# What-Meets-Where: Unified Learning of Action and Contact Localization in Images

**Yuxiao Wang[1], Yu Lei[2], Wolin Liang[1], Weiying Xue[1], Zhenao Wei[3], Nan Zhuang[4], Qi Liu[1]***

[1]School of Future Technology, South China University of Technology
[2]School of Information Science & Technology, Southwest Jiaotong University
[3]Computing Science and Artificial Intelligence College, Suzhou City University
[4]School of Software Technology, Zhejiang University
ftwangyuxiao@mail.scut.edu.cn, drliuqi@scut.edu.cn

## Abstract

People control their bodies to establish contact with the environment. To comprehensively understand actions across diverse visual contexts, it is essential to simultaneously consider **what** action is occurring and **where** it is happening. Current methodologies, however, often inadequately capture this duality, typically failing to jointly model both action semantics and their spatial contextualization within scenes. To bridge this gap, we introduce a novel vision task that simultaneously predicts high-level action semantics and fine-grained body-part contact regions. Our proposed framework, PaIR-Net, comprises three key components: the Contact Prior Aware Module (CPAM) for identifying contact-relevant body parts, the Prior-Guided Concat Segmenter (PGCS) for pixel-wise contact segmentation, and the Interaction Inference Module (IIM) responsible for integrating global interaction relationships. To facilitate this task, we present PaIR (Part-aware Interaction Representation), a comprehensive dataset containing 13,979 images that encompass 654 actions, 80 object categories, and 17 body parts. Experimental evaluation demonstrates that PaIR-Net significantly outperforms baseline approaches, while ablation studies confirm the efficacy of each architectural component.

**Code** — https://drliuqi.github.io/

**Extended version** — https://arxiv.org/abs/2508.09428

## Introduction

Individuals interact with the environment through both behavioral intent and physical contact, whether sitting on a chair or lifting a cup. This dual understanding, combining semantic goals with physical contact mechanisms, is fundamental to human actions. Neuroscience supports this view, showing that actions like "pushing a door" involve not only action recognition but also verification of contact between the hand and the door handle (Bicchi and Kumar 2000).

However, existing benchmarks fail to capture both aspects simultaneously. While prior works focus on specific body-part contact (e.g., hand-object (Shan et al. 2020; Cui et al. 2023), foot-ground (Tripathi et al. 2023), or human-object contact (Chen et al. 2023a)), they are often limited to localized contexts and lack full-body interaction modeling. In contrast, action recognition methods typically emphasize "**what**" action occurs, neglecting the "**where**" of physical contact, thus hindering their applicability in complex, real-world scenarios (Wang et al. 2025b, 2024b,a).

As shown in Figure 1, a single object (e.g., a cake or cup) can correspond to different actions based on contact regions. For instance, "eating" involves the hand and head, while "holding" requires only the hand. This underscores the need to jointly model **what** the action is and **where** contact occurs. Beyond action classification (Wang, Li, and Lei 2022), models should also localize body parts involved (e.g., "buttocks contacting chair" for "sitting"), essential for fine-grained understanding (Wang et al. 2025b) and applications like robotic planning, AR/VR assessment and imitation learning (Westermeier et al. 2024; Zare et al. 2024).

To address this, we introduce a new task that jointly infers action semantics and physical contact segmentation. The task poses two interconnected questions: **What** is the person doing, and **where** is the body in contact with the object to perform the action? This requires a dataset that captures both high-level intent and low-level grounding. **We propose PaIR (Part-aware Interaction Representation), a dataset of 13,979 real-world images with 654 action types, 80 object categories, and 17 contactable body parts.** Each sample includes annotations of $\langle$*person*, *verb*, *object*, *contact part*$\rangle$, 2D contact masks, and part labels, enabling joint learning of interaction semantics and contact.

We propose PaIR-Net, a unified framework that jointly models action recognition and contact segmentation. PaIR-Net consists of three key components: the Contact Prior Aware Module (CPAM) predicts interacting body parts; the Prior-Guided Concat Segmenter (PGCS) segments contact regions; and the Interaction Inference Module (IIM) integrates spatial and semantic cues to infer interaction types. Additionally, two mechanisms enhance performance: the H-O RoI Enhancer for guiding PGCS with bounding boxes and the Mask-Guided RoI Feature module leverages contact masks for better recognition.

## Related Works

**Action Recognition**. Standard object detectors (Dai et al. 2016; Ren et al. 2015) localize instances but fail to capture

