

# Panoptic Scene Graph Generation with Semantics-Prototype Learning

Li Li<sup>1</sup>, Wei Ji<sup>1\*</sup>, Yiming Wu<sup>2</sup>, Mengze Li<sup>3\*</sup>, You Qin<sup>1</sup>, Lina Wei<sup>4</sup>, Roger Zimmermann<sup>1</sup>

<sup>1</sup>National University of Singapore

<sup>2</sup>The University of Sydney

<sup>3</sup>Zhejiang University

<sup>4</sup>Hangzhou City University

lili02@u.nus.edu, weiji0523@gmail.com, mengzeli@zju.edu.cn, e0962995@u.nus.edu

## Abstract

Panoptic Scene Graph Generation (PSG) parses objects and predicts their relationships (predicate) to connect human language and visual scenes. However, different language preferences of annotators and semantic overlaps between predicates lead to biased predicate annotations in the dataset, i.e. different predicates for the same object pairs. Biased predicate annotations make PSG models struggle in constructing a clear decision plane among predicates, which greatly hinders the real application of PSG models. To address the intrinsic bias above, we propose a novel framework named ADTrans to adaptively transfer biased predicate annotations to informative and unified ones. To promise consistency and accuracy during the transfer process, we propose to observe the invariance degree of representations in each predicate class, and learn unbiased prototypes of predicates with different intensities. Meanwhile, we continuously measure the distribution changes between each presentation and its prototype, and constantly screen potentially biased data. Finally, with the unbiased predicate-prototype representation embedding space, biased annotations are easily identified. Experiments show that ADTrans significantly improves the performance of benchmark models, achieving a new state-of-the-art performance, and shows great generalization and effectiveness on multiple datasets. Our code is released at <https://github.com/lili0415/PSG-biased-annotation>.

## Introduction

Panoptic Scene Graph Generation (PSG) (Yang et al. 2022) aims to simultaneously detect instances and their relationships within visual scenes (Chang et al. 2023). Instead of coarse bounding boxes used in Scene Graph Generation (SGG) (Li et al. 2021b; Xu et al. 2017; Lin et al. 2020; Li et al. 2021a; Chen et al. 2020b; Yao et al. 2022, 2021), PSG proposed to construct more comprehensive scene graphs with panoptic segmentation (Kirillov et al. 2019). PSG methods have the potential to bridge the gap between visual scenes and human languages and thus has the ability to contribute to related vision-language tasks, such as image retrieval (Johnson et al. 2015; Wang et al. 2022a,b; Lv et al. 2023), image captioning (Chen et al. 2020a; Wei et al. 2022),

\*Corresponding Author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

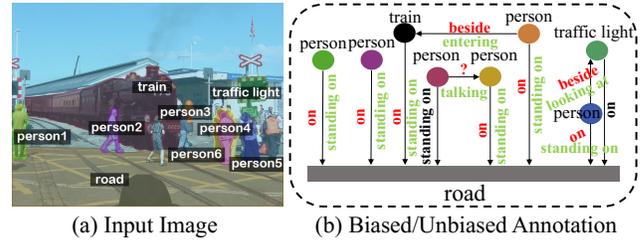


Figure 1: (a) Exemplar panoptic segmentation results of an input image. (b) present annotation transfer process. Our proposed method promotes the original dataset (annotations in red) by identifying biased annotation and potentially positive samples, and then adaptively and accurately transferring them to target triplet pairs (annotation in green).

and visual question answering (Teney, Liu, and van den Hengel 2017; Li et al. 2022b; Antol et al. 2015; Xiao et al. 2022; Fang et al. 2023).

However, PSG methods currently suffer from suboptimal performance due to biased and noisy information in generated scene graphs, stemming from the problem of *biased annotations*.

Exploring the inference mechanism of PSG and SGG models, it is translating visual scenes to linguistic descriptions, i.e., mapping visual instances to subjects/objects, and their relationships to predicates. We regard the problem above as the semantic ambiguity between predicates, and the contradictory mappings from visual to linguistics. There are a lot of semantic overlaps and hierarchical relationships among predicates, e.g., the superclass predicate *on* for its subclass predicate *standing on*. Because of the semantic overlaps and the inconsistent preferences of annotators, contradictory mappings from visual to linguistics unavoidably exist in the training dataset, and deteriorate the long-tail distribution problem of the training dataset. As shown in Fig. 1, with the difficulty of applying a unified standard for annotation, annotators tend to annotate general predicate labels (e.g., “on” and “beside”) for simplicity instead of informative ones (e.g., “standing on” and “looking at”). As a result, models cannot learn a consistent mapping from visual to semantics, instead, they entangle predicate labels and the prior knowledge of long-tail distribution, leading to serious harm

for model training stage.

Previous works (Zellers et al. 2018; Yu et al. 2020; Yao et al. 2022; Xu et al. 2017; Tang et al. 2019, 2020; Lin et al. 2020) exploit numerous model architectures to alleviate the bias problem, but these models trained by biased datasets achieve relatively limited performances, and cannot fundamentally solve the problem. Zhang et al. (2022) have proposed to enhance the training dataset by a data transfer framework, which transfers head predicate labeled samples to tail predicate labeled ones. However, their framework inaccurately transfers a significant number of samples, leading to imbalanced performance among predicates.

To alleviate the biased annotation problem, we propose to construct a promised and reasonable dataset, which includes plentiful samples with consistent predicate annotation. Specifically, there exist two types of predicates that need to be refined: indistinguishable triplet pairs with semantic ambiguity and potentially positive samples missed by annotators (Zhang et al. 2022). We transfer these two kinds of biased annotations to high-quality consistent predicate annotations (Zhang et al. 2023; Lv et al. 2022).

