Pose-aware Multi-level Feature Network for Human Object Interaction Detection: Supplementary Material

Bo Wan* Desen Zhou* Yongfei Liu Rongjie Li Xuming He
ShanghaiTech University, Shanghai, China
{wanbo, zhouds, liuyf3, lirj2, hexm}@shanghaitech.edu.cn

In this supplementary material, we first describe the implementation details, and then provide more visualization results for V-COCO and HICO-DET datasets. We organize the qualitative results into two parts: 1) Same person interacting with different objects and 2) Multiple interaction categories with different human-object pairs. All examples are highlighted by semantic part attention heatmap.

1. Implementation Details

We use Faster R-CNN [6] as object detector and CPN [1] as pose estimator, which are pre-trained on the COCO train2017 split. Each human pose has a total of $K = 17$ keypoints as in COCO dataset. Note that for V-COCO dataset, part of the test images come from COCO train2017 split, we thus train the object detector and human pose estimator on train2017 split while removing those images.

Our backbone module uses ResNet-50-FPN [4] as feature extractor, and we crop RoI features from the highest resolution feature map in FPN [4]. The size of our spatial configuration map M is set to 64. The RoI-Align in holistic module has a resolution $R_h = 7$, while in zoom-in module, the size of human parts is $\gamma = 0.1$ of human box height and all the features are rescaled to $R_p = 5$.

Our holistic module uses multiple two fully-connected layers to embed human, object, union, spatial features into a feature space with 256 dimension separately. Features of multiple branches are then concatenated into a 1024-dimension feature vector.

Our zoom-in module uses a two-layer fully-connected network with hidden dimension d=64 to predict a 17-dimension semantic part attention from spatial configuration map. A sigmoid function is applied to each dimension of semantic attention to normalize its range to $(0, 1)$. The attention enhanced part-level feature f_{att} is fed to two fully-connected layers to get a 256-dimension feature Γ_{loc}.

We freeze ResNet-50 backbone and train the parameters of FPN component. We use SGD optimizer for training with initial learning rate 4e-2, weight decay 1e-4, and momentum 0.9. The ratio of positive and negative samples is 1:3. Batch size is set to 4. For V-COCO [3], we reduce the learning rate to 4e-3 at iteration 24,000, and stop training at iteration 48,000. For HICO-DET [5], we reduce the learning rate to 4e-3 at iteration 250,000 and stop training at iteration 300,000. During testing, we use object proposals from [2] for fair comparison. Following [2], we discard the human / object proposals whose detection scores are lower than 0.5 / 0.4.

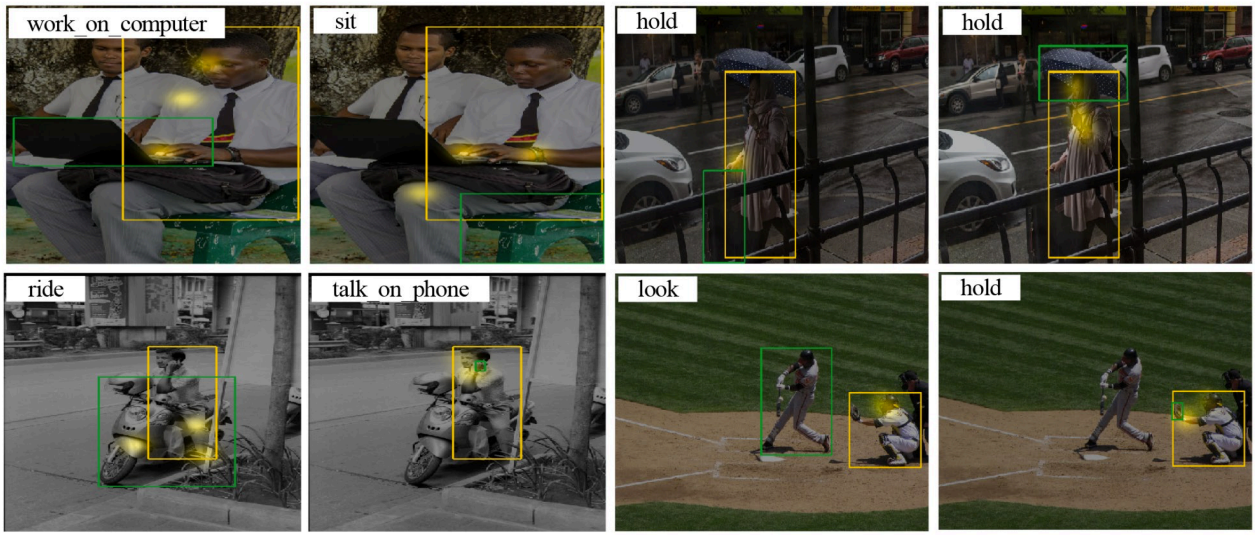
Our implementation code is available at https://github.com/bobwan1995/PMFNet.

