PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

DecAug: Augmenting HOI Detection via Decomposition

Hao-Shu Fang,[1*] **Yichen Xie,**[1*] **Dian Shao,**[2] **Yong-Lu Li,**[1] **Cewu Lu**[1†]
[1] Shanghai Jiao Tong University, China [2] The Chinese University of Hong Kong
fhaoshu@gmail.com, xieyichen@sjtu.edu.cn, sd017@ie.cuhk.edu.hk, yonglu_li@sjtu.edu.cn, lucewu@sjtu.edu.cn

Abstract

Human-object interaction (HOI) detection requires a large amount of annotated data. Current algorithms suffer from insufficient training samples and category imbalance within datasets. To increase data efficiency, in this paper, we propose an efficient and effective data augmentation method called **DecAug** for HOI detection. Based on our proposed object state similarity metric, object patterns across different HOIs are shared to augment local object appearance features without changing their states. Further, we shift spatial correlation between humans and objects to other feasible configurations with the aid of a pose-guided Gaussian Mixture Model while preserving their interactions. Experiments show that our method brings up to **3.3 mAP** and **1.6 mAP** improvements on V-COCO and HICO-DET dataset for two advanced models. Specifically, interactions with fewer samples enjoy more notable improvement. Our method can be easily integrated into various HOI detection models with negligible extra computational consumption. Our code will be made publicly available.

1 Introduction

Human-object interaction (HOI) detection aims to localize humans and objects as well as infer their interaction categories in a still image. For each interaction, a triplet of $\langle subject, predicate, object \rangle$ should be retrieved. As a subtask of visual relationship detection, HOI detection pays attention to human-centric interactions with objects. It plays an essential role in the understanding of scenes, which facilitates many other fields like activity understanding (Caba Heilbron et al. 2015; Pang et al. 2020), image captioning (Li et al. 2017) and robot learning (Argall et al. 2009).

Along with the recent achievements computer vision has reached, many exciting deep neural network (DNN) models for HOI detection have been developed. They took various types of features into account such as visual features (Gupta and Malik 2015), spatial location (Chao et al. 2018; Xu et al. 2019), human poses (Yao and Fei-Fei 2010; Gkioxari et al. 2018b) or text corpus (Liang, Guan, and Rojas 2020). However, the progress of HOI detection is still slower compared with the achievement in other tasks like object detection and instance segmentation. There are currently two main hindrances to further performance gains. For one thing, HOI detection depends on a better understanding of contextual information. It calls for a large amount of high quality data. However, large datasets are not easily accessible due to the labor intensity of annotation. For another thing, an apparent imbalance inevitably exists between different interaction categories in current large datasets (Gupta and Malik 2015; Chao et al. 2018; Zhuang et al. 2017). Some interactions naturally have much more positive samples than others, such as *look at*, *sit on* and *stand on*, which causes a serious long-tail issue.

(a) Instance-Level Augmentation Example: *heatmap-guided instaboost* (Fang et al. 2019) (left: original, right: augmented)

(b) Our Approach: local object appearance augmentation (left) and global spatial correlation augmentation (right)

Figure 1: (a) shows the result of *heatmap-guided instaboost*. The left is the original image while the right has been augmented. The board is moved far away, which has no negative effect on object detection or instance segmentation. However, it devastates the relationship between the human and object. In contrast, (b) shows the two steps of our DecAug. Local object appearance is changed in the left image. Then, global spatial correlation augmentation is applied in the right one. The human-object interaction remains distinguishable in both.

[*]Equal contribution. Names in alphabetical order.
[†]Cewu Lu is corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute.
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

To tackle such problems, a natural idea is to resort to data augmentation, whose power has been witnessed in many other tasks of computer vision (Cubuk et al. 2019; Simard et al. 2003; Hinterstoisser et al. 2019; Liu et al. 2016; Peng et al. 2018; Jaderberg et al. 2015; Fang et al. 2019). Unfortunately, previous research in cognition (Baldassano, Beck, and Fei-Fei 2017) demonstrated the difficulty of data augmentation for the task of HOI detection. Specifically, image-level random cropping cannot improve the diversity of interactions while instance movement damages the spatial correlation between humans and objects. As shown in Fig. 1(a), it is hard to identify the interaction in the images using such simple augmentation.

In this paper, we propose a novel data augmentation method named **DecAug**. Aiming to improve the diversity of interactions without semantic loss, **DecAug** mainly includes two components: *local object appearance augmentation* and *global spatial correlation augmentation*.

To elaborate, for local object appearance, we propose a simple but effective cross-image instance substitution technique to increase the generalization ability of models towards entity concepts instead of object patterns. An object state similarity metric is also introduced to justify the replacement of an object with another based on their state coherency.

