2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI) | 979-8-3315-3399-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCBD-AI65562.2024.00088

DeHOT: Reconstructing Pseudo-3D Scenes for Human-Object Contact Detection

1st Yuxiao Wang
South China University of Technology
GuangZhou, China
ftwangyuxiao@mail.scut.edu.cn

2nd Yu Lei
Southwest Jiaotong University
ChengDu, China
leiyu1117@my.swjtu.edu.cn

3rd Qiwei Xiong
South China University of Technology
GuangZhou, China
xiongqiwei@vip.qq.com

4th Weiying Xue
South China University of Technology
GuangZhou, China
202320163283@mail.scut.edu.cn

5th Qi Liu
South China University of Technology
GuangZhou, China
drliuqi@scut.edu.cn

6th Zhenao Wei∗
South China University of Technology
GuangZhou, China
wza@scut.edu.cn

***Abstract*—The goal of human-object contact detection (HOT) is to identify the contact areas in human-object interactions. Currently, existing work mainly focuses on segmenting the human-object contact regions without considering the spatial relationships among the human, the object, and the background, which leads to poor segmentation performance. To address this issue, we propose a method that utilizes depth maps to accomplish the HOT task. Specifically, depth maps are used to reconstruct pseudo-3D information about the scene, thereby obtaining the relative spatial relationships between people, objects, and the background. Subsequently, the combination of depth information with the actual input can further assist in determining the image segmentation areas. It is worth noting that, to address the overfitting issue in the HOT dataset, a comprehensive data augmentation strategy is proposed, which significantly improves the model's accuracy.**

***Index Terms*—human-object contact, semantic segmentation, deep learning**

I. Introduction

Human-object contact (HOT) is used to detect the contact of human-object areas, allowing the computer to focus on the interaction center [1]. Specifically, given an image, the model outputs a segmentation map that distinguishes between non-contact and contact regions between the human and the object. Furthermore, based on the segmentation of the human-object contact area, the contact region is further subdivided to determine which body part is in contact with the object. HOT can be applied to various fields, including human-object interaction (HOI), autonomous driving, virtual reality, and more.

HOT is a further refinement of the HOI domain. HOI aims to detect human-object interactions by identifying the positions of the human and the object and classifying the type of interaction, enabling the computer to understand where the interaction occurs and what specific interaction is taking place. HOI methods can be categorized into two-stage [2] and one-stage [3] approaches. The two-stage approach first detects the

∗ Corresponding author: Zhenao Wei.

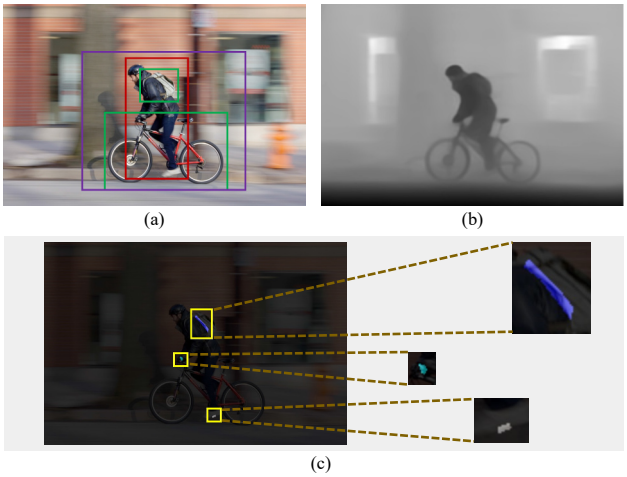

Fig. 1: Subfigure (a) shows an image of a person riding a bicycle, where the red box represents the person, the green box represents the object (the bicycle), and the purple box represents the background. From a 2D perspective, the person, the bicycle, the backpack, and the wall appear to be on the same plane. This could lead to an incorrect assessment of human-object contact, mistakenly identifying contact between the person and the background. Subfigure (b) is the depth map corresponding to subfigure (a). The depth map reveals that the person, the bicycle, and the backpack are similar in color, indicating that they are on or near the same plane, while the person and the wall are on different planes. When a person interacts with an object, they are likely to be on the same plane as the object. Therefore, the depth map makes it easier to exclude the background. Subfigure (c) shows the ground truth of the human-object contact, with the yellow box highlighting the area of contact between the person and the object. By incorporating the depth map, it is possible to reconstruct the pseudo-3D information of the scene, thereby better assisting the model in determining human-object contact.

