Appendix

This supplementary document is organized as follows:

- Section A shows more detailed comparison results under the **Mean Recall** metric between GPS-Net and state-of-the-art methods on the VG database. Results are summarized in Table 7. Meanwhile, we also treat the number of relationship predictions per object pair (k) as a hyper-parameter on VRD, and report Recall with respect to different k in Table 8.
- Section B first provides the qualitative comparisons between GPS-Net and a strong baseline named MOTIFS [7] under the SGDET protocol in Figure 6. Then, attention maps of different MP modules are visualized in Figure 7.

A. Quantitative Analysis

A.1. Mean Recall for Scene Graph on VG

As shown in Table 7, GPS-Net shows the best performance under all protocols. In particular, GPS-Net outperforms one very recent work named VCTREE-HL [2] by 2.3% on average over the three protocols.

A.2. Performance Comparison on VRD with various k

As revealed in previous works [42, 43, 6], each object pair may be described by several plausible predicates. In other words, it should have been formulated as a multi-label classification problem. Therefore, evaluation metrics based on the top-1 prediction (k=1) per object pair only may be unreasonable. Following [42, 43, 6], we further report recall with respect to different k (k=1, 70) and compare with state-of-the-art methods. As shown in Table 8, GPS-Net consistently achieves the best performance among state-of-the-art methods.

B. Qualitative Analysis

B.1. Generated Scene Graph

Figure 6 illustrates qualitative comparisons between GPS-Net and MOTIFS [7]. In Figure 6(a), it is shown that for nodes with low priority and relationships with high frequency, GPS-Net still makes better predictions than MOTIFS. Therefore, we owe this performance gain to the DMP module that encodes edge direction information and provide node-specific context. In Figure 6(b), it is shown that GPS-Net makes fewer mistakes than MOTIFS for nodes of high priority. We give this credit to the NPS-loss. Finally, in Figure 6(c), it can be observed that GPS-Net makes outstanding improvement in predicting low-frequency relationships, *e.g.*, *walking on* and *wearing*, via the help of the ARM module.

B.2. Attention Maps of Different MP Modules

We make qualitative comparisons between GCMP, S-GCMP, and DMP in Figure 7. Figures 7(a) and (b) show ground-truth object regions and the ground-truth relationship matrix. More specifically, in the relationship matrix, yellow cube denotes one relationship is presented, and purple cube represents the opposite. Figures 7(c)(d)(e) show the attention maps produced by GCMP, S-GCMP, and DMP respectively. It is clear that GCMP and S-GCMP produce very similar context for each node (elements in each column are similar). Only DMP obtains node-specific context (elements in each column are diverse). Furthermore, the attention map produced by DMP is highly consistent with the ground-truth relationship matrix in Figures 7(b). Therefore, our proposed DMP module plays a key role in SGG, helping GPS-Net to achieve state-of-the-art performance.

Model	SGDET			SGCLS			PREDCLS			Mean
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	
IMP$^\diamond$ [5]	-	3.8	4.8	-	5.8	6.0	-	9.8	10.5	6.8
FREQ$^\diamond$ [7]	4.5	6.1	7.1	5.1	7.2	8.5	8.3	13.0	16.0	9.8
MOTIFS$^\diamond$ [7]	4.2	5.7	6.6	6.3	7.7	8.2	10.8	14.0	15.3	9.6
KERN$^\diamond$ [23]	-	6.4	7.3		9.4	10.0	-	17.7	19.2	11.7
Chain$^\diamond$ [2]	4.6	6.3	7.2	6.3	7.9	8.8	11.0	14.4	16.6	10.2
Overlap$^\diamond$ [2]	4.8	6.5	7.5	7.2	9.0	9.3	12.5	16.1	17.4	11.0
Multi-Branch$^\diamond$ [2]	4.7	6.5	7.4	6.9	8.6	9.2	11.9	15.5	16.9	10.7
VCTREE-SL$^\diamond$ [2]	5.0	6.7	7.7	8.0	9.8	10.5	13.4	17.0	18.5	11.7
VCTREE-HL$^\diamond$ [2]	5.2	6.9	8.0	8.2	10.1	10.8	14.0	17.9	19.4	12.2
GPS-Net$^\diamond$	**6.9**	**8.7**	**9.8**	**10.0**	**11.8**	**12.6**	**17.4**	**21.3**	**22.8**	**14.5**

Table 7: Mean recall (%) of various methods across all the 50 relationship categories. All methods in this table adopt the same Faster-RCNN detector.