Furthermore, we try to augment the global spatial correlations between humans and objects without contextual loss. According to (Knill, Kersten, and Yuille 1996), the perceptual inference of human derives from information available to observers and some *empirical knowledge* of the world. Intuitively, reasonable placement of objects could also be obtained with prior knowledge from the whole dataset. Inspired by the strong correlation between human pose and HOI (Yao and Fei-Fei 2012), we build a probability distribution of object location for each training sample, which comes from the spatial relationship of other samples with similar human poses. With this distribution aware augmentation, we are able to improve the diversity within each interaction without damaging their semantic meanings.

We conduct extensive experiments on two mainstream datasets: V-COCO (Gupta and Malik 2015) and HICO-DET (Chao et al. 2018). After augmentation, the performance of two advanced open-source models: iCAN (Gao, Zou, and Huang 2018) and Transferable Interactiveness Network (Li et al. 2019c) can be improved by a large margin (**3.3** and **2.6 mAP** on V-COCO; **1.6** and **1.3 mAP** on HICO-DET). Same object detection proposals are used to ensure the improvements come from interaction recognition instead of object detection. Specifically, for those interactions with fewer positive samples, the improvement is more notable, suggesting our method helps tackle the long-tail issue. Our code will be made publicly available.

2 Related Work

2.1 Visual Relationship Detection

Visual relationship detection (Lu et al. 2016; Xu et al. 2017; Gkioxari et al. 2018a; Zellers et al. 2018; Zhang et al. 2017) needs to not only find objects location in an image but also detect the relationships between them. These relationships includes actions (Shao et al. 2020), interactions (Gkioxari et al. 2018a) or other more general relationships (Lu et al. 2016; Zhang et al. 2017). Different from object detection or instance segmentation, visual relationship detection requires to exploit more semantic information (Baldassano, Beck, and Fei-Fei 2017) like the spatial positions of humans and objects (Chao et al. 2018). Since such semantic information is difficult to extract, enough training samples are necessary for these models. Requirement for maintaining the semantic information also poses an extra challenge to data augmentation.

2.2 Human-Object Interaction Detection

Human-object interaction (HOI) detection task is significant for understanding human behavior with objects. Some early work (Gupta and Malik 2015) tried to detect humans and objects separately, which led to limited performance. Christopher *et al.* (Baldassano, Beck, and Fei-Fei 2017) proposed that rather than the sum of parts, more information should be taken into consideration. Gao *et al.* (Gao, Zou, and Huang 2018) proposed an instance-centric attention module to enhance regions of interest. Chao *et al.* (Chao et al. 2018) added the relative spatial relationship between humans and objects into the input of CNN. The significance of pair spatial configuration was further emphasized by Ulutan *et.al.* and Wang *et.al.* (Ulutan, Iftekhar, and Manjunath 2020; Wang et al. 2020), which helped associate the interacted humans and objects. Some recent works (Fang et al. 2018a; Wan et al. 2019; Qi et al. 2018a; Li et al. 2019c, 2020; Zhou et al. 2020) also thought of human poses as a crucial indicator of interaction.

More information means a higher requirement for data amount. There exist some popular datasets for this task such as V-COCO (Gupta and Malik 2015), HICO-DET (Chao et al. 2018), HAKE (Li et al. 2019b) and HCVRD (Zhuang et al. 2017). However, these datasets suffer from internal imbalance between different interaction categories, which is the so-called long-tail issue. Some interaction categories lack positive samples, which encumbers the overall performance. By increasing the diversity of data, data augmentation may help to solve this problem.

2.3 Data Augmentation

Data augmentation has been widely used in many tasks in the field of computer vision, such as image classification (Cubuk et al. 2019; Simard et al. 2003; Krizhevsky, Sutskever, and Hinton 2012), object detection (Hinterstoisser et al. 2019; Liu et al. 2016), and pose estimation (Peng et al. 2018). By generating additional training data, these methods helped to improve performance of various data-hungry models. Specifically, one branch of data augmentation focused on the instance-level, which fully exploited the fine-annotated segmentation of instances. Transformation applied on instances included scaling, rotation (Jaderberg et al. 2015), jitter (Fang et al. 2019), pasting (Kisantal et al. 2019) and affine transform (Khoreva et al. 2019). However, all these above just utilized the information in a single image instead of the whole dataset. Some other work (Choi, Kim, and Kim 2019; Qi et al. 2018b; Liu, Breuel, and Kautz 2017) generated new images with Generative Adversarial Networks

(GAN). Despite the impressive results, GAN needs plentiful extra training data, which is not applicable for current HOI datasets.