979-8-3315-3399-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICCBD-AI65562.2024.00088

human and the object and then classifies the interaction. In contrast, the single-stage approach directly inputs information about the human, the object, and the interaction category.

Unlike HOI, HOT focuses more on identifying which specific body part is interacting with the object, thereby providing a more precise localization of the human-object interaction. However, current HOT methods are limited to simple semantic segmentation tasks and do not fully consider the spatial relationships among the human, the object, and the background. For example, in a 2D perspective, the human, the object, and the background may all appear on the same plane, but in a 3D perspective, the human and the object may be on the same level, while the human and the background could be far apart. Without considering 3D information, the model might mistakenly identify the human and the background as an interaction area, leading to a decline in performance.

To address this issue, we design a method named DeHOT, which utilizes depth maps to reconstruct pseudo-3D information, thereby enabling the model to analyze the problem from a 3D perspective. Our main contributions are as follows:

- We integrate depth maps into the analysis of human-object contact, effectively solving the problem of erroneous segmentation caused by the lack of 3D information in existing models.
- We develop a comprehensive data augmentation strategy to overcome the issue of overfitting in the dataset.
- Extensive experiments demonstrate that our method achieves significant performance improvements on two benchmark datasets: HOT-Annotated and HOT-Generated.

II. Related Work

Hand-object interaction is a form of human-computer interaction. As an intuitive means of communication, it is one of the most natural ways for humans to convey information. Consequently, hand analysis plays a significant role in daily life, technological advancement, and various other fields, holding profound research and application value. There are studies focused on detecting hands in images [4]. Moreover, various technologies such as 3D-CNNs [5] and Transformers have been employed for video interaction analysis, but handling fine details remains challenging. To address this challenge, some studies focus on estimating 3D hand gestures in hand-object interactions. The aforementioned methods explore more complex scenarios, reconstructing interactions between two hands. Furthermore, inspired by diffusion models, researchers are beginning to explore the use of language models to generate hand movements. Contact is closely related to Human-Object Interaction (HOI), as humans perform tasks by making contact with objects. Recently, Chen et al. [6] address the gap in the task of human-object contact.

III. Method

To address the issue of incorrect human-object contact detection in 2D scenes, we propose a method that incorporates depth maps to accomplish the HOI task. The network architecture is illustrated in Figure 2. In simple terms, the input image is converted into a depth map to obtain pseudo-3D scene information. Next, the depth map undergoes a normalization process to accelerate the network's convergence. Finally, the features from the depth map are fused with the features from the original image, resulting in a predicted human-object contact map.

A. Backbone

The backbone is used for feature extraction from the input image. Currently, many mainstream frameworks can be selected as the backbone, such as ResNet, ViT, and Swin Transformer. To achieve efficient training and inference, ResNet-50 is chosen as the backbone for image feature extraction in this work. Specifically, given an image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times C}$, the features $\boldsymbol{F}_b \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2048}$ are obtained after extraction by the backbone, as described by:

$$\boldsymbol{F}_b = \texttt{ResNet-50}(I). \tag{1}$$

This process allows the model to capture essential features from the image, which are then used in subsequent stages of the network for further processing and prediction. In this context, H, W, and C represent the height, width, and channels of the image, respectively.

