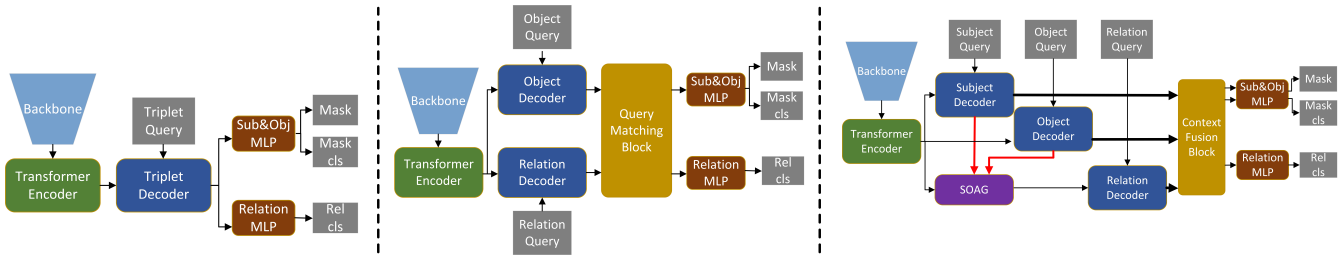
Contextual Associated Triplet Queries for Panoptic Scene Graph Generation

Jingbin Xu
The University of Electro-Communications, Tokyo, Japan
xu-j@mm.inf.uec.ac.jp

Junwen Chen
The University of Electro-Communications, Tokyo, Japan
chen-j@mm.inf.uec.ac.jp

Keiji Yanai
The University of Electro-Communications, Tokyo, Japan
yanai@cs.uec.ac.jp



(a) PSGTR (b) PSGFormer (c) CATQ

Figure 1: The comparison of PSG task baseline methods (a), (b), and our method (c). In CATQ we introduce the SOAG module and Context Fusion block. The red line represents the attention map of cross-attention in the decoder.

ABSTRACT

The Panoptic Scene Graph generation (PSG) task aims to extract the triplets composed of subject, object, and relation based on panoptic segmentation. For one-stage methods, PSGTR predicts the subject, object, and relation by one query. However, the integrated query is too implicit to simultaneously ascertain pairs of instances and relations. In PSGFormer, it learns instances and relation queries separately and establishes matches between subject-relation and object-relation pairs by employing the relation as an index. Nevertheless, this method could potentially impede the accurate determination of the optimal match. To address the aforementioned issues, we propose a new one-stage method, **C**ontextual **A**ssociated **T**riplet **Q**ueries (CATQ), which employs three branches to decode subject, object, and relation features separately. Additionally, we leverage instance information to guide the relation decoding process. Furthermore, we introduce the triplet context fusion block to enable the extraction of more comprehensive instance pairs and triplet relations. Our proposed method achieves 34.8 Recall@20 and 20.9 mRecall@20 respectively and surpasses the state-of-the-art baseline method by 22.5% and 26.0% with half of the training session.

MMAsia '23, December 06–08, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12…$15.00
https://doi.org/10.1145/3595916.3626745

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Panoptic Scene Graph Generation, transformer, guide, relation

ACM Reference Format:
Jingbin Xu, Junwen Chen, and Keiji Yanai. 2023. Contextual Associated Triplet Queries for Panoptic Scene Graph Generation. In *ACM Multimedia Asia 2023 (MMAsia '23), December 06–08, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3595916.3626745

1 INTRODUCTION

The Scene Graph Generation (SGG) task [5] aims to generate graph-structured representations by localizing objects and their pairwise relationships. The subjects and objects are represented using bounding boxes. However, the utilization of bounding box representation suffers from two issues. Firstly, the coarse positioning of bounding boxes can introduce noise, especially when boxes of different categories are interleaved. Secondly, bounding boxes do not cover stuff regions such as the background. To address these issues, a novel variant of the SGG task is proposed, namely Panoptic Scene Graph generation (PSG) task. For the PSG task, a new dataset, OpenPSG is introduced. Different from the SGG task, the annotation is a pixel-level segmentation mask instead of the rigid bounding box.

