GraphFPN: Graph Feature Pyramid Network for Object Detection (*Supplementary Material*)

Gangming Zhao [†1,2,3], Weifeng Ge [*1,2], and Yizhou Yu [*3]

[1]Nebula AI Group, School of Computer Science, Fudan University
[2]Shanghai Key Lab of Intelligent Information Processing
[3]Department of Computer Science, The University of Hong Kong

1. Network Architecture and Visual Results

We provide the network architecture of the backbone (ResNet-101 [2]) and FPN used in the proposed pipeline in Table 1. The network architecture of our GraphFPN is given in Table 2. For our GraphFPN, the feature dimension F of every graph node is always set to 256 in all experiments reported in this paper. In GraphFPN, the first group of three layers are contextual layers, the second group of three layers are hierarchical layers, and the last group of three layers are contextual layers again. As mentioned in the paper, the graphs in all these layers have identical sets of nodes (distributed in five levels), but contextual and hierarchical layers have different sets of graph edges. Each of these layers has three attention modules, a spatial self-attention module, a local channel-wise attention module and a local channel self-attention module. Note that the number of graph nodes in each layer of GraphFPN is $(N + \frac{N}{4} + \frac{N}{16} + \frac{N}{64} + \frac{N}{256})$, where N is the number of superpixels in the finest level of a superpixel hierarchy.

Figures 2, 3, and 4 show sample superpixel hierarchies based on hierarchical image segmentation algorithm COB [5]. Starting from the finest partition $\mathcal{S}^{l_1}$, superpixels are recursively merged according to contour strengths to generate a set of partitions and form a superpixel hierarchy $\left\{\mathcal{S}^{l_1}, \mathcal{S}^{l_2}, \mathcal{S}^{l_3}, \mathcal{S}^{l_4}, \mathcal{S}^{l_5}\right\}$. Input images are taken from the MS COCO 2017 dataset [4].

Figure 1 shows sample detection results from FPN [3], FPT [7], and our GraphFPN based method. Input images are taken from the MS COCO 2017 validation set [4]. Figures 5 and 6 show additional sample detection results from our GraphFPN based method. Images are taken from the MS COCO 2017 validation set [4].

2. Experiments on Semantic Segmentation

To demonstrate the effectiveness of our method in capturing intrinsic image structures, we further apply our method to semantic segmentation. In our experiments in semantic segmentation, we test the performance of UFP + GraphFPN and compare its results with unscathed feature pyramid networks(UFP [6]) and feature pyramid transformer. Table 3 shows experimental results on the Cityscapes [1] dataset, which contains 19 classes and includes 2,975,500 images for training and validation. The settings of this experiment are the same as in [7]. We also adopt Unscathed Feature Pyramid (UFP) [6] as the feature pyramid construction module. From the experimental results shown in Table 3, it can be found out that our proposed method achieves clearly better performance, which also demonstrates the applicability of our method.

† This work is done when Gangming Zhao is a visiting student at Fudan University. *Corresponding authors: wfge@fudan.edu.cn and yizhouy@acm.org

Stage	Layer Name	Output Size	Kernels, #channels
C_1	conv1	$W \times H$	7×7, 64,stride 2
C_2	conv2_x	$\frac{W}{2} \times \frac{H}{2}$	3×3 max pool, stride 2 $\begin{bmatrix} 1\times 1, 64 \\ 3\times 3, 64 \\ 1\times 1, 256 \end{bmatrix} \times 3$
C_3	conv3_x	$\frac{W}{4} \times \frac{H}{4}$	$\begin{bmatrix} 1\times 1, 128 \\ 3\times 3, 128 \\ 1\times 1, 512 \end{bmatrix} \times 23$
C_4	conv4_x	$\frac{W}{8} \times \frac{H}{8}$	$\begin{bmatrix} 1\times 1, 256 \\ 3\times 3, 256 \\ 1\times 1, 1024 \end{bmatrix} \times 4$
C_5	conv5_x	$\frac{W}{16} \times \frac{H}{16}$	$\begin{bmatrix} 1\times 1, 512 \\ 3\times 3, 512 \\ 1\times 1, 2048 \end{bmatrix} \times 3$
P_1	-	$W \times H$	3×3, 256
P_2	-	$\frac{W}{2} \times \frac{H}{2}$	3×3, 256
P_3	-	$\frac{W}{4} \times \frac{H}{4}$	3×3, 256
P_4	-	$\frac{W}{8} \times \frac{H}{8}$	3×3, 256
P_5	-	$\frac{W}{16} \times \frac{H}{16}$	3×3, 256