*Corresponding author

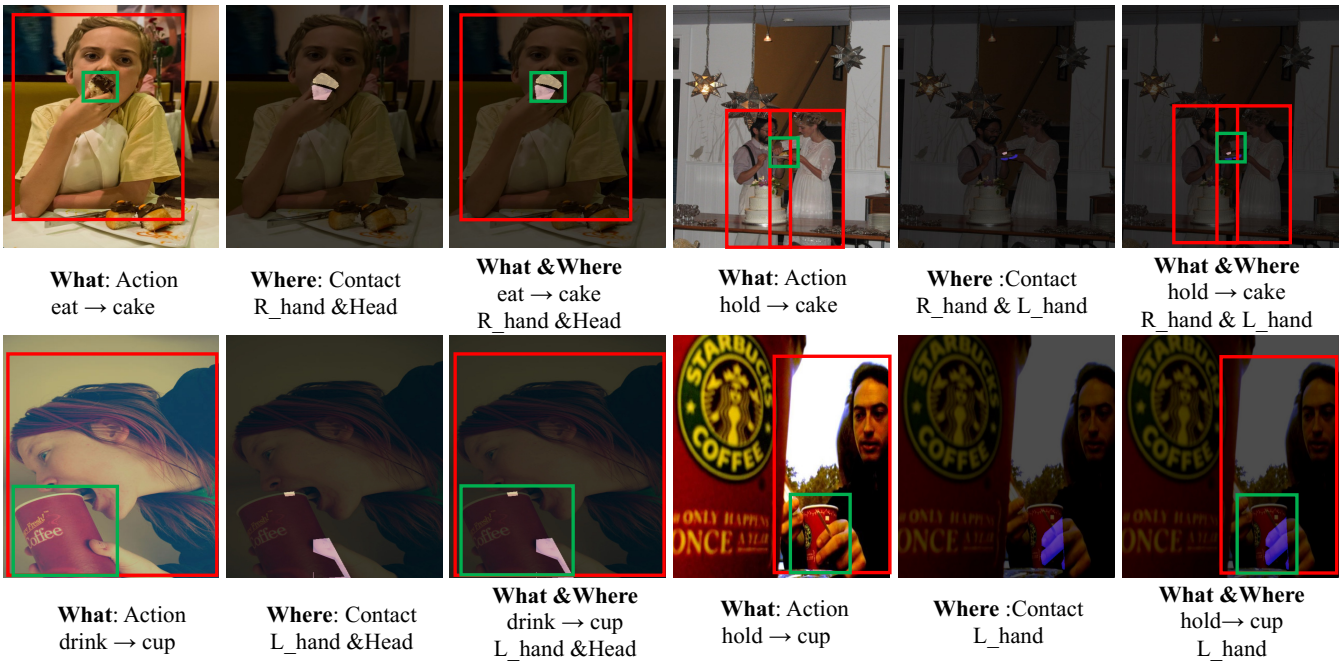


Figure 1: The same object (e.g., cake or cup) can imply different actions depending on contact regions. For example, "eating" involves both hand and head, while "holding" involves only the hand. To bridge this gap, we propose joint modeling of **What** (action & object) and **Where** (contact body part).

inter-instance interactions. Early works address this using pose cues (Desai and Ramanan 2012; Yao and Fei-Fei 2010), while visual phrase models (Sadeghi and Farhadi 2011) detect interacting pairs. Recent human-object interaction methods (Liao et al. 2022; Yang et al. 2024) improve fine-grained recognition by identifying specific human-object pairs, and others (Wang, Liu, and Lei 2024; Wang, Teng, and Wang 2024) incorporate context and language priors.

**Contact Perception**. Recent studies explore contact modeling at specific body parts, such as hands (Shan et al. 2020; Darkhalil et al. 2022) and feet (Tripathi et al. 2023), revealing the value of contact-aware understanding. However, these methods often focus on isolated parts or rely on specific modalities like depth or video. Although human-object contact (Chen et al. 2023a) introduces a full-body contact dataset to detect multiple contact points, it lacks integration with semantic action reasoning. In contrast, our approach unifies semantic action recognition and dense contact localization from a single 2D image.

**Datasets**. Several recent datasets focus on physical contact between the human body and the environment. VISOR (Darkhalil et al. 2022) serves as a dataset for pixel-level annotations and a benchmark for segmenting hands (Ma and Damen 2022) and active objects in egocentric video (Girdhar et al. 2019). PressurePose (Clever et al. 2020) provides 206K synthetic pressure images with paired 3D pose and shape. RICH (Huang et al. 2022) captures real-world multi-view interactions in 3D scenes, while ContactPose (Brahmbhatt et al. 2020) and ARCTIC (Fan et al. 2023) focus on hand or whole-body contact. HOT (Chen et al. 2023b) enables contact detection from images, yet these datasets primarily emphasize contact perception and lack integration with action semantics (Wang et al. 2025a). Conversely, action recognition datasets such as imSitu (Yatskar, Zettlemoyer, and Farhadi 2016), HICO (Chao et al. 2015), HICO-Det (Chao et al. 2018), and V-COCO (Gupta and Malik 2015) provide ⟨human, verb, object⟩ triplets but omit contact-specific annotations. To bridge this gap, we introduce PaIR, the first dataset to provide ⟨*human*, *verb*, *object*, *contact part*⟩ annotations at the image level.

## Method

In this paper, we propose a novel framework, PaIR-Net, to jointly model action semantics and contact regions. Specifically, PaIR-Net can identify which human-object pairs are interacting in an image, classify the type of interaction, and segment the specific body parts involved in the contact.

### Overall Architecture

The architecture of PaIR-Net is shown in Figure 2. For input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, we apply data augmentation techniques (Wang, Liu, and Lei 2024) before using ResNet (He et al. 2016), or Swin Transformer (Liu et al. 2021) as backbone to extract feature $\boldsymbol{F}_B \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times C}$, where $s = 32$ and $C$ is the channel dimension. Then a CPAM network is employed to focus on 17 body parts that contact objects. The $\boldsymbol{F}_B$ also feeds into both the PGCS and the IIM to generate contact region segmentation and interaction actions, respectively. For effective collaboration between these modules, we introduce the H-O RoI Enhancer, which guides PGCS to