Another challenge rises about the placement of segmented instances on augmented images. Dvornik *et al.* (Dvornik, Mairal, and Schmid 2018) placed objects on the background according to the context. However, extra model needed to be trained beforehand. Fang *et al.* (Fang et al. 2019) replaced the offline trained model with online context comparison. Yet, such a method does not preserve the visual relation information between instances inside an image.

Due to the difficulty in context preservation, there exists no effective data augmentation approach to generate extra training samples for visual relation detection tasks. Some prior effort (Bansal et al. 2019; Hou et al. 2020) generated new interaction patterns based on word embedding but these could hardly improve visual diversity in training samples. In contrast, we develop a novel data augmentation method to visually boost data diversity for HOI detection. It makes use of information across the whole dataset as well as reserves visual relationships between humans and objects.

3 Methods

3.1 Overview

For the task of human-object interaction detection, we need to identify the interacting human-object pair, localize their positions and recognize their interaction category. In this paper, we focus on the interaction identification and recognition parts. Given detected humans and objects, a classifier f needs to capture the very subtle details in the image to recognize the relationship R. A human-object interaction can be decomposed into the **background I**, the **human state h** including human appearance, pose, parsing, shape, gaze, etc., the **object state o** including category, 6D pose, occlusion, functionality, etc., and the **spatial relationship s** between the human and object. Mathematically, we have

$$R = f(\mathbf{I}, \mathbf{h}, \mathbf{o}, \mathbf{s}). \tag{1}$$

In this paper, we mainly augment the object state and spatial correlations, coherent with the human perception process. This is nontrivial, since R is very sensitive to the object state and spatial relations. We must find a manifold space in pixel level that could augment the object appearance while preserving the object state. In Sec. 3.2, we introduce our local object appearance augmentation where an *object state similarity metric* is proposed. Meanwhile, to find feasible spatial configurations for global spatial correlation augmentation, we propose the *pose-guided probability distribution map* in Sec. 3.3. An overview of our method is shown in Fig. 2.

3.2 Local Object Appearance Augmentation

When recognizing the HOI, the state of an object is far more important than its texture pattern. For example, when identifying the interaction of *holding a mug*, the standing pose and the occlusion with hands are more important than the mug's color and texture. Thus, we propose to augment the local object appearance features to improve the generalization ability of the network, helping it pay more attention to the crucial object state instead of appearance. The key of such augmentation is to preserve the object state as much as possible. Meanwhile, patterns of augmented objects should be photo-realistic to avoid too many artifacts. Naturally, we can utilize the same category objects from the dataset during training i.e. we replace the object with suitable same category instances in other images. We then explain our principle for objects appearance replacement as follows.

Whether to Replace an Object We first judge whether an object can be substituted or not. Some objects are not suitable to be replaced if they interlock with its neighbours too tightly. In this case, adjacent humans or objects are likely to overlap with each other. As a consequence, it is difficult to find a proper replacement to maintain this interaction.

Intuitively, tightly interlocked instances share a long common borderline. Therefore, we develop a metric called *instance interlocking ratio* measuring the interlocking extent between two adjacent instances in the same image.

As shown in Fig. 4, we define $\mathcal{C}_i$ as the contour of instance O_i and $\mathcal{M}_i$ as the mask of this instance. The contour $\mathcal{C}_i$ serves as the outline of the mask with width w. For two adjacent instances O_i, O_j in the same image, we define their interlocking area as $U_{i,j}$ and their union contour area as $V_{i,j}$:

$$U_{i,j} = S(\mathcal{M}_i \cap \mathcal{C}_j) + S(\mathcal{C}_i \cap \mathcal{M}_j) \tag{2}$$

$$V_{i,j} = S(\mathcal{C}_i \cup \mathcal{C}_j), \tag{3}$$

where $S(\mathcal{A} \cap \mathcal{B})$ denotes the intersection area of $\mathcal{A}$ while $\mathcal{B}$ and $S(\mathcal{A} \cup \mathcal{B})$ denotes the union area of $\mathcal{A}$ and $\mathcal{B}$

Further, the *instance interlocking ratio* between instance O_i, O_j is defined as $r_{i,j}$:

$$r_{i,j} = \frac{U_{i,j}}{V_{i,j}} \in [0, 1]. \tag{4}$$

If two adjacent instances have a high interlocking ratio, chances are that they seriously overlap with each other. As a result, neither of them will be replaced. Thus, objects in image $\mathcal{I}$ that can be replaced are selected from the following set:

$$\mathbf{O}' = \{O_i | O_i \in \mathcal{I}, \forall O_j \in \mathcal{I}, j \neq i : r_{i,j} < t\}, \tag{5}$$

where t is a hyper-parameter as a threshold. We empirically set it to 0.1.