Similar to the SGG task, the PSG task can be categorized into two paradigms: one-stage and two-stage methods. Specifically, the two-stage methods [10, 13, 15, 17] are derived from the SGG task and are adapted to the PSG task. With the advancements of transformer

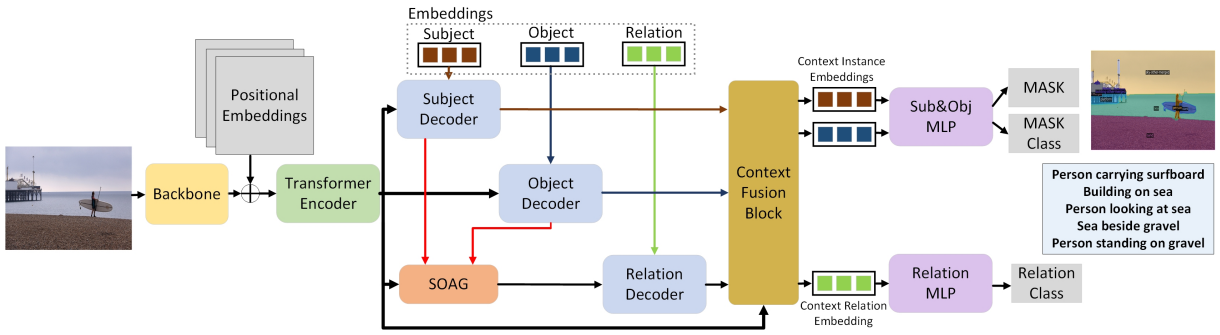


Figure 2: This figure illustrates the overall architecture of the proposed method, CATQ.

encoder-decoder architecture [14] in object detection, transformer-based detection method DETR [2] is adapted to two kinds of one-stage PSG task baselines PSGTR [16] and PSGFormer [16].

As shown in Figure 1(a), PSGTR utilizes highly integrated triples, which are represented by a single type query. Predictions for both instances and relations are made through this query. However, this implicit approach incurs instability in the learning process of each query. For example, the same query may correspond to different triples at each iteration, which may lead to slow convergence. In Figure 1(b), PSGFormer extends explicit relation modeling with a query-matching mechanism. It employs a parallel method to independently learn instance and relation queries. Then, an instance-relation matching block is used to pair subjects and objects with relations and then computes the cosine similarity to find the optimal results. PSGFormer simply matches subject-relation and object-relation pairs using the relation as an index, thus potentially hindering the identification of the best match. In this paper, we focus on improving the architecture of baseline one-stage methods to achieve higher accuracy and faster training accuracy.

For a more precise division of labor and to increase the predictive influence of objects on relationships, we propose a new one-stage method CATQ, which is illustrated in Figure 1(c). In summary, the contributions of our work are threefold.

- We propose a new one-stage framework for PSG, CATQ, which leverages contextually associated triplet quires to better extract the semantic information in the context.
- We introduce a novel attention operation module to enhance the global context features with the guide from attention maps of multiple branches.
- With our proposed triplet query architecture and global feature enhancement module, our method, CATQ outperforms baseline methods with large margins and half of the training schedule.

2 RELATED WORK

One-stage Scene Graph Generation. SGTR [8] stands as the pioneering one-stage Scene Graph Generation method that adopts a transformer architecture. Analogous to PSGFormer, SGTR also disentangles the decoding of instances and relations. However, in contrast to PSGFormer, SGTR places its emphasis on relationship prediction by leveraging ample object context information. To achieve this, SGTR introduces two types of object decoders. The first decoder solely relies on image features for decoding, while the second decoder employs iterative decoding in conjunction with relation prediction. This approach ensures that relations are predicted within a rich object context. RelTR [4] introduces a methodology where the embedding of triples is coupled with the subject and object embeddings. These coupled embeddings are then decoded using a combination of coupled and decoupled attention modules. Finally, the triplets are directly predicted utilizing the triplet embedding. This approach introduces object pairs contextual information into the triplet query during the pairing and decoupling procedure.

In the SGTR and RelTR, decoding instances and relations are executed distinctly. Inspired by the multi-branch design of SGG approaches, our method adopts a three-decoder architecture with triplet queries to better extract features of specific targets.

