Tri-modal Confluence with Temporal Dynamics for Scene Graph Generation in Operating Rooms

Diandian Guo*[1,2], Manxi Lin*[3], Jialun Pei [1,2](✉), He Tang[4], Yueming Jin[5], Pheng-Ann Heng[1,2]

[1]Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
[2]Institute of Medical Intelligence and XR,
The Chinese University of Hong Kong, Hong Kong, China
[3]Technical University of Denmark, Kongens Lyngby, Denmark
[4]Huazhong University of Science and Technology, Wuhan, China
[5]National University of Singapore, Singapore, Singapore
jialunpei@cuhk.edu.hk

Abstract. A comprehensive understanding of surgical scenes allows for monitoring of the surgical process, reducing the occurrence of accidents and enhancing efficiency for medical professionals. Semantic modeling within operating rooms, as a scene graph generation (SGG) task, is challenging since it involves consecutive recognition of subtle surgical actions over prolonged periods. To address this challenge, we propose a **Tri**-modal (*i.e.*, images, point clouds, and language) confluence with **Temp**oral dynamics framework, termed TriTemp-OR. Diverging from previous approaches that integrated temporal information via memory graphs, our method embraces two advantages: 1) we directly exploit bi-modal temporal information from the video streaming for hierarchical feature interaction, and 2) the prior knowledge from Large Language Models (LLMs) is embedded to alleviate the class-imbalance problem in the operating theatre. Specifically, our model performs temporal interactions across 2D frames and 3D point clouds, including a scale-adaptive multi-view temporal interaction (ViewTemp) and a geometric-temporal point aggregation (PointTemp). Furthermore, we transfer knowledge from the biomedical LLM, LLaVA-Med, to deepen the comprehension of intraoperative relations. The proposed TriTemp-OR enables the aggregation of tri-modal features through relation-aware unification to predict relations to generate scene graphs. Experimental results on the 4D-OR benchmark demonstrate the superior performance of our model for long-term OR streaming. Codes are available at https://github.com/RascalGdd/TriTemp-OR.

Keywords: Surgical scene understanding · Scene graph generation · Temporal OR interaction · Multi-modality learning.

* Equal contribution.

1 Introduction

With the growing complexity and diversity of modern surgical procedures, the automated scene understanding of the operating room (OR) using computer-assisted systems has become a critical necessity [28,6]. Compared to specific surgical assistance interventions, *e.g.*, surgical phase recognition [9], instrument segmentation [21], and anatomy tracking [10], holistic scene modeling of operating theatres [28,27,6] better facilitates coordination and communication among surgical teams, and embraces more organized and efficient surgical process optimization in ORs. Scene graph generation (SGG) [2,25,3], by condensing image content through nodes and edges along with their relationships, enables effective monitoring of surgical procedure and detailed guidance for operating activities [28]. However, generating scene graphs within the complex OR environment is a challenging task that requires relational associations between subjects and objects (*e.g.*, surgeons, patients, and medical equipment) in ORs and involves continuous follow-up of operation details throughout the surgical period.

To address this challenge, the pioneering work 4D-OR [28] adopts a multi-stage framework, initially identifying medical staff and scene objects via a 3D pose estimation model and an object detector, followed by a modified 3DSGG network [24] to generate scene graphs. Nevertheless, this OR scene graph generation (OR-SGG) approach neglects to explore temporal information to enhance graph construction. LABRAD-OR [27] introduces a memory scene graph to represent dynamic temporal characteristics and combines visual information extracted from 2D and 3D modalities to achieve more consistent predictions. However, the multi-stage spatial-temporal fusion, *i.e.*, first generating a coarse graph and then conducting sequential interaction, diminishes the sensitivity to fine-grained visual features and exploitation of spatial-temporal information. For better utilizing temporal dynamics, we suggest directly leveraging bi-modal temporal cues for finer-grained hierarchical feature interaction. Since the process of surgery exhibits wide variations in the duration of activities at different phases, it is inevitable to confront a significant imbalance of classes between frequent ones (*e.g.*, 'Lying on') and rare ones (*e.g.*, 'Cutting').

