Relational Context Learning for Human-Object Interaction Detection
-Supplementary Material-

Sanghyun Kim Deunsol Jung Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
{sanghyun.kim, deunsol.jung, mscho}@postech.ac.kr
http://cvlab.postech.ac.kr/research/MUREN

A. Appendix

In this supplementary material, we provide additional experimental results and analysis to support our method including qualitative results.

A.1. Performance Comparison with Deeper Backbone

Table A1 compares recent HOI methods with ours using deeper backbone network on HICO-DET [2] and V-COCO [4] benchmarks. We observe that our proposed method with Resnet-50 backbone outperforms previous transformer-based methods with Resnet-101 backbone. Especially, MUREN gains significant improvement over UPT [9] with Resnet101-DC5 by 7.5%p and 3.9%p on two scenarios in V-COCO, respectively. MUREN with Resnet-101 backbone further increases this gap to 8.3%p and 5.3%p. It shows the effectiveness of MUREN.

Method	Backbone	HICO-DET		V-COCO	
		Default	Known Object	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
HoiTrans [11]	R101	26.61	29.13	-	-
QPIC [6]	R101	29.90	32.38	58.3	60.7
CDN [8]	R101	32.07	35.05	63.9	65.9
UPT [9]	R101	32.31	35.65	60.7	66.2
UPT [9]	R101-DC5	32.62	31.41	61.3	67.1
MUREN	R50	32.87	35.52	68.8	71.0
MUREN	R101	**33.28**	**35.85**	**69.6**	**72.4**

Table A1. Performance comparison with deeper backbone. We report mAP on Full split for HICO-DET [2] and role average precision (AP_{role}) under two scenarios for V-COCO [4].

A.2. Model Complexity Analysis

In Table A2, we report FLOPS and the number of parameters to analyze model complexity compared with previous transformer-based methods. Following DisTR [10], we compute average FLOPS over the first 100 images in the V-COCO test set with the tool flop_count_operators from Detectron2 [7]. We observe that MUREN has comparable FLOPS compared with existing HOI methods. MUREN only introduces 4.9% extra FLOPS compared with DisTR, a state-of-the-art two-branch method, although MUREN has one more branch than two-branch architecture. MUREN-M and MUREN-S are the same MUREN model used in the experiments of our main paper but with smaller numbers of parameters (we make these smaller models simply by decreasing the number of branch layer L). The results show that all our models clearly outperform previous methods in both scenarios, $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$, with comparable or less FLOPS and inference time. In particular, the smallest of ours, MUREN-S, performs better than the state-of-the-art two-branch method, DisTR, by 1.1%p in both scenarios, while it consumes 2.3M smaller parameters and 3.1G less FLOPS than DisTR.

Method	Backbone	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Params (M)	FLOPS(G)
QPIC [6]	R50	58.8	61.0	41.68	87.87
QPIC [6]	R101	58.3		60.62	156.18
AS-Net† [3]	R50	53.9	-	52.75	88.86
HOTR† [5]	R50	55.2		51.41	88.78
HOITrans [11]	R101	52.9		60.62	156
CDN† [8]	R50	63.9	64.4	51.14	93.19
DisTR† [10]	R50	66.2		57.31	94.23
MUREN	R50	68.8	71.0	69.3	98.7
MUREN-M	R50	68.3	70.6	59.6	93.6
MUREN-S	R50	67.3	69.6	55.0	91.1

Table A2. Model complexity comparison. † indicates two-branch methods. Following DisTR [10], we report role average precision on V-COCO test set, the number of parameters, and FLOPS.

