symmetry

MDPI

Article

Human–Object Interaction Detection with Ratio-Transformer

Tianlang Wang, Tao Lu *, Wenhua Fang and Yanduo Zhang

Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430000, China
* Correspondence: lut@wit.edu.cn

Abstract: Human–object interaction (HOI) is a human-centered object detection task that aims to identify the interactions between persons and objects in an image. Previous end-to-end methods have used the attention mechanism of a transformer to spontaneously identify the associations between persons and objects in an image, which effectively improved detection accuracy; however, a transformer can increase computational demands and slow down detection processes. In addition, the end-to-end method can result in asymmetry between foreground and background information. The foreground data may be significantly less than the background data, while the latter consumes more computational resources without significantly improving detection accuracy. Therefore, we proposed an input-controlled transformer, "ratio-transformer" to solve an HOI task, which could not only limit the amount of information in the input transformer by setting a sampling ratio, but also significantly reduced the computational demands while ensuring detection accuracy. The ratio-transformer consisted of a sampling module and a transformer network. The sampling module divided the input feature map into foreground versus background features. The irrelevant background features were a pooling sampler, which were then fused with the foreground features as input data for the transformer. As a result, the valid data input into the Transformer network remained constant, while irrelevant information was significantly reduced, which maintained the foreground and background information symmetry. The proposed network was able to learn the feature information of the target itself and the association features between persons and objects so it could query to obtain the complete HOI interaction triplet. The experiments on the VCOCO dataset showed that the proposed method reduced the computational demand of the transformer by 57% without any loss of accuracy, as compared to other current HOI methods.

Keywords: human–object interaction; end-to-end; attention mechanism; transformer; symmetry; sampler; VCOCO

Citation: Wang, T.; Lu, T.; Fang, W.; Zhang, Y. Human–Object Interaction Detection with Ratio-Transformer. *Symmetry* **2022**, *14*, 1666. https://doi.org/10.3390/sym14081666

Academic Editors: Basil Papadopoulos and Dumitru Baleanu

Received: 16 June 2022
Accepted: 9 August 2022
Published: 11 August 2022

1. Introduction

The human–object interaction (HOI) task has become a research hotspot in the field of computer vision. It aims to detect the interactions that occur between persons and objects in an image in order to then generate a person–object-verb triplet.

In terms of specific tasks, HOI tasks are similar to image comprehension tasks [1], target-detection tasks [2], and action recognition tasks [3]. However, compared to image comprehension tasks, which describe the relationship between targets and targets in an image along with the related scene information, HOI tasks focus on human-based interactions that have not been influenced by other scenes. As compared to object detection tasks, which determine the prediction class and the prediction box for the object, each interaction triplet of HOI tasks must contain the respective class and box of the person and the object while also determining their interaction. As compared to action recognition tasks, which predict action duration according to a temporal sequence of actions via video-captured data, HOI tasks rely on data captured through still images rather than videos; therefore, it cannot predict the duration of interactions. However, in terms of prediction accuracy, action recognition tasks may detect the wrong action category when multiple

similar actions occur or when the motion amplitude is low, whereas HOI tasks incorporate the category and position data of the interacting objects to determine the most reasonable action; therefore, HOI detection accuracy is not impaired by motion amplitude and other similar category actions.

Previous algorithms utilized two-stage methods [4–6] for HOI detection, which require the integration of multiple detectors to complete the task. First, different detectors separately identify the persons and the objects in the image, and then the detected persons and objects are matched to determine whether an association occurs between them. Finally, the interaction triplets in which an association exists are then subjected to the category of interactive actions to form a complete interaction triplet. This method is able to train different detectors separately, which increases the efficiency of the training process. However, there are two disadvantages when using two-stage methods. First, all the subjects and objects must be matched to determine whether an interaction occurred; thus, the speed and accuracy of detection is greatly impacted by how many objects are in an image rather than the total number of interaction triplets. Second, the matching process refers to the results of other networks instead of the original information in the image; this results in the loss of a large amount of correlated data, which can reduce the detection accuracy.