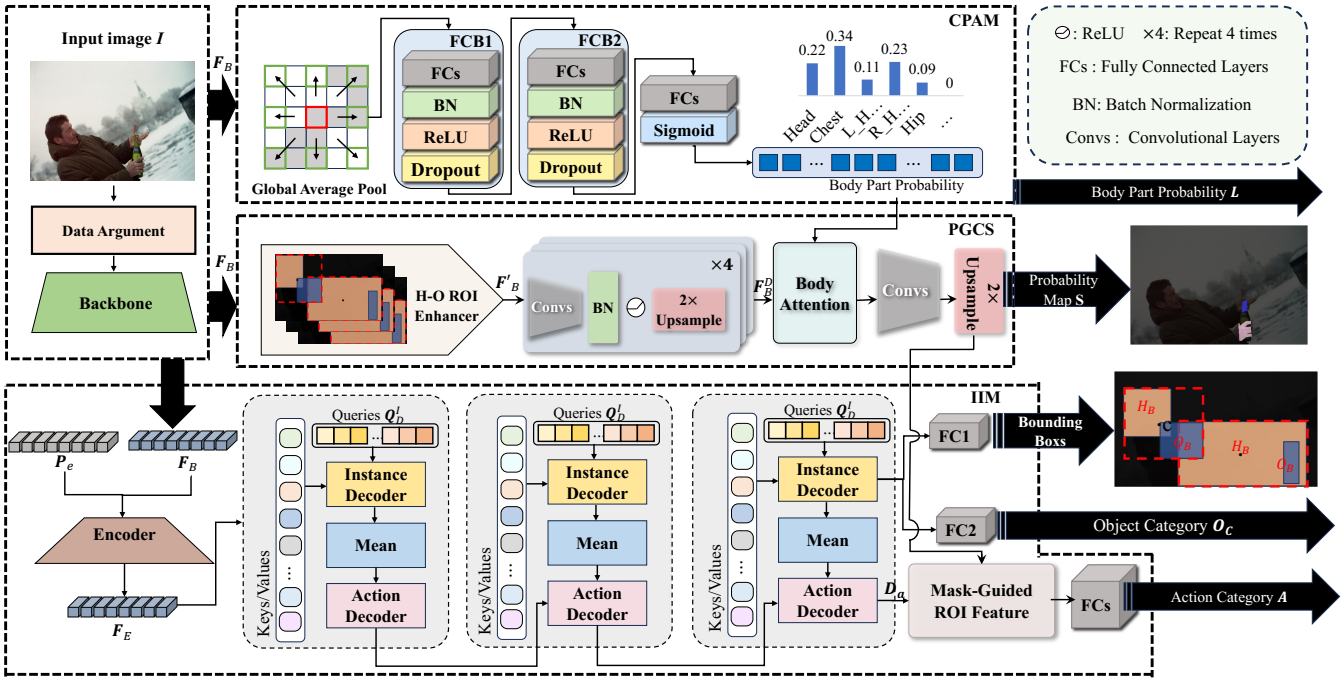


Figure 2: The overall workflow of PaIR-Net. It comprises three branches: CPAM for multi-label body part contact prediction (upper part of Figure), PGCS for outputting contact region segmentation (middle part of Figure), and IIM for detecting human-object pairs and identifying interaction categories (lower part of Figure). To facilitate effective collaboration between contact understanding and action recognition, we design two key modules: the H-O RoI Enhancer and the Mask-Guided RoI Feature.

focus on potential interaction areas, and the Mask-Guided RoI Feature module, which leverages the segmentation results of PGCS to enhance interaction recognition.

## Contact Prior Aware Module (CPAM)

To guide PGCS toward interaction-relevant regions, CPAM is designed. The module first applies global average pooling (GAP) to $\boldsymbol{F}_B$, ensuring consistent dimensions for subsequent processing. The pooled features pass through fully connected blocks (FCB)—each containing a Fully Connected (FC) layer, Batch Normalization (BN), ReLU, and Dropout—to enhance semantic representation. Then, a sigmoid activation is used to output the contact probability for each body part, helping PGCS focus on body parts likely involved in interactions. Specifically, given features $\boldsymbol{F}_B \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times C}$, we apply GAP to obtain $\boldsymbol{F}_G \in \mathbb{R}^{1\times1\times C}$, which passes through two FCB to produce the final contact prediction $\boldsymbol{L} \in \mathbb{R}^{N_c}$, where $N_c$ represents the number of contact categories. The process can be formulated as:

$$\boldsymbol{L} = \text{Sigmoid}(\text{FC}\,(\text{FCB}_2\,(\text{FCB}_1\,(\text{GAP}(\boldsymbol{F}_B))))), \quad (1)$$

where $\text{FCB}_i(\cdot)$ represents the $i$-th fully connected block. Since the CPAM performs a multi-label classification task, $\boldsymbol{L}$ is compared with the ground-truth contact labels using the Binary Cross-Entropy (BCE) loss.

## Prior-Guided Concat Segmenter (PGCS)

The PGCS produces pixel-level segmentation maps of contact regions through two key submodules: the H-O ROI Enhancer Module that refines region-level features, and the Body Attention mechanism that provides body part relevance priors for improved contact localization.

**H-O RoI Enhancer Module**. Since physical contact is inseparable from interaction areas, inspired by this, we propose the H-O RoI Enhancer module to enhance features of potential contact regions, guiding the PGCS network to focus on interaction-related areas. As shown in Figure 3(a), the module takes feature map $\boldsymbol{F}_B$, human bounding boxes $\boldsymbol{H}_B = [\boldsymbol{h}_{b1}, \boldsymbol{h}_{b2}, \ldots, \boldsymbol{h}_{bn}]$ and object bounding boxes $\boldsymbol{O}_B = [\boldsymbol{o}_{b1}, \boldsymbol{o}_{b2}, \ldots, \boldsymbol{o}_{bn}]$ predicted by the IIM module. Each bounding box $\boldsymbol{h}_{b_i}$ or $\boldsymbol{o}_{b_i}$ is defined by $(x_1, y_1, x_2, y_2)$, representing the top-left and bottom-right coordinates. For each pair of the sets of $\boldsymbol{H}_B$ and $\boldsymbol{O}_B$, we compute the minimum enclosing rectangle $\boldsymbol{E}_b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ to determine the feature region for enhancement, via:

$$x_{\min} = \min\left\{\min_{1\le i\le n} x_1(\boldsymbol{h}_{b_i}),\ \min_{1\le j\le m} x_1(\boldsymbol{o}_{b_j})\right\}, \quad (2)$$

$$y_{\min} = \min\left\{\min_{1\le i\le n} y_1(\boldsymbol{h}_{b_i}),\ \min_{1\le j\le m} y_1(\boldsymbol{o}_{b_j})\right\}, \quad (3)$$

$$x_{\max} = \max\left\{\max_{1\le i\le n} x_2(\boldsymbol{h}_{b_i}),\ \max_{1\le j\le m} x_2(\boldsymbol{o}_{b_j})\right\}, \quad (4)$$

$$y_{\max} = \max\left\{\max_{1\le i\le n} y_2(\boldsymbol{h}_{b_i}),\ \max_{1\le j\le m} y_2(\boldsymbol{o}_{b_j})\right\}, \quad (5)$$