A.3. Impact of Sequential Embedding

We utilize fine-grained context information (unary and pairwise relation contexts) to enrich holistic context information (triple relation context). For this, we embed the unary and the pairwise relation contexts to the triplet relation context in a sequential manner. To investigate the im-

pact of sequential embedding, we replace sequential embedding with parallel embedding. Specifically, the input query of cross attention in Eq. 9 is replaced with the results of Eq. 2. Then, we combine unary-embedded and pairwise-embedded triplet context with MLP to generate a multiplex relation context as follows:

$$\tilde{\mathbf{f}}_i^{HOI} = \text{CrossAttn}(\mathbf{f}_i^{HOI}, U_i), \quad \text{(A1)}$$

$$\hat{\mathbf{f}}_i^{HOI} = \text{CrossAttn}(\mathbf{f}_i^{HOI}, P_i), \quad \text{(A2)}$$

$$m_i = \text{CrossAttn}(\text{MLP}([\tilde{\mathbf{f}}_i^{HOI}; \hat{\mathbf{f}}_i^{HOI}]), \mathbf{X}). \quad \text{(A3)}$$

As shown in Table A3, the performance drops by 1.33%p and 1.28%p in the two scenarios. Additionally, we change the unary-pairwise embedding order to pairwise-unary. We observe that the performance drops by 1.51%p and 1.41%p in the two scenarios.

Method	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
MUREN	**68.8**	**71.0**
w.o sequential embedding	67.4	69.7
w. pairwise-unary order	67.2	69.6
w.o sequential embedding	64.7	67.1

Table A3. The impact of sequential embedding.

A.4. Impact of intermediate loss

As shown in Table A4, we observe that removing the intermediate loss from MUREN decreases the performance by 4.1%p in scenario #1 and 3.9%p in scenario #2. As observed in DETR [1], the intermediate supervision tends to improve detecting the correct number of object of each class. Similarly, DETR-like HOI methods learn to detect the correct number of HOI instance of each HOI class with the intermediate supervision.

Method	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
MUREN	**68.8**	**71.0**
w.o intermediate loss	64.7	67.1

Table A4. Impact of intermediate loss.

A.5. Additional Qualitative Results

In this section, we show additional qualitative results to analyze MUREN and compare MUREN with CDN [8]. CDN adopts two-branch architecture where one is responsible for human-object pair detection (*i.e.*, human-object branch) and the other for interaction classification (*i.e.*, interaction branch). As shown in Figure A1, CDN fails to find HOI instances in complicated scenes. In contrast, MUREN successfully detects HOI instances as we utilize the multiplex relation context in an HOI instance for relational reasoning. Moreover, we observe that the human-object branch of CDN, which is responsible for human-object pair detection, tends to focus only on an object to detect human-object pairs in Figure A2a. It might lead to inaccurate localization of human since they ignore human context information. However, we have designed three-branch architecture to focus on each sub-tasks. Therefore, MUREN properly localizes the human and the object as shown in Figure A2b. In Figure A3, we also show more visualization of cross-attention maps in MUREN. Our proposed method focuses on the region that contains the context information for each sub-task (**b-d** column in Figure A3). Moreover, the multiplex relation embedding module (MURE) focuses on the region that contains the context information about all sub-task for relational reasoning (**e** column in Figure A3). Especially, our proposed method properly attends to the regions for discovering the HOI instance in complicated scenes (**2**,**3** row in Figure A3). More qualitative results for detected HOI instances can be found in Figures A4 and A5.

