Focusing on Flexible Masks: A Novel Framework for Panoptic Scene Graph Generation with Relation Constraints

Jiarui Yang[†]

Chuan Wang[‡*]
yngjiarui@iie.ac.cn
wangchuan@iie.ac.cn
SKLOIS, Institute of Information Engineering, CAS
China

Zeming Liu
cstgnlp@gmail.com
School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics
China

Jiahong Wu
wujiahong@pku.edu.cn
Kuai Shou
China

Dongsheng Wang
wangdsh@cupl.edu.cn
The Department of Science and Technology Teaching, China University of Political Science and Law
China

Liang Yang
yangliang@vip.qq.com
School of Artificial Intelligence, Hebei University of Technology
China

Xiaochun Cao
caoxiaochun@mail.sysu.edu.cn
School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-Sen University
China

ABSTRACT

Panoptic Scene Graph Generation (PSG) presents pixel-wise instance detection and localization, leading to comprehensive and precise scene graphs. Current methods employ conventional Scene Graph Generation (SGG) frameworks to solve the PSG problem, neglecting the fundamental differences between bounding boxes and masks, i.e., bounding boxes are allowed overlap but masks are not. Since segmentation from the panoptic head has deviations, non-overlapping masks may not afford complete instance information. Subsequently, in the training phase, incomplete segmented instances may not be well-aligned to annotated ones, causing mismatched relations and insufficient training. During the inference phase, incomplete segmentation leads to incomplete scene graph prediction. To alleviate these problems, we construct a novel two-stage framework for the PSG problem. In the training phase, we design a proposal matching strategy, which replaces deterministic segmentation results with proposals extracted from the off-the-shelf panoptic head for label alignment, thereby ensuring the all-matching of training samples. In the inference phase, we present an innovative concept of employing relation predictions to constrain segmentation and design a relation-constrained segmentation algorithm. By reconstructing the process of generating segmentation results from proposals using predicted relation results, the algorithm recovers more valid instances and predicts more complete scene graphs. The experimental results show overall superiority, effectiveness, and robustness against adversarial attacks. Code is available at https://github.com/flyfaerss/RCpsg.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

Panoptic Scene Graph Generation, Instance Matching, Relation Constraints, Scene Graph Recovery

ACM Reference Format:
Jiarui Yang, Chuan Wang, Zeming Liu, Jiahong Wu, Dongsheng Wang, Liang Yang, and Xiaochun Cao. 2023. Focusing on Flexible Masks: A Novel Framework for Panoptic Scene Graph Generation with Relation Constraints. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612544

1 INTRODUCTION

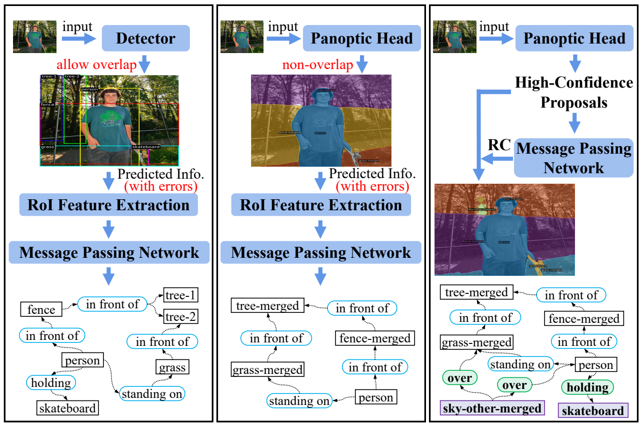
A scene graph (SG) provides a comprehensive representation that encompasses all instances and their interrelationships. By condensing visual information into a graph structure, scene graphs facilitate the effective solution of complex computer vision tasks, such as visual question answering (VQA)[39, 53], image retrieval[15, 43], image captioning[26, 47] and image generation[1, 14].

Scene graph generation (SGG) is often implemented in two stages, the first one conducts the instance detection, and the second one predicts relations between composited object pairs. [36] designs a unified two-stage SGG codebase (Figure 1(a)), which becomes the

[†] School of Cyberspace Security, CAS
[‡] Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security
[*] corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612544



(a) Common SGG framework (b) Same framework as (a) for PSG (c) Our proposed framework for PSG

Figure 1: (a) The pipeline of the common two-stage SGG framework[36]. Non-instance for background causes incomplete scene graph prediction. (b) The pipeline of current two-stage PSG framework. Incomplete segmentation results from the panoptic head provide limited instances and incomplete scene graph prediction. (c) Our proposed framework for PSG. We focus on the flexibility of masks thus generating representative instance proposals. Relation-Constrained (RC) reasoning provides a more complete scene graph (*e.g.* complemented instances and relations (green and purple entities) in the scene graph).

most popular SGG framework. Firstly, the model obtains the detection results through the detector and extracts the initial features of proposals covering instances and relations from prediction results. Secondly, the initial features are passed through a message-passing network[16, 31–33, 37, 49], to get the scene graph prediction, incorporated with a de-biased module[3, 10, 22, 25, 52]. Recent work[46] presents a novel scene graph generation task focusing on panoptic scene understanding, named Panoptic Scene Graph Generation (PSG). It provides more accurate instance locations via pixel-wise form and all-contents scene understanding including backgrounds like *tree-merged* and *sky-other-merged* in Figure 1(b)(c).

To implement scene graph generation under panoptic scenes, existing works[42, 46] accordingly use the common two-stage SGG framework to solve the PSG problem. That is, they first use a panoptic head (*e.g.* Panoptic FPN[17]) for segmentation then extract initial features from the segmentation results, which are fed into a message-passing network for relation prediction. However, this framework is not suitable for the PSG task. As shown in Figure 1 (b), due to *the non-overlap requirement of the masks*, it is difficult for the panoptic head to predict complete segmentation results. In one case, smaller targets can be easily covered by larger targets with higher confidence (*e.g. sky-other-merged*). In another case, hard samples might be discarded due to low-confidence (*e.g. skateboard*). Under this circumstance, in the training phase, incomplete segmentation results impede the model's ability to align with all ground truth instances, causing unaligned instances to be discarded. Consequently, associated relation samples are excluded from the training process, leading to a waste of training samples. In the inference phase, incomplete segmentations prevent the model from predicting relations related to the undetected instances, resulting in incomplete scene graph predictions.

