Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection

Xubin Zhong[1]   Xian Qu[1]   Changxing Ding[1,2*]   Dacheng Tao[3]

[1] South China University of Technology   [2] Pazhou Lab, Guangzhou   [3] The University of Sydney

{eexubin, eequxian.scut}@mail.scut.edu.cn, chxding@scut.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Modern human-object interaction (HOI) detection approaches can be divided into one-stage methods and two-stage ones. One-stage models are more efficient due to their straightforward architectures, but the two-stage models are still advantageous in accuracy. Existing one-stage models usually begin by detecting predefined interaction areas or points, and then attend to these areas only for interaction prediction; therefore, they lack reasoning steps that dynamically search for discriminative cues. In this paper, we propose a novel one-stage method, namely Glance and Gaze Network (GGNet), which adaptively models a set of action-aware points (ActPoints) via glance and gaze steps. The glance step quickly determines whether each pixel in the feature maps is an interaction point. The gaze step leverages feature maps produced by the glance step to adaptively infer ActPoints around each pixel in a progressive manner. Features of the refined ActPoints are aggregated for interaction prediction. Moreover, we design an action-aware approach that effectively matches each detected interaction with its associated human-object pair, along with a novel hard negative attentive loss to improve the optimization of GGNet. All the above operations are conducted simultaneously and efficiently for all pixels in the feature maps. Finally, GGNet outperforms state-of-the-art methods by significant margins on both V-COCO and HICO-DET benchmarks. Code of GGNet is available at `https://github.com/SherlockHolmes221/GGNet`.

1. Introduction

Human-Object Interaction (HOI) detection is one of the fundamental tasks in human-centric scene understanding. It involves not only detecting persons and objects in an image, but also the interactions (verbs) between each human-object pair. The output of an HOI detection model can be represented as a set of triplets in the form of $<human$

(a) InteractNet
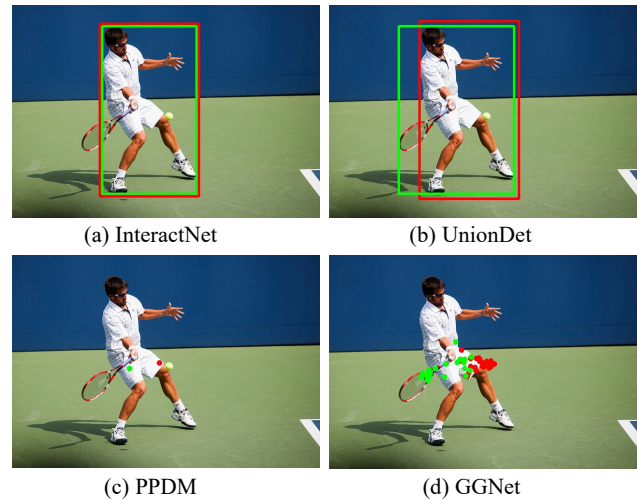
(b) UnionDet

(c) PPDM

(d) GGNet

Figure 1. Comparisons of interaction area definition. Green boxes or points represent the interaction area for "hold tennis_racket", while the red ones stand for "hit sports_ball". (a) InteractNet [3] uses the same human bounding box to represent the interaction area for all interactions pertaining to the person. (b) UnionDet [1] adopts the union box of one human-object pair to represent their interaction area. (c) PPDM [10] leverages the middle point of one human-object pair to represent their interaction area. (d) GGNet employs a single set of dynamic points to adaptively capture informative areas for the interaction between each human-object pair.

$interaction$ $object>$, and each triplet is also referred to as a single HOI category. For example, there are two HOI categories in Figure 1, i.e. $<human$ $hold$ $tennis\_racket>$ and $<human$ hit $sports\_ball>$.

According to the order in which object detection and interaction detection is performed, modern HOI detection methods can be divided into one-stage and two-stage approaches. Two-stage methods must perform the object detection first and then identify the interactions between each possible human-object pair. However, because the two stages are separated in this approach, these methods are usually inefficient. By contrast, one-stage methods can per-

---

*Corresponding author

form object detection and interaction detection in parallel by first defining interaction areas (e.g. a union box or a single point). Generally speaking, one-stage methods tend to be more efficient and structurally elegant, but two-stage methods are more accurate at present.

One of the key issues for one-stage methods is the way to represent the "interaction area" for each human-object pair [1]. Existing approaches usually define this area artificially and the interaction area often face the semantic ambiguity problem. For example, as shown in Figure 1(a), Interact-Net [3] utilizes a human bounding box to represent the area of all interactions involving the person, meaning that object-specific information is ignored. UnionDet [1] addresses this problem by utilizing the union box for each human-object pair as their interaction region. However, union boxes may overlap significantly with each other (Figure 1(b)), which introduces ambiguity between pairs. For its part, PPDM [10] utilizes the middle point of each human-object pair as the interaction point (Figure 1(c)). Although interaction points are less likely to overlap with each other under this approach, a single interaction point is often vague to represent complex interactions between a human-object pair.

With the predefined interaction areas discussed above, existing one-stage methods usually attend to the interaction area only once to predict the interaction categories. Recent works [24], [26] have revealed that the eyes of human beings usually move around an object to discover more cues regarding its location. Similarly, when it comes to HOI detection, people often first **glance** at the scene to identify possible human-object pairs with any interaction; they then search for cues around each pair, and finally **gaze** at discriminative areas to identify the interaction class.