Table 1. Network architecture of the backbone (ResNet-101) and convolutional FPN used in the proposed pipeline. Residual building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of 2. W and H are the input width and height.

Stage	Layer Name	#Node	#Feature Channel
CGL-1	CL-1 CL-2 CL-3	$N + \frac{N}{4} + \frac{N}{16} + \frac{N}{64} + \frac{N}{256}$	256
HGL	HL-1 HL-2 HL-3	$N + \frac{N}{4} + \frac{N}{16} + \frac{N}{64} + \frac{N}{256}$	256
CGL-2	CL-4 CL-5 CL-6	$N + \frac{N}{4} + \frac{N}{16} + \frac{N}{64} + \frac{N}{256}$	256

Table 2. Network architecture of our GraphFPN. "CGL-1" stands for the first group of contextual layers, "HGL" stands for the group of hierarchical layers, and "CGL-2" stands for the second group of contextual layers. Each "CL" or "HL" layer has three attention modules. Note that the number of graph nodes in each layer is $N + \frac{N}{4} + \frac{N}{16} + \frac{N}{64} + \frac{N}{256}$, where N is the number of superpixels in the finest level of a superpixel hierarchy.

Methods	Train.mIoU	Val.mIoU	Params	GFLOPs
UFP [6]	86.0	79.1	71.3 M	916.1
UFP+FPS [7]	87.4	81.7	127.2 M	1063.9
UFP+GraphFPN	**88.4** (↑1.0)	**83.2**(↑1.5)	130.1 M	1104.2

Table 3. Comparison with state-of-the-art semantic segmentation methods on the Cityscapes validation set [1].