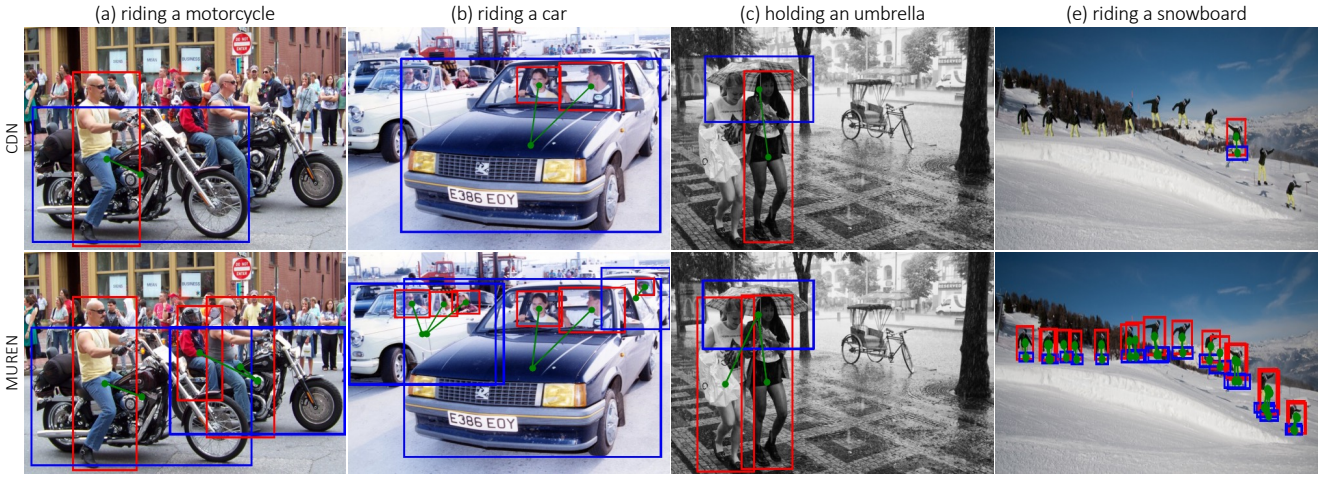
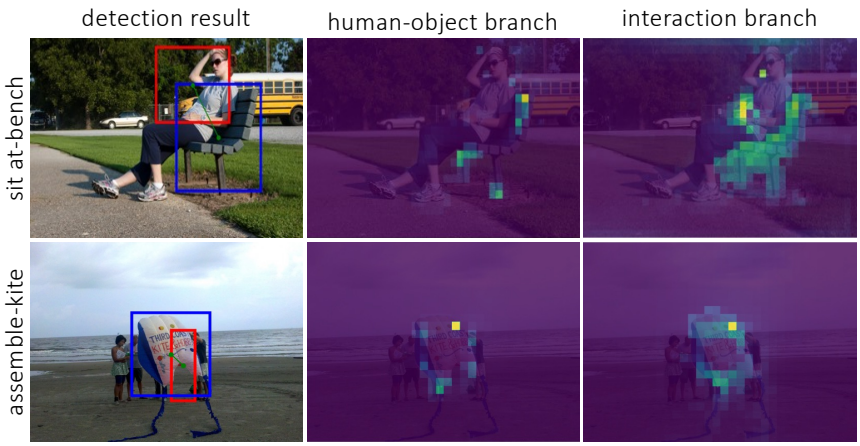
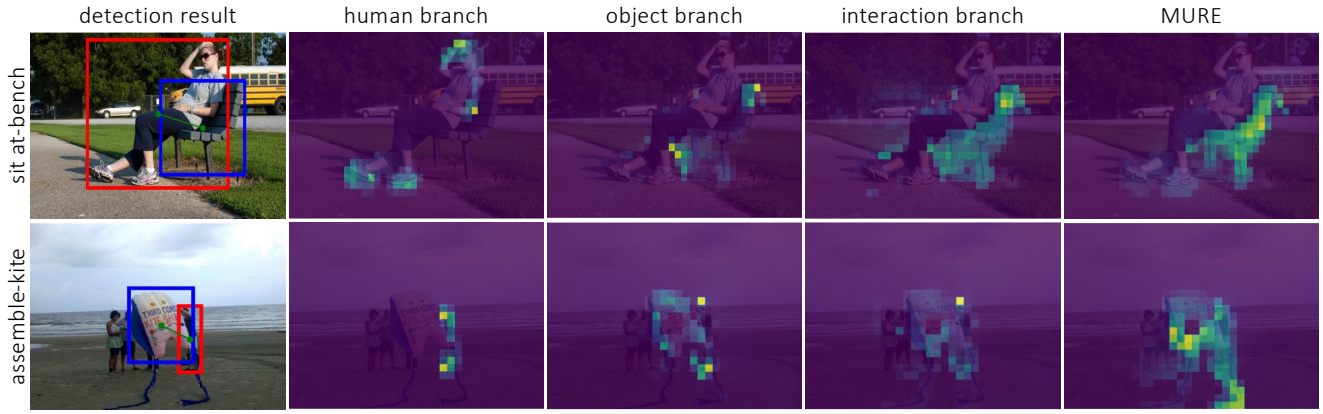


Figure A1. Visualization of HOI detection results. First row: HOI detection results from CDN [8]. Second row: HOI detection results from MUREN. Red boxes, blue boxes and green lines indicate humans, objects, and interactions, respectively. Our proposed method successfully detects HOI instances with the relation context information in complicated scenes.



