HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation

Zijian Zhou
Department of Informatics
King's College London
zijian.zhou@kcl.ac.uk

Miaojing Shi✉
College of Electronic and Information Engineering
Tongji University
mshi@tongji.edu.cn

Holger Caesar
Intelligent Vehicles Lab
Delft University of Technology
h.caesar@tudelft.nl

Abstract

Panoptic Scene Graph generation (PSG) is a recently proposed task in image scene understanding that aims to segment the image and extract triplets of subjects, objects and their relations to build a scene graph. This task is particularly challenging for two reasons. First, it suffers from a long-tail problem in its relation categories, making naive biased methods more inclined to high-frequency relations. Existing unbiased methods tackle the long-tail problem by data/loss rebalancing to favor low-frequency relations. Second, a subject-object pair can have two or more semantically overlapping relations. While existing methods favor one over the other, our proposed HiLo framework lets different network branches specialize on low and high frequency relations, enforce their consistency and fuse the results. To the best of our knowledge we are the first to propose an explicitly unbiased PSG method. In extensive experiments we show that our HiLo framework achieves state-of-the-art results on the PSG task. We also apply our method to the Scene Graph Generation task that predicts boxes instead of masks and see improvements over all baseline methods. Code is available at https://github.com/franciszzj/HiLo.

1. Introduction

Scene Graph Generation (SGG) [49] is a crucial task in image scene understanding that extracts triplets in the form of subjects, objects and their relations to build a scene graph. Subjects and objects are represented with bounding boxes. Since this task links vision and text, it holds great potential for a variety of applications, including visual question answering [27], image captioning [21, 8], image retrieval [32, 53, 51] and visual reasoning [1, 55].

Recently a novel variant of SGG was proposed, which is Panoptic Scene Graph generation (PSG) [62]. Subjects and objects are represented with panoptic segmentation [35] masks. Unless stated otherwise, in this work we focus on PSG, since it is pixel-level accurate and also covers background classes and their relations with foreground objects.

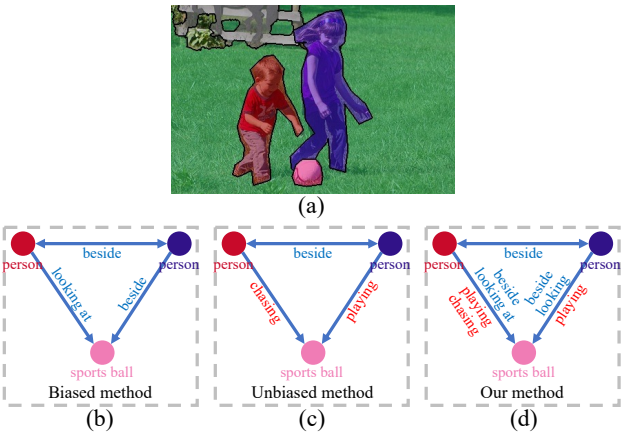

Figure 1. An example of the PSG task. We compare the predicted object relations of different approaches. a) An image and its panoptic segmentation. Two persons and a sports ball are shown in different colors. b) A biased method predicts mostly high frequency relations. c) An unbiased method predicts mostly low frequency relations. d) Our method predicts both low and high frequency relations, as well as more relations in total.

The performance of the PSG model is affected by a *long-tail problem* in its relation categories. For instance, relations such as *over*, *in front of* and *holding* occur tens of thousands of times in the PSG dataset [62], while others like *swinging* and *kissing* occur only a few dozen times. This severe class imbalance in the relation categories can lead to model predictions that are more inclined to high-frequency relations, which poses significant challenges to the application of panoptic scene graphs in real-world scenarios.

Previous methods [12, 18, 39, 17] have often treated the long-tail problem of the PSG task as equivalent to the long-tail problem in object-centric tasks such as classification [15, 5, 29] or semantic segmentation [14, 2]. Consequently,

✉ Corresponding author.

these methods have employed re-balancing techniques to address class imbalance, either through re-sampling the data [39] or by using a class-balanced loss [33] that assigns different weights to different relation categories.

In contrast, in the relation-centric PSG task, a subject-object pair can have multiple relations that exhibit *relational semantic overlap*, such as being partially or fully overlapping. For example, in Fig. 1, there are multiple relations between the boy and the sports ball, such as *beside*, *looking at*, *playing* and *chasing*. For regular biased models [61, 65, 60, 46], the results are dominated by high-frequency relations (*beside*, *looking at*). For specifically unbiased models [66, 62], the results are dominated by low-frequency relations (*playing*, *chasing*). However, since the low frequency relations can be more specific (e.g. *on* and *standing on*) or only partially overlapping with high frequency relations (e.g. *looking at* and *chasing*), it is crucial to include both to fully understand the image. We found that relational semantic overlap occurs in large numbers in the PSG dataset [62] and that current methods do not effectively address it. This is reflected in the increase in the category-averaged *mean recall* metric of unbiased methods, at the cost of the decrease in global *recall* (see Sec. 4.4).

To address the long-tail problem of scene graphs under relational semantic overlap, we introduce the HiLo framework. This framework simultaneously learns the high and low frequency relations in different network branches and unifies their strengths with the help of two novel consistency loss functions. We apply our framework on top of a novel baseline. This baseline uses a recent transformer-based approach [10] for panoptic segmentation and adapts triplet queries [62] and masked attention [10] for the PSG task. In summary, we make the following contributions:

- We identify the long-tail problem with relational semantic overlap in the PSG task and propose the HiLo framework to address this problem. The framework is general and can be applied to any PSG method.
- We propose a powerful and efficient one-stage end-to-end baseline. This baseline enhances the interaction between mask and relation prediction in the transformer decoder layer.
- We conduct extensive experiments to demonstrate the effectiveness of our framework and baseline. Our results outperform the state-of-the-art in both recall and mean recall on the PSG dataset and show systematic improvements on the VG dataset.