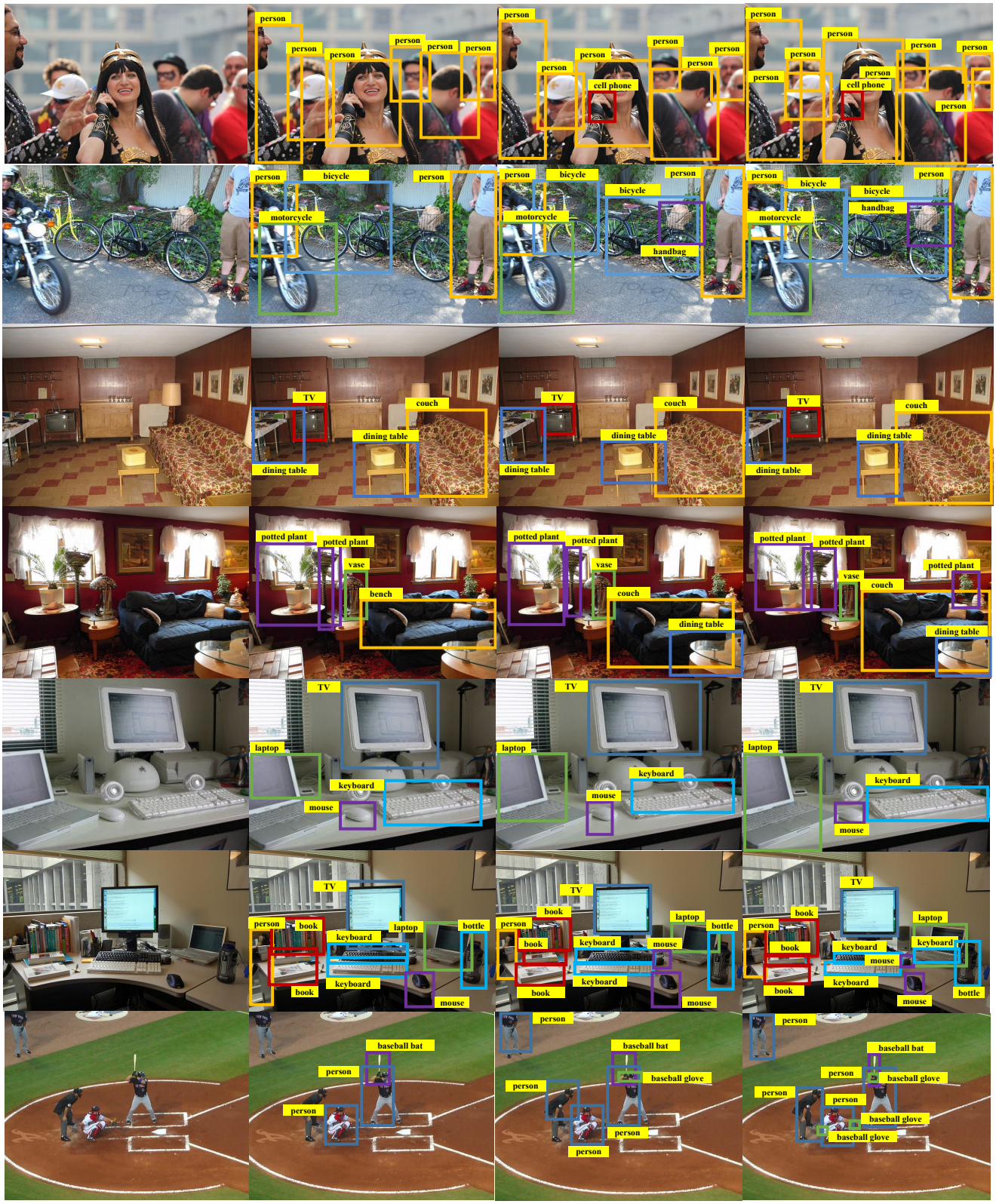


Figure 1. Sample detection results from FPN [3], FPT [7], and our GraphFPN based method. Images are from the MS COCO 2017 validation set [4].

Figure 2. Sample result of superpixel hierarchy. Each superpixel hierarchy consists of 5 levels, $\{\mathcal{S}^{l_1}, \mathcal{S}^{l_2}, \mathcal{S}^{l_3}, \mathcal{S}^{l_4}, \mathcal{S}^{l_5}\}$. Images are from the MS COCO 2017 dataset [4].

Figure 3. Sample result of superpixel hierarchy. Each superpixel hierarchy consists of 5 levels, $\{\mathcal{S}^{l_1}, \mathcal{S}^{l_2}, \mathcal{S}^{l_3}, \mathcal{S}^{l_4}, \mathcal{S}^{l_5}\}$. Images are from the MS COCO 2017 dataset [4].