To alleviate these problems caused by the masks, we design a new two-stage PSG framework. In this framework, we design two different pipelines for the training and the inference phases. During the training phase, we propose a proposal matching scheme for label alignment. We extract high-confidence proposals from the existing off-the-shelf panoptic head, and then using them for matching. Since the number of proposals is much larger than the number of ground truth instances, this ensures full utilization of training samples. Additionally, with the help of the complete proposal matching scheme, we can further simplify the initial feature extraction. We directly extract the proposal features instead of the RoI features, which not only reduces a large amount of computational overhead but also alleviates the cumulative errors of the two-stage model, making the model training more effective and reliable. In the inference phase, we creatively propose a realization of using relation predictions to constrain segmentation results, thereby obtaining more complete segmentation results and scene graph predictions. First, we perform pairwise relation prediction for all high-confidence proposals. Then, we calculate the mask coverage priority based on the highest relation confidence for each proposal, and forcibly recover proposals with high-confidence relation. This relation-priority masking method allows the model to focus more on instances with interaction information, making it more capable of recovering reliable small targets or low-confidence proposals, and thus obtaining more complete scene graph predictions.

The main contributions are summarized as follows:

- We analyze the incompatibility of the common two-stage SGG framework on solving PSG problem, indicating problems of the waste of training samples and incomplete scene graph prediction caused by the non-overlap masks.
- We proposed a method of using the off-the-shelf panoptic head's proposals instead of deterministic segmentation results for label alignment and RoI feature extraction. It effectively utilizes the representative pre-trained features of dedicated models, meanwhile ensuring the full utilization of the training samples.
- We make the first step in considering the guidance role of the relations for instance understanding. A novel algorithm is proposed to simultaneously achieve panoptic segmentation and scene graph prediction under relation constraints.
- The proposed framework achieves outstanding superiority on the PSG dataset.

2 RELATED WORK

Scene Graph Generation. SGG aims to use a comprehensive graph structure to represent a scene image. Early works pay more attention to exploring different networks, such as GNN[27, 34], CRF[7, 8], and RNN/LSTM[37, 49], to model the message passing mechanisms between the entities and predicates. Follow-up works focus on extracting more powerful global contextual information. [35] constraints the rationality of the distribution of the scene graph structure by introducing a global energy value. [24, 31, 32] are to deeply mine the attributes of nodes on a graph-based network and fully utilize their contextual information to achieve better

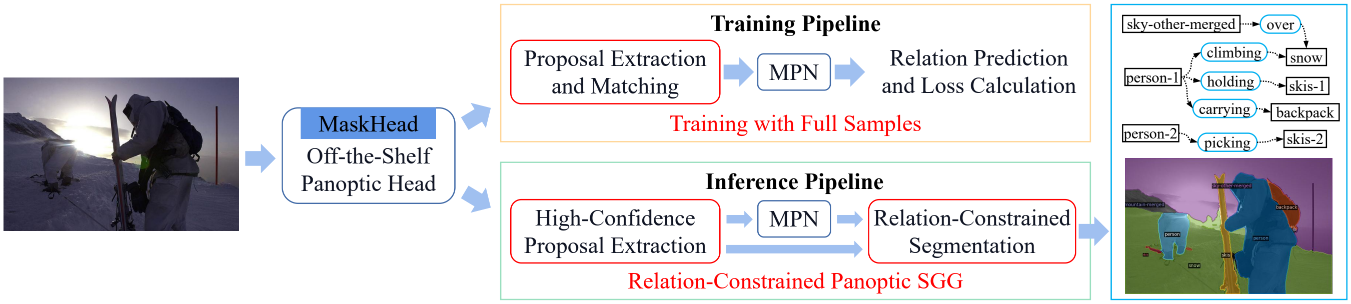
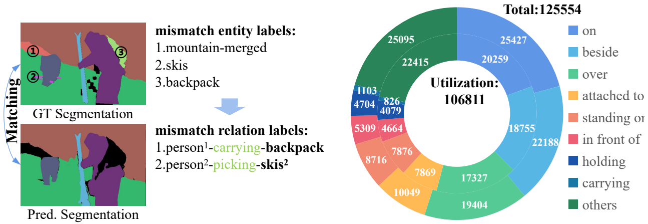

MaskHead
Off-the-Shelf Panoptic Head
Training Pipeline
Proposal Extraction and Matching
MPN
Relation Prediction and Loss Calculation
Training with Full Samples
Inference Pipeline
High-Confidence Proposal Extraction
MPN
Relation-Constrained Segmentation
Relation-Constrained Panoptic SGG
sky-other-merged
over
snow
climbing
person-1
holding
skis-1
carrying
backpack
person-2
picking
skis-2

Figure 2: Overview of the proposed framework. We design different pipelines for training and inference. 1) In the training, we show a specific Proposal Extraction and Matching for PSG to extract proposals from the panoptic head for label alignment, so as to fully utilize the training samples; 2) In the inference, we reconstruct the process of segmentation generation from proposals with relation constraints, predicting more complete scene graphs by recovering more valid instances.



(a) An example of cascade mismatching caused by direct matching scheme.

(b) Number of relation training samples with Mask2Former. Outer: total; Inner: utilization.

Figure 3: Mismatching caused by directly using panoptic head’s deterministic segmentation results.

feature representations. Recently, more research has shifted the attention to the severe long-tail problem of SGG datasets, like Visual Genome[19], Open Image[21] and GQA[13]. [36] proposes the first solution for unbiased SGG model prediction. Most works mainly utilize re-sample[9, 12] or re-weight[45, 48], and their variants[10, 38] to alleviate biased prediction. Yang et al.[46] introduces a novel SGG paradigm grounded in panoptic segmentation, named PSG. The enhanced accuracy in localization and comprehensive instance detection lead to a more complete scene graph. Nevertheless, the conventional SGG framework is not directly applicable to PSG. Although Wang et al.[42] provides a tailored solution that adeptly leverages the feature extraction potential of the off-the-shelf panoptic head, they have yet to address the problem of non-overlap masks. We consider the challenges posed by these masks and design a new PSG framework, which significantly improves model performance.

