HUMAN-OBJECT RELATION NETWORK FOR ACTION RECOGNITION IN STILL IMAGES

*Wentao Ma and Shuang Liang**

School of Software Engineering, Tongji University, China
{wentao.ma, shuangliang}@tongji.edu.cn.

ABSTRACT

Surrounding object information has been widely used for action recognition. However, the relation between human and object, as an important cue, is usually ignored in the still image action recognition field. In this paper, we propose a novel approach for action recognition. The key to ours is a human-object relation module. By using the appearance as well as the spatial location of human and object, the module can compute the pair-wise relation information between human and object to enhance features for action classification and can be trained jointly with our action recognition network. Experimental results on two popular datasets demonstrate the effectiveness of the proposed approach. Moreover, our method yields the new state-of-the-art results of 92.8% and 94.6% mAP on the PASCAL VOC 2012 Action and Stanford 40 Actions datasets respectively. Ablation study and visualization confirm the proposed method can model and utilize the human-object relation for action recognition.

***Index Terms*—** Action Recognition, Attention Mechanism, Convolutional Neural Network

1. INTRODUCTION

Action recognition is one of the core topics in computer vision with a wide range of applications, such as robotics, health, and security. Compared to video action recognition, identify actions from a single image is a more challenging problem due to the lack of temporal information. Hence, researchers [1–6] have combined different cues with the human body features in still images to characterize actions better. Surrounding object information is one of the vital cues and has been widely used in many methods [1–5].

However, if we do not consider the relation between human and object, employing object features can also mislead action recognition. As Fig.1 shows, the bikes around people may cause the person who is "standing" to be misclassified as "riding bike" in Fig.1(a), and the skateboard and the nearby person may cause the person who is "sitting" on the ground to be misclassified as "riding skateboard" in Fig.1(b).

Although some early action recognition works [7, 8] model human-object relations with traditional methods, to the best of our knowledge, there is no prior work utilizes such relations to improve the performance of deep learning-based methods for action recognition in still images. One possible reason is that, as backbone networks for recent action recognition methods, modern convolutional neural networks (CNNs) [9–11] need regular input, which are hard to handle surrounding objects with arbitrary numbers, scales, and locations for different images.

(a) (b)

Fig. 1. Examples of surrounding objects may mislead action recognition without the consideration of the human-object relation. The people in question are shown with blue boxes, while the surrounding objects with yellow boxes.

Recently, the self-attention module [12] gives promising results in the natural language processing field and has been adapted to improve the performance of object detection systems [13]. The module can mine the relevant features from a set of elements (corresponding to words or region proposals for natural language processing or object detection respectively) for each element by applying learnable weights.

In this paper, motivated by our observation and the recent success of the self-attention module, we propose an action recognition model with our human-object relation module for recognizing actions in still images. By using the appearance as well as the spatial location of human and object, the module can compute pair-wise relation information between human and object to enhance features for action classification. Besides, it can be trained jointly with our action recognition framework. In Fig.2, we present an overview of our proposed method. An apparent distinction between ours and previous self-attention based modules [12, 13] is that the relation between human and object, or the learned weight, is applied to two rather than one set of elements, human features and ob-

*Corresponding author.

978-1-7281-1331-9/20/$31.00

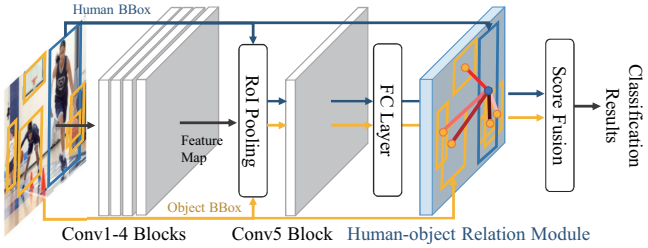


Fig. 2. Overview of our proposed method for action recognition in still images.

ject features. The reason is that not only human features but also object features enhanced by human-object relations can help recognize actions, especially for some actions such as "riding a bicycle" and "riding a motorcycle" which cannot be distinguished without object information.