Find Objects with Similar States Despite the same category, objects show various states including pose variance, shape variance, occlusion variance, etc. Objects to be substituted should be matched with others with similar states. Otherwise, the interaction may be affected. Fortunately, we find that the mask of an object can serve as an indicator of the object state. As the projection of an object on the camera plane given a specific pose, instance mask implicitly encodes the shape and 6D pose of the object. Same category objects may share similar shapes and 6D poses if they have similar masks. What's more, an object's occlusion state can also be reflected from the combination of its own and its neighbours' masks. Thus, we build our object state descriptor based on the object mask.

Figure 2: Overview of our method: the first image is the original input (red box: human, blue box: object). The second image is the result of local object appearance augmentation (Sec. 3.2). The third and forth images show the *pose-guided probability distribution map* and the result of global spatial correlation augmentation (Sec. 3.3).

Figure 3: Left images are original ones. We replace the objects (blue boxes) with instances from the middle images (blue boxes). The rightmost images are the augmentation results.

Figure 4: In the middle image, light blue region shows the object mask while dark blue denotes the contour. In the right image, for the two instances O_i, O_j, $U_{i,j}$ is colored in green and $V_{i,j}$ is composed of the green, dark yellow and dark blue regions.

For object O_i with a $W \times H$ bounding box $\mathcal{X}_i$, we divide $\mathcal{X}_i$ into three parts: object mask $\mathcal{M}_i$, background $\mathcal{B}_i$ and adjacent mask $\mathcal{A}_i$. Based on that, we construct the corresponding *object state matrix* $\mathbf{E}_i \in \mathbb{R}^{W \times H}$ for each instance i. Each element in this matrix corresponds with a pixel in the bounding box of instance i. The mapping is shown as follows:

$$\mathbf{E}_i^{x,y} = \begin{cases} 1 & I_{x,y} \in \mathcal{M}_i \\ 0 & I_{x,y} \in \mathcal{B}_i \\ -1 & I_{x,y} \in \mathcal{A}_i \end{cases} \quad (6)$$
$$x \in \{1, \cdots, W\}, y \in \{1, \cdots, H\}$$

where $I_{x,y}$ denotes the pixel with coordinate (x, y) in the bounding box. This matrix $\mathbf{E}_i$ serves as a descriptor of the shape, 6D pose and overlapping condition of instance O_i.

With such descriptor, for objects O_i and O_j with state matrix $\mathbf{E}_i \in \mathbb{R}^{W \times H}$ and $\mathbf{E}_j \in \mathbb{R}^{W' \times H'}$, we define their object state distance $D(i, j)$ as

$$D(i,j) = \frac{\sum_{x,y} |\mathbf{E_i} - \mathbf{E_j'}|}{W \times H}, \quad (7)$$
$$x \in \{1, 2, \cdots, W\}, y \in \{1, 2, \cdots, H\}$$

where $\mathbf{E_j'}$ is the resized matrix of $\mathbf{E_j}$ with same size with $\mathbf{E_i}$.

In the training period, when we process a replaceable object instance O_i in a given image, we randomly select 20 same category objects from other images and calculate their object state distance to O_i. Object with the smallest state distance is selected to replace O_i. Fig. 5 shows some positive or negative examples for replacement.

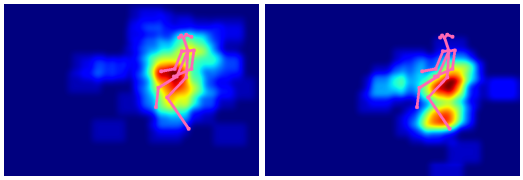
(a) (b)

Figure 5: In (a) and (b), the left is the original image with blue box showing the object. The right above (green box) two images show instances which have high similarity with original object while the right below two (yellow box) in each sub-figure are with low similarity.

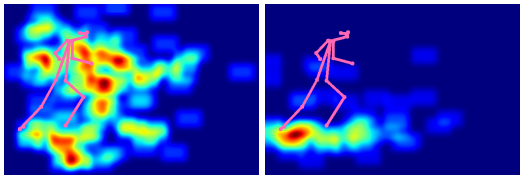
Object Replacement After finding substitution candidate O_s for object O_i, we extract both instances from background using instance masks. For datasets without ground-truth segmentation annotations (like HICO-DET), we generate instance masks with Deep Mask (Pinheiro, Collobert, and Dollár 2015). Matting (He et al. 2011) with alpha channel is adopted to extracted instances so that smoother outlines are acquired. At the same time, we conduct inpainting with Fast Marching (Bertalmio, Bertozzi, and Sapiro 2001) to fill the hole of O_i in the background, which ensures the continuous distribution of the raw image. Finally, we resize object O_s to have the same bounding box size as O_i and paste the segmented instance O_s to the original location of object O_i.