Panoptic Segmentation. The panoptic segmentation unifies semantic segmentation and instance segmentation tasks for holistic scene understanding. Some works treat this problem as a joint task that combines the best of specialized semantic and instance segmentation architectures into a single framework[4, 17, 44]. However, these methods will bring unnecessary model complexity by solving surrogate sub-tasks to achieve the target task. Recently, researchers make effort on a unified panoptic segmentation framework. Several works[5, 6, 41, 51] use query to represent thing and stuff through universal architectures based on Transformer[40] like DETR[2].

3 APPROACH

3.1 Problem Setting and Overview

Problem Setting. Given an image I, the task of panoptic scene graph generation is to parse I into a scene graph $\mathcal{G} = \{\mathcal{E}_{sub}, \mathcal{P}, \mathcal{E}_{obj}\}$, where $\mathcal{E}_{sub}$ and $\mathcal{E}_{obj}$ denote the set of subject and object entities and $\mathcal{P}$ represents the set of predicates. Typically, we can use two-stage methods to solve the PSG problem: first conduct panoptic segmentation, then detect relations between instances. In our framework, the goal of panoptic segmentation is to extract instance proposals from the off-the-shelf panoptic head, *e.g.* Mask2Former[5]. Each instance proposal provide four key information: *mask confidence* (m^c), *mask prediction* (m), *label prediction* (l) and *proposal feature* ($\mathbf{q}$). Then, in the second stage, all predicted instance proposals should be matched to generate relation candidates and align corresponding relation annotations for relation prediction. Finally, combining proposal segmentation results and relation prediction results, we obtain a scene graph prediction.

Overview. In this paper, we design a new panoptic scene graph generation framework to adapt to the flexibility and the non-overlap of the masks, as shown in Figure 2. To this end, we divide the entire framework into two parts: the training pipeline and the inference pipeline, which have a little differences. In the training pipeline, an image first goes through the off-the-shelf panoptic head, where instance proposals are extracted from the *MaskHead*. Afterward, these instance proposals are matched with ground truth instances for label alignment. The matched instances are fed together into a message-passing network (MPN), resulting in fine-tuned instance features and relation features. Finally, classification results are obtained through a classifier, followed by loss calculation and training. In the inference pipeline, we obtain instance proposals using the same operations as the training process. Then we extract the high-confidence parts and input them into the message-passing network to obtain prediction for instances and relations. Finally, we use the predicted relation results to constrain the process of generating panoptic segmentation results from instance proposals, ultimately producing a more complete scene graph.

3.2 Difference Analysis between SGG and PSG

We start by analyzing the differences between SGG and PSG tasks to understand the reasons behind our framework design. Obviously, compared to the traditional SGG task, PSG uses *masks* to represent the regions of each instance. This difference in region representation has inspired the design of the framework.

Instance Matching in the Training Pipeline. Direct matching is mostly common used in the two-stage SGG framework, which

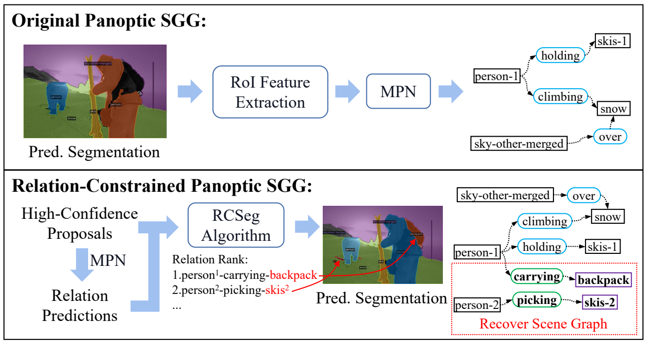


Figure 4: Differences between panoptic SGG schemes. The relation-constrained method obtains more fine-grained segmentations and complete scene graph predictions.

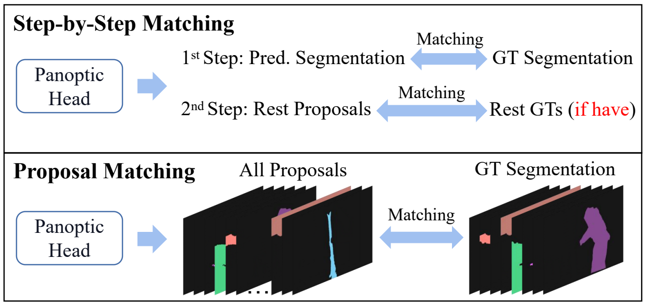


Figure 5: Diagram of two instance all-matching strategies.

directly employs the deterministic prediction results acquired from the detector to align the ground truth labels. However, this matching scheme is unsuitable for PSG. Owing to the non-overlap requirement of the masks, the segmentation results predicted by the panoptic head might be incomplete, resulting in a greater number of ground truth instances than predicted ones. In this case, not only can the unmatched ground truth instances not participate in training, but the relation annotations associated with these instances also cannot be involved in the training process, resulting in the waste of training samples. As shown in Figure 3 (left), the panoptic head (Mask2Former, the same below) can't predict *mountain-merged*, *skis*, and *backpack*, where *skis* and *backpack* generate relation annotations through interactions with other instances, *i.e.*, *person-picking-skis* and *person-carrying-backpack*. In this way, during the training stage, these relation samples cannot be matched and thus are discarded. We statistic the sample utilization of direct matching under the high-accuracy panoptic head, Mask2Former in Figure 3 (right), we still find that a large proportion of relation samples are discarded in training, which hurt the model performance.