Accordingly, inspired by the above observation, we herein propose a novel model, named Glance and Gaze Network (GGNet), which adaptively infers a set of action-aware points (ActPoints) to represent the interaction area (Figure 1(d)). GGNet mimics the two steps taken by humans to identify human-object interactions: Glance and Gaze. First, GGNet quickly determines whether each pixel in the feature maps is an interaction point; we call it the glance step. Based on the feature maps in the glance step, the subsequent gaze step searches for a set of ActPoints around each pixel. This step then progressively proceeds to refine the location of these ActPoints. In brief, this step comprises two sub-steps, in which the coarse location and location residuals of ActPoints are inferred, respectively. Finally, GGNet aggregates features of the refined ActPoints to predict interaction categories at the interaction points.

We further propose an action-aware point matching (AP-M) approach designed to match each interaction with its associated human-object pair. This matching process specifies the location of both the human and object instances for each interaction. Existing interaction point-based methods

tend to employ a single location regressor shared by all interaction categories [10], [20]; however, we observe that the interaction category affects the spatial layout of one human-object pair. We accordingly propose to assign each interaction category a unique location regressor, which is proven in the experimentation section to be a more effective approach.

Finally, we propose a novel focal loss, namely Hard Negative Attentive (HNA) loss, to further promote the performance of GGNet. As there are massive numbers of negative samples for each interaction classifier of the interaction point-based methods [10], [20], a serious imbalance problem exists between the positive and negative samples for each interaction category. We thus develop an efficient approach to address this problem by inferring and highlighting hard negative samples. Hard negatives are inferred between meaningful HOI categories containing the same object. For example, we can infer a hard negative sample "repair bicycle" according to the labeled positive sample "carry bicycle", unless "repair bicycle" is labeled as positive; in this way, the decision boundary between easily confused interaction categories can be clarified.

Both APM and HNA loss can be readily applied to other interaction point-based HOI detection methods. We conduct extensive experiments on the two most popular HOI detection databases, i.e. V-COCO [31] and HICO-DET [30]. Experimental results demonstrate that our proposed GGNet consistently outperforms start-of-the-art methods.

2. Related Works

Two-Stage Methods. Most existing HOI detection methods belong to the two-stage category. These methods typically perform object detection first, then pair the human and object proposals for interaction recognition. Various types of features are utilized to promote the detection accuracy: for example, human-object spatial feature encoded using their bounding box locations are widely adopted in two-stage methods [5], [6], [7], [11], [34], [13], [12]. In line with this, Liu *et al.* [8] introduced a mechanism that encodes the fine-grained spatial layout of human-object pairs. The spatial layout is obtained using human parsing and object segmentation tools. Wan *et al.* [7] and Zhou *et al.* [16] utilized key points on the human body to crop human part features. Different model structures have also been developed to promote HOI detection performance. For example, several recent works have utilized graph convolutional networks to integrate human and object appearance features [11], [19], [21], [9]. Moreover, Gao *et al.* [12] and Wang *et al.* [13] made use of attention mechanisms to capture context information. Other works have also investigated the semantic meaning of verbs and the relationships between them: for example, Zhong *et al.* [14, 38] promoted HOI detection accuracy by overcoming the so-called verb polysemy problem, while Kim *et al.* [15] modeled the relation-
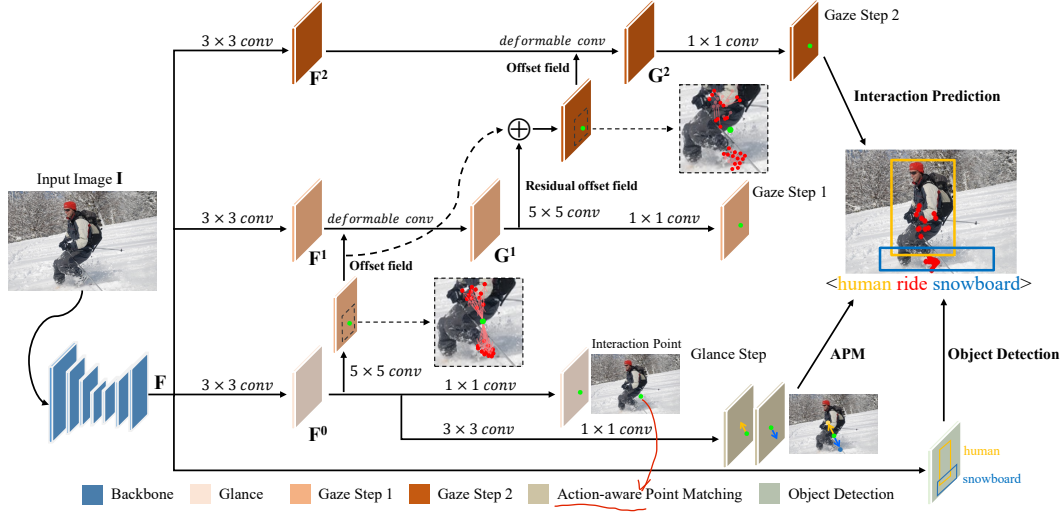
Figure 2. Overview of GGNet in the training stage. GGNet includes three main tasks, namely interaction prediction, human-object pair matching, and object detection. The three tasks share the same backbone model. The interaction prediction task includes one Glance step and two Gaze steps. The two Gaze steps infer a set of ActPoints for each pixel in feature maps. By aggregating their features into the interaction point, the second Gaze step is able to predict interactions more robustly. The human-object pair matching task is realized by the Action-aware Point Matching (APM) module, which bridges the interaction prediction and object detection tasks. In the testing stage, the glance step and the first gaze step are only utilized to infer ActPoints, while the other layers of the two steps are removed. $\oplus$ denotes the element-wise addition operation. Best viewed in color.

ship between interaction categories by means of action co-occurrence matrices, which were then utilized to promote HOI detection performance. Although two-stage methods are flexible to include diverse features, they divide object detection and interaction prediction into two sequential steps, which is typically very time-consuming.

