Body Part Information Additional in Multi-decoder Transformer-Based Network for Human Object Interaction Detection

Zihao Guo[1][a], Fei Li[1], Rujie Liu[1], Ryo Ishida[2] and Genta Suzuki[2]

[1]*Fujitsu Research & Development Center Co., Ltd., Beijing, China*

[2]*Fujitsu Research, Fujitsu Limited, Kawasaki, Japan*

Keywords: Human Object Interaction Detection, Transformer, Multi-decoder, Body Part Information, Channel Attention.

Abstract: Human Object Interaction Detection is one of the essential branches of video understanding. However, many complex scenes exist, such as humans interacting with multiple objects. The whole human body as the subject of interaction in the complex interaction environment may misjudge the interaction with the wrong objects. In this paper, we propose a Transformer based structure with the body part additional module to solve this problem. The Transformer structure is applied to provide powerful information mining capability. Moreover, a multi-decoder structure is adopted for solving different sub-problems, enabling models to focus on different regions to provide more powerful performance. The most important contribution of our work is the proposed body part additional module. It introduces the body part information for Human-Object Interaction(HOI) detection, which refines the subject of the HOI triplet and assists the interaction detection. The body part additional module also includes the Channel Attention module to ensure the balance between the information, preventing the model from paying too much attention to the body part or the Human-Object pair. We got better performance than the State-Of-The-Art model.

1 INTRODUCTION

Human Object Interaction Detection(HOID) means detecting 'human is doing something to the object' from an image or a video. It has been one of the cornerstones of image or video understanding. HOID includes the branches of image-based and video-based. Many papers, such as Gkioxari et al. (2015b), Ma et al. (2022) and Ji et al. (2021), Sunkesula et al. (2020) have contributed to the above two aspects, respectively. However, the majority of scholars pay more attention to the case of instance-based HOID (Gao et al., 2018; Li et al., 2020; Liao et al., 2020; Tamura et al., 2021; Zhang et al., 2021a; Zhou et al., 2022), which means that when given a single-frame picture, it is not only to detect the interactive information in the picture like the image-based HOID but also to find the position of the Human-Object pair accurately.

The instance-based HOID could be practically applied in various situations. For example, this technology could be used to determine if an athlete is committing a foul on the field of play and could be deployed at supermarket self-checkout machines to detect theft. Nevertheless, all these application scenarios have a common problem: most of the video frames captured from the real scenes show some complex situations rather than a clear composition. The people and multiple interactive objects always stack on top of each other, and even multiple people and objects interact simultaneously. These application scenarios bring a dilemma to the application of traditional HOI technology. It is difficult to judge the correct interaction pairs in two-dimensional images without depth information. Some works also involve additional information, such as language(Yuan et al., 2022; Li et al., 2022b) and graph(Zhang et al., 2021b), to increase performance. However, the information mentioned above could not directly solve the problem of the application in complex situations.

Our work aims to accurately predict the correct HOI when a person interacts with multiple objects simultaneously. Considering the complex HOI situation, the whole human body is too large for the subject to determine the interaction. Therefore, interaction detection with the fine-grained body part could detect interactive actions more accurately. For example, the body part of interaction in 'hold something' should be the hand, and the body part in 'kicking the ball' should be the foot. It is not enough to consider only part of the human body or the Human-Object pair be-

[a] https://orcid.org/0000-0001-9432-2227

Guo, Z., Li, F., Liu, R., Ishida, R. and Suzuki, G.
Body Part Information Additional in Multi-decoder Transformer-Based Network for Human Object Interaction Detection.
DOI: 10.5220/0011755300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 221-229
ISBN: 978-989-758-634-7; ISSN: 2184-4321

cause both should be used to distinguish many complex actions considering the part and Human-Object pair. Therefore, we also need to integrate the character of the interacting Human-Object pairs and consider it comprehensively.

Overall, the main contributions of our works are:

- The concept of body part detection is introduced into the detection model of human interaction to assist HOI detection;
- While introducing the body part, the whole body and objects are considered comprehensively with the body part to improve the performance, keeping the balance of attention between information characteristics of the Human-Object Pair and the body part.