The end-to-end method uses transformer networks [7] to directly detect the complete human–object-verb interaction triplets according to queries. This allows for the transformer to spontaneously locate associations between multiple objects in the image, since each queried triplet contains two different object types, as well as the association information between the objects. Compared to a two-stage method that matches all detected objects in the image, the number of queries performed by the transformer is essentially constant, which can improve efficiency in complex scenes. However, a transformer using multi-head attention can significantly increase computational demand, which should be optimized for the end-to-end approach.

Inspired by a previous method used in [8], we proposed a new method for HOI tasks that adjusted the ratio of the foreground and background information for symmetry, which substantially improved the detection speed without decreasing accuracy. The main contributions can be summarized as follows:

1. We proposed an input-feature-controlled transformer, named a ratio-transformer, to boost the performance of HOI tasks. This method filtered the input information to effectively decrease the computational demand. This was able to pre-process the input feature map by reducing the feature information to achieve symmetry between foreground and background information, and ensured the best results for the trade-off between computational performance and computational demand.
2. We innovated the HOI prediction heads to decode the query results of the ratio-transformer. This could convert the HOI query results into corresponding classes and detection boxes, and calculate the matching loss for each query.
3. The results of the comparison experiment on VCOCO [9] showed that using the ratio-transformer did not introduce any loss of accuracy and used only 43% of the computational demand of a previous transformer.

2. Related Work

HOI tasks have involved two methods: a two-stage method and an end-to-end method. The two-stage method decomposes the HOI task into two subtasks to solve: an object-detection task and an object-matching task, while the end-to-end method directly detects the HOI interaction triplets in the image. In two-stage methods, Instance-centric attention network (iCAN) [4] proposed an instance-based attention module that could selectively aggregate HOI-related information in an image in order to generate HOI detection results. InteractNet [10] determined interaction relationships based on the detection of human appearance features and their specific action density with the object of interest. Parallel point detection and matching (PPDM) [6] designed a parallel detection framework for performing interaction point detection and classification tasks. It also defined each HOI

point as a triad, human–interaction–object, for matching. This removed the risk of isolated detection frames participating in the matching of HOI triples, effectively reducing the computational demand. In end-to-end methods, an HOI transformer [8] introduced a multi-layer perception (MLP) module to replace the original feed-forward network (FFN), based on DETR [2], and designed a matching loss calculation for HOI interaction triplets. Query-Based Anchors for human–object interaction (QAHOI) [11] used a framework based on a deformable DETR for detection, which was able to extract and merge feature information from different scales and identify interaction features that were overlooked by single-scale methods, thus greatly improving its detection accuracy. An Object-guided Cross-modal Calibration Network (OCN) [12] introduced additional semantic information to guide HOI detection, and proposed a verb-semantic model (VSM) to generate semantic features and incorporate both visual and semantic features into the reference for detection results.

As compared to the fast-RCNN [13] network used in previous HOI detection tasks, transformer requires a longer training period to achieve convergence. Since the attention module projects the attention weights equally on all the pixels of the feature map during initialization, this process can require a longer training time to define meaningful locations. Furthermore, the overall computational demands of the network become more important when the input features are more precise, due to the complicated encoder model in DETR [2]. To improve the computational speed of the transformer, two main approaches have been investigated: optimizing the self-attention model and ameliorating the allocation of attention during training. To optimize the self-attention model, line-former [14] replaced the random matrix formed by self-attention with a low-rank matrix, which reduced the time and space complexity for the self-attention mechanism from $O(n^2)$ to $O(n)$, effectively improving the network efficiency. Litetransformer [15] replaced the self-attention model with a long short-range attention (LSRA) model, which guaranteed accuracy in multiple tasks using only 40% of the computational demand of the previous transformer. For attention allocation, deformable DETR [16] adopted a deformable attention mechanism, which enabled the attention during training to be distributed among the reference points and sped up the convergence of the model. PNP-DETR [17] divided the input features into foreground and background features and processed them separately, increasing the proportion of foreground information in the input transformer and reducing irrelevant information in the input.