One-Stage Methods. Some early HOI detection works [2], [3] devised end-to-end models based on the Faster R-CNN object detector [4]. Although these methods are more efficient than most two-stage methods, they adopt the same human appearance features for predicting different interactions, meaning that object-specific information in different human-object pairs is ignored. Recent works, moreover, have proposed to perform object detection and interaction prediction in parallel branches. For example, Liao et al. [10] and Wang et al. [20] defined the middle point of each human-object pair as their interaction point, which is capable of roughly capturing both human and object appearance features. These authors then detect interaction points in the interaction prediction branch. Kim *et al*. [1] represented the interaction location of one human-object pair using the union box, then detect union boxes using networks that are similar to anchor-based object detection models [4], [25], [32]. These recent one-stage methods can substantially improve HOI detection efficiency; however, as illustrated in Figure 1, their definitions of the interaction areas or points remain relatively coarse and may thus introduce ambiguity into HOI detection.

In this paper, to handle the semantic ambiguity problem associated with interaction areas, we propose to represent these area using a set of dynamic ActPoints. These ActPoints are adaptively inferred around each interaction point, after which their features are aggregated to improve the interaction recognition accuracy.

3. Methods

3.1. Overview

The architecture of GGNet is illustrated in Figure 2. Similar to PPDM [10], GGNet breaks HOI detection into three main tasks, namely interaction prediction, human-object pair matching, and object detection. These three tasks all share the same backbone. The second task bridges the first and the third tasks via associating each detected interaction with a single human-object pair. The object detection task is realized according to [10]. GGNet improves the interaction prediction task with a novel glance-and-gaze strategy. It also promotes the accuracy of the second task with an Action-aware Point Matching (APM) module.

3.2. Glance and Gaze Network

As illustrated in Figure 2, given an input image $\mathbf{I} \in R^{H \times W \times 3}$, the output feature maps of the backbone can be expressed as $\mathbf{F} \in R^{\frac{H}{d} \times \frac{W}{d} \times C}$, where d denotes the output stride of backbone and C denotes the number of channels.

In this paper, we adopt the same definition of interac-

tion point as [10], [20], which is also illustrated in Figure 1(c). GGNet handles the semantic ambiguity problem of the single interaction point by further inferring a set of action-aware points (ActPoints). ActPoints adaptively capture more contextual information, and their features are aggregated to predict the interaction category. The set of ActPoints around one interaction point can be represented as:

$$\mathcal{P} = \{(x_k, y_k)\}_{k=1}^{n}, \tag{1}$$

where n is the total number of ActPoints sampled for an interaction. We empirically set n as 25. The location of ActPoints are inferred progressively by mimicking human's visual system [24] with one glance step and two gaze steps.

Glance Step. In this step, GGNet quickly determines whether each pixel in $\mathbf{F}$ is an interaction point. As shown in Figure 2, this step is realized using a 3×3 convolutional (Conv) layer with ReLU, followed by a 1×1 Conv layer and a sigmoid layer. The size of heatmaps produced by the sigmoid layer is $\frac{H}{d} \times \frac{W}{d} \times V$, where V denotes the number of interaction categories. We apply a V-dimensional element-wise focal loss [10] to the heatmaps as supervision for the inference of interaction categories. Due to this supervision, the feature maps output by the 3×3 Conv layer, i.e. $\mathbf{F}^0$ in Figure 2, are action-aware.

Gaze Step. This step infers the location of ActPoints via a two sub-steps, which are referred to as Gaze Step 1 and Gaze Step 2, respectively. In Gaze Step 1, $\mathbf{F}^0$ is employed to predict the coarse location of n ActPoints for each pixel, since features in $\mathbf{F}^0$ have been action-aware. Moreover, as the discriminative power of each ActPoint varies with respect to the target interaction, GGNet also predicts a weight for each ActPoint. Both location and weight prediction are achieved by a 5×5 Conv layer. Next, we aggregate features of ActPoints as well as their weights using one deformable Conv layer [27]. The 5×5 offset field is determined by the number of ActPoints, which are also verified in Table 2. To ensure that the predicted ActPoints are reasonable, the feature maps produced by this sub-step are also used for interaction prediction with the V-dimensional element-wise focal loss as supervision.

On its own, the above step cannot always obtain precise locations of ActPoints; this is because the above 5×5 Conv operation has a fixed field of view, while the location of the human and object instances in one pair can vary dramatically. To address this problem, Gaze Step 2 is introduced to refine the location of ActPoints. In more detail, we aggregate the features in $\mathbf{F}^1$ of the coarse ActPoints using one deformable Conv layer, of which the output feature maps are denoted as $\mathbf{G}^1$. Now, each pixel in $\mathbf{G}^1$ has a larger field of view; $\mathbf{G}^1$ is then send to another 5×5 Conv layer to predict the residual offsets of ActPoints locations, along with their new weights. As shown in Figure 2, the final position of ActPoints are obtained by summing their coarse location
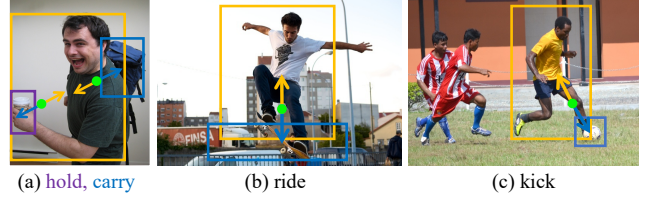

(a) hold, carry   (b) ride   (c) kick

Figure 3. Human-object pairs with different interaction categories present different spatial characteristics, i.e. the relative location between one human-object pair. Green points represent interaction points, each yellow arrow specifies the human instance in the pair, while each blue arrow indicates the object instance in the pair.

and residual offsets. Finally, we aggregate the features of the refined ActPoints as well as their weights using another deformable Conv layer to predict interaction categories for each pixel in $\mathbf{F}^2$. Similar to Gaze Step 1, the V-dimensional element-wise focal loss is adopted as supervision.