3 METHOD

The architecture of CATQ is shown in Figure 2, we adopt a separating instance decoding strategy to enable more expressive decoding for the subject and object instance in a triplet. This separation allows the subject and object to focus on distinct instances as much as possible. To leverage the contextual information from instances during relation prediction, we introduce a Subject-Object Attention Guide(SOAG) module, which extracts the visual information from entities and utilizes it to guide the learning process of relation queries. Lastly, we incorporate a context fusion module to enrich the completely separated triple queries with additional contextual information.

3.1 Image Encoding

Following the DETR [2] encoding process, we employ a transformer encoder with a CNN backbone as the feature extractor to obtain global visual features from the input images. Given an input image $X \in \mathbb{R}^{H \times W \times 3}$, the CNN backbone is employed to extract an image feature map $f \in \mathbb{R}^{K_f \times \frac{H}{32} \times \frac{W}{32}}$. This feature map is subsequently tokenized by flattening and feed into the transformer encoder to produce the global visual feature $V \in \mathbb{R}^{T \times D}$, where T and D are the numbers of image tokens and channel dimension.

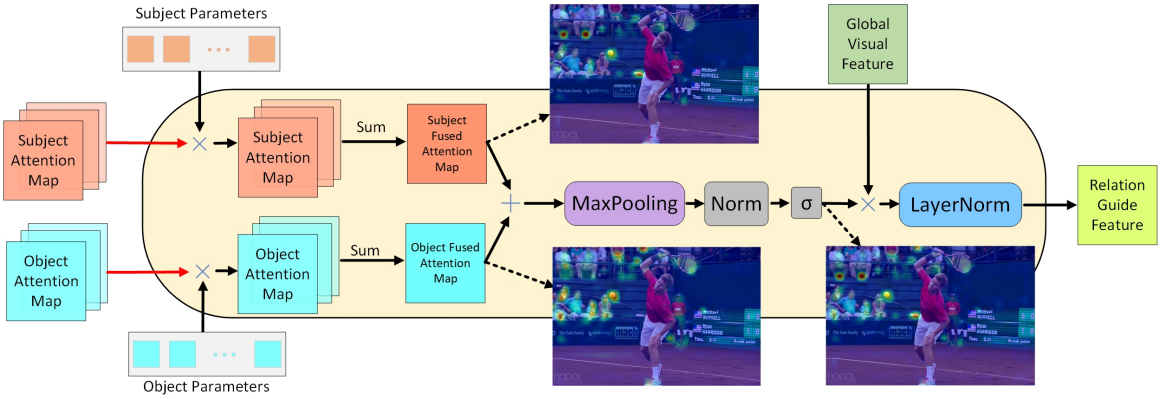


Figure 3: This figure illustrates the architecture of the SOAG module. The subject and object parameters are learnable weights and are used to balance the attention of different attention maps.

3.2 Triple Decoder

We divide the decoding process into three branches corresponding to the decoding of each element of the triplet. As for the inputs, we utilize the global visual features V obtained from the encoder as the source input for both the subject and object branches. Additionally, we employ completely independent learnable queries $Q \in \mathbb{R}^{N_q \times D}$ as the embedding input for both the subject and object branches. The embedding input of the relation branch remains the same as that of the instance branches. The source input for the relation branch is obtained from the SOAG module discussed in Section 3.3. The outputs of the three branches, $Q^S, Q^O, Q^R \in \mathbb{R}^{N_q \times D}$, correspond to the embeddings of the subject, object, and relation, respectively.

3.3 Subject-Object Attention Guide (SOAG)

In CATQ, the source inputs for two instance branches are directly from the feature extractor, while the relation branch uses a refined source input. This disparity arises from our intention to facilitate the learning of relations, wherein we aim to guide the relation acquisition using object-specific information as prior knowledge. This strategy avoids spending attention on unnecessary objects during the relation decoding. Hence, we introduce the Subject-Object Attention Guide (SOAG) module, the architecture of SOAG is in Figure 3. The Guided feature V_g is calculated as:

$$V_g = LN(M_{s,o}^{max} \circ V) \tag{1}$$

$$M_{s,o}^{max} = \sigma(\text{Norm}(\text{MaxPooling}(\{M_s^{max}, M_o^{max}\}))) \tag{2}$$

$$M_k^{max} = \sum_{i=1}^{N_q} M_k^{(i)} \circ W_k^{(i)}, k \in \{s, o\} \tag{3}$$

where the attention maps $M_s, M_o \in \mathbb{R}^{N_q \times T}$ are derived from the cross-attention of the subject and object decoder in the last layer. W_s and W_o are learnable parameters initialized to shape $\mathbb{R}^{N_q}$.

