

Precision-Enhanced Human-Object Contact Detection via Depth-Aware Perspective Interaction and Object Texture Restoration

Yuxiao Wang^{1*}, Wenpeng Neng^{2*}, Zhenao Wei¹, Yu Lei³, Weiying Xue¹, Nan Zhuang⁴, Yanwu Xu¹, Xinyu Jiang¹, Qi Liu^{1†}

¹School of Future Technology, South China University of Technology

²Ningbo Konfoong Bioinformation Tech Co., Ltd.

³School of Information Science & Technology, Southwest Jiaotong University

⁴School of Software Technology, Zhejiang University

ftwangyuxiao@mail.scut.edu.cn, nengwenpeng@gmail.com, drliuqi@scut.edu.cn

Abstract

Human-object contact (HOT) is designed to accurately identify the areas where humans and objects come into contact. Current methods frequently fail to account for scenarios where objects are frequently blocking the view, resulting in inaccurate identification of contact areas. To tackle this problem, we suggest using a perspective interaction HOT detector called PIHOT, which utilizes a depth map generation model to offer depth information of humans and objects related to the camera, thereby preventing false interaction detection. Furthermore, we use mask dilatation and object restoration techniques to restore the texture details in covered areas, improve the boundaries between objects, and enhance the perception of humans interacting with objects. Moreover, a spatial awareness perception is intended to concentrate on the characteristic features close to the points of contact. The experimental results show that the PIHOT algorithm achieves state-of-the-art performance on three benchmark datasets for HOT detection tasks. Compared to the most recent DHOT, our method enjoys an average improvement of 13%, 27.5%, 16%, and 18.5% on SC-Acc., C-Acc., mIoU, and wIoU metrics, respectively.

Code — <https://drliuqi.github.io/>

Extended version — <https://arxiv.org/abs/2412.09920>

Introduction

Human-object contact (HOT) (Chen et al. 2023) refers to the process where humans come into proximity with and touch objects, a regular part of everyday life that helps us connect with the world around us. In Figure 1, as a person skateboards, the area where their feet touch the skateboard creates a connection between the person and the skateboard. Research on HOT detection within the field of human-object interaction (HOI) focuses on accurately identifying the areas where humans and objects come into contact, allowing machines to precisely detect interactions between them. This technology has a wide range of potential applications and

can be used in various areas such as human-computer interaction (Wang, Liu, and Lei 2024), virtual reality (Illahi et al. 2023), gesture recognition (Cui et al. 2023), among others, providing new opportunities in these fields.

At present, certain studies are mainly concentrated on particular areas such as hand-object contact (Yang et al. 2024; Shiota et al. 2024), or they only take into account interactions between feet and the ground (Rempe et al. 2021). They rely on contact information as previous knowledge, which only identifies interactions between certain body parts and objects. Other studies focus on identifying interaction in particular settings (Huang et al. 2022; Shimada et al. 2022), but the applicability of these methods is restricted and may not be relevant to more intricate situations.

Current methods used for HOT fail to take into account the spatial relationship between humans and objects, resulting in possible misidentification of contact areas. In Figure 1, in a two-dimensional view, the overlap of the human, skateboard, and wall may appear like the human is touching the wall, leading to a possible misinterpretation. Moreover, the current HOT method has not effectively utilized the interaction data between humans and objects. This approach (Chen et al. 2023) simplifies HOT to a semantic segmentation task and does not take into account the frequent instances of occlusion by humans or objects. Occlusion makes it difficult for them to identify the shape of the hidden area, which in turn impacts the accuracy of determining the boundary of contact. The bench’s shape is lost because the human body covers it, as illustrated in Figure 2(a). Figure 2(b) displays the real contact labels, but the human completely obscures the segmentation boundary and object in Figure 2(c), presenting challenges in the HOT task. The intrinsic quality of the 2D image makes it challenging to deduce the meaning of the covered area based on the image’s surroundings. The occlusion phenomenon limits the visibility of important edge information in the approach, resulting in inaccuracies when determining the contact area’s location. Additionally, because the 2D image is just a projection from the 3D space, it lacks important spatial information in terms of depth. The spatial relationship is important evidence for determining if actual contact occurs in occlusion scenarios, as illustrated in Figure 2(d).

*These authors contributed equally.

†Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

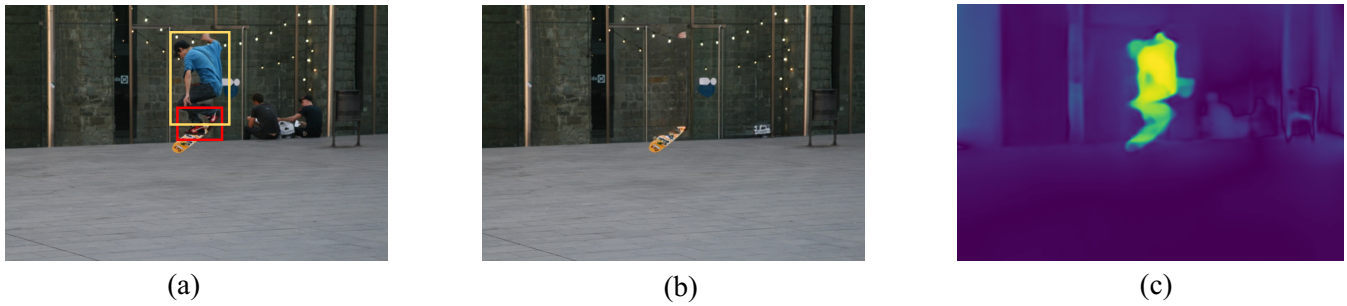


Figure 1: In Figure (a), there are overlaps between the feet and the skateboard (highlighted by the red rectangle) as well as between the body and the wall (highlighted by the yellow rectangle) when viewed from a 2D plane perspective. In 3D space, only the feet are touching the skateboard. To gather information in 3D space, it is essential to reconstruct the objects, as illustrated in Figure (b), and then extract and compare their depth maps to determine the relative spatial positions in 3D space, as depicted in Figure (c). In Figure (c), it is evident that the person and the skateboard are situated on the same level in 3D space, whereas the person and the wall are on different levels. As a result, only the person and the skateboard are touching. The proposed method is founded on this principle for detecting human-object contact.

As a result, our proposal PIHOT utilizes depth and spatial awareness to investigate the relationship between humans and objects for interactive perspectives. PIHOT restores the shape of the original object by filling in the parts that are hidden and uses depth information to analyze the spatial relationship between humans and objects, leading to a deeper understanding of the scene. Specifically, PIHOT utilizes mask dilation to correct inaccuracies resulting from human errors in mask annotation and recovers object texture information lost in occluded areas through object inpainting to improve the reliability of inputs for the HOT task. By incorporating depth maps, spatial positions, and object information, PIHOT can create a more detailed feature representation near the contact area, allowing for a comprehensive understanding and accurate segmentation of the contact area.

In conclusion, this work’s primary contributions include:

- 1) A proposed technique using image inpainting aims to address the obscured issue of objects touching each other by restoring the shape contours and texture information in the occluded regions.
- 2) We aim to capture the distribution of depth information from objects in the camera’s field of view to address the challenge of interpreting 3D information from 2D images. It leverages disparities in spatial distance between humans and objects to discern HOT interactions.
- 3) A spatial awareness method is suggested to improve the integration of the inpainting image with the depth map.
- 4) Extensive experiments demonstrate that the proposed method achieves state-of-the-art (SOTA) performance on three benchmark datasets.

Related Work

Like HOI, both HOI and HOT are necessary for identifying interactions between humans and objects. Nevertheless, there are some distinctions: HOT focuses mainly on precisely segmenting the particular contact areas between humans and objects, whereas HOI concentrates on identifying

the interaction behaviors between humans and objects. Thus, HOT can be viewed as a more detailed breakdown of the HOI. To achieve this, the research will commence with the analysis of the HOI and explore current interaction models in more detail.

Human-Object Interaction

HOI detection mainly focuses on identifying how humans interact with objects in pictures or videos. Existing HOI detection models can be divided into twofolds: two-stage pipeline (Liu, Yuan, and Chen 2020; Gao et al. 2021; Kim et al. 2021; Wang et al. 2021) versus one-stage solution (Zou et al. 2021; Liao et al. 2022; Wang, Liu, and Lei 2024; Lin et al. 2023; He et al. 2023). The first step in two-stage HOI detection involves using a pre-trained object detection model to identify both humans and objects, and then categorizing the interactions between each human and object. Separating object detection and interaction classification using these approaches results in decreased efficiency and accuracy (Liao et al. 2022; Wang, Liu, and Lei 2024). As a result, one-step methods have been developed to tackle these challenges. Contrary to the two-stage approach, the one-stage model directly converts input images into HOI triples (Zou et al. 2021). TED-Net (Wang, Liu, and Lei 2024) featured a unique triple stream designed to concentrate on the context between humans and objects, enabling direct output of the positions of humans and objects as well as their interaction classification. Nevertheless, the HOI models mentioned above do not take into account if the interaction necessitates physical contact or specify which body parts of the human are being used in the interaction.

Contact Modeling

Several projects have been suggested for studying interactions between hands and objects (Narasimhaswamy, Nguyen, and Nguyen 2020; Shan et al. 2020; Hasson et al. 2019; Zhang et al. 2022; Yang et al. 2024; Shiota et al. 2024). The study by (Hasson et al. 2019) examined how to effectively apply constraints on reconstructing both hands

