**Spatial-Net for Human-Object Interaction Detection**

**AHMED E. MANSOUR**, 1 AMMAR MOHAMMED, 2 HUSSEIN ABD EL ATTY ELSAYED, 1 (Member, IEEE), AND SALWA ELRAMLY, 1 (Life Senior Member, IEEE)

1Electronics and Electrical Communication Department, Faculty of Engineering, Ain Shams University, Cairo 11517, Egypt
2Computer Science Department, Faculty of Graduate Studies of Statistical Researches, Cairo University, Giza 12613, Egypt

Corresponding author: Ahmed E. Mansour (dotahmed@aol.com)

**ABSTRACT** Human-object interaction (HOI) detection is the detection of a human’s relationship with an object in still images and videos. The majority of HOI detection methods rely on appearance features as the primary feature for detecting the relationship between humans and objects. Furthermore, the model’s performance is affected by the abundance of false-positive pairs generated by the image’s non-interactive human-object pairs and human-object mis-grouping. In this paper, we propose “Spatial-Net”, a new HOI detection approach in still images. In the proposed approach, the HOI problem is divided into two main tasks, namely pair-prediction and global-rejection. In the pair-prediction task, the spatial relationship is adopted to predict the human-object interaction for each human-object pair using spatial features that contains spatial map which is a single channel image that represents human-object pairs including body parts and object masks, relative geometry features such as relative size, relative distance, and intersection-over-union between body part and objects, and weighted distance that is used as body part attention deterministic model. In the global-rejection task, an augmented model is employed to reject false positive pairs. We use the Hungarian matching technique to assign human-object pairs for each action and human-centric model to reject the non-interaction human-object pairs according to semantic co-occurrence between human and object. The experimental results on the V-COCO dataset demonstrate that the proposed Spatial-Net outperforms many state-of-the-art HOI models with less inference time.

**INDEX TERMS** Artificial neural network, deep neural network, learning challenges, visual relationship.

**I. INTRODUCTION**

Object detection is the technique of identifying instances of semantic objects (such as persons, buildings, or airplanes) in digital images and videos [1], [2], [3]. Deep learning has made major strides in a variety of computer vision applications, including object detection, action recognition, motion tracking, and human pose estimation [4], [5], [6]. Although deep learning techniques [7], [8], [9] achieve a great performance in detecting objects in complex scenes, object detection alone is not considered the key factor for understanding surrounding environments and improve, machine intelligence. To improve the capabilities of the machine to understand the environment, visual relationship detection is used to give a richer semantic understanding in images and videos by predicting the relationship between objects in form on (actor, action, object). Human-Object interaction (HOI) detection [10] is a visual relationship task aiming at detecting human and objects instances generating a relationship in form of (human, action, object). For instance (human, ride, bike) and (human, drink, cup). The main difference between visual relationship detection [11], [12], [13] and HOI is that we have human as a fixed actor in the relationship. HOI detection is crucial in high-level semantic understanding of human-centric scenes that is playing an important role in many applications such as intelligent surveillance systems and image analysis. As a result, the problem of HOI detection has piqued the interest of several researchers, and various
techniques have been proposed to solve it. However, the majority of the proposed HOI models perform poorly in comparison with object detection models. HOI detection is a challenging computer vision task due to the variation in human poses while performing the same actions, and humans might interact with the object with many actions, such as (human, hold, phone) and (human, talk on, phone) hold and talk are interacting with the same object “phone”. Another challenging task is detecting and recognizing fine-grained human-centric actions (e.g., eat cake and cut cake) where more than one human is interacting with the same object (e.g., throw and catch a ball) and more than one human is performing the same action (e.g., humans riding bikes in a race). Another additional challenge is the generation of false-positive pairs that are generated due to mis-grouping between human-object pairs and the existence of many non-interacting objects in the image.

Traditionally, the main methods used in HOI detection problems are sequential and parallel detectors. Both methods use appearance features as the main feature for detecting HOI which needs several CNN layers that consume time and resources. Sequential detectors performs the detection in two-stages [14], [15], [16], [17], [18]. In the first stage, all humans and objects are detected. The second stage is responsible to infer the relations between human-object pairs. These two stages, however, are considered time consuming processes and computationally expensive. Parallel detectors, on the other hand, use one-stage methods [16], [19], [20] to infer the interaction and hence are considered faster than sequential detectors, but their performance is low.

In this paper, we propose a new two-stage HOI approach in still images called “Spatial-Net”, which divides the HOI problem into two main tasks. The first task is to predict the HOI action between human-object pairs, while the second task is to reject non-interacting pairs in order to reduce false positives. The proposed Spatial-Net approach detects HOI interactions using only spatial relationships between human and object and achieves high inference speed when compared to one-stage models.