(a) Visualization of HOI detection results and cross-attention maps from CDN [8]



(b) Visualization of HOI detection results and cross-attention maps from MUREN

Figure A2. Visualization of HOI detection results and cross-attention maps. (a) human-object and interaction branch columns indicate the cross-attention map from human-object branch and interaction branch, respectively. (b) visualization of the cross-attention maps in each branch and the multiplex relation embedding module (MURE). All cross-attention maps come from the last layer of each branch. Red boxes, blue boxes and green lines indicate humans, objects, and interactions, respectively.

Figure A3. Visualization of HOI detection results and cross-attention maps in each branch and the multiplex relation embedding module (MURE). All cross-attention maps come from the last layer of each branch. Red boxes, blue boxes and green lines indicate humans, objects, and interactions, respectively.

Figure A4. Visualization of HOI detection results on HICO-DET [2]. Red boxes, blue boxes and green lines indicate humans, objects, and interactions, respectively.

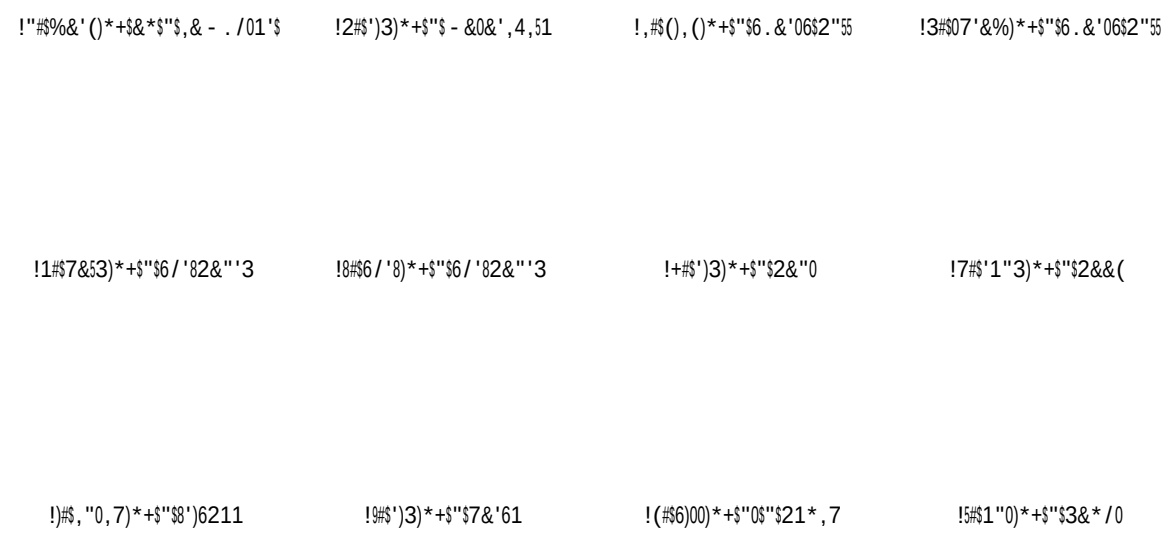


Figure A5. Visualization of HOI detection results on V-COCO [4]. Red boxes, blue boxes and green lines indicate humans, objects, and interactions, respectively.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1, 4

[3] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 1

[4] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 5

[5] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 1

[6] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1

[7] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[8] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 1, 2, 3

[9] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 1

[10] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 1

[11] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021. 1