2 RELATED WORKS

2.1 Review of Transformer Based HOID

Thanks to the success of Transformer models in the object detection area, i.e. DETR(Carion et al., 2020) and the other relative models(Zhu et al., 2020; Dai et al., 2021), and the powerful information mining capabilities it provides, there have been a lot of HOID models built on Transformer in recent years. The Transformer network model can analyse the relationship between all pixels in the whole image rather than be limited to a particular part, which is more suitable for HOID tasks. QPIC(Tamura et al., 2021) and HOTR(Kim et al., 2021) algorithms get good performance by directly transforming the set prediction of DETR into the prediction of HOI and cleverly setting up the loss function. By referring to the idea of Deformable-DETR(Zhu et al., 2020), the use of a deformable attention mechanism in MSTR(Kim et al., 2022) can noticeably improve the defect of the long training time of Transformer Based model, but the model accuracy is not satisfactory.

Many scholars have modified the model structure based on the characteristics of HOID tasks. AS-Net(Chen et al., 2021) uses a Transformer structure with parallel instances and interactive branches, achieving good performance and laying a foundation for developing the CDN(Zhang et al., 2021a) model with cascade structure. In addition, CDN reveals the difference between the task in HOI and the traditional target detection, and shows the advantages brought by the different work of multiple decoders. On this basis, Zhou et al. (2022) continuously increases the number of decoders and encoders, getting some good results. However, the performance growth can only partially compensate for the rapid increase in algorithm complexity, and compared with these models, we think CDN is a simple and prospective algorithm.

The CDN model achieves noticeable performance improvements with the same magnitude of parameters as the original Transformer structure. It analyses and excavates the advantages and disadvantages of one-stage and two-stage structures, whose main difference is whether the HOI is predicted once or not. CDN has extracted the essence of both one and two-stage model structures. The HOID task is divided into object detection and action classification, and different decoders are assigned to different characters for calculation, thus achieving performance improvement.

In the Transformer based architecture, the decoders can first calculate the relationship between the query vectors by the self-attention module and then find the relationship between the query vectors and the features extracted from the image by the cross-attention module. This cross-attention weight should be understood as the model's attention to some specific pixels in the image, which is also visualized several times for intuitive understanding(Carion et al., 2020; Zhang et al., 2021a), exposing the model's attention and improvements in an explainable way. In Section 4, we will visualise the cross-attention weight for qualitative analysis.

2.2 Part Information Involved Models

Body part information has been introduced into the HOID domain for a long time. Gkioxari et al. (2015a) has verified that body parts can work on action recognition effectively, and Fang et al. (2018) shows a correlation between multiple body parts corresponding to activities. The above two papers are based on the traditional CNN network structure, which may have a priori bias, and their information mining ability is poor than that of the Transformer-based structures. The model will pay more attention to the area near the convolution kernel. However, in the HOID task, its complex interaction background leads to the predominance of the Transformer structure model that can mine the relationship between pixels.

Besides, Li et al. (2020) shows that the sub-actions of each body part can be spliced into the whole person's actions, but it converts the actions of each part into entries and then deduces the whole body actions through language knowledge. The construction of this algorithm is tedious, and the training time is extended. Therefore, we propose a Transformer-based algorithm that does not require additional language information and introduces body part information to assist HOI detection.

