Human-Object Interaction Detection Based on Multi-scale Attention Fusion

Qianling Wu and Yongzhao Zhan(✉)

School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China
yzzhan@ujs.edu.cn

Abstract. Human-object interaction detection is one of the key issues of scene understanding. It has widespread applications in advanced computer vision technology. However, due to the diversity of human postures, the uncertainty of the shape and size in objects, as well as the complexity of the relationship between people and objects. It is very challenging to detect the interaction relationship between people and objects. To solve this problem, this paper proposes a multi-scale attention fusion method to adapt to people and objects of different sizes and shapes. This method increases the range of attention which can more accurately judge the relationships between people and objects. Besides, we further propose a weighting mechanism to better characterize the interaction between people and close objects and express people's intention of interaction. We evaluated the proposed method on HICO-DET and V-COCO datasets, which has verified its effectiveness and flexibility as well as has achieved a certain improvement in accuracy.

Keywords: Interaction detection · Multi-Scale attention fusion

1 Introduction

Human-object interaction detection which promoted the ability of machines to understand the visual world has made tremendous progress. It is applied in many fields, such as security monitoring, service robots, sports training and image retrieval. In some cases, action recognition [1] is considered to be similar to human-object interaction detection. However, there are substantial differences between them. Action recognition, which mainly focuses on the simple classification of individual instance actions in images [2] or video clips [3], is not sufficient to describe complex visual scenes in the real world because of not consider the type of interaction between them. In contrast, human-object relationship detection provides a more specific and comprehensive description of the objective context for human activities.

Generally speaking, human-object detection first performs object detection to identify humans and objects in the images. Then we can infer the predicates in the triple group of <human, predicate, object> which is based on the detected people and objects. Finally, we can obtain the relationship between people and objects. However, there may

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 280–292, 2021.
https://doi.org/10.1007/978-3-030-87355-4_24

be multiple interaction possibilities between people and the same object. These complex and diverse scenes have brought major challenges to the detection of human-object interaction relationships.

The main task of human-object interaction detection is to obtain high-level semantic information of entities from complex semantic scenes. Specifically, due to differences in human-object instances and contexts, the visual patterns in the same human-object interaction category may be very different. In addition, since many interactions involve subtle movements of certain body parts, the appearance deviations between different categories are small. In response to these problems, although some existing methods have also made some progress, there are some shortcomings. The recently proposed methods ICAN [4] and TIN [5] both use attention mechanisms, but neither took into account the shape and size of the object nor the influence of the distance between people and objects on the detection of interaction.

To cope with the above challenges, this paper proposes a multi-scale attention fusion method, which is based on the different shapes and sizes of people and objects in images. This method introduces multi-scale attention. After obtaining the features of people and objects, this multi-scale attention is used to obtain the attention range of people and objects with different sizes and shapes which increases the receptive field. In the final interaction fusion, we increase the weight of the object, which emphasizes the interaction between people and close objects. The main contributions of this paper are as follows:

(1) Considering that the shape and size of people and objects in the image have a certain influence on the judgment of the interaction relationship, multi-scale attention is proposed to obtain different attention ranges and improve the accuracy of interaction relationship detection.
(2) We propose a weighting mechanism that emphasizes the interaction between people and objects that close to people and can better express the intention of interaction.
(3) Compared with recent relative methods, our approach achieves superior performance on both V-COCO and HICO-DET benchmarks.

2 Related Work

Object Detection. In recent years, due to the development of deep convolutional neural networks, significant progress has been made in the field of object detection. Generally speaking, object detection methods can be divided into single-stage [6–9] and two-stage [10–13]. Usually, two-stage object detection methods first generate candidate object proposal boxes and then classify and regress these proposals in the second stage. The single-stage object detection methods directly classify and regress the default anchor box at each position. Two-stage methods are generally more accurate, while single-stage methods are relatively faster.