More formally, $(x_k^{(t)}, y_k^{(t)})$ is updated as follows:

$$(x_k^{(t)}, y_k^{(t)}) = (x_k^{(t-1)}, y_k^{(t-1)}) + (\Delta x_k^{(t)}, \Delta y_k^{(t)}), \tag{2}$$

$$(\Delta x_k^{(t)}, \Delta y_k^{(t)}) = T_{offset}^{(t)}(\mathbf{G}^{t-1}), \tag{3}$$

where t stands for the t-th gaze step, while $(\Delta x_k^{(t)}, \Delta y_k^{(t)})$ denotes the predicted offset with respect to the k-th ActPoint's location in the last step. Moreover, we set the initial location of all n ActPoints as $(0,0)$ and set $\mathbf{G}^0$ as $\mathbf{F}^0$. $T_{offset}^{(t)}$ is a Conv layer, whose kernel size is equal to the square root of n. More details of the glance and gaze network can be confirmed from our open-source project.

Action-aware Point Matching. To compose one HOI instance, each detected interaction point is associated with one human-object pair. Existing works [10], [20] adopt a single regressor shared by all interaction categories for this association process. However, as shown in Figure 3, human-object pairs with different interactions also differ in terms of spatial characteristics. Therefore, we propose the action-aware point matching (APM) module that assigns a unique location regressor for each interaction category.

As illustrated in Figure 2, APM is attached to $\mathbf{F}^0$. It includes one 3×3 and one 1×1 Conv layers. The latter layer acts as regressors. Each regressor outputs a 2-dimensional offset to human point (object point) with respect to the interaction point. Therefore, the output dimension of this layer is $\frac{H}{d} \times \frac{W}{d} \times 4V$. In line with [10], we utilize these predicted offsets to match the target human (object) proposal during inference. This process will be detailed in Section 3.4.

3.3. Hard Negative Attentive Loss

Recent one-stage methods [10], [20] adopt the element-wise focal loss [35], [33] on the output heatmaps as super-

vision to train their models. However, the heatmap size for each interaction category is $\frac{H}{d} \times \frac{W}{d}$, which is often a large number. Therefore, there are massive negative samples in each heatmap, which brings in the problem of imbalance between positive and negative samples. Moreover, because of the long-tailed distribution of interaction categories [30], [31], some interaction categories have very limited positive samples, which further exacerbates the imbalance problem. Based on the above observation, we propose a novel hard negative attentive (HNA) loss to guide the model to focus more on hard negative samples for each respective interaction category.

Definition of Hard Negatives. We infer hard negatives between meaningful HOI categories that share the same object class. For example, we can infer a hard negative sample $<human\ repair\ bicycle>$ for the "repair" category according to a labeled positive sample $<human\ carry\ bicycle>$, if $<human\ repair\ bicycle>$ is indeed not labeled as positive. We do not infer hard negatives from meaningless HOI categories that have never appeared in the training set of each database, e.g. $<human\ eat\ bicycle>$ and $<human\ drink\ bicycle>$. The inferred sample of $<human\ repair\ bicycle>$ can be highlighted as a hard negative in the interaction heatmap for "repair" category.

Loss Formulation. First, we introduce the Gaussian heatmap masks $\mathbf{M} \in [-1,1]^{\frac{H}{d} \times \frac{W}{d} \times V}$, which are used to mark both positive samples and hard negative samples. For a ground truth HOI sample (v_i, o_i) with an interaction point located at (x_i, y_i), $M_{x_i y_i v_i}$ is set to 1 and it is also used as the center of a Gaussian distribution in the v_i-th channel of $\mathbf{M}$. The value of the elements in this distribution is within $[0, 1]$. The interaction category and object category for this HOI sample are denoted as v_i and o_i, respectively.

Second, we infer a set of HOI samples $\{(v_j, o_i)\}$ as hard negatives for the v_j-th interaction category, with the help of the labeled positive sample (v_i, o_i). If (v_j, o_i) is not labeled as a positive sample, then we set $M_{x_i y_i v_j}$ as -1 and it is also used as the center of another Gaussian distribution in the v_j-th channel of $\mathbf{M}$. The value of the elements in this distribution is within $[-1, 0]$.

We repeat the above two operations for each ground truth sample. The value of remaining elements in $\mathbf{M}$ is set to 0. Finally, the HNA loss can be represented as follows:

$$
\mathcal{L} = -\frac{1}{N} \sum_{xyv}
\begin{cases}
(1 - P_{xyv})^\alpha \log(P_{xyv}) & \text{if } M_{xyv} = 1, \\
(1 - M_{xyv})^\beta (P_{xyv})^\alpha \\
\quad \log(1 - P_{xyv}) & \text{if } M_{xyv} = -1, \\
(1 - M_{xyv})^\gamma (P_{xyv})^\alpha \\
\quad \log(1 - P_{xyv}) & \text{otherwise,}
\end{cases}
\tag{4}
$$

where N is the number of ground truth interaction points in an image. P_{xyv} is the predicted score for the interaction category v at point (x, y). β is set as 7. α, γ, and the pa-

rameters of the Gaussian distribution are set following [10].