We introduce a new adaptive data transfer framework named ADTrans for PSG. Our framework emphasizes consistency during the data transfer process, and performs the data transfer process adaptively and accurately. Besides the prior knowledge of dataset distribution, we believe textual information alignment helps in building consistency during the data transfer process. A general way is leveraging language models to extract semantic embeddings, however, words embedding vectors generated by language models often have high similarity because of the broad class intersection of the open world, leading to the misalignment of the textual domain and the relationship domain. Thus, we propose a prototype-based predicate representation learning method. Unbiased predicate representations are expected to share invariant features within each predicate class. Thus, we employ contrastive learning to increase intra-class cohesion and inter-class separation, while focusing more on hard samples (visually similar predicates). Meanwhile, we observe the invariance degree of representations in each predicate class, and learn predicate prototypes with dynamic intensities. We continuously measure the distribution changes between each presentation and its prototype, and constantly screen potentially biased data. Finally, with the unbiased predicate representation embedding space, biased annotations are easily identified and transferred.

In summary, the following contributions are made in this paper:

- A novel, plug-and-play framework named ADTrans is proposed, which aims at adaptively and accurately performing data transfer to promise a reasonable dataset with informative and standard-unified labels, and more solid training samples.
- We propose a new prototype-based predicate representation learning method, aiming at a reasonable information alignment process between the textual domain and the relationship domain, to promise consistency during the data transfer process.

- Comprehensive experiments demonstrate that the proposed method shows validity on two datasets, and significantly enhances the performance of benchmark models, achieving new state-of-the-art performances.

## Related Work

**Panoptic Scene Graph Generation.** SGG has gained increasing attention from the computer vision community for its promising future in high-level vision-language tasks (Teney, Liu, and van den Hengel 2017; Li et al. 2022b; Antol et al. 2015; Chen et al. 2020a; Johnson et al. 2015). Early two-stage methods divide the whole task into objects locating process and relationships prediction process, and struggle for a better feature extraction network (Xu et al. 2017; Zellers et al. 2018; Tang et al. 2019; Lin et al. 2020). More recently, a novel task named Panoptic Scene Graph Generation (PSG) (Yang et al. 2022), which points out that it will contain noisy and ambiguous pixels if only coarse bounding boxes are provided, and aims at constructing more comprehensive scene graphs with panoptic segmentation rather than coarse bounding boxes. In addition, the provided ground truth panoptic segmentation can also significantly promote the performance of even the most classic SGG method, IMP (Xu et al. 2017).

**Towards Debiasing Scene Graph Generation.** (Zellers et al. 2018) first introduces the biased prediction problem in SGG. (Tang et al. 2019) and (Chen et al. 2019) provide a more reasonable metric (Mean Recall) aiming at calculating the recall of each predicate label independently. To directly face the problem, causal inference framework (Tang et al. 2020; Zhang and An 2022) is applied to alleviate data bias during the inference process, and CogTree (Yu et al. 2020) is designed to train models with the ability to make informative predictions on predicate labels. More recently, (Zhang et al. 2022; Li et al. 2023a) argues that performance could be promoted if there is a reasonable and sound dataset. However, the data transfer method in (Zhang et al. 2022) is efficient but so rigid and inflexible that a huge number of positive samples will be wrongly transferred due to its simple and coarse transfer method, which will lead to a remarkable decline in recall rate of head predicate labels. In this paper, we provide a new approach for SGG and PSG datasets debiasing.

## Method

In this section, we first introduce the biased annotation identifying method. Then we introduce relation representation extraction. After that, we provide the detailed semantics-prototype learning method. Finally, the data transfer method and resampling method are introduced.

**Target Identifying.** Following Zhang et al. (2022), we identify indistinguishable triplet pairs by checking the inconsistency between the model’s predictions and the ground truth labels. To be more specific, we use a pre-trained model (e.g. VCTree(Tang et al. 2019)) to predict predicate labels for every pair of ground truth subject and object pairs in the PSG training dataset and identify possible indistinguishable predicate labels. For potentially positive samples, we also use a pre-trained model to predict predicate labels for every pair of