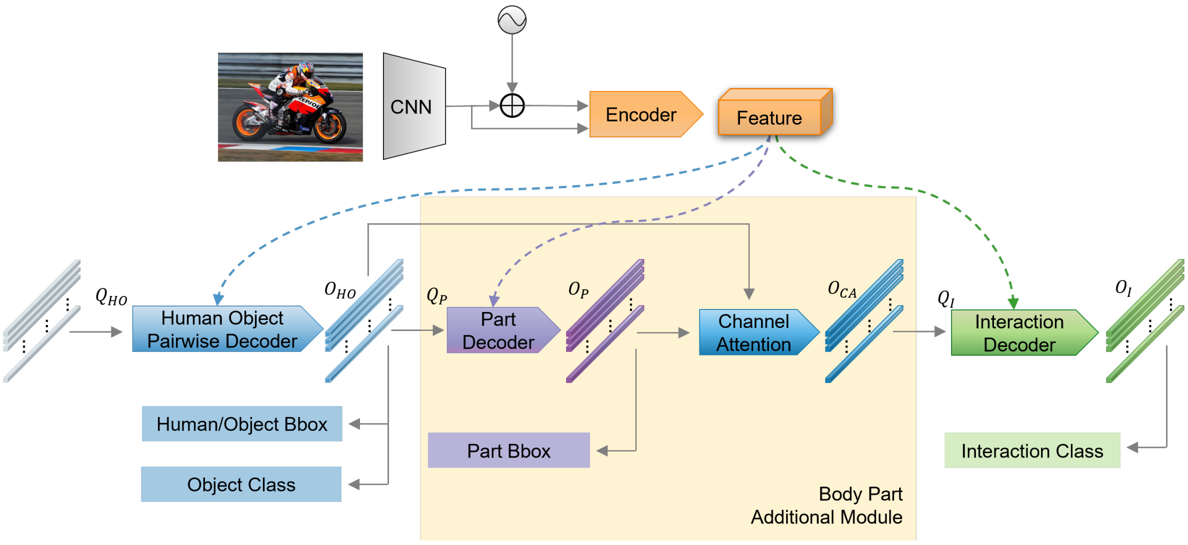


Figure 1: The framework of our model. HOI prediction is obtained by backbone, encoder and decoders from images. Different colours represent different modules. Q_{HO}, Q_P and Q_I mean the query vector input for HOPD, Part decoder and Interaction decoder separately. O_{HO}, O_P and O_I mean the output of each decoder, used to predict the HOI triplet through the Feed-Forward Network(FFN). O_{CA} is obtained by the O_{HO} and O_P processing with the Channel Attention module. Furthermore the previous module outputs O as the input Q for the next decoder.

3 METHODS

In this section, the model structure and the details of our method will be presented. Section 3.1 illustrates the frame of the model architecture. The Body Part Additional Module, which involves the body part information in the model to refine the subject of HOI, and balances the attention weight of the Human-Object pair and body part, will be revealed in Section 3.2. Moreover, the other implementation details will be introduced in Section 3.3.

3.1 Overview

The overview of our proposed model is illustrated in Figure 1. A CNN backbone and the Encoder model extract the visual feature from the input images, co-operating with the position embedding to distinguish different pixels. Different kinds of decoders for each task could achieve better performance than the single decoder for all tasks(Zhang et al., 2021a). Therefore, we apply several decoders to focus on various interested regions for mining information. The HOI prediction tasks are finished by three decoders: 1) the Human Object Pairwise Decoder(HOPD) for the human and object bounding box detection and the object classification; 2) the Part Decoder for detecting the part bounding box; 3) the Interaction Decoder for classifying the interaction. Moreover, the image feature will work in each decoder for the cross-attention module.

Under the premise of deepening the number of model layers, it is imperative to transfer the information between different modules. The information transformation method among modules could connect different modules and find more helpful information. Therefore, in the Body Part Additional Module, we adopt a Channel Attention(CA) module to combine and enhance the valuable information for the final interaction decoder. The output of the previous decoders will be used as the query vector input for the next decoder.

3.2 Body Part Additional Module

One of the main contribution of our work is the Body Part Additional Module. This module introduces information about the body part and provides guiding concerns for HOI predictions throughout the model architecture. It also ensures that both the characteristics of the Human-Object pair and the body parts will be considered through Channel Attention mechanisms rather than only one of them.

3.2.1 Body Part Information

The body part information is included in the model by the additional part decoder, which refines the subject of the HOI triplet. This decoder has the same layer

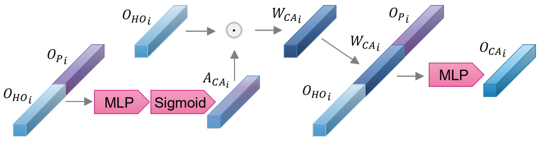


Figure 2: Details of Channel Attention module. It shows how to use i-th output sub-vector of HOPD(O_{HO_i}) and part decoder(O_{P_i}) to calculate the Channel Attention output O_{CA_i} that takes into account both information as the input of interaction decoder Q_{I_i}. A_{CA_i} represents the channel attention weight. $\odot$ means the multiplying the corresponding elements. And W_{CA_i} represents the weighted O_{HO_i}, which is gotten from multiplying the corresponding elements of O_{HO_i} and A_{CA_i}. The vectors, O_{HO_i}, W_{CA_i} and O_{P_i}, are listed together to represent *Concatenate*. The colours used here are the same as in Figure 1, and the colour changes show the fusion process.

numbers and the inside architecture as the other decoders. As it is shown in Figure 1, the output of the HOPD(O_{HO}) will be regarded as the part decoder's input, and the output of this decoder(O_P) will be sent to the body part bounding box prediction FFN. Because the input of the Part Decoder is the information used to predict the HO Pair, and the HOPD and Part Decoder share the sequence number of the Query Vector, the part location information corresponding to each HO Pair can be predicted. O_P and O_{HO} will be used as the input of the Channel Attention(CA) module. The part decoder will be used primarily to predict the location of body parts which are related to the activities, guiding overall HOI detection.

3.2.2 Channel Attention

The primary design purpose of our Channel Attention module is to make the model balance the attention of the Human-Object pair and the body part while adding the body part decoder. This module structure is inspired by Zhou et al. (2022), and the details of the Channel Attention module is shown in Figure 2.