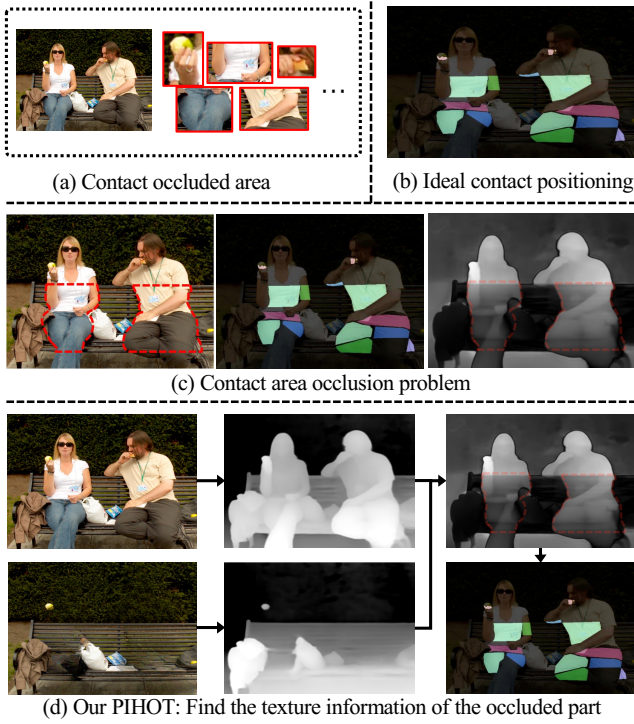


Figure 2: Description of the issue with occlusion in the contact area. (a) The red rectangle on the right highlights the area where the contact is obscured in the input image. This phenomenon is natural when looking at a 2D image, which leads to reduced visibility of important edge details. As a result, the model finds it difficult to accurately identify contact areas. (b) Ideal contact positioning. (c) The problem of contact area. (d) Our proposed method aims to first fix the obstructed objects by restoring their original appearances. Next, it analyzes the depth features of the original and repaired images to gather their spatial positioning data. The depth map has successfully recovered the edge details of the obstructed objects.

and objects together. Tekin et al. (Tekin, Bogo, and Pollefeys 2019) performed the joint estimation of 3D hand and object pose, along with the recognition of object and action categories. Nevertheless, these methods focus solely on the viewpoint of the first person and cannot determine the hand’s contact status accurately. The closely related solutions to this issue are the works by (Narasimhaswamy, Nguyen, and Nguyen 2020; Shan et al. 2020). In the video-frame dataset, Shan et al. (Shan et al. 2020) labeled the positions of hands, sides of hands, states of contact, and locations of objects in contact. In addition, Narasimhaswamy et al. (Narasimhaswamy, Nguyen, and Nguyen 2020) used two attention mechanisms to understand details about hands and surrounding objects, which helped them determine the contact states of the hands.

Chen et al. (Chen et al. 2023) presented a new challenge involving detecting human contact at the full-body level and developed the HOT dataset, which consists of 2D

heatmaps showing contact areas and labels for corresponding human body parts. They take the contact regions between humans and objects into consideration and rely on a semantic segmentation model to segment the contact areas directly, which ignores the fact that contact positions are frequently occluded by objects. Consequently, the network has difficulty accurately segmenting the contact regions, resulting in blurred contact boundaries.

Method

We aim to utilize texture and spatial position information of both humans and objects to address the existing challenges of the HOT model, particularly its underperformance in complex scenes. To do that, we suggest using PIHOT to utilize depth image and repair images for detecting HOT when objects are obscured. Specifically, PIHOT initially retrieves object details using human masks. It uses the difference in depth between the original image and the repaired object image to calculate the physical distance between humans and objects. Moreover, the mechanisms of Instances Perspective Interaction (IPI) and Instances Depth Space Interaction (IDSI) are intended to help the network prioritize occluded object information, improve the contact boundaries between humans and objects, and offer depth information of humans and objects related to the camera within the scene. The network framework is illustrated in Figure 3.

Given an image I , x is extracted after passing through the backbone network. Subsequently, x is fed into the contact branch to extract contact attention feature x_c . Meanwhile, the repaired object image I_o is obtained by the object inpainting module, and feature x_o is extracted through the backbone network. The parameters of these two backbone networks are shared. The space perception operation (SPO) module extracts depth maps from the original image I and the repaired object image I_o , resulting in d_i and d_o , respectively. By computing the difference between d_i and d_o , the spatial relative position feature map d_s between humans and objects is obtained, which is then passed into the proposed IDSI mechanism to get the output d_a . The contact perception operation (CPO) module combines x_c , x_o , d_s , and d_a for final contact feature extraction, thereby generating HOT segmentation map.