Firstly, we calculate the Hadamard product ($\circ$) on each attention map M_k^i with a learnable parameter W_k^i and then add the N_q attention maps together to get $M_s^{max}, M_o^{max} \in \mathbb{R}^T$. Then, we stack M_s^{max} and M_o^{max} and apply max pooling on the result matrix to obtain the instance maximum attention $M_{s,o}^{max}$. After normalization and activation, we calculate the Hadamard product between the instance maximum attention $M_{s,o}^{max}$ and the global visual feature matrix V. Finally, the result is passed through a LayerNorm operation [1], to obtain the output relation guide feature, denoted as V_g. This output serves as the refined source input for the relation decoder.

3.4 Context Fusion Block

For the architecture of multiple interrelated decoders, recent studies [6], [3], [18] have found that the use of an attention-based embedding fusion method can effectively improve the performance of the final prediction heads. Inspired by [6], we incorporate a context fusion module that effectively adds contextual relationships to each embedding from the triple decoder.

The outputs of the three branches can be considered as unary relations, where each relation only contains its specific information and lacks contextual information. Above this, we can also define the pairwise and ternary relations, enabling contextual relationships between elements within triples. Specifically, the unary embedding is defined as:

$$f_{unary} = \{Q^S, Q^O, Q^R\} \tag{4}$$

the pairwise embedding is calculated as:

$$f_{pairwise} = \{MLP([Q^S; Q^O]), MLP([Q^S; Q^R]), MLP([Q^O; Q^R])\} \tag{5}$$

the ternary embedding is initialized as:

$$f_{ternary} = MLP([Q^S; Q^O; Q^R]) \tag{6}$$

Then the self-attention is applied to $f_{unary}, f_{pairwise} \in \mathbb{R}^{3 \times N_q \times D}$, respectively, resulting in the generation of $\hat{f}_{unary}$ and $\hat{f}_{pairwise}$.

$$\hat{f}_{unary} = \text{SelfAttn}(f_{unary}), \hat{f}_{pairwise} = \text{SelfAttn}(f_{pairwise}) \tag{7}$$

Subsequently, we apply the result of cross-attention between $\hat{f}_{unary}$ and $f_{ternary} \in \mathbb{R}^{N_q \times D}$ through cross-attention with $\hat{f}_{pairwise}$, resulting in the generation of $f_{context}$.

$$f_{context} = \text{CrossAttn}(\text{CrossAttn}(\hat{f}_{unary}, f_{ternary}), \hat{f}_{pairwise}) \tag{8}$$

Method	Backbone	*Recall@20*	*Recall@50*	*Recall@100*	*mRecall@20*	*mRecall@50*	*mRecall@100*	*PQ*
Two-stage								
IMP [15]	ResNet-50	16.5	18.2	18.6	6.5	7.1	7.2	40.2
MOTIFS [17]	ResNet-50	20.0	21.7	22.0	9.1	9.6	9.7	40.2
VCTree [13]	ResNet-50	20.6	22.1	22.5	9.7	10.2	10.2	40.2
GPSNet [10]	ResNet-50	17.8	19.6	20.1	7.0	7.5	7.7	40.2
One-stage								
PSGTR† [16]	ResNet-50	28.4	34.4	36.3	16.6	20.8	22.1	13.9
PSGFormer† [16]	ResNet-50	18.0	19.6	20.1	14.8	17.0	17.6	36.8
CATQ	ResNet-50	**34.8**	**39.7**	**40.3**	**20.9**	**24.9**	**25.2**	35.9

Table 1: A comparison between CATQ and PSG baseline methods on the OpenPSG dataset. † denotes the model fine-tuned with 60 epochs.

