. 2

[21] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *ICCV*, pages 6469–6479, 2019. 1, 3

[22] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 3

[23] Donald D Hoffman and Whitman Richards. Parts of recognition. *Cognition*, 1983. 2

[24] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 2, 4, 5, 6, 7, 8

[25] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, pages 5375–5384, 2016. 3

[26] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002. 1, 3

[27] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 1, 3

[28] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, pages 234–251, 2018. 2, 4

[29] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *CVPR*, 2020. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018. 3

[32] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 2

[33] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009. 1, 2

[34] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, pages 10166–10175, June 2020. 2

[35] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019. 2, 3, 5

[36] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 1, 2, 6, 7

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[38] 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 *ECCV*, pages 740–755. Springer, 2014. 5

[39] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 1, 2, 3

[40] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016. 5

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 2

[42] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2

[43] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, October 2019. 1, 2, 4

[44] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, pages 401–417, 2018. 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3, 4

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016. 3

[47] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019. 1, 2, 3

[48] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, pages 1568–1576. IEEE, 2018. 1, 2, 4, 6

[49] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017. 1

[50] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *ICCV*, pages 6372–6381, 2019. 2

[51] Oytun Ulutan, ASM Iftekhar, and BS Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 2

[52] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 1, 2

[53] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, pages 9469–9478, 2019. 2, 6, 7

[54] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*, pages 11652–11661, 2020. 2

[55] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, pages 5694–5702, 2019. 2

[56] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2, 6, 7

[57] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, pages 7278–7286, 2018. 1, 2, 3, 4

[58] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NIPS*, 2017. 1, 2, 3

[59] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5, 8

[60] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9):2251–2265, 2018. 3

[61] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 1, 2, 3, 4

[62] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 1, 2, 4, 7

[63] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *CVPR*, pages 5128–5137, 2019. 5

[64] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 4

[65] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, pages 5409–5418, 2017. 3

[66] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019. 2