$$\boldsymbol{E}_b = (x_{\min}, y_{\min}, x_{\max}, y_{\max}), \quad (6)$$

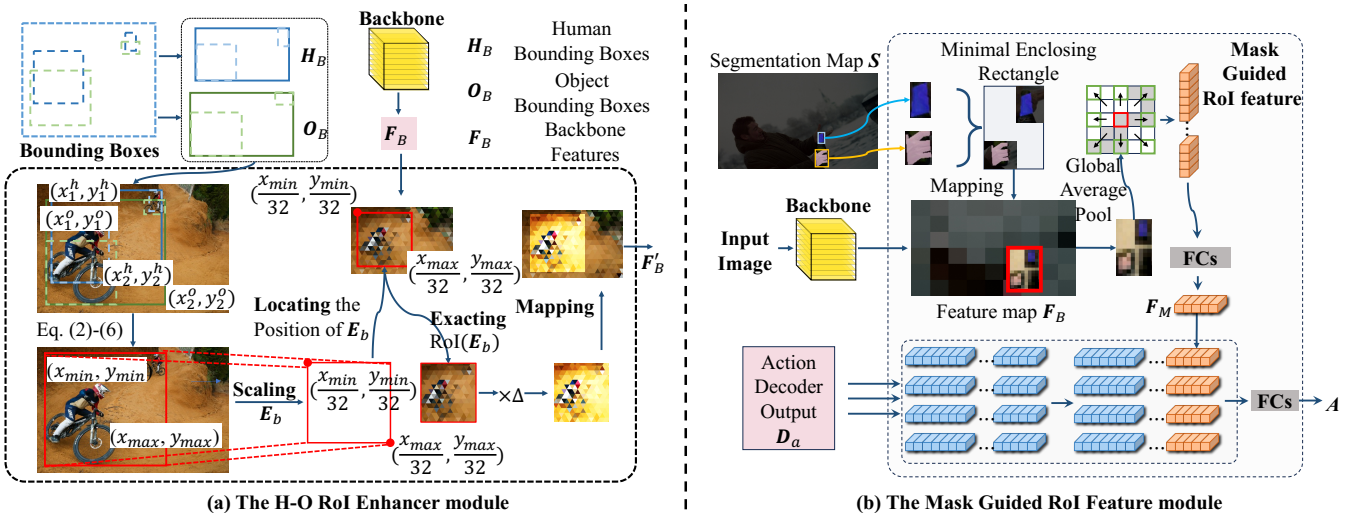


Figure 3: (a) The H-O RoI Enhancer module. It computes the minimum enclosing rectangle based on the human and object bounding boxes, and enhances the feature $\boldsymbol{F}_B$ responses within this region. (b) The structure of the Mask-Guided RoI Feature module. It utilizes $\boldsymbol{S}$ to extract the minimum enclosing contact region, crops the corresponding region from $\boldsymbol{F}_B$, and generates the contact feature encoding $\boldsymbol{F}_M$ through GAP and FC layers. Finally, $\boldsymbol{F}_M$ is fused with $\boldsymbol{D}_a$ to assist action classification.

where $n$ and $m$ represent the number of human and object bounding boxes, respectively. Functions $x_1(\cdot)$ and $y_1(\cdot)$ extract the x and y coordinates of the top-left corner, while $x_2(\cdot)$ and $y_2(\cdot)$ extract the bottom-right corner coordinates. The resulting $\boldsymbol{E}_b$ is the minimal enclosure rectangle that completely covers all the human and object bounding boxes. To ensure spatial correspondence between RoI regions $\boldsymbol{E}_b$ and backbone features $\boldsymbol{F}_B$, $\mathbf{E}_b$ is downscaled by a factor of 32. Subsequently, we amplify the feature values within this region by a factor of $\Delta$, thereby enhancing the PGCS's attention to the instance interaction areas, via:

$$\boldsymbol{F}'_B(i,j) = \begin{cases} \Delta \times \boldsymbol{F}_B(i,j), & \text{if } (i,j) \in \text{RoI}(\boldsymbol{E}_B) \\ \boldsymbol{F}_B(i,j), & \text{otherwise} \end{cases}, \quad (7)$$

where RoI($\boldsymbol{E}_B$) denotes the set of all pixel coordinates inside the scaled bounding box $\boldsymbol{E}_B$. $\Delta$ is a learnable parameter, which is initialized to 1.0.

Subsequently, the $\boldsymbol{F}'_B$ are fed into a multi-stage decoder consisting of convolutional layers, BN, ReLU activation functions, and $2\times$ upsampling layers. As a result, the decoded features $\boldsymbol{F}_B^D \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 64}$ are obtained.

**Body Attention**. Given the $\boldsymbol{F}_B^D$ and the $\boldsymbol{L} \in \mathbb{R}^{N_c}$ (predicted by the CPAM module), the body attention mechanism enhances the features through the following steps:

$$\boldsymbol{L}' = \text{Sigmoid}(\text{FC}(\text{ReLU}(\text{FC}(\boldsymbol{L})))), \quad (8)$$

$$\boldsymbol{F}_B^{D'} = \boldsymbol{F}_B^D \odot \boldsymbol{L}'. \quad (9)$$

The two FC layers expand the feature representation to obtain $\boldsymbol{L}' \in \mathbb{R}^{1\times 1\times 64}$. $\odot$ denotes element-wise multiplication along the channel dimension. $\boldsymbol{F}_B^{D'}$ will pass through a convolutional layer followed by a $2\times$ upsampling layer, producing the segmentation probability map $\boldsymbol{S} \in \mathbb{R}^{H\times W\times(N_c+1)}$ for contact regions, where the additional dimension represents the background.

## Interaction Inference Module (IIM)

To comprehensively infer interactions, we design an IIM based on an encoder-decoder architecture. In the encoder, parameters are initialized using DETR (Carion et al. 2020; Liao et al. 2022). Specifically, $\boldsymbol{F}_B$ is first embedded with positional encoding $\boldsymbol{P}_e$, then processed by the encoder to produce encoded features $\boldsymbol{F}_E$:

$$\boldsymbol{F}_E = \text{Encoder}(\boldsymbol{F}_B + \boldsymbol{P}_e). \quad (10)$$

Subsequently, we initialize two query matrixes for humans and objects: $\boldsymbol{Q}_h \in \mathbb{R}^{N_q\times C_Q}$ and $\boldsymbol{Q}_o \in \mathbb{R}^{N_q\times C_Q}$, where $C_Q$ is the feature dimension and $N_q$ is the maximum number of instances. These are concatenated into a single query set $\boldsymbol{Q}_D^I \in \mathbb{R}^{2N_q\times C_Q}$, which will input to the instance decoder, with $\boldsymbol{F}_E$ serving as both key $\boldsymbol{K}_D^I$ and value $\boldsymbol{V}_D^I$. The instance decoder outputs instance features $\boldsymbol{D}_i$, which are split into human features $\boldsymbol{D}_h$ and object features $\boldsymbol{D}_o$:

$$\boldsymbol{D}_i = \text{InstanceDecoder}(\boldsymbol{Q}_D^I, \boldsymbol{K}_D^I, \boldsymbol{V}_D^I), \quad (11)$$

$$\boldsymbol{D}_h, \boldsymbol{D}_o = \text{Split}(\boldsymbol{D}_i), \quad (12)$$

where $\boldsymbol{D}_h, \boldsymbol{D}_o \in \mathbb{R}^{N_q\times C_q}$, and Split$(\cdot)$ divides the input matrix into two parts along the row dimension. To understand interaction between each human-object pair, we add$\boldsymbol{D}_h$ and $\boldsymbol{D}_o$ and average them to obtain the query vector $\boldsymbol{Q}_D^A$. This query feeds into the Action Decoder, with $\boldsymbol{F}_E$ serving as both key ($\boldsymbol{K}_D^A$) and value ($\boldsymbol{V}_D^A$). The action decoder then outputs action features $\boldsymbol{D}_a$.

