Geometric Features Enhanced Human-Object Interaction Detection

Manli Zhu, Edmond S. L. Ho, Shuang Chen, Longzhi Yang, *Senior Member, IEEE*, Hubert P. H. Shum†, *Senior Member, IEEE*

***Abstract*—Cameras are essential vision instruments to capture images for pattern detection and measurement. Human-object interaction (HOI) detection is one of the most popular pattern detection approaches for captured human-centric visual scenes. Recently, Transformer-based models have become the dominant approach for HOI detection due to their advanced network architectures and thus promising results. However, most of them follow the one-stage design of vanilla Transformer, leaving rich geometric priors under-exploited and leading to compromised performance especially when occlusion occurs. Given that geometric features tend to outperform visual ones in occluded scenarios and offer information that complements visual cues, we propose a novel end-to-end Transformer-style HOI detection model, i.e., geometric features enhanced HOI detector (GeoHOI). One key part of the model is a new unified self-supervised keypoint learning method named UniPointNet that bridges the gap of consistent keypoint representation across diverse object categories, including humans. GeoHOI effectively upgrades a Transformer-based HOI detector benefiting from the keypoints similarities measuring the likelihood of human-object interactions as well as local keypoint patches to enhance interaction query representation, so as to boost HOI predictions. Extensive experiments show that the proposed method outperforms the state-of-the-art models on V-COCO and achieves competitive performance on HICO-DET. Case study results on the post-disaster rescue with vision-based instruments showcase the applicability of the proposed GeoHOI in real-world applications.**

***Index Terms*—Human-object Interaction, Object Keypoints, Interactiveness Learning, Graph Convolutional Network, Attention Mechanism.**

I. Introduction

Cameras, as predominant vision instruments, are extensively employed in methods that rely on visual measurements [1], such as human and object pose estimation [2], [3], [4]. Human-object interaction (HOI) detection is one of the most popular pattern detection approaches for captured human-centric visual scenes. It involves identifying and localizing interactive human-object pairs while predicting the specific interactions between them within an image, yielding HOI triplets $\langle human, interaction, object \rangle$. It plays an important role in numerous applications, such as action recognition [5] and surveillance event detection [6], [7].

M. Zhu and L. Yang are with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK. Emails: {manli.zhu, longzhi.yang}@northumbria.ac.uk

E. S. L. Ho is with the School of Computing Science, University of Glasgow, Glasgow, UK. Email: shu-lim.ho@glasgow.ac.uk

S. Chen and H. P. H. Shum are with the Department of Computer Science, Durham University, Durham, UK. Emails: {shuang.chen, hubert.shum}@durham.ac.uk

†Corresponding author: H. P. H. Shum

The existing HOI detection methods generally fall into two-stage or end-to-end approaches. Two-stage approaches [8], [9], [10] typically take advantage of off-the-shelf object detectors like Fast R-CNN [11]. They first detect all instances (i.e., humans and objects) in an image, and then the interaction classification is carried out on every human-object pair. These methods may lead to sub-optimal HOI detections due to the independent optimization of two sub-problems [12], i.e., object detection and interaction classification. In contrast, end-to-end approaches detect the components of an HOI triplet all at once [13]. In the earlier end-to-end attempts [14], [15], interaction points and object proposals are detected simultaneously. The interactions are then associated with each human-object pair. However, in still images of complex scenes, such as crowded areas with interaction points overlapping among different human-object pairs, these methods could lead to inaccuracies and misinterpretations [16], [13].

End-to-end Transformer-based models [17], [12], [16] have been proposed to overcome these limitations, achieving state-of-the-art performance. Inspired by the Transformer object detector DETR [18], these approaches frame the HOI detection as a set prediction problem, using a bipartite matching loss to align interaction queries with ground-truth HOI triplets. While successful, rich prior knowledge (e.g., the semantic features and structure information) is under-exploited due to the random initialization of parametric interaction queries. To address this limitation, [13], [19], [20] explored semantics, spatial features, and structure information. Nevertheless, the spatial features including instance bounding boxes and human-object layout employed in these works are too coarse to capture fine-grained relationships between human body parts and object parts. The fine-grained geometric features, such as human pose and object structure have proven to be highly effective in two-stage methods [8], [10], [21]. However, they remain under-explored in existing Transformers due to their one-stage paradigm of HOI detection. In this work, we investigate how to enrich HOI representations with fine-grained geometric features in an end-to-end Transformer framework.

To this end, we propose a **Geo**metric features enhanced **H**uman-**O**bject **I**nteraction detection model (GeoHOI). Given that geometric features tend to outperform visual features on datasets with heavy occlusion [22] and offer information that complements visual cues, our idea is to learn fine-grained geometric features (i.e., keypoints) to facilitate interactiveness prediction of human-object pairs and to enhance interaction query representation. In detail, GeoHOI improves the Transformer-based framework of STIP [19] by introducing three novel

components. First, a keypoints detection module unifies the keypoint detection across different object categories, including humans, and is integrated into GeoHOI for end-to-end HOI detection. It simplifies the appearance distribution of different object classes by reconstructing object segmentation masks instead of their RGB images, allowing the network to focus on learning different shapes and enabling it to learn keypoints for arbitrary objects. As a result, it generates consistent and robust keypoint representation across different object categories. Second, a keypoint-aware interactiveness prediction module employs a graph convolutional network, capturing the holistic cues (i.e., cross-instance features) between humans and objects that complement pairwise features to effectively predict the interactiveness of human-object pairs. Third, a part attention module intends to identify informative local cues since specific interaction types are defined with detailed local information of human and object parts. This enhances the representation of interaction queries in the HOI Transformer for effectively classifying specific interactions. Thus, we exploit a self-attention mechanism to produce part-level attention, with keypoint positions serving as positional encodings. This allows the HOI classifier to focus on specific local regions that are informative to each interaction type.

We evaluate our model on two HOI benchmarks V-COCO [23] and HICO-DET [24]. The proposed GeoHOI achieves superior results on both datasets. Source codes are available at https://github.com/zhumanli/GeoHOI. Our contributions are:

- We introduce GeoHOI, a geometric features enhanced human-object interaction detection approach, facilitating pattern detection and measurement in images captured by vision instruments.
- We present a self-supervised keypoints learning method (UniPointNet) to detect keypoints for different object categories including humans in a unified manner. To the best of our knowledge, this is the first attempt that unifies keypoints detection across different object classes in HOI.
- We design a keypoint-aware interactiveness prediction module that incorporates holistic relationships between humans and objects. The geometric keypoint features are exploited to measure the likelihood of human-object interactions, boosting the interactiveness prediction of human-object pairs.
- We propose a part attention module that refines interaction query representation using self-attention, enhancing specific interaction prediction by identifying informative human and object parts.
- We demonstrate the effectiveness of our proposed GeoHOI by conducting experiments in public HOI detection benchmark datasets, outperforming state-of-the-art methods by a large margin of 3.4 mAP on V-COCO and 3.76 mAP on HICO-DET. We further conduct a real-world application case of post-disaster with UAVs, and GeoHOI outperforms all the baselines in terms of AP and recall.

II. Related Work

A. Two-stage Methods

1) Multi-stream Approaches: Early HOI detection models are typically implemented with a two-stage framework. In the first stage, an object detector such as Fast R-CNN [11] is used to localize instances. In the second stage, a classifier is trained to predict human-object interactions. Two-stage methods use pre-trained object detectors to simplify HOI detection, achieving a good trade-off between performance and complexity [10]. Earlier works focus on designing multi-branch HOI classifiers with convolutional neural networks modelling human and object appearance features and spatial layout. Gkioxari et al. [25] extended Fast R-CNN by introducing a human-centric branch to predict interactions at each target object location. Chao et al. [24] proposed a three-branch framework to model pairwise human-object appearance features and their spatial relations. Hou et al. [26] presented a five-branch framework with a novel fabricated compositional branch targeting the issue of long-tailed distributions of HOI interactions. These methods mainly focus on exploring the pairwise human and object features, overlooking the holistic features that could complement the pairwise ones.

Some works have exploited graph convolutional networks (GCNs) to model the relationships between humans and objects from a global perspective. Qi et al. [27] proposed a fully connected graph with humans and objects as nodes, and the adjacency matrix was inferred by their proposed link function. Ulutan et al. [9] introduced a visual-spatial-graph network to model structural connections between instances. Similar to Qi et al., they model humans and objects as nodes. Instead of a fully connected graph, they only build connections between inter-class instances, omitting unnecessary human-human and object-object pairs. Their adjacency matrix is predicted by the visual branch. Zhang et al. [28] presented a spatially conditioned graph with a multi-branch fusion module computing the adjacency structure and refining graph features. GCN-based HOI methods have shown that the modelling of intra-level and inter-level HOI representations can significantly improve HOI detection performance [29]. The reason is that GCNs not only capture pairwise features but also infer holistic cross-instance cues, which are useful for HOI reasoning. We leverage its advantage by fusing both pairwise features and cross-instance cues to enhance HOI prediction.

2) Geometric Features Informed Approaches: Geometric features such as human pose and object structure provide fine-grained spatial information and have been proved to be effective in improving HOI detection performance in two-stage methods. Fang et al. [30] and Wan et al. [8] explored the semantic cues of human body parts with an attention module that effectively identifies the most informative body parts for HOI recognition. Wu et al. [21] proposed to extract cross-person cues for body parts, which afford useful and supplementary information for the discovery of interactiveness. Park et al. [31] designed a graph with a pose-conditioned self-loop structure, allowing the human node embedding to be updated based on the local features of human joints. As discussed, the human pose has been well-studied in HOI detection, while the geometric features of objects such as keypoint positions are less explored. To overcome this, Zheng et al. [32] proposed to model the interactions between human joints and object keypoints using a graph network for capturing fine-grained spatial relationships in HOI detection. Nevertheless,

the representation of object keypoints in their work (i.e., two corner points of the object bounding box) is too simple to capture object shapes or structures as it considers only the rectangular spatial scope of an object.