*Authors contributed equally and are listed in alphabetical order.

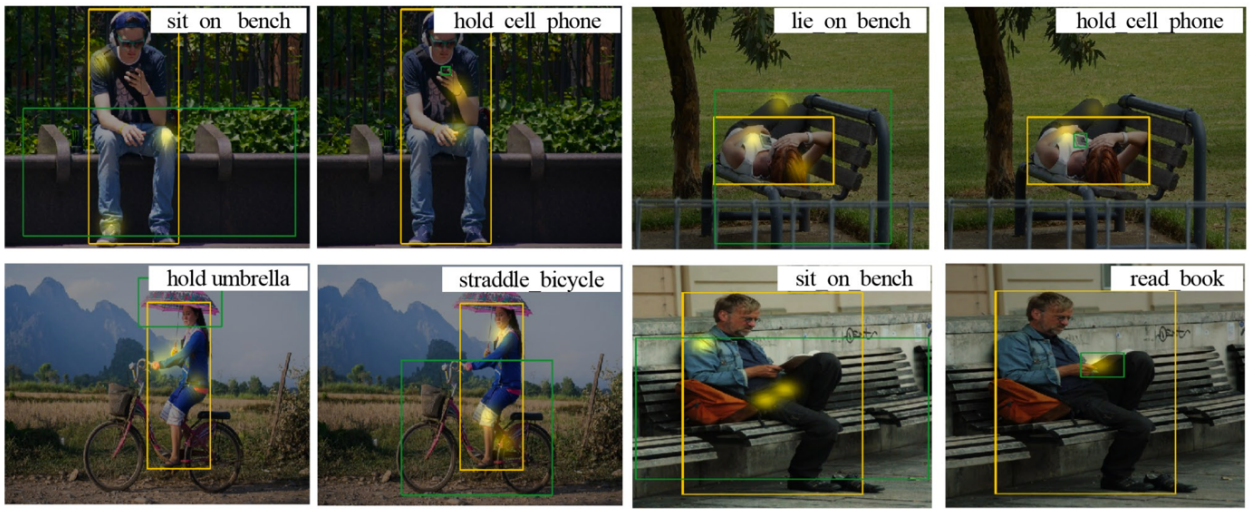
2. Visualization

2.1. Interaction with different objects.

For each dataset, we show four groups of output examples, each of which includes a pair of images showing a person interacting with different target objects. Our semantic-part attention module can automatically focus on different human parts that are strongly related to interaction type.



(a) Qualitative results on V-COCO dataset.