2. Related Work

2.1. Scene Graph Generation

The Scene Graph Generation (SGG) [49] task plays a crucial role in connecting vision and language, and has received widespread attention in the computer vision community. Many methods have been proposed to improve the performance of SGG, which can be classified into three categories [70]. The first is to introduce multi-modal information, such as appearance [52], space [72], depth [54], and segmentation [34]. The second is to introduce prior information and commonsense knowledge, such as statistical [3, 16, 65, 9] and language prior knowledge [49, 43, 68, 31, 20]. The third category involves designing different model structures, such as message passing [41, 16, 42, 65, 24, 64], attention mechanisms [69, 50], tree structures and visual translation [67, 30]. However, most of these methods are two-stage methods that cannot learn scene graphs end-to-end. In contrast, to improve the learning ability of SGG models, several methods based on transformers have been proposed, including SGTR [38], RelationFormer [56] and RelTR [13]. These are end-to-end trainable in a single stage.

2.2. Unbiased Scene Graph Generation

Solving the long-tail problem in the SGG task has attracted considerable attention from researchers, and several unbiased methods [59, 63, 12, 18, 25, 39, 19, 22, 37, 40, 17, 66] have been proposed. These methods typically improve the model from the perspective of data re-sampling [39] or a class-balanced loss [33]. BGNN [39] uses a two-layer re-sampling strategy to provide a more balanced data distribution during training, while CogTree [63] proposal exploits the semantic relation between different predicate classes to design a novel CogTree loss. HML [17] improves the model's ability to solve long-tail problems by designing a staged training process, and IETrans [66] proposes an internal and external transfer method to transfer high frequency relations to low frequency relations and recover missing relations to train the unbiased model. Dong *et al*. [19] propose to group relations by their frequency and train specialized relation encoders for each group. While these methods are able to mitigate the long-tail problem, they do not address relational semantic overlap.

2.3. Panoptic Scene Graph Generation

Contrary to SGG, the PSG task [62] uses panoptic segmentation masks instead of bounding boxes to represent objects, resulting in a more comprehensive scene graph. PSGTR [62], an end-to-end method based on the DETR structure, was proposed to construct a transformer-based PSG model. PSGFormer [62] further improved on PSGTR by introducing Object & Relation Query Learning Blocks and Query Matching Blocks. Their strong performance on most relation classes indicates that their method is implicitly unbiased. In contrast, our approach is the first to explicitly bias a PSG model, creating separate branches for low and high frequency relations, which are then fused together.

Since the scene graph in PSG is built upon the sub-

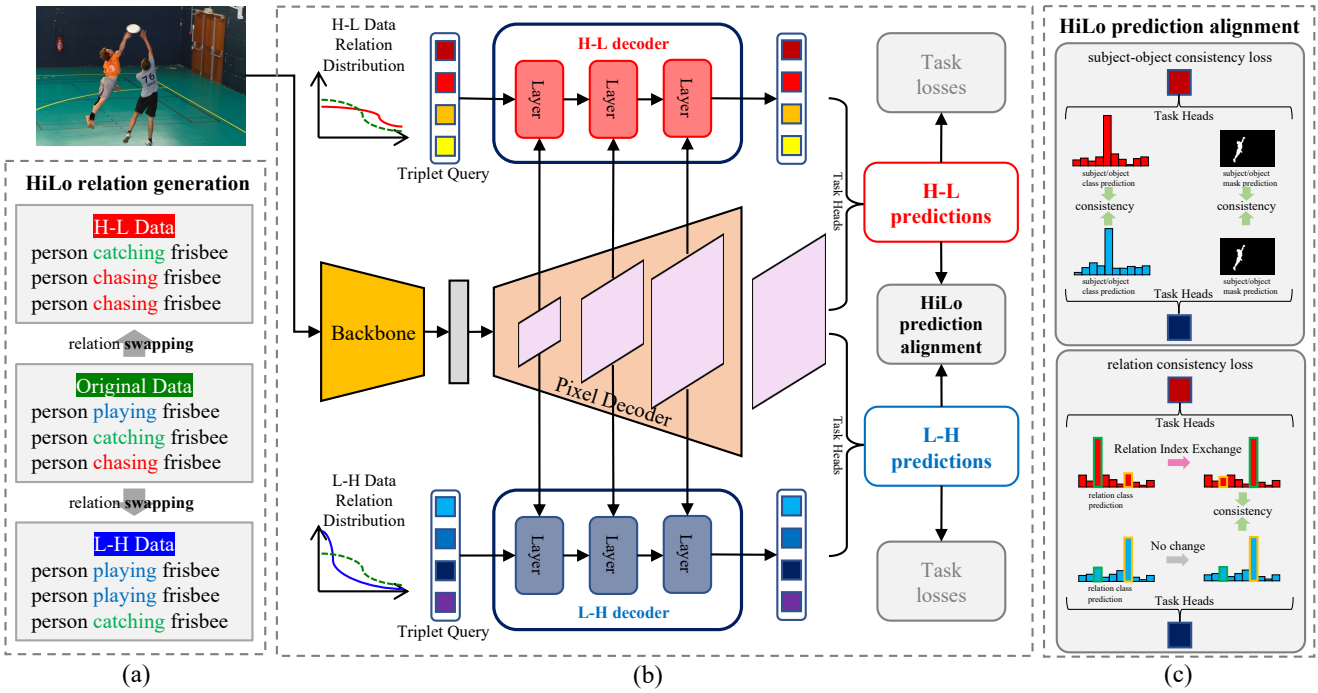

Figure 2. An overview of our HiLo framework with HiLo baseline. a) HiLo relation swapping module swaps the multiple relations in the subject-object pair to obtain H-L Data and L-H Data respectively. b) Input data into our HiLo framework with HiLo baseline model, there are two branches, namely H-L decoder and L-H decoder, which learn H-L Data and L-H Data respectively. c) In addition to task losses for PSG, we propose HiLo prediction alignment, which includes subject-object consistency loss and relation consistency loss, so that the parallel branch can be better optimized.

jects, objects and their relations, strong panoptic segmentation [35] is crucial for PSG. Several DETR-based [6] methods such as Deformable-DETR [71], Segmenter [57], MaskFormer [11] and Mask2Former [10] have recently pushed the envelope in panoptic segmentation. These methods have introduced deformable transformer encoders and decoders, as well as pixel decoders and multi-scale information to improve model performance and speed up convergence. Our proposed baseline builds upon these techniques to further enhance the PSG performance and speed up model convergence.