The first step in human and object interaction detection is to correctly detect people and objects. Recently, some object detection frameworks such as R-CNN [14], Faster R-CNN [10], YOLO [6], feature pyramid network [15] and SSD [7] models can robustly detect multi-scale targets in images. We use a pre-trained Fast R-CNN model to detect people and objects. In addition, we take advantage of the idea of a Faster R-CNN region

proposal network, then we extend it to interaction detection to infer whether there is interaction in a human-object combination.

Attention. Attention has been extensive used in image captioning [31, 32, 39, 40], fine-grained classification [33], pose estimation [34], action recognition [2, 16] and human-object interaction tasks [17, 18]. The attention mechanism helped to highlight the global and local key areas in the image. In recent years, methods based on end-to-end trainable attention have been proposed to improve the performance of action recognition [19] or image classification [21]. However, these methods were designed for image-level classification tasks. Our work is based on the latest developments in attention technology and then we extend it to instance-level human-object recognition tasks, which can adapt to the difference in the size of objects in the image.

Human-Object Interaction. Among the existing human-object interaction detection methods, [21] was the first method to explore the problem of visual semantic role labeling, which located people and objects and detected the interaction between them. [22] introduced a human-centered approach, which extended the Faster R-CNN framework and added a branch to learn the interaction-specific density map at the target location. Qi et al. [23] proposed a method that treats HOI as a graph structure optimization problem by graph convolutional neural networks. Chao et al. [24] established a multi-stream network that is based on human-object regions of interest and paired interaction branches. The input of multi-stream architecture is a pre-trained detector (FPN [15]) which predicted the bounding box of the original image. Subsequent researches have extended the above-mentioned multi-stream architecture, such as the introduction of instance-centric attention [4], gesture information [5], appearance features based on context awareness [25] and deep context attention. Liang et al. [26] proposed a human-object interaction detection model, which used vision, space and graphics. They made use of graph convolution to simulate the interaction between pairs. Xu et al. [27] proposed a new region proposal network for human-object interaction detection tasks, which applied human visual cues to find objects.

3 Methods

In this section, we elaborate on the proposed multi-scale attention fusion mechanism for human-object interaction detection. The overall framework is illustrated in Fig. 1. We outline the details of each part in Sect. 3.1, then we describe our multi-scale attention fusion method in detail in Sect. 3.2.

3.1 Framework

In the traditional human-object interaction detection [4, 5], the range of attention was the same. However, people and objects have different shapes and sizes. Therefore, they should have different attention ranges, which can better characterize the interaction between people and objects. To solve this problem, we propose a multi-scale attention fusion. To accomplish interaction detection, we first perform object detection and pose

estimation to obtain human features, object features, spatial maps and human pose maps. Then human features and object features are sent to the multi-scale attention to locate the key area of interaction between people and objects, which can extract more fine-grained attention features. Finally, the attention features and appearance features of people and objects are embedded, then which are sent to the human stream and the object stream respectively. After obtaining the features, we use the interactive network to filter the human-object pair which does not have interaction first. Then the classification network combines the output of the interactive network to obtain the final interactive detection result. The interactive network and the classification network share the weight in the above process. The interactive network mainly suppresses non-interactive pairs based on visual appearance, spatial location and human posture. It also suppresses pairs of humans and objects that do not have interactive behaviors, which can reduce the resource consumption of interactive relationship detection and classification. The classification network mainly combines the non-interaction suppression of the interactive network, and then it uses a method based on a multi-scale attention fusion mechanism to classify the results of human and object interaction detection. Our overall framework is shown in Fig. 1.

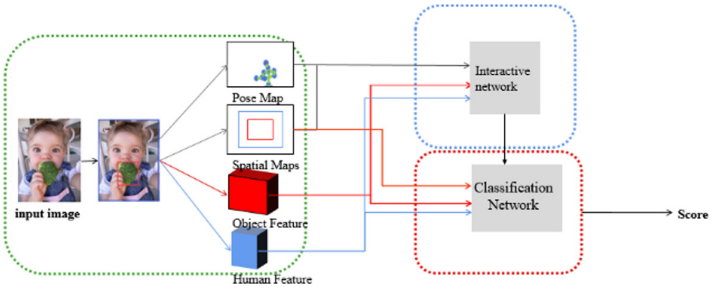

Fig. 1. Our overall network framework.