$$\boldsymbol{D}_a = \text{ActionDecoder}(\frac{\boldsymbol{D}_h + \boldsymbol{D}_o}{2}, \boldsymbol{K}_D^A, \boldsymbol{V}_D^A). \quad (13)$$

After three consecutive decoder stages, we obtain new $\boldsymbol{D}_h$, $\boldsymbol{D}_o$, and $\boldsymbol{D}_a$. $\boldsymbol{D}_h$ and $\boldsymbol{D}_o$ pass through an FC1 network to predict human bounding boxes $\boldsymbol{H}_B \in \mathbb{R}^{N_q\times 4}$ and

| Methods | Params | Times (ms) | Backbone | mAP↑ | SC-Acc.↑ | C-Acc.↑ | mIoU↑ | wIoU↑ |
|---|---|---|---|---|---|---|---|---|
| Trans (2021)+UNet (2015) | 74.0M | 88.7 | ResNet-50 | 13.63 | 14.80 | 26.66 | 9.33 | 15.10 |
| RCLTrans (2023)+UNet (2015) | 107.7M | 123.8 | ResNet-50 | 24.80 | 17.59 | 33.43 | 11.63 | 15.92 |
| STTrans (2022)+UNet (2015) | 87.2M | 114.9 | ResNet-50 | 28.48 | 15.04 | 27.16 | 9.37 | 13.04 |
| DisTrans (2024)+UNet (2015) | 79.8M | 102.3 | ResNet-50 | 31.23 | 16.67 | 30.58 | 11.39 | 15.58 |
| DisTrans (2024)+LinkNet (2017) | 120.4M | 83.9 | ResNet-50 | 31.17 | 20.34 | 35.79 | 11.78 | 15.51 |
| DisTrans (2024)+MANet (2020) | 236.6M | 89.5 | ResNet-50 | 31.01 | 22.91 | 44.32 | 12.72 | 17.36 |
| DisTrans (2024)+HOT (2023a) | 97.5M | 230.4 | ResNet-50 | 31.19 | 23.24 | 38.72 | 13.30 | 18.00 |
| DisTrans (2024)+PIHOT (2025b) | 158.6M | 257.5 | ResNet-50 | 31.45 | 26.35 | 45.16 | 14.53 | 18.54 |
| Ours | **55.6M** | **75.2** | ResNet-50 | 35.09 | 29.93 | 50.87 | 17.79 | 22.23 |
| Ours | 74.6M | 91.4 | ResNet-101 | 36.24 | 31.01 | 52.35 | 18.89 | 22.59 |
| Ours | 73.6M | 74.3 | Swin-S | 36.85 | 33.47 | 53.83 | 19.33 | 23.04 |
| Ours | 224.2M | 119.6 | Swin-L | **38.07** | **37.14** | **56.98** | **20.95** | **24.12** |

Table 1: Performance comparisons on PaIR-1 dataset.

object bounding boxes $\boldsymbol{O}_B \in \mathbb{R}^{N_q \times 4}$, respectively. Additionally, $\boldsymbol{D}_o$ feeds into an FC2 network to predict object category $\boldsymbol{O}_C \in \mathbb{R}^{N_q}$. $\boldsymbol{D}_a$ are input to our mask-guided RoI feature module, which integrates the segmentation map $\boldsymbol{S}$ from PGCS to predict action categories $\boldsymbol{A} \in \mathbb{R}^{N_q}$.

**Mask-guided RoI Feature Module**. Contact regions between humans and objects provide key action recognition cues—sitting involves buttocks-chair contact, while holding requires hand-bottle contact. Therefore, the segmentation map $\boldsymbol{S}$ assists action classification. As shown in Figure 3(b), we extract all non-background contact regions from $\boldsymbol{S}$ and compute their bounding boxes. The minimal enclosing rectangle is determined, and the corresponding sub-region is cropped from $\boldsymbol{F}_B$ to obtain focused interaction features. We then apply GAP followed by a fully-connected layer to generate a compact feature $\boldsymbol{F}_M \in \mathbb{R}^{10}$, which serves as the encoded representation of contact regions.

$$\boldsymbol{F}_M = \text{FC}(\text{GAP}(\text{Crop}(\boldsymbol{F}_B, S))). \tag{14}$$

$\boldsymbol{F}_M$ and $\boldsymbol{D}_a$ are concatenated and passed through FC layers to output the final action category predictions $\boldsymbol{A} \in \mathbb{R}^{N_q}$.

$$\boldsymbol{A} = \text{FC}([\boldsymbol{D}_A; \boldsymbol{F}_M]). \tag{15}$$

## Loss Function

Followed by the query-based object detection method (Kuhn 1955; Liao et al. 2022; Wang, Liu, and Lei 2024), the matching loss is designed in IIM between the predicted interaction pairs $\boldsymbol{Y} = [\boldsymbol{H}_B, \boldsymbol{O}_B, \boldsymbol{O}_C, \boldsymbol{A}]$ and the ground-truth pairs $\boldsymbol{GT}_{pair}$, as follows:

$$\mathcal{L}_{m\_l} = \sum_{i}^{N_q} \mathcal{L}_m(\boldsymbol{GT}^i_{pair}, \boldsymbol{Y}^i), \tag{16}$$

$$\mathcal{L}_m = \sum_{p \in \boldsymbol{O}_C, \boldsymbol{A}} \mathcal{L}^p_{cls} + \sum_{q \in \boldsymbol{H}_B, \boldsymbol{O}_B} \mathcal{L}^q_{box} + \sum_{r \in \boldsymbol{H}_B, \boldsymbol{O}_B} \mathcal{L}^r_{iou}. \tag{17}$$

For each sample, the total matching loss $\mathcal{L}_{\text{m\_l}}$ is defined as the sum of individual matching losses $\mathcal{L}_{\text{m}}$. Specifically, $\mathcal{L}_{\text{m}}$ consists of: (1) classification loss $\mathcal{L}_{\text{cls}}$ for accurate interaction recognition; (2) bounding box regression loss $\mathcal{L}_{\text{box}}$ for localization refinement; and (3) IoU loss $\mathcal{L}_{\text{iou}}$ to further improve prediction-ground truth alignment. The total loss $\mathcal{L}_t$ is computed via:

$$\mathcal{L}_t = \alpha \mathcal{L}_{m\_l} + \beta(\mathcal{L}_{BCE}(\boldsymbol{GT}_L, \boldsymbol{L}) + \mathcal{L}_{CE}(\boldsymbol{GT}_s, \boldsymbol{S})), \tag{18}$$

where BCE measures the contact prediction $\boldsymbol{L}$ against $\boldsymbol{GT}_L$. The Cross Entropy (CE) loss evaluates the accuracy of $\boldsymbol{S}$ compared to $\boldsymbol{GT}_S$. Hyperparameters $\alpha$ and $\beta$ balance the segment and interaction tasks.