Scene Graph Prediction in the Inference Pipeline. In the two-stage SGG framework, compared to the training pipeline, the inference process simply removes the instance matching process. The message-passing network just predicts pairwise relations for all instance proposals generated by the detector. In the PSG task, due to the non-overlapping masks, this relation prediction mode is severely limited by the panoptic segmentation results. As shown in Figure 4 (top), due to the incomplete segmentation results obtained by the panoptic head, instances such as *skis* and *backpack* are not included in the input of the message-passing network, making it impossible to predict relations like *person-picking-skis* and *person-carrying-backpack*. Since we cannot perform ground truth matching during the inference phase, this prediction mode will inevitably lead to incomplete scene graph predictions.

Baseline Network	s / image (SGG task)	s / image (PSG task)
Motif	1.000	0.102
VCtree	1.690	0.132
SGTR / PSGFormer*	0.350	0.175

Table 1: Inference speed comparison of two-stage and one-stage methods under both the SGG and the PSG tasks. * denotes the one-stage method. SGTR[23] and PSGFormer[46] have similar backbones and relation prediction paradigms.

Inference Speed between Two-stage and One-stage Methods. The improvement in inference speed brought by the non-overlapping masks is a reason why we focus on the two-stage framework design. In the SGG task, since the bounding boxes of instances are allowed to overlap, we can generate a large number of instance proposals during the inference phase, which is exact the fundamental reason why the inference speed of the two-stage method in SGG tasks is much slower than that of the one-stage method. As shown in Table 1, the one-stage method SGTR is significantly faster than the two-stage methods, Motif and VCTree. However, in the PSG task, the situation is reversed. Due to the non-overlapping masks, the number of instances is inherently limited. This constraint ensures that the message-passing network consistently maintains a smaller magnitude of relation candidates for prediction. In contrast, the one-stage method still needs to process a specific number of instance queries. As a result, Table 1 demonstrates that the inference speed of the Motif and VCTree methods outperform the PSGFormer in the PSG task.

3.3 Training with Full Samples

In this section, we strive to address the waste of training sample brought about by direct instance matching. To avoid this problem, we further utilize the panoptic head rather than only using deterministic segmentation results. We extract more instance proposals from the *MaskHead*, a common component in most panoptic heads, which conducts mask feature extraction and mask prediction for each proposal. Then we filter useless proposals to reduce model computation according to the mask confidence:

$$P = \{ p_i \mid m_i^c > \alpha \}. \qquad (1)$$

We use the mask Hungarian algorithm[20] to assign labels for some proposals with the loss item defined as:

$$\mathcal{L}_m = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{cate}\mathcal{L}_{cate}, \qquad (2)$$

where $\mathcal{L}_{cls}$ is the CrossEntropy loss between predicted and truth labels, and $\mathcal{L}_{dice}$ is the dice loss between predicted and target masks. $\mathcal{L}_{cate}$ is a 0-1 loss, which is used to distinguish foreground-background to exclude matchings between thing and stuff. λ_{cls}, λ_{cls}, and λ_{cls} are the weight coefficients for each loss item.

By introducing more instance proposals, the number of matched instances is increased. According to the different utilization ways of these additional proposals, we design two matching strategies.

Step-by-Step Matching. As illustrated in Figure 5 (top), we first conduct direct matching for predicted segmentation results. If there exist some truth instances not matched, continual matching from the remaining proposals is conducted until all-matched. This scheme ensures full utilization of training samples while maintaining the priority of prediction instances, making it a reliable matching scheme. However, this approach requires the panoptic head not only to obtain the segmentation results but also to retain the remaining proposals, which will lead to an increase in computational and storage resources.

Proposal Matching. In this matching scheme, we abandon the segmentation results predicted by the panoptic head and only use the instance proposals for matching, as shown in Figure 5 (bottom). Since the number of proposals is much larger than the number of ground truth instances, this method ensures all-matching. With proposal confidence constraints, this approach can also be a reliable matching scheme. Since this scheme is not required segmentation results calculated by an additional model component *Panoptic Fusion Head*, the computation is reduced.

Under the confidence constraint (Eq. 1), the proposal matching strategy is a simpler and more effective matching method. With this approach, we can further simplify the acquisition of initial features for the input of the message-passing network. Since the proposals extracted from the *MaskHead* inherently contain feature information of instances, we can directly input these features into the message-passing network. This method has three advantages: 1) *Low computation*: In proposal matching, we replace the predicted instances with proposals, and similarly, we can also replace RoI features with proposal features, thus significantly reducing the computation; 2) *Low cumulative error*: Additional feature extraction modules are often based on segmentation results, which inevitably leads to cumulative error. However, proposal features are directly extracted from the feature maps in the panoptic head, without any predictions, thus alleviating the cumulative error problem; 3) *High quality*: As a model dedicated to panoptic segmentation, the proposal features extracted from the panoptic head should contain more accurate instance representation information.

Concretely, for any panoptic heads, we extract proposal features from the *MaskHead*, which contain mask and visual information. Then, we use multi-layer perceptron to de-noise these features.

$$\mathbf{e}_i = MLP(\mathbf{q}_i),\ \mathbf{r}_{i\rightarrow j} = MLP(\mathbf{e}_i \oplus \mathbf{e}_j), \tag{3}$$

where $\mathbf{q}_i$ represents the proposal feature of entity i. $\mathbf{e}_i$ and $\mathbf{r}_{i\rightarrow j}$ denote the initial entity and relation features respectively.

Through Eq. 3, we can easily unify the feature representation of instances and relations, and input these initial features into some popular message-passing networks, *e.g.* VCTree and Transformer, to obtain predictions and conduct training. In summary, within the training pipeline, we merely adjust the instance matching scheme and the method for acquiring initial features, enabling us to fully and effectively harness the samples for model training.

3.4 Relation-Constrained Panoptic Scene Graph Generation

In section 3.2, we explain that the original framework would lead to incomplete scene graph predictions during inference, and the key to alleviating this problem is to recover more valid instances. As shown in Figure 3 (left), for the panoptic head, incomplete segmentation mainly stems from two reasons: 1) It is covered by other instances with higher confidence, *e.g. skis*; 2) The confidence of the instance mask is lower than the designed threshold, so it is discarded, *e.g. mountain-merged* and *backpack*. However, due to the lack of extra attributes, the panoptic head is hard to deal with mask selection in these cases. In order to achieve this goal, we propose a method that uses the relation prediction results to recover the rest valid instances, thus generating a more complete scene graph.

