Distance Matters in Human-Object Interaction Detection

Guangzhi Wang
guangzhi.wang@u.nus.edu
Institute of Data Science, National University of Singapore

Yangyang Guo∗
guoyang.eric@gmail.com
School of Computing, National University of Singapore

Yongkang Wong
yongkang.wong@nus.edu.sg
School of Computing, National University of Singapore

Mohan Kankanhalli
mohan@comp.nus.edu.sg
School of Computing, National University of Singapore

ABSTRACT

Human-Object Interaction (HOI) detection has received considerable attention in the context of scene understanding. Despite the growing progress, we realize existing methods often perform unsatisfactorily on distant interactions, where the leading causes are two-fold: 1) Distant interactions are by nature more difficult to recognize than close ones. A natural scene often involves multiple humans and objects with intricate spatial relations, making the interaction recognition for distant human-object largely affected by complex visual context. 2) Insufficient number of distant interactions in datasets results in under-fitting on these instances. To address these problems, we propose a novel two-stage method for better handling distant interactions in HOI detection. One essential component in our method is a novel Far Near Distance Attention module. It enables information propagation between humans and objects, whereby the spatial distance is skillfully taken into consideration. Besides, we devise a novel Distance-Aware loss function which leads the model to focus more on distant yet rare interactions. We conduct extensive experiments on HICO-DET and V-COCO datasets. The results show that the proposed method surpass existing methods significantly, leading to new state-of-the-art results.

CCS CONCEPTS

• Computing methodologies → Scene understanding.

KEYWORDS

Human-Object Interaction Detection, Scene Understanding

ACM Reference Format:
Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. 2022. Distance Matters in Human-Object Interaction Detection. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), Oct. 10–14, 2022, Lisbon, Portugal. ACM, New York, NY, USA, 9 pages.
https://doi.org/10.1145/3503161.3547793

1 INTRODUCTION

Given a natural image, the task of Human-Object Interaction (HOI) detection is to localize all humans and objects, and recognize the interaction between each human-object pair. It constitutes an important step towards high-level scene understanding, and has benefited many multimedia applications, including visual question answering [2, 13], image captioning [45] and surveillance event detection [1].

This task is by its nature challenging, as visual recognition and spatial relation understanding are both required. Existing efforts can be mainly categorized as two-stage and one-stage methods based on their detection strategy. Two-stage methods [4, 10, 11, 51] initially employ an off-the-shelf detector (e.g., Faster R-CNN [38]) to detect all humans and objects in the image. Thereafter, the detected humans and objects are exhaustively paired, followed by another network for interaction classification. By contrast, one-stage methods [21, 27, 57] attempt to solve this problem in a single stage. In particular, with the success of the Transformer [44] architecture, many approaches [22, 41, 54, 58] adopt the attention-based DETR [3] framework in HOI detection and have achieved better performance than their two-stage counterparts. However, the slow convergence of DETR results in high memory cost and increasing training overhead. By this reason, Zhang et al. [56] proposed to revitalize two-stage approaches with the help of DETR. For this method, the object detection is achieved with a fixed DETR in the first stage, and another attention-like mechanism is adopted for predicting the interactions.
In fact, performing interaction recognition expects accurate understanding about the spatial relation between human and objects, where the distance serves as one fundamental characteristic. Nevertheless, predicting interactions with different human-object distances is of distinct difficulties. Take Fig. 1 (a) as an example, the distant interaction (person1, block, frisbee) is harder to recognize than (person2, hold, frisbee), even for us humans. This observation is further reflected in Fig. 1 (b), where state-of-the-art method UPT [56] performs much worse on distant interactions. We attribute this problem to two inherent reasons. Firstly, distant interactions are intrinsically more difficult to recognize, owing to the scene complexity. A natural image often includes multiple humans and objects, while some of them overlap and entangle with each other. As a result, predicting interaction for distant human-object pairs is largely affected by their noisy context. Furthermore, distant interactions usually involve small objects, on which the detection backbone struggles due to low-resolution and occlusion. Secondly, the number of interactions with respect to human-object center distance demonstrates a long-tail distribution (see Fig. 1 (b)). Thus, it is unfavorable to fit distant human-object pairs in the training process, resulting in sub-optimal performance. Nevertheless, to the best of our knowledge, such an important problem remains unexplored by the HOI detection literature.