Efforts have been made to improve the representation of object structure in HOI detection. Zhu et al. [10] proposed a deterministic method for representing object keypoints which encapsulate the underlying structure of an object. They extracted an object skeleton from its segmentation mask using a morphological skeletonization algorithm and obtained its keypoints by applying the K-means clustering to the set of key points on the skeleton. This kind of non-probabilistic method is less robust in handling various object shapes, particularly when dealing with non-articulated objects, making it difficult to accurately detect keypoints across different objects. Bar et al. [33] exploited transfer learning to estimate animal keypoints with a pre-trained object keypoints detector and adopted the interest point detection in geometry with bin girding to obtain keypoints for artefacts such as beds and computers. Ito [34] proposed a human and object keypoint-based extension module to improve conventional HOI detection models such as [9]. However, the different representations of human and object keypoints presented in these frameworks are less consistent and difficult to maintain across different objects, making them less applicable to real-world applications. In this work, we explore a self-supervised framework for learning keypoints of both humans and objects, which is a more versatile and robust approach for keypoint estimation.

B. End-to-end Transformer-based Methods

Transformers have shown superior performance in many fields including HOI detection due to their advanced network architecture and high capacity. They are first adopted in [17], [12], [16] by utilizing the vanilla Transformer architecture [18] to map the parametric interaction queries into a set of HOI predictions with a bipartite matching loss. Later, Kim et al. [35] introduced a multi-scale Transformer architecture to boost HOI detection. Recently, a multiplex relation network that disentangled Transformer decoders to encourage rich context exchange was proposed in [36]. Unlike two-stage methods that optimize instance detection and interaction detection in separate stages, these end-to-end frameworks infer human-object relationships from a global contextual perspective. They predict all elements of HOI triplets directly, significantly surpassing the performance of existing two-stage approaches. Nevertheless, the rich prior knowledge, such as spatial features [9], are not exploited in the above Transformer-based attempts.

Some studies have attempted to inject prior knowledge into Transformer architectures, to address the aforementioned limitation. Iftekhar et al. [13] proposed to utilize the semantic features (i.e., text embeddings) and the spatial features (i.e., the relative spatial configuration of human and object bounding box locations) to enhance the query representations of decoders. Zhang et al. [19] exploited the inter-interaction semantic structure and intra-interaction spatial structure over interaction proposals (i.e., human-object pairs) to strengthen HOI predictions. Xie et al. [20] proposed a novel category query learning approach where interaction queries are explicitly associated with specific and fixed image categories, facilitating HOI detection. We observe that spatial features, such as instance bounding boxes and human-object layout used in these works are too coarse to capture fine-grained relationships between human body parts and object parts, which have been demonstrated to be beneficial in existing two-stage HOI models [10], [33], [34]. In this paper, we leverage the geometric keypoint features to facilitate HOI classification in an end-to-end Transformer-based framework.

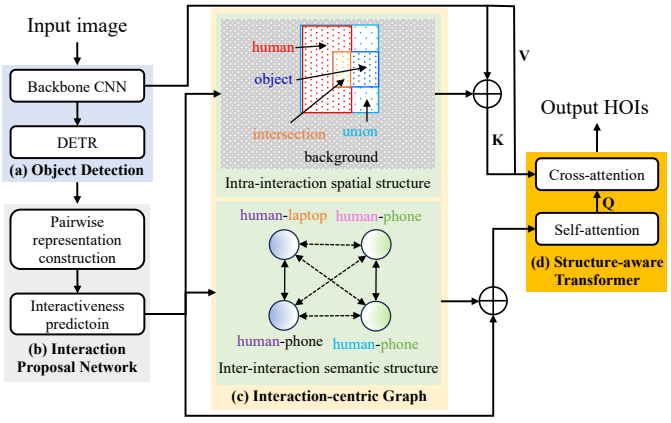

Fig. 1. Simple illustration of STIP. The solid bi-directional arrow means whether or not two HOI triplets share the same human or object, and the dashed bi-directional arrow denotes they do not share anything. (a) Given an input image, DETR is used to detect humans and objects. (b) By constructing all possible human-object pairs, the interaction proposal network uses pairwise features to filter non-interactive ones. (c) Next, an interaction-centric graph is built to inject rich inter-interaction semantic structure and intra-interaction spatial structure. (d) Finally, a structure-aware Transformer is utilized to output a set of HOI predictions.

III. Overview of GeoHOI

This work aims to improve end-to-end Transformer-based HOI detection networks with fine-grained geometric features of humans and objects. To this end, we propose GeoHOI. It utilizes learnable fine-grained geometric features (i.e., keypoint positions) to facilitate the interactiveness prediction of human-object pairs and to enhance interaction query representations. Inspired by the process of HOI detection with prior knowledge, we improve the structure-aware Transformer over interaction proposals (STIP [19]) by using keypoint features. As shown in Fig. 1, STIP is an improved network over the vanilla Transformer with prior knowledge of inter-interaction (i.e., whether or not two HOI triplets share the same human or object) and intra-interaction (i.e., the layout of human and object) structure. It becomes a natural backbone of GeoHOI due to its decompose-style design of HOI predictions, i.e., interaction proposals are first generated, followed by interaction classification. Such design allows us to explore rich geometric features for effective interaction proposal generation and non-parametric interaction query representation.

Specifically, our framework introduces three novel components to STIP, i.e., keypoints detection with our novel UniPointNet, keypoint-aware interactiveness prediction module for predicting interactive human-object pairs, and part

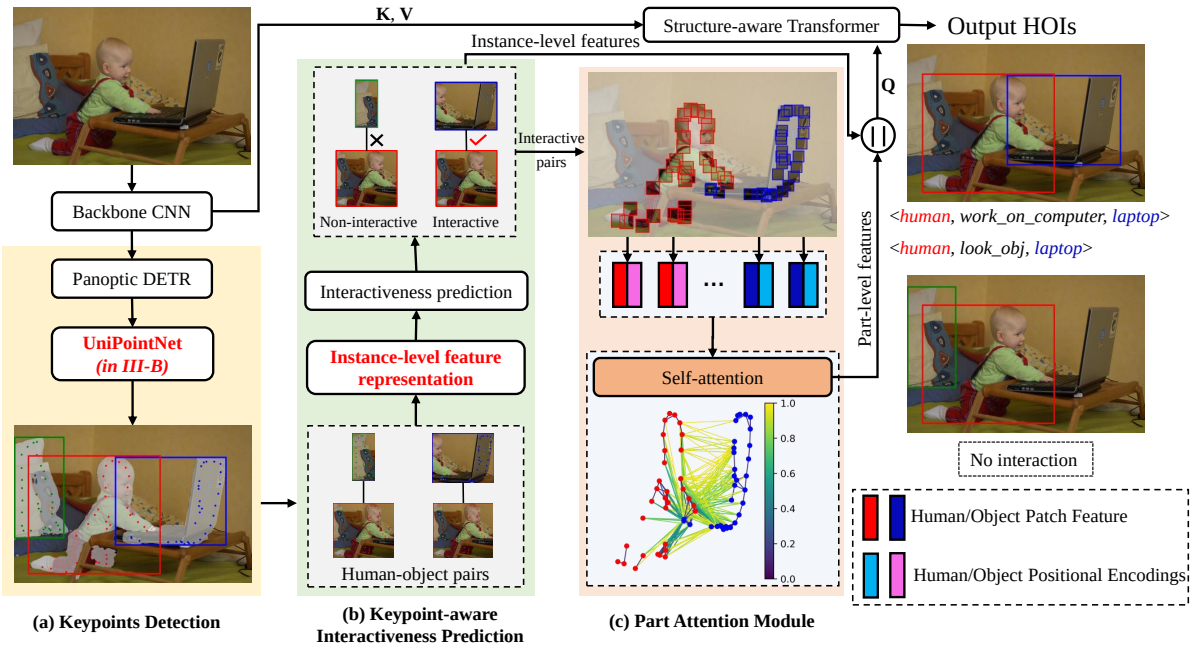

Fig. 2. An overview of our GeoHOI framework, in which ⓘ denotes the concatenation operation. (a) Given an image, we adopt the off-the-shelf Panoptic DETR [18] to detect the human and object instances within this image, generating their bounding boxes and segmentation masks. Based on the masks, we use our proposed UniPointNet to detect keypoints for all instances. (b) With the detected instances, the keypoint-aware interactiveness prediction module enumerates all possible human-object pairs. It then generates interactive ones with the highest interactiveness scores using coarse instance-level features, including pairwise and holistic graph features. (c) By taking all the interactive human-object pairs, we enhance their representations with human and object local patches, which are attended by self-attention. This encourages each interaction query to focus on informative human and object parts. The final concatenated representations serve as interaction queries which are then fed into the structure-aware Transformer [19] to output a set of HOI predictions.

attention module to enhance interaction query representation with informative human and object local parts. We start with GeoHOI architecture (Section III-A), then introduce the keypoint-aware interactiveness prediction module (III-B) and the part attention module (III-C). Section IV details the keypoints detection network (UniPointNet).

A. Architecture of GeoHOI

An overview of GeoHOI is shown in Fig. 2. Given an input image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the image height, width and channels, accordingly, GeoHOI first extracts the image feature map $\boldsymbol{F_x} \in \mathbb{R}^{H'\times W'\times d}$ with a CNN backbone of ResNet. $\boldsymbol{F_x}$ is then sent to Panoptic DETR [18] to obtain instance detections including bounding boxes and segmentation masks. Next, the segmentation masks are fed into UniPointNet to obtain keypoints for each instance. After that, the keypoint-aware interactiveness prediction module constructs pairwise and holistic graph features for instance-level feature presentation. It then predicts and outputs interactive pairs, enhanced by local cues from keypoints in the part attention module. Finally, the structure-aware Transformer generates HOI predictions. Details are introduced in the following sections.

B. Keypoint-aware Interactiveness Prediction Module

The Keypoint-aware Interactiveness Prediction (KIP) module aims to suppress non-interactive human-object pairs using coarse instance-level features. It transforms the random parametric interaction queries in the vanilla Transformer to non-parametric interaction proposals equipped with prior knowledge (e.g., instance visual features and their spatial layout), facilitating relational reasoning among interactions in HOI set prediction [19]. When learning the interactiveness of a human-object pair, visual cues can be explored not only from the targeted human and object but also from other humans and objects in the scene [21], providing a more comprehensive understanding of the scene. However, previous works such as [37], [19] only consider target pairwise features, failing to effectively extract interactive pairs. As a potential solution, mining cues from a global cross-instance perspective, i.e., using other humans and objects as a reference, would offer helpful and supplementary information for interactiveness inference. Therefore, in addition to pairwise features, we incorporate graph features using keypoint positions measuring the geometric distance with a graph convolutional network from a global perspective, extracting cross-instance cues. We first enumerate all human-object pairs using the detected instances by Panoptic DETR, and the KIP then estimates the likelihood of interaction for each pair based on both pairwise features and holistic graph features through a multi-layer perceptron (MLP). Finally, the KIP module outputs the top-K human-object pairs with the highest probability scores.