We evaluate our approach on two public datasets, the PASCAL VOC 2012 Action [14] and the Stanford 40 Actions [15]. Experimental results show that our approach achieves state-of-the-art results on both datasets with 92.8% and 94.6% mAP, respectively. Ablation study and visualization also prove the effectiveness of our proposed human-object relation module.

2. RELATED WORK

2.1. Action recognition in still images

As an actively studied topic, there is a variety of work for recognizing actions in still images. Early works [7, 8] normally rely on a structured graphical model with hand-crafted features to infer actions. Recent state-of-the-art methods [1–5] use CNNs to automatically extract both human features and surrounding object information for action classification. Gkioxari et al. [1] apply selective search [16] to generate object proposals, predict the classification score for each proposal, and fuse the object's scores with human's directly. Zhang et al. [4] propose a method that combines both human features and object features around human key points to classify actions. Fang et al. [5] utilize human body-part features and object features to characterize actions better. However, such state-of-the-art methods directly use object features for score fusion and ignore the important cue of the relation between human and object. Our approach enhances action-related object features and reduces the noise of irrelevant information by using our human-object relation module.

2.2. Visual relationship

Visual relationship, which aims to model the relation between objects or entities, has been previously studied in the computer vision community [17, 18]. Recently, detecting and recognizing human-object interactions (HOI) [19, 20] has attracted much attention. Gkioxari et al. [19] propose a method to estimate interactive object locations based on human appearance. Chao et al. [20] introduce a human-object region-based CNN with three streams, human, object, and pairwise for HOI detection. However, methods for detecting human-object interactions usually requires additional exhaustive annotations of each interaction between human and object, which are only provided in a few specific datasets [20, 21]. Our proposed human-object relation module does not require any annotations of human-object interactions and can be easily trained on current action recognition datasets.

3. PROPOSED METHOD

In this section, we present our human-object relation network for action recognition in still images.

3.1. Network overview

The overall architecture of the proposed method is illustrated in Fig.2. We integrate the human-object relation module with a popular feature extraction network, ResNet [10], in this paper for illustration and experiment, but we do not see any reason preventing it be used in other backbone networks [9, 11]. Since we also want to consider spatial relations when modeling human-object relations, we need the human and object bounding boxes information as additional input. Human bounding boxes are provided in the datasets [14, 15], object bounding boxes are detected by Faster R-CNN [22].

We first extract the image-level feature map using the first four convolution blocks. Next, we apply RoI pooling [23] to obtain the instance-level features according to human and object bounding boxes. From then on, our model becomes a two-stream network. By passing the instance-level features through a weight-shared convolution block and a fully connected layer, we get appearance features for human and object instances, f_h and f_o. Our human-object relation module uses f_h, f_o and bounding boxes to compute relations (will be detailed in Section 3.2). At last, we fuse the classification scores of features enhanced by human-object relations and output the final result (will be detailed in Section 3.3).

3.2. Human-object relation module

Our goal is to use the relations to enhance human features to characterize actions better. We also hope to strengthen object features so that the network can distinguish similar actions by different related objects (e.g. "brushing teeth" and "blowing bubbles"). Thus, as Fig.3 shows, our human-object relation module outputs two relation-enhanced features, human-object feature f_{ho} and object-human feature f_{oh}, which is the most apparent distinction between ours and previous related modules [12, 13]. Besides, since the relation between two individuals is a naturally mutual and equivalent connection, we share the learned relation weight w_{ho} and apply it to both human and object features (with using transpose on one side

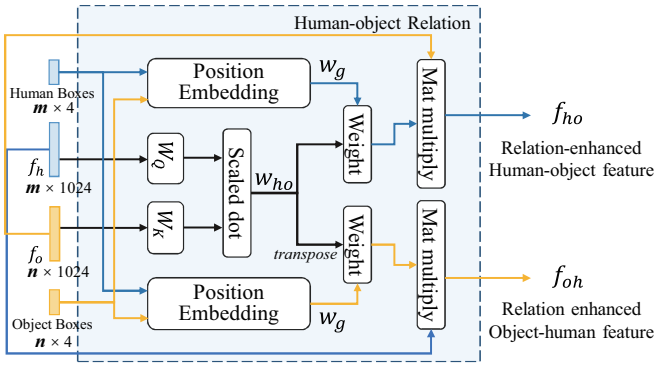


Fig. 3. Detailed computation process of the proposed human-object relation module.