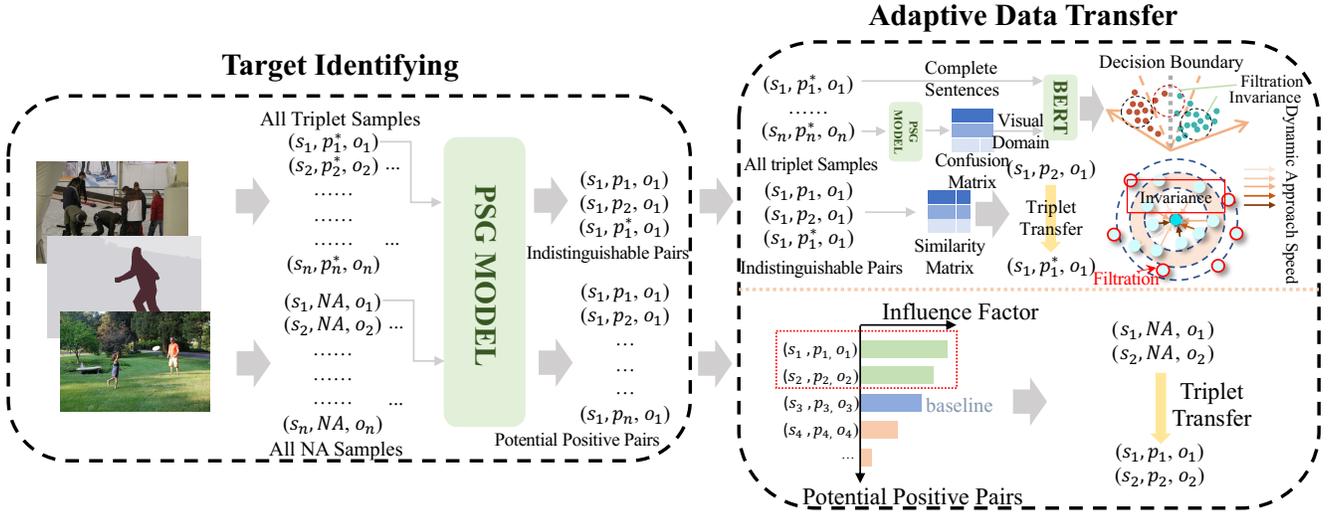


Figure 2: Illustration of the overall pipeline. It learns unbiased semantics-prototypes and the learned prototypes help to promise the consistency during data transfer process.

ground truth subject and object labels that have not yet been annotated with predicate labels, also known as NA samples.

### Relation Representation Extraction

To make the language model more sensitive to predicate semantics, we fine-tune the language model with contrastive relation representation training (Li et al. 2022a).

**Robust Contrastive Training.** We first collect all of the triplets that appear in the training set. Each triplet will be converted to a sentence for language model processing. For example,  $\langle$  person, standing on, snow  $\rangle$  will be converted to *The person is standing on the snow.*

Formally, given a sentence  $s_i$  with predicate  $p_i$  in the batch  $S = \{s_k\}_{i=k}^N$ , we can construct its positive set  $PS_i = \{s_k | p_i = p_k\}_{k \neq i}$  and negative set  $PN_i = \{s_k | p_i \neq p_k\}_{k \neq i}$ . With the training data, we use a InfoNCE loss to optimize the language model:

$$L_{lm} = \sum_{i=1}^N -\log \frac{f_{pos}}{f_{pos} + f_{neg}}, \quad (1)$$

where:

$$f_{pos} = \sum_{s_j \in PS_i} e^{\sin(h_i, h_j)/T}, \quad (2)$$

$$f_{neg} = \sum_{s_g \in PN_i} e^{\sin(h_i, h_g)/T}. \quad (3)$$

$T$  is a temperature hyper-parameter,  $N$  is batch size,  $h_{i,j,g}$  are language model generated sentence representations for  $s_{i,j,g}$ , and  $sim(h_i, h_j)$  is the cosine similarity.

To further boost the sensitivity to predicate similarity, we additionally introduce an angular margin  $m$  for the positive pairs. Formally, instead of using the previously defined  $e^{\sin(h_i, h_j)/T}$ , we use:

$$e^{\sin(h_i, h_j)/T} \rightarrow e^{\cos(\theta_{i,j}+m)/T}, \quad (4)$$

where  $\theta_{i,j}$  is the arc-cosine similarity between  $i$  and  $j$ . Thus,  $f_{pos}$  becomes:

$$f_{pos} = \sum_{s_j \in PS_i} e^{\cos(\theta_{i,j}+m)/T}. \quad (5)$$

**Alignment with Visual Domain.** There is typically a domain gap between the visual domain and the textual domain, i.e., the textual similar predicates are not visually similar. To align the language model with the visual domain (Li et al. 2023b; Fang et al. 2022), we try to incorporate some visual prior knowledge into the language model fine-tuning.

Specifically, we take advantage of the confusion matrix  $C \in \mathbb{R}^{Q \times Q}$  generated by a pre-trained VCTree model, where  $Q$  is the number of predicates in the dataset, and  $C_{i,j}$  denotes the averaged prediction score for predicate  $p_j$  on all examples annotated with  $p_i$ . When  $C_{i,j}$  is near to 1 (very high),  $p_i$  and  $p_j$  are visually similar and when  $C_{i,j}$  is near to 0 (very low),  $p_i$  and  $p_j$  are visually different.

For the language model training, we expect the language model to be aligned with the visual domain similarity judgment. Thus, the language model should also distinguish the visually different predicate pairs but avoid the visually similar predicate pairs from being too distant in the feature space. The metric  $1 - C_{i,j}$  can satisfy our requirement.

Formally, we add  $1 - C_{i,j}$  as a weight for negative pairs:

$$e^{\cos(\theta_{i,j})/T} \rightarrow (1 - C_{i,j}) \times e^{\cos(\theta_{i,j})/T}. \quad (6)$$

Then, the  $f_{neg}$  becomes:

$$f_{neg} = \sum_{s_g \in PN_i} (1 - C_{i,g}) e^{\cos(\theta_{i,g})/T}. \quad (7)$$

**Invariant Representation Exploration.** We propose the invariant representation exploration to further promise the unbiased representation of predicates.

Specifically, we say that an unbiased representation  $\Phi : X \rightarrow H$  elicits an invariant predictor  $\omega$  across positive set  $\mathcal{E}$

if there is an optimizer  $\omega : H \rightarrow Y$  simultaneously optimal for all samples from the positive set. The learning objective can be formulated as:

$$\omega \in \operatorname{argmin}_{\omega: H \rightarrow Y} O^e(\omega, \Phi). \quad (8)$$

Eq. 8 tries to learn a feature representation from  $\Phi(\cdot)$  that can induce an optimizer  $\omega(\cdot)$  which is simultaneously optimal for all  $e \in \varepsilon$ . Thus, we propose the invariant representation regularization which can be formulated as:

$$L_{irm} = \sum_{i=1}^N \lambda \operatorname{Var}(L^i), \quad (9)$$

where  $L^i = \{L_{lm}(j) | p_i = p_j\}$  denotes the loss values of samples from the positive set, and  $\lambda$  is a hyper-parameter. The minimization of variances of loss values encourages unbiased representation learning for each predicate class (Arjovsky et al. 2020).