The pair prediction part adopted in this paper is done by combining relative spatial relations between human and object with a body parts spatial map that represents all of the body parts and the object in a single channel image, which represents the interaction map between the human body parts and objects. Furthermore, we employ statistical analysis to build a weighted distance vector for each action, which is used as an attention vector for action prediction. We use Hungarian Matching and “human actions co-occurrence” semantics to all generated pairings in the rejection model to reject non-interactive pairs. The relative size and distance between each body part and the object result in an approximate 3D (x, y, z axis) relationship between the human and the object. We adopt the relative distance to estimate the relationship in x, y dimensions and the relative size to estimate the relationship in z dimension. Furthermore, we estimate attention for parts that are involved in the action using Jaccard similarity (intersection-over-union (IoU)) between body part bounding boxes and object bounding boxes. The features that represent HOI are generated in a simple form and achieve good representation for HOI detection.

In FIGURE.1 (a), if we query action drink between person and wineglass, according to distance, it’s acceptable to drink from wineglass in the red box but not on wineglass in the blue box which estimates the relationship in x, y. However, in FIGURE.1 (b) if we query action ride between person and the red mask bike, we find that the relative size between person in the blue box and bike is accepted but the relative size between the person in the yellow box and bike is not acceptable because he is far from the bike which estimates the relationship is z-direction. FIGURE. 1 (c) illustrates the attention between human body parts and objects. If we query skateboard action, we find that in most cases there are IoU between the feet bounding box and the skateboard bounding box, but it’s not mandatory to find IoU between other body parts bounding boxes such as hands and the skateboard bounding box.

The interaction map between human and object represents reduced features that includes all body parts and objects as numeric values in a single channel image. The combination of the spatial interaction map and spatial geometry is used as a feature for a multi-modal machine learning model to predict HOI pairs individually in the image. To increase pair prediction confidence, we determine a weighted distance between human and object that gives each action a weight value according to its relationship with human body parts inferred from the relative distance analysis from the training data that increases model attention for the predicted action.

In the Global-Rejection part, in order to reduce the number of false positive pairs, we use Hungarian Matching [21] to filter out non-interacting objects with humans performing the same action in the image. Then, to reject non-interacting pair actions with different actions, we train a human-centric model with all of the predicted pairs associated with the human to reject the non-interacting objects according to
actions co-occurrence semantics. For instance, if the interacting objects with the human are skateboard as the highest scoring action and sit as a second action with another object, the human cannot perform both actions simultaneously. As a result, the human-centric model has been trained to reject the sit action while retaining the skateboard action.

This paper’s primary contributions are as follows: First, we split the HOI detection problem into two parts: pair prediction and global rejection. Second, we propose a new two-stage approach for HOI pair prediction that uses only spatial features in predicting HOI by leveraging the performance of three sub-models: spatial map, relative geometry, and weighted distance to produce human-object pair actions with a shorter inference time than one-stage models. Third, we employ a global rejection model to address the issue of false positive pairs created by incorrectly grouping human-object pairs and surrounding objects. Finally, we apply the proposed method on the V-COCO dataset [22]. In terms of performance and inference time, the experimental results reveal that the suggested technique outperforms most state-of-the-art models.

II. RELATED WORKS
Traditionally, there are two main approaches for HOI detection: a two-stage “sequential detectors” and one-stage “parallel detectors”. In two-stage methods, object detection is performed first to detect human-object pairs and then the interactions between them are inferred. However, in one-stage methods, object detection and interaction between humans and objects are predicted in parallel using parallel detectors. This section highlights the related work on those kinds of HOI detection methods and the features that are adopted to perform HOI detection. In addition, it focuses on the body part-based HOI methods and spatial features in HOI that are close to our proposed approach.

A. TWO-STAGE “SEQUENTIAL” METHODS
In most modern two-stage HOI detection systems, the first stage includes an object detector and the second stage includes an interaction classifier. In the first stage, a fine-tuned object detector is employed to obtain the bounding boxes and class labels for humans and objects. A multi-stream architecture is utilized in the second stage to infer the interactions between each human and object pair. In the previously mentioned multi-stream interaction classifier, there are typically three streams: human stream, object stream, and paired stream. Visual features for human and object boxes are normally encoded in both human and object streams [18] while visual features of action box are encoded in pair stream. The authors in [23] argued that the visual appearance of the object is typically not necessary for the interaction category. So, they replaced object visual features with word embedding. Where authors in [24] used word embedding in the human stream for feature augmentation in addition to visual features. To gain linguistic prior-guided channel attention and feature augmentation, PDNet [25] introduced word embedding for all streams. In multi-stream architecture, many researchers contributed to paired stream because it encodes the relationship visual features between humans and objects. Graph models also achieved good results in HOI prediction. For instance, GPNN model [26] used message passing in graph neural networks. To express the spatial relation, iCAN [18] proposed a two-channel binary image representation. The spatial relationship is enhanced in graph neural network models [27], [28], [29] to explicitly describe human-object interactions, which significantly increased the model’s representation capability. Although the two-stage methods achieve good results, they are normally time consuming due to the two sequential steps that are used to detect HOI.