3. Method

3.1. Problem Setting

The Panoptic Scene Graph generation (PSG) task aims to generate a panoptic scene graph $\mathcal{G}$ for a given image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, where $\mathcal{G}$ contains an object set $\mathcal{O}$ and a relational triplet set $\mathcal{T}$, denoted by $\mathcal{G} = (\mathcal{O}, \mathcal{T})$. For the i-th object o_i in $\mathcal{O} = \{o_i\}_{i=1}^{N}$, we use m_i to represent the object's mask and c_i the object's category, *i.e.* $o_i = (m_i, c_i)$. For the j-th triplet in $\mathcal{T} = \{t_j\}_{j=1}^{M}$, we use s_j to represent the subject, o_j the object, and r_j their relation, *i.e.* $t_j = (s_j, o_j, r_j)$. There are in total C object categories and R relation categories. For the k-th relation, we denote its frequency in the training set as f_k.

3.2. HiLo Baseline

High-quality panoptic segmentation is crucial for achieving good PSG performance. We build our method upon the latest advances in DETR-based panoptic segmentation, Mask2Former [10]. Below we present its structure, as well as our proposed modifications for the PSG task.

Mask2Former. This method comprises three key parts: A backbone (CNN-based or transformer-based) followed by a pixel decoder, a transformer module with a transformer decoder and a task-specific module with different task heads. Specifically, the backbone takes an input image $\mathcal{I}$ and generates an image feature F. The pixel decoder then gradually upsamples F to produce multi-scale features $\widetilde{\mathcal{F}} = \{\widetilde{F}_i\}_{i=1}^{4}$. The transformer decoder takes a set of queries $\mathcal{Q}$ of size N and multi-scale features $\widetilde{\mathcal{F}}$ as input and outputs a set of mask features $\mathcal{X}$ of the same size with $\mathcal{Q}$.

On top of the transformer decoder, there are two task heads including a linear classifier that predicts the class probability for each mask; and a multi-layer perceptron (MLP) that uses the mask features $\mathcal{X}$ to generate the mask embedding $\mathcal{E}$. The mask prediction is obtained by taking the dot product of the mask embedding $\mathcal{E}$ with the scale feature of the highest resolution in the multi-scale features $\widetilde{\mathcal{F}}$.

Triplet queries. The original query in Mask2Former is to predict the object. In order to predict both subject, object

and their relation, we develop the triplet queries $\mathcal{Q}^t$ inspired by [62] into our baseline. Each query predicts a triplet that includes subject, object, and relation. Accordingly, our task heads comprise three linear classifiers. Each classifier is responsible for predicting the class probability for subject, object or relation, respectively. Moreover, we devise two MLPs to generate mask embeddings for the subject and object, denoted by $\mathcal{E}^s$ and $\mathcal{E}^o$. We use them to compute the dot product with the scale feature of the highest resolution in multi-scale feature $\widetilde{\mathcal{F}}$ and obtain the mask prediction for the subject and object, respectively.

Masked relation attention. [10] proposes a transformer decoder variant with masked attention, which extracts spatial features by adding the predicted mask of the object from the previous decoder layer's mask prediction. It makes the model focus on the object-related area in the feature map. To adapt this scheme to the PSG task, for each relation, we extend the masked attention to take the union of the binary masks of the subject and object as input, which represents the pixels corresponding to the relation.

Network training. We adopt the same losses as PSGTR [62], including cross-entropy loss $\mathcal{L}_{so\_cls}$ for object classification of subject-object pairs and a combination $\mathcal{L}_{so\_mask}$ of focal loss [44] and dice loss [58] to jointly supervise mask learning. To supervise the relations, we use the cross-entropy loss $\mathcal{L}_{rel\_cls}$. The baseline loss with ($\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 4$) is thus:

$$\mathcal{L}_{baseline} = \lambda_1 \cdot \mathcal{L}_{so\_cls} + \lambda_2 \cdot \mathcal{L}_{so\_mask} + \lambda_3 \cdot \mathcal{L}_{rel\_cls} \quad (1)$$

3.3. HiLo Framework

The key insight of our HiLo is to build a model that can take into account both high frequency and low frequency relations, and effectively improve the performance on low frequency relations without degrading the performance on high frequency relations.

3.3.1 HiLo relation generation

In the PSG task, multiple relations can be used to describe the connection between a subject and an object from different perspectives, such as *spatial* relations, *actions* and *prepositions*. Since the subject, object and their relative position in the image are fixed, all these relations share the same visual information. This is reflected in the PSG dataset [62] where many subject-object pairs are annotated with multiple relation labels. These relation labels are of different frequencies in the dataset. In this section we introduce the HiLo relation generation module to prepare two sets of training data, biasing towards high and low-frequency relations respectively.

Relation augmentation. Similar to [22], we observe that there are many missing relation annotations in the PSG dataset [62]. To add the missing relation annotations, we design a relation augmentation scheme that is inspired by IETrans [66], which converts high-frequency relations to low-frequency relations and adds a relation to a subject-object pair that has no relation. We adapt it by first training our baseline as a biased model using the original annotated data. For every subject-object pair in the training set, we use this model to predict the relation scores for all predefined relation categories:

- If this subject-object pair has annotated relation labels, we pick the one with the highest predicted score and use the score as a threshold. For other relations (except for already labeled ones) whose predicted scores are greater than the threshold, we add them as the relation labels of this subject-object pair.
- If this subject-object pair has no annotated relation labels, we use the predicted score of the *no-relation* class as a threshold. For other relations whose predicted scores are greater than this threshold we add them as the relation labels of this subject-object pair.

This operation allows us to significantly augment the relation labels for the subject-object pairs in the training data, which is especially relevant for pairs with zero or no annotated relations.

Relation swapping. We swap the relation labels for each subject-object pair. Specifically, given a subject-object pair (s, o) with K relation labels, we have K triplets $(s, o, r_1), ..., (s, o, r_K)$, sorted by their relation frequency in descending order, $f_1 > ... > f_K$. We denote by H-L and L-H the swapping of high-frequency relations with low-frequency relations and vice versa. This creates two sets of data:

H-L Data. Given a triplet (s, o, r_k), we replace its relation label r_k with that of the next triplet r_{k+1} with lower frequency. We sequentially process all triplets from (s, o, r_1) until (s, o, r_{K-1}), keeping the last triplet unchanged.

