Supplementary Materials for Target Adaptive Context Aggregation for Video Scene Graph Generation

Yao Teng[1] Limin Wang[1] ✉ Zhifeng Li[2] Gangshan Wu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Tencent AI Lab, Shenzhen, China
tengyao19980325@gmail.com, {lmwang, gswu}@nju.edu.cn, michaelzfli@tencent.com

A. Quantitative Analysis

The partial results on mean Recalls [4] of various methods are shown in Table 1.

Method	PredCls		SGCls		SGDet	
	mR@20	mR@50	mR@20	mR@50	mR@20	mR@50
Freq Prior [6]	58.79	70.50	36.21	40.28	25.46	35.86
G-RCNN [5]	59.61	67.39	37.80	41.43	28.61	37.06
RelDN [7]	71.27	80.68	41.79	45.23	30.97	41.42
Ours	**73.60**	**82.67**	**42.69**	**46.32**	**31.31**	**41.82**

Table 1. Mean recall [4] (%) of various models with ResNet-101 [2] on all images in AG. The number of triplets per frame is set to a limit of 50 and top 7 predictions for each pair are kept when evaluating.

B. Qualitative Analysis

B.1. Visualization of the per-class performance

As shown in Figure 1, we compare our method to RelDN with per-class analysis for frame-level VidSGG. Our method performs well on motion-related classes such as *lying on*, *wiping* etc.

B.2. Visualization of 2D relation feature

In Figure 3, we provide the origin image and its feature map produced by 2D ResNet-50 [2]. The highlighted areas represent the high activation. In this figure, the activation of the man with the bicycle is far greater than that of the background, which illustrates that our model obtains the direct interaction between the visual entities.

B.3. Visualization of 3D relation feature

In Figure 2, we provide the feature maps produced by an I3D ResNet-50 [1] of the short video clip whose center frame is the image mentioned above. We also provide another frame ahead of it in temporal dimension. In this figure, the bicycle with the man is the entities with the most motion information. The I3D appropriately obtains the movement changes in the video clip and the high activation in the feature map from the I3D indicates the movement.

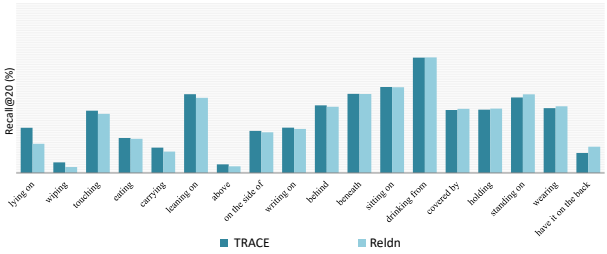


Figure 1. The per-class performance of our model compared to the state-of-the-art.

B.4. Visualization of Hierarchical Relation Tree

In Figure 4, we provide our Hierarchical Relation Tree in one frame. In this figure, the tree is built in a bottom-up manner and gradually expands the scope of its spatial coverage. Therefore, at the lower levels of the tree, our framework has the potential to obtain fine-grained relation feature, while at the top it obtains coarse-grained and long-distance relation feature.

B.5. Visualization of the results

In Figure 5, we provide the examples of frame-level video scene graphs generated by our model and RelDN [7] in AG dataset [3]. In Figure 6, we provide the examples of frame-level video scene graph generation on frames sampled from the same video clip.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE Computer Society, 2017. 1, 2

✉: Corresponding author.

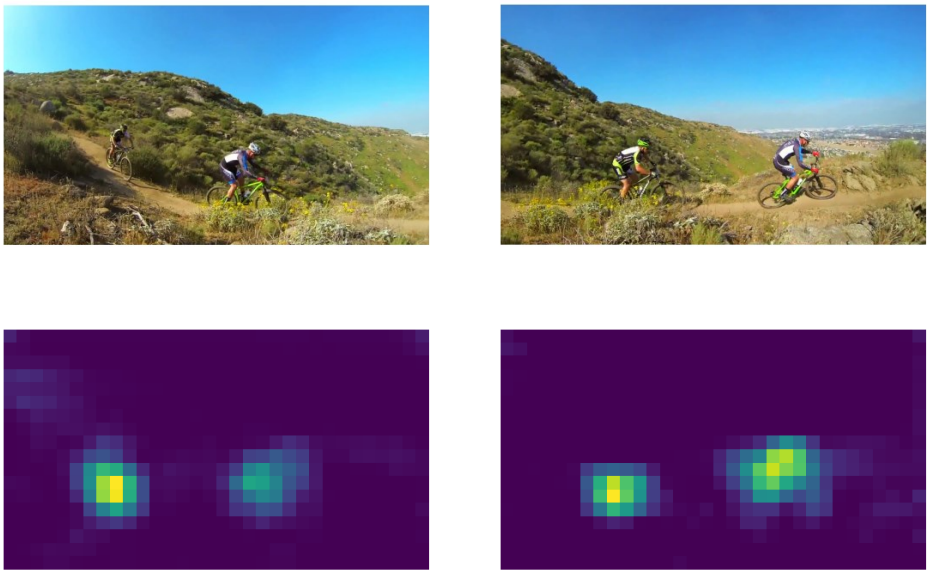


Figure 2. The visualization of the frames and the corresponding spatio-temporal feature map. The center frame and another frame ahead of it in temporal dimension are presented on the top line. The corresponding feature maps produced by an I3D ResNet-50 [1] of the short video clip are presented on the bottom line across the temporal dimension.