Several natural scene graph models [13,26,5] employ knowledge distillation from large language models (LLMs) [17,1,22] pre-trained on large amounts of textual data to provide enriched and subtle semantic representations to improve the prediction accuracy for rare classes. While LLMs have shown considerable effectiveness in various semantic scene modeling tasks [15,8,18] depending on the broad understanding of language and context, the knowledge gap between open vocabulary and medical semantics hinders the application of LLMs in the OR-SGG domain. In this regard, it is appropriate to utilize LLMs fine-tuned on biomedical data to mitigate the class-imbalance problem.

In this work, we propose an end-to-end tri-modal OR-SGG model, named TriTemp-OR. Our model first performs hierarchical feature interaction with 2D and 3D temporal information in video streaming through a scale-adaptive multi-view temporal interaction (ViewTemp) and a geometric-temporal point aggregation (PointTemp). ViewTemp embraces a scale-adaptive feature partition strat-

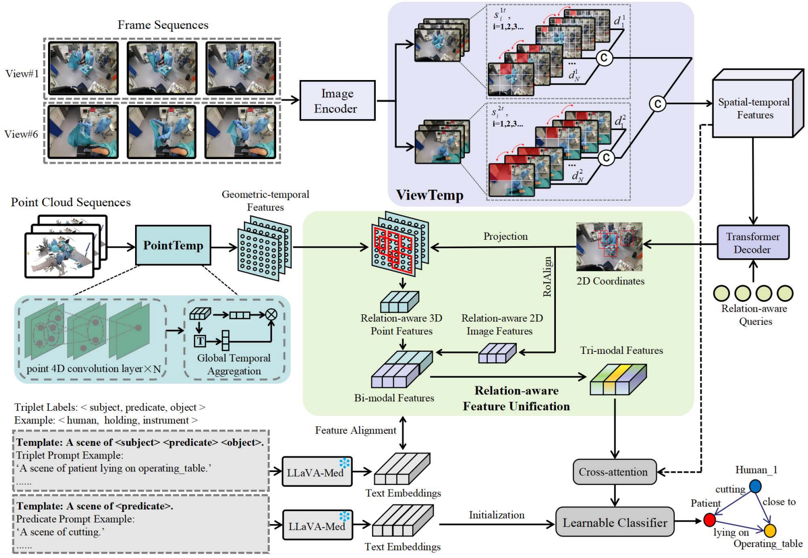


Fig. 1: Overview of the proposed TriTemp-OR for scene graph generation in ORs.

egy based on the receptive fields from different views, to encourage multi-scale feature identification in multi-view 2D video streaming. In parallel, PointTemp captures 3D point cloud temporal information through point 4D convolutions and global temporal aggregation with self-attention [7,23]. Concentrating on subject-object interaction spaces, we also introduce a relation-aware feature unification operation to aggregate 2D spatial-temporal and 3D geometric-temporal information. More importantly, the proposed TriTemp-OR aligns multi-view image and point cloud features with textual semantic features distilled from the biomedical pre-trained LLM (*i.e.*, LLaVA-Med [14]) during training, yielding effective relation-aware tri-modal representations for predicting subject-object relations and generating scene graphs. Experimental results demonstrate that TriTemp-OR achieves superior performance on the 4D-OR dataset, indicating the potential application value in surgical intervention and assistance.

2 Method

Fig. 1 illustrates our tri-modal confluence with temporal dynamics framework TriTemp-OR for generating scene graphs in operating rooms. We first utilize a ResNet-50 [12] to extract spatial features from multi-view video frames. Then, we leverage the proposed scale-adaptive multi-view temporal interaction (ViewTemp) and geometric-temporal point aggregation (PointTemp) to facilitate temporal interactions for 2D multi-view features and 3D point clouds. The spatial-temporal features derived from ViewTemp are then fed into the transformer decoder with relation-aware queries to predict 2D subject-object coordinates. The workflow provides efficient bi-modal feature representations in relation-aware locations. Furthermore, we reformat triplets <subject, predicate, object> from the graph into a proper template for input to the frozen LLaVA-Med [14] to distill text embeddings, which can enhance the understanding of the intraoperative activities by alignment with bi-modal features. After that, the relation-aware feature unification is introduced to aggregate 2D and 3D features into subject-object representations for paired relationship prediction. Finally, we interact relation-aware tri-modal representations (*i.e.*, images, point clouds, and language) with spatial-temporal features via cross-attention, embracing a learnable classifier initialized by predicate text embeddings from LLaVA-Med, to generate scene graphs.