Object Inpainting Module. During OI, the main objective is to eliminate humans from the background by utilizing mask information. The dataset labels typically include this information about the masks. However, it is worth mentioning that manually annotated mask information may contain some inaccuracies, such as labeling errors. Errors in the human body mask annotations are present for each image in the dataset (Chen et al. 2023). As shown in Fig. 4(b), the original annotations (the baby blue area) do not fully cover the human body (the red dashed area). Using a dilatation rule can help extend the human body annotations to correct any small errors and ensure the entire object is captured more accurately. As a result, the mask is resized through a dilatation operation to expand the area to encompass all crucial information about humans. The dilatation is specifically de-

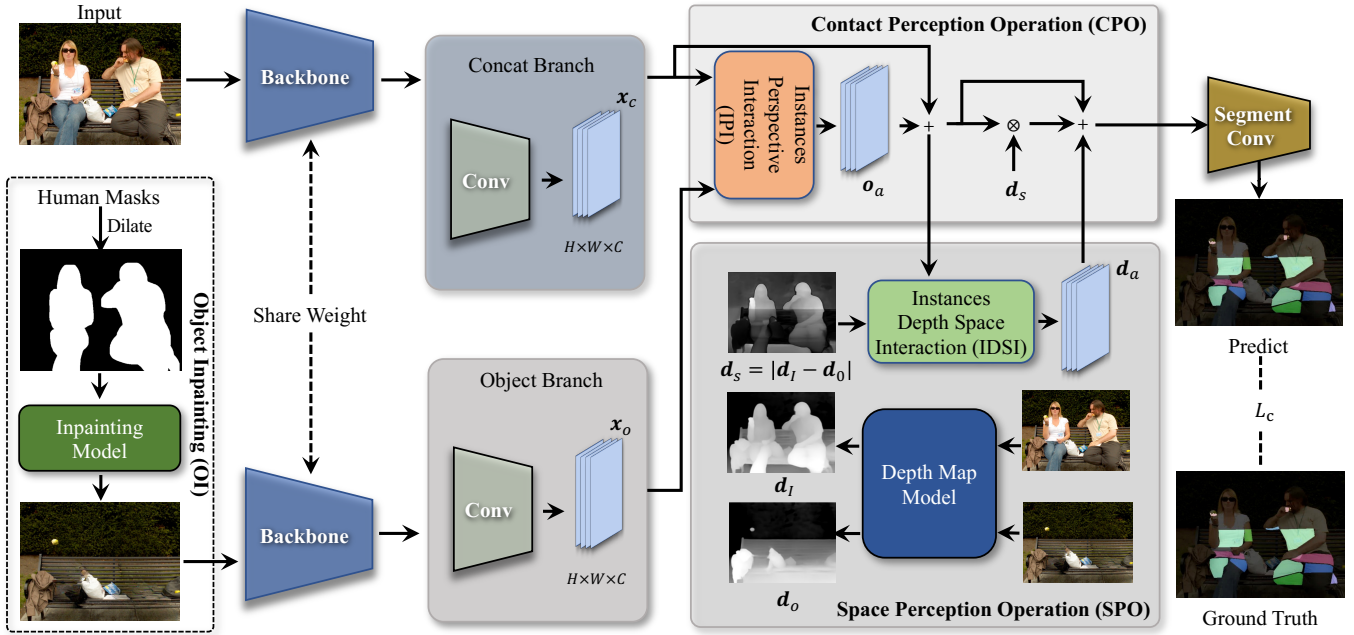


Figure 3: The overall design of the PIHOT that is being proposed. PIHOT is primarily comprised of three main components: the object inpainting (OI) module, the contact perception operation (CPO) module, and the space perception operation (SPO) module. The OI module uses the human mask to restore the texture of occluded objects. The CPO module specifically targets the contact regions of the body parts in the initial image. The SPO module extracts depth maps from the original image and the repaired object image to encode how they are positioned with each other.

terminated by the following formulas:

$$\mathbf{k} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}^{N \times N}, \quad (1)$$

$$\mathbf{m} = \text{conv2d}(\mathbf{m}, \mathbf{k}), \quad (2)$$

where N represents the row or column numbers of the matrix, conv2d denotes the 2D convolution operation, \mathbf{m} represents the human mask, and \mathbf{k} is the convolution kernel of size $N \times N$ with all elements being 1. The next step involves constructing and initializing the OI model with the parameters from the LaMa model (Suvorov et al. 2022). This step removes any human presence from the image, leaving only object information.



Figure 4: Description of human-annotated masks.

Instances Perspective Interaction Mechanism. Merging the characteristics of \mathbf{x}_c and \mathbf{x}_o in the channel is a basic method of fusion, but it does not effectively leverage object

information to restrict the features obtained from the original image. As a result, we develop the IPI mechanism. The structure diagram can be found in the supporting materials. Specifically, the feature \mathbf{x}_o undergoes expansion into a 2D matrix through a convolutional layer, leading to a matrix \mathbf{Q} . Next, the characteristics of \mathbf{x}_c are inputted into two separate convolutional layers to generate matrices \mathbf{K} and \mathbf{V} . The importance of each position, or the weights, can be obtained by multiplying matrix \mathbf{Q} by the transpose of matrix \mathbf{K} . Lastly, when these weights are multiplied by matrix \mathbf{V} , the resulting feature matrix is \mathbf{o}_a .

$$\begin{aligned} \mathbf{Q} &= \mathbf{E}(\text{conv2d}(\mathbf{x}_b)), \\ \mathbf{K} &= \mathbf{E}(\text{conv2d}(\mathbf{x}_c)), \\ \mathbf{V} &= \mathbf{E}(\text{conv2d}(\mathbf{x}_c)), \end{aligned} \quad (3)$$

$$\begin{aligned} \mathbf{o}_a &= \text{FeatureAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \end{aligned} \quad (4)$$

where $\mathbf{E}(\cdot)$ represents the expand operation. d denotes the number of columns in the matrix \mathbf{Q} , \mathbf{K} , or \mathbf{V} . The weights for each conv2d operation are randomly initialized.