After reviewing the previous studies, we found that the current HOI task detection network based on a transformer still had challenges to overcome, such as a long training time and high computational resource demands. By introducing a method to improve the efficiency of the transformer operations into an HOI task detection network, we effectively resolved these issues.

3. Method

Our method is shown in Figure 1. The network was divided into four parts, according to different tasks: feature extraction, feature mapping, ratio-transformer, and HOI matching. The feature map of the input image would be extracted by the feature-extraction module, a ResNet50 network-based module [18]. The feature map would then be reshaped and flattened in the feature-mapping module, before being used as the input for the ratio-transformer. The ratio-transformer divided the input information into foreground and background information, compressed and fused the background information with the foreground information to reduce the total amount of input feature information. The encoder–decoder queried the processed feature information to obtain the HOI interaction triplets in the image. HOI matcher decoded the query results into categories, prediction boxes, and verbs. Finally, the prediction results were obtained and filtering by confidence levels.

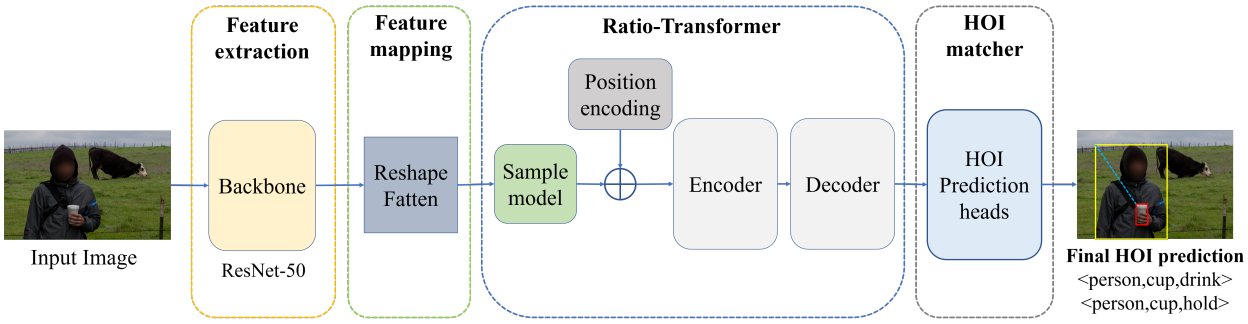

Figure 1. Overall architecture of our proposed model.

3.1. Sample Model

In an object-detection task, the attention of the trained detection network is usually focused on specific regions, as shown in Figure 2.

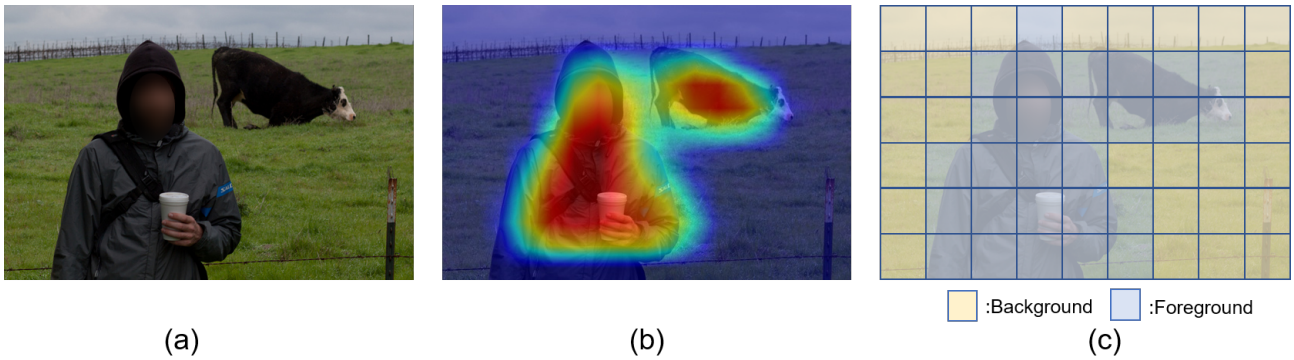

Figure 2. (**a**) Original image. (**b**) Heat map obtained after ResNet50 processing. (**c**) Segmentation of foreground and background areas based on objects.