In this paper, we make contributions from the following two aspects to address the aforementioned problem. Firstly, we shed light on the two-stage training scheme in HOI detection and propose a novel two-stage method, dubbed Spatially Differentiated Transformer (SDT). Inspired by [56], we employ DETR to detect humans and objects (i.e., tokens), the representations of which are then enriched via an intra-class diversification module and spatial fusion. Then, we design a novel Far-Near Distance Attention (FNDA) mechanism to enable improved modeling for distant token pairs. Specifically, FNDA allows one token to propagate information with two glances: the first glance involves only far away tokens while the second glance focuses solely on near tokens. In this way, the model is able to better attend to distant tokens without the influence of nearby ones and vice versa. Thereafter, we pair the obtained features of each human and object, and employ another self-attention module for iterative context aggregation. At last, each human-object pair is classified into candidate interactions.

Furthermore, as the number of interactions manifests a long-tail distribution with respect to human-object distances, we therefore design a novel Distance-Aware (DA) loss function to re-weight each human-object pair during training. In particular, DA loss adaptively adjusts the weights for each human-object pair, wherein relatively higher and lower weights are assigned to distant and close interactions, respectively. It is expected that with our DA loss, distant interactions are treated with more importance, thereby alleviating the effect of dominance of close interactions during training.

To validate the effectiveness of our proposed method, we conduct extensive experiments on two challenging benchmarks, namely HICO-DET [4] and V-COCO [15]. The results show that our method outperforms existing approaches by a large margin, resulting in new state-of-the-art performance on the two benchmarks. Besides, additional analysis and visualizations further demonstrate that our method is able to better model distant interactions.

To summarize, the contributions of this paper are three-fold:

- We propose a novel two-stage method – Spatially Differentiated Transformer for HOI detection, wherein a Far-Near Distance Attention (FNDA) is devised to effectively model distant interactions.
- To balance the learning of interactions with different human-object distances, we design a novel Distance-Aware (DA) loss function to dynamically adjust the weight of each human-object pair according to their center distance.
- Extensive experiments on two benchmarks demonstrate that the proposed method surpasses existing approaches by a large margin, achieving new state-of-the-art results.\(^1\)

2 RELATED WORK

2.1 Human-Object Interaction Detection

Existing approaches for HOI detection can be mainly categorized as two-stage and one-stage methods, according to their detection strategy. Two-stage methods firstly adopt an off-the-shelf detector (like Faster-RCNN [38]) to detect all humans and objects in the image. Afterwards, the humans and objects are exhaustively paired and fed into a downstream network for interaction recognition. These approaches mostly focus on improving the interaction recognition capacity of the downstream network. Early work [4, 40] often leverages the visual appearance and spatial relation to identify the interactions. In addition, more features such as human pose [16, 46], gaze [50], 3D representations [24] and word embeddings [23, 32], have also been exploited. Different from these methods, some studies utilize the graph structure for information propagation between the detected humans and objects. For example, [36] builds a fully-connected graph while [10, 55] apply a bipartite graph for humans and objects. Despite their effectiveness, these methods usually suffer from sub-optimal detection performance [10, 25, 55].

By contrast, one-stage methods tackle the task in an end-to-end manner. Early one-stage approaches often detect the interaction points [27, 47] or the human-object union regions [21] as interaction clues. Recently, with the success of DETR in the object detection domain, some studies attempt to employ this Transformer-based architecture for HOI detection. For instance, [22, 41, 58] append more heads on the decoder to recognize the interaction. [6] applies another decoder for decoupled object detection and interaction recognition, while [54] performs the two steps in a cascaded way.

2.2 Single- and Cross-Modal Attention

The past few years have witnessed the rapid development of attention mechanism due to its effectiveness in a variety of tasks. Among the dedicated efforts, the Multi-Head Self-Attention (MHSA) [44] has recently attracted increasing research interest. In addition to its wide application in the language domain, some researchers attempt to apply MHSA to the vision tasks, such as image classification [7], semantic segmentation [33], point cloud analysis [8, 9] and video understanding [35]. Besides, some studies have been conducted to improve the efficiency of the attention mechanism. For example, [5, 14, 48] present to reduce the computational cost or model bi-level relations by splitting attention into two groups of different