3.4. Training and Inference

Training. The overall loss function for GGNet can be represented as follows:

$$
\mathcal{L}_{hoi} = \mathcal{L}_{gaze2} + \lambda_1(\mathcal{L}_{glance} + \mathcal{L}_{gaze1} + \mathcal{L}_m) + \mathcal{L}_d, \tag{5}
$$

where

$$
\mathcal{L}_m = \mathcal{L}_{mh} + \mathcal{L}_{mo}, \tag{6}
$$

$$
\mathcal{L}_d = \mathcal{L}_h + \mathcal{L}_o + \lambda_2 \mathcal{L}_{wh} + \mathcal{L}_{off}. \tag{7}
$$

λ_2 is set to 0.1 following [10]; λ_1 is set to 0.1, which is analyzed in the supplementary file. $\mathcal{L}_{glance}$, $\mathcal{L}_{gaze1}$ and $\mathcal{L}_{gaze2}$ denote the HNA loss for the glance step, and each of two gaze steps, respectively. $\mathcal{L}_{mh}$ and $\mathcal{L}_{mo}$ stand for the L1 loss for matching human and object points in the APM module, respectively. $\mathcal{L}_d$ stands for the object detection loss. $\mathcal{L}_h$ and $\mathcal{L}_o$ are the focal loss functions to predict human and object locations. $\mathcal{L}_{wh}$ and $\mathcal{L}_{off}$ denote the L1 loss for object size and center offset predictions, respectively. $\mathcal{L}_m$ and $\mathcal{L}_d$ are realized in the same way as [10].

Inference. We use the output of Gaze Step 2 to obtain interaction points; while the final score is the multiplication result of interaction prediction score from Gaze step 2 and object detection scores of the associated human-object pair. The following point matching process is the same as that in [10]. First, a set of interaction $\hat{\mathcal{S}}^i$, human $\hat{\mathcal{S}}^h$ and object points $\hat{\mathcal{S}}^o$ are respectively selected based on their prediction confidence scores. The number of points in $\hat{\mathcal{S}}^i$, $\hat{\mathcal{S}}^h$, and $\hat{\mathcal{S}}^o$ is set to 100, respectively. Second, we associate each predicted interaction point $(\hat{x}^i, \hat{y}^i) \in \hat{\mathcal{S}}^i$ with a human point $(\hat{x}^h_{opt}, \hat{y}^h_{opt}) \in \hat{\mathcal{S}}^h$ and an object point $(\hat{x}^o_{opt}, \hat{y}^o_{opt}) \in \hat{\mathcal{S}}^o$ according to the predicted subject and object offsets by APM, respectively:

$$
(\hat{x}^h_{opt}, \hat{y}^h_{opt}) = \underset{(\hat{x}^h, \hat{y}^h) \in \hat{\mathcal{S}}^h}{\arg\min} \frac{1}{P^h_{(\hat{x}^h, \hat{y}^h)}}
$$
$$
(|(\hat{x}^i, \hat{y}^i) - (\hat{d}^{hx}_{(\hat{x}^i, \hat{y}^i)}, \hat{d}^{hy}_{(\hat{x}^i, \hat{y}^i)}) - (\hat{x}^h, \hat{y}^h)|),
\tag{8}
$$

where $P^h_{(\hat{x}^h, \hat{y}^h)}$ denotes the object detection score for human point $(\hat{x}^h, \hat{y}^h)$; $(\hat{d}^{hx}_{(\hat{x}^i, \hat{y}^i)}, \hat{d}^{hy}_{(\hat{x}^i, \hat{y}^i)})$ denotes the predicted offset from the interaction point to the human point by APM. The optimal object point $(\hat{x}^o_{opt}, \hat{y}^o_{opt})$ can be inferred similarly.

4. Experimental Setup

4.1. Datasets and Evaluation Metrics

V-COCO. V-COCO was constructed based on the MS-COCO database [29]. Its training and validation sets contain 5,400 images, while its testing set includes 4,946 images. It covers 80 object categories, 26 interaction categories and 234 HOI categories. The mean average precision of Scenario 1 role (mAP_{role}) [31] is used for evaluation.

Table 1. Ablation studies on each component of GGNet. DLA-34 is adopted as backbone for experiments on HICO-DET.

Methods	Components				mAP	
	HNA Loss	Gaze # 1	Gaze # 2	APM	V-COCO	HICO-DET (DT)
Our Baseline	-	-	-	-	51.06	20.16
Incremental	✓	-	-	-	52.51	20.58
	✓	✓	-	-	54.03	21.05
	✓	✓	✓	-	54.43	21.54
Drop-one-out	-	✓	✓	✓	53.55	21.03
	✓	-	-	✓	52.60	21.19
	✓	✓	-	✓	54.21	21.65
	✓	✓	✓	-	54.43	21.54
GGNet	✓	✓	✓	✓	**54.72**	**22.03**

Table 2. Comparisons with variants of the Gaze Step on V-COCO.

# ActPoints	# Gaze Step	Sharing	AP_{role}
9	1	-	53.54
25	1	-	54.32
49	1	-	53.81
25	1	-	54.21
25	2	-	**54.72**
25	3	-	54.34
25	2	✓	54.43

HICO-DET. HICO-DET [30] is a large-scale HOI detection dataset with more than 150,000 annotated instances. It contains 38,118 and 9,658 images for training and testing, respectively. There are a total of 80 object categories, 117 verb categories, and 600 HOI categories. Those HOI categories with fewer than 10 training samples are referred to as "rare" categories and the remaining ones are called as "non-rare" categories; specifically, there are 138 rare and 462 non-rare categories in total. There are two modes of mAP on HICO-DET, namely the Default (DT) mode and the Known-Object (KO) mode. In DT mode, each HOI category is evaluated on all testing images; while in KO mode, one HOI is only evaluated on images that contain its associated object category.