Concretely, for each human-object pair, the human visual feature $\boldsymbol{f}_h^v$, object visual feature $\boldsymbol{f}_o^v$, spatial feature $\boldsymbol{f}_s^u$, union feature $\boldsymbol{f}_u^v$ are represented as 256-dimensional vectors, while the object's semantic feature $\boldsymbol{f}_o^c$ (the embedding of the object class label) is a 300-dimensional vector. We refer to these as pairwise features. For the graph representation, we model humans and objects as nodes, connecting each human to all

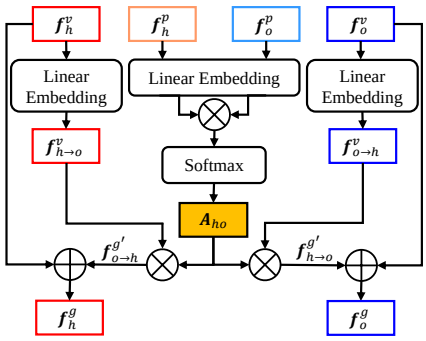

Fig. 3. Illustration of the graph convolution layer, in which $\otimes$ represents the tensor product, and $\oplus$ is the residual connection. The output graph features encode relationships between all humans and objects from a global perspective, with keypoint similarity measuring their connectivity.

objects and each object to all humans. As shown in Fig. 3, the node features are represented with visual features $\boldsymbol{f}_h^v$, $\boldsymbol{f}_o^v$. We use the similarities between human keypoints and object keypoints to define the adjacency matrix $\boldsymbol{A}$. This highlights that closer keypoints between a human and an object indicate a higher likelihood of interaction. Unlike the implicit adjacency matrix representation (i.e., it is predicted from instance visual features) in [9], the keypoints similarity explicitly captures the geometric distance prior knowledge between a human and an object, resulting in effective interactiveness prediction. As depicted in Fig. 3, given the keypoint features $\boldsymbol{f}_h^p \in \mathbb{R}^{N\times 2}$ of a human and $\boldsymbol{f}_o^p \in \mathbb{R}^{N\times 2}$ of an object, they are first embedded by a linear layer with 128-dimensional vectors and the similarity between them is served as their edge weight $\boldsymbol{A}_{ho} \in \boldsymbol{A}$, which is expressed as follows:

$$\boldsymbol{A}_{ho} = \phi(\boldsymbol{f}_h^p) \otimes \phi(\boldsymbol{f}_o^p), \tag{1}$$

where ϕ is implemented with a linear layer to encode keypoint positions, $\otimes$ denotes the dot product. Note that the edge weight between a human and an object is symmetric, i.e., $\boldsymbol{A}_{ho} = \boldsymbol{A}_{oh}$. The graph features $\boldsymbol{f}_h^g$ and $\boldsymbol{f}_o^g$ are then defined as follows:

$$\boldsymbol{f}_h^g = \boldsymbol{f}_h^v + \sum_{o=1}^{\hat{O}} \boldsymbol{A}_{ho} \boldsymbol{f}_{o\to h}^v, \tag{2}$$

$$\boldsymbol{f}_o^g = \boldsymbol{f}_o^v + \sum_{h=1}^{\hat{H}} \boldsymbol{A}_{oh} \boldsymbol{f}_{h\to o}^v, \tag{3}$$

where $\hat{O}$ and $\hat{H}$ are the numbers of humans and objects, $\boldsymbol{f}_{o\to h}^v$ is the projection of object visual feature $\boldsymbol{f}_o^v$ in the human space, and $\boldsymbol{f}_{h\to o}^v$ is the projection of human visual feature $\boldsymbol{f}_h^v$ in the object space.

Finally, the instance-level features for interactiveness prediction of each human-object pair in this module are obtained as the concatenation of all the features as follows:

$$\boldsymbol{f}_{ho} = \boldsymbol{f}_h^v \textcircled{\scriptsize ||} \boldsymbol{f}_o^v \textcircled{\scriptsize ||} \boldsymbol{f}_h^g \textcircled{\scriptsize ||} \boldsymbol{f}_o^g \textcircled{\scriptsize ||} \boldsymbol{f}_u^s \textcircled{\scriptsize ||} \boldsymbol{f}_o^c \textcircled{\scriptsize ||} \boldsymbol{f}_u^v. \tag{4}$$

C. Part Attention Module

While the instance-level features provide coarse information for interactions, specific interaction types are defined with fine-grained details. They highlight local information on human and object parts that are unlikely to be captured in instance-level features [8]. In addition, the fine-grained correlations among human body parts and object parts (e.g., the spatial layout between the human hands and the laptop keyboard shown in Fig. 2) implicitly depict the consistent spatial, scale, and co-occurrence relationships between humans and objects, providing a finer granularity context information of an image [33]. However, existing works [38], [21] only consider human body parts while overlooking object structure parts.

To address the aforementioned limitation, we introduce a Part Attention Module (PAM) designed to identify the most relevant parts of both humans and objects for detecting a specific interaction category. We use self-attention to learn the part-level features of a given human-object pair, enabling each part to aggregate information from all other parts, regardless of their distance or position. This allows the network to extract richer and more comprehensive context features, leading to a deeper understanding of the scene. This module serves to enhance the interaction query representation of each selected interactive human-object pair, improving the effectiveness of classifying particular interactions.

In detail, given the human keypoints $\boldsymbol{f}_h^p = \left\{\boldsymbol{f}_h^{p1} \ldots, \boldsymbol{f}_h^{pN}\right\}$, we define a local region $\boldsymbol{x}_{pi} \in \mathbb{R}^4$ for each keypoint $\boldsymbol{p}_i^h$, it is centered at $\boldsymbol{p}_i^h$ and has a size γ proportional to the size of the human bounding box. We adopt RoI-Align [39] to generate local patch features and rescale them to a resolution of $R_p \times R_p$. We apply the same operations to object keypoints $\boldsymbol{f}_o^p = \left\{\boldsymbol{f}_o^{p1} \ldots, \boldsymbol{f}_o^{pN}\right\}$ to generate their local patch features as well. For the sake of simplicity, we denote the extracted patch features of humans and objects as $\boldsymbol{f}^{p'} = \left\{\boldsymbol{f}^{p'1} \ldots, \boldsymbol{f}^{p'2N}\right\}$. In addition, we embed each keypoint as positional encodings to its corresponding patch. By doing this, the model can capture more detailed spatial relationships and configurations within each human-object pair. It also ensures a richer representation of the data, allowing the model to make more context-aware predictions of specific interaction types.

We then represent patch features of each human-object pair, integrated with their corresponding positional encodings, as a sequence of queries $\hat{\boldsymbol{q}} = (\hat{\boldsymbol{q}}_1, \ldots, \hat{\boldsymbol{q}}_{2N})$, keys $\hat{\boldsymbol{k}} = \left(\hat{\boldsymbol{k}}_1, \ldots, \hat{\boldsymbol{k}}_{2N}\right)$, and values $\hat{\boldsymbol{v}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_{2N})$. Following the self-attention mechanism [40], each patch is computed by aggregating all values weighted with attention, and an attended patch feature is represented as follows:

$$\boldsymbol{f}_i^{\hat{p}} = \sum_j \alpha_{ij} \left(\boldsymbol{W}_{\hat{v}} \hat{\boldsymbol{v}}_j\right), \tag{5}$$

where each $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_j \exp(e_{ij})}$ is the normalized attention weight with softmax. Here the primary attention weight e_{ij} is the scaled dot-product between each key $\hat{\boldsymbol{k}}$ and query $\hat{\boldsymbol{q}}$:

$$e_{ij} = \frac{\left(\boldsymbol{W}_{\hat{q}} \hat{\boldsymbol{q}}_i\right)^T \left(\boldsymbol{W}_{\hat{k}} \hat{\boldsymbol{k}}_j\right)}{\sqrt{d_{key}}}, \tag{6}$$

note that $\boldsymbol{W}_{\hat{q}}$, $\boldsymbol{W}_{\hat{k}}$, $\boldsymbol{W}_{\hat{v}}$ are learnable embedding matrices, and d_{key} is the embedding dimension of keys.

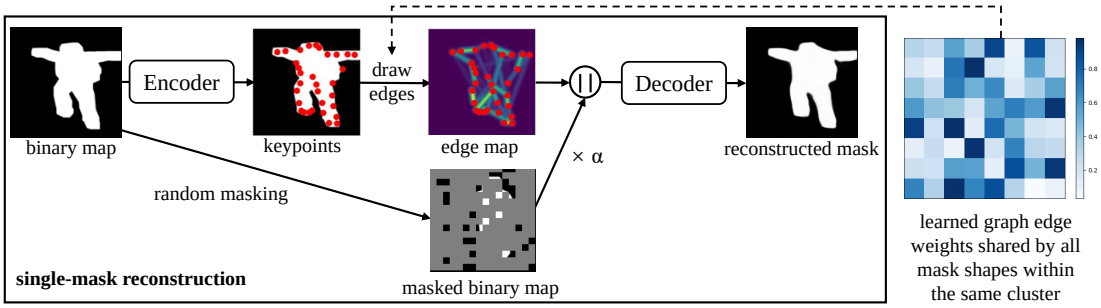

Fig. 4. Overview of the self-supervised keypoints learning framework (UniPointNet). Given an object segmentation, we detect the keypoints with learnable graph edge weights by reconstructing its binary mask. The edge weights are represented by a color matrix and are shared across segmentation masks within clusters of similar shapes and structures. The masked segmentation binary map provides minimal appearance information, forcing the network to focus on learning keypoints that are important for representing the structure and shape of an object.