L-H Data. Given a triplet (s, o, r_k), we replace its relation label r_k with that of the previous triplet r_{k-1} with higher frequency. We sequentially process all triplets from (s, o, r_K) until (s, o, r_2), keeping the first triplet unchanged. Hence, we obtain two new sets of triplets, denoted by $\mathcal{T}^{\text{H-L}}$ and $\mathcal{T}^{\text{L-H}}$. We devise two parallel decoders from the shared encoder of our backbone. They are learned with the H-L and L-H data, respectively. The H-L and L-H decoders favour the predictions for low and high-frequency relations, respectively. Despite their difference, for the corresponding triplet query in the two decoders, their predictions are highly correlated: on one hand, the subject and object predictions should be the same; on the other hand, the distribution of relation predictions should be overlapping. Below we first build the query correspondence in the two decoders and then introduce the HiLo subject-object and relation consistency loss to align the predictions from two the decoders.

3.3.2 HiLo prediction alignment

Training two different relation distributions simultaneously confuse the model. To make the model differentiate between the two branches, we built a HiLo prediction alignment module, including triplet query correspondence, subject-object consistency loss and relation consistency loss respectively.

Triplet query correspondence. In order to construct the subject-object consistency loss and relation consistency loss, we first need to construct the correspondence between the triplet queries in the H-L and L-H decoders. The same query index in the two decoders is not naturally matched. In order to find the query correspondence, we need to rely on the ground truth assignment: we use Hungarian matching to assign the triplet label to the corresponding triplet query, and record the relation label index corresponding to the triplet query. This label index allows us to construct the correspondence between the triplet queries in both decoders. We calculate the consistency loss for the triplet query prediction with the same relation label index in the two decoders.

Subject-object consistency loss. Having the corresponding predictions from the two decoders, both their subjects and objects have the same ground truth and should be equal. We propose a subject-object consistency loss $\mathcal{L}_{obj}$ to minimize the mean squared error (MSE) of the corresponding predictions from the two decoders.

$$\mathcal{L}_{obj} = \mathcal{L}_{cls} + \mathcal{L}_{mask} \tag{2}$$

$$\mathcal{L}_{cls} = \|softmax(p_c^{\text{H-L}}) - softmax(p_c^{\text{L-H}})\|^2 \tag{3}$$

$$\mathcal{L}_{mask} = \|sigmoid(p_m^{\text{H-L}}) - sigmoid(p_m^{\text{L-H}})\|^2 \tag{4}$$

Here $\mathcal{L}_{cls}$ and $\mathcal{L}_{mask}$ represent the MSE losses for class prediction and mask prediction, respectively.

$p_c^{\text{H-L}}$ and $p_c^{\text{L-H}}$ are the class prediction logits from the H-L and L-H decoders. After a softmax, they are C-dimensional probability vectors. After a sigmoid, p_m^{H-L} and p_m^{L-H} are the mask prediction logits for the H-L and L-H decoders.

Relation consistency loss. Given a pair of subject and object, we have previously swapped the high-low frequency relation labels to create data for the H-L and L-H decoders. For H-L, the prediction of the low-frequency relation logit is of high value; while for L-H, the prediction of the high-frequency relation logit is of high value. For the predictions on the rest logits, they should be similar, since it is the same pair of subject and object for the two decoders. Based on this observation, we introduce the relation consistency loss.

Specifically, for a pair of subject and object, we use $p_r^{\text{H-L}}$ to denote the predicted relation logits from the H-L decoder, $p_r^{\text{L-H}}$ the predicted relation logits from the L-H decoder. They are R-dimensional probability vectors after softmax. We can map between the distributions of the two vectors by swapping the logit values between the high- and low-frequency relation indices, which is named Relational Index Exchange operation. We use RIE$(\cdot)$ to denote this operation. RIE$(\cdot)$ includes a stop gradient operation, which creates copies from original predictions of two branches, enabling value exchanges. For example, for relations r_k and r_{k+1}, RIE$(\cdot)$ exchanges the values between $p_{r,k}$ and $p_{r,k+1}$ in p_r. We can therefore compute the distance between $p_r^{\text{H-L}}$ and its mapped counterpart from $p_r^{\text{L-H}}$, vice versa:

$$\begin{aligned}\text{Dist}_{\text{HiLo}} =& \|softmax(p_r^{\text{H-L}}) - RIE(softmax(p_r^{\text{L-H}}))\|^2 \\ &+ \|RIE(softmax(p_r^{\text{H-L}})) - softmax(p_r^{\text{L-H}})\|^2\end{aligned} \tag{5}$$

The relation consistency loss is defined to minimize $\text{Dist}_{\text{HiLo}}$ with a margin of m:

$$\mathcal{L}_{rel} = max(\text{Dist}_{\text{HiLo}} - m, 0), \tag{6}$$

where m is a small constant. Adding m is due to the fact that the high- and low frequency relations might be only partially semantically overlapping.

Network training. Our subject-object consistency loss and relation consistency loss can seamlessly integrate with the losses of any baseline method to jointly supervise the training of the entire model. Notably, we supervise the output of each transformer decoder layer to ensure effective learning. The final loss $\mathcal{L}$ is thus:

$$\mathcal{L} = \mathcal{L}_{baseline} + \mathcal{L}_{obj} + \mathcal{L}_{rel} \tag{7}$$

3.3.3 HiLo inference fusion

The H-L and L-H decoders favour low-frequency relation prediction and high-frequency relation prediction, respectively. To combine the strength of both during inference, we introduce the HiLo inference fusion module. Specifically, we denote by $\mathcal{G}^{\text{H-L}}$ and $\mathcal{G}^{\text{L-H}}$ the predicted panoptic scene graphs from H-L and L-H decoders, respectively. There are N_1 triplets in $\mathcal{G}^{\text{H-L}}$ and N_2 triplets in $\mathcal{G}^{\text{L-H}}$.

- First, we merge the triplets in $\mathcal{G}^{\text{H-L}}$ and $\mathcal{G}^{\text{L-H}}$ and sort them in the descending order according to their relation scores. We obtain a list of $N_1 + N_2$ triplets.
- Second, starting from the first triplet, we de-duplicate the triplet list. For the i-th triplet $\mathcal{T}_i$, we identify its duplicated versions from the $(i+1)$-th triplet until the end of the list. Remove any follow-up triplet from the list if it 1) has the same subject, object and relation classes to that in T_i and 2) has a mask IoU greater than a threshold, *e.g.* 0.5, between the predicted subject/object and the corresponding subject/object in T_i.
- Third, after deduplication, for each triplet in the list, we multiply the relation, subject and object scores as an overall score for it. We sort the triplet list according to this score in descending order to obtain the final panoptic scene graph $\mathcal{G}^{\text{HiLo}}$.