Two ways of channel attention mechanisms are applied in this module. Firstly, the attention weight between HOPD and the output of the part decoder is computed, and the former result weights the HOPD output. Then, the weighted output between the weighted HOPD output and the output of the two decoders is calculated. However, the main difference is that Zhou et al. (2022) connects each layer of two parallel decoders through the channel attention module to enhance the capability of one of the decoders. In comparison, we take the output of the last layer of two decoders as the input and use the output for the next decoder's query vector. The formula of the Channel Attention Module is shown below:

$$O_{CA_i} = \mathrm{MLP}\left(Concat\left(O_{HO_i}, O_{P_i}, W_{CA_i}\right)\right) \tag{1}$$

$$W_{CA_i} = O_{HO_i} \cdot \mathrm{MLP}\left(\sigma\left(Concat\left(O_{HO_i}, O_{P_i}\right)\right)\right) \tag{2}$$

where O means the output of each module, and the subscript i represents the i-th sub-vector. W_{CA} means weighted HOPD output O_{CA}. $Concat\left(\cdot\right)$ means concatenating these vectors, and $\mathrm{MLP}\left(\cdot\right)$ means the vector will be calculated by the Multi-layer Perceptron. $\sigma\left(\cdot\right)$ means the sigmoid activation function, which could be able to limit the attention weight range between 0 and 1.

3.3 Implementation Details

3.3.1 Learning

We use ResNet-50(He et al., 2016) as the CNN backbone here. Only one specific body part will assist the part-relative HOI prediction in our model. Following the learning method of set-based prediction in object detection task(Carion et al., 2020), we adopt the bipartite matching before the loss calculation, which could make the set-based prediction result match the most relevant ground truth. The HOI loss function is similar to the one in QPIC(Tamura et al., 2021). For the original HOI loss, it is compute composing with the L1 loss $\mathcal{L}_b$ and GIoU loss(Rezatofighi et al., 2019) $\mathcal{L}_u$ for bounding box, cross-entropy loss $\mathcal{L}_c$ for object classification and focal loss(Lin et al., 2017) $\mathcal{L}_a$ for the activity classification. We add the bounding box location loss $\mathcal{L}_P$ for the body part during training:

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_u \mathcal{L}_u + \lambda_c \mathcal{L}_c + \lambda_a \mathcal{L}_a + \lambda_P \mathcal{L}_P \tag{3}$$

$$\begin{aligned} \lambda_P \mathcal{L}_P = \frac{1}{|\Phi|} \sum_{i=1}^{N_q} 1_{\{i \notin \Phi\}} \Big\{ \left\| \widehat{b}_i - b_{\widehat{m}(i)} \right\| \cdot \lambda_{P_1} + \\ \left[1 - \mathrm{GIoU}\left(\widehat{b}_i, b_{\widehat{m}(i)} \right) \right] \cdot \lambda_{P_2} \Big\} \end{aligned} \tag{4}$$

where the λ_b, λ_u, λ_c, λ_a and λ_P are the hyper-parameters for the balance of L1 loss, GIoU loss, cross-entropy loss, focal loss and the body part loss, respectively. λ_P is composed by $\lambda_{P_1}, \lambda_{P_2}$ for the part bounding box L1 and GIoU loss separately. Φ represents the empty set, which means this body part does not exist. b means the bounding box of the part, $\widehat{b}$ means the prediction location result and $\widehat{m}$ means the matched ground truth index. This loss function could make the body part information location loss only calculated when the specific body part exists in the image and remain the original loss function when the body part does not exist.

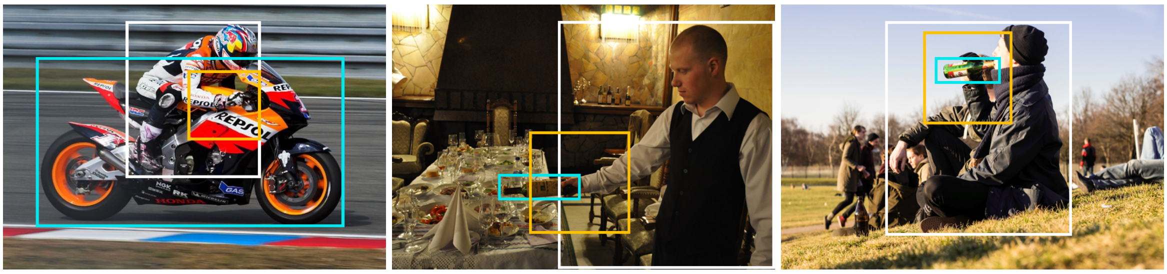


Figure 3: Some examples of the HICO-Hand-DET dataset. The colourful bounding boxes show the Human, Hand and Object ground truth in white, yellow and cyan separately. The actions shown in the figure are all 'hold', which are not marked for better effect. The best colour for visualisation.