Instances Depth Space Interaction. We have developed a mechanism called IDSI to make the most of the feature information provided by the depth map. It aims to utilize the proximity between humans and objects to reach the point of contact. The structure of IDSI can be found in the supporting materials. Specifically, we subtract the depth feature map

of the object d_o from the depth feature map of the original image d_i to extract the feature information related to relative spatial positions d_s . It is important to apply normalization to ensure that the depth map information falls within a specific range, which is accomplished by:

$$d_s = |d_i - d_o|, \quad (5)$$

$$d_s = \frac{d_s - \min(d_s)}{\max(d_s) - \min(d_s)}, \quad (6)$$

where $|\cdot|$ represents the absolute value operator, and \max and \min indicate taking the maximum and minimum values of the vector, respectively. This step enforces the relative spatial distance d_s to be positive and maintains the accuracy of the depth map data. In an ideal situation, by subtracting the depth maps (d_s), we would be left with only the human and the occluded object.

Nevertheless, the repaired image shows a noticeable difference in depth maps compared to the original image generated by the model, with the body appearing brighter, the background and object parts darker, and occlusions having intermediate brightness, as depicted in the right subfigure of Figure 2(c). Therefore, we are motivated to separate the occlusive and occlusion-free parts due to the increasing visibility of the boundary of occlusion.

Next, the output o_a of the IPI module is fed to the convolutional layer, and then combined with d_s to input the feature attention module, via:

$$d_a = \text{Feature_Attention}(S), \\ S = d_s, E(\text{conv2d}(o_a)), E(\text{conv2d}(o_a)) \quad (7)$$

where Feature_Attention is identical to the Eq. 4. The IDSI mechanism enables the model to incorporate depth information more thoroughly. It can enhance the understanding of spatial relationships between humans and objects, as well as improve the accuracy and reliability of human object tracking.

Contact Perception Operation Module. This module is intended for receiving the contact feature x_c , the IPI feature o_a , and the depth feature d_a to be utilized in the final segmentation feature extraction process. It specifically adheres to the following equations:

$$x = (x_c + \alpha \times o_a) \times d_s + (x_c + \alpha \times o_a) + \beta \times d_a, \quad (8)$$

where d_s represents the normalized result of subtracting the object-repaired depth map d_o from the original image depth map d_i . Values of d_s are expected to be near 0 for matching areas in the original and object-repaired images, and above 0 for areas that do not match. Therefore, by using $(x_c + \alpha \times o_a) \times d_s + (x_c + \alpha \times o_a)$, the feature $(x_c + \alpha \times o_a)$ are highlighted for the differing parts between the original and object-repaired images. Subsequently, the result is combined with the IDSI d_a to produce the comprehensive output. α and β are weight hyperparameters. x is forwarded to the segment conv module to generate the features for the HOT segmentation map. This module consists of three sets of 1×1 convolutions followed by batch normalization and ReLU activation functions. Finally, the output passes through a 1×1 convolutional layer with 18 output

channels to obtain the final segmentation probability map. These 18 channels represent different HOT categories, with a total of 17 classes excluding the background. The detailed introduction of the specific HOT categories can be found in the supporting materials.

Loss. Given image I , PIHOT will predict a probability map $\hat{y} \in \mathbb{R}^{H \times W \times C_y}$, where C_y represents the number of HOT classes, and H and W are the height and width of the probability map, respectively. \hat{y} is compared with the ground truth y to compute the cross-entropy loss. Specifically, each pixel of \hat{y} denotes a classification example, and the cross-entropy loss is computed for each corresponding pixel in y , via:

$$L_c = - \sum_{i=1}^H \sum_{j=1}^W \sum_{k=1}^{C_y} y_{i,j}^k \log \hat{y}_{i,j}^k + (1 - y_{i,j}^k) \log(1 - \hat{y}_{i,j}^k) \quad (9)$$