3.3 Global Spatial Correlation Augmentation

In Sec.3.2, the substituted object is pasted at the original position. Although it augments the object appearance, the variance in the image is too slight to cover other unobserved situations. As a supplement, movement with longer distance



(a) *hold* vs *sit*



(b) *look* vs *kick*

Figure 6: For same atomic pose, object placement of different interactions has distinct probability distribution. As shown above, objects of *holding* are close to hands, *sitting* close to buttock or legs, *kicking* close to feet, while objects of *looking* extends from eyes and dispersed extensively.

can effectively improve the performance. In the meantime, such movement should not pose damage to the contextual information.

Therefore, we develop a *pose-guided probability map* to obtain feasible positions of an object. To get the pose data, we follow (Li et al. 2019c) to employ AlphaPose (Fang et al. 2017; Li et al. 2019a) on each human. The generated pose data $\mathbf{K}$ is in COCO (Lin et al. 2014) format with 17 keypoints of each person.

For each human-object interaction category, the relative spatial correlation between the human and object can be described with a 2-dimension vector $\mathbf{v}_{sp}$.

$$\mathbf{v}_{sp} = \mathbf{c}_o - \mathbf{c}_h \tag{8}$$

where $\mathbf{c}_h = (x_{c,h}, y_{c,h})$, $\mathbf{c}_o = (x_{c,o}, y_{c,o})$ are the torso center of human and bounding box center of object respectively.

We perform normalization to deal with different scales of instances and images. Specifically, torso centers of human poses are set as the origins and torso lengths are normalized to one. Also, the relative spatial position vector $\mathbf{v}_{sp}$ is normalized by dividing the torso length. We denote the normalized pose as $\hat{\mathbf{K}}$ and the normalized offset as $\hat{\mathbf{v}}_{sp}$.

To get feasible configurations to augment spatial correlations between human-object pairs, we model the object location $\mathbf{L}$ as a conditional probability distribution *w.r.t* normalized human pose $\hat{\mathbf{K}}$. Considering the proper object location distribution differs across different HOI categories, we learn the conditional distribution for each HOI category separately. Given category $\mathbf{h}$, we model $p(\mathbf{L}|\hat{\mathbf{K}}, \mathbf{h})$ as a mixture of Gaussian distribution. Mathematically, we have

$$p(\mathbf{L}|\hat{\mathbf{K}}, \mathbf{h}) = p(\hat{\mathbf{v}}_{sp}|\mathbf{h}) = \sum_{j=1}^{N_G} \omega_j \, \mathbb{N}(\hat{\mathbf{v}}_{sp}; \mu_j, \sigma_j), \tag{9}$$

where N_G denotes the number of Gaussian distributions, ω_j is the combination weight for the j-th component, $\mathbb{N}(\hat{\mathbf{v}}_{sp}; \mu_j, \sigma_j)$ denotes the j-th multivariate Gaussian distribution with mean μ_j and covariance σ_j. Following (Andriluka et al. 2014; Fang et al. 2018b), we set N_G as the number of atomic poses in the dataset, which is 42 in practice. By enforcing the probability distributions independent among each HOI category, we can ensure the object location coherence within each distribution.

We learn the Gaussian mixture distribution $p(\mathbf{L}|\hat{\mathbf{K}}, \mathbf{h})$ efficiently using an EM algorithm, where the E-step estimates the combination weights ω and M-step updates the Gaussian parameters μ and Σ. To simplify the learning process, we utilize K-means clustering to group the pose data in different HOI categories and initialize the parameters as a warm start. Our learned Gaussian Mixture Model (GMM) constitutes the prior knowledge of relative spatial position distribution of the object. The learned mean μ_j of each Gaussian represents the average of a group of similar 2D poses, which is referred to as atomic pose. Some atomic poses and their corresponding object placement distribution are visualized in Figure 6.

When augmenting an HOI sample in category h given a human pose $\hat{\mathbf{K}}$, we determine the new relative spatial position vector $\mathbf{v}'_{sp}$ by sampling the distribution $p(\mathbf{L}|\hat{\mathbf{K}}, \mathbf{h})$. The augmentation process was illustrated in Fig. 2. Objects are more likely to be placed in a relative spatial position with more prior samples of current interaction type, where they share human poses of the same cluster. With our pose-guided probability map, we are able to augment the spatial correlations between humans and objects in an effective manner.

Table 1: **Results on V-COCO**: Original models' results come from their papers.