3.3.2 Inference

The inference post-processing will fuse the outputs of each FFN to form an additional HOI set, which is composed of the location of the human and object, the object and verb class and the confidence score, as the following form $<\widehat{b}^h, \widehat{b}^o, \widehat{c}^o, \widehat{c}^v, \widehat{s}>$. The $\widehat{b}^h$, $\widehat{b}^o$, $\widehat{c}^o$, $\widehat{c}^v$ and $\widehat{s}$ mean the prediction of the human bounding box, object bounding box, object classification, action classification, confidence score separately. The confidence score is obtained by multiplying the classification score of the object and the action.

3.3.3 Auxiliary Loss

Carion et al. (2020) has pointed out that using the outputs of the decoder's each layer to predict the bounding box and calculate the loss will increase the performance. Recently, most of the Transformer based HOID models, such as Tamura et al. (2021), Zhang et al. (2021a) and Zhou et al. (2022) have followed, extending this auxiliary loss to all of the predictions, and we will also follow this setting.

4 EXPERIMENT

In this section, extensive experimentation will prove the role of body part information and channel attention. We will first introduce the dataset we used in Section 4.1. Then, the experiment setup, including the criterion metrics and the hyper-parameters setting situation, will be illustrated in Section 4.2. In Section 4.3, we will compare with another model, followed by the ablation study, which reveals the detailed improvement of each step.

4.1 HICO-Hand-DET

In order to validate the theory that body part information will directly increase performance, we focus on the hands and the hand-relative activities. We have conducted extensive experiments on a sub-dataset of the widely-used open-source dataset, HICO-DET(Chao et al., 2018).

For the hands' location, thanks to the contribution of the HAKE(Li et al., 2022a, 2020; Lu et al., 2018), the human keypoints detection algorithm has been adopted to the original HICO-DET dataset as the first step. Then, the hand location bounding boxes are drawn based on a specific ratio of the other body parts and the predicted wrist keypoints. If the head and pelvis keypoints are reliable, the side length of hand bounding box is based on the detected distance between them. If not, it will base on the distance from the wrist to the elbow. A matching algorithm is applied to ensure that each detected hand is correctly associated with the original HOI triplet labels. New quadruplets, which means $\langle Human, Hand, Object, Action \rangle$, are labelled as shown in Figure 3 for training.

As for the hand relative activities, we manually selected 50 kinds of verbs from the original 117 verbs in the HICO-DET dataset, which could be directly associated with the hand in most cases, such as 'catch, hold'. Based on the build-up methods introduced above, we composed the HICO-Hand-DET dataset with 22154 images for training and 6096 images for testing.

4.2 Setup

4.2.1 Criterion Metrics

Following the metric construction in the Chao et al. (2018) that publishes the HICO-DET dataset, we use the mean Average Precision(mAP) as the critical evaluation indicator.

Our model divides HOI into different sub-tasks in multi-decoders for prediction. In order to distinctly feel the performance of each sub-task, we build a new criterion metric, the HO mAP, to quantify the detection accuracy of the Human-Object Pair. As for the