### Semantics-prototype Learning

The extracted relation representation may be suffered from possibly biased annotations. Thus, we utilize dynamic prototype learning to maximize the discriminative power between predicate classes. We measure the invariance within each predicate class, and discard biased data by stage. This promise unbiased predicate representation embedding space for accurate data transfer.

**Dynamic Prototype Updating.** To build an unbiased embedding space for predicates, we further propose to construct the prototype space for predicates. Specifically, we specify the total prototype space  $P_{type} \in \mathbf{R}^{L \times Q}$ , where  $L$  is the same as the size of semantic embedding. Then we dynamically update the prototype space depending on the degree of invariance during the robust contrastive training process.

Given a batch  $S$ , we can construct multiple positive sets  $PS = \{ps_1, ps_2, \dots, ps_P\}$  with different predicates, where  $P$  is the number of different predicates in the batch  $S$ .

For every positive set in the  $PS$  with predicate  $p_i$ , we obtain its average predicate representation embedding as follows:

$$H_{aver}^{p_i} = \frac{1}{N^{p_i}} \sum_{s=1}^{N^{p_i}} h_i, \quad (10)$$

where  $N^{p_i}$  denotes the number of samples with predicate  $p_i$  in the batch  $S$ , and  $h_i$  denotes the predicate representation embedding. We average the summation of all predicate representations with the same predicate in the batch to get the average feature embedding.

With the help of observed invariant representation, we update the prototype space for predicate  $p_i$  with a moving average approach:

$$P_{type}^{p_i} = \beta P_{type}^{p_i} + (1 - \beta) \underbrace{\frac{1}{\gamma \operatorname{Var}(L^i) N^{p_i}}}_{\text{Approach Speed}} (H_{aver}^{p_i} - P_{type}^{p_i}), \quad (11)$$

where  $\beta$  and  $\gamma$  are hyper-parameters.

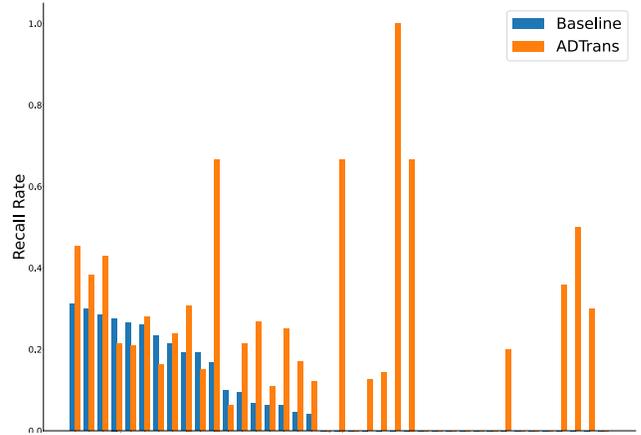


Figure 3: R@100 for predicates under SGDet task among plain PSGTR, and PSGTR with ADTrans. ADTrans achieves more balanced and effective predicate discrimination among predicates with different frequencies than plain PSGTR (The horizontal axis, moving from left to right, illustrates predicates arranged in order of high frequency to low frequency).

**Multistage Data Filtration.** Biased and noisy samples in the training dataset are certain to influence the unbiased predicate representation learning process. Thus, We design the multistage data filtration to multistage-ly filter out these bad samples. Specifically, we take advantage of the invariant representation regularization and the sample-prototype distribution shift as the measurements for the sample’s quality.

For every training epoch, we collect  $V \in \mathbf{R}^G$  from Eq. 9, which denotes the variance of loss values of every training sample in the training dataset with  $G$  samples. Then we average the collected variances on predicate labels, getting  $V_{aver} \in \mathbf{R}^Q$ . For every sample  $S_i$  with predicate label  $p_i$  and variance  $V_i$  in the training dataset, we judge whether it is part of potentially biased and noisy samples, which can be formulated as:

$$P_{bn} = \{S_i | V_i > \mu V_{aver}^i (H_{aver}^{p_i} - P_{type}^{p_i})\}, \quad (12)$$

where  $V_{aver}^i$  denotes the averaged variance on predicate label  $p_i$ , and  $\mu$  is a hyper-parameter. We further sort  $P_{bn}$  by the loss value derived from Eq. 1 and drop out the top  $D\%$  of training data. If there are fewer than 100 samples in a predicate class, we do not drop out any more samples from it.

The multistage data filtration avoids the influence of a large number of biased annotations (outlier noise). Thus, the whole unbiased predicate representation learning process promises the unbiased representation of predicates.

### Data Transfer

As a result, a similarity matrix  $S \in \mathbf{R}^{Q \times Q}$  can be generated by calculating the cosine similarities between all prototypes.

For indistinguishable triplets, we directly use the similarity score as an adaptive transfer ratio.