Method	Backbone	Scene Graph Detection					
		R@20	mR@20	R@50	mR@50	R@100	mR@100
IMP [61]	R50	16.5	6.5	18.2	7.1	18.6	7.2
MOTIF [65]	R50	20.0	9.1	21.7	9.6	22.0	9.7
VCTree [60]	R50	20.6	9.7	22.1	10.2	22.5	10.2
GPSNet [46]	R50	17.8	7.0	19.6	7.5	20.1	7.7
PSGTR [62]	R50	28.4	16.6	34.4	20.8	36.3	22.1
PSGFormer [62]	R50	18.0	14.8	19.6	17.0	20.1	17.6
HiLo (ours)	R50	34.1	23.7	40.7	30.3	43.0	33.1
HiLo (ours)	Swin-B	38.5	28.3	46.2	35.3	49.6	39.1
HiLo (ours)	Swin-L	**40.6**	**29.7**	**48.7**	**37.6**	**51.4**	**40.9**

Table 1. Comparison between our HiLo and other methods on the PSG dataset. Our method shows superior performance compared to all previous methods.

4. Experiments

4.1. Datasets

Panoptic Scene Graph Generation (PSG) [62]. This is the first Panoptic Scene Graph generation dataset. It has a total of 48,749 labeled images including 2,177 test images and 46,572 training images. The object categories comprise 80 thing classes and 53 stuff classes, which is the same as the COCO [45] and COCO-Stuff datasets [4]. The relation categories comprise 56 classes, including positional relations, common object-object relations, common actions, human actions, actions in the traffic scene, actions in the sports scene and interactions between backgrounds [62].

Visual Genome (VG) [36]. VG is a widely used benchmark dataset for Scene Graph Generation. Following previous work [65, 7], we adopt the widely accepted split, VG-150, which contains 150 object categories and 50 relation categories. The object categories cover a wide range of classes, such as *animals*, *vehicles* and *household items*. The relation categories include both spatial and semantic classes, such as *on*, *in* and *wearing*.

4.2. Tasks and Metrics

Three subtasks have been proposed for the SGG and PSG tasks, which are *Predicate Classification*, *Scene Graph Classification* and *Scene Graph Detection* [61]. We focus on Scene Graph Detection for both datasets, since it is the most comprehensive and addressed by [62]. This subtask requires the model to first localize the objects and then predict the object classes and relations. Note that Scene Graph Detection in PSG includes the detection on stuff classes, while in SGG it does not.

Following previous work [60, 63, 62], we use Recall@K (R@K) and mean Recall@K (mR@K) as our metrics, where the former metric is dominated by high-frequency relations, while the latter assigns equal weight to all relation classes.

4.3. Implementation details

In our experiments, we follow the training strategy of PSGTR [62]. We use the AdamW optimizer [48], with a learning rate of $1e^{-4}$ and weight decay of $1e^{-4}$, except for the backbone, which is trained with a learning rate of $1e^{-5}$. For initialization, we use Mask2Former [10] pre-trained on COCO [45] to initialize our backbone and pixel decoder. Following Mask2Former [10], we use 100 triplet queries for the H-L and L-H decoders respectively. Additionally, both the H-L and L-H decoders are initialized with Mask2Former's transformer decoder. To ensure consistent comparison with PSGTR, we adopted the same data augmentation settings. Our model is trained for 12 epochs with a step scheduler at epoch 10, taking approximately 18 hours to train on four A100 GPUs with a batch size of 1 for each GPU.

4.4. Comparison to the state-of-the-art

PSG. Tab. 1 reports the performance of our method compared to the state-of-the-art on the PSG dataset [62]. We separate the methods into two groups. The first are two-stage methods consisting of a separate segmentor and relation predictor, which are modified for the PSG task in [62]. The second are one-stage end-to-end methods, which are able to simultaneously predict panoptic segmentation and relations. Our method belongs to the second category. For a fair comparison between the different methods, we use the same Resnet-50 [26] backbone. Our method shows superior performance compared to all previous methods. Particularly, it outperforms the previous best-performing method PSGTR [62] by a large margin, *i.e.* +6% in R@100 and +11% in mR@100. Our model is able to converge within only 12 epochs of training, whereas PSGTR [62] is trained for 60 epochs. We also evaluate our method using pre-trained transformer-based backbones, *i.e.* Swin-B and Swin-L [47], with the latter being a bigger model. Our results are consistently improved over all metrics due the powerful feature representation ability of the transformer. We also conducted a visual comparison, as shown in Fig. 3.

Method	Scene Graph Detection	
	R/mR@50	R/mR@100
MOTIF [65]	**31.0** / 6.7	**35.1** / 7.7
+IETrans [66]	26.4 / 12.4	30.6 / 14.9
+HiLo (ours)	26.2 / **14.7**	30.3 / **17.7**
VCTree [60]	**30.2** / 6.7	**34.6** / 8.0
+IETrans [66]	25.4 / 11.5	29.3 / 14.0
+HiLo (ours)	27.1 / **12.9**	29.8 / **15.2**
Transformer [59]	**30.0** / 7.4	**34.3** / 8.8
+IETrans [66]	25.5 / 12.5	29.6 / 15.0
+HiLo (ours)	25.4 / **14.6**	29.2 / **17.6**
GPSNet [46]	**30.3** / 5.9	**35.0** / 7.1
+IETrans [66]	25.9 / 14.6	28.1 / 16.5
+HiLo (ours)	25.6 / **15.8**	27.9 / **18.0**

Table 2. Comparison between our HiLo framework and other methods on the VG-150 dataset. Similar to [66], we apply IETrans and our own method on top of four leading baselines.

SGG. Here we study whether our approach that was developed for the PSG task can also be applied to the SGG task. In Tab. 2, we conduct experiments on the VG-150 dataset. Following IETrans [66], we extend our *unbiased* HiLo framework (Sec. 3.3) without our PSG-specific baseline (Sec. 3.2) to four state-of-the-art *biased* SGG methods, namely MOTIF [65], VCTree [60], Transformer [59] and GPSNet [46], using the implementation of [66]. Compared to the unbiased IETrans [66] method, our method improves mean recall without sacrificing recall, demonstrating its effectiveness in enhancing the performance of low-frequency relations. Compared to the biased baselines [65, 60, 59, 46], our method achieves a significantly larger mean recall, while still maintaining an acceptable recall. This indicates that our method can improve the performance of low-frequency relations while also taking into account the performance of high-frequency relations. It also shows that the HiLo framework is a general technique that yields systematic improvements in both the PSG and SGG tasks.