A detection network based on a CNN will focus attention around the target, which takes less time. Based on the position of the heat map and the target box, the foreground regions related to the detection target and the irrelevant background regions can be divided in the image. However, there is typically far more background information than relevant foreground information in an image. In comparison, a transformer-based DETR network requires a longer time to converge due to the initial attention on each image being equally distributed during training [16,19,20]. If we predicted regions of background information in the input images before training and then reduced the feature input in these regions, we could effectively decrease the training time while reducing the computational demands.

Our goal was to reduce the irrelevant input information as much as possible by adding a pooling layer to the network; however, reducing the feature data would also significantly reduce the detection accuracy. Previous experimental results and related studies have shown that there is a strong correlation between detection accuracy and the amount of input feature information in HOI tasks. Some methods [11,12] used deeper backbone networks to replace the ResNet50 backbone, such as Swin-B [21] and ResNet101, which improve detection accuracy while significantly increasing the computational demand. Therefore, adding a downsampling model alone would not achieve our goal.

After referring to related studies, we realized that different processing strategies could be adopted for different regions when reducing the input features to avoid pool sampling in the target region as well. We built a new sampling network based on previous study [17]. The entire sample model is shown in Figure 3.

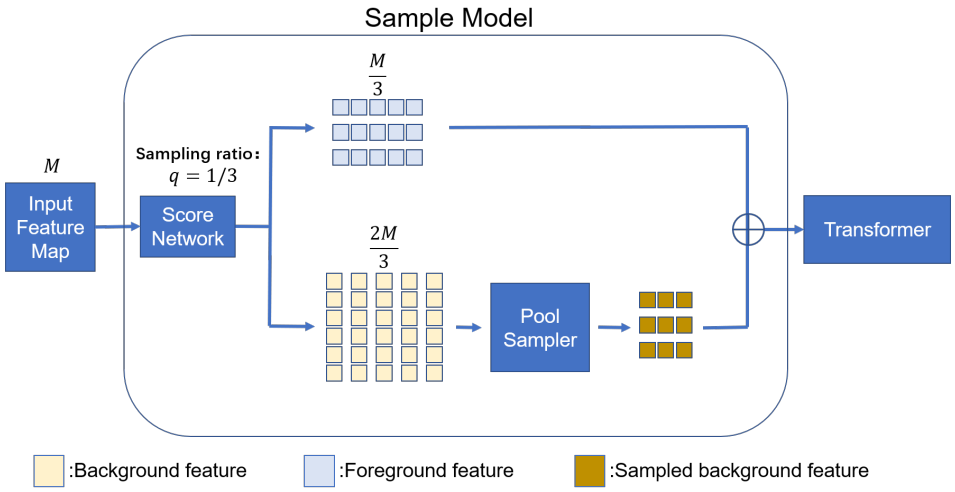

Figure 3. The overall structure of the sample model.

As shown in Figure 3, the scoring network scored and ranked the information blocks according to the amount of information they contained, as shown in (1) and (2).

$$s_{ij} = ScoringNet(f_{ij}, \theta_s), \tag{1}$$

$$[s_l, |l = 1, 2, \ldots, L], \aleph = Sort(\{s_{ij}\}). \tag{2}$$

Equation (1) uses a score network to score each block according to its information content: more information will receive a higher score, and Equation (2) ranks all blocks according to their scores, from highest to lowest. Assuming that the input features can be divided into M blocks and the sampling ratio is q, the $M * q$ blocks with the highest scores will be classified as foreground information, while the rest will be classified as background information. The foreground features will retain the original feature information, and the background features will be processed by the pool-sampler module to reduce the total feature information. The processed feature information will be used as the input for the transformer network. The sampling ratio q controls the ratio of foreground and background information. As the sampling ratio q increases, there will be more feature blocks surrounding the target, and the filtered blocks will gradually contain more obvious background information.

3.2. HOI Matcher

We referred to DETR in the matching method and the calculation of the matching loss. Due to the difference between the object-detection task and the HOI detection task, each query would receive three prediction classes and two prediction boxes, so the matching loss would consider the loss of each subclass, and we improved the original loss calculation function so that it could be applied to an HOI task.