Results

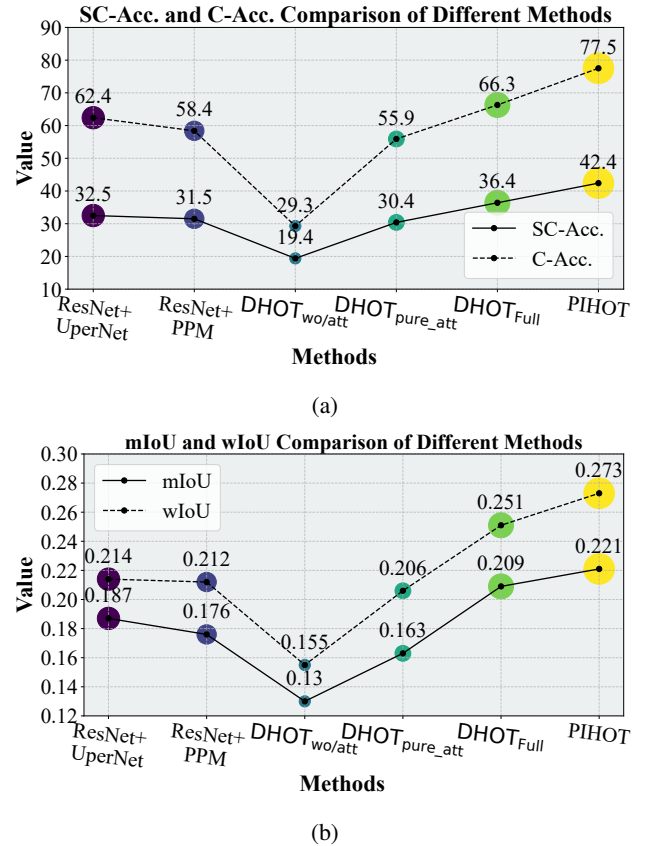


Figure 5: (a) compares the SC-Acc. and C-Acc. metrics across different methods. (b) compares the mIoU and wIoU metrics across different methods. The performance of the models is represented by the size of the dots. The proposed PIHOT method outperforms the current HOT model.

Model	HOT-Annotated				HOT-Generated			
	SC-Acc.	C-Acc.	mIoU	wIoU	SC-Acc.	C-Acc.	mIoU	wIoU
ResNet+UperNet (Xiao et al. 2018)	35.1	62.6	0.195	0.227	21.1	42.7	0.080	0.116
ResNet+PPM (Zhao et al. 2017)	34.6	61.1	0.201	0.233	21.2	41.1	0.075	0.119
DHOT _{wo/att} (Chen et al. 2023)	24.1	42.8	0.148	0.187	12.0	24.6	0.051	0.099
DHOT _{pure_att} (Chen et al. 2023)	33.8	58.4	0.189	0.237	20.3	40.1	0.077	0.113
DHOT _{Full} (Chen et al. 2023)	40.7	70.7	0.215	0.260	30.4	54.3	0.139	0.167
PIHOT	45.3 (+11%)	80.7 (+14%)	0.236 (+10%)	0.286 (+10%)	34.9 (+15%)	76.3 (+41%)	0.169 (+22%)	0.212 (+27%)

Table 1: Assessment of contact detection accuracy using the HOT-Annotated dataset.

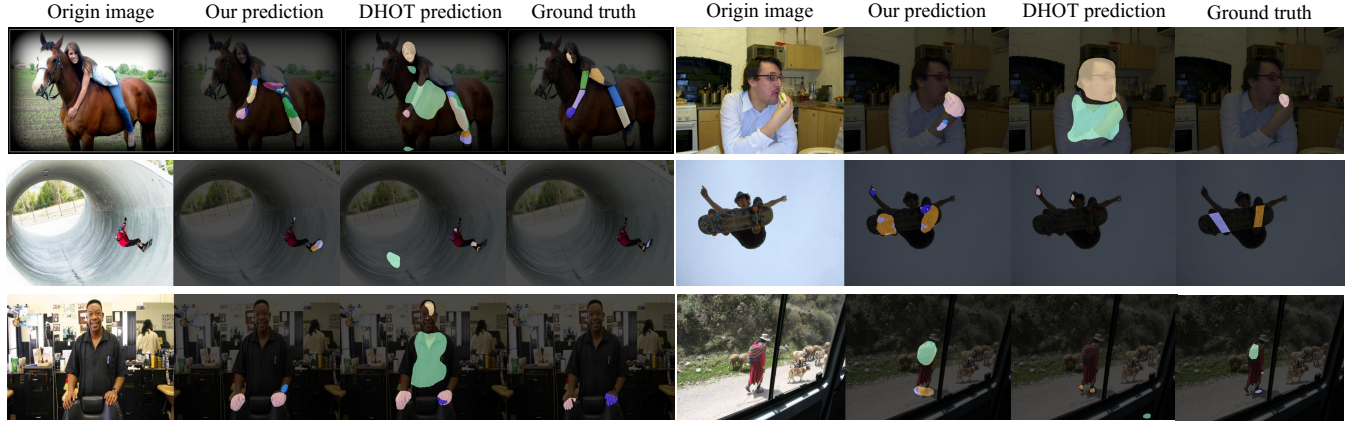


Figure 6: Visualizing the segmentation results of HOT on a portion of the dataset. Each group of pictures is separated into four sections: the original image, our method’s predicted results, the predicted results from DHOT, and ground truth.