4.2. Implementation Details

We use the Hourglass-104 model [28], pre-trained on the MS-COCO [29], as the backbone of GGNet. Moreover, we adopt a lightweight network, named DLA-34 [36], as the backbone to perform ablation studies on HICO-DET for shortening experimental cycle. GGNet is trained using the Adam optimizer with an initial learning rate of 1.5e-5 (1.5e-4) and batch size of 7 (23) on V-COCO (HICO-DET) for 120 epochs. For all the experiments, the learning rate is reduced by multiplying 0.1 at the 90th epoch. Resolution of input images is 512×512 and the output stride d of backbone is set to 4.

5. Experimental Results and Discussion

5.1. Ablation Studies

We perform ablation studies on both V-COCO and HICO-DET datasets to demonstrate the effectiveness of each component of GGNet. Our baseline is constructed by removing the gaze steps from GGNet, and replacing the H-NA loss and APM module with their counterparts in PPDM [10]. Experimental results are tabulated in Table 1.

Effectiveness of the HNA Loss. As analyzed in Section 3.3, a serious problem of imbalance exists between positive and negative samples for each interaction category. We therefore propose the HNA loss to handle this problem by highlighting the hard negative samples. The adoption of H-NA loss promotes the performance of our model by 1.45%

(0.42%) mAP on V-COCO (HICO-DET).

We further evaluate the optimal value of the hyper-parameters β in in the HNA loss. The experimental results are provided in the supplementary material.

Effectiveness of the Glance-Gaze Strategy. The glance-and-gaze strategy is developed to infer ActPoints to represent the interaction area for one human-object pair. The Gaze Step 1 is found to promote the HOI detection performance by 1.52% and 0.47% in terms of mAP on V-COCO and HICO-DET. Moreover, when the ActPoints is refined by Gaze Step 2, the performance is further improved by 0.40% and 0.49% mAP on V-COCO and HICO-DET, respectively.

Effectiveness of APM. As the interaction category affects the human-object spatial layout in an HOI instance, we propose the APM module that assigns each interaction category a unique location regressor to facilitate matching of both the human and object points. The adoption of APM promotes the performance of our model by 0.29% and 0.49% in terms of mAP on V-COCO and HICO-DET, respectively.

Drop-one-out Study. We further perform a drop-one-out study in which each proposed component is removed individually. In particular, as Gaze Step 2 is built based on Gaze Step 1, we remove both Gaze Steps 1 and 2 in the experiment where "Gaze # 1" is dropped out. These experimental results further demonstrate that each proposed component is indeed effective at promoting HOI detection performance.

5.2. Comparisons with Variants of GGNet

Comparisons with Variants of the Gaze Step. We compare the performance of gaze step with some possible variants by changing the number of ActPoints, the number of gaze steps, and whether layers are shared in different gaze steps. Experimental results are summarized in Table 2.

First, we set the number of gaze steps to 1 and change the number of ActPoints. As the results show, our model achieves the best performance when the number of Act-Points is 25; this may be because a small number of Act-Points is insufficient to cover the entire interaction area, while too many ActPoints will increase the complexity of searching their locations.

Second, we compare the performance of different numbers of gaze steps. The number of ActPoints is set to 25 here. When the number of gaze steps increases from 1 to 2, our model is promoted by 0.51% in terms of mAP; no-

Table 3. Comparisons with variants of feature aggregation on V-COCO. "I", "H", and "O" denote the features of interaction point, human center point, and object center point, respectively.

Methods	AP_{role}
I (our baseline)	51.06
I + H	51.84
I + O	51.71
I + H + O	52.32
ActPoints w/o glance and gaze	51.66
ActPoints + glance and gaze	**53.25**

Table 4. Performance comparisons on V-COCO. ° denotes methods that are reproduced using their open-source codes. 'A', 'P', 'S', and 'L' represent the appearance feature, human pose feature, spatial feature, and language feature, respectively.

	Methods	Feature	AP_{role}
Two-Stage	iCAN [12]	A+S	45.3
	RPNN [16]	A+S+P	47.5
	TIN (RP$_D$C$_D$) [6]	A+S+P	47.8
	VCL [17]	A+S	48.3
	C-HOI [37]	A+S	48.3
	DRG [19]	A+S+L	51.0
	VSGNet [11]	A+S	51.7
	PMFNet [7]	A+S+P	52.0
	PD-Net [14]	A+S+P+L	52.6
	ACP [15]	A+S+P+L	52.9
	FCMNet [8]	A+S+P+L	53.1
	ConsNet [18]	A+S+L	53.2
One-Stage	InteractNet [3]	A	40.0
	UnionDet [1]	A	47.5
	IP-Net [20]	A	51.0
	PPDM-Hourglass° [10]	A	51.1
	GGNet-Hourglass	A	**54.7**

tably, the performance is not further promoted through the addition of more gaze steps, which may be because this increases the difficulty of model optimization.

Third, we try sharing the parameters of the two 3×3 Conv layers that generate $\mathbf{F}^1$ and $\mathbf{F}^2$ in Figure 2. As Table 2 shows, the performance of our model decreases by 0.29% in terms of mAP. One reason for this is that Gaze Step 2 captures more fine-grained features for interaction prediction; therefore, it is better for the two gaze steps to adopt independent Conv layers.

Comparisons with Variants of Feature Aggregation In two-stage methods, it is a common practice to aggregate features of the human instance, the object instance, and their union box for interaction prediction purposes [7], [17], [14]. Accordingly, in this experiment, we compare the performance of the above feature aggregation approach with our proposed glance-and-gaze strategy. The experimental results are tabulated in Table 3. All methods in this table are constructed based on our baseline model. Here, "I + H" ("I + O") means that we concatenate the features of each pixel in $\mathbf{F}^0$ (in Figure 2) with those of one human (object) center point. "I + H + O" denotes that both features of the human and object center points are aggregated. The human and object center points are obtained via the point matching strategy in [10]. Moreover, "ActPoints w/o glance and gaze" means that we adopt one deformable Conv layer to aggregate features of the predicted coarse ActPoints on $\mathbf{F}^0$ for interaction prediction. The structure of these models is outlined in more detail in the supplementary material.