2.1 Multi-view Image and Point Cloud Temporal Dynamics

ViewTemp. Temporal dynamics of 2D frames can facilitate the consistency of spatial features at consecutive time points. To this end, we introduce the scale-adaptive multi-view temporal interaction (ViewTemp) for spatial-temporal integration of 2D image features across various viewpoints. Given six viewpoints in the 4D-OR dataset, we select view#1 to offer a broader field of view and view#6 to allow for finer-grained observation of the surgery. ViewTemp operates varying-scale convolution kernels depending on the granularity of each viewpoint, which explores finer features in view#1 and interacts with macro-level cues in view#6. Specifically, we predefine a view-specific set of N various-size kernels as $\mathbb{M}_k \coloneqq \left\{m_i^k\right\}_{i=1}^N$ for the k-th viewpoint, where m_i^k is determined according to the feature granularity. Given the feature sequence from ResNet-50 [12] encoder for the k-th viewpoint, we extract a group of multi-scale features $\mathbb{S}_t^k \coloneqq \left\{s_i^{kt}\right\}_{i=1}^N$ for each feature map f_t^k at time point t by applying various-size kernels:

$$s_i^{kt} = f_t^k * m_i^k, \tag{1}$$

where i represents the i-th convolution kernel in $\mathbb{M}_k$ and $*$ denotes the convolution operation. Afterward, for each feature scale, the image features s_i^{kt} at different time points t are interacted through the local cross-attention [23] for hierarchical feature extraction. We denote a set of the same-scale features from consecutive video frames as $\mathbb{U}_i^{kt} \coloneqq \left\{s_i^{kj}\right\}_{j=t-l+1}^{j=t}$, where l is the number of consecutive frames in one batch. Here, s_i^{kt} is used as queries and $\mathbb{U}_i^{kt}$ serves as keys and values for local interaction:

$$Q = \mathcal{W}_q(s_i^{kt}), \quad K = \mathcal{W}_k(\mathcal{C}(\mathbb{U}_i^{kt})), \quad V = \mathcal{W}_v(\mathcal{C}(\mathbb{U}_i^{kt})), \tag{2}$$

$$d_i^k = CA(Q, K, V), \tag{3}$$

where $\mathcal{W}(\cdot)$ represents the fully connected layer and $\mathcal{C}(\cdot)$ is the concatenation operation. $CA(\cdot)$ means the local cross-attention operation, where only feature points with consistent locations are interacted for temporal integration. d_i^k denotes the temporal-fused features of the i-th scale. Then, we merge the features of different scales in each viewpoint by MLP to attain the multi-scale temporal representation. The proposed ViewTemp adaptively aggregates multi-level granularity spatial-temporal features across various viewpoints.

PointTemp. The point cloud provides a detailed 3D perspective of dynamic interactions, which is essential for recognizing subject-object relations and describing global geometric dynamics in ORs. Thus, we design the geometric-temporal point aggregation (PointTemp) to explore temporal dynamics in 3D point clouds for more consistent alignment with 2D multi-view spatial-temporal features. As illustrated in Fig. 1, PointTemp consists of point 4D convolution layers and a global temporal aggregation [7]. The 4D convolution layer perceives the local geometric-temporal structures within the point cloud, while the global temporal aggregation layer performs temporal interactions from various time points via self-attention [23] to capture geometric-temporal correlations.