Algorithm 1 Relation-Constrained Segmentation Algorithm

Input: m^c: mask confidence; m: mask prediction; r^c: relation confidence; l: label prediction;

Output: $M \in \mathbb{R}^{w\times h}$: Segmentation Result;

```
1:  Initial $M = 0$, $A = [0]$. $A$ is used to statistic each instance's mask area.
2:  Extract the highest relation confidence for each instance proposal: $r_i^{c*} = max\{r_{k\rightarrow j}^c | k == i \ or \ j == i\}$
3:  Calculate the coverage confidence of each instance proposal: $c_i = m_i^c * r_i^{c*}$, and then sorted $c$
4:  for $c_i$ in $c$ do
5:      $a_i = m_i \ \& \ (M == 0)$
6:      if $\frac{area(a_i)}{area(m_i)} > \epsilon$ then
7:          $M[a_i] = l_i$
8:          $A[i] = area(a_i)$
9:      else if $r_i^{c*} > \delta$ then
10:         $m_i - a_i \rightarrow a_i^{\mathcal{B}}$, $\mathcal{B}$ is the set of covered instances
11:         for $j \in \mathcal{B}$ do
12:             if $\frac{area(a_i^j)}{A[j]} < \tau$ then
13:                 $A[j] = A[j] - area(a_i^j)$
14:                 $a_i = a_i \ \& \ a_i^j$
15:             else
16:                 continue
17:             end if
18:         end for
19:         $M[a_i] = l_i$
20:     end if
21: end for
22: return $M$
```

We first extract high-confidence proposals from the panoptic head through Eq. 1. Each proposal typically contain four types of information: *mask confidence* (m^c), *mask prediction* (m), *label prediction* (l), and *proposal feature* ($\mathbf{q}$). Then, we input the *proposal features* into a message-passing network to obtain *relation confidence* (r^c) and *relation prediction* (r). Finally, we input all instance prediction information and relation prediction information into a unify relation-constrained segmentation algorithm framework to get a relation-prior segmentation results, as shown in Figure 2.

Relation-Constrained Segmentation Algorithm. We present our proposed algorithm in Algorithm 1, a mask-wise merging method that accepts m^c, m, l, and r^c as input. First, we extract the relation candidate with the highest confidence corresponding to each instance proposal (r^{c*} in Line 2) and combine it with the

Panoptic Head	**Model**	SGDet									PQ	Inference Time
		R@20	R@50	R@100	mR@20	mR@50	mR@100	hR@20	hR@50	hR100		
PSGTR†[46]		25.4	27.6	27.7	15.2	16.8	16.8	19.0	20.9	20.9	34.0	0.230
PSGFormer†[46]		17.7	19.3	19.6	14.4	16.6	16.9	15.9	17.9	18.2	41.2	0.175
Panoptic FPN [17]	VCTree	20.6	22.1	22.5	9.7	10.2	10.2	13.2	14.0	14.0	40.3	0.132
	VCTree+Ours	22.3$_{+1.7}$	24.2$_{+2.1}$	24.6$_{+2.1}$	**11.0**$_{+1.3}$	11.5$_{+1.3}$	11.8$_{+1.6}$	**14.7**$_{+1.5}$	15.5$_{+1.5}$	15.9$_{+1.9}$	40.7$_{+0.4}$	0.142
	Transformer	20.3	21.9	22.5	9.2	10.4	10.7	12.7	14.1	14.5	40.3	**0.099**
	Transformer+Ours	**22.7**$_{+2.4}$	**24.3**$_{+2.4}$	**25.0**$_{+2.5}$	10.9$_{+1.7}$	**12.0**$_{+1.6}$	**12.4**$_{+1.7}$	14.7$_{+2.0}$	**16.0**$_{+1.9}$	**16.6**$_{+2.1}$	**40.8**$_{+0.5}$	0.118
Panoptic SegFormer [28]	VCTree	26.4	28.7	29.4	12.3	13.2	13.4	16.8	18.1	18.4	49.4	0.179
	VCTree+Ours	**29.1**$_{+2.7}$	**30.9**$_{+2.2}$	31.4$_{+2.0}$	13.8$_{+1.5}$	14.5$_{+1.3}$	14.7$_{+1.3}$	18.7$_{+1.9}$	19.7$_{+1.6}$	20.0$_{+1.6}$	**49.7**$_{+0.3}$	0.201
	Transformer	26.3	28.4	29.1	12.4	13.2	13.4	16.9	18.0	18.4	49.4	**0.159**
	Transformer+Ours	28.7$_{+2.4}$	30.9$_{+2.5}$	**31.6**$_{+2.5}$	**14.2**$_{+1.8}$	**14.9**$_{+1.7}$	**15.0**$_{+1.6}$	**19.0**$_{+2.1}$	**20.1**$_{+2.1}$	**20.3**$_{+1.9}$	49.5$_{+0.1}$	0.181
Mask2Former [5]	VCTree	27.0	29.2	29.9	13.5	14.3	14.5	18.0	19.2	19.6	50.8	0.189
	VCTree+Ours	**30.8**$_{+3.8}$	**33.5**$_{+4.3}$	**34.4**$_{+4.5}$	15.4$_{+1.9}$	16.3$_{+2.0}$	16.6$_{+2.1}$	20.5$_{+2.5}$	21.9$_{+2.7}$	22.4$_{+2.8}$	50.9$_{+0.1}$	0.197
	Transformer	27.1	29.4	30.0	13.1	14.2	14.5	17.7	19.1	19.5	50.8	**0.141**
	Transformer+Ours	29.7$_{+2.6}$	32.4$_{+3.0}$	33.3$_{+3.3}$	**16.5**$_{+3.4}$	**17.6**$_{+3.4}$	**17.9**$_{+3.4}$	**21.2**$_{+3.5}$	**22.8**$_{+3.7}$	**23.3**$_{+3.8}$	**51.0**$_{+0.2}$	0.155