# Experiment

## Datasets

To advance research on interaction and fine-grained contact region segmentation, we introduce a high-quality dataset unifying interaction pair recognition and pixel-level contact annotation in real-world images. We integrate diverse sources from COCO (Lin et al. 2014), HICO (Chao et al. 2015), HAKE (Li et al. 2020), Watch-n-Patch (Wu et al. 2015), and task-specific datasets like HICO-Det (Chao et al. 2018), V-COCO (Gupta and Malik 2015), and HOT (Chen et al. 2023a; Wang et al. 2025b). Through cross-source integration and re-annotation, we build a dataset with strict interaction semantics and high-quality contact segmentation labels.

Our dataset, PaIR, is the first to establish pixel-level contact annotation guided by interaction semantics in real-world images. After reviewing tens of thousands of interaction instances across nearly 100,000 images, we selected 13,979 images for annotation. The final dataset includes **45,103** instances, **46,616** interaction pairs, and **32,301** contact regions, covering **654** action categories and **80** object categories. The dataset is divided into two subsets: **PaIR-1** (8,591 images, 30,309 instances, 22,896 interaction pairs, 21,312 contact regions, 430 action categories) and **PaIR-2** (5,388 images, 14,794 instances, 23,720 interaction pairs, 10,989 contact regions, 224 action categories). PaIR-1 covers diverse action categories with broader interaction range and more challenging tasks, while PaIR-2 focuses on common daily object interactions, suitable for evaluating models in general-purpose scenarios.

| Methods | Params | Backbone | mAP↑ | SC-Acc.↑ | C-Acc.↑ | mIoU↑ | wIoU↑ |
|---|---|---|---|---|---|---|---|
| Trans (2021)+UNet (2015) | 74.0M | ResNet-50 | 47.93 | 12.64 | 24.99 | 7.84 | 13.65 |
| RCLTrans (2023)+UNet (2015) | 101.5M | ResNet-50 | 52.89 | 12.93 | 26.24 | 7.96 | 14.04 |
| DisTrans (2024)+UNet (2015) | 79.6M | ResNet-50 | 53.59 | 12.52 | 24.79 | 8.30 | 13.79 |
| DisTrans (2024)+LinkNet (2017) | 120.2M | ResNet-50 | 53.17 | 14.98 | 28.25 | 10.74 | 18.00 |
| DisTrans (2024)+MANet (2020) | 236.5M | ResNet-50 | 53.03 | 15.78 | 31.76 | 10.81 | 17.45 |
| DisTrans (2024)+HOT (2023a) | 97.3M | ResNet-50 | 53.45 | 17.45 | 32.24 | 11.54 | 16.56 |
| DisTrans (2024)+PIHOT (2025b) | 158.4M | ResNet-50 | 53.71 | 17.03 | 35.87 | 11.98 | 18.30 |
| Ours | **55.5M** | ResNet-50 | **57.60** | **19.80** | **42.40** | **13.25** | **20.37** |
| Ours | 74.4M | ResNet-101 | 59.06 | 21.30 | 45.42 | 14.11 | 21.50 |
| Ours | 73.5M | Swin-S | 59.43 | 22.41 | 47.30 | 15.97 | 22.31 |
| Ours | 224.1M | Swin-L | **61.32** | **24.01** | **49.62** | **17.11** | **23.79** |

Table 2: Performance comparisons on PaIR-2 dataset.

| Module | mAP | SC-Acc. | C-Acc. | mIoU | wIoU |
|---|---|---|---|---|---|
| baseline | 32.08 | 23.80 | 43.66 | 13.33 | 17.10 |
| +CPAM | 32.27 | 26.04 | 46.16 | 15.94 | 19.16 |
| +H-O | 32.63 | 29.21 | 50.01 | 17.37 | 21.92 |
| +M-G | **35.09** | **29.93** | **50.87** | **17.79** | **22.23** |

Table 3: Performance of each component when using ResNet-50 backbone. The **+** indicates incremental addition of modules. The H-O and M-G denote the H-O RoI enhancer and mask-guided ROI feature modules, respectively.

| $\alpha$ | $\beta$ | mAP | SC-Acc. | C-Acc. | mIoU | wIoU |
|---|---|---|---|---|---|---|
| 0.1 | 1.0 | 34.59 | 29.02 | 50.43 | 17.45 | 21.96 |
| 0.5 | 1.0 | 34.38 | 28.41 | 49.84 | 17.24 | 21.73 |
| 1.0 | 1.0 | 34.06 | 28.48 | 48.36 | 16.96 | 21.25 |
| 0.1 | 0.5 | **35.09** | **29.93** | **50.87** | **17.79** | **22.23** |
| 0.5 | 0.1 | 33.81 | 27.78 | 48.09 | 16.35 | 20.97 |
| 1.0 | 0.1 | 33.39 | 27.09 | 48.53 | 16.02 | 20.19 |

Table 4: The impact of different values of $\alpha$ and $\beta$ in Eq. 18 on the experimental results when using ResNet-50 backbone.

## Implementation Details

The weights of CPAM and PGCS are initialized using the method proposed by He et al. (He et al. 2015). For the encoder part of the IIM, we directly load the pre-trained DETR (He et al. 2015). The AdamW optimizer is used to optimize the network with an initial learning rate of $1 \times 10^{-4}$. In the loss function, the weight coefficients are set to $\alpha = 0.1$ and $\beta = 0.5$. The entire model is trained on 8 NVIDIA A6000 (48G) GPUs, with a batch size of 4 per GPU. Our implementation is based on PyTorch 1.7.1 and TorchVision 0.8.2, running on Ubuntu 22.04.

## Evaluation Metric

For contact region segmentation evaluation, we adopt four metrics proposed by Chen et al. (Chen et al. 2023a): SC-Acc., C-Acc., mIoU, and wIoU. SC-Acc. measures the proportion of pixels correctly identified as "in contact" and correctly associated with corresponding body part labels. C-Acc. evaluates binary classification accuracy at pixel level for contact. mIoU is the average IoU across all categories, while wIoU is the class-weighted average IoU based on pixel proportions. For interaction detection, we adopt mAP as the evaluation criterion (Wang, Liu, and Lei 2024).