Object Detection and Pose Estimation. The prerequisite for the judgment of the human-object relationship is to know the location of the human and the object instance first. Therefore, we must detect the human instance and the object instance in the image first. The object detector used in this paper is Faster R-CNN and combines with the use of ResNet50-FPN as a feature extraction network to obtain feature maps of people and objects. In addition to convolutional features, we also extract a set of geometric features to encode the spatial configuration of each person and object instance. We start with the two binary masks proposed by people and objects, and capture the spatial configuration of the object level in their joint space, as in [4, 25]. In addition, we obtain fine-grained spatial information of humans and objects through pose maps with predicted poses. Given the joint box of each person and their counterparts, we use pose estimation [28, 29] to estimate the 17 key points of humans. Then we link the key points with lines of different gray values ranging from 0.15 to 0.95 to represent different body parts, which can implicitly encode pose features.

Interactive Network. The interactive network is mainly based on visual appearance, spatial location and human posture information to determine whether it is an interactive pair. It is composed of a human stream, object stream and spatial posture stream. The spatial posture stream is comprised of spatial graphs and posture graphs. Human stream and object stream consists of residual blocks, global average pooling layer, fully connected layer and Sigmoid function to obtain human, object and context features. The spatial posture stream including a convolutional layer, a max-pooling layer and a fully connected layer represents the positional relationship between people and objects in space and can predict actions based on people's postures. The outputs of the three streams are merged, and then we use two fully connected layers to perform the interactive discrimination. Finally, non-interaction suppression is used to determine whether there is an interaction between the human and the object, which can filter out the pairs that do not have interaction.

Classification Network. The classification network is mainly to obtain the result of the final interaction relationship detection. Figure 2 shows the classification network. It consists of pairwise stream, human stream and object stream. The human stream and object stream here are identical to those in the interactive network. The pairwise stream uses the channel attention mechanism to highlight the key areas of people and objects, which can obtain attention features. Then the channel attention features are used to encode the spatial layout between the bounding boxes of people and objects.

In real scenes, the interaction between people and close objects has more possibilities. To better characterize the interaction between people and close objects, we propose a human-object weighting mechanism for interaction relationship detection. The main idea is to increase the weight of the object score when the human stream, the object stream and the pairwise stream are fused. The interaction relationship detection final score is shown as follows:

$$S(h,o) = S_{sp} \times \left(S_h + \left(1 + (IoU(h,o))^2\right) \times S_o\right) \tag{1}$$

where S_o is the output score of the object stream, S_h is the output score of the human stream and S_{sp} is the output score of the pairwise stream. IoU represents the relation between human and object, which is the ratio of the intersection and union of the "predicted bounding box" and the "ground truth bounding box".

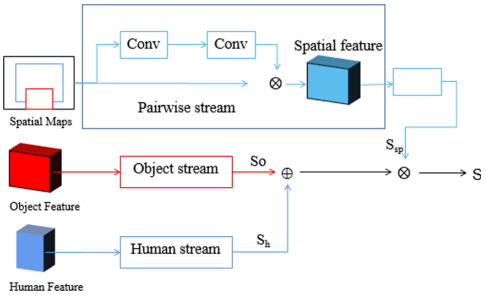

Fig. 2. Classification network

3.2 Multi-scale Attention Fusion Mechanism

In the previous detection of the interaction relationship between people and objects, the attention mechanism usually uses 1 × 1 convolution to locate the key areas of people and objects. However, it has some problems. For instance, for some relatively large objects, the range of the extracted attention features is relatively small. So the range of the receptive field mapped to the image is relatively limited and some important areas that affect the detection result of the interaction relationship may be ignored, which causes the result of the interaction relationship detection inaccurate. Therefore, we propose a multi-scale attention fusion in human stream and object stream. It will use a 1 × 1, a 1 × 3 and a 3 × 1 convolution to label the key areas of people and objects. Then, at the feature fusion stage, the attention features of different scales are fused with appearance features to obtain the multi-scale attention fusion features. The features of multi-scale attention fusion can be expressed as the following:

$$att\_o = (((head\_phi \odot fc1) \odot head\_g) \odot head\_h) \tag{2}$$

where head_h denotes features from the 3 × 1 convolution, head_g denotes feature from the 1 × 3 convolution, head_phi denotes features from the 1 × 1 convolution, fc1 denotes appearance feature of the object and ⊙ represents Hadamard product.

This multi-scale attention fusion can flexibly adapt to the shape and size of the object bounding box, which can improve the accuracy of the interaction relationship detection between people and objects. As shown in Fig. 3, the upper branch uses the pooling layer, the residual blocks, the global average pooling layer and the fully connected layer to extract the appearance features of humans and objects. The following branch uses three different convolutions to achieve the features based on the multi-scale attention fusion. Then the appearance feature and the features from the three convolutions are merged. After that, the final attention feature is obtained through 1 × 1 convolution and the fully connected layer. Finally, the obtained attention feature and the appearance feature are connected to acquire the final feature.

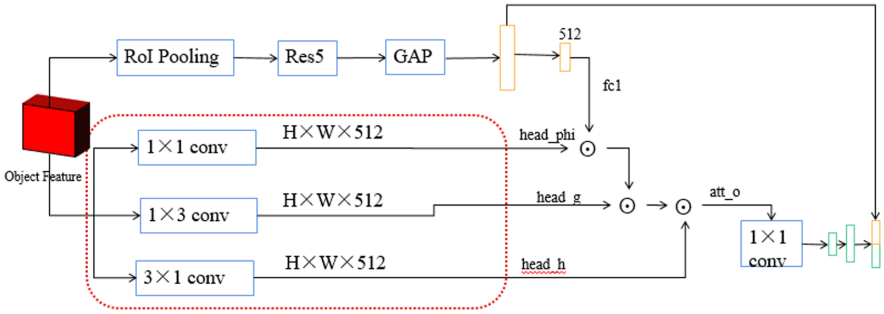

Fig. 3. Multi-scale attention fusion mechanism

4 Experimental

In this section, we first introduce the datasets and metrics, then give the implementation details of the framework. Next, we will make a quantitative and qualitative comparison with the most advanced methods. After that, we report our HOI test results. Finally, we conduct an ablation study to verify the effectiveness of the components in our framework.

4.1 Datasets and Metrics

We use two datasets HICO-DET [34] and V-COCO [21] for the detection of the human-object interaction relationship. HICO-DET [34] contains 47,776 images (38,118 in the training set and 9,658 in the test set), 600 human-object categories on 80 object categories and 117 verbs. Meanwhile, it provides more than 150k annotated instances of human-objects pairs. V-COCO [21] provides 10,346 images (2,533 images for training, 2,867 images for verification and 4,946 images for testing) and 16,199 human examples. Each person has an annotation for 29 action categories (5 of which have no matching objects).

Metrics. We use the role mean average precision [21] to measure performance. That is, when referring to the ground-truth label, and the IoU of the bounding box of the human and the object are greater than 0.5, the prediction is positive, and the HOI classification result is accurate.