Table 2: Comparison on the PSG dataset with existing baseline networks based on ResNet-50 backbone. † means the result is produced on unique segmentation results rather than single triplet for equal comparison. The best is highlighted in bold.

mask confidence to obtain the *coverage confidence* (c) for segmentation (Line 3). This step enhances the coverage priority of instance proposals with high relation confidence. Then, the algorithm conducts mask-wise merging with coverage confidence priority. We calculate the disjoint region between the proposal mask and current mask results and perform mask coverage only if the region area is close to the original predicted mask area, controlled by the threshold ϵ (Line 6). Moreover, we introduce an additional decision branch to prevent instances from being covered by their interactive ones. We enforce recovery of instances with high-confidence relation, *e.g.* $r_i^{c*} > \delta$ in Line 9. To ensure that the compulsory recovery instances do not damage the original segmentation results, in Line 12, we set a threshold to further constrain the recovery conditions. We only execute the coverage if the ratio of the covered part to the existing mask area of the instance does not exceed τ. Although our algorithm appears complex, it has proven to be an effective and general operation for various panoptic heads in our experiments. Since most relation predictions tend to be redundant, the thresholds set in this algorithm are crucial for reducing noise in the relations.

After obtaining a panoptic segmentation result under relation-constrained, some proposals are discarded, so we need to eliminate the relations containing useless ones, so as to obtain the final scene graph. In Figure 4 (bottom), we show the advantages of relation-constrained segmentation. In the panoptic head, *skis* shows lower confidence than surrounding instances *snow*, which exactly covers the region of *skis*. At this time, by interacting with the surrounding instances, *person-skis* produces a high-confidence relation prediction, so we can recover *skis* to achieve better segmentation and more complete scene graph prediction.

4 EXPERIMENTS

4.1 Experimetal Setup

Dataset. We evaluate our proposed framework on the only large-scale PSG dataset[46]. PSG conducts relation annotation based on the COCO[30] dataset, which contains 133 entity classes, including 80 thing and 53 stuff classes. PSG has 47874 images with relation annotations, including 45697 for training and 2177 for testing.

Tasks and Evaluation Metrics. We focus on solving the **SGDet** task, which detects the whole scene graph from scratch. To estimate the performance, we use **mean Recall@K (mR@K)**, **Recall@K (R@K)** and **harmonic Recall (hR@K)**[16, 50] as relation evaluation metrics and **PQ** as segmentation evaluation metric. The **hR@K** is defined as the harmonic mean of mR@K and R@K, which has a healthy balance between the head and the tail class performance. We also estimate the inference speed of the model for each image.
Implementation Details: We test three different types of panoptic heads to verify the effectiveness of our proposed framework, including Panoptic FPN[17], Panoptic SegFormer[28], and Mask2Former[5], whose parameters are frozen during training. For all panoptic heads, We set $\alpha = 0.5, 0.7, 0.3$ to filter the proposals for three panoptic heads, respectively. In proposal matching, we set $\lambda_{cls} = 1.0$, $\lambda_{dice} = 2.0$ and $\lambda_{cate} = 1.0$. In relation-constrain segmentation algorithm, ϵ, δ, and τ are set to 0.9, 0.01, and 0.1, respectively. The whole training contains 15 epochs. The batch size is set to 8 and the initial learning rate is 3.0×10^{-2} with being decayed by a factor of 10 at the 9^{th} epoch and 12^{th} epoch. All our experiments are conducted using 2 RTX A5000 GPUs.

4.2 Performance Comparisons

In this section, we perform the quantitative comparison with existing frameworks including PSGTR and PSGFormer[46]. We experiment with three different panoptic heads, including Panoptic FPN[17], Panoptic SegFormer[28], and Mask2Former[5], to verify the universality and effectiveness of our proposed framework.

Table 2 shows the superiority of our framework under two common-used baseline networks, *i.e.*, VCTree[37] and Transformer[40]. When VCTree is used as the message passing network, our framework has an average increase of 10.6%, 12.9%, and 11.9% across R@K, mR@K, and hR@K on three panoptic heads. Similar improvements also take place on the Transformer, which well explain the general effectiveness of our proposed framework. It is worth noting that, although our framework provides more proposals during the relation prediction, it shows close inference speed compared to traditional ones. This is mainly because our

Module			SGDet			PQ	Inference Time
PM	PF	RC	R@20	mR@20	hR@20		
Baseline			27.1	13.1	17.7	50.8	0.141
✓			28.0	13.6	18.3	50.8	0.141
✓	✓		28.1	16.0	20.4	50.8	**0.106**
✓	✓	✓	**29.7**	**16.5**	**21.2**	**51.0**	0.155

Table 3: Ablation study of the framework components with Mask2Former as panoptic head and Transformer as baseline network. *Proposal matching* **(PM) only occurs in the training phase, so it does not impact PQ and inference speed.**

Matching Methods	SGDet		
	R@20 / 100	mR@20 / 100	hR@20 / 100
Direct Matching	27.1 / 30.0	13.1 / 14.5	17.7 / 19.6
Step by Step	**28.7** / **31.3**	14.6 / 15.5	19.4 / 20.7
Proposal Matching	28.1 / 30.9	**16.0** / **16.9**	**20.4** / **21.8**

Table 4: Comparison with different instance matching methods without relation-constrained.

α	SGDet			PQ	Inference Time
	R@20	mR@20	hR@20		
None	28.11	15.95	20.35	50.85	0.106
0.5	**30.06**	**16.68**	**21.45**	50.46	0.199
0.6	29.95	16.52	21.29	50.75	0.188
0.7	29.67	16.46	21.17	**51.05**	0.155
0.8	28.99	16.22	20.80	50.84	**0.145**

Table 5: Performance of different value choices of α to control the number of proposals in the inference stage based on Mask2Former. “None" means relation-constrained method is unused but remaining proposal matching scheme.

framework designs a more simple and effective feature extraction step, which does not require extra RoI calculation and has a lower feature dimension of proposals. Meanwhile, the framework does not require segmentations in the panoptic head, instead conducts scene graph inference and panoptic segmentation simultaneously thus significantly improving the efficiency of model inference.