Model	**DecAug**	**mAP**$_{role}$
iCAN (Gao, Zou, and Huang 2018)		44.7
iCAN	✓	**48.0**
Improvement		**3.3**↑
TIN (**RP**$_D$**C**$_D$) (Li et al. 2019c)		47.8
TIN (**RP**$_D$**C**$_D$)	✓	**50.4**
Improvement		**2.6**↑

4 Experiments

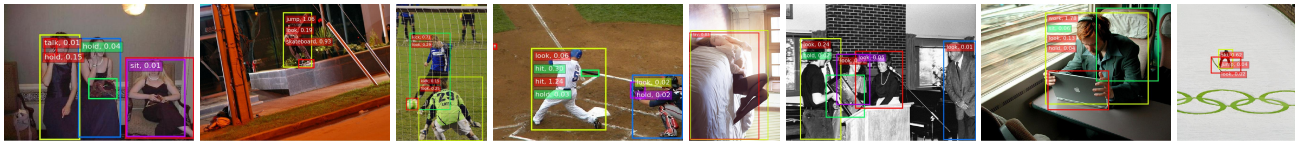
In this section, we first describe the datasets and metrics. We then introduce the base models on which DecAug is performed, including other implementation details. Next, improvements brought by our method is revealed. In the Analysis part, we show that our methods alleviate the long-tail issue. Detailed ablation studies are also conducted.

4.1 Dataset and Metric

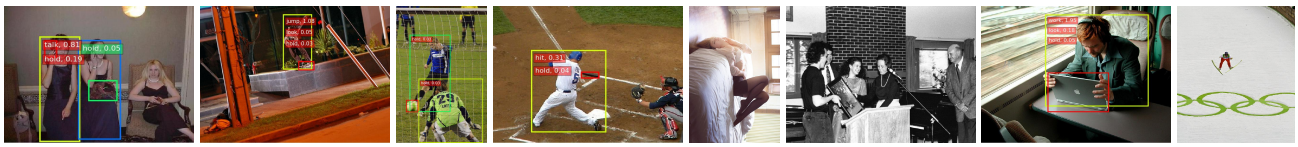
Dataset We evaluate our methods on two mainstream benchmarks: **V-COCO** (Gupta and Malik 2015) and **HICO-DET** (Chao et al. 2018). **V-COCO** is a subset of COCO dataset (Lin et al. 2014) annotated with HOI labels. It includes 10,346 images (2,533 for training, 2,867 for validating and 4,946 for testing) and 16,199 human instances. Each

Table 2: **Results on HICO-DET**: Original models' results come from their papers.

Model	DecAug	Default			Known Object		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN (Gao, Zou, and Huang 2018)		14.84	10.45	16.15	16.26	11.33	17.73
iCAN	✓	**16.39**	**12.23**	**17.63**	**17.85**	**13.68**	**19.10**
Improvement		**1.55**↑	**1.78**↑	**1.48**↑	**1.59**↑	**2.35**↑	**1.37**↑
TIN (**RP**$_D$**C**$_D$) (Li et al. 2019c)		17.03	13.42	18.11	19.17	15.51	20.26
TIN (**RP**$_D$**C**$_D$)	✓	**18.38**	**14.99**	**19.39**	**20.50**	**16.93**	**21.57**
Improvement		**1.35**↑	**1.57**↑	**1.28**↑	**1.33**↑	**1.42**↑	**1.31**↑



(a) iCAN with DecAug



(b) iCAN w/o DecAug

Figure 7: HOI detection results of iCAN trained with (top) and w/o (bottom) DecAug. DecAug brings more accurate detection.

person is annotated with 29 action types, 5 of which have no object. The objects are split into two types: *object* and *instrument*. **HICO-DET** is a subsect of HICO (Chao et al. 2015) dataset annotated with bounding boxes. It contains 47,776 images (38,118 for training and 9,658 for testing), 600 HOI categories over 80 object types and 117 verbs.

Metric We apply the mainstream metric for HOI detection: role mean average precision (role mAP). A prediction is true positive only when 1) HOI classification is correct, and 2) both the IoUs between the predicted bounding boxes of human and object v.s. the ground truth > 0.5.

4.2 Implementation Details

Models We apply DecAug to the following two representative HOI detection models: iCAN (Gao, Zou, and Huang 2018) and Transferable Interactiveness Network (TIN) (Li et al. 2019c). Same object proposals are applied so that we can ensure the performance gain comes from interaction recognition instead of object detection. Baseline results are those reported in their published papers.

Hyper-parameters We adopt stochastic gradient descent in training. All hyper-parameters strictly follow the original setting of our baseline models including iteration number, learning rate, weight decay, backbones and so on.