## Results

We evaluate our method on both PaIR-1 and PaIR-2 datasets. As shown in Table 1, on PaIR-1, we compare against some interaction detection methods (Trans (Tamura, Ohashi, and Yoshinaga 2021), RCLTrans (Kim, Jung, and Cho 2023), STTrans (Zhang et al. 2022), DisTrans (Wang, Liu, and Lei 2024)) and contact segmentation models (UNet (Ronneberger, Fischer, and Brox 2015), LinkNet (Chaurasia and Culurciello 2017), MANet (Fan et al. 2020), HOT (Chen et al. 2023a), PIHOT (Wang et al. 2025b)). With ResNet-50 backbone, our approach achieves best performance across all metrics—mAP (35.09), SC-Acc. (29.93), C-Acc. (50.87), mIoU (17.79), and wIoU (22.23)—surpassing the second-best method by a notable margin. **Moreover, our model maintains a compact size of only 55.6M parameters, significantly fewer than comparable methods, and achieves the fastest inference time of 75.2 ms, demonstrating superior efficiency-performance trade-off.** Performance further improves with stronger backbones like ResNet-101, Swin-S (Liu et al. 2021), and Swin-L (Liu et al. 2021).

Table 2 presents results on PaIR-2, where our method again outperforms all baselines across all evaluation metrics using ResNet-50, while retaining the lowest parameter count (55.5M), reinforcing its effectiveness and efficiency.

## Ablation Study

To evaluate the contribution of each proposed component, we conduct an ablation study on PaIR-1, with results summarized in Table 3. The baseline configuration includes only the PGCS module (without the H-O RoI Enhancer) and the interaction reasoning module (without the Mask-Guided RoI Feature). Under this setup, the model achieves 32.08 mAP,

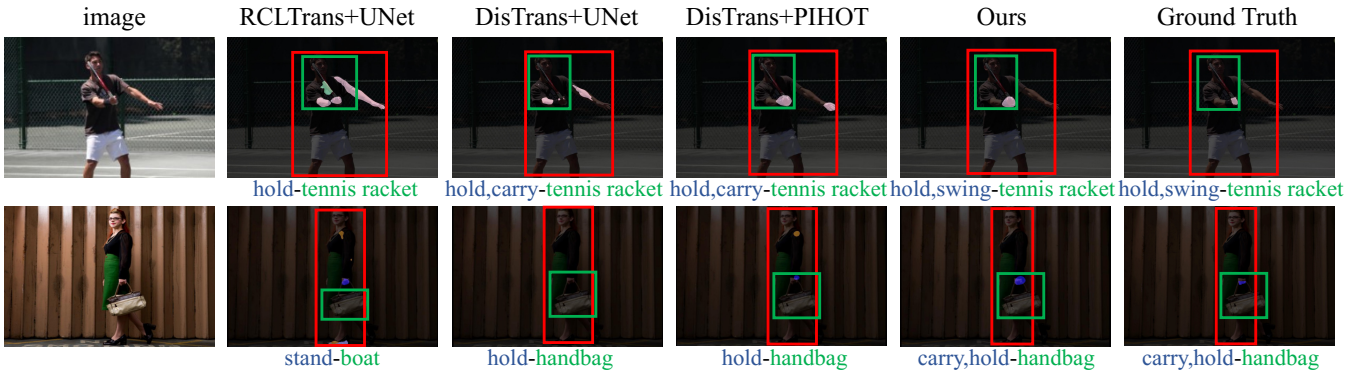


Figure 4: Visualization results. Red and green bounding boxes represent the human and object, respectively. Blue text indicates the action category, and green text indicates the object category.

| Module | mAP | SC-Acc. | C-Acc. | mIoU | wIoU |
|---|---|---|---|---|---|
| Trans (2021) | 13.70 | 29.31 | 50.19 | 17.57 | 21.98 |
| RCLTrans (2023) | 25.85 | 29.17 | 51.01 | 17.59 | 22.04 |
| STTrans (2022) | 30.21 | 27.16 | 49.85 | 17.33 | 22.17 |
| DisTrans (2024) | 31.92 | 29.76 | 50.74 | 17.52 | 21.88 |
| IIM | **35.09** | **29.93** | **50.87** | **17.79** | **22.23** |

Table 5: Experimental results of replacing different interaction detection modules when using ResNet-50 backbone. Note that only the proposed IIM is replaced, while all other components remain unchanged.

| Module | mAP | SC-Acc. | C-Acc. | mIoU | wIoU |
|---|---|---|---|---|---|
| UNet (2015) | 34.85 | 15.67 | 27.15 | 9.93 | 16.18 |
| LinkNet (2017) | 34.52 | 22.02 | 36.91 | 12.85 | 16.95 |
| MANet (2020) | 35.01 | 24.29 | 46.18 | 13.41 | 17.99 |
| HOT (2023a) | 34.78 | 24.86 | 41.63 | 14.04 | 18.70 |
| PIHOT (2025b) | 34.69 | 26.89 | 47.58 | 15.18 | 19.74 |
| PGCS | **35.09** | **29.93** | **50.87** | **17.79** | **22.23** |

Table 6: Performance of replacing different contact segmentation modules when using ResNet-50 backbone. All other components of the proposed framework remain unchanged.

23.80 SC-Acc., 43.66 C-Acc., 13.33 mIoU, and 17.10 wIoU. Incorporating the CPAM module yields noticeable gains in contact segmentation by enhancing the PGCS branch. Further addition of the H-O RoI Enhancer (+H-O) leads to improved contact performance, achieving 29.21, 50.01, 17.37, and 21.92 across the respective metrics. Finally, integrating the Mask-Guided RoI Feature (+M-G) substantially enhances interaction detection, resulting in the highest overall performance across all evaluation criteria.

We further analyze the sensitivity of the model to the hyperparameters $\alpha$ and $\beta$ in Eq. 18, as shown in Table 4. The best performance is achieved when $\alpha = 0.1$ and $\beta = 0.5$. Notably, $\alpha$ is smaller than $\beta$, primarily because the loss computed for interaction detection tends to be larger than that for segmentation. Therefore, to maintain a balanced contribution between the two tasks, it is necessary to assign a smaller weight to $\alpha$ than to $\beta$.