B. Depth Map Generation

The Encoder and Decoder are used to perform encoding and decoding operations on the input image, respectively, to obtain the depth map $\boldsymbol{D} \in \mathbb{R}^{H\times W}$. This process can be represented as:

$$\boldsymbol{D} = \texttt{Decoder}(\texttt{Encoder}(\boldsymbol{I})), \tag{2}$$

where $\boldsymbol{D}$ is the resulting depth map, and $\boldsymbol{I}$ is the input image. The Encoder compresses the input image into a lower-dimensional representation, capturing essential features, while the Decoder reconstructs this representation into the depth map, which provides pseudo-3D information about the scene.

To enable the network to better fit the data, we need to normalize the depth map $\boldsymbol{D}$ to get $\boldsymbol{D}_{norm}$. This normalization ensures that the depth values are scaled between 0 and 1, which helps in stabilizing and accelerating the training process. Afterward, we perform an 8x downsampling on the depth map to obtain $\boldsymbol{D}_{c\_norm} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}}$.

C. Contact Map Prediction

The feature vector $\boldsymbol{F}_b$ obtained from the backbone is further passed through a convolutional layer to reduce the number of channels, resulting in $\boldsymbol{F}_c \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 512}$. Afterward, the normalized depth map feature D_{c_norm} is divided by 3 and then multiplied with $\boldsymbol{F}_c$. The result is then added to $\boldsymbol{F}_c$ to obtain the fused feature $\boldsymbol{F}_f \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 512}$, as described by:

$$\boldsymbol{F}_f = \boldsymbol{F}_c + \left(\boldsymbol{F}_c \times \frac{\boldsymbol{D}_{c\_norm}}{3}\right).$$

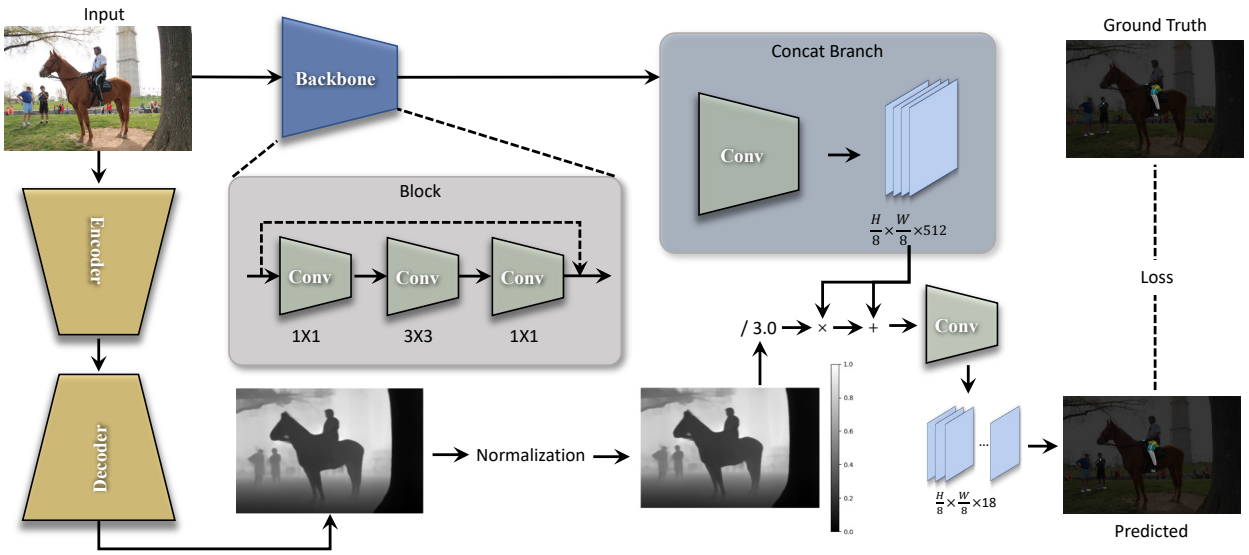

Fig. 2: Overall network architecture diagram. The input is a single image $\boldsymbol{I}$, and the output is a contact map $\boldsymbol{C}$.