Pretrained	Model	Pre. k=1		Rel. k=1		Rel. k=70		Phr. k=1		Phr. k=70	
		R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Unknown	VRD-Full [16]	47.9	47.9	16.2	17.0	-	-	13.9	14.7	-	-
	PPRFCN [35]	47.4	47.4	19.6	23.2	-	-	14.4	15.7	-	-
	VTranse [37]	44.8	44.8	19.4	22.4	-	-	14.1	15.2	-	-
	Vip-CNN [39]	-	-	17.3	20.0	-	-	22.8	27.9	-	-
	VRL [40]	-	-	18.2	20.8	-	-	21.4	22.6	-	-
	KL distilation[43]	55.2	55.2	19.2	21.3	22.7	31.9	23.1	24.0	26.3	29.4
	MF-URLN [44]	58.2	58.2	23.9	26.8	-	-	31.5	36.1	-	-
ImageNet	Zoom-Net [42]	50.7	50.7	18.9	21.4	21.4	27.3	24.8	28.1	29.1	37.3
	CAI+SCA-M [42]	56.0	56.0	19.5	22.4	22.3	28.5	25.2	28.9	29.6	18.4
	RelDN[6]	-	-	19.8	23.0	21.5	26.4	26.4	31.4	28.2	25.4
	GPS-Net	**58.7**	**58.7**	**21.5**	**24.3**	**23.6**	**28.9**	**28.9**	**34.0**	**30.4**	**38.2**
COCO	RelDN [6]	-	-	25.3	28.6	28.2	33.9	31.3	36.4	34.5	42.1
	GPS-Net	**63.4**	**63.4**	**27.8**	**31.7**	**30.6**	**37.0**	**33.8**	**39.2**	**36.8**	**44.5**

Table 8: Performance comparison with state-of-the-art methods on the VRD dataset. Pre., Phr., and Rel. represent predicate detection, phrase detection, and relation detection, respectively. $-$ denotes that the result is unavailable.

Figure 6: Qualitative comparisons between GPS-Net and MOTIF with R@20 in the SGDET setting. Green boxes are detected objects with IOU larger than 0.5 with the ground-truth. Green edges are predictions of relationships that are consistent with the ground-truth. Yellow boxes (edges) denote reasonable detections of objects (relationships), but are not annotated in the database. Red boxes (edges) represent ground-truth objects (relationships) that have no match with the detection results by the algorithm.

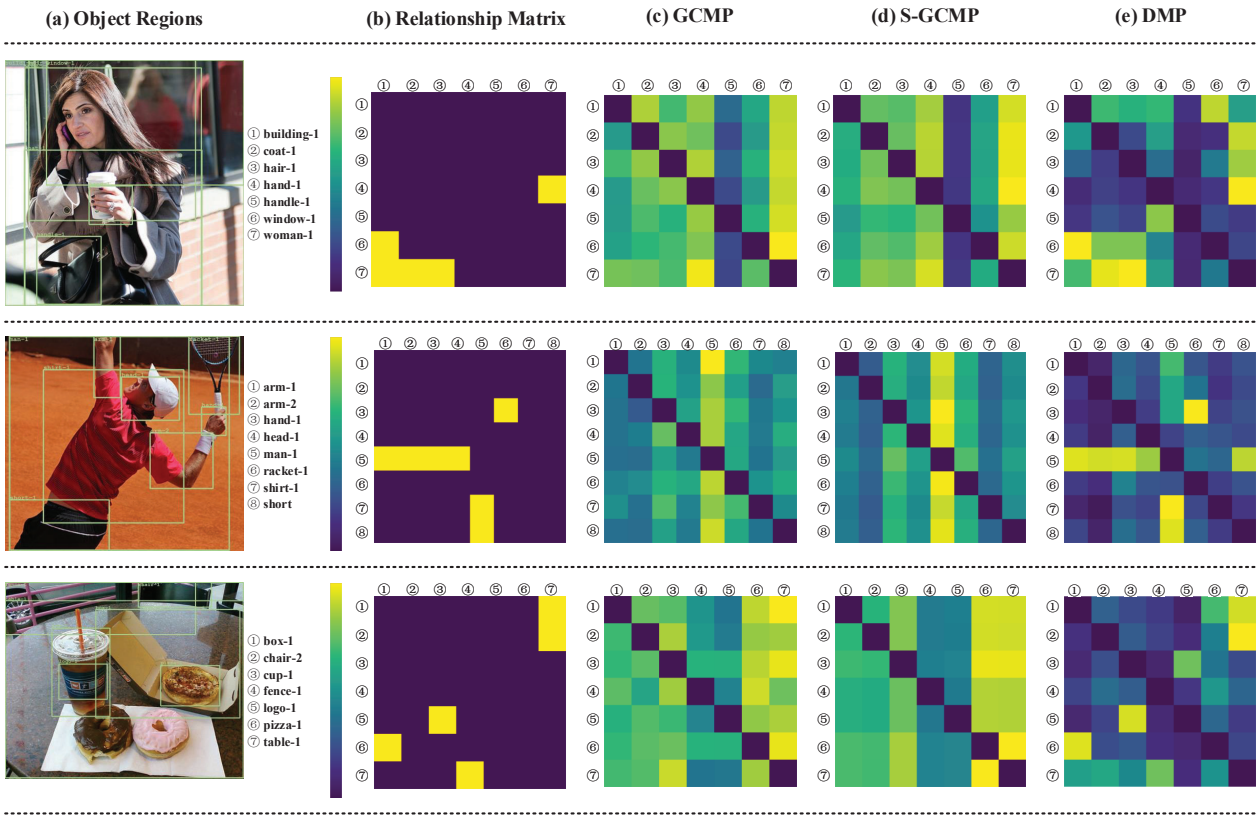

Figure 7: Attention Maps of Different MP Modules.