DETR uses a query to obtain the results of each prediction: suppose it makes N queries for each image, and there are n targets in the ground truth of the image. If a single query does not successfully match, the query will be classified as "no object", and if it successfully, matches the prediction result will be output. As compared to the previous two-stage detection method, each query theoretically has a unique target, and the targets of multiple queries will not overlap. Suppose N queries are performed, the obtained data are the prediction set containing N queries, and the matching cost of the whole query is Equation (3):

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{argmin} \sum_{i}^{N} \mathcal{L}_{match}\left(\mathcal{Y}_i, \hat{\mathcal{Y}}_{\sigma(i)}\right), \tag{3}$$

where $y_i = (c_i, b_i)$ and $\hat{y}_{\sigma(i)} = \left(\hat{c}_{\sigma(i)}, \hat{b}_{\sigma(i)}\right)$. c_i, b_i denote the category and box of ground truth, respectively, while $\hat{c}_{\sigma(i)}$ and $\hat{b}_{\sigma(i)}$ denote the category and box of prediction, respectively. Therefore, the matching cost of DETR can be shown as Equation (4):

$$\mathcal{L}_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log_{\hat{p}_{\hat{\sigma}(i)}}(c_i) + \mathcal{L}_{box}\left(b_i, \hat{b}_{\hat{\sigma}}(i)\right) \right], \tag{4}$$

where $\hat{p}_{\hat{\sigma}(i)}(c_i)$ denote the category cost between c_i and $\hat{c}_{\sigma(i)}$, and $\mathcal{L}_{box}\left(b_i, \hat{b}_{\hat{\sigma}}(i)\right)$ represent the box cost between b_i and $\hat{b}_{\sigma(i)}$.

In contrast with the target-detection task, HOI detection requires the provision of a triad, including object, subject, and the verb class linking object and subject. According to the public dataset evaluation approach [5,22], the most influential of the HOI evaluation metrics were subject-class, object-class, and verb-class; each class error would lead to an overall error in the HOI instance. In object-class and verb-class, each class error would lead to an overall error in the HOI instance. The second influence on the evaluation metrics was the interaction-over-union (IOU) between the prediction frame and the ground truth. Therefore, each query in the detection network should be considered as a whole, and the matching loss of the HOI task would need to balance the weight of each loss. The matching process of the HOI-matcher task is shown in Figure 4.

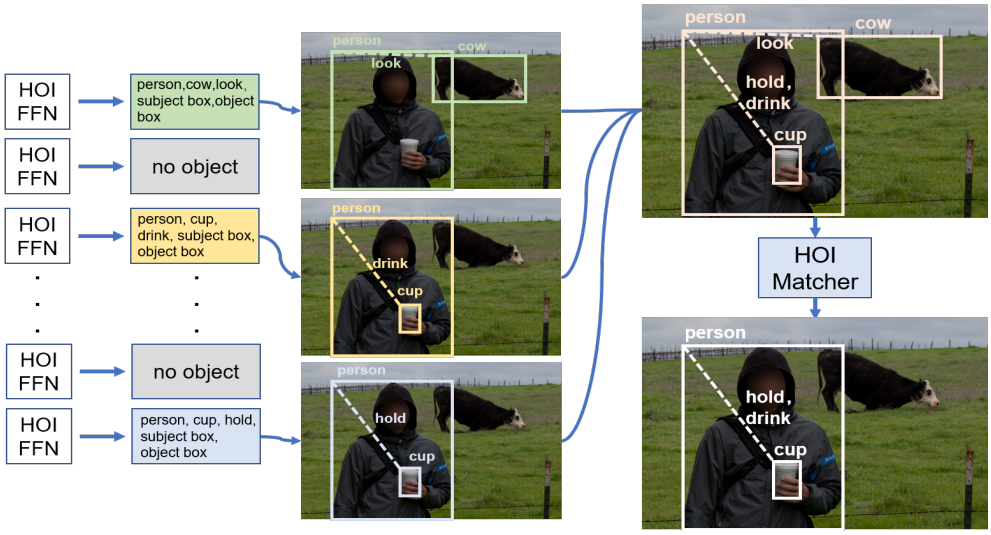

Figure 4. HOI matching process.