From the above, we can make the following observations. First, performance is promoted when either human or object features are aggregated. However, one person may interact with different objects, while a single object may also be interacted by multiple persons. Therefore, features for the center points of human and object instances may lack specific information for each human-object pair. By contrast, our proposed ActPoints adaptively capture features in the discriminative area for each human-object pair; therefore, it outperforms "I + H + O" by 0.93% mAP. Second, when the glance and gaze steps are omitted, the performance of ActPoints drops by 1.59% mAP. These comparisons demonstrate the effectiveness of our proposed glance-and-gaze strategy.

5.3. Comparisons with State-of-the-Art Methods

We compare the performance of GGNet with state-of-the-art methods on both V-COCO and HICO-DET.

As shown in Table 4, GGNet outperforms all state-of-the-art methods by significant margins on V-COCO. In particular, with the same backbone model, GGNet outperforms one of the most recent one-stage methods, i.e. PPDM [10] by a large margin of 3.6% in terms of mAP. Moreover, although recent two-stage methods adopt various types of features to promote HOI detection performance, GGNet still outperforms all of these by at least 1.5% in terms of mAP while utilizing appearance features only.

As shown in Table 5, on HICO-DET, GGNet still outperforms state-of-the-art methods by clear margins in different settings of "Backbone Sharing". In particular, GGNet outperforms PPDM [10], by 1.74% (2.78%), 2.70% (3.58%), and 1.50% (2.64%) mAP in DT (KO) mode for the full, rare and non-rare HOI categories, respectively. Moreover, GGNet outperforms one of the best two-stage methods, i.e. PD-Net [38], by 1.10% mAP in DT mode. Again, it is worth noting that PD-Net utilizes four types of features while G-GNet employs the appearance feature only.

As Table 5 shows, several recent two-stage methods have adopted two separate backbones for object detection and interaction prediction. In particular, two approaches have fine-tuned both backbones on the HICO-DET databases, thereby achieving superior performance [19], [17] (marked with ° in Table 5). For its part, however, GGNet has a single shared backbone for object detection and interaction prediction. To facilitate fair comparison, we further test the performance of GGNet with the same object detection results provided by the object detector in DRG [19]; this setting is denoted as GGNet[‡]. As shown in Table 5, GGNet[‡] outperforms DRG[°‡] [19] by a large margin of 4.64% (5.52%) mAP in DT (KO) mode.

Table 5. Performance comparisons on HICO-DET. "Backbone Sharing" represents methods that use the same feature backbone for object detection and interaction detection. $\diamond$ denotes two-stage methods that first pre-train their object detectors on COCO, and then further fine-tune the object detectors on HICO-DET. $\ddagger$ denotes methods that use the same object detection results during inference.

	Methods	Backbone Sharing	Feature	DT Mode			KO Mode		
				Full	Rare	Non-Rare	Full	Rare	Non-Rare
Two-Stage	No-Frills [5]	✓	A+S+P	17.18	12.17	18.68	-	-	-
	DRG [19]	-	A+S+L	19.26	17.74	19.71	23.40	21.75	23.89
	Peyre et al. [23]	✓	A+S+L	19.40	14.60	20.90	-	-	-
	VCL [17]	-	A+S	19.43	16.55	20.29	22.00	19.09	22.87
	FCMNet [8]	-	A+S+P+L	20.41	17.34	21.56	22.04	18.97	23.12
	ACP [15]	✓	A+S+P+L	20.59	15.92	21.98	-	-	-
	PD-Net [14]	✓	A+S+P+L	20.81	15.90	22.28	24.78	18.88	26.54
	DJ-RN [22]	-	A+S+P+L	21.34	18.53	22.18	23.69	20.64	24.60
	ConsNet [18]	✓	A+S+L	22.15	17.12	23.65	-	-	-
	PD-Net [38]	✓	A+S+P+L	22.37	17.61	23.79	26.86	21.70	28.44
	VCL$^\diamond$[17]	-	A+S	23.63	17.21	25.55	25.98	19.12	28.03
	DRG$^{\diamond\ddagger}$ [19]	-	A+S+L	24.53	19.47	26.04	27.98	23.11	29.43
One-Stage	Shen et al. [2]	✓	A+P	6.46	4.24	7.12	-	-	-
	InteractNet [3]	✓	A	9.94	7.16	10.77	-	-	-
	UnionDet [1]	✓	A	17.58	11.72	19.33	19.76	14.68	21.27
	Our Baseline-Hourglass	✓	A	21.43	13.46	23.81	24.29	16.40	26.65
	PPDM-Hourglass [10]	✓	A	21.73	13.78	24.10	24.58	16.65	26.84
	GGNet-Hourglass	✓	A	**23.47**	**16.48**	**25.60**	**27.36**	**20.23**	**29.48**
	PPDM-Hourglass‡ [10]	-	A	26.50	19.35	28.63	29.24	22.18	31.34
	GGNet-Hourglass‡	-	A	**29.17**	**22.13**	**30.84**	**33.50**	**26.67**	**34.89**

The above comparisons also show that the adoption of a separate object detector can significantly promote GGNet's performance. This may be because the object detection and interaction prediction tasks conflict with each other, as they require different features. It is also worth noting that adopting two backbones does not alter the one-stage nature of GGNet, as the object detection and interaction prediction tasks still run in parallel, regardless of whether or not they adopt the same backbone. The above experiments further justify the effectiveness of GGNet.