B. ONE-STAGE “PARALLEL” METHODS
Interact-Net [30] is one of the first models that adopts one-stage method in prediction HOI but it needs cascaded inference. By adopting a novel concept of interaction point, PPDM [31] and IPNet [19] addressed HOI as a detection point problem and directly detected interactions in a one-stage method. Moreover, PPDM adopted the concept of Corner-Net [32] to perform both object detection and HOI detection in a unified method but they have a problem with detection ambiguity because the interaction points aren’t always the same distance from each other. In addition, they need a post-processing stage that is made by hand to connect the predicted interactions with the output from the object detector. The authors in UnionDet [20] adopted the concept of RetinaNet [33] to solve HOI detection problem as union box detection by adding an extra union branch for detecting union box inserted parallel to the standard object detection branch. One-stage methods pipeline becomes simpler, faster, more efficient, and easier to deploy for real-world applications as compared to two-stage techniques. One-stage approaches, on the other hand, still necessitate significant post-processing to group object detection results and interaction predictions. The attention mechanism and transformers are considered great revolutions that enhance the machine learning models’ performance in solving natural language processing and computer vision problems. In HOI detection, the transformer is used in several models, such as work in [34] which proposed an end-to-end HOI detection approach. However,
the authors in [35] enhanced HOI detection by upgrading the vanilla transformer with additional encoders for the holistically semantic structure among interaction proposals and the locally spatial structure of human-object within each interaction proposal respectively. The authors in [36] combined natural language supervision of interactions and embed them into a joint visual and text space to perform zero-shot HOI interaction detection.

C. HOI MODELS FEATURES
Most HOI detection models depend mainly on the visual features of the human, object, and interaction bounding boxes to train and infer the interaction between humans and different objects [20], [26], [30], [31]. However, visual features alone do not achieve good enough discrimination between different actions to solve the HOI detection problem. Because of this, more features are added to the HOI models in order to improve the detection results, such as human pose estimation and spatial correlation.

Pose estimation adds a discriminative feature to HOI model [16], [17], [23], [29], [37] to reduce the ambiguity of actions. For instance, ride action should be associated with the sitting pose. However, kick action should be associated with stand pose. The authors in [38] presented the human and object as volumes. The object is represented as a ball of a certain volume and, using SMPLify-X [39] 3D model body pose, head pose, and facial expression for humans are generated. The volume model achieves good results, but it is computationally expensive. Linguistic prior models [24], [27], [38], [40] use the word similarity to estimate a prior probability for human and object possible interactions models such as Word2Vec [41] and Glove [42]. For instance, the sentence person eats cake has a high probability to eat and cake while the sentence person eats laptop has very low probability between eat and laptop to occur.

D. BODY PARTS BASED HOI
Holistic human detection could be improved if detailed body parts are detected. The relationship and state of the body parts give more understanding of the human pose and interaction pattern with objects. The body parts are nodes in a graphical model as in [43] where authors used the graphical model with object features and pose to predict HOI. Where authors in [44] combined the whole body and parts features to improve action classifier accuracy. Moreover, the proposed method in [45] learned to focus on crucial parts, and their correlations for HOI recognition, and they achieved noticeable improvement in HOI results. Furthermore, by adopting fine-grained action semantics, in [46] human parts states are determined to estimate the activities based on part-level semantics.

E. SPATIAL FEATURES IN HOI
Spatial correlation adds a good cue for HOI detection [18], [47], [48], [49]. For instance, in the riding bike example shown in FIGURE1b, the human should be above the bike. However, if a human is talking on the phone, it should be close to his head.

Abstract spatial-semantic representation was aggregated with scene contextual features in [27] to solve HOI detection using dual graph neural networks.

Course layout and fine-grained layout spatial features were exploited in [23] to predict HOI, and the authors argued that the appearance features of objects did not affect the HOI prediction performance. Spatial clues for scene understanding are investigated in [50] and propose canonical spatial representation templates that indicate the power of spatial features in visual relationship applications and outperform many HOI state-of-the-art models.

III. PROPOSED SPATIAL-NET APPROACH
In this section, we explain the proposed Spatial-Net approach architecture and features. Spatial-Net approach as illustrated in FIGURE3 comprises two main sub-models: (1) pair prediction model and (2) global-rejection model. First, for a given image I, we apply feature extraction using ResNet-101 [51] that are shared backbone for instance segmentation model and body parts encoder model. Instance segmentation model, uses end-to-end object detection with transformers (DETR) [52] to retrieve objects masks $O = \{o_{m0}, o_{m1}, \ldots, o_{mn}\}$ and objects bounding boxes $O_b = \{o_{b0}, o_{b1}, \ldots, o_{bn}\}$ from the image where $n$ is the total number of objects and $o_b \in \mathbb{R}^4$. In parallel, we run human body parts encoder [53] to retrieve humans bounding boxes $H_b = \{h_{b1}, h_{b2}, \ldots, h_{bm}\}$ where $m$ is the total number of humans and $h_b \in \mathbb{R}^4$. For each human, 25 body parts masks $P_m = \{p_{m0}, p_{m1}, \ldots, p_{m24}\}$ and bounding boxes $P_b = \{p_{b0}, p_{b1}, \ldots, p_{b24}\}$ are generated. The index 0 represents the whole human mask and box, where the remaining indexes from 1 to 24 represent individual body parts. In a pair prediction model, the proposed HOI are generated by enumerating all pairs of candidate human and object bounding boxes. The generated output from spatial maps and geometric features is combined with the weighted distance in the fusion model to predict the final pair scores.