The interaction triplets with query results would be decoded into corresponding categories, boxes, and actions, and queries without matching results would be classified as no-object. All queried pairs would then be filtered based on confidence, and those with high confidence would be retained. In the matching process, we calculated the class and the box separately, and the overall matching cost was determined according to Equation (5):

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{argmin} \sum_{i}^{N} \left(\beta \sum_{j \in h,o,r} \alpha_j \mathcal{L}_{cls}^{j} + \gamma \sum_{k \in h,o} \mathcal{L}_{box}^{k} \right). \tag{5}$$

In Equation (5), $\mathcal{L}_{cls}$ denote the cost between the predicted category and ground truth, and $j \in h, o, r$ denote human-class, object-class, and interaction-class. The $\mathcal{L}_{box}$ denote the cost of box, and $k \in h, o$ represent human-box and object-box loss, respectively. β and γ are hyperparameters that control the proportion of category and box in the total matching cost.

4. Experiments

4.1. Experimental Setting

We conducted experiments on the VCOCO dataset, a subset of the MS-COCO [23] dataset. The images containing interactions between people and objects in COCO2014 were selected for secondary annotation. The dataset annotations included 90 target detection classes (i.e., human class), and 29 different action classes (five of these actions were human-based and did not require interactive objects) were included in the annotations.

We followed the recommended evaluation method for VCOCO datasets as the evaluation criteria. For HOI interaction triplets with objects, the evaluation method used Average Precision(AP) as the evaluation criteria. It required that each prediction accurately identified the verb, subject, and object classes, and that the IOU in the subject-prediction box, the object-prediction box, and the ground truth was greater than 0.5 to be considered correct.

We choose ResNet50 as the backbone for the feature extraction, and the transformer network part was performed according to a DETR setting, where both the encoder and decoder layers were set at six layers. The number of queries for the HOI interaction triplets on each image was set at 100. We set the HOI interaction triplet query to 300 but found no significant improvement in accuracy, and the training and testing speeds both decreased. We inferred that the HOI task may be different from the target-detection task, where there may have dozens of targets per graph in a VCOCO dataset. According to the deformable DETR research, the number of queries must far exceed the number of targets. In contrast, in an HOI task, the maximum number of HOI interaction triplets per graph would not exceed five, so it was not reasonable to substantially increase the number of queries.

For training, we used the pre-trained model provided by DETR. All training was performed on a server with two 2080TI GPUs. During the training process, the batch size of each GPU was set at two, and the total batch size at four. The training was performed for a total of 150 epochs, with a learning rate decrease at epoch 120. The HOI transformer was trained under the same environment. The test dataset was selected according to the test dataset that was delineated in the VCOCO dataset.

4.2. Results and Ablation Study

This section compares the effectiveness of our network proposed in this paper with other networks for HOI task detection on a VCOCO dataset.

Table 1 shows the results of our method on a VCOCO dataset, as compared to mainstream HOI detection methods, which include various methods for both two-stage and end-to-end.

Table 1. Comparison of computational accuracy on VCOCO dataset.

	Backbone	**AP**
Two-stage methods		
VSRL [9]	ResNet-50-FPN	31.8
TIN(RCT) [24]	ResNet-50	38.5
InteractNet [10]	ResNet-50-FPN	40.0
TIN(RCD) [24]	ResNet-50	43.2
BAR-CNN [25]	ResNet-50-FPN	43.6
GPNN [26]	ResNet-152	44.0
iCAN [4]	ResNet-50	44.7
End-to-End methods		
HOITransformer [8]	ResNet-50	45.4
ours	ResNet-50	45.6