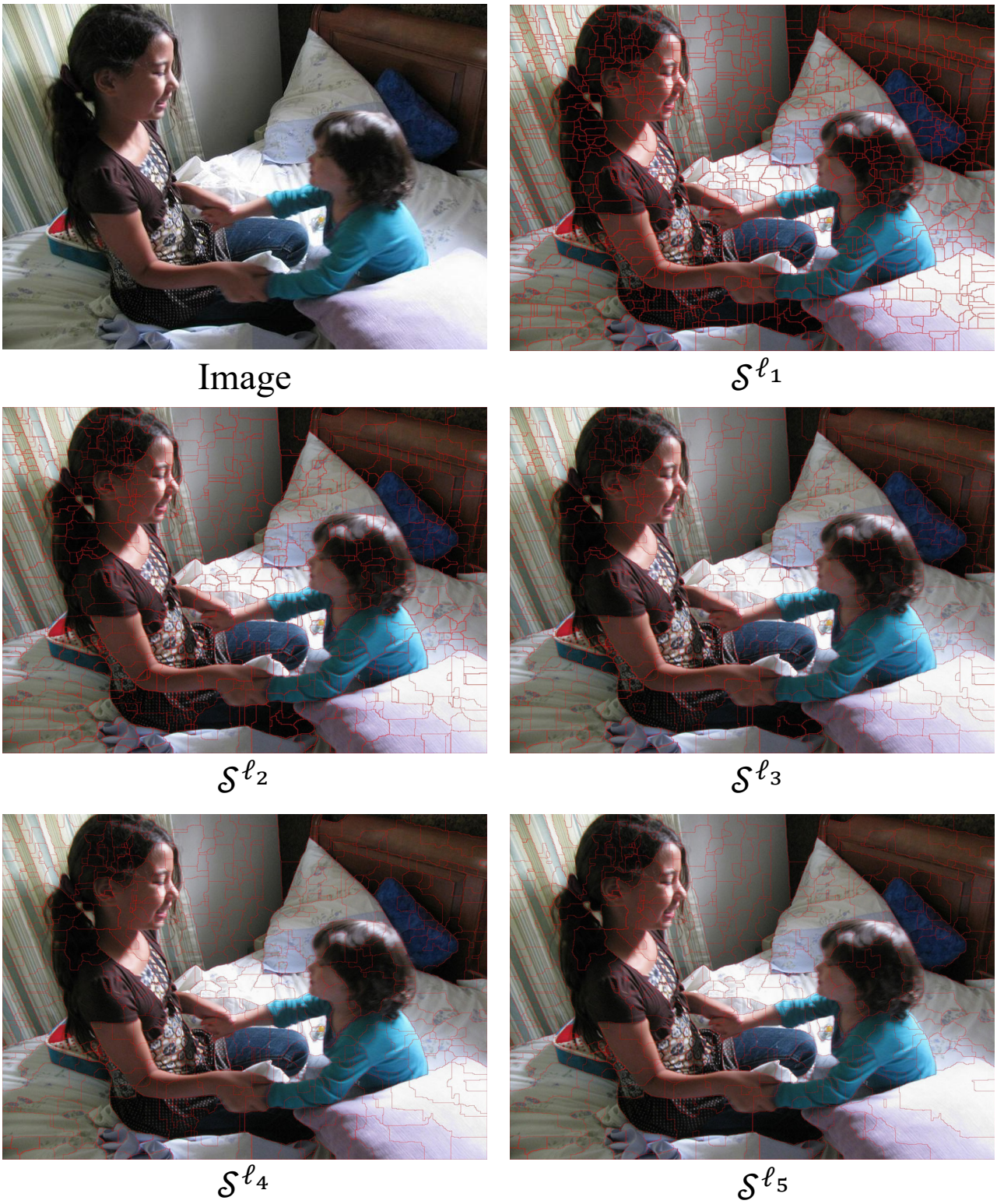
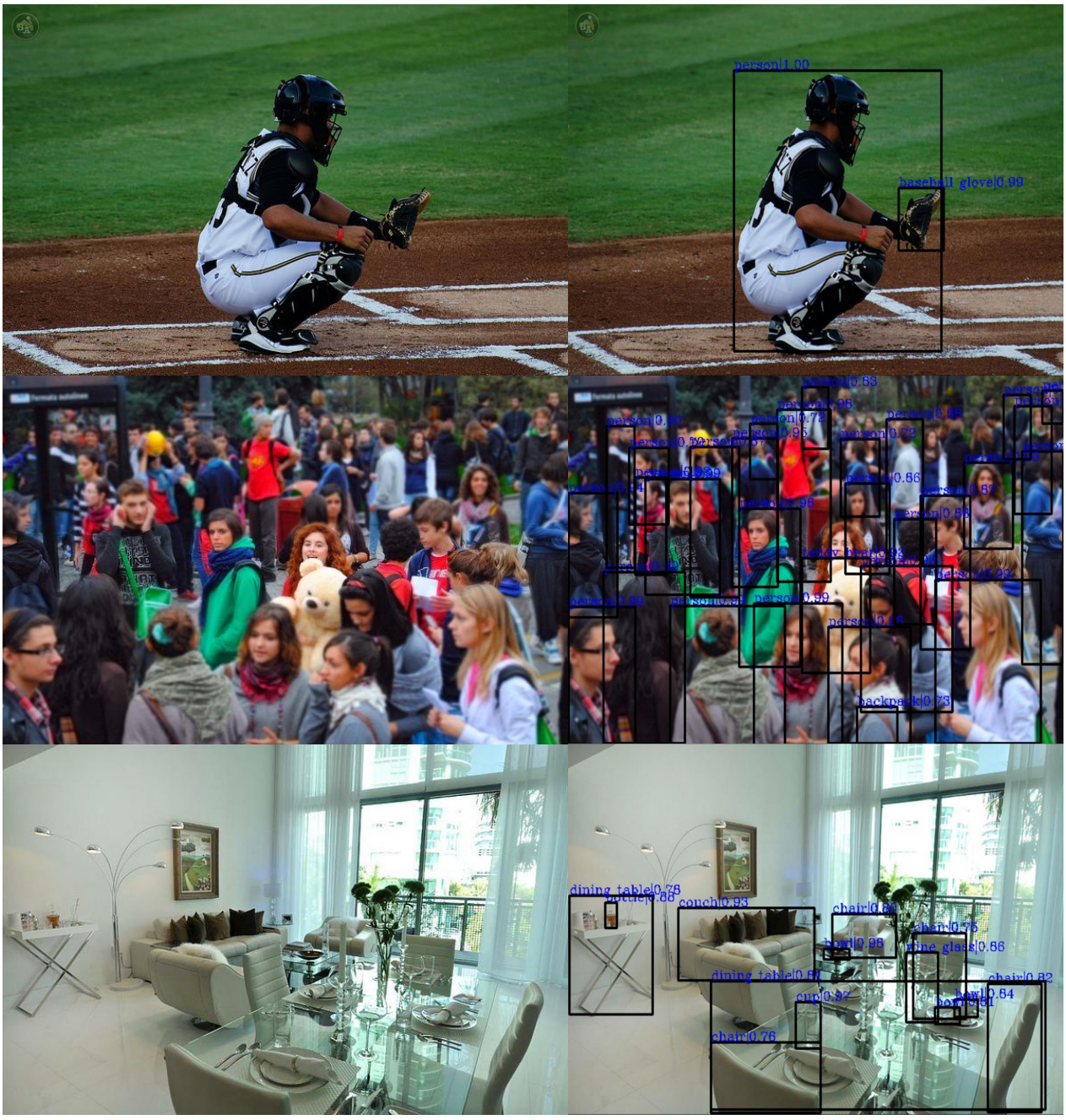


Figure 4. Sample results of superpixel hierarchy. Each superpixel hierarchy consists of 5 levels, $\{\mathcal{S}^{l_1}, \mathcal{S}^{l_2}, \mathcal{S}^{l_3}, \mathcal{S}^{l_4}, \mathcal{S}^{l_5}\}$. Images are from the MS COCO 2017 dataset [4].



(a) Image **(b) Result**

Figure 5. Sample detection results from our GraphFPN based method. Images are sampled from the MS COCO 2017 validation set [4].

(a) Image **(b) Result**

Figure 6. Sample detection results from our GraphFPN based method. Images are sampled from the MS COCO 2017 validation set [4].

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] 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*, pages 2117–2125, 2017. 1, 4

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 4, 5, 6, 7, 8, 9

[5] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):819–833, 2017. 1

[6] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 1, 3

[7] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *European Conference on Computer Vision*, pages 323–339. Springer, 2020. 1, 3, 4