4.5. Ablation Studies

Consistent with the paper, we use HiLo with a Resnet-50 backbone to perform ablation experiments on the PSG dataset.

HiLo framework for different baselines. In this section we investigate whether our HiLo framework (Sec. 3.3) yields improvements for other baselines, rather than just the one presented in Sec. 3.2. Tab. 3 shows the results for two baselines, with and without the HiLo framework. We observe that our biased baseline outperforms the previous PSGTR [62] method on all metrics. Furthermore, by applying the HiLo framework, we can substantially improve the performance over both baselines. It is worth mentioning that the HiLo framework improves recall and mean recall simultaneously, whereas other methods typically improve one metric at the cost of the other [59, 63].

HiLo relation augmentation. We observe that out of 260,296 labeled triplets in the PSG dataset, only about 10% of subject-object pairs have multiple relations, for which we can apply relation swapping (Sec. 3.3.1). After applying our proposed relation augmentation technique (Sec. 3.3.1), this ratio significantly increases to 40%. Our experimental results in Tab. 4 demonstrate that only applying the HiLo framework on 10% already gives an improvement over the baseline from Tab. 3. As the number of swappable triplets increases due to augmentation, the model's performance is further enhanced, highlighting the potential of our method.

Baseline	HiLo	R/mR@20	R/mR@50	R/mR@100
HiLo baseline	✓	**34.1 / 23.7**	**40.7 / 30.3**	**43.0 / 33.1**
HiLo baseline	-	32.6 / 20.9	38.0 / 27.4	38.9 / 28.4
PSGTR [62]	✓	**30.1 / 20.2**	**36.6 / 23.9**	**38.3 / 24.5**
PSGTR [62]	-	28.4 / 16.6	34.4 / 20.8	36.3 / 22.1

Table 3. Comparison of different baselines, with and without HiLo framework. Using the HiLo framework, we see significant improvements on both metrics.

Relation Aug.	Multiple relations	R/mR@50	R/mR@100
✓	40%	**40.7 / 30.3**	**43.0 / 33.1**
-	10%	40.1 / 28.1	42.8 / 32.5

Table 4. Ablation study for HiLo relation augmentation. Relation augmentation affects the ratio of subject-object pairs with multiple relations. The larger this ratio, the more relations can be swapped, which leads to better results.

HiLo prediction alignment. We conduct ablation experiments on the subject-object consistency loss and relation consistency loss (Sec. 3.3.2), which are used to align the predictions from the high and low frequency branches. The results, as presented in Tab. 5, demonstrate that using both losses yields the best performance. It is worth mentioning that we have explored the margin in the relation consistency loss and found that setting the margin to zero leads to a small performance degradation. This finding confirms that there is a partial semantic overlap between swapped relations, indicating that they are not entirely consistent.

To investigate the impact of relational index exchange (RIE) on the relation consistency loss, we conducted experiments to verify the effect of omitting RIE. In the absence of RIE, we solely compute the consistency loss for relation categories that are not involved in relation swapping and exclude the swapping component from the calculation of the consistency loss. The outcomes of this experiment are presented in Tab. 6, and demonstrate a notable reduction in the mean recall and a decline in the model's performance for relations with relational semantic overlap when RIE is not utilized.

HiLo inference fusion. We ablate the inference fusion (Sec. 3.3.3) and evaluate the performance of each branch's output separately. In Tab. 7, experimental results suggest that fusion can effectively leverage the uniqueness of the high and low frequency branch predictions to achieve comprehensive improvements on all metrics.

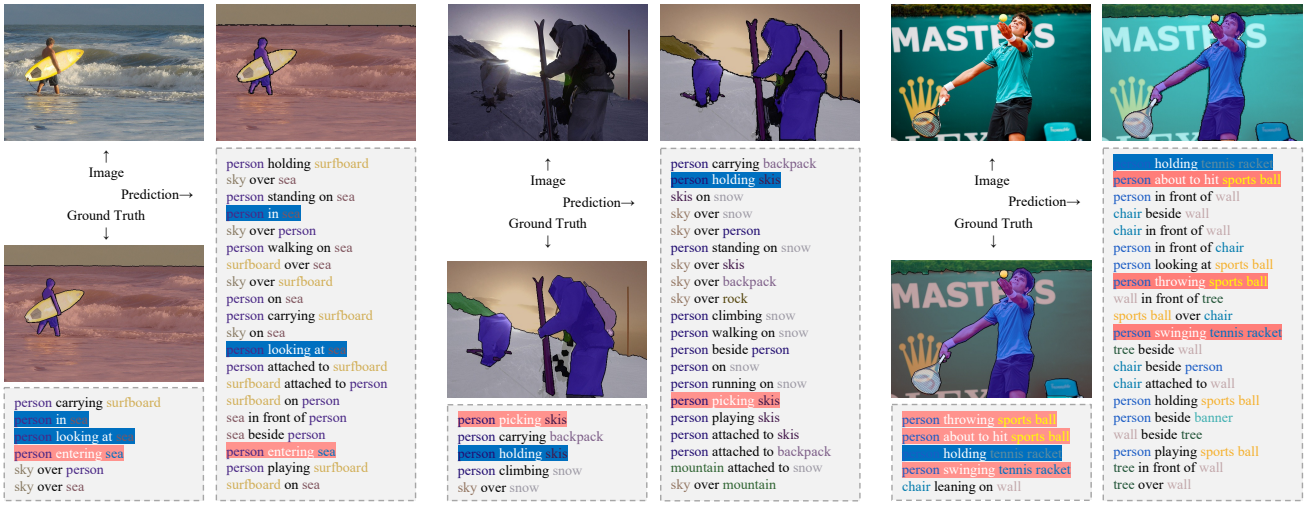

Figure 3. Visualization of panoptic segmentations and the top 20 predicted triplets compared with ground truth. The upper left is the original image, the lower left is the ground truth and on the right are the predictions. The highlighted triplets represent the subject-object pairs with multiple relations, where the blue highlights represent the high frequency relations and the red highlights represents the low frequency relations. The visualization results show that our method can predict both high frequency and low frequency relations.