2.2 Relation-aware Feature Unification

2D image and 3D point cloud feature representations are crucial for semantic modeling in ORs [28]. To facilitate feature unification, we employ 2D spatial coordinates to integrate relation-aware bi-modal features. In detail, spatial-temporal features from ViewTemp are passed through a transformer decoder to predict the locations of all entities. Then, we acquire 2D relation-aware features in the target locations by RoIAlign [11]. Meanwhile, we project the 2D coordinates into the 3D point cloud domain by applying the camera projection matrix for spatial translation. For each subject-object pair, we extract the relation-aware point cloud features from geometric-temporal features through a max-pooling operation among neighborhood points. Finally, the relation-aware bi-modal features are aggregated into subject-object representations for paired relationship prediction. The proposed relation-aware feature unification utilizes 2D images and 3D point clouds accompanied by temporal dynamics for efficient bi-modal feature integration to enrich the representation of subject-object relationships.

2.3 Knowledge Transfer from LLaVA-Med

Since different phases in surgery pose varying durations, we consider exploiting the rich semantic representations and the broad understanding of language and context from LLMs to alleviate the class imbalance in ORs, *i.e.*, frequent classes versus rare classes. However, the knowledge gap between the open vocabulary and medical scenarios poses challenges for the application of LLMs in the OR-SGG task. We alleviate this issue with the novel biomedical LLM, LLaVA-Med [14], in two steps, including knowledge distillation for each triplet in the graph and knowledge transfer for each predicate. As shown in Fig. 1, we first regularize the relation-aware bi-modal features using the corresponding text embeddings from LLaVA-Med. Before training, we prompt all triplets in the graph with the template 'A scene of a/an [subject][predicate] a/an [object]' and store the average of the frozen LLaVA-Med token features from the last layer. Then, the bi-modal features are aligned with the transferred token features using L1 loss. After aggregating the bi-modal features with textual features via relation-aware feature unification, the unified tri-modal features are passed through a cross-attention operation followed by a learnable classifier. Notably, the classifier is initialized with text embeddings generated by LLaVA-Med from the template 'A scene of [predicate]' to predict relationships. Considering the fixed triplet combinations, we extract token features from the pre-trained LLaVA-Med offline before training. Additionally, the alignment between bi-modal features and text embeddings is performed during training. Overall, we efficiently transfer prior knowledge from medical LLMs into our model by imposing additional constraints on relation-aware features, alleviating the class imbalance in ORs.

Accordingly, the total loss function of our model is expressed as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{coord} + \lambda_c \mathcal{L}_{cls} + \lambda_t \mathcal{L}_{text}, \tag{4}$$

where λ_c and λ_t are weighting factors. The coordinate loss $\mathcal{L}_{coord}$ includes L1 loss and GIoU loss [20] for coordinate prediction and focal loss [16] for predicted subject/object scores. The relation classification loss $\mathcal{L}_{cls}$ can be computed as

$$L_{cls} = \frac{1}{\sum_{i=1}^{N} \sum_{z=1}^{Z} y_{iz}} \sum_{i=1}^{N} \sum_{z=1}^{Z} F(\hat{y}_{iz}, y_{iz}), \tag{5}$$

where $F(\cdot)$ denotes focal loss and Z is the number of relation classes. N denotes the number of predicted subject-object pairs. $y_{iz} \in \{0, 1\}$ indicates whether the i-th prediction contains the z-th relation class. $\hat{y}_{iz}$ is the predicted score of the z-th relation class. $\mathcal{L}_{text}$ is the L1 loss between bi-modal features and text embeddings from LLaVA-Med for alignment.

3 Experiments

3.1 Dataset and Evaluation Metrics

We evaluate the proposed framework on the public benchmark 4D-OR [28]. The dataset consists of 10 knee replacement surgery videos, involving 6,734 frames captured by RGB-D cameras. Each frame includes RGB images with a size of 1536×2048 from six different viewpoints and 3D point clouds computed from the depth map. We follow the same data partitioning as in [28], where 6 videos (4,024 scenes) are for training, 2 videos (1,332 scenes) for validation, and 2 videos (1,378 scenes) for online testing. The annotations contain 12 subject/object classes and 14 relation classes in ORs. For evaluation, we use precision, recall, and F1 metrics in line with previous methods [28,27].