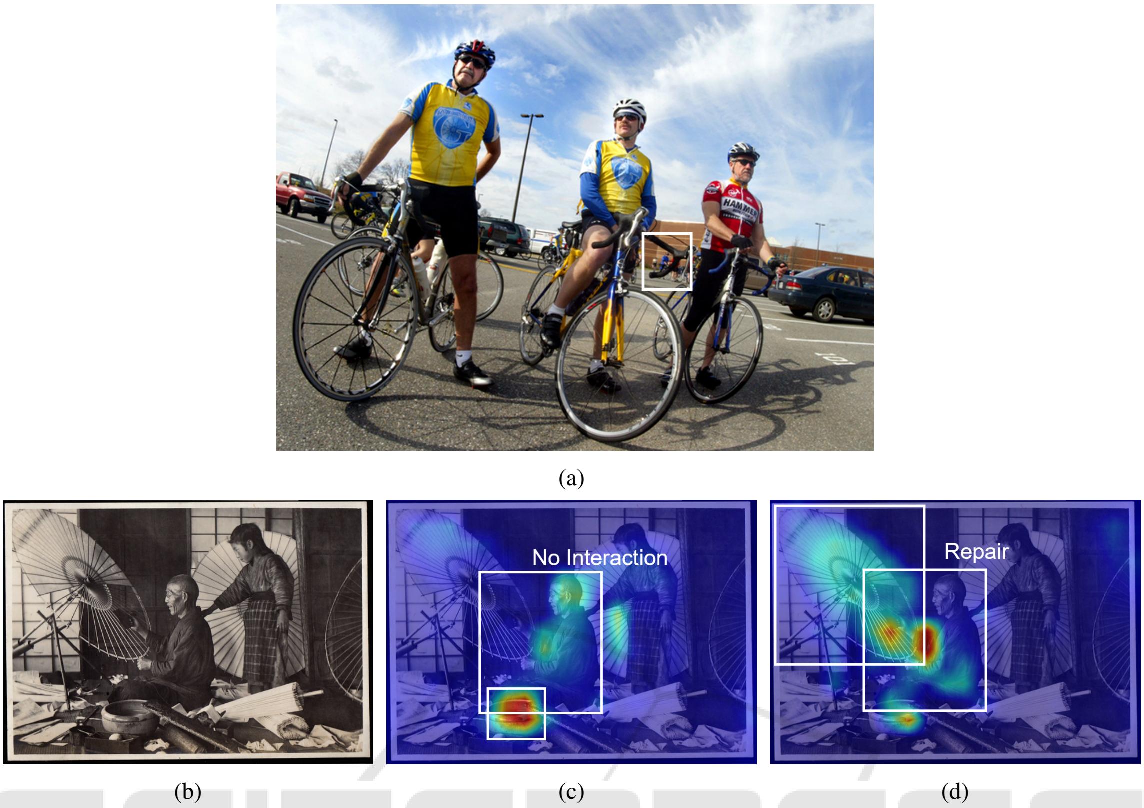


Figure 4: Improvement by body part information. (a) The prediction result without body part information. (b) The original image whose the ground truth is the older repairing the broken umbrella. (c) The cross-attention weight visualisation of the interaction decoder's last layer on CDN-S. (d) The cross-attention weight visualisation of the interaction decoder's last layer on the model involving part information.

detection accuracy of activities, since the action classification is actually based on the detection accuracy of HO, we use the correspondence between the overall mAP and HO mAP to map indirectly. The HO pair prediction will be considered positive when:

- The Intersection over Union(IoU) between the predicted and ground truth bounding box, including the human and object, is larger than 0.5;
- The predicted object category is the same as the one of ground truth.

As for the whole HOI triplet, each HOI triplet will be considered positive when the HO pair and the predicted verb category are correct. It will be used to calculate the overall mAP. Following the setting in QPIC(Tamura et al., 2021), we will only consider the HOI triplets categories introduced during training. Pair-wise non-maximal suppression(PNMS)(Zhang et al., 2021a) will be applied before the final evaluation. In contrast to the training period, the prediction result of the Part Bbox FFN will not be considered as the criterion.

4.2.2 Hyper-Parameters

The learning rate is set to 10^{-5} for the backbone and 10^{-4} for the primary model. We train this model for 90 epochs and the learning rate drops 10 times after 60 epochs. The loss function balanced weight $\lambda_b, \lambda_u, \lambda_c, \lambda_a, \lambda_{P_1}$ and λ_{P_2} are equal to 2.5, 1, 1, 1, 1, 2.5 respectively.

4.3 Comparison and Ablation Study

In order to decrease the model structure complexity, we use the original CDN-S model, which has only three layers for each decoder, as the baseline in the experiments. We train the CDN-S model on our composed HICO-Hand-DET, and the result is shown in the first row of Table 1. It is lower than the result

Table 1: Comparison and analysis the improvement of each optimization step.

Strategy	Full	Rare	Non-Rare
CDN-S	30.58	28.80	31.09
+Body Part	31.02	27.99	**31.81**
+Channel Attention	**31.43**	**30.09**	31.78

trained in HICO-DET because the hand-relative verbs may be more challenging than the other activities. According to the second row of Table 1, we could find that when only adding the Part decoder after the HOPD and summarising the outputs of the former decoders as the interaction decoder's query, the overall performance has increased by around 0.44(1.4%) from the baseline. When we used the Channel Attention(CA) to enhance the feature extracted from the former decoder to give the interaction decoder a better prior query, the overall performance could increase by 0.41 again. In these ways, the Full mAP could increase by 0.85, which means over 2.77% rise from the baseline. There is a significant increase on the Rare set, rising by 4.5% to 30.09, which is 2.1 higher than without the Channel Attention module and 1.29 higher than the baseline.