The last stage in the pair prediction model is action prior, which rejects the actions according to the object type. For instance, a human can ride bike but he cannot ride cell phone. So, after the last prediction, the action prior function [18] is applied to the final result to set the unaccepted action according to the object type to zero. The final predicted HOI pairs are processed by Global-Rejection Model to produce the final HOI proposals.

On the other hand, in the global-rejection model, the whole HOI predicted pairs in an image are processed to assign objects to humans according to prediction score by using bipartite matching and the human-centric rejection model.
FIGURE 3. Spatial-Net model architecture. For a given image, body parts encoder and instance segmentation model shared the backbone to generate spatial map and geometry features, then the features are fused with fusion model to generate HOI pairs. Global-Rejection model filter the HOI pairs using bipartite matching and human-centric rejection model to produce the final prediction.

FIGURE 4. Spatial map for sample actions.

1) SPATIAL ENCODER AND PREDICTOR
The spatial encoder adopts body parts detector and instance segmentation to create a spatial map that contains numeric values for each body part from 1 to 24 and the object value equals the object index in COCO dataset, [54] plus 25 (the max of body parts index), while 0 is reserved to map background. The intuition behind the spatial map is that the human body parts and object's shape, boundary, and value in the map can be used to discriminate between the different HOI interaction classes, and each interaction class has an approximately similar spatial layout. In addition, it does not include a large size of features. This map shows where each body part mask is in relation to other body parts. This shows the human pose that keeps human appearance features in a reduced form.

FIGURE 4 illustrates some actions that produced by the spatial map encoder.

The generated spatial map then processed spatial predictor that contains 5 CNN-layers. The input layer size (1 × 128 × 128) and the model output $f_{sm} \in \mathbb{R}^A$ is the spatial map output scores that represent the spatial map features. The spatial map features is given by:

$$f_{sm} = \sigma((f_{SM}^{h,o}))$$

where $\sigma$ is sigmoid activation function.

2) RELATIVE GEOMETRY ENCODER AND PREDICTOR
Relative geometry generates the the relative size, distance and IoU between human body parts bounding boxes and object bounding box. Relative size is determined as the size function that get size for a given mask. $mask_0$ to $mask_{24}$ are the human and his body parts masks and $S \in \mathbb{R}^{25}$.

Relative distance is the normalized Euclidean distance between body part bounding box center and object bounding box center. The distance is normalized related to width and height of the image. The relative distance body part and object is given by:

$$d_p = \sqrt{(\frac{W}{x_o - x_{p_i}})^2 + (\frac{H}{y_o - y_{p_i}})^2)}$$

where $(x_o, y_o)$ is the center of object bounding box and $(x_{p_i}, y_{p_i})$ is the center of $i^{th}$ human part bounding box, $W$ and
The model output is a `f`× `try` predictor to output the geometry features `f` where they concatenate relative size, distance, and an IoU vector, features are used by the geometry attention. In this case, the occluded part geometry where IoU the object bounding box and `iou`, where `iou` is the IoU between the human bounding box and object, and `IoU` is the relative overlap between the part and object that add discrimination feature to the action performed with the same part as shown in FIGURE. 5.

The overlap between bounding box `B_1` and bounding box `B_2` is determined by IoU between them as following [55]:

\[
IoU = \frac{area(B_1 \cap B_2)}{area(B_1 \cup B_2)}
\]

The IoU vector contains all IoU values for a human and his individual body parts as follows:

\[
IoU = \{iou_0, iou_1, \ldots, iou_{24}\}
\]

where `iou_0` is the IoU between the human bounding box and the object bounding box and `iou_1` to `iou_{24}` is the body parts where `IoU \in \mathbb{R}^{25}`.

In most images, not all body parts appear in the image due to crops or occlusion. In this case, the occluded part geometry features are zeros.

The final output `G` of a relative geometry encoder is the concatenation of relative size, distance, and an IoU vector, where `G \in \mathbb{R}^{75}`.

The generated geometry features `G` are used by the geometry predictor to output the geometry features `f_{geo}`. The geometry predictor is 5 linear layers, the input layer is `(1 \times 75)` and the model output is `f_{geo}^h \in \mathbb{R}^A`. The geometry feature is given by:

\[
f_{geo} = \sigma(f_{geo}^h)
\]

where `wd_{a_o}` is the weighted distance for all action object pair `p_{a,o}` in the training set and `D` is the relative distance between human and object pair in `p_{a,o}` that determined in III-A2. After generating the weighted distance vectors for all action-object pairs, we create an action-object weighted distance matrix that contains action-object pairs for all actions. During inference, we determine the weighted distance `wd_{a_o}` for HOI proposal and multiply it with the weighted distance for each action to get the weighted distance for the proposed pair `wd_{a_o} \in \mathbb{R}^A`.