3.2 Implementation Details

We train the proposed model for 60 epochs with an AdamW optimizer. The batch size is 4, and the initial learning rate is 5e-5 with a weight decay of 0.0001. The loss weight λ_c and λ_t are set to 1 and 0.1, respectively. We adopt ResNet-50 [12] pre-trained on ImageNet [4] as the image encoder. In our experiments, two typical viewpoints (view#1 and view#6) are selected as 2D image inputs to our model, and the number of consecutive frames l is 3. During training, we apply color jitter, random resize, and random horizontal flip to augment 2D images. According to the feature map size, we employ kernel size combination $\{1, 3, 5, 7\}$ for view#1, and $\{3, 5, 7, 9\}$ for view#6 in ViewTemp.

Table 1: Detailed comparisons with existing OR-SGG models on 4D-OR test set.

Methods	Params	Metrics	Assist	Cement	Clean	CloseTo	Cut	Drill	Hammer	Hold	LyingOn	Operate	Prepare	Saw	Suture	Touch	Avg
4D-OR [28]	84.8M	Precision	0.42	0.78	0.53	**0.97**	0.49	0.87	0.71	0.55	**1.00**	0.55	0.62	0.69	0.60	0.41	0.68
		Recall	**0.93**	0.78	0.63	0.89	0.49	**1.00**	0.89	**0.95**	0.99	**0.99**	**0.91**	0.91	**1.00**	0.69	0.87
		F1	0.58	0.78	0.57	0.93	0.49	0.93	0.79	0.70	0.99	0.71	0.74	0.79	0.75	0.51	0.75
LABRAD-OR [27]	85.8M	Precision	0.60	0.96	0.86	0.96	**0.91**	**1.00**	0.93	0.71	**1.00**	0.85	0.77	0.78	**1.00**	0.71	0.87
		Recall	0.86	0.93	0.72	**0.94**	0.68	0.94	**0.95**	**0.95**	**1.00**	**0.99**	**0.91**	0.93	**1.00**	**0.72**	**0.90**
		F1	0.71	0.94	0.78	**0.95**	0.78	**0.97**	0.94	0.81	**1.00**	0.91	0.84	0.85	**1.00**	0.71	0.88
TriTemp-OR	**67.1M**	Precision	**0.74**	**1.00**	**0.92**	**0.97**	0.86	0.98	**0.96**	**0.84**	**1.00**	**0.93**	**0.86**	**0.95**	0.96	**0.82**	**0.91**
		Recall	0.79	**0.95**	**0.88**	**0.94**	**0.85**	0.94	**0.95**	0.82	**1.00**	0.90	0.88	**0.99**	0.95	**0.72**	**0.90**
		F1	**0.76**	**0.97**	**0.89**	**0.95**	**0.85**	0.96	**0.95**	**0.83**	**1.00**	**0.92**	**0.87**	**0.97**	0.95	**0.77**	**0.90**

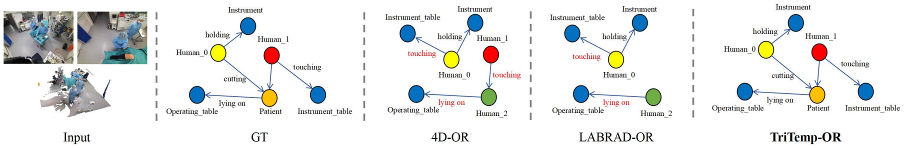


Fig. 2: Qualitative results of our TriTemp-OR and existing OR-SGG methods on the 4D-OR validation set. Erroneous predicted relations are shown in red.