Object	Relation	Margin	R/mR@50	R/mR@100
✓	✓	0.5	**40.7 / 30.3**	**43.0 / 33.1**
-	✓	0.5	40.6 / 29.7	42.8 / 32.8
-	✓	0.0	40.5 / 29.5	42.7 / 32.8
✓	-	-	40.4 / 29.0	42.8 / 32.2
-	-	-	39.7 / 28.6	42.4 / 32.0

Table 5. Ablation study for different losses in HiLo prediction alignment. *Object* refers to the subject-object consistency loss and *relation* refers to the relation consistency loss. The margin parameter is defined in Eq. 6.

Whether to use RIE	R/mR@20	R/mR@50	R/mR@100
✓	**34.1 / 23.7**	**40.7 / 30.3**	**43.0 / 33.1**
-	33.5 / 22.3	40.3 / 29.0	42.6 / 32.3

Table 6. Ablation study for relation consistency loss in HiLo prediction alignment.

We also attempted to average the tensor generated by the two branches and obtain the PSG result through post-processing. However, we found that this approach leads to a substantial drop in performance, as evident in Tab. 8. This can be attributed to the inconsistent prediction results of the two branches for the same query index. These findings validate that the inference fusion method effectively merges the results from the two branches. Furthermore, our experimental results demonstrate that the query associated with the identical index in two branches does not predict the same subject-object pair. Thus, directly averaging the tensor produced by the two branches results in prediction ambiguity, ultimately leading to a substantial decline in performance. This observation underscores the necessity of conducting triplet query correspondence when performing prediction alignment. In particular, due to the inconsistent query prediction content for the corresponding index in the two branches, a one-to-one correspondence must be constructed based on the label assigned by each query to achieve prediction alignment.

H-L Result	L-H Result	R/mR@50	R/mR@100
✓	✓	**40.7 / 30.3**	**43.0 / 33.1**
✓	-	38.8 / 29.9	39.8 / 30.9
-	✓	38.5 / 26.5	39.5 / 27.8

Table 7. Ablation study for HiLo inference fusion.

Fusion method	R/mR@20	R/mR@50	R/mR@100
inference fusion	**34.1 / 23.7**	**40.7 / 30.3**	**43.0 / 33.1**
average tensor	19.6 / 13.1	23.1 / 15.7	23.9 / 16.3

Table 8. Ablation study for HiLo inference fusion. *Inference fusion* refers to the method proposed in the paper to fuse the results of two branches. *Average tensor* refers to the fusion method that directly averages the tensor output by the two branches.

Attention focus	R/mR@20	R/mR@50	R/mR@100
subject-object	**32.6 / 20.9**	**38.0 / 27.4**	**38.9 / 28.4**
full image	30.4 / 19.3	36.5 / 25.9	37.1 / 26.8

Table 9. Ablation study for masked relation attention.

Masked relation attention. We investigate the impact of different mask input types for cross-attention on the HiLo baseline (Sec. 3.2) performance. Specifically, we compare two different attention focus regions, namely the subject-object region and the full image. The results are shown in Tab. 9. Focusing on the full image presents a more challenging optimization task for the model since no target region is specified. Consequently, we observe a drop of 1.8% in R@100 and 1.6% in mR@100. This shows that it is crucial to apply masked relation attention to the subject and object.

Method	All	Rare	All	Overlap
PSGTR	22.1	6.2	22.1	23.5
HiLo (ours)	33.1 (+11.0)	20.3 (+14.1)	33.1 (+11.0)	38.8 (+15.3)

Table 10. Verification of the improvements in the long-tail problem and relational semantic overlap. *All* refers to testing on the whole test set. *Rare* refers to testing only on rare relations. *Overlap* refers to testing only on data with relation semantic overlap.

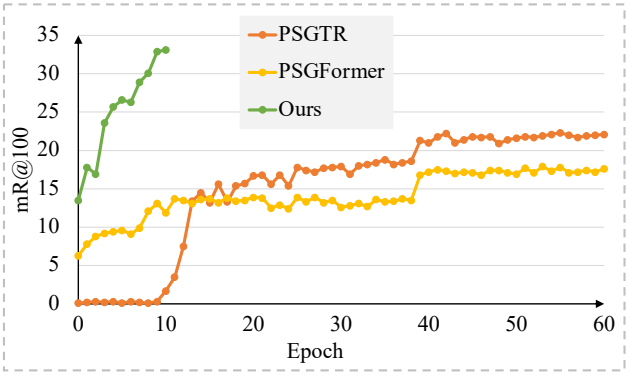

Figure 4. Convergence speed analysis of different methods. Our method converges significantly faster than previous methods.

4.6. Analysis

To further demonstrate the efficacy of our method, we conduct the following analysis on HiLo.

Long-tail problem and relational semantic overlap. To verify whether the problems of long-tail and relational semantic overlap have been tackled, we conduct experiments shown in Tab. 10. For the *long-tail problem*, we consider relations appearing less than 500 times (28 relations) in the PSG dataset (relations appear on average 4787 times) as rare relations, and report their mR@100. There is a 14% improvement for rare relations and an 11% improvement for all relations. This suggests that our method addresses the long-tail problem.

For the *relational semantic overlap*, we select all test images that have this problem (927 images) and report their mR@100. Our method shows a 15.3% improvement over PSGTR on images with semantic overlap relations, exceeding the overall improvement across all test images. This suggests that our approach addresses the problem of relational semantic overlap.

Convergence speed and time cost analysis. We evaluate the convergence of our model by assessing its performance on the validation set at various epochs, as illustrated in Fig. 4. Our analysis reveals that our proposed method outperforms prior methods [62] both in terms of final performance and convergence speed. Specifically, PSGTR [62] achieves negligible performance in the initial 12 epochs, requiring 60 epochs to converge, as per the authors. In contrast, our HiLo method achieves better results in just 12 epochs, indicating its superior convergence speed.