\(^1\)Code available: https://github.com/daoyuan98/SDT-HOI.
While above methods focus only on single modality information, encoding e.g. first fed into a backbone network, superior results in object detection. In this step, an input image is detection with a Transformer-based DETR [3], which demonstrates Stage 1: Token Detection. Inspired by [56], we perform token detection with a Transformer-based DETR [3], which demonstrates superior results in object detection. In this step, an input image is first fed into a backbone network, e.g., ResNet-50 [17], to obtain the visual representation \( g \in \mathbb{R}^d \). With the addition of positional encodings, the visual features are then inputted to \( N \) Transformer encoder layers and \( N \) Transformer decoder layers sequentially. Thereafter, we have a set of token representations, corresponding to the token queries inputted to the decoder layers. Finally, these token representations are fed into two separate Multi-Layer Perceptrons (MLPs) for object classification and bounding box regression, respectively.

Stage 2: Interaction Recognition. In the second stage, we filter the tokens according to their confidence scores, and keep \( n \) most confident ones. Each token is represented as a feature vector \( t_i \in \mathbb{R}^d \), normalized bounding box \( b_i \in \mathbb{R}^4 \), confidence score \( s_i \), and class \( c_i \), where \( i = 1, ..., n \). With the detected tokens, we first post-process these tokens to make them more compatible with HOI detection. Then, these tokens are enabled to propagate information with distance discrimination by our Far-Near Distance Attention. Afterwards, we obtain the interaction representation by combing each human and object tokens with the global context. Finally, the interaction representations are fed into an MLP for final prediction.

3 METHODOLOGY

3.1 Preliminary

Method Intuition. Human-Object Interaction (HOI) detection is to detect and predict a set of \( \langle \text{human}, \text{verb}, \text{object} \rangle \) interaction triplets in an image. In this work, we attempt to address the problem that distant human-object often lead to inferior interaction recognition results. In particular, we propose a novel two-stage method named Spatially Differentiated Transformer (SDT), which is designed to flexibly model both distant and close interactions. As shown in Fig. 2, our method is composed of two stages: The first stage detects human and object (token) with an object detector, followed by the second stage of interaction recognition on the detected tokens.

Stage 1: Token Detection. Inspired by [56], we perform token detection with a Transformer-based DETR [3], which demonstrates superior results in object detection. In this step, an input image is first fed into a backbone network, e.g., ResNet-50 [17], to obtain the receptive fields. Qin et al. [37] decreases the quadratic computation complexity by eliminating the softmax layer. Other methods also consider the relative position between inputs by taking it as learned relative bias embedding [39] or part of the attention weights [33, 48]. While above methods focus only on single modality information, the effectiveness of the attention mechanism on cross-modality tasks has also been extensively studied. One typical operation is to replace the query matrix with the features from another modality, and the cross-modality information exchange is accordingly enabled. It demonstrates improved results on various multi-modal tasks, including visual question answering [31], image-text matching [30, 49], zero-shot learning [52] and audio-visual active speaker detection [42].

Stage 2: Interaction Recognition. In the second stage, we filter the tokens according to their confidence scores, and keep \( n \) most confident ones. Each token is represented as a feature vector \( t_i \in \mathbb{R}^d \), normalized bounding box \( b_i \in \mathbb{R}^4 \), confidence score \( s_i \), and class \( c_i \), where \( i = 1, ..., n \). With the detected tokens, we first post-process these tokens to make them more compatible with HOI detection. Then, these tokens are enabled to propagate information with distance discrimination by our Far-Near Distance Attention. Afterwards, we obtain the interaction representation by combing each human and object tokens with the global context. Finally, the interaction representations are fed into an MLP for final prediction. In the rest of this section, we will sequentially elaborate the token post-processing step, the Far-Near Distance Attention, our Distance-Aware loss function, and training and inference procedures.

3.2 Token Post-Processing

The token representations \( \{t_i\}_{i=1}^n \) obtained from DETR contains discriminative information for classification. However, they can be of low quality and oblivious to the spatial relations with other tokens. Therefore, we first post-process these tokens to enrich their representation for better interaction recognition.

Intra-Class Diversification. It is reasonable that the token representations are often of limited diversity. In particular, tokens representing small size objects often suffer from occlusion and low-resolution due to the scene complexity, and is thus less informative.
To this end, we propose an Intra-Class Diversification (ICD) module to enrich the token representations with the help of other instances from the same class.