3.3 Comparison with State-of-the-art Methods

Table 1 compares the experimental results of our method with two existing OR-SGG models, 4D-OR [28] and LABRAD-OR [27]. Both of them use multi-view images together with point cloud samples as input. Furthermore, LABRAD-OR and our TriTemp-OR also utilize temporal information to enhance the predictions in consecutive frames. As observed in Table 1, TriTemp-OR shows promising performance with only two views as input, especially in the average precision and F1 score of 0.91 and 0.90, respectively. Fig. 2 exhibits the scene graphs generated by existing OR-SGG methods on the 4D-OR validation set. Benefiting from the transferred knowledge of LLaVA-Med and temporal interaction, our model also mitigates the class-imbalance problem. Even compared to LABRAD-OR, our model demonstrates a remarkable advantage in recognizing less-frequent relations such as 'Saw' and 'Clean', with improvements of 12% and 11%.

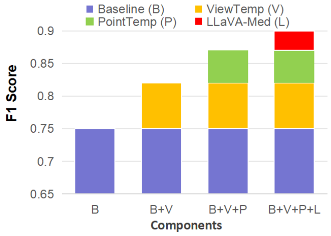
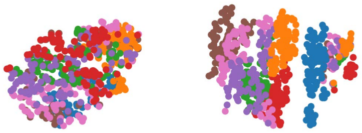


Fig. 3: Ablation study for the contribution of each component, where baseline uses only encoded image features without 2D and 3D temporal information.

View#1/View#6	P ↑	R ↑	F1 ↑
$\{1,3,5,7\}$ / $\{1,3,5,7\}$	0.89	0.88	0.88
$\{1,3,5,7\}$ / $\{3,5,7,9\}$	**0.91**	**0.90**	**0.90**
$\{1,3,5,7\}$ / $\{5,7,9,9\}$	0.90	0.87	0.88
$\{3,5,7,9\}$ / $\{3,5,7,9\}$	0.88	0.86	0.87
$\{5,7,9,9\}$ / $\{3,5,7,9\}$	0.85	0.82	0.83
$\{3,3,3,3\}$ / $\{3,3,3,3\}$	0.87	0.84	0.85

Table 2: Parameter search on kernel size combinations in ViewTemp. P and R represent Precision and Recall.



Embeddings	P ↑	R ↑	F1 ↑
No Embeddings	0.88	0.86	0.87
LLaVA-Med [14]	**0.91**	**0.90**	**0.90**
CLIP [19]	0.87	0.86	0.86

Fig. 4: **Left:** t-SNE of CLIP latent space. **Middle:** t-SNE of LLaVA-Med latent space. **Right:** Ablations on the impact of the knowledge transfer from different pre-trained LLMs. P and R represent Precision and Recall.

3.4 Ablation Study

Ablation Study for Each Component. We study the contribution of each component to the model performance in Fig. 3. Results show that all components, including ViewTemp, PointTemp, as well as the knowledge transfer from LLaVA-Med, contribute to our final result. Among these components, ViewTemp has the most remarkable impact, bringing in an increase of 7% in average F1 score. This shows that multi-view temporal information is pivotal for the spatial modeling of the OR environment. PointTemp comes next, reflecting the importance of dynamic information provided by temporal 3D point cloud features in surgical action recognition. The LLaVA-Med text embedding also improves the prediction, which proves that the transferred knowledge from medical LLMs enhances the comprehension of intraoperative actions.

t-SNE of CLIP and LLaVA-Med. Fig. 4 shows the 3D t-SNE results of different sentences in the latent space from CLIP [19] and LLaVA-Med [14]. The points represent the text embeddings of the prompted sentences. Sentences with the same intraoperative actions are denoted with the same color. As illustrated in the figure, points of different colors from CLIP embeddings intermingle indiscriminately. In contrast, within the LLaVA-Med embedding space, points sharing the same color coalesce into clusters, demonstrating that LLaVA-Med has better discrimination for different relations in surgery compared to CLIP.

Knowledge Transfer from LLMs. In our experiments, we use LLaVA-Med [14] pre-trained on large-scale biomedical data to perform knowledge transfer to enhance the context awareness of our model and alleviate the relation class imbalance. To validate the effectiveness of pre-trained LLMs to the OR-SGG task, we display the results of TriTemp-OR with and without text embeddings from different LLMs in Fig. 4. Notably, the text embeddings from the text encoder of CLIP [19] bring no benefit to our model, which may be caused by the prior knowledge gap between the open vocabulary and the biomedical semantics.