Subsequently, $\boldsymbol{F}_f$ passes through three convolutional layers. The input and output channels for each convolutional layer are as follows: $\langle 512, 512\rangle$, $\langle 512, 256\rangle$, $\langle 256, 128\rangle$, and $\langle 128, 18\rangle$. The final output is the contact map $\boldsymbol{C} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 18}$, which represents the network's final prediction. Here, 18 indicates the categories of human body parts involved in the human-object contact.

D. Loss

The contact map $\boldsymbol{C}$ is compared with the ground truth $\boldsymbol{G} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 18}$ to calculate the loss. In this work, we use cross-entropy loss to measure the difference between $\boldsymbol{C}$ and $\boldsymbol{G}$.

IV. Experiments

A. Datasets

We use the HOT-Annotated and HOT-Generated datasets to validate our proposed method. The HOT-Annotated dataset contains 10,482 images for training, 2,300 images for validation, and 2,300 images for testing. The HOT-Generated dataset includes 14,144 images for training, 3,031 images for validation, and 3,030 images for testing.

B. Training Details

The network is optimized using the AdamW optimizer, with the learning rate set to 10^{-5}. The PyTorch version used is 1.11.0, and the Python version is 3.8.18. The experiments were conducted on a server with 8 A6000 GPUs (48GB each), running Ubuntu 22.04. The model is trained for 10 epochs.

C. Metrics

Chen et al. [1] proposed using four metrics to evaluate model performance: SC-Acc., C-Acc., mIoU, and wIoU. SC-Acc. represents the accuracy of human-object contact body parts, while C-Acc. indicates the contact accuracy. mIoU stands for mean Intersection over Union, and wIoU represents weighted Intersection over Union.

D. Data Augmentation

To solve the problem of network overfitting, we have designed a complete set of data augmentation methods, as follows:

- Randomly rotates the image by 90 degrees.
- Randomly flips the image, either horizontally, vertically, or both.
- Transposes the image, swapping the rows and columns, thereby changing the directionality of the image.
- The image is randomly shifted by 20% along the x and y axes and rotated by ± 45 degrees. The border area's fill value is white. There is a 75% probability of applying this transformation.
- Randomly adjusts the brightness and contrast of the image. Randomly changes the hue, saturation, and brightness of the image. Randomly shifts the red, green, and blue channels of the image. Randomly adjusts the gamma value of the image, within the range of 70 to 150.
- Blurs the image using a Gaussian filter.
- Adds Gaussian noise to the image.
- Randomly reduces the image resolution to simulate a low-quality image.
- Resizes the image to the specified width and height.
- Randomly crops a specified area from the image.

E. Results

Our proposed method, DeHOT, achieved state-of-the-art (SOTA) performance on the HOT-Annotated dataset, as shown in Table I. On this dataset, DeHOT outperformed the next best method with improvements of 10% in SC-Acc., 8% in C-Acc., 3% in mIoU, and 6% in wIoU, achieving performance scores of 44.7, 76.6, 0.222, and 0.276, respectively. As shown in Table I, on the HOT-Generated dataset, our model also attained SOTA performance in SC-Acc., C-Acc., and wIoU, with scores of 32.3, 67.1, and 0.186, respectively.

TABLE I: The results of HOT-Annotated and HOT-Generated dataset.

Model	HOT-Annotated				HOT-Generated			
	SC-Acc.	C-Acc.	mIoU	wIoU	SC-Acc.	C-Acc.	mIoU	wIoU
ResNet+UperNet	35.1	62.6	0.195	0.227	21.1	42.7	0.080	0.116
ResNet+PPM	34.6	61.1	0.201	0.233	21.2	41.1	0.075	0.119
$DHOT_{wo/att}$	24.1	42.8	0.148	0.187	12.0	24.6	0.051	0.099
$DHOT_{pure\_att}$	33.8	58.4	0.189	0.237	20.3	40.1	0.077	0.113
$DHOT_{Full}$	40.7	70.7	0.215	0.260	30.4	54.3	**0.139**	0.167
DeHOT	**44.7**	**76.6**	**0.222**	**0.276**	**32.3**	**67.1**	0.133	**0.186**