For qualitative analysis, the benefits of involving the part information and the channel attention into the HOID model, we infer the images and visualise the prediction bounding box. To find out the main attention changes after the optimisation, we also visualise the cross-attention weights of the last layer in decoders.

4.3.1 Body Part Information

The visualisation result is shown in Figure 4. According to the images, the body part information involved in the model structure could increase the interaction detection performance based on hands in two ways.

Firstly, it could suppress irrelevant interactions, solving the problem of false combining the non-interaction Human-Object pair, especially in crowd objects and multi-people situations. For example, Figure 4a shows a person who stands far away from the bicycle. Nevertheless, from the angle of the camera, the person seems to be next to the bike due to the lack of depth information. In this situation, the basic CDN-S will detect this human-bicycle pair and predict that the human is holding the bicycle, even if it is almost impossible in our minds. In contrast, when we include the part decoder in the model structure, this misleading HOI will be suppressed.

Secondly, it will also draw the attention to the regions associated with hands. As we could see in Figure 4b, an old person interacts with a broken umbrella. Suppose we visualise the attention weight of the interaction decoder. In that case, we could find in Figure 4c that the model only focuses on the whole body rather than the specific part interacting with objects, so the correct interaction could not be detected. However, Figure 4d shows that when the part information is involved in the model structure, hand relative area will be paid more attention than other parts, increasing the interaction detection accuracy.

4.3.2 Channel Attention

This section compares the predicted results with or without Channel Attention. The visualisation results are shown in Figure 5, and the quantitative analysis result is shown in Table 2. According to the results, the Channel Attention module could balance the attention weight between the Human-Object pair and the body part information.

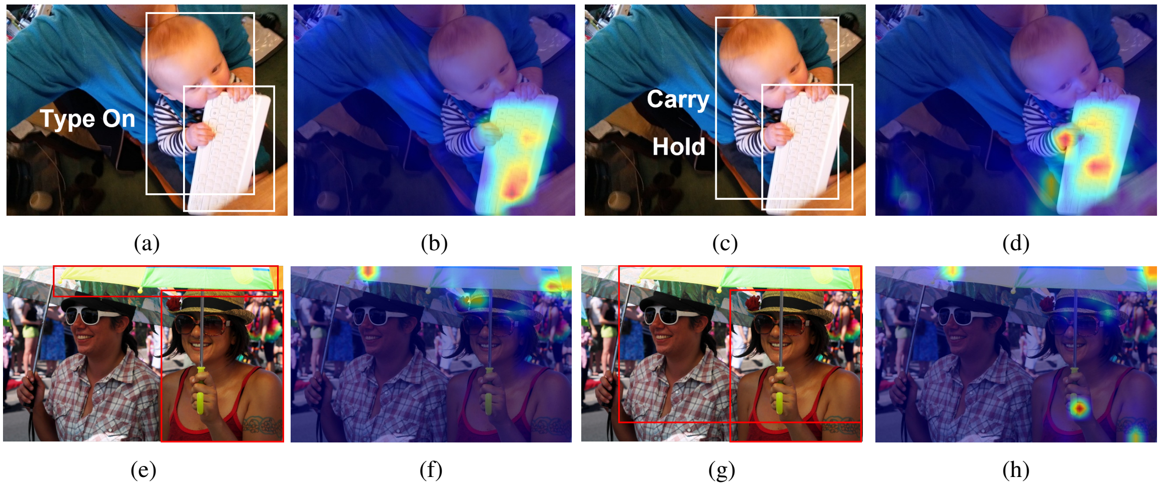
Table 2: Comparison and analysis the HO mAP of each optimization step.

Strategy	HO mAP
CDN-S	34.42
+Body Part	34.08
+Channel Attention	**34.43**

The Channel Attention module could make the model consider both the characteristics of the HO pair and the hands rather than only considering one of them. As shown in the first row of Figure 5, the interaction prediction results and the cross-attention weight of the interaction decoder of the model with or without Channel Attention reveal the improvement. As we can see in Figure 5b, the model without Channel Attention module mainly concentrates on the keyboard itself. In contrast, the model with Channel Attention also focuses on the relative position relationship between the human, hands and objects which is shown in Figure 5d. These attention weight differences lead to different action prediction results, which are wrong to predict as 'type on' in Figure 5a and correct to predict as 'hold, carry' in Figure 5c, increasing the activities' prediction accuracy.