4.2 Implementation Details

We use Faster R-CNN as the object detection framework and ResNet-50-FPN as the feature extractor. Using RMPE [28] and CrowdPose [29] as a pose estimator, each human pose consists of 17 key points. Then we link the key points with lines of different gray values ranging from 0.15 to 0.95 to represent different body parts. The size of the pose map is 64×64. The interactive network mainly contains three streams: human stream, object stream and spatial pose stream. Human stream and object stream are composed of residual blocks, a global average pooling layer, two 1024-size fully connected layers and a Sigmoid function. The spatial pose stream includes two convolutional layers, a max-pooling layer and two 1024-size fully connected layers. The outputs of the three streams are connected through a post-fusion strategy and interactive discrimination is performed through two 1024-size fully connected layers. The interactive network combines non-interaction suppression to determine whether there is an interaction between the human and the object. It can filter out pairs that do not have interaction between people and objects. The classification network is composed of pairwise stream, human stream and object stream. The human stream and object stream are the same as those in the interactive network. The pairwise stream uses the channel attention mechanism to highlight the key areas of people and objects to obtain attention features, which are used to encode the spatial layout between the bounding boxes of people and objects.

We conduct experiments on two datasets for a fair comparison, setting the initial learning rate to 0.001, weight decay to 0.0001 and momentum to 0.9. We use the stochastic gradient descent algorithm in the experiment. For V-COCO datasets, we set the human threshold to 0.6 and the object threshold to 0.4. For HICO-DET datasets, we set the human threshold to 0.8 and the object threshold to 0.3.

4.3 Results and Comparisons

We compare our method with the latest method on two datasets. Table 1 and Table 2 show our quantitative results on V-COCO and HICO-DET datasets respectively. From Table 1, we can see that on V-COCO datasets, the AP_{role} of our method is best and is 1.05 higher than the AP_{role} of the HBP method.

Table 1. Performance comparison on V-COCO test set.

Methods	AP_{role}
Interact [22]	40.0
GPNN [23]	44.0
ICAN [4]	45.3
TIN [5]	47.8
Cascade [36]	48.9
HBP [35]	49.05
Our method	50.1

From Table 2 we can see that the AP_{role} obtained from the default Object and the known Object on HICO-DET datasets has also been improved in most cases. Therefore, it can be said that our method based on a multi-scale attention fusion has a certain degree of feasibility and a certain degree of improvement in accuracy on both HICO-DET and V-COCO datasets.

Table 2. Performance comparison on HICO-DET test set.

Methods	Default object			Known object		
	Full	Rare	Non-rare	Full	Rare	Non-rare
Interact [22]	9.94	7.16	10.77	–	–	–
GPNN [23]	13.11	9.34	14.23	–	–	–
ICAN [4]	14.84	10.45	16.15	16.26	11.33	17.73
TIN [5]	17.22	13.51	18.32	19.38	15.38	20.57
PMFNet [37]	17.46	15.65	18.00	20.34	17.47	21.20
Wang.et [38]	19.56	12.79	21.58	22.05	15.77	23.92
Our method	19.90	14.67	21.35	22.30	15.80	24.50

Figure 4 shows our visualization results on HICO-DET. Figure 5 shows the visualization results on V-COCO. The test results prove that our method is reasonable and satisfactory.