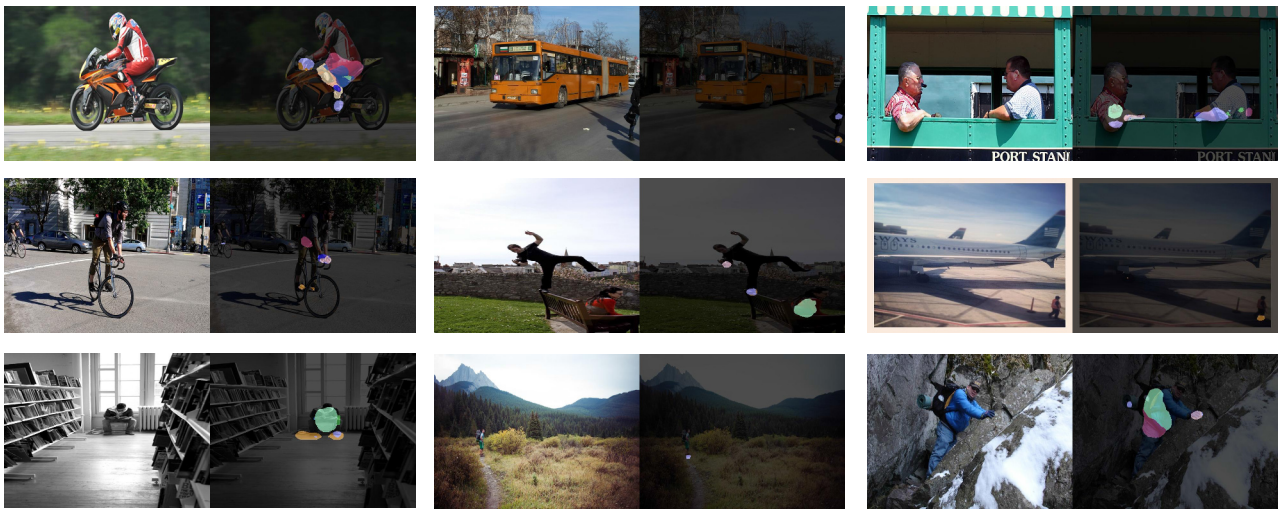

Fig. 3: Visualization of the proposed model's performance on the dataset. In each group of images, the left side shows the original image, while the right side displays the contact map output by the method.

F. Visualization

To better visualize the segmentation performance of the proposed model, we conducted a visual analysis on the dataset, as shown in Figure 3. In each group of images, the left side represents the original input, while the right side shows the contact map produced by the proposed method. It is evident that the proposed model effectively segments the human-object interaction regions. Moreover, the model does not incorrectly classify the background as human-object contact, further demonstrating the effectiveness of our approach.

V. Conclusion

In this work, we use depth maps to reconstruct pseudo-3D information of the scene, addressing the issue of incorrect segmentation of human-object contact regions in existing methods. Specifically, by feeding the input image into a depth map generation model, we obtain a pseudo-3D human-object interaction scene, capturing the spatial location information of the human, object, and background. Subsequently, to incorporate the depth map into our method, we performed a series of operations on the depth map, such as normalization and downscaling, to enable the network to quickly converge. Finally, we addressed the overfitting problem by applying data augmentation techniques. Extensive experimental results demonstrate the effectiveness of our proposed method on two datasets, achieving state-of-the-art performance.

References

[1] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas, "Detecting human-object contact in images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 100–17 110.

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

[3] Y. Wang, Q. Liu, and Y. Lei, "TED-Net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[4] L. Yang, S. Chen, and A. Yao, "Semihand: Semi-supervised hand pose estimation with consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 364–11 373.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[6] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas, "Detecting human-object contact in images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 100–17 110.