Method		Scene Graph Generation								
		R@20	R@50	R@100	mR@20	mR@50	mR@100	PR@20	PR@50	PR@100
Two-Stage	IMP (Xu et al. 2017)	16.5	18.2	18.6	6.52	7.05	7.23	12.9	13.7	13.9
	+IETrans (Zhang et al. 2022)	14.5	15.9	16.4	10.2	11.0	11.3	14.5	15.4	15.7
	+ADTrans	15.0	16.5	17.0	<b>12.5</b>	<b>13.5</b>	<b>14.0</b>	<b>16.0</b>	<b>17.1</b>	<b>17.5</b>
	VCTree (Tang et al. 2019)	20.6	22.1	22.5	9.70	10.2	10.2	16.0	16.8	16.9
	+IETrans (Zhang et al. 2022)	17.5	18.9	19.3	17.1	18.0	18.1	19.6	20.5	20.7
	+ADTrans	17.9	19.5	19.9	<b>18.0</b>	<b>18.9</b>	<b>19.0</b>	<b>20.2</b>	<b>21.2</b>	<b>21.4</b>
	MOTIFS (Zellers et al. 2018)	20.0	21.7	22.0	9.10	9.57	9.69	15.5	16.3	16.5
	+IETrans (Zhang et al. 2022)	16.7	18.3	18.8	15.3	16.5	16.7	18.2	19.4	19.7
	+ADTrans	17.1	18.6	19.0	<b>17.1</b>	<b>18.0</b>	<b>18.5</b>	<b>19.4</b>	<b>20.4</b>	<b>20.8</b>
	GPSnet (Lin et al. 2020)	17.8	19.6	20.1	7.03	7.49	7.67	13.6	14.4	14.7
+IETrans (Zhang et al. 2022)	14.6	16.0	16.7	11.5	12.3	12.4	15.3	16.2	16.5	
+ADTrans	17.8	19.2	19.5	<b>16.5</b>	<b>17.5</b>	<b>17.6</b>	<b>19.3</b>	<b>20.3</b>	<b>20.5</b>	
One-Stage	PSGTR (Yang et al. 2022)	28.4	34.4	36.3	16.6	20.8	22.1	21.9	26.3	27.6
	+IETrans (Zhang et al. 2022)	25.3	28.8	29.2	23.1	27.2	27.5	24.9	28.4	28.7
	+ADTrans	26.0	29.6	30.0	<b>26.4</b>	<b>29.7</b>	<b>30.0</b>	<b>27.1</b>	<b>30.2</b>	<b>30.5</b>

Table 1: The results (R@K, mR@K and PR@K) on SGDet task of our method and other baselines on PSG dataset. IETrans and ADTrans denote models equipped with different dataset-enhancement methods.

For potentially positive samples, we further define an influence factor, where the intuition is to transfer more data for the scarce relation triplets with low NA scores.

Formally, the influence factor is:

$$E_{(s_i, p_i, o_i)} = \sqrt{-\log(NA_{score}) \times c_{(s_i, o_i)} \times c_{p_i}}, \quad (13)$$

where  $-\log(NA_{score})$  is the NA score,  $c_{(s_i, o_i)}$  is the scarcity of triplets with subject  $s_i$  and  $o_i$ , and  $c_{p_i}$  is the scarcity of triplets with predicate  $p_i$ .

In practice, we further normalize  $c_{(s_i, o_i)}$  with softmax.

To judge whether a NA sample should be transferred, we rank all potential target triplet pairs according to their influence factors, and conduct the transfer of top  $K_g\%$  pairs.

## Resampling

Without conflicts, we can directly integrate the transferred indistinguishable pairs and potentially positive pairs. Furthermore, a special re-sampling method is introduced on the integrated dataset to enhance it further. We propose a new repeat factor for the task. For every triplet  $(s_i, p_i, o_i)$  in each image, we calculate its repeat factor as:

$$R = \max(1, t \times c_{(s_i, o_i)} \times c_{(p_i)}), \quad (14)$$

where  $t$  is a hyper-parameter controlling the possible repeat times. The maximum value of the repeat factor within each image is then selected.

## Experiment

### Dataset and Evaluation Metrics

**Dataset.** We evaluate our method on Visual Genome (Krishna et al. 2017) and PSG dataset (Yang et al. 2022).

**Evaluation Metric.** Following previous works (Tang et al. 2020; Zellers et al. 2018), we take recall@K (R@K) and mean recall@K (mR@K) (Tang et al. 2019; Chen et al. 2019)

Method	Predicate Classification					
	mR@20	@50	@100	F@20	@50	@100
MOTIFS	11.7	15.2	16.2	19.5	24.5	26.0
+ADTrans	<b>29.0</b>	<b>36.2</b>	<b>38.8</b>	<b>36.1</b>	<b>41.7</b>	<b>43.5</b>
VCTree	14.0	16.3	17.7	22.7	26.0	28.0
+ADTrans	<b>30.0</b>	<b>32.9</b>	<b>35.5</b>	<b>37.2</b>	<b>40.5</b>	<b>42.5</b>
GPSnet	13.2	15.0	16.0	21.7	24.4	25.8
+ADTrans	<b>27.3</b>	<b>32.0</b>	<b>34.7</b>	<b>34.8</b>	<b>40.2</b>	<b>42.1</b>

Table 2: The results (mR@K and F@K) on PREDCLS task of our method and other baselines on VG dataset.

as evaluation metrics. We also adopt a new evaluation metrics named percentile recall (PR), which can be formulated as  $PR = 30\%R + 60\%mR + 10\%PQ$ , where PQ measures the quality of a predicted panoptic segmentation relative to the ground truth (Kirillov et al. 2019). For conventional SGG tasks on the VG dataset, we also adopt an overall metric F@K (Zhang et al. 2022), which is the harmonic average of R@K and mR@K.

## Tasks and Implementation Details

**Tasks.** We evaluate our method on three classic SGG tasks: Predicate Classification (PREDCLS), Scene Graph Classification (SGCLS) and Scene Graph Generation (SGDET).