Augmentation Pipeline During training, the proposed local and global augmentation strategies are incorporated simultaneously since they are complimentary. Each input image will be augmented with a probability of 0.5.

4.3 Results and Comparison

The HOI detection results are evaluated by following the detailed metrics defined by each specific dataset. Results of all the experiments verify the effectiveness and generality of the proposed DecAug.

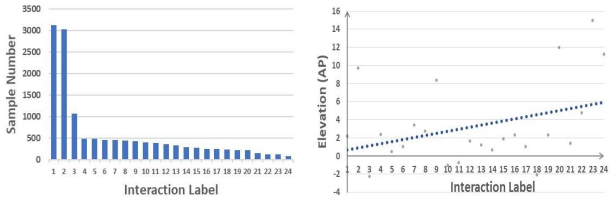
For **V-COCO**, we evaluate mAP_{role} in Tab. 1. We can see that substantial improvements (3.3 mAP) are achieved by applying DecAug.

For **HICO-DET**, we evaluate mAP_{role} of Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) interactions of two different settings: Default and Known Object. Results are shown in Tab. 2. Unsurprisingly, notable performance gain is also achieved (1.6 mAP), indicating the effectiveness of our methods on large datasets without ground-truth segmentation or keypoints.

In Fig 7, we show some visualized results trained with and w/o DecAug. We can see examples that our DecAug compensates for some ignorance and corrects some detection mistakes, as it makes full use of the information within the whole dataset.

4.4 Analysis

Long-tail Issue is a pervasive problem in HOI datasets. In Fig. 8(a), we plot the number of samples from each interaction categories in V-COCO dataset. Severe data imbalance could be observed. Fig. 8(b) then shows the effectiveness of DecAug, from which we can clearly see that more remarkable



(a) sample number (b) improvement

Figure 8: (a) shows the training sample number of each interaction category in V-COCO dataset. Interaction names are ignored for clarity. Grey points in (b) show the AP_{role} improvement of each interaction category (corresponding with (a)). The blue dotted line in (b) reveals the fitted trend line of AP_{role} improvement. We can see that the elevation increases as the sample number decreases.

improvement could be made for interaction categories with fewer training samples. This is because DecAug could make full use of favourable information (e.g. object appearance, spatial locations) across the whole dataset.

Training Efficiency As a data augmentation method, DecAug can be embedded into various existing models conveniently with negligible offline data preprocessing. During training, it could generate augmented samples online without burdening GPUs. As shown in Tab. 3, when applying multi-threads data loader, the training efficiency almost remains unaffected.

4.5 Ablation Study

In this part, the impact of 1) local object appearance augmentation (LOA), and 2) global spatial correlation augmentation (GSC) in DecAug is separately analyzed. The results are shown in Tab. 3. We can see that both strategies contribute notably to the final performance. Next, we evaluate the effectiveness of some key techniques in each strategy.

Table 3: **Ablation Study by Removing Either Component**: LOA denotes local object appearance augmentation and GSC denotes global spatial correlation augmentation.

Model	LOA	GSC	Train Rate (s/it)	mAP$_{role}$
iCAN			0.183	44.7
	✓		0.191	46.8
		✓	0.190	47.2
	✓	✓	0.193	**48.0**

Local Object Appearance Augmentation Here we evaluate the two key components in LOA, *instance interlocking ratio (IIR)* and *object state matrix (OSM)*, by replacing them with other possible metrics. For IIR, we try other two possible choices: simply replacing all objects (replace all) and applying bbox IoU between neighbours as the metric (bbox IoU). For OSM, we also select other four alternatives: random selection, chamfer distance, instance mask distance and l_2 distance of the image inside resized bounding boxes. In Tab. 4(a), results show apparent degradation using other metrics, verifying the significance of our proposed metric.

Global Spatial Correlation Augmentation Global spatial correlation augmentation can greatly increase the data diversity without harming the context. We exhibit its value by comparing our results with the other two possible choices: random placement and appearance consistent metric *heatmap* in (Fang et al. 2019). Tab. 4(b) reveals that performance drops notably with the other alternatives, further proving the power of our pose-guided method.

Table 4: **Ablation Study of Object Appearance and Spatial Correlation Augmentation**

(a) **Local Object Appearance Augmentation Ablation Study**: Apply other alternative interchangeability metrics or object similarity metrics. IIR and OSM denote *instance interlocking ratio* and *object state matrix* respectively

Interchangable	Similarity	mAP$_{role}$
IIR	random	46.6
IIR	chamfer distance	47.2
IIR	mask distance	47.5
IIR	bbox l_2 distance	47.1
replace all	OSM	47.1
bbox IoU	OSM	47.5
IIR	OSM	**48.0**

(b) **Global Spatial Correlation Augmentation**: we compare three placement metrics: random, heatmap (Fang et al. 2019) and our pose-guided GMM.