Next, we compute the cross-attention between the global visual feature V and $f_{context}$.

$$\hat{f}_{context} = \text{CrossAttn}(V, f_{context}) \tag{9}$$

After obtaining the $\hat{f}_{context}$ with visual information, we add this context information to the output of each branch like Eq.10.

$$\hat{f}^{k}_{context} = MLP(f_{context}, Q^{k}), k \in \{S, O, R\} \tag{10}$$

3.5 Training and Inference

During the training process, we extend the Hungarian matching technique used in DETR [2] for the triplet matching. The matching cost between the ground-truth triplets and the predicted triplets is determined by considering both segment matching, denoted as $\mathcal{H}^{seg}$, and class matching, denoted as $\mathcal{H}^{cls}$:

$$C_m(\mathcal{T}_i, \mathcal{G}_i) = \sum_{cls \in \{S,O\}} \mathcal{H}^{cls}_i + \sum_{seg \in \{S,O,R\}} \mathcal{H}^{seg}_i \tag{11}$$

Here, $\mathcal{T}$ represents the predicted triplets, while $\mathcal{G}$ represents the ground truth triplets. After the matching process, the final loss is computed by applying the Focal loss [9] for class labels and the DICE loss [12] for segmentation.

4 EXPERIMENTS

4.1 Dataset

We conducted the experiments on the OpenPSG dataset, which has a total of 48,749 labeled images including 2,177 test images and 46,572 training images with panoptic segmentation and scene graph annotation. The object categories comprise 80 thing classes and 53 stuff classes. The relation categories comprise 56 classes.

4.2 Implementation Details

The weights of the subject and object decoder are initialized from DETR pre-trained on the MS-COCO dataset. The AdamW [11] optimizer is used with a learning rate of 10^{-4} and a weight decay of 10^{-4}. The training is performed on 8 NVIDIA A6000 GPUs, with a batch size of 8 (1 image per GPU).

SOAG	*Recall@20*	*Recall@100*	*mRecall@20*	*mRecall@100*
✓	**34.75**	**40.26**	**20.87**	**25.19**
-	33.44	38.89	20.23	24.20

Table 2: Ablation comparison of the SOAG module.

4.3 Comparison to State-of-the-Arts

Table 1 presents a comprehensive comparison of our method with the current PSG approaches on the OpenPSG dataset. We adopt Recall@K and mean Recall@K (mRecall@K) as our primary evaluation metrics and PQ [7] is used as the evaluation metric for panoptic segmentation. Our method achieves remarkable performance, surpassing the best-performing PSGTR model by 22.5% in terms of Recall@20 and 26.0% in terms of mRecall@20. Additionally, our proposed method achieves convergence in just 30 epochs.

4.4 Ablation Studies

Contributions of SOAG In Table 2, we compare the presence and absence of the SOAG module. It clearly indicates that the inclusion of the SOAG module leads to improved performance compared to its absence, *i.e.*+4% in R@100 and +4% in mR@100. This finding provides evidence supporting the effectiveness of the SOAG module in improving the performance of the relation prediction task.

5 CONCLUSION

In this paper, we propose a novel one-stage approach for Panoptic Scene Graph generation. We separate the triple branches used to explicitly learn the elements of the triplets. To enhance relation learning, we incorporate an attention guidance module called SOAG, which leverages object-specific visual attention to guide relation prediction. Furthermore, we integrate a context fusion module to incorporate contextual information into the three parallel queries that have been comprehensively learned in the subtask. The experimental results presented in this study showcase the remarkable performance of our proposed method on the OpenPSG dataset.

Acknowledgments: This work was supported by JSPS KAKENHI Grant Numbers, 21H05812, 22H00540, 22H00548, and 22K19808.

REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *ArXiv Preprint arXiv:1607.06450* (2016).

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. Springer, 213–229.

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

[4] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2023. RELTR: Relation transformer for scene graph generation. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[5] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[6] Sanghyun Kim, Deunsol Jung, and Minsu Cho. 2023. Relational Context Learning for Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2925–2934.

[7] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6399–6408.

[8] Rongjie Li, Songyang Zhang, and Xuming He. 2022. SGTR: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19486–19496.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[10] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. 2020. GPS-NET: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3746–3753.

[11] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[12] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-NET: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*. Ieee, 565–571.

[13] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6619–6628.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[15] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5410–5419.

[16] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *Proceedings of the European Conference on Computer Vision*. Springer, 178–196.

[17] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. NEURAL MOTIFS: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

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