The attended local part feature for a human-object pair is calculated by concatenating all patches:

$$\boldsymbol{f}^{\hat{p}} = \boldsymbol{f}_1^{\hat{p}} \textcircled{u} \boldsymbol{f}_2^{\hat{p}} \textcircled{u} \ldots \textcircled{u} \boldsymbol{f}_{2N}^{\hat{p}}. \tag{7}$$

Finally, each interaction query $\boldsymbol{q} \in Q$ is represented by the fusion of instance-level interactiveness features and the attended part features:

$$\boldsymbol{q} = \boldsymbol{f}_{ho} \textcircled{u} \boldsymbol{f}^{\hat{p}}. \tag{8}$$

They are fed into the structure-aware Transformer [19] for HOI classification.

D. Training and Inference

We follow the training and inference procedure of the STIP [19]. The KIP module is optimized with focal loss (FL) [41]:

$$L_{interactiveness} = \frac{1}{\sum_{i=1}^{\hat{N}} z_i} \sum_{i=1}^{\hat{N}} FL\left(\hat{z}_i, z_i\right), \tag{9}$$

where $\hat{N}$ is the number of sampled human-object pairs, $z_i \in \{0, 1\}$ denotes the existence of ground-truth interaction, and $\hat{z}_i$ is the predicted interactiveness score. For each of the output human-object pair of KIP, the focal loss is also used as the multi-label classification loss to train the possible interactions:

$$L_{class} = \frac{1}{\sum_{i=1}^{\hat{N}} \sum_{j=1}^{\hat{C}} y_{ij}} \sum_{i=1}^{\hat{N}} \sum_{j=1}^{\hat{C}} FL\left(\hat{y}_{ij}, y_{ij}\right), \tag{10}$$

where $\hat{C}$ is the number of interaction classes, $y_{ij} \in \{0, 1\}$ indicates the ground-truth interaction class, and $\hat{y}_{ij}$ is the predicted probability of j-th interaction class. The overall training objective of our GeoHOI integrates the above interactiveness loss and the interaction classification loss:

$$L_{GeoHOI} = L_{interactiveness} + L_{class}. \tag{11}$$

IV. Self-Supervised Object Keypoint Detection

As a core of our model, we propose leveraging keypoints as fine-grained geometric features of both humans and objects, to facilitate HOI prediction, but it is essential to detect these keypoints before utilizing them. Existing keypoints detection models typically focus on a single object class (e.g., human pose estimation [42]) rather than common objects. The difficulties lie in the complexity of distinct spatial structures and appearance distributions exhibited by various objects and the limited annotation availability. In addition, there is very limited work in detecting keypoints across different objects in HOI [10] due to a large number of object categories (e.g., 80 common object categories in MS-COCO [43]) and occlusions.

To address this challenge, we propose UniPointNet which can detect keypoints for arbitrary objects. We employ the self-supervised keypoints learning framework of AutoLink [44]. While AutoLink was proposed to learn keypoints for single object classes, our goal is to detect keypoints across all classes present in the HOI task. To this end, we make two key changes to AutoLink. First, we feed object segmentation masks into the network instead of RGB images. This eliminates the appearance variations across different object classes, simplifying their appearance distribution. As a result, the network can focus on learning object shapes and structures. Second, instead of using an individual edge graph with shared graph weight to align all samples, we opt for a set of edge graphs with different graph weights, aligning samples within their respective clusters. This design accommodates object masks with significant variations, thus allowing the network to detect keypoints across a diverse range of object categories.

Using such a network to detect keypoints for humans and all the other object classes is advantageous. First, it unifies keypoints detection for different object classes within a single network, which is more applicable in real-world applications in which diverse object types are often involved. Second, it ensures the consistency of keypoints distribution across different object categories including humans, resulting in a unified and consistent keypoints representation that facilitates network learning. Third, unlike the common keypoints representation in occluded cases (e.g., zeros for occluded or invisible joints of a human), all the detected keypoints in our UniPointNet contribute to the representation of an object's shape. This guarantees a more robust keypoints representation when objects are partially visible.

A. Architecture of UniPointNet

An overview of UniPointNet is shown in Fig. 4. Given an object segmentation binary map $\boldsymbol{B} \in \mathbb{R}^{H \times W \times 1}$ with a height

of H and a width of W, our goal is to learn a set of keypoints $\kappa = \left\{\boldsymbol{k}_i | i = 1, 2, 3, \ldots, N; \boldsymbol{k}_i \in [0,1] \times [0,1] \subset \mathbb{R}^2\right\}$, where N is the number of keypoints. As per [44], keypoints are detected by the encoder with ResNet and upsampling, and each pair of keypoints is connected with a differentiable edge [45]. This kind of graph connectivity defines a unique structure of a group of objects with similar shapes that share the same cluster label, learned in a self-supervised manner. The edge map $\boldsymbol{E} \in \mathbb{R}^{H \times W}$ is concatenated with the masked binary map $\boldsymbol{B}_m \in \mathbb{R}^{H \times W \times 1}$ along the channel dimension, and fed into the decoder to obtain the reconstructed segmentation binary map $\boldsymbol{B'}$. Detailed encoder and decoder network architectures can be referred to [44].

B. Segmentation Structure Representation

Here, we introduce keypoints representation and the edge map generation. $\mathcal{H} = \left\{\boldsymbol{h}_i | i = 1, 2, 3, \ldots, N; \boldsymbol{h}_i \in \mathbb{R}^{H \times W}\right\}$ is the N heatmaps generated by the Encoder from the input mask. The keypoint $\boldsymbol{k}_i$ is obtained by the differentiable soft-argmax function,

$$\boldsymbol{k}_i = \sum_{\boldsymbol{p}} \psi(\boldsymbol{h}_i)\boldsymbol{p}, \tag{12}$$

where $\psi(\boldsymbol{h}_i)$ is the $Softmax$ operation on a single heatmap $\boldsymbol{h}_i$, defined as,

$$\psi(\boldsymbol{h}_i) = \frac{exp(\boldsymbol{h}_i(\boldsymbol{p}))}{\sum_{j=1}^{N} exp(\boldsymbol{h}_j(\boldsymbol{p}))}, \tag{13}$$

where $\boldsymbol{p}$ is normalized pixel coordinates.

According to [44], a differentiable edge map $\boldsymbol{E}_{ij}$ is generated for any two keypoints $\boldsymbol{k}_i$ and $\boldsymbol{k}_j$, by assigning a value of 1 to pixels on the edge connecting the keypoints. For other pixels, their values decrease exponentially based on their distance to the line. The edge map $\boldsymbol{E}_{ij}$ is a Gaussian that extends along the line [45], and it is formally expressed as,

$$\boldsymbol{E}_{ij}(\boldsymbol{p}) = \exp\left(d_{ij}^2(\boldsymbol{p})/\sigma^2\right), \tag{14}$$

where the hyperparameter σ controls the thickness of the line, and $d_{ij}^2(\boldsymbol{p})$ is the L_2 distance between the pixel $\boldsymbol{p}$ and the line from keypoints $\boldsymbol{k}_i$ and $\boldsymbol{k}_j$. According to the location of the pixel $\boldsymbol{p}$, i.e., before the starting keypoint $\boldsymbol{k}_i$, between the starting keypoint $\boldsymbol{k}_i$ and the ending keypoint $\boldsymbol{k}_j$, or after the ending keypoint $\boldsymbol{k}_j$, it is defined as,

$$\begin{aligned} \boldsymbol{d}_{ij}(\boldsymbol{p}) &= \begin{cases} \|\boldsymbol{p} - \boldsymbol{k}_i\|_2 & \text{if } t \leq 0, \\ \|\boldsymbol{p} - (\boldsymbol{k}_i + t\boldsymbol{k}_j)\|_2 & \text{if } 0 < t < 1, \\ \|\boldsymbol{p} - \boldsymbol{k}_j\|_2 & \text{if } t \geq 1, \end{cases} \\ where \quad t &= \frac{(\boldsymbol{p} - \boldsymbol{k}_i) \cdot (\boldsymbol{k}_j - \boldsymbol{k}_i)}{\|\boldsymbol{k}_i - \boldsymbol{k}_j\|_2^2}. \end{aligned} \tag{15}$$

The final edge map $\boldsymbol{E} \in \mathbb{R}^{H \times W}$ is obtained by taking the maximum at each pixel of all the heatmaps,

$$\boldsymbol{E}(\boldsymbol{p}) = \max_{ij} w_{ij} \boldsymbol{E}_{ij}(\boldsymbol{p}), \tag{16}$$

where w_{ij} is a learnable edge weight. As explained in [44], opting for the maximum value at each pixel helps untangle the edge weights from the convolution kernel weights and generates better performance.

C. Segmentation Reconstruction

The masked segmentation $\boldsymbol{B}_m$ is obtained by randomly masking out 90% of the input segmentation. It is then concatenated with the edge map and is fed into the decoder to reconstruct the original segmentation,

$$\boldsymbol{B'} = Decoder(\alpha \boldsymbol{B}_m \textcircled{\scriptsize ||} \boldsymbol{E}), \tag{17}$$

where $\textcircled{\scriptsize ||}$ denotes concatenation along the channel dimension and the parameter α is a learnable factor that adjusts for the variation in edge weight magnitude during training and is initialized to 1. The L_1 loss and VGG perceptual loss [46] are used to minimize the difference between the original segmentation and the reconstructed one,

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} (|B_i - B_i'| + \|\Gamma(B_i) - \Gamma(B_i')\|_2^2), \tag{18}$$

where M represents the total number of examples, and Γ indicates the feature extractor, i.e., the VGG network.

V. Experiments

In this section, we introduce HOI benchmark datasets V-COCO [23] and HICO-DET [24], followed by experimental settings and implementation details. We then evaluate our proposed model against state-of-the-art approaches and provide insights on per-class performance by comparing it with the backbone STIP. Finally, we present ablation studies on the selection of the number of keypoints and the impact of individual component designs of our model.

A. Datasets

V-COCO is a popular HOI detection dataset and is a subset of MS-COCO [43] including 29 different action classes. It consists of 10,346 images, with 2533 images for training, 2867 images for validating, and 4946 images for testing. Following the settings in previous works [17], [19], [10], we apply the Average Precision (AP_{role}) metric over 24 interactions for the evaluation. Five actions are omitted as one of them has limited samples and the other four have no object associated with humans. Two types of AP_{role} (i.e., $AP_{role}^{\#1}$ and $AP_{role}^{\#2}$) are reported under different scenarios with different scoring criteria for cases where objects are occluded. Concretely, in the scenario of $AP_{role}^{\#1}$, the occluded object bounding box must be predicted as empty, i.e., $[0, 0, 0, 0]$. In contrast, in scenario $AP_{role}^{\#2}$, the occluded object is ignored. A human-object pair is considered a true positive if the predicted bounding boxes for both the human and the object have an Interaction-over-Union (IoU) ratio greater than 0.5 with their corresponding ground-truth annotation and the interaction category is accurate.