4.3 Ablation Study and Model Analysis

We further conduct a detailed ablation study over components in our framework and show the performance gain in Table 3. **Baseline** denotes the common PSG framework with Mask2Former as panoptic head and Transformer as baseline network. “PM" and “PF" are the abbreviation of *Proposal Matching* and *Proposal Feature*, respectively. “RC" represents the *Relation-Constrained segmentation*.

Analysis on components. As shown in Table 3, we quantitatively verify the effectiveness of each component. Compared to the baseline, utilizing the proposed PM scheme achieves an improvement of 3.5% on average. It convinces that the proposed PM alleviates the problem of mismatching caused by incomplete segmentations. We further use proposal features to replace the RoI features obtained from RoI feature eatraction module. Compared with the RoI features (i.e., baseline in the 1^{st} row), PF shows a huge performance improvement with 9.8% in relation prediction and 24.8% in inference speed. We argue that extracting extra features in the framework is proven not only to be unnecessary for better scene graph generation but also introduces large requirements on computation resources. Furthermore, We also emphasize the auxiliary role of relations for segmentation. Based on the designed relation-constrained segmentation algorithm, the performance achieves further improvements of 5.7%, 3.1%, and 0.4% on R@K, mR@K, and PQ, separately. While the improvement in PQ is not obvious, the recall has achieved great progress. Relation constrain may recover some unremarkable instances, which still present large effects on segmentation performance. For the relation prediction, the increase in the number of segmentation instances directly increases the integrity of relation prediction, consequently improving R@K and mR@K.

δ	SGDet			PQ
	R@20	mR@20	hR@20	
0.001	29.67	**16.48**	**21.19**	50.98
0.01	**29.67**	16.46	21.17	**51.05**
0.1	29.58	16.42	21.12	50.74
0.2	29.23	16.18	20.83	50.47

Table 6: Performance of different value choices of δ to control the number of compulsory recovery instances.

Discussion on matching scheme. We compare three instance matching strategies in Table 4. Due to the waste of training samples, *direct matching* shows the worst performance. Although *step-by-step matching* scheme shows some improvement, it is not as significant as the *proposal matching* scheme. This scheme defaults to the fact that the matching priority of segmentations is always higher than the rest proposals. So when the panoptic head produces error segmentation results, this scheme will confuse matching.

Value choice of proposal confidence α. In the training phase, the selection of proposal confidence α, which controls the number of instances proposals in the message-passing network, is unnecessary due to the ground truth constraint. But in the inference phase, the value of α is quite important. We conduct a discussion on values of α based on Mask2Former. As shown in Table 5, a small value generates too many proposals, which seriously affects the inference speed of the model and hurts the segmentation results. A large value means fewer instance proposals, resulting in unobvious benefits. We find that $\alpha = 0.7$ has a better overall performance. Due to the difference between panoptic heads, in practice, we choose $\alpha = 0.5$ for Panoptic FPN and 0.3 for Panoptic SegFormer.

Relation confidence δ for instance recovery. In the relation-constrained panoptic segmentation algorithm, we design a threshold δ to control the number of compulsory recovery instances. We experiment with different value choices of δ. As shown in Table 6, $\delta = 0.01$ shows the best overall performance. Small values will recover many unnecessary instances, which hurts the segmentation performance. Large values will cause some instances to be

Metric / Model	Perturb.	No Perturb.		FGSM on Panoptic Head				FGSM on SAM	
		PQ	hR@20	PQ	hR@20	PQ	hR@20	PQ	hR@20
Mask2Former [5]	P	50.8	18.9	45.6	16.4	-	-	43.6	15.9
	P+RC	**51.0**	**19.8**	**46.3**	**17.1**	-	-	**44.1**	**16.7**
Panoptic SegFormer [28]	P	49.4	16.6	-	-	44.0	13.7	41.7	13.3
	P+RC	**49.8**	**17.4**	-	-	**44.7**	**14.6**	**42.5**	**13.9**

Table 7: Performance against adversarial attacks under motifs[49] baseline models. P = PM+PF.

unable to recover, so all related relations can't be predicted, resulting in a significant decline in both the relation prediction and the segmentation performance.

Robustness against segmentation noises. The robustness of model against input noisy consistently attacks significant attentions, especially regarding its resilience against adversarial attacks [29]. To discuss the robustness of our proposed method, we utilize the FGSM [11] adversarial attack method to generate noisy images based on the MaskFormer and SegFormer models, separately. These noisy images are then inputted into the corresponding attack models to generate scene graphs. Additionally, we perform scene graph generation using the noisy images obtained through adversarial attacks on SAM. Our observations reveal that while the adversarial attack does diminish the segmentation (PQ) and PSG (hR@20) performance, the RC consistently mitigates the impact to some extent, demonstrating resilience against adversarial attacks. Specifically, in the case of the large model SAM [18], although experiencing a more substantial decline in performance, the proposed model still exhibits the pull-back effect and showcases robustness for the large model (as shown in Table 7).

4.4 Qualitative Analysis

We visualize several panoptic scene graph prediction results in Figure 6, which show that our proposed framework achieves more complete predictions. In each example, the above and below scene graphs show the results predicted by the original and our proposed framework, respectively. By comparing these results, it is obvious that our framework realizes a more complete panoptic segmentation and scene graph prediction from two aspects: 1) **Recovering instances covered by high-confidence ones**. As shown in the top image, *paper-merged* is covered by *refrigerator* in the original segmentation results. However, in our framework, we input *paper-merged* as a proposal into the message passing network and generate a reliable relation prediction by interacting with *refrigerator*. Therefore, the relation-constrained segmentation algorithm recovers *paper-merged*, thus increasing the triplet *paper-merged-on-refrigerator*, achieving a more complete scene graph prediction; 2) **Retaining un-mask instances with high-confidence relation prediction**. Still using the top image as an example, *bottle* is a low-confidence proposal in the panoptic head, which is discarded in the original segmentation results. However, in our framework, we take relation as another basis for whether to mask, and *bottle-on-cabinet-merged* produces a high-confidence relation prediction, so *bottle* is retained. Other images illustrate the same phenomenon, and the region marked in red shows the ground truth relations that the original framework can't predict.