**Implementation Details.** For the pre-trained language model, we use pre-trained BERT-base (Devlin et al. 2019). The decision margin  $m$  is set to 10 degrees, the temperature hyper-parameter  $T$  is set to 0.05, and we use an AdamW (Loshchilov and Hutter 2017) optimizer with a learning rate  $2e-5$ . The hyper-parameter  $\lambda$  is set to 0.3,  $\beta$  is set to  $5e5$ , and  $\gamma$  is set to 1.5. The  $D$  in multistage data filtration is set to 50. For NA sample transfer, the value of  $K_g$  is set to 0.05, and for the re-sampling process, the value of  $t$  is set to  $3e7$ .

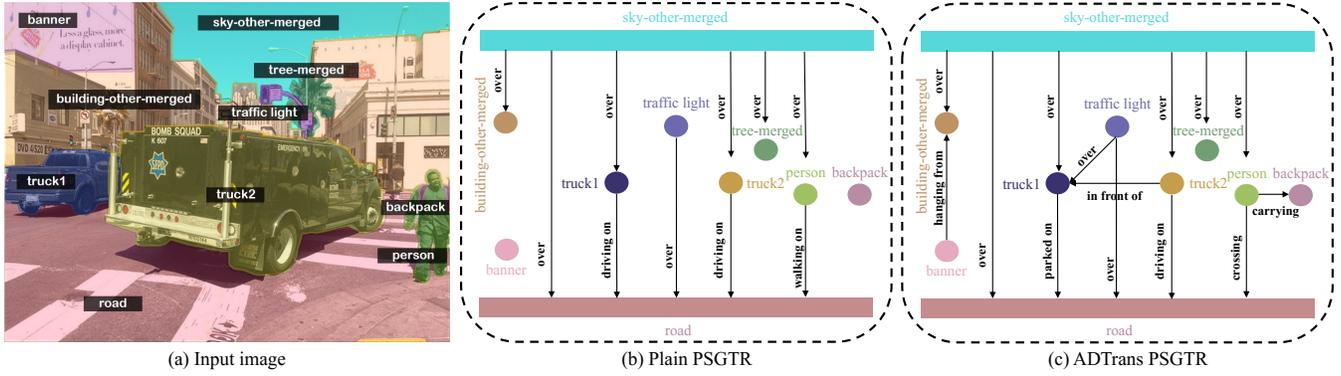


Figure 4: Visualization of plain PSGTR model and PSGTR equipped with our ADTrans. PSGTR with ADTrans can predict relationships between instances with greater accuracy and also select predicates that better match the visual scene.

Method	Scene Graph Classification					
	mR@20	@50	@100	F@20	@50	@100
MOTIFS	6.0	8.0	8.5	10.1	13.1	13.8
+ADTrans	<b>14.8</b>	<b>17.0</b>	<b>17.8</b>	<b>20.2</b>	<b>22.5</b>	<b>23.7</b>
VCtree	6.3	7.5	8.0	10.7	12.5	13.3
+ADTrans	<b>16.0</b>	<b>19.0</b>	<b>19.8</b>	<b>20.3</b>	<b>23.7</b>	<b>24.5</b>
GPSnet	10.0	11.8	12.6	15.7	17.9	18.9
+ADTrans	<b>15.5</b>	<b>18.2</b>	<b>18.8</b>	<b>19.9</b>	<b>22.5</b>	<b>23.7</b>

Table 3: The results (mR@K and F@K) on SGCLS task of our method and other baselines on VG dataset.

### Qualitative Analysis

As shown in Figure 4, we can compare results predicted by plain PSGTR and PSGTR equipped with our ADTrans. Obviously, PSGTR with ADTrans can predict more accurate relationships between instances and also predict predicate that better fits the scene. We believe our method helps construct a more comprehensive scene graph.

### Comparison with State-of-the-Art Methods

In this section, we report the results for ADTrans on different datasets, tasks, and baseline methods. All models use ResNet-50 (He et al. 2016) as their backbones.

**Effectiveness of ADTrans.** From the results, we observe that our method can effectively improve the performance of baseline networks in nearly all metrics. Fig. 3 presents a comparison between our method and plain PSGTR of the detailed recall@100 of SGDet task on part of predicate classes. ADTrans performs better than plain PSGTR on almost all the above predicate classes. When compared to IETrans (Zhang et al. 2022), our method shows significant improvements in both recall and mean recall on all baseline models, indicating that our method can enhance the training dataset more effectively, avoiding noisy or redundant transfer processes. When it comes to PR, which takes into account both recall and mean recall, our method outperforms all the original models by significant margins. This suggests that our method not only improves the recall of the models but also balances the performance across different predicate

Method	Scene Graph Generation					
	mR@20	@50	@100	F@20	@50	@100
MOTIFS	4.8	6.2	7.1	8.0	10.3	11.8
+ADTrans	<b>10.6</b>	<b>15.5</b>	<b>18.1</b>	<b>13.4</b>	<b>18.9</b>	<b>22.0</b>
VCtree	5.2	6.7	7.9	8.7	11.0	13.0
+ADTrans	<b>9.7</b>	<b>12.5</b>	<b>16.9</b>	<b>12.2</b>	<b>16.3</b>	<b>20.3</b>
GPSnet	5.2	5.9	7.1	8.6	9.9	11.8
+ADTrans	<b>12.3</b>	<b>15.8</b>	<b>19.2</b>	<b>15.1</b>	<b>18.6</b>	<b>21.9</b>

Table 4: The results (mR@K and F@K) on SGDET task of our method and other baselines on VG dataset.