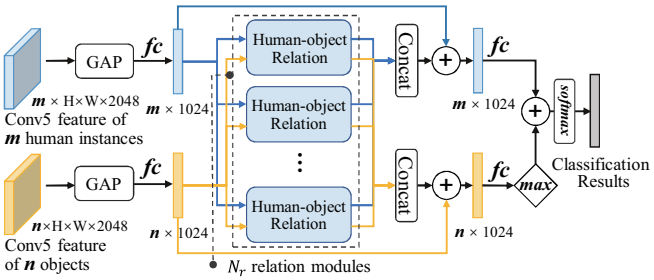


Fig. 4. Detailed architecture of our proposed method after the Conv5 block. GAP and fc represent a global pooling layer and a fully connected layer respectively.

to meet the dimensional requirement). The effectiveness of these changes is demonstrated in our ablation study in Section 4.3.

We now describe the computation process of our human-object relation. Given the input consist of human features f_h and object features f_o, we first adopt the scaled dot formula [12] to compute the relation weight w_{ho},

$$w_{ho} = softmax(\frac{W_Q f_h \cdot W_K f_o}{\sqrt{d_k}}), \tag{1}$$

where d_k represents the number of dimensions of $W_K f_o$ that is used as a scaling factor to have more stable gradients in the training phase.

Because the relation between human and object is not only related to appearance features but also to spatial locations, we also compute the geometry weight w_g corresponding to position embedding in Fig.3 as follow:

$$w_g = fc(\varepsilon_g(b_h, b_o)), \tag{2}$$

where fc indicates a single fully connected layer, b_h and b_o represent human and object bounding boxes. The operation ε_g is an embedding function used in [2], which computes cosine and sine function of different wavelengths. To make the geometry weight invariant to translation and scaling of bounding boxes, we compute the relative position of b_h and b_o as [13]: $(log(\frac{|x_h - x_o|}{w_h})), log(\frac{|y_h - y_o|}{h_h}), log(\frac{w_o}{w_h}), log(\frac{h_o}{h_h})$, and use it rather than original bounding boxes coordinates as the input in Eq.2.

Given the relation weight w_{ho}, the geometry weight w_g and object features f_o, the relation-enhanced human-object feature f_{ho} is computed as:

$$f_{ho} = fc(\sum_{i=1}^{n}((w_g w_{ho}) \cdot f_o)). \tag{3}$$

In our implementation, instead of relying a singe human-object relation module to generate d_k-dimensional relation-enhanced features, we follow the multi-head attention [12] form, use N_r modules in parallel as illustrated in Fig.4. Multiple relation modules enable the network to jointly attend to information from different representation sub-spaces. Specifically, $d_k * N_r$ equals to the dimension of input human and object features, f_h and f_o (in Fig.4, the dimension is 1024). By concatenating N_r features $f_{ho}^{1-N_r}$, we have the final relation-enhanced human-object feature f_{ho}^*. The computation of a single relation-enhanced object-human feature f_{oh} and the concatenated feature f_{oh}^* are similar to f_{ho} and f_{ho}^* respectively, except that the relation weight w_{ho} is transposed and the input is replaced by human features f_h.

Throughout the computation process of our human-object relation module, we do not need to know the specific values of the number of human instances m and object instances n, which is suitable for action recognition in still images where both m and n in the input is uncertain.

3.3. Score fusion and loss function

As Fig.4 shows, after employing multiple human-object relation modules to obtain concatenated relation-enhanced human-object and object-human features, f_{ho}^* and f_{oh}^* , we add f_{ho}^* and f_{oh}^* back to f_h and f_o. The obtained features are directly passed through a fully connected layer as a classifier to compute the scores S_h^a and S_o^a for each action category a. Since we only want to fuse the most action-related object feature, a max operation is adopted to fuse human and object scores as follow:

$$Score(a; h) = S_h^a + \max\{S_{o(1)}^a, ..., S_{o(n)}^a\}, \text{for } n \text{ objects}. \tag{4}$$

At last, we apply the $softmax$ operation to convert the fused score to the estimated probability for predicting actions.

