OED: Towards One-stage End-to-End Dynamic Scene Graph Generation Supplementary Material

Guan Wang[1] Zhimin Li[2] Qingchao Chen[3] Yang Liu[1*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Tencent Inc. [3]National Institute of Health Data Science, Peking University
w.g@stu.pku.edu.cn zhiminli.cn@outlook.com {qingchao.chen, yangliu}@pku.edu.cn

Our supplementary material presents long-tail related performance comparison in section 1 and some qualitative cases to demonstrate different effects of our spatial context aggregation and temporal context aggregation in section 2.

1. Long-tail Problem

Visual relation detection presents a significant challenge due to its long-tailed distribution. To thoroughly assess our methodology, we conduct a comparison against sota methods that are publicly available in SGDET task, utilizing metrics related to the long-tail issue as detailed in Table 1, in line with the approach taken by TEMPURA [1]. Our method consistently outperforms across all evaluated metrics, achieving notable improvements in mean Recall@50 (mR@50) by 9.3% and 5.9% over the most competitive baseline within the *With Constraint* and *No Constraints* scenarios, respectively.

Additionally, we delve into the performance across different class segments—head, body, and tail—using mR@50 for a more granular analysis. The results demonstrate our method's superior performance across all metrics, establishing a significant lead over other approaches, despite TEMPURA being specifically designed to address the long-tail challenge.

2. Qualitative Results

2.1. Spatial Context Aggregation

As can be seen from Fig. 1, we visualize the decoder attention map for the predicted scene graph. The pair-wise instance heatmaps generated by pair-wise instance decoder highlight both the subject and object area, as shown in Fig. 1 (a), meaning that our model reasons for pair-wise instance from a long-range global image context. Meanwhile, as shown in Fig. 1(b), some areas with relatively higher attention weight indicate the relational region in pair-wise relation heatmaps generated by pair-wise relation decoder. It is obvious that the decoder has the ability to find the discriminative part for pair-wise instance and relation information in our one-stage approach.

2.2. Temporal Context Aggregation

As illustrated in Fig. 2, we present several scene graph predictions to demonstrate the distinct effects of Temporal Context Aggregation (TCA). Given a target frame, we attempt to generate scene graphs using our model with only Spatial Context Aggregation (SCA) and a complete framework incorporating both SCA and TCA.

SCA can provide sufficient information for relatively static subject-object pairs and predicates with weak or no temporal dependency, as shown in Fig. 2(a), e.g., $<$ cup $-$ in front of $-$ person $>$, $<$person $-$ look at $-$ floor$>$. However, spatial context struggles to handle cases involving blurred subject-object pairs and predicates with strong temporal dependencies. In the first row, the model with only SCA fails to detect the blurred broom. In the second row, the SCA model is unable to recognize the predicate *drink from* due to the lack of temporal dependency. In contrast, our model with Spatial-Temporal Aggregation effectively captures both the temporal dynamics of object motion and the dependency of predicates, resulting in more accurate scene graph prediction, as shown in Fig. 2(b).

References

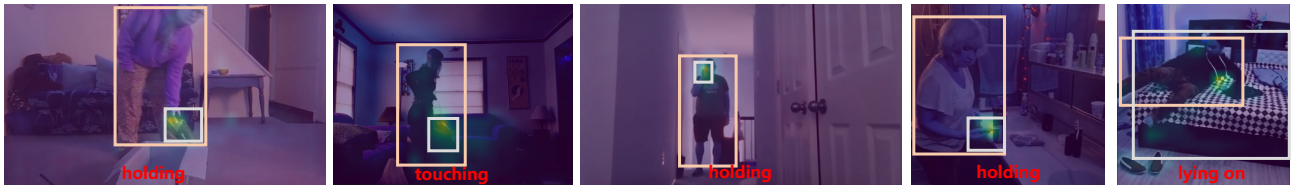
[1] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22803–22813, 2023. 1

*Corresponding author

Table 1. Performance comparison under both Recall and Mean Recall Metrics and Head-body-tail classes with mR@50.