HICO-DET is a larger HOI detection dataset consisting of 47,051 images with 37,535 training and 9,515 testing images. It has 600 annotated human-object interactions and covers the same 80 object categories in MS-COCO [43]. We follow previous works [12], [19] and report in two different settings, i.e., *Default* and *Known Object*. The *Default* setting represents the evaluation of AP across all testing images, whereas the *Known Object* setting calculates the AP of each object solely

TABLE I
PERFORMANCE COMPARISON WITH END-TO-END METHODS OF MAP ON V-COCO AND HICO-DET. THE BEST RESULTS ARE MARKED IN **BOLD** AND THE SECOND BEST RESULTS ARE MARKED WITH UNDERLINE.

Method	Published In	Backbone	V-COCO		HICO-DET					
					Default			Known Object		
			$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare	Full	Rare	Non-Rare
UnionDet [47]	ECCV 2020	R50-FPN	47.5	56.2	17.58	11.72	19.33	19.76	14.68	21.27
IPNet [14]	CVPR 2020	HG-104	51.0	-	19.56	12.79	21.58	22.05	15.77	23.92
GGNet [48]	CVPR 2021	HG-104	54.7	-	29.17	22.13	30.84	33.50	26.67	34.89
HOTR [17]	CVPR 2021	R50	55.2	64.4	23.46	16.21	25.60	-	-	-
QPIC [16]	CVPR 2021	R50	58.8	61.0	29.07	21.85	31.23	31.68	24.14	33.93
DSSF [49]	IEEE TIM 2022	HG-104	57.6	-	25.23	18.72	27.17	28.53	21.68	30.57
MSTR [35]	CVPR 2022	R50	62.0	65.2	31.17	25.31	32.92	34.02	28.83	35.57
ERNet [29]	IEEE TIP	EfficientNet	64.2	-	31.57	26.76	33.10	-	-	-
MUREN [36]	CVPR 2023	R50	68.8	71.0	32.87	28.67	34.12	35.52	30.88	36.91
STIP [19]	CVPR 2022	R50	66.0	70.7	28.81	27.55	29.18	32.28	31.07	32.64
STIP* [19]	CVPR 2022	R50	-	-	32.22	28.15	33.43	35.29	31.43	36.45
GeoHOI (Ours)	-	R50	67.8	73.3	30.07	29.72	30.13	33.36	32.97	33.43
GeoHOI* (Ours)	-	R50	**69.4**	**74.4**	**35.05**	**33.01**	**35.71**	**37.12**	**34.79**	**37.97**

TABLE II
PERFORMANCE COMPARISON WITH TWO-STAGE METHODS OF MAP ON V-COCO AND HICO-DET. THE NOTATION IS THE SAME AS IN TABLE I.

Method	Published In	Backbone	V-COCO		HICO-DET					
					Default			Known Object		
			$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare	Full	Rare	Non-Rare
InteractNet [25]	CVPR 2018	R50-FPN	40.0	48.0	9.94	7.16	10.77	-	-	-
TIN [37]	CVPR 2019	R50	48.7	-	17.22	13.51	18.32	19.38	15.38	20.57
DRG [50]	ECCV 2019	R50-FPN	51.0	-	24.53	19.47	26.04	27.98	23.11	29.43
FCMNet [51]	ECCV 2020	R50	53.1	-	20.41	17.34	21.56	22.04	18.97	23.12
IDN [52]	NeurIPS 2020	R50	53.3	60.3	23.36	22.47	23.63	26.43	25.01	26.85
iHOI [53]	IEEE TMM 2020	R50-FPN	45.8	-	13.39	9.51	14.55	-	-	-
ACP++ [54]	IEEE TIP 2021	R152	53.2	-	18.90	16.80	19.52	24.78	23.87	25.05
HRNet [55]	IEEE TIP 2021	R152	53.1	-	21.93	16.30	23.62	25.22	18.75	27.15
IPGN [56]	IEEE TIP 2021	R50-FPN	53.8	-	21.26	18.47	22.07	-	-	-
CP-HOI [57]	IEEE TPAMI 2022	R50	50.4	-	19.42	13.98	20.91	22.01	15.73	22.80
UPT [58]	CVPR 2022	R101-DC5	61.3	67.1	32.62	28.62	33.81	36.08	31.41	37.47
ViPLO [31]	CVPR 2023	ViT	62.2	68.0	**37.22**	**35.45**	**37.75**	**40.61**	**38.82**	**41.15**
GeoHOI (Ours)	-	R50	67.8	73.3	30.07	29.72	30.13	33.36	32.97	33.43
GeoHOI* (Ours)	-	R50	**69.4**	**74.4**	35.05	33.01	35.71	37.12	34.79	37.97

on images that contain that object class. We report the AP for each setting over three different sets of HOI categories based on the number of training samples, i.e., **Full** (all 600 HOI categories), **Rare** (138 HOI categories that have less than 10 training samples), and **Non-Rare** (462 HOI categories with at least 10 training samples).

B. Implementation Details

To train UniPointNet, we extract object segmentation masks from the COCO dataset [43]. Masks with a ratio less than 0.2 relative to the image are discarded, since we aim to learn object shapes and structures and these tiny masks do not contain enough pixels to compute shape features. We finally collected a total of 50,238 training samples. We then apply ResNet to group all samples into 100 clusters with K-means clustering. Each cluster is associated with a unique set of graph weights during training, aligning the shape of samples within that cluster. Following AutoLink [44], the network is trained for 20k iterations with Adam optimizer, a learning rate of 10^{-4}, a batch size of 64, and the edge thickness of $\sigma^2 = 5e-5$. During inference, an input sample is assigned a cluster label based on its distance to each cluster centroid. Subsequently, its keypoints are detected using the corresponding graph weights. The number of keypoints can range from 4 to 48.

We adopt the object detector Panoptic DETR [18] pre-trained over MS-COCO, for both object bounding box detection and segmentation. It provides segmented inputs to UniPointNet, thereby enabling us to seamlessly incorporate UniPointNet into GeoHOI. The UniPointNet is then utilized as a pre-trained component of GeoHOI, detecting HOIs in an end-to-end manner. The backbone RestNet-50 is used for image feature extraction. We present results for two variations of our proposed method: GeoHOI and GeoHOI*. GeoHOI is trained for the HOI detector only with frozen parameters in Panoptic DETR, whereas GeoHOI* is trained with joint fine-tuning of both the object detector and HOI detector in an alternate manner. In the experiments, same as STIP, we output top-32 interactive human-object pairs of the Keypoints-aware Interactiveness Prediction module. Following the previous practice in [8], the RoI-Align in PAM is set to a resolution of $R_p = 5$, and the size of human and object patches is $\gamma = 0.1$ of their respective instance bounding box height and width. Then all the patch features are scaled to 5×5. The whole architecture is trained for 30 epochs over a single NVIDIA A100 GPU with a mini-batch size of 6, initial learning rate 5×10^{-5}, and AdamW optimizer.

C. Comparisons with State of the Art

We evaluate the performance of GeoHOI and compare it with state-of-the-art models, including methods that use geometric features of both humans and objects.

Table I shows the performance comparison with end-to-end methods. For V-COCO, our method beats all existing end-to-end methods by a large margin in both scenarios. In particular,

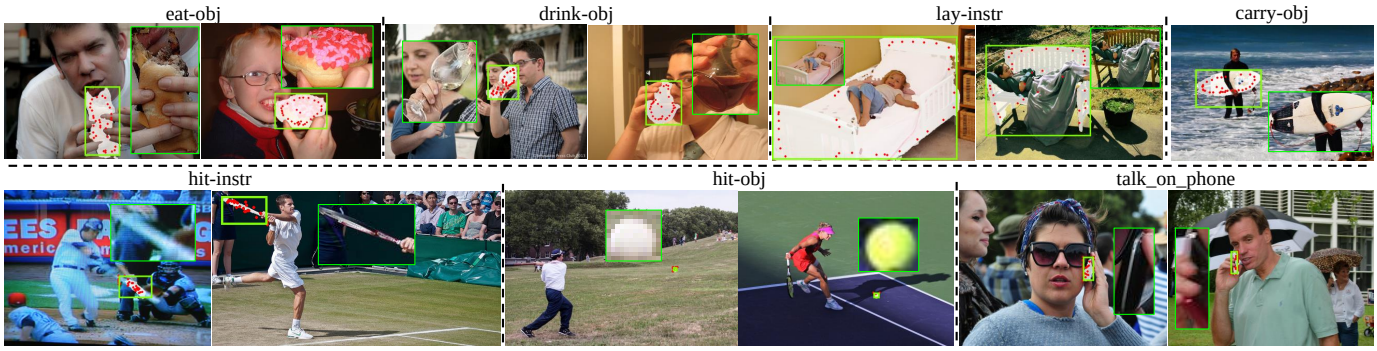

Fig. 5. Qualitative results. The upper row showcases the effectiveness of the keypoints representation, while the lower row depicts failure cases.

compared with MUREN [36], which is the previous state-of-the-art method, GeoHOI achieves a significant performance gain of 0.6 mAP in $AP^{\#1}_{role}$and 3.4 mAP in $AP^{\#2}_{role}$. For HICO-DET, GeoHOI achieves consistent performance gains and surpasses all the previous state-of-the-art methods. These results indicate our method's effectiveness in capturing the holistic cross-instance cues between humans and objects using their keypoints through graph convolutional networks and enhancing interaction query representations with local patches.