The Channel Attention module could make the object's boundaries complete. As we can see in the second line of Figure 5, paying less attention to the object may decrease the object's integrity. Like the cross-attention weight of HOPD shown here, the model without the Channel Attention module, which could not enhance the Human-Object pair information, will only focus on some part of the whole object. In contrast, the Channel Attention module could complete the detected object bounding box. It could also be noticed from the visualisation of the cross-attention. The model only looks at the surface of the umbrella in Figure 5f, while the model involves the umbrella's handle in Figure 5h. Therefore, the prediction result of the umbrella shown in Figure 5g is more integrated than the one in Figure 5e.

After the quantitative analysis, we found that the Channel Attention module could increase the HO mAP, further proving the abovementioned deduction.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5: Improvement by Channel Attention module. (a)(e) The prediction result of the model w/o CA; (b) The cross attention weight visualisation of interaction decoder's last layer on the model w/o CA; (c)(g) The prediction result of the model w/ CA; (d) The cross attention weight visualisation of interaction decoder's last layer on the model w/ CA; (f) The cross attention weight visualisation of HOPD's last layer on the model w/o CA; (h) The cross attention weight visualisation of HOPD's last layer on the model w/ CA.

Table 2 illustrates that the HO mAP decreases by about 0.34 after involving the body part information in the model. The reduction may be due to the model paying more attention to the body part during training and the backpropagation period. When we add the Channel Attention module to balance the attention weight, we can find that the HO mAP rises to the same level as the baseline model. Under the increasing overall performance, it reveals that the model can improve the performance of interactive detection on the premise of ensuring HO accuracy.

5 CONCLUSIONS

We have proposed a Transformer based HOID model, which involves the body part information as the assistant and uses the Channel Attention module to make the model attention balance between the Human-Object pair and the body part. The body part information could refine the subject of interaction detection and the balancing mechanism could dynamically adjust the importance weight of the two kinds of feature information in the same channel. Complicated experiments have verified that the body part information could suppress the irrelevant interaction and draw attention to the part's relative area. The Channel Attention module could complete the object's boundaries and make the model consider both the characteristics of the HO pair and the hands rather than only considering one of them, increasing the accuracy of the activities' prediction. Our proposed method could achieve a better performance comparing the State-Of-The-Art baseline model. However, we only use the specific body part and the relative activities for training and testing. We plan to automatically find the most relevant body parts during HOI prediction in the future.

REFERENCES

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*, pages 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)*, pages 381–389. IEEE.

Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., and Qian, C. (2021). Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013.

Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., and Zhang, L. (2021). Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997.

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

Gao, C., Zou, Y., and Huang, J.-B. (2018). ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Gkioxari, G., Girshick, R., and Malik, J. (2015a). Actions and attributes from wholes and parts. In *Proceedings* *of the IEEE international conference on computer vision*, pages 2470–2478.

Gkioxari, G., Girshick, R., and Malik, J. (2015b). Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088.

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*, pages 770–778.

Ji, J., Desai, R., and Niebles, J. C. (2021). Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116.

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

Kim, B., Mun, J., On, K.-W., Shin, M., Lee, J., and Kim, E.-S. (2022). Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587.

Li, Y.-L., Liu, X., Wu, X., Li, Y., Qiu, Z., Xu, L., Xu, Y., Fang, H.-S., and Lu, C. (2022a). Hake: A knowledge engine foundation for human activity understanding.

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 *CVPR*.

Li, Z., Zou, C., Zhao, Y., Li, B., and Zhong, S. (2022b). Improving human-object interaction detection via phrase learning and label composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1509–1517.

Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., and Feng, J. (2020). Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lu, C., Su, H., Li, Y., Lu, Y., Yi, L., Tang, C.-K., and Guibas, L. J. (2018). Beyond holistic object recognition: Enriching image understanding with part states. In *CVPR*.

Ma, X., Nie, W., Yu, Z., Jiang, H., Xiao, C., Zhu, Y., Zhu, S.-C., and Anandkumar, A. (2022). Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Sunkesula, S. P. R., Dabral, R., and Ramakrishnan, G. (2020). Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 691–699.

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*, pages 10410–10419.

Yuan, H., Wang, M., Ni, D., and Xu, L. (2022). Detecting human-object interactions with object-guided cross-modal calibrated semantics. *arXiv preprint arXiv:2202.00259*.

Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., and Li, X. (2021a). Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220.

Zhang, F. Z., Campbell, D., and Gould, S. (2021b). Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327.

Zhou, D., Liu, Z., Wang, J., Wang, L., Hu, T., Ding, E., and Wang, J. (2022). Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.