During the training phase, our entire network only needs action categories for human instances as the supervision information. We use the cross entropy loss as our network loss function, which is defined as:

$$L = -log(\frac{exp(Score(a^{gt}; h))}{\sum_a exp(Score(a; h))}), \tag{5}$$

Table 1. AP (%) on the PASCAL VOC 2012 Action validation set.

Method	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	mAP
R*CNN [1]	88.9	79.9	95.1	82.2	96.1	97.8	87.9	85.3	94.0	71.5	87.9
Attention [2]	87.8	78.4	93.7	81.1	95.0	97.1	96.0	85.5	93.1	73.4	87.1
Part Action [4]	**89.6**	86.9	94.4	**88.5**	94.9	97.9	91.3	87.5	92.4	76.4	90.0
Ours	89.2	**89.8**	**96.5**	87.6	**98.2**	**99.1**	**92.3**	**91.6**	**95.2**	**79.2**	**91.9**

Table 2. AP (%) on the PASCAL VOC 2012 Action test set.[1]

Method	jumping	phoning	playing instrument	reading	riding bike	riding horse	running	taking photo	using computer	walking	mAP
VGG 16&19 [9]	89.3	71.3	94.7	71.3	97.1	98.2	90.2	73.3	88.5	66.4	84.0
R*CNN [1]	**91.5**	84.4	93.6	83.2	96.9	98.4	93.8	85.9	92.6	81.8	90.2
Attention [2]	92.7	86.0	93.2	83.7	96.6	**98.8**	93.5	85.3	91.8	80.1	90.2
Ours[2]	91.1	**89.8**	**95.4**	**87.7**	**98.6**	**98.8**	**95.4**	**91.4**	**95.8**	**84.3**	**92.8**

[1] The offical leaderboard: http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_noeq.php?challengeid=11&compid=10
[2] Our anonymous result link: http://host.robots.ox.ac.uk:8080/anonymous/GWUICU.html

where $Score(a; h)$ is the fused score of the action category a for the human instance h in Eq.4, and a^{gt} represents the ground-truth action category.

4. EXPERIMENTS

In this section, we evaluate our proposed approach on two action recognition benchmarks and perform ablation analysis and visualization of our method.

4.1. Implementation details

We use ResNet-50 [10] as the backbone network of our proposed method and implement on MXNet [24] based on GluonCV [25]. Same as most action recognition methods [1–6], the ImageNet pre-trained weight is used to initialize parameters. We use SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is set to 3×10^{-5} and drops to 1×10^{-6} in 10 epochs by using cosine annealing. The hyperparameter N_r, the number of human-object relation module, is set to 16, so d_k is equal to 64. The human bounding boxes are provided by datasets [14, 15], and the object bounding boxes are detected by Faster R-CNN [22]. We conduct all the experiments on an NVIDIA GTX1080 GPU with the same settings on two datasets if not explicitly stated.