FIGURE. 6 illustrates samples of weighted distance for some actions from V-COCO dataset. For instance, in FIGURE. 6 (a), we find that for action `drink`, the highest values are for `right hand and torso`. However, for `skateboard` action FIGURE. 6 (b) shows that the maximum values for `left`
and right feet. On the other hand, action talk on phone gets the highest values for head and hand. The advantage of using weighted distance in addition to parts IoU with the object in the spatial geometry encoder is to keep the body part attention in case the part is occluded or not detected due to human part detector error.

4) FUSION MODEL

For each HOI proposal, the spatial map and geometric features are generated by the body parts encoder and geometry encoder, respectively. The generated features are fed to the spatial map predictor and geometry predictor that generate the spatial map features \( f_{sm} \) and geometric features \( f_{geo} \), then \( f_{sm} \) and \( f_{geo} \) are concatenated together to form spatial features \( f_{sp} \) where \( f_{geo}, f_{sm} \) and \( f_{sp} \) \( \in \mathbb{R}^A \), \( A \) is the is the number of action classes. Weighted distance is determined for each HOI proposal as explained in section III-A3 and produce weighted distance features \( f_{wd} \) \( \in \mathbb{R}^A \). Fusion model combines \( f_{sp} \) and \( f_{wd} \) to generate the pair prediction results. In order to reject the unacceptable pairs, action prior is applied to the final HOI predicted pairs to ignore the unmatched action-object pairs. Additionally, if object-action occurrence is not exist, the associated weighted distance for the predicted pair will be zero because no action-object pairs are exist in the dataset and it’s used as a key indicator for the action semantics with different objects.

The Pair prediction model produces the final score for each HOI proposal according to the following:

\[
s_i = s_o * s_h * s_a
\]

where \( s_i \) is the pair score and \( s_o, s_h, s_a \) are object, human, and action scores respectively.

B. GLOBAL-REJECTION MODEL

Global-Rejection model is a human-centric model to reject false positive interactions for predicted pairs. Normally, the pair prediction model detects all the HOI between all human instances \( H = \{ h_1, h_2, \ldots, h_m \} \) and every object instances \( O = \{ o_1, o_2, \ldots, o_n \} \) in the image resulting image pairs \( P \in \mathbb{R}^{m \times n}, P = \{ p_{11}, p_{12}, \ldots, p_{mn} \} \). Then, all predicted pairs are evaluated for all humans and objects in the image by applying Hungarian matching [21] to all pairs in \( P \) according to the pair scores that assign a group of objects to each human and remove duplicated objects for humans. Then, we apply the human-centric rejection to each human and his associated objects to generate the final prediction according to co-occurrence action semantics.

1) HUNGARIAN MATCHING

The Hungarian matching algorithm is used to solve the bipartite graph matching problem according to cost or profit value between two sets [21]. We adopt the Hungarian matching to pairs scores as used in [56] to filter the pair predictions and keep the highest scoring prediction pairs for each human and object in the image that involved in the same action class. After pair prediction, we get the prediction score of \( S_p \) for each predicted pair in \( P \), where and as HOI proposals are multi-label classification problem. According to \( S_p, H, \) and \( O \), we generate for each action \( a \in A \) a matrix \( M_{ao}(m, n) \) includes the index for all human, objects and action score for each pair. Hungarian matching is a one-to-one matching algorithm but according to the action annotations, we found that one object can be interacting with multiple humans, such as sit action more than one person can sit on a single bench or couch. To overcome this problem, we assign cardinally \( C_{ao} \) value for each action-object pair and apply the Hungarian algorithm recursively according to the cardinality value. The majority of action-objects have \( C_{ao} = 1 \) but some action-object pairs have more cardinal values, such as sit on couch and look at TV. FIGURE. 7 illustrates an example of the Hungarian matching for drink action and the resulting pairs according to the best score for each human.

2) REJECTION MODEL

The rejection model is very crucial in HOI detection because many invalid actions are detected for the same human. We cannot reject the undesired pairs according to the low score alone, because some actions have a low score due to occluded human parts or objects. In the rejection model, after matching the HOI proposals using Hungarian matching, we train the rejection model to select the best pairs associated with each human in the image and reject the other pairs. We train the model with the annotated pairs and mark them as true and all other pairs are used as negative pairs, marked as false. FIGURE. 8 illustrates an example of the rejection model. Predicted pairs dominant action in the image are \{ work on computer, sit, hold, drink, read, look \} with 6 different objects. The human-object semantic co-occurrence for the human in the red box indicates that the human can sit while he is working on a computer, but he can not read or look at another object while working on it, and work on computer action score is higher than read and look scores. As a result, the interactions with paper and apple are rejected. Additionally, there were no co-occurrence actions between
work on computer action and drink action. So, drink action with the cup is rejected too.