Experimental settings

Datasets. Three datasets, namely HOT-Annotated, HOT-Generated, and Full Set, are used to evaluate the effectiveness of the proposed method. Specifically, within the HOT-Annotated subset, 5,235 images and 20,273 contact areas are sourced from V-COCO (Gupta and Malik 2015), 9,522 images and 45,645 contact areas from HAKE (Li et al. 2020), and 325 images and 1,170 contact areas from the Watch-n-Patch (Wu et al. 2015). HOT-Generated originates from data collected from the PROX (Hassan et al. 2019) and SMPL-X (Pavlakos et al. 2019). HOT-Generated includes 95,179 contact areas in 20,205 images. The Full Set dataset consists of the union of HOT-Annotated and HOT-Generated.

Evaluation metric. We follow the metrics proposed by Chen et al. (Chen et al. 2023) to evaluate the proposed method, i.e., Semantic contact accuracy (SC-Acc.), Contact accuracy (C-Acc.), Mean IoU (mIoU), and Weighted IoU (wIoU). During inference, the human mask can be generated using existing human segmentation models as an alternative to human annotation masks.

Effectiveness for regular HOT

The performance of different methods on the HOT-Annotated and HOT-Generated datasets is shown in Table 1. For HOT-Annotated, PIHOT surpasses other methods on this dataset, achieving state-of-the-art performance. Specifically, PIHOT outperforms DHOT by 11%, 10%, and 10% on

metrics SC-Acc., mIoU, and wIoU, respectively, with a notable improvement of 14% observed in metric C-Acc.. This indicates that PIHOT is more effective in identifying object contact areas. Additionally, on the HOT-Generated dataset (as shown in the right of Table 1), PIHOT achieves performance metrics of 34.9, 76.3, 0.169, and 0.212, outperforming DHOT by 15%, 41%, 22%, and 27%, respectively.

For the combined dataset Full Set, the experimental performance is illustrated in Figure 5. Figure 5(a) shows the performance of various methods on SC-Acc. and C-Acc., while Figure 5(b) shows the performance on mIoU and wIoU. We use the size of the dots to indicate performance levels to make it easier for readers to understand. It is evident that PIHOT significantly outperforms the existing HOT model, achieving SOTA performance. Specifically, PIHOT improves the metrics by 16%, 17%, 6%, and 9% compared to the second-best method, achieving 77.5, 42.4, 0.273, 0.221 in SC-Acc., C-Acc. mIoU and wIoU, respectively.

Ablation study

Table 2 illustrates the impact of different components on the performance of the proposed method. For the HOT-Annotated dataset, when using the baseline model, the results are only 40.5, 73.8, 0.214, and 0.263. When the OI module is added, there is an improvement of 7%, 1%, 0%, and 2% in the four metrics, respectively. Further performance enhancement is observed with the addition of IPI,

Row	Baseline	OI	IPI	SPO	IDSI	HOT-Annotated			
						SC-Acc.	C-Acc.	mIoU	wIoU
1	✓					40.5	73.8	0.214	0.263
2		✓				42.4(+7%)	74.8(+1%)	0.214(+0%)	0.269(+2%)
3		✓	✓			44.2(+4%)	79.4(+6%)	0.221(+3%)	0.265(-1%)
4		✓	✓	✓		44.6(+1%)	79.7(+0.4%)	0.231(+5%)	0.280(+6%)
5		✓	✓	✓	✓	45.3(+2%)	80.7(+1%)	0.236(+2%)	0.286(+2%)

OI stands for the object inpainting module, IPI represents the instances perspective interaction mechanism, SPO denotes the space perception operation, and IDSI indicates the instances depth space interaction mechanism. The baseline refers to the simple model without these four modules.

Table 2: Ablation study 1: different components on HOT-Annotated dataset.

Row	Baseline	OI	IPI	SPO	IDSI	HOT-Generated			
						SC-Acc.	C-Acc.	mIoU	wIoU
1	✓					27.7	57.1	0.135	0.177
2		✓				30.9(+12%)	67.3(+18%)	0.128(-7%)	0.191(+8%)
3		✓	✓			31.3(+1%)	71.6(+6%)	0.134(+5%)	0.176(-8%)
4		✓	✓	✓		34.4(+10%)	71.7(+0.1%)	0.169(+24%)	0.202(+15%)
5		✓	✓	✓	✓	34.9(+1%)	76.3(+6%)	0.169(+0%)	0.212(+5%)

Table 3: Ablation study 2: different components on HOT-Generated dataset.

Method	SC-Acc.	C-Acc.	mIoU	wIoU
MAE+MDENet	43.4	75.6	0.279	0.265
Lama+ZoeDepth	45.3	80.7	0.236	0.286

Table 4: The performance of different depth models and restoration models on HOT-Annotated dataset.

resulting in 4%, 6%, and 3% improvements in SC-ACC., C-ACC., and mIoU, respectively. However, there is a decrease of 1% in wIoU. The influence of incorporating SPO module on the experimental results is demonstrated in the fourth row of Table 2, showing an improvement of 1%, 0.4%, 5%, and 6%. Upon integrating the IDSI mechanism, the results reach optimum values at 45.3, 80.7, 0.236, and 0.286.