4.2. Comparison with existing methods

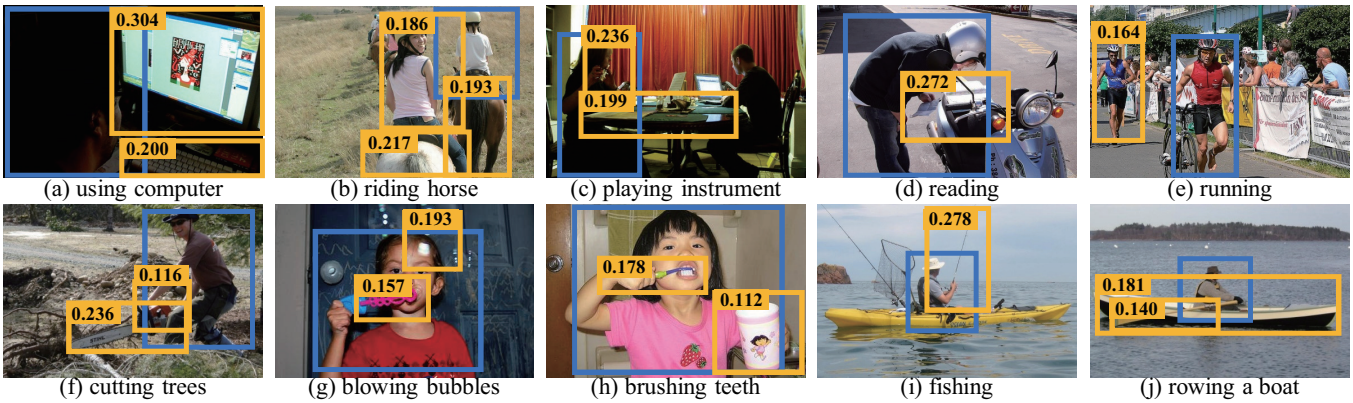
To evaluate the proposed method, we conduct experiments on two popular image datasets for action recognition, the PASCAL VOC 2012 Action dataset [14] and the Stanford 40 Actions dataset [15]. The PASCAL VOC dataset consists of 10 different categories and 9157 images, each category contains 300-500 images for training and validation, and the rest is used for the test. The Stanford 40 dataset contains 9532 images in 40 action classes. For each class, 100 images are used for training, and the others are used for the test.

Table 3. Mean AP on the Stanford 40 Actions test set.

Method	Feature Backbone	Mean AP (%)
Attention [2]	VGG 16	90.7
R*CNN [1]	VGG 16	90.9
Part Action [4]	ResNet-50	91.2
Human Mask [6]	ResNet-50	91.1
Human Mask [6]	Inception-ResNet-v2	94.1
Ours	ResNet-50	93.1
Ours	ResNet-101	**94.6**

PASCAL VOC 2012 dataset. Table 1 shows the comparison with the state-of-the-art methods. The results on the test set are shown in Table 2, which are obtained by submitted to the official PASCAL VOC evaluation server. In this dataset, our proposed method outperforms others 1.9% and 2.6%, respectively. In the test set, our human-object relation network reaches the best results for 9 of 10 categories, especially in categories that people may have various poses (e.g. standing, squatting or lying) and objects and human-object relations become the main factors for identifying actions, such as "phoning" (+2.8%), "reading" (+4.0%), "taking photo" (+5.5%) and "using computer" (+3.2%).

Stanford 40 dataset. Table 3 shows the results on the Stanford 40 dataset. Using ResNet-50 [10] as the feature backbone, our method achieves a promising result 93.1% on the Stanford 40 dataset. It is worth noting that the previous method [6] uses a much larger backbone network, Inception-ResNet-v2 [11]. However, due to the lack of the



(a) using computer (b) riding horse (c) playing instrument (d) reading (e) running

(f) cutting trees (g) blowing bubbles (h) brushing teeth (i) fishing (j) rowing a boat

Fig. 5. Visualization of our network predictions on the PASCAL VOC 2012 validation set (the first row) and the Stanford 40 test set (the second row). The human instances to be classified is blue. The objects with a high relation weight (shown on the top-left) is yellow. The predicted action label is listed below each sample.

Table 4. Ablation study on the VOC 2012 validation set.

No.	HBox	OBox	f_{ho}	f_{oh}	Mutual Relation	Score Fusion	mAP
a)	✓	-	-	-	-	-	88.5
b)	✓	✓	-	-	-	✓	90.2
c)	✓	✓	✓	-	-	-	90.2
d)	✓	✓	✓	-	-	✓	91.1
e)	✓	✓	-	✓	-	✓	90.7
f)	✓	✓	✓	✓	-	✓	91.0
g)	✓	✓	✓	✓	✓	✓	**91.9**

pre-trained weight, we can only evaluate our method with ResNet-101 network. Although the top-1 error rate on ImgaeNet of ResNet-101, 19.9% [10] is lower than 18.7% [11] of Inception-ResNet-v2, our proposed method still refreshes the state-of-the-art result with 94.6% mAP.