- VSRL (Visual Semantic Role Labeling) [9]—A Fast-RCNN-based HOI detection network.
- TIN(RCT) (Transferable Interactiveness Network) [24]—Using interactivity networks to learn interactivity information from multiple HOI datasets. RCT stands for "representation extractor", "interaction classifier" and "Training with addition datasets" respectively.
- InteractNet [10]—InteractNet predicts the position of objects based on human appearance information.
- TIN(RCD) (Transferable Interactiveness Network) [24]—The same as TIN, where D stands for "interactiveness discriminator".
- BAR-CNN (Box Attention Relational CNN) [25]—BAR-CNN uses box attention to model interactions between objects.
- GPNN (Graph Parsing Neural Networks) [26]—GPNN uses Graph Parsing Neural Network to detect the presence of HOI interactions in images and videos.
- iCAN (Instance-Centric Attention Network) [4]—iCAN proposes an instance-centric attention mechanism that can highlight the region in the image where each instance is located and better identify HOI-related features.
- HOITransformer [8]—HOITransformer firstly used the Transformer to build an end-to-end network to solve HOI tasks.

We compared the computational demand using a ratio-transformer and transformer to solve the HOI task, with all other conditions being equal. As shown in Table 2, using the ratio-transformer effectively reduced the overall computational demand by 57% as compared to the transformer. This was due to the sampling and compression process during the sample mode, which reduced the total amount of information entered into the encode module, resulting in reductions in the computational demands in subsequent processes.

Table 2. Comparison of Transformer computational cost.

	Encoder FLOPs [1] (G)	**Decoder FLOPs(G)**	**Transformer FLOPs(G)**
Transformer [2]	9.53	1.89	11.42
Ratio-Transformer	3.61	1.35	4.96

[1] Floating point operations(FLOPs).

We also explored the impact of different sampling ratios on the computational accuracy and demand. As the results of Figure 5 show, the computational accuracy did not significantly improve with an increase in the sampling rate, but the computational demand significantly increased. We speculated that this could have been due to the proportion of the object region in the input image being far smaller than the overall image; the detection object would be unlikely to appear in the background region with low information content. As the sampling ratio increased, a portion of the existing background information could be converted into foreground information, which could provide a better detection reference but could also result in a false detection. Therefore, the detection accuracy fluctuated as the adoption ratio increased.

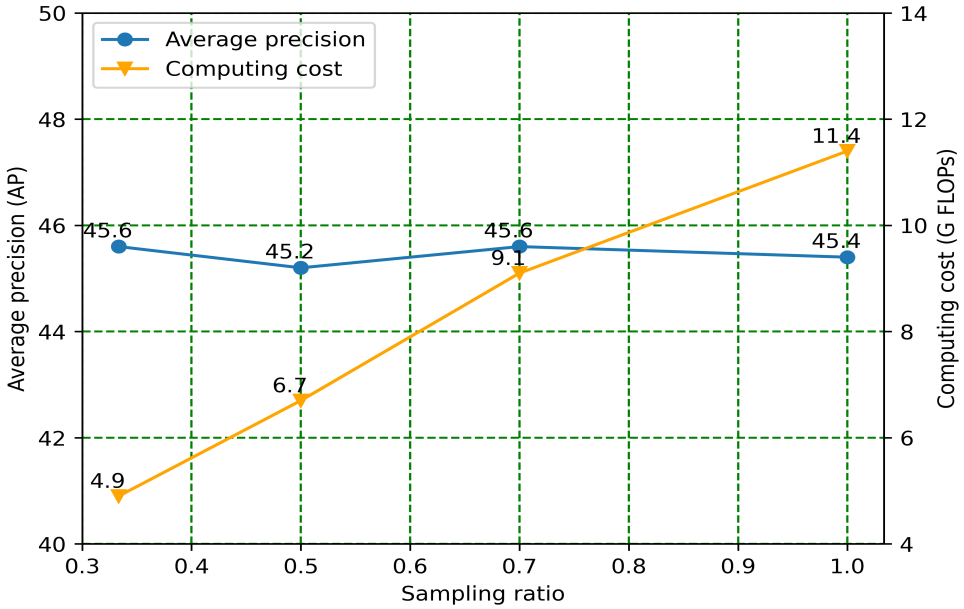

Figure 5. Computational accuracy/cost at different sampling ratios.