The key to our ICD module is an object-wise memory, which stores token representations of high confidence score for each class. During training, given a token with representation \( t_i \), we take it as a query and randomly sample \( l \) features \( \{ t_k \}_{k=1}^{l} \) from the memory cell corresponding to its class for key and value computation. Then, the intra-class diversification is implemented through the cross-attention mechanism [44]:

\[
\begin{align*}
A_{ik} &= \frac{t_i W_Q^T W_k^T t_k}{\sqrt{d}} \\
B_{ik} &= \frac{e^{A_{ik}}}{\sum_{k'=1}^{n} e^{A_{ik'}}} \\
t_i' &= t_i + \sum_{k=1}^{l} B_{ik} t_k W_V
\end{align*}
\]

where \( W_Q, W_K, W_V \in \mathbb{R}^{d \times d} \) are learnable transformation matrices. This mechanism allows each token to adaptively aggregate representation from high quality tokens in the same class, thereby improving token diversity and enhancing model’s generalizability.

**Spatial Fusion.** Besides the visual representation, the bounding box of a detected token also serves as a key attribute for describing the spatial information. For understanding the relation between two tokens, it is more important to focus their relative spatial relations. Thus, inspired by [55, 56], we first obtain the pairwise spatial relation \( p_i = f(b_i, \{ b_j \}_{j=1}^{n}) \), which represents the spatial relation between the \( i \)-th and all the other tokens. The spatial relation vector \( p_i \) can be regarded as a high-level positional embedding, which better represents the detected tokens. Afterwards, we fuse the pairwise spatial relation \( p_i \) into the token representations:

\[
t_i = \text{FFN}(t_i; p_i),
\]

where \( \text{FFN} \) denotes a feed forward network. With the combination of spatial relations, the processed tokens are more informative, and thus more compatible for HOI detection.

### 3.3 Far-Near Distance Attention

After fusing the spatial embeddings, one common way to propagate information between tokens is to leverage the Multi-Head Self-Attention (MHSA) mechanism in Transformer [44]. Nonetheless, it is sub-optimal to directly apply MHSA for HOI detection. Broadly speaking, one can interact with multiple objects and vice versa, making MHSA struggle on dense scenes. Furthermore, interaction prediction for tokens with far and near distances is often of distinct difficulties. Current MHSA tackles all token-pairs without discrimination, which may unexpectedly impede the modeling for distant relations.

To this end, we design a novel Far-Near Distance Attention (FND) mechanism, which models the relation between tokens in two alternative steps, according to their center distance. Specifically, given the bounding boxes of the detected tokens, we first compute \( L_2 \) center distance for each token pair, and obtain the pairwise

\[
M^f = \begin{cases} 1, & D_{ij} > \text{Med}(d_i) \text{ or } i = j \\ 0, & \text{otherwise} \end{cases}
\]

\[
M^n = \begin{cases} 1, & i = j \\ 1 - M^f_{ij}, & \text{otherwise} \end{cases}
\]

where \( \text{Med}(d_i) \) denotes the median of the \( i \)-th row in matrix \( D \), which serves as the threshold for differentiating far and near distances. In this way, \( M^f \) allows the model to attend to far tokens only, which prohibits the intervention from near ones, and vice versa. Then, we perform the MHSA operation with the distance-guided masks:

\[
A_{ij} = \frac{i_i W_Q^T W_k^T i_j}{\sqrt{d}} \\
\tilde{A} = A \odot M^n \\
B_{ik} = \frac{e^{A_{ik}}}{\sum_{k'=1}^{n} e^{A_{ik'}}} \\
t_i' = t_i + \sum_{k=1}^{l} B_{ik} t_k W_V
\]

where \( \odot \) denotes element-wise product, \( W_Q^T, W_K^T, W_V^T \in \mathbb{R}^{d \times d} \) are learnable transformation matrices, \( M^n \) indicates either \( M^f \) or \( M^n \). In our model, \( M^f \) and \( M^n \) are alternatively applied, so that each token can iteratively update their representations with tokens of different distances, thereby improving the model’s capacity for

---

\(^1\)We detail the computation of \( f \) in supplementary material due to space limit.
distant relation modeling. It is worth noting that our FNDA can replace MHSA in the Transformer encoder layer with few efforts, resulting in our token encoder layer. The detailed operation is illustrated in Fig. 3.