The rejection model keeps the highest scoring actions and removes the lowest scores according to the human co-occurrence knowledge gained from the training data. The rejection model predicts $K$ HOI pairs that contains for each pair $p_i$ the relative distance $D_i$ as in equation (Equation 4) concatenated with the action score $s_i$ that produced from Hungarian matching where $K$ is an arbitrary number that is larger than the typical number of pairs associated with a single human and we set it to 20 empirically. The input data is given by:

$$R = \left\| \sum_{i=0}^{K} [D_i s_i] \right\|$$  \hspace{1cm} (10)

where $R \in \mathbb{R}^{25+|A|}$ ($25$ body parts and $|A|$ actions scores) and $\left\| \right\|_{2}$ denotes the vertical stack of pairs vectors that are padded with $\phi$ (zeros vector) and zero labels of size $K$. The predicted output $O \in \mathbb{R}^K$ is multi-label vector which includes 1 if there is any interaction between the pair, 0 otherwise.

The rejection model is a linear NN model that generate output label 0 (if $O < 0.5$) or 1 (if $O \geq 0.5$) to determine the non-interactive and interactive pairs respectively.

IV. TRAINING SPATIAL-NET

Spatial-Net as explained above includes two models, the pair prediction model and, the global-rejection model. The Pair prediction model includes three sub-models, spatial map, geometry, and weighted distance predictors, while the global-rejection model includes the Hungarian matching and human-centric rejection model. All the sub-models are trained in a fully supervised fashion using multi-label binary cross-entropy loss. In the pair prediction training phase, each HOI proposal is assigned to the multi-label vector of length $|A|$ including one for the interactive class and zero for non-interactive. Spatial map and geometry are encoded using the spatial and relative geometry encoders and the output score for each model is fused using the fusion model. As we mentioned above, the HOI problem is a multi-label classification problem, we use binary cross-entropy loss and Adam optimizer [57] in model training. Binary cross-entropy loss determine the loss for each class individually and the total loss is given by:

$$L(\hat{y}, y) = -\sum_{j=1}^{c} y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j)$$

$$BCELoss = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}_i, y_i)$$  \hspace{1cm} (11)

where $\hat{y}, y$ are the predicted score and ground truth label respectively, $c$ is the number of classes and $m$ number of samples. Loss of pair prediction model is given by:

$$\mathcal{L}_p = \sum_{i=1}^{K} \mathcal{L}_{ap}(a_{pi}, \hat{a}_{um}) + \mathcal{L}_{sp}(a_{pi}, \hat{a}_{geo}) + \mathcal{L}_f(a_{pi}, \hat{a}_{sp})$$  \hspace{1cm} (12)

where $\mathcal{L}(a, b) = BCELoss(a, b)$ and $\mathcal{L}_{ap}, \mathcal{L}_{geo}$ and $\mathcal{L}_f$ are spatial map, geometry, and fusion loss respectively, $a_{pi}$ is the ground truth action, $\hat{a}_{um}, \hat{a}_{geo}, \hat{a}_{sp}$ are spatial predictor, geometry predictor and fusion model scores respectively.

Global-rejection model uses the results that are provided by the pair prediction model and reject false positive pairs by applying Hungarian matching which is a deterministic algorithm, then provide the rejection model with relative distance concatenated with the prediction score $\Delta_i = [D_{pi}, S_{pi}, \ldots, D_{pi}, S_{pi}]$ to filter the object for each human individually. To construct the rejection data, we apply the pair prediction model and Hungarian matching to the training HOI proposals and label the ground truth pairs with 1 and 0 otherwise. Rejection model loss is calculated as $\mathcal{L}_r = BCELoss(\Delta_i, \Delta_i)$.

V. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we will describe the implementation details for our proposed approach as well as the dataset (V-COCO dataset) that we use as benchmarks [22]. Then we demonstrate how Spatial-Net successfully captures HOI proposals. Finally, we compare the results of our proposed Spatial-Net to those of some state-of-the-art HOI models.

A. IMPLEMENTATION DETAILS

For HOI prediction, we first detect the object and human body parts for each HOI proposal in the image and create a $(1 \times 128 \times 128)$ spatial map. Then, we retrieve the geometric data for each HOI proposal. We use PyTorch [58] framework for our model implementation. We employ data augmentation using Albumentations [59] (random flip and random rotate) and we use the Adam optimizer [60] for loss function optimization during training. We use batch-size of 16 on a single Nvidia RTX 2080 Super GPU and a learning rate of $1.10^{-4}$ for 300 epochs. We set the object detector threshold to 0.4 for human and 0.3 for object.
B. DATASET AND METRICS

We test our model in V-COCO dataset that provides verb annotations for MS-COCO [54]. The dataset contains 10,346 images split into 3 sets, training (2,533 images), validation (2,867 images) and test (4946 images). It contains 16,199 HOI proposals that are annotated with 26 action labels. It also contains three actions (cut, eat, hit) that are annotated with two types of targets (object and instrument). The V-COCO dataset contains the same 80 objects as the MS-COCO dataset.

There are many other datasets that are used in HOI detection in images such as HICO-Det [14] and SWIG-HOI [61] that contain more interaction classes but they don’t include the required annotation for our proposed approach.

