HR-RCNN: Hierarchical Relational Reasoning for Object Detection

supplementary material

Hao Chen, Abhinav Shrivastava
{chenh,abhinav}@cs.umd.edu
University of Maryland
College Park, USA

Thanks for viewing the supplementary material. Section 1 provides more results for ablation study, *i.e.*, different combinations of relation graphs, additional results with other backbones, more ablation studies about group number and temperature. Section 2 provides more discussion about experimental analysis and visualization.

1 Ablation Study

Graph Combinations We study how to build relational graphs in using three heterogeneous graphs: pixel graph, scale graph, and RoI graph. We explore three basic ways to combine them: sequential, parallel, and joint (illustrated in Fig 1). For sequential combinations, we simply stack the reasoning modules sequentially and accumulate relationships gradually as information flows from one module to the next (illustrated in Fig 1(a)). For a parallel combination, every reasoning module has its box head and work separately. Finally, we propose a joint combination strategy where reasoning modules have separate branches but share the same box head. In such a case, our joint reasoning can leverage the advantage of both parallel reasoning (encode different relationships without being interrupted) and sequential reasoning (fuse the relationships in a single model using multitask relational reasoning). With this joint combination, we encode heterogeneous relations implicitly while retaining a smaller model footprint compared to other contemporary approaches.

It proves to be the most efficient and effective combination strategy for hierarchical reasoning. where a final average operation is used on their outputs to fuse the relation information. In contrast to sequential combination, parallel combination puts different reasoning modules in parallel, and As for joint combination, relation modules still work on separate branches but all branches share the same copy of parameter weights. Through joint training By parameter sharing, the box head implicitly encodes heterogeneous relationships during the training time and we can fuse the hierarchical relationships in such case.

Table 1 shows the results of different combination strategies: sequential, parallel, and joint. For this ablation, we utilize the pixel and RoI relations. Joint combination of reasoning components achieves the best mAP while retaining the model size, and is thus being default choice.

Other backbones. We show main results introduced by hierarchical reasoning based on Faster RCNN. Large improvements can be seen across different backbones, which clearly

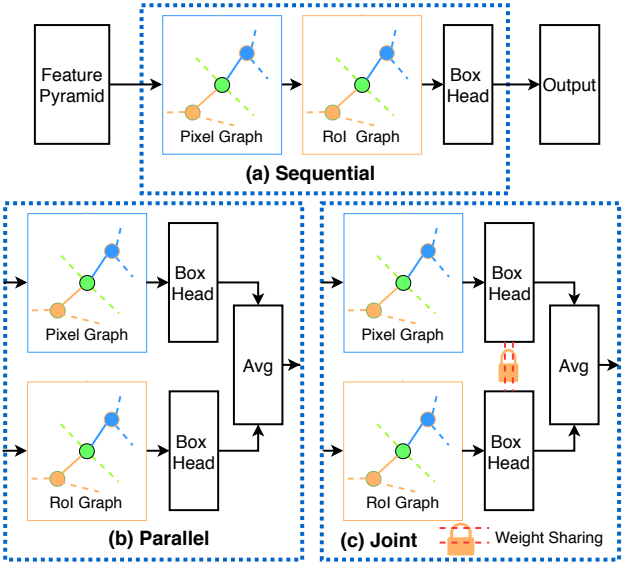


Figure 1: Different relation graph combination strategies. You can replace 'pixel graph' with 'scale graph' for its combination with RoI graph.

Table 1: Ablation results for different combination strategies. P: pixel relation, S: scale relation, R: RoI relation.

		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Sequential	(P→R)	38.8	60.1	42.4	23.1	42.2	50.9
Parallel	(P‖R)	39.2	60.3	22.5	22.9	42.3	**51.2**
Joint	(P+R)	**39.5**	**60.5**	**43.1**	**23.7**	**42.9**	51.1

Table 2: **More Main Results** on COCO validation set. The impact of using HR-RCNN with different backbones. All methods are based on Faster RCNN with feature pyramid network.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50 [□]	38.0	58.6	41.4	22.1	41.8	48.8
HR-RCNN	**41.6**	**61.8**	**45.2**	**25**	**45.1**	**54.2**
ResNet101 [□]	40.2	61.2	43.8	24.1	43.8	52.1
HR-RCNN	**42.8**	**63.1**	**46.3**	**25.5**	**46.4**	**55.8**
DCN-V2 [□]	40.8	62.0	44.5	24.2	44.0	54.0
HR-RCNN	**42.9**	**63.2**	**46.6**	**26.2**	**45.9**	**57.1**
VoVNetV2-39 [□]	39.8	61.1	43.1	24.7	43.0	50.1
HR-RCNN	**41.1**	**61.1**	**44.1**	**25.8**	**43.7**	**52.4**
MobileNet-V2 [□]	29.4	48.7	30.8	16.6	21.0	38.0
HR-RCNN	**34.5**	**53.4**	**36.5**	**19.8**	**36.3**	**45.4**

shows the effectiveness of our HR-RCNN.