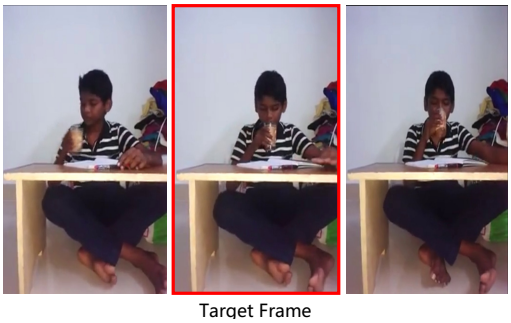
Method	With Constraint									No Constraints								
	R@10	R@20	R50	mR@10	mR@20	mR50	Head	Body	Tail	R@10	R@20	R50	mR@10	m20	mR50	Head	Body	Tail
STTran[4]	25.2	34.0	36.9	16.5	20.8	22.2	39.00	25.53	12.78	24.5	36.1	48.8	20.9	29.6	39.1	45.39	52.32	26.67
DSG-DETR[8]	30.4	34.9	36.0	18.0	21.3	22.0	37.91	25.84	12.62	32.3	40.9	48.2	23.6	30.1	36.5	44.76	52.65	21.09
TEMPURA[19]	28.0	33.3	34.8	18.4	22.5	23.6	36.53	23.79	18.15	29.8	38.0	46.3	24.5	33.8	43.6	42.50	50.66	38.86
Ours	**33.5**	**40.9**	**48.9**	**20.9**	**26.9**	**32.9**	**51.44**	**36.38**	**22.72**	**35.3**	**44.0**	**51.8**	**26.3**	**39.5**	**49.5**	**48.75**	**56.12**	**44.89**

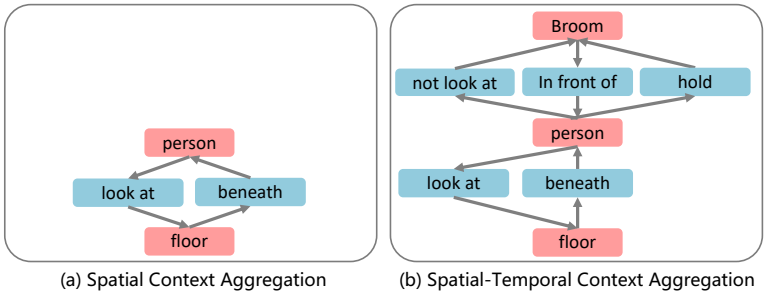
(a) Attention map of pair-wise instance feature



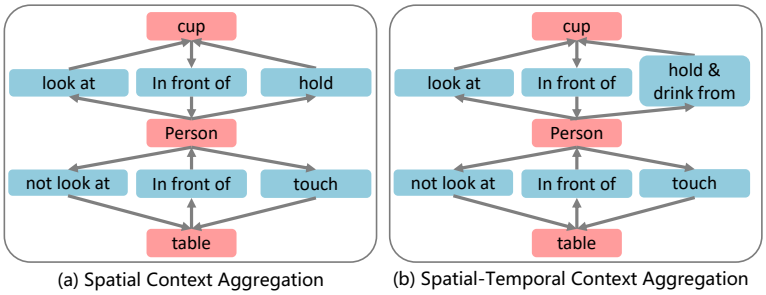
(b) Attention map of pair-wise relation feature

Figure 1. Visualization of attention maps of pair-wise instance feature and pair-wise relation feature.





(a) Spatial Context Aggregation (b) Spatial-Temporal Context Aggregation



(a) Spatial Context Aggregation (b) Spatial-Temporal Context Aggregation

Figure 2. Part of predictions of our Spatial Context Aggregation and Spatial-Temporal Context Aggregation under *with constraint* setting, which match with ground truth.