We follow the standard evaluation metric used in PASCAL VOC [55] that reports the role mean average precision (role mAP). The mAP is computed based on both recall and precision, which is appropriate for the HOI detection task to test not only the model prediction ability but also the rejection ability. Average precision (AP) for a given action class represented by the area under the precision/recall curve that is calculated using a method’s ranked output for a specific task and class. The proportion of all positive examples ranked above a specific rank is known as recall. Precision is the percentage of positive examples in all examples above that rank. AP is calculated for each class \( a \in A \) as following:

\[
AP(a) = \sum_{k=1}^{N} p(k) \Delta r(k)
\]

where, \( p \) and \( r \) are the precision and recall respectively, \( N \) is the number of samples in the action class \( a \). The model mAP for all action classes is given by:

\[
mAP = \frac{1}{|A|} \sum_{a \in A} AP(a)
\]

The predicted pair is considered a true positive if the predicted actions are correct and the IoU between human and object predicted bounding boxes, and ground truth boxes \( \geq 0.5 \).

C. RESULTS AND COMPARISONS

In this section, we compare Spatial-Net performance with some of the state-of-the-art methods. We follow the evaluation technique that is explained in V-B.

As illustrated in TABLE. 1, our approach outperforms many HOI models and achieves comparable results with the state-of-the-art methods, even though we are using spatial features only for HOI detection. In addition, the inference speed of our model is 2ms after object detection, which outperforms many of the state-of-the-art HOI models.

D. QUALITATIVE VISUALIZATIONS

As illustrated in FIGURE 9, we show some HOI detection results on the V-COCO dataset. In the first row of the figure, we show actions that are detected correctly, which shows the model’s ability to detect actions with multiple humans and objects in the image and assign each object correctly to the human, even when there is existing overlap between human boxes as in (c) and (d). In (c), the correct association occurs due to the high weight that the model gives to the object near the head and hand, while in (e) the correct association occurs due to the high weight to object (horse) near the legs and torso. In the second row, we show wrong HOI predictions caused by miss association and wrong prediction. The miss association between woman and laptop in FIGURE 9 occurs because the laptop is closer to the woman than the right laptop. However, in (e), the chair is close to the person’s legs and torso and no other objects are associated with the human, so the model predicts sit action with the chair. In (h), the model predicts action as catch instead of throw because the human pose and distances from body parts are similar in some poses for both actions (catch, throw).

E. ABLATION STUDY

In this section, we explore how each element in Spatial-Net affects the performance of the final result. HOI interaction detection using spatial map only achieves 44.37 mAP on V-COCO test set role which is considered the model baseline. To improve the baseline prediction results, we add the geometry data (relative size, relative distance, and parts

| Method       | Features                                      | Role(mAP) | Inference time (ms) |
|--------------|-----------------------------------------------|-----------|---------------------|
| VSRL [22]    | V                                             | 31.80     | -                   |
| InteractNet [44] | V                                              | 40.00     | 50                  |
| BAR-CNN [62] | V + S                                         | 43.60     | 110                 |
| GPNN [26]    | V                                             | 44.00     | 40                  |
| ICAN [18]    | V + S                                         | 45.30     | 90                  |
| DCA [47]     | V + S                                         | 47.30     | 130                 |
| UnionDet [37] | V                                             | 47.50     | 9                   |
| TIN [16]     | V + S + P                                     | 47.80     | 70                  |
| VCL [58]     | V + S                                         | 48.30     | -                   |
| DRG [27]     | V + S                                         | 51.00     | -                   |
| PastaNet [63] | V + S + P + L                                 | 51.00     | -                   |
| IPNet [19]   | V                                             | 51.00     | 5                   |
| VSGNet [28]  | V + S                                         | 51.80     | -                   |
| PD-Net [25]  | V + S + P + L                                 | 52.00     | -                   |
| ACP [37]     | V + S + P                                     | 53.00     | -                   |
| FCNNet [23]  | V + S + P                                     | 53.30     | -                   |
| IDN [64]     | V + S                                         | 53.30     | 60                  |
| HOTR [65]    | V                                             | 55.20     | 1                   |
| Spatial-Net  | \( + P \)                                      | 55.25     | 2                   |

TABLE 2. Ablation study on both V-COCO test set role mAP. (SM: Spatial map, GEO: Geometry data, WD: weighted distance, H Matching: Hungarian matching and G_rejection: Human-centric rejection.)
of IoU) to the baseline spatial map which increases the model performance to 47.68 mAP. To enhance the results, we add the weighted distance as attention values based on dataset statistics to focus on the parts that are involved in the action. Adding weighted distance increases the performance to 48.69 mAP. To remove the repeated pairs and assign objects to humans according to the prediction score, we use the Hungarian matching with the whole predicted pairs in the image, which increases the performance to 50.13 mAP. The most important model that elevates the model performance is the human-centric rejection model, which removes the false positive HOI pairs according to the score or remaining pairs after applying Hungarian matching. The human-centric rejection model keeps the high-scoring actions and rejects the low-scoring actions according actions co-occurrence semantics, which filters the entire HOI pairs on the image. After applying human-centric rejection, the performance is increased to 55.25 mAP. In TABLE. 2 model summary, results using each element are illustrated.