Table 3: Temperature ablation results

T	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	41.3	61.3	44.6	**25.2**	44.7	53.6
2	**41.6**	**61.8**	**45.2**	25	**45.1**	**54.2**
3	41.2	61.7	43.9	**25.2**	44.5	53.8

Table 4: Group size ablation results

Groups	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	40.8	60.9	43.9	24.7	44	53.2
2	**41.6**	61.8	**45.2**	25	**45.1**	**54.2**
4	**41.6**	**61.9**	45	**25.5**	45	54.1
8	40.6	60.7	43.6	24	43.3	54.3

Table 5: Ablation results for refinement in HR-RCNN.

	Refinement	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
HR-RCNN	✓	**41.6**	**61.8**	**45.2**	**25**	**45.1**	**54.2**
w/o Refine		40.1	61	43	24.2	43.8	52.4

Temperature ablation In Tab. 3, we provide ablation results when changing the softmax temperature (L202-204 in the main paper) .

$$w_{ij} = \text{softmax}_j(\alpha_{ij}/T) = \frac{\exp(\alpha_{ij}/T)}{\sum_{k\in N(i)} \exp(\alpha_{ij}/T)}, \tag{1}$$

where w_{ij} is the normalized attention weights, $N(i)$ is the neighbor nodes for query node i. The final performance is robust to temperature settings and we use 2.0 as the default setting since it leads to the best performance.

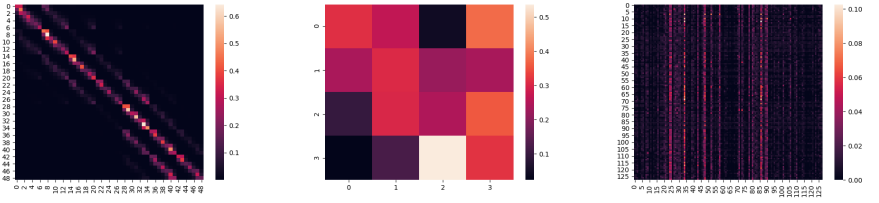
Groups number ablation In Tab. 4, we provide ablation results when given different groups for semantic distance. When groups number is 2 or 4, it leads to the best performance and we use 2 as the default setting in all experiments.

Refinement ablation In Tab. 5, we provide ablation results for refinement in HR-RCNN. For method without refinement, we utilize the region proposals from region proposal network (RPN) instead of the prediction of 1*st* stage at evaluation time. Without refinement, AP of our HR-RCNN drops by 1.5 points.

2 Results Analysis and visulization

Attention weights We plot the attention weights of pixel reasoning, scale reasoning, and RoI reasoning in Fig. 2 (a), (b), and (c) respectively. For pixel reasoning, most pixels have the highest attention weight from itself. For scale reasoning, feature maps from higher levels provide more enhancement for the query elements. For RoI reasoning, feature enhancement mainly comes from a few key region proposals regardless of the query proposals.

Detection Results In Fig 3, we shows some detection results by Faster RCNN and HR-RCNN. Due to hierarchical relation reasoning, HR-RCNN can find overlooked objects by local and global context (*e.g.*, the snowboard for 1st image, bottle for 3rd image, car for 4th image), and reject unreasonable predictions (*e.g.*, refrigerator prediction for 2nd image, sports ball for the 4th image).



(a) Pixel reasoning (b) Scale reasoning (c) RoI reasoning

Figure 2: Attention weights for visual reasoning

Figure 3: Detection results **Top**: Faster RCNN; **Bottom**: HR-RCNN. Via hierarchical reasoning, HR-RCNN can find overlooked objects (e.g., the snowboard for 1st image, bottle for 3th image, car for 4th image) and reject reject unreasonable predictions (e.g., refrigerator prediction for 2nd image, sports ball for the 4th image).

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[2] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. 2020.

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[4] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.