Similarly, results on the HOT-Generated are listed in Table 3. Extensive experiments indicate that the optimal performance is achieved when OI, IPI, OI, and IDSI are simultaneously utilized, resulting in 34.9, 76.3, 0.169, and 0.212 on HOT-Generated. The performances of different components on the Full Set are written in supporting materials.

Before finalizing the Inpainting Model and Depth Map model, we conducted experiments with other similar models and ultimately selected the LaMa and ZoeDepth models (Suvorov et al. 2022; Bhat et al. 2023), which achieved state-of-the-art performance. Other models (as shown in Table 4), such as MAE and MDENet (He et al. 2022), demonstrated performance scores of 43.4, 75.6, 0.279, and 0.265 on HOT-Annotation, falling short of the selected models by 4%, 6%, 6%, and 2% across the four metrics, respectively.

Visualization

To visually compare the performance of the proposed method against the SOTA method, we conducted visualizations on some images from the dataset. Figure 6 respectively

demonstrate the result images on the HOT dataset. The proposed method accurately segments the contact area between the human and the object, even in cases where the object is partially hidden from view. In the same way, our approach surpasses the most recent DHOT method in segmentation effectiveness.

Limitations and future directions

PIHOT may be limited to extreme occlusion scenarios or dynamic scenes. Future directions may lie on a unified multi-task design to jointly optimize human prediction, depth estimation, and HOT detection, or a novel HOT detection framework for 3D real-time video streams.

Conclusion

This study introduces a technique called PIHOT that aims to solve the problem of inaccurate segmentation in the HOT task caused by occlusion between humans and objects. Initially, PIHOT uses human masks to identify the areas where humans are present and then restores the obstructed parts of the objects. Second, it uses the depth map to capture the spatial relationship between humans and objects, allowing the method to have a clear understanding of their spatial positioning. Moreover, IPI and IDSI mechanisms are developed to emphasize crucial contact areas between humans and objects by incorporating object information and spatial position features. Extensive experiments have shown that PIHOT outperforms all other methods on the benchmark datasets for HOT detection tasks. The concept of HOT was first introduced by (Chen et al. 2023), and it has since become a vibrant and rapidly growing research area. Our research has injected significant momentum into the development of this field.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, in part by The Taihu Lake Innocation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611004, and in part by Guangdong Science and Technology Department Grant 2024A1313010012.

References

- Bhat, S. F.; Birkel, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.
- Chen, Y.; Dwivedi, S. K.; Black, M. J.; and Tzionas, D. 2023. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17100–17110.
- 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.
- Gao, Y.; Kuang, Z.; Li, G.; Zhang, W.; and Lin, L. 2021. Hierarchical Reasoning Network for Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 30: 8306–8317.
- Gupta, S.; and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.
- Hassan, M.; Choutas, V.; Tzionas, D.; and Black, M. J. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2282–2292.
- Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11807–11816.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- He, T.; Gao, L.; Song, J.; and Li, Y.-F. 2023. Toward a Unified Transformer-Based Framework for Scene Graph Generation and Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 32: 6274–6288.
- 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.
- Illahi, G. K.; Vaishnav, A.; Kämäräinen, T.; Siekkinen, M.; and Di Francesco, M. 2023. Learning to Predict Head Pose in Remotely-Rendered Virtual Reality. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, 27–38.
- Kim, D.-J.; Sun, X.; Choi, J.; Lin, S.; and Kweon, I. S. 2021. ACP++: Action Co-Occurrence Priors for Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 30: 9150–9163.
- 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, W.-K.; Zhang, H.-B.; Fan, Z.; Liu, J.-H.; Yang, L.-J.; Lei, Q.; and Du, J. 2023. Point-Based Learnable Query Generator for Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 32: 6469–6484.
- Liu, Y.; Yuan, J.; and Chen, C. W. 2020. ConsNet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4235–4243.
- Narasimhaswamy, S.; Nguyen, T.; and Nguyen, M. H. 2020. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33: 7841–7851.
- Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.
- Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11488–11499.
- 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.
- Shimada, S.; Golyanik, V.; Li, Z.; Pérez, P.; Xu, W.; and Theobalt, C. 2022. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision*, 516–533. Springer.
- Shiota, T.; Takagi, M.; Kumagai, K.; Seshimo, H.; and Aono, Y. 2024. Egocentric action recognition by capturing hand-object contact and object state. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6541–6551.
- Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.
- Tekin, B.; Bogo, F.; and Pollefeys, M. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and in-

teractions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4511–4520.

Wang, H.; Jiao, L.; Liu, F.; Li, L.; Liu, X.; Ji, D.; and Gan, W. 2021. IPGN: Interactiveness Proposal Graph Network for Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 30: 6583–6593.

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*.

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.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Zhang, J.; Li, J.; and Lu, C. 2024. Learning a contact potential field for modeling the hand-object interaction. *IEEE transactions on pattern analysis and machine intelligence*.

Zhang, L.; Zhou, S.; Stent, S.; and Shi, J. 2022. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, 127–145. Springer.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. 2021. End-to-end human object interaction detection with HOI transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11825–11834.