Iterative Context Aggregation. Thereafter, we pair each human token and object token, and take them as the initial interaction representation for each candidate pair. We also fuse the global context feature $g$ into the representations, which can provide contextual clues for interaction recognition:

$$h_{ij} = [i; j] \odot \text{FFN}(g),$$

where $\odot$ denotes element-wise addition, $i$ implies the index for human and $i \neq j$. We then employ $L_i$ interaction encoder layers to perform self-attention [44] to iteratively update the interaction representations with the global context. After that, an MLP is used to predict the final interactions for each human-object pair.

### 3.4 Distance-Aware Loss

Following previous work [4, 32, 54, 55], we formulate interaction recognition with an multi-label classification objective, since there can be multiple interactions (e.g., read book and hold book) within one human-object pair. In this way, the full model can be optimized with the Binary Cross Entropy (BCE) loss. However, distant interactions often manifest rare in benchmark datasets, which leads to underfitting on these interactions. Therefore, we propose a Distance-Aware (DA) loss to adaptively assign higher weights and lower weights to distant and close interactions, respectively. In formulation, the interaction recognition model is optimized with:

$$L_{DA} = \sum_{i, j \neq j} w_{ij} \sum_{c \in C} y^c_{ij} \log \hat{y}^c_{ij} + (1 - y^c_{ij}) \log (1 - \hat{y}^c_{ij}),$$

where $\delta_{ij} = \sigma(\text{MLP}(h_{ij}))$ represents the verb scores transformed by the sigmoid function $\sigma(\cdot)$, $C$ is the number of classes and $y^c_{ij}$ indicates the ground-truth label of class $c$. We implement $w_{ij}$ as,

$$w_{ij} = \sigma(\alpha \cdot D_{ij} + \beta),$$

where $D_{ij}$ is the $L_2$ center distance between the $i$-th human and the $j$-th object, and $\alpha$ and $\beta$ are both learnable parameters. In this way, relatively higher weights are assigned to more distant interactions, so as to increase their importance during training. Notably, the proposed DA loss acts on each instance.

### 3.5 Training and Inference

**Training.** We first train the object detector, and then freeze it to train the interaction recognition model. Pertaining to the latter training phase, focal loss [28] has been proven effective to tackle the class-imbalance problem, which is shown to be influential to the performance in HOI detection [19, 55, 56]. Therefore, we combine our DA loss with the focal loss to optimize our model. Besides, since our method breaks HOI detection into two stages, after detection, infeasible verb-object combinations can be filtered out in advance.

**Inference.** During inference, the final interaction score $z$ is calculated as the multiplication of human confidence score $s_i$, object confidence score $s_j$ and their verb score $\delta_{ij}^v$:

$$z_{ij} = (s_i)^{\lambda} \cdot (s_j)^{\lambda} \cdot \delta_{ij}^v,$$

where $s_i$ and $s_j$ are obtained from DETR and $\lambda \geq 1$ is a constant to suppress overconfident objects [55, 56]. Infeasible verb-object combinations are also removed according to the object label output by DETR.

### 4 EXPERIMENTS

#### 4.1 Experimental Setup

**Datasets.** We conducted experiments on two benchmark datasets, namely HICO-DET [4] and V-COCO [15]. HICO-DET involves 80 COCO objects and 117 verb classes, resulting in a total of 600 interaction classes. There are 38,118 and 9,658 images for training and test in this dataset, respectively. V-COCO is built upon the MS-COCO [29] dataset. It covers 24 action categories with 80 COCO objects, and contains 2,533, 2,867 and 4,946 images for training, validation and test, respectively.

**Evaluation Protocol.** We adopted mean Average Precision (mAP) as the evaluation metric. A detection result is regarded as true positive if (1) the predicted human and object bounding box have IoUs larger than 0.5 with corresponding ground-truth boxes, and (2) the predicted action class is correct. We computed this metric for each interaction class in HICO-DET and verb class in V-COCO.

Following [4, 54, 56], we provide results under default setting and known-object setting on HICO-DET. For the first setting, the APs are calculated on the basis of all test images, while the second setting calculate APs based on images that contain the object corresponding to each class. Under both settings, the result under full (a total of 600 interaction classes), rare (less than 10 training instances) and non-rare (10 or more training instances) classes are all reported.