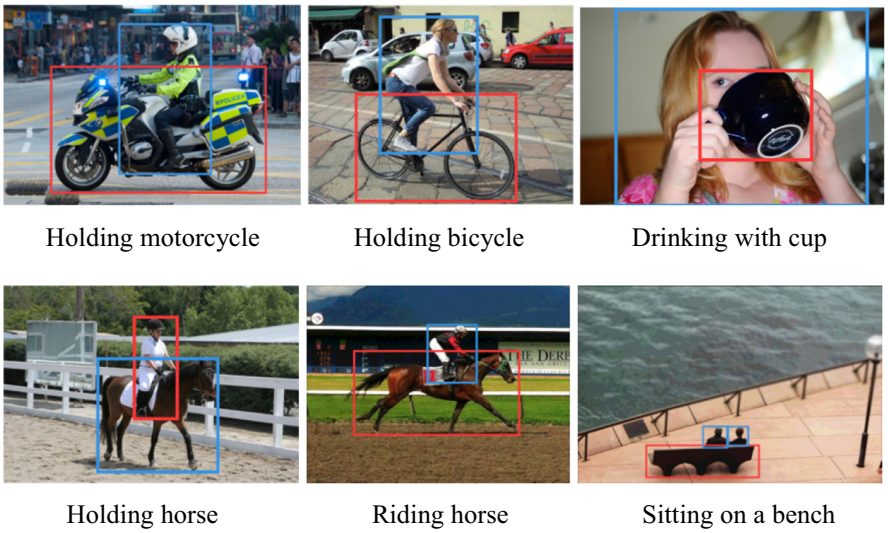

Fig. 4. Results on HICO-DET

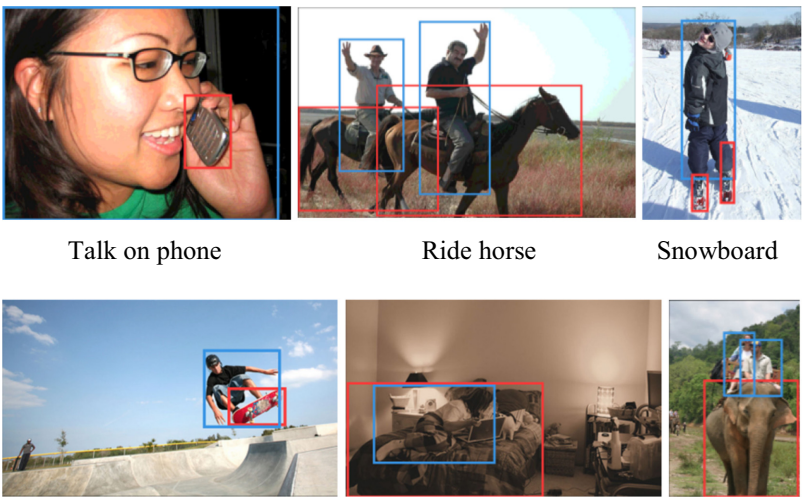

Fig. 5. Results on V-COCO

4.4 Ablation Studies

In this part, we implement several experiments to prove the effectiveness of every part of our method on V-COCO and HICO-DET datasets. Results are shown in Table 3 and Table 4.

Multi-Scale Attention Fusion Mechanism. The multi-scale attention fusion uses a 1 × 1, a 1 × 3 and a 3 × 1 convolution to locate the key areas of people and objects. Then when in the feature fusion stage, the features of different scales are dot-multiplied, which can obtain the finally multi-scale attention fusion feature. This part can flexibly adapt to the shape and size of the object, thereby improving the accuracy of interaction detection between people and objects.

Fusion. The interaction between people and close objects is more possible. If the distance between the object and the human is closer, the weight of the object should be greater. To better express people's intention of interaction and characterize the interaction between people and close objects, we also propose a method of fusion that is based on a weighting mechanism. From Table 4, we can see that our method gets the best performance.

Table 3. Ablation study on V-COCO datasets

Components	AP_{role}
Baseline	47.8
Multi-scale attention	49.6
Our method	50.1

Table 4. Ablation study on HICO-DET datasets

Methods	Default object			Known object		
	Full	Rare	Non-rare	Full	Rare	Non-rare
Baseline	17.22	13.51	18.32	19.38	15.38	20.57
Attention	19.78	14.21	20.82	21.54	15.47	23.20
Our method	19.90	14.67	21.35	22.30	15.80	24.50

5 Conclusions

In this paper, we propose a multi-scale attention fusion method for human-object interaction detection. Our main idea is that for people and objects of different sizes in the image, we use different convolutions to adapt to different people and objects, which can increase the range of attention, and more accurately determine the interaction between people and objects. Meanwhile, the interaction between people and close objects has more possibilities. To better characterize the interaction between people and close objects, we propose a weighting mechanism. It can better express people's intention of interaction

and improve the accuracy of interaction relationship detection. We verify the effectiveness of the method on two datasets and achieved excellent performance. In the future, we will study the interaction detection of the relationship between people and objects in the video.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61672268).