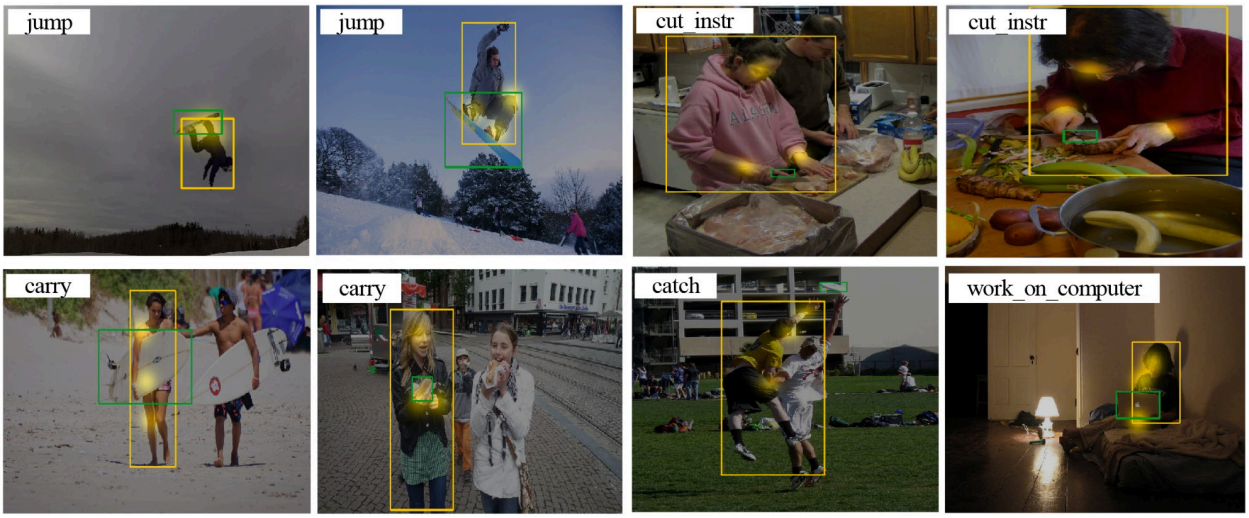


(b) Qualitative results on HICO-DET dataset.

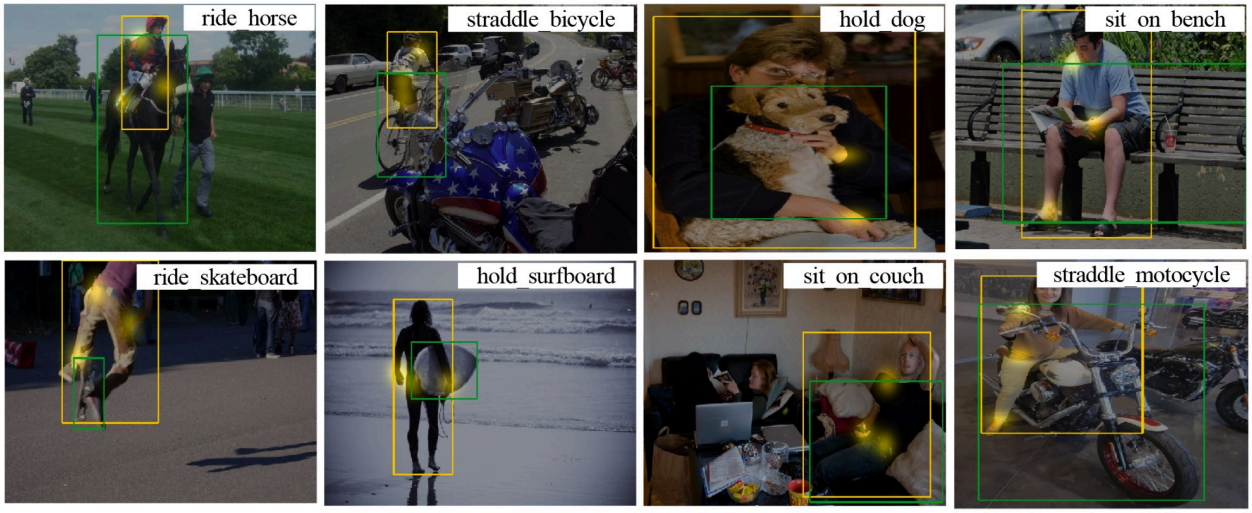
Figure 1: Interaction with different objects.

2.2. Interaction on different human-object pairs.

From each dataset, we show more examples of our predictions on multiple interaction categories. These qualitative results illustrate the variance of the part attention for different instances of human object interactions (HOIs).



(a) Qualitative results on V-COCO dataset.



(b) Qualitative results on HICO-DET dataset.

Figure 2: Interaction on different human-object pairs.

References

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[2] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference (BMVC)*, 2018.

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

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[5] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[6] 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 (NeuIPS)*, pages 91–99, 2015.