For V-COCO, we provide results under two evaluation settings: Scenario 1 and Scenario 2. In the former setting, the detector is required to report an empty box when no object is involved in the interaction, while the object box can be ignored in the latter one.

**Implementation Details.** We first fine-tuned DETR for 30 epochs on the two datasets, which has been pre-trained on MS-COCO. Following [56], some data augmentation techniques were applied in the detector fine-tuning process: We scaled the images such that the shorter side is between 480 to 800 pixels while the longer side is at most 1,333 pixels. Furthermore, each image was cropped with a probability of 0.5 to a random rectangle with each side between 384 to 600 pixels before scaled. Besides, we also applied color jittering augmentation, where brightness, contrast and saturation are randomly selected between 0.6 to 1.4. After that, the DETR was frozen in the next stage for interaction recognition.

For each given image, the fine-tuned DETR first perform object detection and generate 100 proposals. Then, we filter out tokens with confidence score less than 0.2, and keep 3–15 human/object tokens with higher confidence, based on which our SDT is trained. We trained the interaction recognition model with the AdamW optimizer [34], which has a learning rate of 2e-4 and weight decay of 1e-4. The SDT is trained for 20 epochs and the learning rate is decayed by 10 at the 10-th epoch. $L_y$ was set to 3 on both datasets.

---

Note that MS-COCO training set contains some images in the V-COCO test set, which should be excluded in the detector pre-training process [41, 56].
Table 1: Results on the HICO-DET and V-COCO datasets. The best results are highlighted in bold while the second best ones are underscored.

| Method       | HICO-DET          | V-COCO          |
|--------------|-------------------|-----------------|
|              | Default Setting   | Known-Object Setting | Scenario 1 | Scenario 2 |
| HO-RCNN [4]  | 7.81 5.37 8.54    | 10.41 8.94 10.85 | -          | -          |
| InteractNet [12] | 9.94 7.16 10.77 | - - - | 40.0 | - |
| GPNN [36]    | 13.11 9.34 14.23  | 19.17 15.51 20.26 | 47.8 | 54.2 |
| TIN [26]     | 17.03 13.42 18.11 | 23.40 21.75 23.89 | 51.0 | - |
| DRG [10]     | 19.26 13.74 19.71 | - - - | 44.0 | - |
| VSGNet [43]  | 19.80 16.05 20.91 | 23.69 20.64 24.60 | 51.8 | 57.0 |
| TIN [24]     | 21.34 18.53 22.18 | 24.81 17.09 27.12 | 47.8 | 54.2 |
| PPDM [27]    | 21.94 13.97 24.32 | 24.81 17.09 27.12 | - | - |
| ConsNet [32] | 22.15 17.55 23.52 | 25.98 19.12 28.03 | 48.3 | 54.2 |
| VCL [18]     | 23.63 17.21 25.55 | 27.89 23.64 29.16 | 53.3 | 60.3 |
| ATL [19]     | 23.81 17.43 27.42 | - - - | 53.2 | - |
| IDN [25]     | 24.58 20.33 25.86 | 27.89 23.64 29.16 | 53.3 | 60.3 |
| HOTR [22]    | 25.10 17.34 27.42 | - - - | 55.2 | 64.4 |
| FCL [20]     | 25.27 20.57 26.67 | 27.71 22.34 28.93 | 52.4 | - |
| HOI-Trans [58] | 26.61 19.15 28.84 | 29.13 20.98 31.57 | 52.9 | - |
| AS-Net [6]   | 28.87 24.25 30.25 | 31.74 27.07 33.14 | 53.9 | - |
| SCG [35]     | 29.26 24.61 30.65 | 32.87 27.89 34.35 | 54.2 | 60.9 |
| QPIC [41]    | 29.90 23.92 31.69 | 32.38 26.06 34.27 | 58.8 | 61.0 |
| OCN [53]     | 30.91 25.56 32.51 | - - - | - | - |
| CDN [54]     | 31.44 27.39 32.64 | 34.09 29.63 35.42 | 61.7 | 63.8 |
| UPT [56]     | 31.66 25.94 33.36 | 35.05 29.27 36.77 | 59.0 | 64.5 |
| SDT (ResNet-50) | 32.45 28.09 33.75 | 35.95 31.30 37.34 | 60.3 | 65.7 |
| SDT (ResNet-101) | 32.97 28.49 34.31 | 36.32 31.90 37.64 | 61.8 | 67.6 |