VI. CONCLUSION
The detection of a human’s relationship with an object in still images and videos is one of the major tasks in computer vision. The great majority of human-object interaction (HOI) recognition methods in images rely on the visual features as their primary component for determining the connection between humans and objects. Solutions to the HOI problem might lead to the formation of false-positive pairs, which is caused by incorrect grouping between human-object pairs as well as the presence of many non-interacting objects in the image. In this paper, we introduced “Spatial-Net”, a HOI approach that adopts spatial relations data between humans and objects to predict HOI between humans and objects in images. We argued in the proposed approach that appearance features are not required in HOI detection and that spatial semantic relations can only compete with the state-of-the-art HOI method. The proposed method is divided into two parts: detecting HOI pairs and rejecting false positive pairs. In the first part, we created a spatial map with masks of body parts and objects. Each mask of a body part has its own unique index. The proposed approach was then supplemented by the addition of relative geometry features between the human and the object. These features include body part and object relative size, distance, and IoU. Furthermore, we used statistical analysis to represent the weighted distance between each body part and object in all actions, which was used to generate an action vector with scores for each action based on the weighted distance. The addition of weighted distance features to the augmented approach improves performance. In the second part, we used Hungarian matching and a human-centric rejection model to eliminate false positive pairs while keeping only true positive pairs. When compared to the most recent HOI models on the V-COCO dataset, the proposed approach achieved state-of-the-art performance. Furthermore, the proposed model’s detection inference speed was 1–2 ms, making the model suitable for deployment on small embedded systems and real-time applications. In our future work, we will create the required annotations for different datasets that contains more HOI classes and test our approach on them.

REFERENCES
[1] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, “A survey of modern deep learning based object detection models,” Digit. Signal Process., vol. 126, Jun. 2022, Art. no. 103514.
[2] Y. Yuan and Y. Zhang, “OLCN: An optimized low coupling network for small objects detection,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
[3] X. Lu, Y. Zhang, Y. Yuan, and Y. Feng, “Gated and axis-concentrated localization network for remote sensing object detection,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 1, pp. 179–192, Jan. 2019.
[4] M. Koohzadi and N. M. Charkari, “Survey on deep learning methods in human action recognition,” IET Comput. Vis., vol. 11, no. 8, pp. 623–632, Dec. 2017.
AHMED E. MANSOUR received the B.Sc. degree from Alexandria University, Alexandria, Egypt, in 1999, the Postgraduate Diploma degree in computer science from the Faculty of Graduate Studies of Statistical Researches, Cairo University, Egypt, in 2012, and the M.Sc. degree in electrical communication from the Electronics and Communications Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt, in 2016, where he is currently pursuing the Ph.D. degree with the Faculty of Engineering. His research interests include radar, electrical communication, deep learning, computer vision applications, and artificial intelligence algorithms.

AMMAR MOHAMMED received the bachelor’s and master’s degrees in computer science from Cairo University, Egypt, and the Ph.D. degree in computer science from the University of Koblenz-Landau, Germany, in 2010. He worked as a Researcher and a Research Fellow at the Artificial Intelligence (AI) Research Group, University of Koblenz-Landau. He is currently a Professor of computer science at Cairo University. He supervised a group of Ph.D. and master’s students and established the Machine/Deep Learning Research Group at the Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University. His research interests include machine and deep learning techniques, methods, algorithms, and their applications in several domains.

HUSSEIN ABD EL ATTY ELSAYED (Member, IEEE) received the B.Sc. and M.Sc. degrees from Ain Shams University, Cairo, Egypt, in 1991 and 1995, respectively, and the Ph.D. degree from the Electrical Engineering Department, City University of New York, New York City, NY, USA, in 2003. He is currently a Professor with the Electronics and Communications Engineering Department, Faculty of Engineering, Ain Shams University. Since then, he served in different positions and built both practical and theoretical skills in the area of communication networks. He has supervised and joined several telecommunication graduation, research, and lab/operational networks projects, including service provider NOC. His research interests include telecommunication networks, ad-hoc networks, cognitive radio networks, wireless sensor networks, medium access control, routing protocols, and mobile systems, where he has a long list of publications.

SALWA ELRAMLY (Life Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Faculty of Engineering, Ain Shams University, Egypt, in 1967 and 1972, respectively, and the Ph.D. degree from Nancy University, France, in 1976. She is currently a Professor Emeritus with the Electronics and Communications Engineering Department, Faculty of Engineering, Ain Shams University, where she was the Head of the Department, from 2004 to 2006. She has published more than 200 scientific papers and has supervised more than 100 M.Sc. and Ph.D. theses. Her research interests include wireless communication systems and signal processing, language engineering, coding, encryption, and radars. She received the Ain Shams Award of Appreciation in Engineering Sciences, in 2010, the Award of Excellence from the Society of Communications Engineers, in 2009, and the Award of Excellence from the Egyptian Society of Language Engineering and the Award of Excellence from the National Radio Science Committee, Egypt. She has participated in the establishment of the Egyptian Society of Language Engineering, in 1996, where she is currently the President. She is the IEEE Signal Processing Chapter Chair in Egypt.