In table II, we compare GeoHOI against two-stage methods. Our GeoHOI outperforms all the existing two-stage methods on V-COCO. Compared with the latest method ViPLO [31], it obtains large performance improvements of 7.2 mAP in $AP^{\#1}_{role}$and 6.4 mAP in $AP^{\#2}_{role}$. This is mainly because most of these two-stage methods use CNNs or vanilla Transformers for HOI classification, leading to limited model capacity or prior knowledge. For HICO-DET, we also achieve comparable performance to previous state-of-the-art methods. Compared to ViPLO, the performance gain is not as noticeable as on V-COCO. Considering the complexity and end-to-end nature of GeoHOI, we follow STIP to use the lightweight ResNet for image feature extraction, while ViPLO employs the advanced Transformer backbone ViT [59] in their first stage of object detection. We believe ViT has a larger capacity than ResNet and is superior for handling the larger HICO-DET.

TABLE III
COMPARISON OF PERFORMANCE WITH OBJECT-STRUCTURE-AWARE METHODS ON V-COCO.

Method	Published In	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
SGCN4HOI [10]	IEEE SMC 2022	53.1	57.9
Liu et al. [60]	Pattern Recognition 2022	52.3	-
HOKEM [34]	IEEE ICIP 2023	54.6	59.7
ObjectPart [33]	Pattern Recognition 2023	62.5	-
GeoHOI (Ours)	-	**67.8**	**73.3**

Table III compares our results with existing methods using human and object keypoints on V-COCO. We omit comparison on HICO-DET because most of these methods did not provide results on this dataset. For fairness, we compare GeoHOI without fine-tuning the object detector against these methods. The proposed GeoHOI outperforms all of them by a marked margin in both $AP^{\#1}_{role}$(5.3 mAP) and $AP^{\#2}_{role}$ (13.6 mAP). Demonstrating the effectiveness of GeoHOI, i.e., by taking advantage of both the advanced Transformer architecture and the fine-grained geometric keypoints, it boosts HOI detection.

TABLE IV
PER-CLASS $AP^{\#1}_{role}$ PERFORMANCE COMPARISONS WITH STIP.

HOI Class	STIP [19]	GeoHOI
hold-obj (#pos = 3608)	56.07	**57.68** (↑ 0.71)
sit-instr (#pos = 1916)	56.44	**58.99** (↑ 2.55)
ride-instr (#pos = 556)	75.35	**75.52** (↑ 0.17)
look-obj (#pos = 3347)	**46.09** (↑ 0.88)	45.21
hit-instr (#pos = 349)	**82.21** (↑ 2.91)	79.30
hit-obj (#pos = 349)	**78.22** (↑ 3.43)	74.79
eat-obj (#pos = 521)	65.48	**74.52** (↑ 9.04)
eat-instr (#pos = 521)	77.44	**80.65** (↑ 3.21)
jump-instr (#pos = 635)	**82.37** (↑ 1.09)	81.28
lay-instr (#pos = 387)	61.50	**69.70** (↑ 8.2)
talk_on_phone (#pos = 285)	**57.30** (↑ 2.75)	54.55
carry-obj (#pos = 472)	39.56	**48.97** (↑ 9.41)
throw-obj (#pos = 244)	55.25	**56.22** (↑ 0.97)
catch-obj (#pos = 246)	**53.59** (0.98)	52.61
cut-instr (#pos = 269)	57.00	**57.57** (↑ 0.57)
cut-obj (#pos = 269)	70.35	**72.33** (↑ 1.98)
work_on_comp (#pos = 410)	75.80	**79.03** (↑ 3.23)
ski-instr (#pos = 424)	**55.03** (↑ 0.83)	54.20
surf-instr (#pos = 486)	79.88	**86.51** (↑ 6.63)
skateboard-instr (#pos = 417)	**92.66** (↑ 1.59)	91.07
drink-instr (#pos = 82)	55.30	**65.20** (↑ 9.9)
kick-obj (#pos = 180)	73.89	**76.20** (↑ 2.31)
read-obj (#pos = 111)	**51.92** (↑ 0.89)	51.03
snowboard-instr (#pos = 277)	82.54	**84.30** (↑ 1.76)
Average	65.89	**67.81** (↑ 1.92)

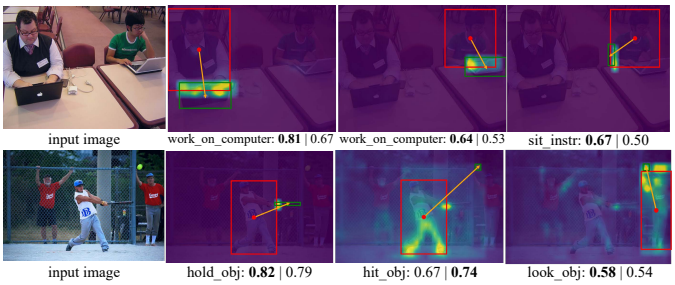

Fig. 6. Visualization of the attention in simple cases. The prediction scores of GeoHOI and STIP are shown (left: GeoHOI, right: STIP).

In Table IV, we report the per-class performance of GeoHOI and compare it with the backbone model STIP on V-COCO. We run the pre-trained checkpoint of STIP to obtain its per-class results since they are not provided in the original paper. We can see that GeoHOI outperforms the backbone STIP in the majority of classes, particularly in the "eat-obj", "lay-instr", "carry-obj", and "drink-instr" classes. From the upper row of Fig. 5, objects in these classes are generally either partially occluded with humans or quite large. For example, the objects in "drink-instr" are often occluded with human

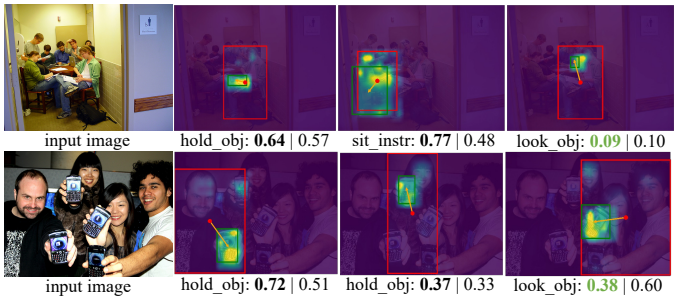

Fig. 7. Visualization of the attention in complex cases. The prediction scores of GeoHOI and STIP are shown (left: GeoHOI, right: STIP). The black color represents true positive interactions and the green color means true negative interactions.

hands. In these cases, we believe that keypoints can provide valuable information on the visible parts of these objects and how they are being interacted with the human, resulting in enhanced performance. Moreover, objects such as beds and surfboards, often associated with the "lay-instr" and "carry-obj" actions, are typically quite large. The detected keypoints can capture their shapes pretty well. As a result, it boosts the performance of interaction detection.

On the other hand, GeoHOI performs much worse than STIP in the "hit-instr", "hit-obj", and "talk_on_phone" classes. First, objects such as baseball bats and tennis rackets shown in the lower row of Fig. 5, typically appear in crowded scenes, leading to inaccurate object detection. Second, the inherent slender shape of baseball bats and the varying perspectives of tennis rackets hinder our UniPointNet from detecting their representative keypoints effectively. Third, balls associated with the "hit-obj" action and cell phones in the "talk_on_phone" action are too small. Thus, their masks have very small areas compared to the entire images, leading to noisy keypoints that harm interaction detection.

Fig. 6 shows qualitative results and compares GeoHOI with the backbone STIP. The top 3 interaction prediction probabilities of GeoHOI are visualized. The images show the variance in object sizes, human visibilities, and different interaction classes. First, the attention maps highlight different local regions for the same interaction category in the same image. For example, the hands are highlighted in different regions for action "work_on_computer" as shown in the first row. Second, when the human and object are far away from each other, they also gather a certain amount of information from their neighbourhood as illustrated in the second row, indicating both local and cross-instance cues are essential for HOI classification. We further showcase crowded scenes with multiple humans and objects in Fig. 7. GeoHOI shows higher confidence for true interaction actions "hold_obj" and "sit_instr" and less confidence for true negative interaction "look_obj", indicating its effectiveness. Overall, GeoHOI improves the backbone of STIP by predicting higher scores in various true interactions and lower scores in negative ones in most cases. More qualitative results of GeoHOI on complex scenes can be found in the supplementary material.

In Table V, we report the performance, number of parameters, and speed for training and inference on V-COCO

TABLE V
COMPARISON OF PERFORMANCE, NUMBER OF PARAMETERS, AND TIME FOR TRAINING AND INFERENCE ON V-COCO. WE CONDUCT BOTH THE TRAINING AND INFERENCE PROCESS WITH A BATCH SIZE OF 1 ON A SINGLE QUADRO RTX 5000 GPU. THE NUMBER OF PARAMETERS IS REPRESENTED IN MILLIONS (M). THE SPEED MEANS THE ELAPSED TIME (MS) FOR PROCESSING ONE IMAGE.

Method	$AP^{\#1}_{role}$ ↑	$AP^{\#2}_{role}$ ↑	#Params↓	Training↓	Inference↓
STIP [19]	66.0	70.7	13.2 M	143.2 ms	117.8 ms
GeoHOI	67.8	73.3	17.8 M	794.4 ms	588.6 ms

of STIP and GeoHOI for an objective comparison. GeoHOI outperforms STIP by a significant margin of 1.8 and 2.6 in terms of mAP in $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$ with a comparable number of parameters. GeoHOI's training and inference time is slower than STIP but remains within an acceptable one-second threshold. Compared to STIP, GeoHOI requires two additional modules (object segmentation and keypoint detection) executed sequentially, resulting in higher time consumption.

D. Ablation Studies

In this section, we analyze each GeoHOI design by discussing its possible variants on V-COCO to provide more insights. All experiments are carried out under the training setting of the pre-trained object detector with frozen weights.

TABLE VI
PERFORMANCE CONTRIBUTION ANALYSIS OF EACH COMPONENT IN GEOHOI ON V-COCO.