Module			Scene Graph Generation		
IT	PT	RE	R/mR@20	R/mR@50	R/mR@100
✗	✗	✗	28.4 / 16.6	34.4 / 20.8	36.3 / 22.1
✓	✗	✗	26.2 / 24.9	30.3 / 28.2	30.7 / 29.2
✓	✓	✗	25.5 / 25.6	29.2 / 29.1	29.7 / 29.6
✓	✓	✓	26.0 / 26.4	29.6 / 29.7	30.0 / 30.0

Table 5: Ablation study on model components. IT: Indistinguishable Triplet Transfer; PT: Potential Positive Transfer; RE: re-sampling.

labels, resulting in a more comprehensive evaluation of the models' performance.

**Expansibility of ADTrans.** Applied with our ADTrans, the performances of all baseline models trained on all the datasets on all the tasks are greatly improved. With the observation that all the baseline models trained on the VG dataset have poorer performances compared to the same baseline models trained on the PSG dataset, VG dataset is more challenging for our ADTrans due to more biased annotations. Our method shows great expansibility on VG dataset, baseline models trained on the dataset achieve competitive performances on all the tasks.

### Ablation Studies

**Different components in ADTrans framework.** We evaluate the importance of each component in our ADTrans. As

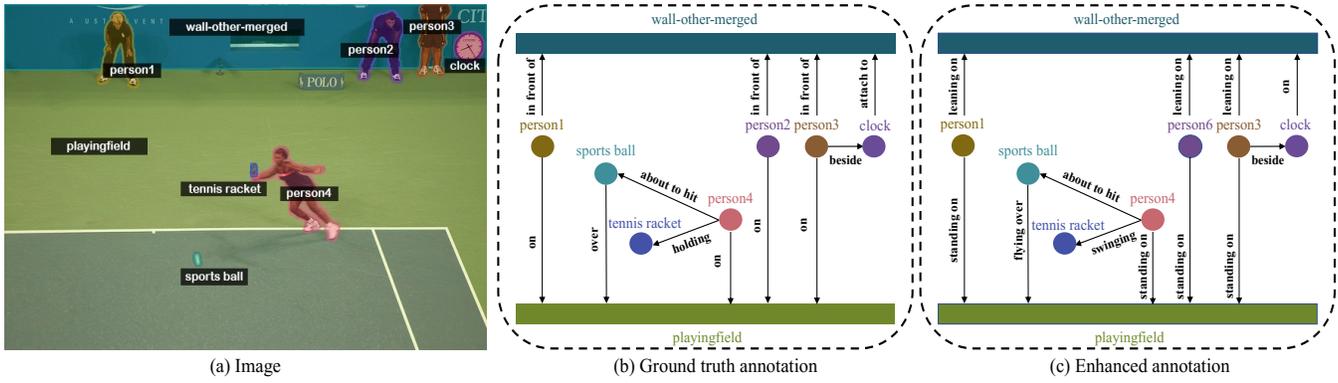


Figure 5: Visualization of the original dataset and a new dataset enhanced by our ADTRans. For the same image, we visualize its original biased annotations, and new annotations enhanced by our method. The enhanced dataset shows more informative annotations than the original one. These informative annotations promise the consistent training of models.

Data Processing Method	SGDet		
	mR@20	mR@50	mR@100
Original	16.6	20.8	22.1
Remove	20.0	24.6	25.3
Triplet Transfer	24.9	28.2	29.2

Table 6: Ablation study on data processing methods. Triplet Transfer: Indistinguishable triplet transfer. Remove: simply removing all indistinguishable triplet pairs. Original: baseline method on the original dataset.

Contrastive Learning Method	SGDet		
	mR@20	mR@50	mR@100
Original	16.6	20.8	22.1
InfoNCE	21.0	25.8	26.3
RCT	26.4	29.7	30.0

Table 7: Ablation study on contrastive learning methods. InfoNCE: the classic contrastive learning method. Original: the performance of original PSGTR. RCT: the proposed robust contrastive training method.

shown in Table. 5, we incrementally add one component to the plain baseline PSGTR (Yang et al. 2022) to validate their effectiveness. The indistinguishable triplet transfer component provides the most promotion on performance. The reason is that, we transfer these inconsistent annotations to informative and consistent ones, so that models can learn a consistent mapping from visual to predicates. The potentially positive transfer component additionally provides performance promotion. Our ADTrans transfers original NA samples, which are probably missed by annotators, to informative annotations. This step provides more reasonable training samples to construct a consistent training dataset.

**Different data processing methods during triplet transfer.** To prove the effectiveness of our indistinguishable triplet transfer method and the harm of biased annotation, we design another simple method to process indistinguishable triplet pairs. Instead of adaptively transferring them to target pairs, we simply remove them from the original training dataset. As shown in Table. 6, when comparing the simply removing method and the original baseline, the simply removing method surprisingly greatly overtakes the baseline model, indicating the serious conflict resulting from biased annotation within the original dataset. These biased annotated samples can not help the model during the training process, but will make it difficult for the model to distinguish each predicate label, resulting in a sharp decline in model performance. When comparing the triplet transfer method and the simply removing method, there are also

great margins between these two methods. With the fact that our triplet transfer method greatly outperforms the simply removing method, we observe a promising textual information alignment process and the effectiveness of our adaptive triplet transfer method is proved.