5.4. Qualitative Visualization Results

Figure 4 visualizes the interaction points and ActPoints predicted by GGNet. Here, three images are randomly chosen from V-COCO. We can observe that the interaction points, i.e. the green points in first column, are often located in the background area; therefore, their own features are vague to represent the interaction category. In comparison, ActPoints capture cues from discriminative object and human parts, both of which are important for interaction prediction. Moreover, the ActPoints refined by Gaze Step 2 are usually located at more important object and human parts than the ActPoints sampled by Gaze Step 1. In the supplementary material, we also present qualitative comparisons between GGNet and PPDM in terms of HOI detection results.

6. Conclusion

Existing one-stage HOI detection methods typically utilize the features of predefined interaction areas for interaction prediction, while these artificially defined areas are usually vague to represent interactions. In this paper, we propose a novel one-stage network, namely GGNet, which

(a) throw frisbee
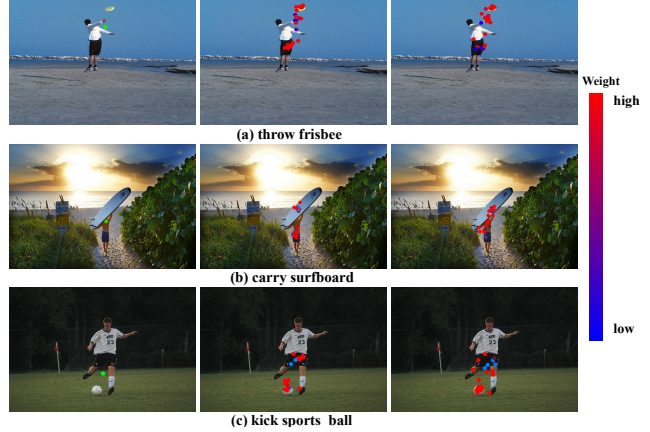
(b) carry surfboard

(c) kick sports ball

Figure 4. Visualization of results in glance and gaze steps. The first column shows the detected interaction point in the Glance Step; the second and third columns visualize the adaptively sampled ActPoints by Gaze Step 1 and Step 2, respectively. The color of each ActPoint reflects its weight, i.e. discriminative power.

adaptively samples a set of action-aware points (ActPoints) via glance and gaze steps. We also propose an action-aware point matching approach that robustly matches target human and objects for each detected interaction. Moreover, a novel hard negative attentive loss is devised to improve the optimization of GGNet. Extensive experiments results show that GGNet outperforms start-of-the-arts on both V-COCO and HICO-DET datasets.

References

[1] B. Kim, T. Choi, J. Kang1y, H. J. Kim. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. In *ECCV*, 2020.

[2] L. Shen, S. Yeung, J. Hoffman, G. Mori, and F. Li. Scaling Human-Object Interaction Recognition Through Zero-Shot Learning. In *WACV*, 2018.

[3] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[5] T. Gupta, A. Schwing, and D. Hoiem. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *ICCV*, 2019.

[6] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *CVPR*, 2019.

[7] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *ICCV*, 2019.

[8] Y. Liu, Q. Chen, and A. Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020.

[9] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[10] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

[11] O. Ulutan, A. Iftekhar, and B. Manjunath. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In *CVPR*, 2020.

[12] C. Gao, Y. Zou, and J. Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[13] T. Wang, R. Anwer, M. Khan, F. Khan, Y. Pang, L. Shao, and J. Laaksonen. Deep Contextual Attention for Human-Object Interaction Detection. In *ICCV*, 2019.

[14] X. Zhong, C. Ding, X. Qu, and D. Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020.

[15] D. Kim, X. Sun, J. Choi, S. Lin, and I. Kweon. Detecting Human-Object Interactions with Action Co-occurrence Priors. In *ECCV*, 2020.

[16] P. Zhou, and M. Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.

[17] Z. Hou, X. Peng, Y. Qiao, and D. Tao. Visual Compositional Learning for Human-Object Interaction Detection. In *ECCV*, 2020.

[18] Y. Liu, J. Yuan, and C. Chen. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In *ACM MM*, 2020.

[19] C. Gao, J. Xu, Y. Zou, and J. Huang. DRG: Dual Relation Graph for Human-Object Interaction Detection. In *ECCV*, 2020.

[20] T. Wang, T. Yang, M. Danelljan, F. Khan, X. Zhang, and J. Sun. Learning Human-Object Interaction Detection using Interaction Points. In *CVPR*, 2020.

[21] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. Kankanhalli. Learning to Detect Human-Object Interactions With Knowledge. In *CVPR*, 2019.

[22] Y. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu. Detailed 2D-3D Joint Representation for Human-Object Interaction. In *CVPR*, 2020.

[23] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019.

[24] H. Deubel, W. Schneider and others. Saccade target selection and object recognition: Evidence for a common attentional mechanism. In *Vision research*, 1996.

[25] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[26] S. Lan, Z. Ren, Y. Wu, L. Davis, and G. Hua. SaccadeNet: A Fast and Accurate Object Detector. In *CVPR*, 2020.

[27] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.

[28] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[29] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[30] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[31] S. Gupta, and J. Malik. Visual semantic role labeling. In *arXiv preprint arXiv:1505.04474*, 2015.

[32] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[33] H. Law, and J. Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018.

[34] X. Lin, C. Ding, J. Zeng, and D. Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020.

[35] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[36] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. In *CVPR*, 2018.

[37] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020.

[38] X. Zhong, C. Ding, X. Qu, and D. Tao. Polysemy Deciphering Network for Robust Human-Object Interaction Detection. In *IJCV*, 2021.