Method	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Baseline	65.2	69.8
+ KIP	66.2 (↑ 1.0)	71.0 (↑ 1.2)
+ PAM (w/o human patch)	66.2 (↑ 1.0)	71.1 (↑ 1.3)
+ PAM (w/o object patch)	66.5 (↑ 1.3)	71.4 (↑ 1.6)
+ PAM (w/o positional encodings)	66.3 (↑ 1.1)	71.1 (↑ 1.3)
+ PAM	66.7 (↑ 1.5)	71.9 (↑ 2.1)
+ KIP + PAM (GeoHOI)	**67.8** (↑ 2.6)	**73.3** (↑ 3.5)

Impact of Individual Components. We conduct ablation experiments by comparing different variants of GeoHOI in Table VI. We start with the baseline model (**Baseline**), which adopts the structure-aware HOI network introduced in [19], but with its object detector replaced by Panoptic DETR. Next, we extend the Baseline model by integrating our keypoint-aware graph convolution network into its interactiveness prediction module, incorporating the holistic graph features, yielding **Baseline + KIP** which demonstrates better performance. After that, we enhance the Baseline model with our Part Attention Module but without human patches. This variant of our model (**Baseline + PAM (w/o human patch)**) achieves better performance than both the Baseline model and Baseline + KIP. Another variant of our model (**Baseline + PAM (w/o object patch)**) shows even better performance. This indicates that the human patch features are more important than the object patch features. We believe this is because the rich poses of humans captured by keypoints are more beneficial for recognizing interactions, which aligns with the findings in [10]. To evaluate the benefits of using keypoints as positional encodings, we create (**Baseline + PAM (w/o positional encodings)**). It outperforms other variants, though

it is slightly worse than **Baseline + PAM** that incorporates both patch features and keypoints. Both Baseline + PAM (w/o positional encodings) and Baseline + PAM demonstrate the effectiveness of using local features with self-attention. Finally, when jointly upgrading the Baseline model with the keypoint-aware interactiveness prediction module and part attention module (i.e., our **GeoHOI**), it results in the best performance.

TABLE VII
PERFORMANCE COMPARISON BY USING THE DIFFERENT NUMBER OF KEYPOINTS (N) ON V-COCO.

# of keypoints (N)	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
4	65.6	70.3
8	66.6	71.5
16	66.8	71.7
32	**67.8**	**73.3**
48	67.0	71.9

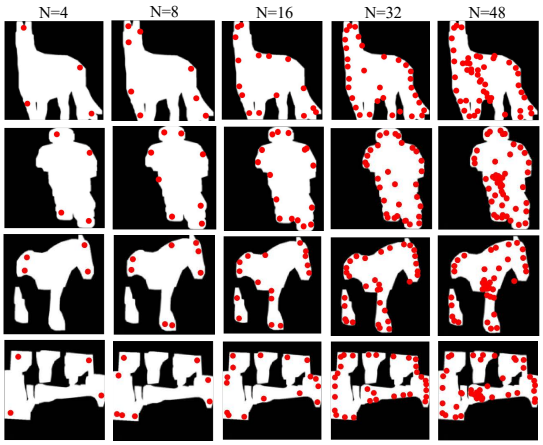

Fig. 8. Examples illustrating different numbers of keypoints. When there are very few keypoints, they only represent the basic shape of an object. For example, with $N = 4$, only edge corners are represented. With abundant keypoints, the representation can be redundant.

Effect of Different Numbers of Keypoints. Here, we vary N from 4 to 48 to demonstrate the relationship between the performance and the select keypoints number N. Table VII shows the quantitative ablation tests, and the best performance is obtained when N is 32. The increasing number of keypoints (until $N = 32$) can generally boost the performance. This is expected since more details can be captured with more keypoints. For example, when $N = 4$, only the upper part of the horse is modelled as shown in the first column of the third row in Fig. 8. In addition, when the object's mask is separated into multiple fragments due to occlusion (as seen in the last two rows in Fig. 8), a higher number of keypoints can span more of these segments, generating a more accurate shape representation of an object. However, when N is greater than 32, i.e., 48, the performance decreases. We speculate that too many keypoints might introduce more noise, leading the model to overfit, which results in affecting its generalization. As such, we have empirically selected N to be 32.

To evaluate the effectiveness of UniPointNet, in Table VIII, we compare it with the existing skeleton-based keypoint representation (Skeletal Keypoint) for HOI detection [10], which also utilizes object segmentation. GeoHOI (UniPointNet) means the keypoints in GeoHOI are detected by our proposed UniPointNet, and GeoHOI (Skeletal Keypoint) represents that the keypoints are obtained from [10]. We can see that UniPointNet surpasses the skeletal keypoints in both $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$, showcasing the effectiveness of the proposed UniPointNet. The Skeletal Keypoint is a skeleton-driven method, it is only robust for articulated objects like humans and dogs, which demonstrate a clear and consistent structure of joints and parts. It is also limited when extracting keypoints from non-articulated objects such as pizzas and phones because of its skeletonization process. In contrast, UniPointNet is a shape-driven representation, making it robust to arbitrary shapes of objects.

TABLE VIII
EFFECT OF UNIPOINTNET IN GEOHOI ON V-COCO.

Method	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
GeoHOI (UniPointNet)	**67.8**	**73.3**
GeoHOI (Skeletal Keypoint [10])	66.7	71.3

E. A Case Study in Post-Disaster Rescue with UAVs

The proposed HOI detector (GeoHOI) has a wide range of applications, including vision-based instrumentation and measurement. To showcase the generalization of our GeoHOI and evaluate its performance in real-world applications relevant to instrumentation and measurement, we conducted a case study in Post-Disaster Rescue with unmanned aerial vehicles (UAVs) on the PDD dataset [61]. It was collected from real-world ruins, including various post-disaster scenes, from multiple angles of UAVs and different distances and resolutions. Disasters include natural calamities such as earthquakes and outdoor rescue scenarios, among others. It consists of 832 training, 100 validation, and 100 testing images.

By conducting the experiments on the PDD dataset, we evaluate our GeoHOI for human detection task and compare it with the baseline methods. GeoHOI is designed for HOI detection, outputting triplets as $\langle human, interaction, object \rangle$. To evaluate it on human detection, we measure its outputs of detected human bounding boxes and ignore interaction and object bounding box predictions. To make a fair comparison between our methods and baselines in [61], we requested the PDD test set (100 images) from the authors for evaluation, and we also used the same evaluation metrics, i.e., average precision (AP), F1 score, recall, and precision. We directly apply the pre-trained GeoHOI and STIP (both are trained on V-COCO) on the test set of the PDD dataset. In addition, we conduct a qualitative analysis of HOI detection to showcase that the proposed method can further facilitate post-disaster rescue with UAVs. For example, detecting individuals in wheelchairs or needing medical assistance allows rescue teams to effectively prioritize rescue efforts such as aid and resources for those who need them most.

In Table IX, we compare the quantitative performance of HOI-based models, including our proposed GeoHOI and its backbone STIP and baselines proposed in [61]. With the default number of 32 output proposals in the Keypoint-aware Interactiveness Prediction (KIP) module, GeoHOI outperforms all the baselines on AP and achieves comparable precision,

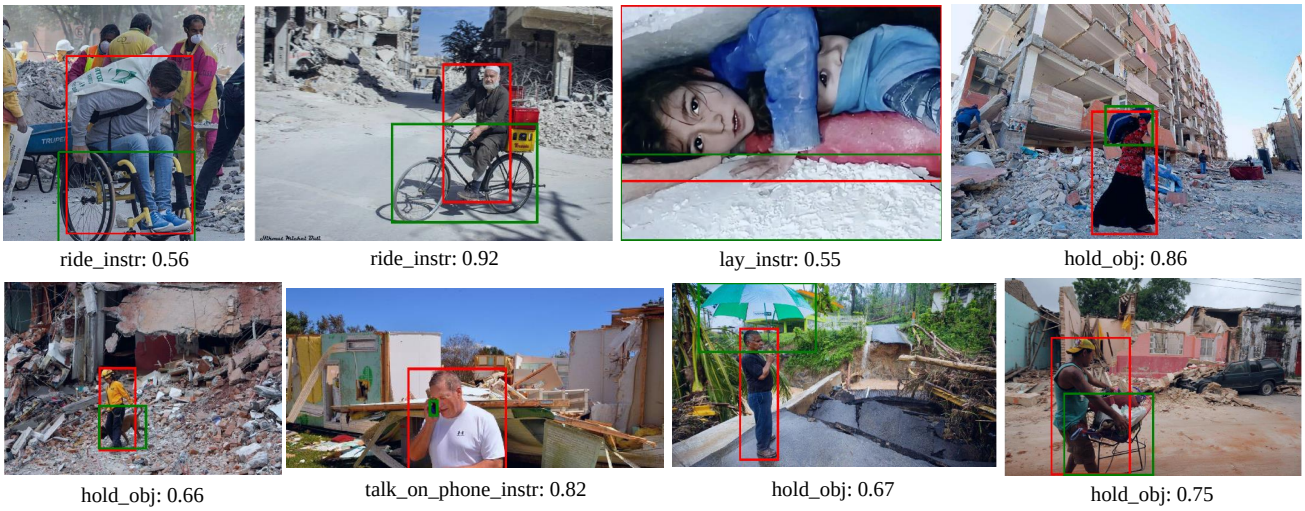

Fig. 9. Case study in post-disaster rescue. Different interactions and scenes are shown, and the top 1 interaction of each image is given.

TABLE IX
PERFORMANCE EVALUATION OF GENERALIZING GEOHOI IN HUMAN DETECTION ON THE PDD TEST SET. "DETECTION-BASED" DENOTES HUMAN DETECTION MODELS USED IN THE PDD DATASET, AND "HOI-BASED" REPRESENTS HOI DETECTION MODELS THAT ARE EVALUATED ON HUMAN DETECTION. NOTE THAT "K" IS THE NUMBER OF OUTPUT PROPOSALS FROM OUR KIP MODULE.