We also evaluate the effect of replacing the proposed IIM and PGCS modules with existing methods. As IIM and PGCS are responsible for interaction detection and contact segmentation, respectively, substituting IIM primarily affects detection performance, while replacing PGCS impacts segmentation. Table 5 reports the results of using alternative interaction detectors. As detector capability improves, overall performance increases. Similarly, Table 6 shows that our PGCS module outperforms other methods, benefiting from the integration of interaction context to improve contact region discrimination.

## Visualization

Figure 4 presents qualitative results of our method. The predicted contact regions are highlighted against a black background. The first row illustrates the action "hold and swing tennis racket", where our approach accurately identifies all interaction instances and correctly segments the contact between the right hand and the racket. In contrast, alternative methods exhibit failures in interaction recognition or contact segmentation. Similar observations in the second row further demonstrate the effectiveness of our method in both interaction understanding and contact localization.

## Conclusion

In this paper, we propose PaIR-Net, a novel and unified framework that jointly performs interaction recognition and pixel-level segmentation of contact regions. To enable effective joint modeling, we introduce several complementary and synergistic modules—CPAM, PGCS, and IIM—which collaboratively fuse high-level interaction semantics with fine-grained structural contact cues. To facilitate research in this domain, we also construct a high-quality composite dataset that, for the first time, provides joint annotations of interacting object pairs and their corresponding contact regions in real-world scenarios. Extensive experiments on benchmark datasets show that our method consistently outperforms state-of-the-art approaches across multiple evaluation metrics, highlighting the effectiveness and robustness of our proposed framework.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the GJYC program of Guangzhou under Grant 2024D01J0081, and in part by the ZJ program of Guangdong under Grant 2023QN10X455, and in part by the Fundamental Research Funds for the Central Universities under Grant 2025ZYGXZR053.

## References

Bicchi, A.; and Kumar, V. 2000. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, volume 1, 348–353. IEEE.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 361–378. Springer.

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

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

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

Chaurasia, A.; and Culurciello, E. 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, 1–4. IEEE.

Chen, Y.; Dwivedi, S. K.; Black, M. J.; and Tzionas, D. 2023a. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17100–17110.

Chen, Y.; Dwivedi, S. K.; Black, M. J.; and Tzionas, D. 2023b. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17100–17110.

Clever, H. M.; Erickson, Z.; Kapusta, A.; Turk, G.; Liu, K.; and Kemp, C. C. 2020. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6215–6224.

Cui, Z.; Lei, Y.; Wang, Y.; Yang, W.; and Qi, J. 2023. Hand gesture segmentation against complex background based on improved atrous spatial pyramid pooling. *Journal of Ambient Intelligence and Humanized Computing*, 14(9): 11795–11807.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29.

Darkhalil, A.; Shan, D.; Zhu, B.; Ma, J.; Kar, A.; Higgins, R.; Fidler, S.; Fouhey, D.; and Damen, D. 2022. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35: 13745–13758.

Desai, C.; and Ramanan, D. 2012. Detecting actions, poses, and objects with relational phraselets. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, 158–172. Springer.

Fan, T.; Wang, G.; Li, Y.; and Wang, H. 2020. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8: 179656–179665.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12943–12954.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 244–253.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, C.-H. P.; Yi, H.; Höschle, M.; Safroshkin, M.; Alexiadis, T.; Polikovsky, S.; Scharstein, D.; and Black, M. J. 2022. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13274–13285.

Kim, S.; Jung, D.; and Cho, M. 2023. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2925–2934.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 382–391.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20123–20132.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft

CoCo: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, J.; and Damen, D. 2022. Hand-object interaction reasoning. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sadeghi, M. A.; and Farhadi, A. 2011. *Recognition using visual phrases*. IEEE.

Shan, D.; Geng, J.; Shu, M.; and Fouhey, D. F. 2020. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9869–9878.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10410–10419.

Tripathi, S.; Müller, L.; Huang, C.-H. P.; Taheri, O.; Black, M. J.; and Tzionas, D. 2023. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4713–4725.

Wang, Y.; Lei, Y.; Wei, Z.; Xue, W.; Jiang, X.; Zhuang, N.; and Liu, Q. 2025a. Prompt Guidance and Human Proximal Perception for HOT Prediction with Regional Joint Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wang, Y.; Lei, Y.; Xiong, Q.; Xue, W.; Liu, Q.; and Wei, Z. 2024a. DeHOT: Reconstructing Pseudo-3D Scenes for Human-Object Contact Detection. In *2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+ AI)*, 497–500. IEEE.

Wang, Y.; Li, K.; and Lei, Y. 2022. A general multi-scale image classification based on shared conversion matrix routing. *Applied Intelligence*, 52(3): 3249–3265.

Wang, Y.; Liu, Q.; and Lei, Y. 2024. TED-Net: Dispersal Attention for Perceiving Interaction Region in Indirectly-Contact HOI Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Y.; Neng, W.; Wei, Z.; Lei, Y.; Xue, W.; Zhuang, N.; Xu, Y.; Jiang, X.; and Liu, Q. 2025b. Precision-enhanced human-object contact detection via depth-aware perspective interaction and object texture restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8187–8195.

Wang, Y.; Teng, Y.; and Wang, L. 2024. CycleHOI: Improving Human-Object Interaction Detection with Cycle Consistency of Detection and Generation. *arXiv preprint arXiv:2407.11433*.

Wang, Y.; Wei, Z.; Jiang, X.; Lei, Y.; Xue, W.; Liu, J.; and Liu, Q. 2024b. FreeA: Human-object interaction detection using free annotation labels. *arXiv preprint arXiv:2403.01840*.

Westermeier, F.; Brübach, L.; Wienrich, C.; and Latoschik, M. E. 2024. Assessing depth perception in vr and video see-through ar: A comparison on distance judgment, performance, and preference. *IEEE Transactions on Visualization and Computer Graphics*.

Wu, C.; Zhang, J.; Savarese, S.; and Saxena, A. 2015. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4362–4370.

Yang, J.; Li, B.; Zeng, A.; Zhang, L.; and Zhang, R. 2024. Open-world human-object interaction detection via multimodal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16954–16964.

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

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5534–5542.

Zare, M.; Kebria, P. M.; Khosravi, A.; and Nahavandi, S. 2024. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*.

Zhang, Y.; Pan, Y.; Yao, T.; Huang, R.; Mei, T.; and Chen, C.-W. 2022. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19548–19557.