4.3. Ablation Study and visualization

To demonstrate the effectiveness of each component in our method, we use the same ResNet-50 as our backbone network with different settings, train them on the PASCAL VOC 2012 train set and evaluate on the validation set. The results are shown in Table 4.

Our baseline is Table 4(b), which also takes the detected object bounding boxes as one of the input and fuses object features. With object score fusion, the baseline model has been dramatically improved compared to Table 4(a) and already performs better than the previous best method in Table 1, which is a strong baseline. Also, if we take the human-object relation into account without score fusion, the model can achieve similar results with baseline as Table 4(c) shows. When we add the relation-enhanced either human-object or object-human features into the baseline model, the performance of the model can be further improved, as shown in Table 4(d)(e). Further, if we use two independent modules to calculate f_{ho} and f_{oh} respectively, and ignore the fact that the relation is naturally mutual like Table 4(f), we can not achieve better results. Finally, our proposed method, Table 4(g), reports the best result among these networks.

In Fig.5, we present predictions of our network on two datasets. For better visualization, we only mark the objects with a high relation weight ($>$0.1) and omit the bounding boxes that overlap with the people to be classified (IoU $>$0.5). The presented samples show that: by employing the relation between human and object, our method can (1) effectively extract action-related information from surrounding objects (e.g. the monitor and the keyboard in Fig.5(a)), (2) discriminate actions with similar poses by focusing on key objects (e.g. the bubble in Fig.5(g) and the toothbrush in Fig.5(h)), (3) and reduce the weights of misleading objects (e.g. the bike around the person in Fig.5(d) and (e)).

5. CONCLUSION

In this paper, we proposed a novel action recognition approach with the human-object relation module. The module can help identify human actions in still images by computing the relation between human and object, the automatically learned weights, according to both human and object appearances as well as their spatial locations. Extensive experiments presented the improvements brought by the relation module and demonstrated the effectiveness of the proposed method, which achieved the state-of-the-art results on two public datasets.

Acknowledgments

This work is supported by Shanghai Natural Science Foundation (No. 19ZR1461200) and National Natural Science Foundation of China (No. 61976159). The authors would also

like to thank the anonymous reviewers for their valuable comments and suggestions.

6. REFERENCES

[1] Georgia Gkioxari, Ross Girshick, and Jitendra Malik, "Contextual action recognition with r*cnn," in *IEEE International Conference on Computer Vision*, 2015, pp. 1080–1088.

[2] Shiyang Yan, Jeremy S Smith, Wenjin Lu, and Bailing Zhang, "Multibranch attention networks for action recognition in still images," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 1116–1125, 2018.

[3] Zhichen Zhao, Huimin Ma, and Xiaozhi Chen, "Semantic parts based top-down pyramid for action recognition," *Pattern Recognition Letters*, vol. 84, pp. 134–141, 2016.

[4] Zhichen Zhao, Huimin Ma, and Shaodi You, "Single image action recognition using semantic body part actions," in *IEEE International Conference on Computer Vision*, 2017, pp. 3391–3399.

[5] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu, "Pairwise body-part attention for recognizing human-object interactions," in *European Conference on Computer Vision*, 2018, pp. 51–67.

[6] Lu Liu, Robby T Tan, and Shaodi You, "Loss guided activation for action recognition in still images," in *Asian Conference on Computer Vision*, 2018, pp. 152–167.

[7] Bangpeng Yao and Li Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 9–16.

[8] Bangpeng Yao and Li Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 17–24.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, "Relation networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[15] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *IEEE International Conference on Computer Vision*, 2011, pp. 1331–1338.

[16] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[17] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*, 2016, pp. 852–869.

[18] Bo Dai, Yuqi Zhang, and Dahua Lin, "Detecting visual relationships with deep relational networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3076–3086.

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He, "Detecting and recognizing human-object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[20] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng, "Learning to detect human-object interactions," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 381–389.

[21] Saurabh Gupta and Jitendra Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[23] Ross Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[24] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[25] Junyuan Xie, Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, and Mu Li, "Bag of tricks for image classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.