References

1. Tian, Y., Ruan, Q., An, G., et al.: Action recognition using local consistent group sparse coding with spatio-temporal structure. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 317–321 (2016)
2. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
3. Xu, B., Ye, H., Zheng, Y., et al.: Dense dilated network for few shot action recognition. In: Proceedings of the ACM on International Conference on Multimedia Retrieval, pp. 379–387 (2018)
4. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437 (2018)
5. Li, Y.L., Zhou, S., Huang, X., et al.: Transferable interactiveness knowledge for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3585–3594 (2019)
6. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
7. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
8. Li, Y., Qi, H., Dai, J., et al.: Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2367 (2017)
9. Law, H., Deng, J.: Cornernet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
10. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
11. He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
12. Chen, K., Pang, J., Wang, J., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974–4983 (2019)
13. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1483–1498 (2019)
14. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
15. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

16. Chéron, G., Laptev, I., Schmid, C.: P-CNN: pose-based cnn features for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3218–3226 (2015)
17. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
18. Mallya, A., Lazebnik, S.: Learning models for actions and person-object interactions with transfer to question answering. In: European Conference on Computer Vision, pp. 414–428. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_25
19. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. arXiv preprint arXiv: 1711.01467 (2017)
20. Jetley, S., Lord, N.A., Lee, N., et al.: Learn to pay attention. arXiv preprint arXiv:1804.02391 (2018)
21. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
22. Gkioxari, G., Girshick, R., Dollár, P., et al.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8359–8367 (2018)
23. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.-C.: Learning human-object interactions by graph parsing neural networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 407–423. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_25
24. Chao, Y.W., Liu, Y., Liu, X., et al.: Learning to detect human-object interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 381–389. IEEE (2018)
25. Wang, T., Anwer, R.M., Khan, M.H., et al.: Deep contextual attention for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5694–5702 (2019)
26. Liang, Z., Liu, J., Guan, Y., et al.: Visual-semantic graph attention networks for human-object interaction detection. arXiv e-prints: arXiv: 2001.02302 (2020)
27. Xu, B., Li, J., Wong, Y., et al.: Interact as you intend: intention-driven human-object interaction detection. IEEE Trans. Multimedia **22**(6), 1423–1432 (2019)
28. Fang, H.S., Xie, S., Tai, Y.W., et al.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
29. Li, J., Wang, C., Zhu, H., et al.: Crowdpose: efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10863–10872 (2019)
30. Xu, K., Ba, J., Kiros, R., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning. PMLR, pp. 2048–2057 (2015)
31. You, Q., Jin, H., Wang, Z., et al.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016)
32. Ji, Z., Fu, Y., Guo, J., et al.: Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Advances in Neural Information Processing Systems, pp. 5995–6004 (2018)
33. Chu, X., Yang, W., Ouyang, W., et al.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1831–1840 (2017)
34. Chao, Y.W., Wang, Z., He, Y., et al.: Hico: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1017–1025 (2015)

35. Kuang, H., Zheng, Z., Liu, X., et al.: A human-object interaction detection method inspired by human body part information. In: 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 342–346. IEEE (2020)
36. Zhou, T., Wang, W., Qi, S., et al.: Cascaded human-object interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4263–4272 (2020)
37. Wan, B., Zhou, D., Liu, Y., et al.: Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9469–9478 (2019)
38. Wang, T., Yang, T., Danelljan, M., et al.: Learning human-object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4116–4125 (2020)
39. Zhao, W., Lu, H., Wang, D.: Multisensor image fusion and enhancement in spectral total variation domain. IEEE Trans. Multimedia **20**(4), 866–879 (2017)
40. Lan, R., Sun, L., Liu, Z., et al.: MADNet: a fast and lightweight network for single-image super resolution. IEEE Trans. Cybern. **51**, 1443–1453 (2020)