Parameter Search on Different Kernel Sizes in ViewTemp. Table 2 compares the effect of kernel size combinations for different viewpoints in ViewTemp. When increasing the kernel sizes from $\{1, 3, 5, 7\}$ to $\{3, 5, 7, 9\}$, view#6 has a positive impact on the accuracy, while view#1 does the opposite. It is attributed to the broader perspective in view#1 while view#6 provides a more focused viewpoint. Our model achieves the optimal results with the combination of $\{1, 3, 5, 7\}$

and $\{3, 5, 7, 9\}$. Besides, using the scale-adaptive partition obtains better results than the single-scale kernel sizes $\{3, 3, 3, 3\}$ across different perspectives.

4 Conclusion

This paper presents TriTemp-OR, an end-to-end tri-modal framework for the OR-SGG task. In one fold, we introduce ViewTemp and PointTemp to capture the temporal dynamics of multi-view images and point clouds. The bi-modal features are subsequently integrated with relation-aware feature unification to predict subject-object relations. On the other fold, we distill the semantic knowledge from the biomedical LLM, obtaining the tri-modal representation and further alleviating the class-imbalance problem. Experimental results illustrate that our method consistently outperforms the state-of-the-art methods, endorsing a potential application for improving surgical procedure efficiency in ORs.

Acknowledgments. The work described in this paper was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.: T45-401/22-N), in part by a grant from Hong Kong Innovation and Technology Fund (Project No.: GHP/060/21GD), and in part by the DIREC project EXPLAIN-ME (9142-00001B).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alayrac, J.B., Donahue, J., Luc, P., et al, A.M.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
2. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: generation and application. IEEE TPAMI **45**(1), 1–26 (2021)
3. Cong, Y., Yang, M.Y., Rosenhahn, B.: Reltr: relation transformer for scene graph generation. IEEE TPAMI **45**(9), 11169–11183 (2023)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE CVPR (2009)
5. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: language-driven open-vocabulary 3d scene understanding. In: IEEE CVPR (2023)
6. Ege Özsoy, Czempiel, T., Evin Pınar Örnek, Eck, U., Tombari, F., Navab, N.: Holistic or domain modeling: a semantic scene graph approach. IJCARS (2023)
7. Fan, H., Yang, Y., Kankanhalli, M.: Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In: IEEE CVPR (2021)
8. Gao, K., Chen, L., Zhang, H., Xiao, J., Sun, Q.: Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In: ICLR (2023)
9. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: MICCAI (2021)

10. Green, O.L., Rankine, L.J., Cai, B., Curcuru, A., Kashani, R., Rodriguez, V., Li, H.H., Parikh, P.J., Robinson, C.G., Olsen, J.R., et al.: First clinical implementation of real-time, real anatomy tracking and radiation beam control. Med. Phys. **45**, 3728–3740 (2018)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR (2015)
13. He, T., Gao, L., Song, J., Li, Y.F.: Towards open-vocabulary scene graph generation with prompt-based finetuning. In: ECCV (2022)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2023)
15. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: Gen-vlkt: simplify association and enhance interaction understanding for hoi detection. In: IEEE CVPR (2022)
16. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE ICCV (2017)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
18. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: IEEE ICCV (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
20. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: IEEE CVPR (2019)
21. Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N.: Fun-sis: A fully unsupervised approach for surgical instrument segmentation. Med. Image Anal. **85**, 102751 (2023)
22. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**, 172–180 (2023)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
24. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: IEEE CVPR (2020)
25. Wald, J., Navab, N., Tombari, F.: Learning 3d semantic scene graphs with instance embeddings. IJCV **130**(3), 630–651 (2022)
26. Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., Sheng, L.: Vl-sat: visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In: IEEE CVPR (2023)
27. Özsoy, E., Czempiel, T., Holm, F., Pellegrini, C., Navab, N.: Labrad-or: lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In: MICCAI (2023)
28. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: semantic scene graphs for or domain modeling. In: MICCAI (2022)