Approach	mAP$_{role}$
random	43.6
heatmap	45.3
pose-guided GMM	**48.0**

5 Conclusion

In this paper, we propose a novel data augmentation method, **DecAug**, for HOI detection, which mainly includes two components: local object appearance augmentation and global spatial correlation augmentation. With negligible cost, our method can be easily combined with various existing models to further improve their performance, and it helps to address the long-tail problem. We hope our DecAug could give a new insight into the data augmentation of visual relationship detection.

Acknowledgement This work is supported in part by the National Key RD Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332, Shanghai Qi Zhi Institute, SHEITC(2018-RGZN-02046) and Baidu Fellowship.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57(5): 469–483.

Baldassano, C.; Beck, D. M.; and Fei-Fei, L. 2017. Human–object interactions are more than the sum of their parts. *Cerebral Cortex* 27(3): 2276–2288.

Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; and Chellappa, R. 2019. Detecting Human-Object Interactions via Functional Generalization.

Bertalmio, M.; Bertozzi, A. L.; and Sapiro, G. 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, I–I. IEEE.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 1017–1025.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6830–6840.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2019. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719* .

Dvornik, N.; Mairal, J.; and Schmid, C. 2018. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 364–380.

Fang, H.-S.; Cao, J.; Tai, Y.-W.; and Lu, C. 2018a. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 51–67.

Fang, H.-S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.-L.; and Lu, C. 2019. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE International Conference on Computer Vision*, 682–691.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2334–2343.

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018b. Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. In *AAAI Conference on Artificial Intelligence*.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437* .

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018a. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8359–8367.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018b. Detecting and Recognizing Human-Object Interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gupta, S.; and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* .

He, K.; Rhemann, C.; Rother, C.; Tang, X.; and Sun, J. 2011. A global sampling method for alpha matting. In *CVPR 2011*, 2049–2056. IEEE.

Hinterstoisser, S.; Pauly, O.; Heibel, H.; Marek, M.; and Bokeloh, M. 2019. An annotation saved is an annotation earned: Using fully synthetic training for object instance detection. *arXiv preprint arXiv:1902.09967* .

Hou, Z.; Peng, X.; Qiao, Y.; and Tao, D. 2020. Visual Compositional Learning for Human-Object Interaction Detection.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Khoreva, A.; Benenson, R.; Ilg, E.; Brox, T.; and Schiele, B. 2019. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision* 127(9): 1175–1197.

Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; and Cho, K. 2019. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296* .

Knill, D. C.; Kersten, D.; and Yuille, A. 1996. Introduction: A Bayesian formulation of visual perception. *Perception as Bayesian inference* 1: 1–21.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019a. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10863–10872.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, 1261–1270.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Chen, M.; Ma, Z.; Wang, S.; Fang, H.-S.; and Lu, C. 2019b. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539* .

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020. PaStaNet: Toward Human Activity Knowledge Engine.

Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; and Lu, C. 2019c. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3585–3594.

Liang, Z.; Guan, Y.; and Rojas, J. 2020. Visual-Semantic Graph Attention Network for Human-Object Interaction Detection. *arXiv preprint arXiv:2001.02302* .

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European conference on computer vision*, 852–869. Springer.

Pang, B.; Zha, K.; Zhang, Y.; and Lu, C. 2020. Further Understanding Videos through Adverbs: A New Video Task. In *AAAI*, 11823–11830.

Peng, X.; Tang, Z.; Yang, F.; Feris, R. S.; and Metaxas, D. 2018. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2226–2234.

Pinheiro, P. H. O.; Collobert, R.; and Dollár, P. 2015. Learning to Segment Object Candidates. *CoRR* abs/1506.06204.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018a. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 401–417.

Qi, X.; Chen, Q.; Jia, J.; and Koltun, V. 2018b. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8808–8816.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020. Intra- and Inter-Action Understanding via Temporal Action Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simard, P. Y.; Steinkraus, D.; Platt, J. C.; et al. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3.

Ulutan, O.; Iftekhar, A. S. M.; and Manjunath, B. S. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 9469–9478.

Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning Human-Object Interaction Detection Using Interaction Points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, B.; Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia* .

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5410–5419.

Yao, B.; and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 17–24. IEEE.

Yao, B.; and Fei-Fei, L. 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE transactions on pattern analysis and machine intelligence* 34(9): 1691–1703.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5532–5540.

Zhou, T.; Wang, W.; Qi, S.; Ling, H.; and Shen, J. 2020. Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4263–4272.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I.; and Hengel, A. v. d. 2017. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892* .