Figure 3. The visualization of the frame and the corresponding feature map produced by 2D ResNet-50 [2].

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 1, 2

[3] Jingwei Ji, Ranjay Krishna, Fei-Fei Li, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10233–10244. IEEE, 2020. 1, 3

[4] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628. Computer Vision Foundation / IEEE, 2019. 1

[5] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, pages 690–706. Springer, 2018. 1

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840. IEEE Computer Society, 2018. 1

[7] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, pages 11535–11543. Computer Vision Foundation / IEEE, 2019. 1, 3

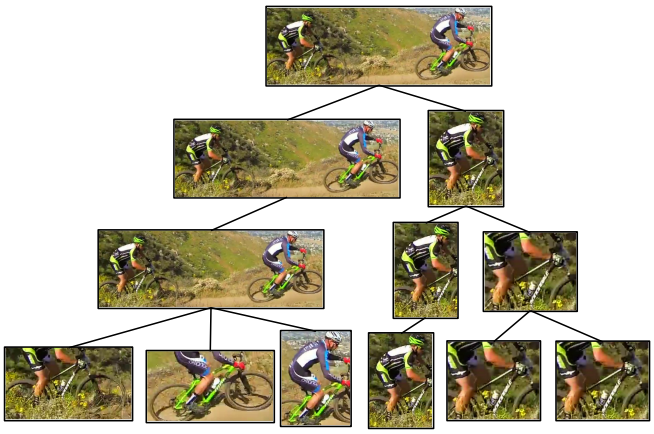
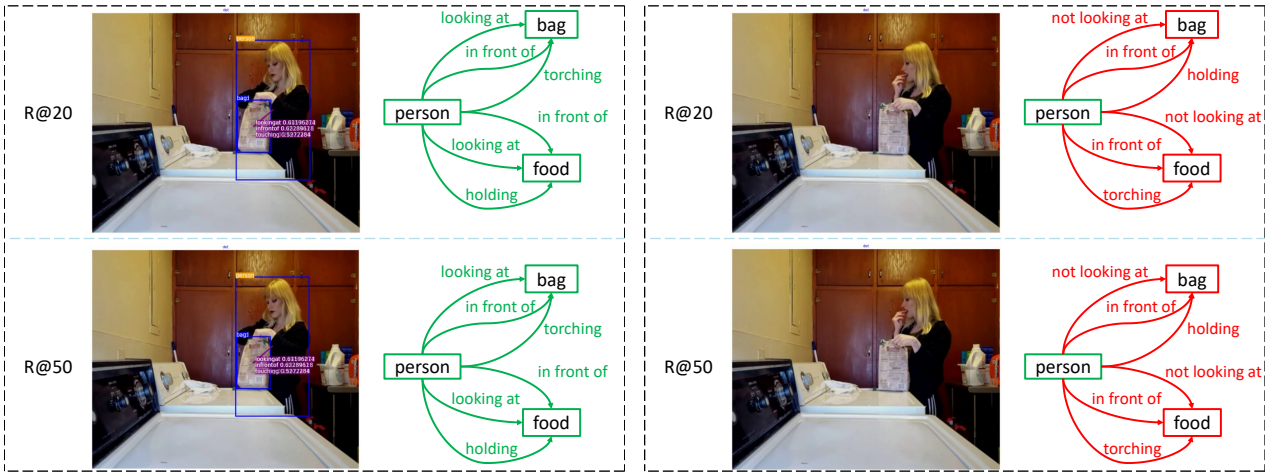


Figure 4. The visualization of Hierarchical Relation Tree. Each node is filled by the feature generated from the corresponding patch in the image. The repeated leaf nodes are attributed to the object detection algorithm.

Figure 5. Qualitative comparisons between our model and RelDN [7] at Recall@20 on SGDet in AG [3]. In each group (black dashed boxes), the top graph is our result and the bottom one is the output of RelDN [7]. In each graph, green boxes are objects which are contained in the predicted triplets and have IOU larger than 0.5 with the ground-truth boxes. Green edges are predicted relations which hit the ground-truth. Red boxes and edges are the ground-truth objects and relations which have no match with the results.



(a) Frame 1.

(b) Frame 2.

(c) Frame 3.

(d) Frame 4.

Figure 6. Qualitative results of our model at Recall@20 and Recall@50 on SGDet in frames sampled from a single video. In each group (black dashed boxes), the top graph is our result at Recall@20 and the bottom one is the output at Recall@50. In each graph, green boxes are objects which are contained in the predicted triplets under each metric and have IOU larger than 0.5 with the ground-truth boxes. Green edges are predicted relations which hit the ground-truth. Red boxes and edges are the ground-truth objects and relations which have no match with the results.