For time cost, we primarily analyze relation augmentation and swapping. 1) *Relation augmentation*, inspired by IETrans [66], involves training a baseline model and then using it to predict relation labels so as to augment the original relation labels. For the PSG dataset, following our experimental setup (see Sec. 4.3), this takes 18 hours. Afterwards, the training of our HiLo model requires 18 hours, which makes the whole process 36 hours. In contrast, the PSGTR model training takes 48 hours. Both our method and PSGTR utilize ResNet-50 as the backbone for fair comparison. Our method's rapid convergence (see Sec. 4.5) reduces training time compared to PSGTR, making the additional time cost for relation augmentation tolerable. 2) *Relation swapping* is a quick operation on relation labels during training and does not significantly contribute to overall time consumption.

Method	GFlops	Param. (M)	Train Mem. (G)	Infer. Time (ms)
PSGTR	461.3	44.2	26.5	140
HiLo (ours)	229.4	58.7	16.1	156

Table 11. Training and inference cost.

Training and inference cost. Training and inference cost is shown in Tab. 11. Despite using two transformer-based decoders, our method requires less resource. Given the same input sizes $(1280, 800)$ and ResNet-50 as backbone, the resource usage is shown in Tab. 11. PSGTR [62] generates mask features for subject and object separately, while our method reduces this computation by only generating mask features once. Our inference time is marginally higher due to more complex post-processing. Using H-L and L-H data only changes the labels between two branches, not increasing resource use.

5. Conclusion

In this work we proposed the HiLo framework to tackle the long-tail problem with relational semantic overlap in Panoptic Scene Graph generation. The HiLo framework simultaneously learns the high and low frequency relations in different network branches and unifies their strengths by aligning their predictions. We also constructed a HiLo baseline to allow high-quality panoptic segmentation to improve PSG performance. Experimental results demonstrate that our method achieves state-of-the-art performance on the PSG dataset, confirming its effectiveness. In future work, we will investigate how knowledge distillation [28, 23] can be used to fuse the high and low branches in our method, as well as its application to downstream tasks such as visual question answering and image captioning.

Acknowledgment

The authors would like to thank Prof. Tomasz Radzik for helpful discussions. Computing resources provided by King's Computational Research, Engineering and Technology Environment (CREATE). This work was supported by the European Union's Horizon 2020 FET Proactive Program under Agreement 101017857 and Fundamental Research Funds for the Central Universities.

References

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45, 2018. 1

[2] Konstantinos Panagiotis Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed instance segmentation using gumbel optimized loss. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 353–369. Springer, 2022. 1

[3] Stephan Baier, Yunpu Ma, and Volker Tresp. Improving visual relationship detection using semantic modeling of scene descriptions. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 53–68. Springer, 2017. 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[7] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4613–4623, 2019. 6

[8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9962–9971, 2020. 1

[9] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 2

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2, 3, 4, 6

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[12] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 1, 2

[13] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[14] Jiequan Cui, Yuhui Yuan, Zhisheng Zhong, Zhuotao Tian, Han Hu, Stephen Lin, and Jiaya Jia. Region rebalance for long-tailed semantic segmentation. *arXiv preprint arXiv:2204.01969*, 2022. 1

[15] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1

[16] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, pages 3076–3086, 2017. 2

[17] Youming Deng, Yansheng Li, Yongjun Zhang, Xiang Xiang, Jian Wang, Jingdong Chen, and Jiayi Ma. Hierarchical memory learning for fine-grained scene graph generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 266–283. Springer, 2022. 1, 2

[18] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021. 1, 2

[19] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 2

[20] Mohammed Haroon Dupty, Zhen Zhang, and Wee Sun Lee. Visual relationship detection with low rank non-negative tensor decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10737–10744, 2020. 2

[21] Lizhao Gao, Bo Wang, and Wenmin Wang. Image captioning with scene-graph based semantic concepts. In *Proceedings of the 2018 10th international conference on machine learning and computing*, pages 225–229, 2018. 1

[22] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 2, 4

[23] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 9

[24] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 1969–1978, 2019. 2

[25] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16383–16392, 2021. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[27] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*, 2020. 1

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 9

[29] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 1

[30] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3820–3832, 2020. 2

[31] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1023, 2018. 2

[32] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[33] Haeyong Kang and Chang D Yoo. Skew class-balanced reweighting for unbiased scene graph generation. *Machine Learning and Knowledge Extraction*, 5(1):287–303, 2023. 2

[34] Siddhesh Khandelwal, Mohammed Suhail, and Leonid Sigal. Segmentation-grounded scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15879–15889, 2021. 2

[35] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1, 3

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[37] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. 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*, pages 18869–18878, 2022. 2

[38] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19486–19496, 2022. 2

[39] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. 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*, pages 11109–11119, 2021. 1, 2

[40] Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. Ppdl: Predicate probability distribution based loss for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2022. 2

[41] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1356, 2017. 2

[42] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 2

[43] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[46] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2, 6, 7

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[49] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In

Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 852–869. Springer, 2016. 1, 2

[50] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2019. 2

[51] Mengshi Qi, Yunhong Wang, and Annan Li. Online cross-modal scene retrieval by binary representation and semantic graph. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 744–752, 2017. 1

[52] Mohammad Amin Sadeghi and Ali Farhadi. *Recognition using visual phrases*. IEEE, 2011. 2

[53] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 1

[54] Sahand Sharifzadeh, Sina Moayed Baharlou, Max Berrendorf, Rajat Koner, and Volker Tresp. Improving visual relation detection using depth maps. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3597–3604. IEEE, 2021. 2

[55] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8376–8384, 2019. 1

[56] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *European Conference on Computer Vision*, pages 422–439. Springer, 2022. 2

[57] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 3

[58] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[59] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 2, 7

[60] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 2, 6, 7

[61] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 2, 6

[62] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 1, 2, 4, 6, 7, 9

[63] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1274–1280. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 2, 6, 7

[64] Xu Chen Ya Zhang Xiao Gu Yue Hu, Siheng Chen. Neural message passing for visual relationship detection. In *ICML Workshop on Learning and Reasoning with Graph-Structured Representations*, Long Beach, CA, June 2019. 2

[65] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 2, 6, 7

[66] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 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*, pages 409–424. Springer, 2022. 2, 4, 7, 9

[67] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 2

[68] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9185–9194, 2019. 2

[69] Sipeng Zheng, Shizhe Chen, and Qin Jin. Visual relation detection with multi-level attention. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 121–129, 2019. 2

[70] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*, 2022. 2

[71] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 3

[72] Yaohui Zhu, Shuqiang Jiang, and Xiangyang Li. Visual relationship detection with object spatial distribution. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 379–384. IEEE, 2017. 2