5. Conclusions

In this paper, we proposed a new ratio-transformer and applied it to an HOI detection task. This could select foreground information related to interaction pairs from the input features, reduce the proportion of irrelevant background information, and achieve symmetry between foreground and background information. As compared to other two-stage methods, our method substantially improved detection accuracy. As compared to current end-to-end methods [8], our method reduced the computational demand of the transformer network by 57%, with guaranteed detection accuracy (+0.2AP). In the ablation study, we verified the effect of different sampling ratios on the overall detection accuracy, as well as the computational demand. The actual results showed that the detection accuracy did not increase with increases in the sampling ratios, and we selected one-third of the sampling rate as the result of our experiments.

However, our method had some limitations. First, when a fixed sampling ratio was used, the proportion of the target regions in a few images far exceeded the sampling rate q, which could lead to a loss of effective information and affect the results. Second, determining the foreground region relied on the amount of information in the region to be divided, which could incorrectly identify irrelevant objects as foreground information, and thus affect the computational efficiency. In future studies, we intend to address and resolve these problems.

Author Contributions: Conceptualization, T.W., W.F., Y.Z. and T.L.; methodology, T.W., W.F. and T.L.; software, T.W. and W.F.; validation, T.W., W.F. and T.L.; formal analysis, T.W. and W.F.; investigation, T.W. and W.F.; writing–original draft preparation, T.W., W.F. and T.L.; writing–review and editing, T.W.; visualization, T.W.; supervision, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (62072350, 62171328), Hubei Technology Innovation Project (2019AAA045), the Central Government Guides Local Science and Technology Development Special Projects (2018ZYYD059), the High value Intellectual Property Cultivation Project of Hubei Province, the Enterprise Technology Innovation Project of Wuhan (202001602011971).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guan, Z.; Liu, K.; Ma, Y.; Qian, X.; Ji, T. Sequential Dual Attention: Coarse-to-Fine-Grained Hierarchical Generation for Image Captioning. *Symmetry* **2018**, *10*, 626. [CrossRef]
2. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
3. Yang, W.; Zhang, J.; Cai, J.; Xu, Z. Relation Selective Graph Convolutional Network for Skeleton-Based Action Recognition. *Symmetry* **2021**, *13*, 2275. [CrossRef]
4. Gao, C.; Zou, Y.; Huang, J.B. ican: Instance-centric attention network for human-object interaction detection. *arXiv* **2018**, arXiv:1808.10437.
5. Gao, C.; Xu, J.; Zou, Y.; Huang, J.B. Drg: Dual relation graph for human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 696–712.
6. Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; Feng, J. 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, Seattle, WA, USA, 14–19 June 2020; pp. 482–490.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 6000–6010.
8. Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. End-to-end human object interaction detection with hoi transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11825–11834.
9. Gupta, S.; Malik, J. Visual semantic role labeling. *arXiv* **2015**, arXiv:1505.04474.
10. Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8359–8367.
11. Chen, J.; Yanai, K. QAHOI: Query-Based Anchors for Human-Object Interaction Detection. *arXiv* **2021**, arXiv:2112.08647.
12. Yuan, H.; Wang, M.; Ni, D.; Xu, L. Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics. *arXiv* **2022**, arXiv:2202.00259.
13. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
14. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv* **2020**, arXiv:2006.04768.
15. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite transformer with long-short range attention. *arXiv* **2020**, arXiv:2004.11886.
16. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
17. Wang, T.; Yuan, L.; Chen, Y.; Feng, J.; Yan, S. PnP-DETR: Towards efficient visual analysis with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4661–4670.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Roh, B.; Shin, J.; Shin, W.; Kim, S. Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity. *arXiv* **2021**, arXiv:2111.14330.
20. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3651–3660.
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
22. Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 381–389.
23. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
24. Li, Y.L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.S.; Wang, Y.; Lu, C. Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3585–3594.
25. Kolesnikov, A.; Kuznetsova, A.; Lampert, C.; Ferrari, V. Detecting visual relationships using box attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
26. Qi, S.; Wang, W.; Jia, B.; Shen, J.; Zhu, S.C. Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 401–417.