while \( L_I \) was 2 and 3 on V-COCO and HICO-DET, respectively. We set \( \lambda \) in Eq. 8 to 1 for training and 2.8 for inference. The token dimension \( d \) is set to 256. For all of the attention mechanism adopted in this paper, we set the number of heads to 8, hidden dimension to 1024 and dropout probability to 0.1. We conducted all experiments on 4 NVIDIA RTX A5000 GPUs with CUDA 11.1, whereby each GPU has a batch of 4 images, resulting in an effective batch size of 16. It takes about 6 hours and 40 minutes to train on HICO-DET and V-COCO, respectively. Besides, we employed two backbone networks, i.e., ResNet-50 and ResNet-101 for global feature extraction, resulting in two variants of SDT.

4.2 Comparison with State-of-the-art Methods

We compared the proposed SDT with state-of-the-art methods, and reported the results in Table 1. It can be observed that, with ResNet-50 as backbone, the proposed method already outperforms existing methods significantly on the two datasets. For example, on HICO-DET, we surpass UPT [56] by about 1 mAP on both default setting and known-object setting. On V-COCO, SDT also achieves consistent improvements. Specifically, under scenario 1, our method can outperform the second-best two-stage method UPT [56] by 1.3 mAP, while under scenario 2, SDT achieves the best performance over all the existing methods. Notably, the proposed method can also benefit from a stronger backbone (i.e., ResNet-101). For example, on V-COCO, we outperform all existing methods under two scenarios. In particular, under scenario 2, we can further improve upon SDT with ResNet-50 by 1.9 mAP, resulting in an improvement of 3.1 mAP upon the runner-up. These results prove the superiority of the proposed method over existing approaches.

4.3 Ablation Studies

To further investigate the effects of each component in the proposed method, we conducted extensive ablation studies on the larger HICO-DET dataset with ResNet-50 as backbone.

Module Effectiveness. We first study the effectiveness of token encoder layers (T-Encoder), interaction encoder layers (I-Encoder) and DA loss, and show the results in Table 2. We can see that all three modules improve the baseline with a clear margin. For example, with T-Encoder, our method surpasses the baseline by more than 2 mAP. Furthermore, any combination of two modules outperforms the variant with a single module, indicating that the modules are in fact complementary to each other. In particular, the combination of T-Encoder and I-Encoder can boost upon the baseline by more than 4 mAP. Finally, with all of the three modules, the proposed method achieves the best results, which leads to an improvement of more than 5 mAP upon the baseline.
Figure 4: Qualitative results from MHSA and the proposed FNDA. (a) Attention value variation of MHSA and FNDA with respect to the normalized center distances. Note that we split the attention values into bins of 0.05 according to the center distance, where the mean and variance in each bin are both shown. (b) Visualization of attention maps in two randomly selected images. The edges in the interaction graphs indicate the existence of interaction between connected nodes.

Table 2: Effectiveness of each module in the proposed SDT.

|     | T-Encoder | I-Encoder | DA Loss | Full  | Rare  | Non-Rare |
|-----|-----------|-----------|---------|-------|-------|----------|
| ✓   | ✓         | ✓         |         | 27.08 | 23.01 | 28.30    |

Table 3: Performance variation with different numbers of layers.

|     | L_T | L_I | Full | Rare | Non-rare |
|-----|-----|-----|------|------|----------|
| 2   | 2   | 31.06 | 25.98 | 32.58 |
| 2   | 3   | 31.40 | 26.73 | 32.79 |
| 2   | 4   | 31.02 | 25.77 | 32.59 |
| 3   | 2   | 31.71 | 26.95 | 33.13 |
| 3   | 3   | 32.45 | 28.09 | 33.75 |
| 3   | 4   | 31.93 | 27.54 | 33.24 |
| 4   | 2   | 30.91 | 25.01 | 32.67 |
| 4   | 3   | 30.82 | 26.08 | 32.23 |
| 4   | 4   | 30.67 | 25.57 | 32.20 |

Quantitative Study on FNDA. To demonstrate the superiority of the proposed FNDA, we compared it with the most widely used Multi-Head Self-Attention (MHSA) [44], and show the results in Table 4. We can see that the plain FNDA exceeds MHSA by more than 1 mAP. In addition, MHSA also benefits from the token post-processing steps, i.e., ICD and spatial fusion, which also proves the validity of these two. After combing our FNDA with these two steps, the performance can be further promoted, making it outperform the MHSA by more than 1.5 mAP.