**Different contrastive learning methods.** As shown in Table. 7, though we observe improvement with the help of the InfoNCE, we can promote the model to a higher level with our method. An simply InfoNCE method increase the intra-class cohesion and inter-class separability, but cannot learn unbiased predicate representations because of biased annotations. As a result, biased annotations are inconspicuous and hard to be identified in a biased predicate representation embedding space. Our method utilizes invariant representation exploration and multistage data filtration to avoid the influence of biased annotations, and performs accurate biased-annotation identification.

## Conclusion

We introduce a novel framework named ADTrans to alleviate the biased annotation problem in SGG. ADTrans transfers indistinguishable samples and potentially positive samples to promise a reasonable training dataset with more informative and standard-unified labels. Experiments demonstrate that ADTrans greatly enhances the models’ performance on two datasets, achieving a new SOTA performance.

## Acknowledgements

This research was supported in part by the National Key R&D Program of China (2022ZD0119103), in part by the National Natural Science Foundation of China (Grant No. 62006123). This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

## References

- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.
- Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2020. Invariant Risk Minimization.
- Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; and Hauptmann, A. 2023. A Comprehensive Survey of Scene Graphs: Generation and Application. *TPAMI*.
- Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020a. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In *CVPR*.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.
- Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019. Knowledge-Embedded Routing Network for Scene Graph Generation. In *CVPR*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
- Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *TMM*.
- Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *CVPR*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.
- Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image Retrieval Using Scene Graphs. In *CVPR*.
- Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollar, P. 2019. Panoptic Segmentation. In *CVPR*.
- Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV*.
- Li, L.; Wang, C.; Qin, Y.; Ji, W.; and Liang, R. 2023a. Biased-Predicate Annotation Identification via Unbiased Visual Predicate Representation. In *ACM Multimedia*.
- Li, M.; Wang, H.; Zhang, W.; Miao, J.; Zhao, Z.; Zhang, S.; Ji, W.; and Wu, F. 2023b. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *CVPR*.
- Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Zhang, W.; Miao, J.; Pu, S.; and Wu, F. 2022a. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *ACM Multimedia*.
- Li, R.; Zhang, S.; Wan, B.; and He, X. 2021a. Bipartite Graph Network With Adaptive Message Passing for Unbiased Scene Graph Generation. In *CVPR*.
- Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022b. Invariant Grounding for Video Question Answering. In *CVPR*.
- Li, Y.; Yang, X.; Shang, X.; and Chua, T.-S. 2021b. Interventional Video Relation Detection. In *ACM Multimedia*.
- Lin, X.; Ding, C.; Zeng, J.; and Tao, D. 2020. GPS-Net: Graph Property Sensing Network for Scene Graph Generation. In *CVPR*.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization.
- Lv, Z.; Wang, F.; Zhang, S.; Kuang, K.; Yang, H.; and Wu, F. 2022. Personalizing Intervened Network for Long-tailed Sequential User Behavior Modeling. *arXiv preprint arXiv:2208.09130*.
- Lv, Z.; Zhang, W.; Zhang, S.; Kuang, K.; Wang, F.; Wang, Y.; Chen, Z.; Shen, T.; Yang, H.; Ooi, B. C.; et al. 2023. DUET: A Tuning-Free Device-Cloud Collaborative Parameters Generation Framework for Efficient Device Model Generalization. In *WWW*.
- Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *CVPR*.
- Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *CVPR*.
- Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-Structured Representations for Visual Question Answering. In *CVPR*.
- Wang, C.; Huang, Y.; Liu, X.; Pei, J.; Zhang, Y.; and Yang, J. 2022a. Global in local: A convolutional transformer for SAR ATR FSL. *GRSL*.
- Wang, C.; Pei, J.; Yang, J.; Liu, X.; Huang, Y.; and Mao, D. 2022b. Recognition in Label and Discrimination in Feature: A Hierarchically Designed Lightweight Method for Limited Data in SAR ATR. *TGRS*.
- Wei, M.; Chen, L.; Ji, W.; Yue, X.; and Chua, T.-S. 2022. Rethinking the Two-Stage Framework for Grounded Situation Recognition. *AAAI*.
- Xiao, J.; Yao, A.; Liu, Z.; Li, Y.; Ji, W.; and Chua, T.-S. 2022. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. In *AAAI*.
- Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene Graph Generation by Iterative Message Passing. In *CVPR*.
- Yang, J.; Ang, Y. Z.; Guo, Z.; Zhou, K.; Zhang, W.; and Liu, Z. 2022. Panoptic Scene Graph Generation. In *ECCV*.

- Yao, Y.; Chen, Q.; Zhang, A.; Ji, W.; Liu, Z.; Chua, T.-S.; and Sun, M. 2022. PEVL: Position-enhanced Pre-training and Prompt Tuning for Vision-language Models. In *EMNLP*.
- Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models.
- Yu, J.; Chai, Y.; Hu, Y.; and Wu, Q. 2020. CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation. In *IJCAI*.
- Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *CVPR*.
- Zhang, A.; Yao, Y.; Chen, Q.; Ji, W.; Liu, Z.; Sun, M.; and Chua, T.-S. 2022. Fine-Grained Scene Graph Generation with Data Transfer. In *ECCV*.
- Zhang, R.; and An, G. 2022. Causal Property based Anti-Conflict Modeling with Hybrid Data Augmentation for Unbiased Scene Graph Generation. In *ACCV*.
- Zhang, W.; Lv, Z.; Zhou, H.; Liu, J.-W.; Li, J.; Li, M.; Tang, S.; and Zhuang, Y. 2023. Revisiting the Domain Shift and Sample Uncertainty in Multi-source Active Domain Transfer. *arXiv preprint arXiv:2311.12905*.