	Model	AP@0.5	F1	Recall	Precision
Detection-based [61]	YOLOv5s	84.15%	0.86	80.31%	93.58%
	YOLOv5m	84.36%	0.87	82.49%	92.58%
	YOLOv5l	84.25%	0.88	83.92%	93.45%
	im-YOLOv5s	80.82%	0.85	78.43%	92.59%
	im-YOLOv5m	83.32%	0.88	82.28%	94.14%
	im-YOLOv5l	84.38%	0.87	82.03%	93.33%
	YOLOv7	85%	0.86	82.56%	90.64%
	YOLOv7x	87.35%	0.89	85.88%	91.63%
	YOLOv8s	85.52%	0.85	83.33%	87.76%
	YOLOv8m	90.78%	0.89	85.66%	91.70%
	YOLOv8l	87.81%	0.88	84.05%	91.53%
	YOLO-NASs	86.08%	0.86	86.61%	85.27%
	YOLO-NASm	85.94%	0.86	86.38%	85.06%
	YOLO-NASl	87.26%	0.86	88.72%	84.13%
	DETR	87.89%	0.88	89.08%	87.24%
	DDETR	76.87%	0.64	84.23%	51.01%
	DAB-DETR	88.78%	**0.91**	88.42%	93.47%
	DN-DETR	87.17%	0.9	86.15%	**94.92%**
	DINO	91.03%	**0.91**	89.66%	92.86%
	faster R-CNN	88.52%	0.87	89.19%	84.62%
HOI-based	STIP	91.01%	0.80	71.51%	91.58%
	GeoHOI ($K = 32$)	**92.32%**	0.81	71.94%	92.63%
	GeoHOI ($K = 64$)	88.65%	0.88	89.41%	88.04%
	GeoHOI ($K = 100$)	84.64%	0.88	**91.48%**	86.49%

demonstrating its effectiveness in detecting humans in post-disaster scenes. STIP obtains similar performance in both AP and precision, and we believe the main reason is that the HOI-based detection systems can enhance human bounding box precision by leveraging contextual information (i.e., interactions between humans and objects) and joint optimization (i.e., optimizing the predictions of humans, interactions, and objects simultaneously). The integrated analysis of humans, objects, and their interactions refines human detection accuracy compared to these baselines designed alone for human detection. The lower performance on the F1 score and recall of GeoHOI and STIP indicate that the HOI-based systems have a higher missed detection rate. We think the KIP module that suppresses non-interactive human-object pairs is the primary cause since it can filter out humans who do not interact with objects, resulting in compromised performance in recall. To verify this, we increase the number of proposals (K) to 64 and 100, respectively. Recall significantly improves with the number of proposals and outperforms all the baselines at $K = 100$. This indicates our model's adaptability in balancing recall and precision by tuning the number of output proposals in our KIP module in practical applications.

In addition, we show the qualitative results of HOI detection in Fig. 9 to provide an in-depth analysis of how HOI detection can facilitate post-disaster rescue. GeoHOI demonstrates a varied performance across different scenarios. For instance, it predicts relatively high confidence scores in recognizing the interactions of "ride_instr" (riding a bicycle), "talk_on_phone_instr", and "hold_obj" where the scenes are less complicated. In contrast, it shows diminished confidence in more complex scenes, such as a person lying down in the rubble, or when the scene is crowded, e.g., the image in the first row and column. This indicates the challenges in detecting interactions in cluttered post-disaster scenes.

The qualitative results show that our proposed GeoHOI is able to detect different human interactions in post-disaster scenes, facilitating search and rescue operations. For instance, identifying individuals in wheelchairs or those lying on the ground enables rescue teams to prioritize medical attention. Observations of people using phones or riding bicycles provide crucial insights into the operational status of communication networks and the accessibility of various areas. Additionally, recognizing survivors holding onto pets or personal belongings allows rescue teams to provide not only necessities like food and water but also support for pet care and the safekeeping of valuables, enhancing the overall rescue operation.

VI. CONCLUSION

In this paper, we have proposed GeoHOI, an end-to-end Transformer-style model for detecting human-object interactions using fine-grained geometric keypoint features of humans and objects. We have also presented UniPointNet, a self-supervised framework that detects keypoints for arbitrary objects and enhances HOI performance. The KIP module

uses keypoints to mine cross-instance cues via a graph network, enhancing pairwise cues for optimizing the prediction of interactive human-object pairs. The PAM module uses self-attention on keypoint patches to discover informative local cues, facilitating the prediction of specific interaction categories. Extensive experimental results have shown that GeoHOI improves the backbone of STIP and achieves superior performance on public HOI benchmarks. We further demonstrated the advantages of using GeoHOI on human-centric applications such as the case study on post-disaster rescue. The presented UniPointNet also facilitates visual measurement tasks, including object pose estimation [4] and 3-D reconstruction [62].

The end-to-end GeoHOI is limited in training and analysis of the geometric features. For future research, the proposed geometric features can be employed in two-stage frameworks such as [31], facilitating more analysis and insights into the geometric context (e.g., relative keypoint distance) in HOI detection. As discussed in the experiments, our UniPointNet struggles with tiny or slender objects due to their limited spatial resolution in images, which hinders accurate shape reconstruction and keypoint detection. Future work could explore better keypoint representation for these objects, such as adaptively selecting the optimal numbers and locations of keypoints to represent objects in different sizes. Additionally, investigating how to incorporate semantic information in keypoint detection and evaluating the effect on HOI detection would also be valuable.

Furthermore, recent advancements in large language models, especially those with integrated vision-language capabilities such as CLIP [63], have demonstrated their effectiveness in zero-shot HOI detection [64], [65]. Given that annotating HOI triplets is challenging and rare HOIs are not learned as effectively as non-rare ones, it is worth further exploring the capabilities of large language models in the future to tackle the long-tail problem and zero-shot learning in HOI detection, facilitating real-world HOI applications.

ACKNOWLEDGEMENT

This research is supported in part by the EPSRC NorthHFutures project (ref: EP/X031012/1).

REFERENCES

[1] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: the rising trend of vision based measurement," *IEEE Instrumentation & Measurement Magazine*, vol. 17, no. 3, pp. 41–47, 2014.

[2] Q. Wu, Y. Wu, Y. Zhang, and L. Zhang, "A local–global estimator based on large kernel cnn and transformer for human pose estimation and running pose measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[3] T. Li and H. Yu, "Visual–inertial fusion-based human pose estimation: A review," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[4] Y. Zhang, G. Tian, and X. Shao, "Safe and efficient robot manipulation: Task-oriented environment modeling and object pose estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[5] Y. Wang, X. Wang, H. Yang, Y. Geng, H. Yu, G. Zheng, and L. Liao, "Mhagnn: A novel framework for wearable sensor-based human activity recognition combining multi-head attention and graph neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[6] M. Dogariu, L. Stefan, M. G. Constantin, and B. Ionescu, "Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios," in *13th International Conference on Communications, COMM 2020, Bucharest, Romania, June 18-20, 2020*. IEEE, 2020, pp. 157–160.

[7] Y. Y. Ghadi, M. Waheed, T. Al Shloul, S. A. Alsuhibany, A. Jalal, and J. Park, "Automated parts-based model for recognizing human–object interactions from aerial imagery with fully convolutional network," *Remote Sensing*, vol. 14, no. 6, p. 1492, 2022.

[8] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9468–9477.

[9] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 614–13 623.

[10] M. Zhu, E. S. L. Ho, and H. P. H. Shum, "A skeleton-aware graph convolutional network for human-object interaction detection," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2022, pp. 275–281.

[11] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[12] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, and J. Sun, "End-to-end human object interaction detection with hoi transformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 820–11 829.

[13] A. S. M. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, "What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5343–5353.

[14] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4115–4124.

[15] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 479–487.

[16] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 405–10 414.

[17] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 74–83.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[19] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 526–19 535.

[20] C. Xie, F. Zeng, Y. Hu, S. Liang, and Y. Wei, "Category query learning for human-object interaction classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 275–15 284.

[21] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu, "Mining cross-person cues for body-part interactiveness learning in hoi detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 121–136.

[22] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 72–90.

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

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

[25] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[26] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 646–14 655.

[27] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 401–417.

[28] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 319–13 327.

[29] J. Lim, V. M. Baskaran, J. M.-Y. Lim, K. Wong, J. See, and M. Tistarelli, "Ernet: An efficient and reliable human-object interaction detection network," *IEEE Transactions on Image Processing*, vol. 32, pp. 964–979, 2023.

[30] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 51–67.

[31] J. Park, J.-W. Park, and J.-S. Lee, "Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 152–17 162.

[32] S. Zheng, S. Chen, and Q. Jin, "Skeleton-based interactive graph network for human object interaction detection," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[33] L. Bai, F. Chen, and Y. Tian, "Automatically detecting human-object interaction by an instance part-level attention deep framework," *Pattern Recognition*, vol. 134, p. 109110, 2023.

[34] Y. Ito, "Hokem: Human and object keypoint-based extension module for human-object interaction detection," *arXiv preprint arXiv:2306.14260*, 2023.

[35] B. Kim, J. Mun, K.-W. On, M. Shin, J. Lee, and E.-S. Kim, "Mstr: Multi-scale transformer for end-to-end human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 578–19 587.

[36] S. Kim, D. Jung, and M. Cho, "Relational context learning for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2925–2934.

[37] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[38] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 843–851.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[42] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

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

[44] X. He, B. Wandt, and H. Rhodin, "Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints," in *NeurIPS*, 2022.

[45] D. Mihai and J. S. Hare, "Differentiable drawing and sketching," *CoRR*, vol. abs/2103.16194, 2021.

[46] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[47] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 498–514.

[48] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 234–13 243.

[49] D. Gu, S. Ma, and S. Cai, "Dssf: Dynamic semantic sampling and fusion for one-stage human–object interaction detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[50] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 696–712.

[51] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 248–265.

[52] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5011–5022, 2020.

[53] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423–1432, 2020.

[54] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Acp++: Action co-occurrence priors for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9150–9163, 2021.

[55] Y. Gao, Z. Kuang, G. Li, W. Zhang, and L. Lin, "Hierarchical reasoning network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8306–8317, 2021.

[56] H. Wang, L. Jiao, F. Liu, L. Li, X. Liu, D. Ji, and W. Gan, "Ipgn: Interactiveness proposal graph network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6583–6593, 2021.

[57] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2827–2840, 2022.

[58] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 104–20 112.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[60] L. Liu and R. T. Tan, "Human object interaction detection using two-direction spatial enhancement and exclusive object prior," *Pattern Recognition*, vol. 124, p. 108438, 2022.

[61] H. Song, W. Song, L. Cheng, Y. Wei, and J. Cui, "Pdd: Post-disaster dataset for human detection and performance evaluation," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–14, 2024.

[62] F. Yu, M. Liu, W. Chen, H. Wen, Y. Wang, and T. Zeng, "Automatic repair of 3-d neuron reconstruction based on topological feature points and an most-based repairer," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[64] S. Ning, L. Qiu, Y. Liu, and X. He, "Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 507–23 517.

[65] F. Zhang, L. Sheng, B. Guo, R. Chen, and J. Chen, "Sqa: Strong guidance query with self-selected attention for human-object interaction detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.