Table 4: Comparison between MHSA and FNDA. ++ICD means composing ICD upon spatial fusion.

| Variant | Full  | Rare  | Non-Rare |
|---------|-------|-------|----------|
| MHSA [44] | 30.66 | 26.06 | 32.03    |
| + Spatial Fusion | 30.80 | 24.78 | 32.60    |
| ++ ICD | 30.90 | 26.74 | 32.14    |
| FNDA (Ours) | 31.70 | 26.26 | 33.32    |
| + Spatial Fusion | 32.10 | 27.02 | 33.62    |
| ++ ICD | 32.45 | 28.09 | 33.75    |

Qualitative Analysis on FNDA. We used all interactive token pairs in the HICO-DET test set, and kept their corresponding attention values in the token encoder layers for both MHSA and our FNDA. We then illustrate these values with respect to normalized human-object center distances in Fig. 4 (a). It can be observed that the attention values generally decline with the increase of distance for both types of attentions, implying that information propagation between distant pairs is more challenging. More importantly, our FNDA assigns higher attention weights on distant interactions than MHSA, which further justifies the superiority of FNDA over MHSA.

We then used two randomly selected cases to illustrate why FNDA outperforms MHSA. The two images in Fig. 4 (b) are selected from HICO-DET test set and the attention map is averaged for far-attention and near-attention. We observe that the strength...
of FNDA comes from two aspects: 1) It can effectively propagate information between distant and interactive human-object pairs, and thus benefit distant interaction recognition. For example, in the first case of Fig. 4 (b), the person in box2 is lassoing the cow in box5. The attention value (2, 5) is minor in MHSA due to the far distance. By contrast, our FNDA can emphasize more on this human-object pair. 2) FNDA is able to constrain non-interactive human/objects, thereby focusing more on informative ones. For instance, in the second case of Fig. 4 (b), the people in box3 and box4 are not involved in any interactions to the people in box1 and box2. Nonetheless, MHSA assigns higher values in (1, 3), (1, 4), (2, 3), (2, 4), which distracts the people in box1 and box2 from attending to truly related objects. On the contrary, our FNDA can alleviate this problem by attending to more informative tokens in the image.

4.4 Qualitative Results

To further understand the effectiveness of the proposed SDT, we randomly selected some images from the HICO-DET test set and visualized SDT prediction results in Fig. 5. Note that we only showed partial interactions for clearer view. It can be observed that our method is capable of handling interaction of diverse distances. For example, distant instances, such as fly kite in Fig. 5 (e) and pull kite in Fig. 5 (f), and close instances, such as ride horse in Fig. 5 (i), can both be successfully recognized by the proposed method. Notably, when the detected object is small (e.g., the frisbee in Fig. 5 (c) and the ball in Fig. 5 (h)), SDT can provide the correct interaction with a high confidence. Furthermore, it also demonstrates an evident advantage on scenes with complex context, where multiple objects interact with the same person. For instance, the man in Fig. 5 (d) is feeding three birds simultaneously, yet our method correctly predicts the interaction for all these three objects.

5 CONCLUSION AND FUTURE WORK

In this work, we propose a novel two-stage method for better modeling distant interactions in HOI detection. Two fundamental components make our method distinguished from existing approaches. First, we design a Far-Near Distance Attention to guide the model attention to flexibly focus on distant and close interactions. Second, a novel Distance-Aware loss is presented by this paper to handle the long-tailed interaction-distance distribution. We conduct extensive experiments on two benchmark datasets and observe that our method achieves a new state-of-the-art on them. Additional analysis further validates that the performance gain is mainly brought by the distant interaction modeling from the proposed method.

As indicated by our empirical findings, a large quantity of distant interactions involve relatively small objects, which impedes the interaction recognition due to the weak detection results. It is thus promising for HOI detection to be equipped with improved detectors on small-sized objects. In addition, the 2D image shows limitation in distance estimation as the human-object relations are physically expressed in the 3D world. In view of this, incorporating the depth information to distance modeling will potentially benefit the learning of distant interactions in HOI detection.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.