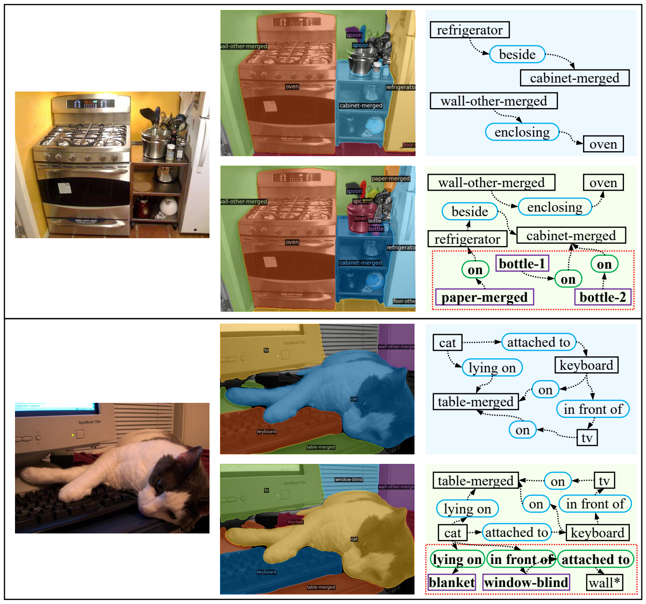


Figure 6: Visualization results of panoptic scene graphs generated by the current PSG framework (above) and our framework (below). Explored extra instances (purple) and relation (green) predictions are highlighted in bold.

5 CONCLUSION

In this paper, we rethink the original SGG framework under panoptic segmentation. By analyzing the differences between the SGG task and the PSG task, we devise a more effective panoptic scene graph generation framework. Considering the non-overlap requirement of the masks, we design distinct pipelines for training and inference. In the training pipeline, we introduce a novel proposal matching strategy that utilizes proposals extracted from the off-the-shelf panoptic head for label alignment instead of using deterministic segmentation results, thereby ensuring the all-matching of training samples. Subsequently, by capitalizing on the proprietary model's exceptional expressiveness, we directly input the proposal features into the message-passing network, consequently reducing the computation required for feature extraction. In the inference pipeline, we develop a versatile and efficient algorithm that reconstructs the process of generating segmentation results from proposals through relation constraints, predicting more comprehensive scene graphs by recovering more valid instances. The results show that our proposed framework not only improves inference efficiency but also achieves outstanding performance.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (Grant No. 2022ZD0118100), in part by National Natural Science Foundation of China (No. 62102413, 61972442, U1936208), in part by the National Science Foundation of Hebei Province of China under Grant F2020202004, and in part by the China PostDoctoral Science Foundation under Grant 2021M703472.

REFERENCES

[1] Oron Ashual and Lior Wolf. 2019. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4561–4569.

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

[3] Chao Chen, Yibing Zhan, Baosheng Yu, Liu Liu, Yong Luo, and Bo Du. 2022. Resistance Training using Prior Bias: toward Unbiased Scene Graph Generation. *arXiv preprint arXiv:2201.06794* (2022).

[4] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. 2020. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12475–12485.

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1290–1299.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 17864–17875.

[7] Weilin Cong, William Wang, and Wang-Chien Lee. 2018. Scene graph generation via conditional random fields. *arXiv preprint arXiv:1811.08075* (2018).

[8] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*. 3076–3086.

[9] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. 2021. Learning of Visual Relations: The Devil is in the Tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15404–15413.

[10] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2022. Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation. *arXiv preprint arXiv:2203.09811* (2022).

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. 1–7.

[12] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. 2021. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16383–16392.

[13] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[14] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1219–1228.

[15] 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/CVF Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[16] Siddhesh Khandelwal and Leonid Sigal. 2022. Iterative Scene Graph Generation. *arXiv preprint arXiv:2207.13440* (2022).

[17] 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.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023), 1–30.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[20] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[22] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. 2022. The Devil is in the Labels: Noisy Label Correction for Robust Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18869–18878.

[23] 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.

[24] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. 2021. Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11109–11119.

[25] Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. 2022. PPDL: Predicate Probability Distribution Based Loss for Unbiased Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19447–19456.

[26] Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia* 21, 8 (2019), 2117–2130.

[27] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*. 1261–1270.

[28] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. 2022. Panoptic SegFormer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1280–1289.

[29] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. 2020. Efficient adversarial attacks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 34–50.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[31] 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.

[32] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. 2022. HL-Net: Heterophily Learning Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19476–19485.

[33] Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. 2022. RU-Net: Regularized Unrolling Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19457–19466.

[34] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. 2019. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3957–3966.

[35] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. 2021. Energy-Based Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13936–13945.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3716–3725.

[37] 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.

[38] Leitian Tao, Li Mi, Nannan Li, Xianhang Cheng, Yaosi Hu, and Zhenzhong Chen. 2022. Predicate correlation learning for scene graph generation. *IEEE Transactions on Image Processing* (2022).

[39] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[40] 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).

[41] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5463–5474.

[42] Qixun Wang, Xiaofeng Guo, and Haofan Wang. 2023. 1st Place Solution for PSG competition with ECCV'22 SenseHuman Workshop. *arXiv preprint arXiv:2302.02651* (2023).

[43] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1508–1517.

[44] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. 2019. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8818–8826.

[45] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. 2020. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 265–273.

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

[47] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10685–10694.

[48] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*. Springer, 606–623.

[49] 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.

[50] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2022. Fine-Grained Scene Graph Generation with Data Transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*. Springer, 409–424.

[51] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. 2021. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 10326–10338.

[52] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. 2023. Prototype-based Embedding Network for Scene Graph Generation. (2023).

[53] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. 1097–1103.