Semi-Supervised Clustering Framework for Fine-grained Scene Graph Generation

Jiarui Yang[1, 2, 3*], **Chuan Wang**[4, 3, 5†], **Jun Zhang**[3], **Shuyi Wu**[6], **Zhao Jinjing**[7], **Zeming Liu**[8], **Liang Yang**[9]

[1]Shanghai Key Lab of Intell. Info. Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Institute of Information Engineering, CAS
[4]School of Computer Science and Technology, Beijing JiaoTong University
[5]Guangdong Provincial Key Lab of Intell. Info. Processing & Shenzhen Key Lab of Media Security, Shenzhen University
[6]Information Research Center of Military Science, PLA Academy of Military Science
[7]National Key Laboratory of Science and Technology on Information System Security, China
[8]School of Computer Science and Engineering, Beihang University
[9]School of Artificial Intelligence, Hebei University of Technology

Abstract

Scene Graph Generation (SGG) aims to detect all objects and identify their pairwise relationships existing in the scene. Considering the substantial human labor costs, existing scene graph annotations are often sparse and biased, which result in confusion training with low-frequency predicates. In this work, we design a *Semi-Supervised Clustering framework for Scene Graph Generation* (SSC-SGG) that uses the sparse labeled data to guide the generation of effective pseudo-labels from unlabeled object pairs, thus enriching the labeled sample space, especially for low-frequency interaction samples. We approach from the perspective of clustering, reducing the problem of confirmation bias in a self-training manner. Specifically, we first enhance the model's robustness to feature extraction via prototype-based clustering, aggregating different relationship augmented features onto the same prototype. Secondly, we design a dynamic pseudo-label assignment algorithm based on a mini-batch, which adjusts the detection sensitivity to different frequency samples from the historical assignment. Finally, we conduct joint training on the pseudo-labels and the labeled data. We conduct experiments on various SGG models and achieve substantial overall performance improvements, demonstrating the effectiveness of SSC-SGG.

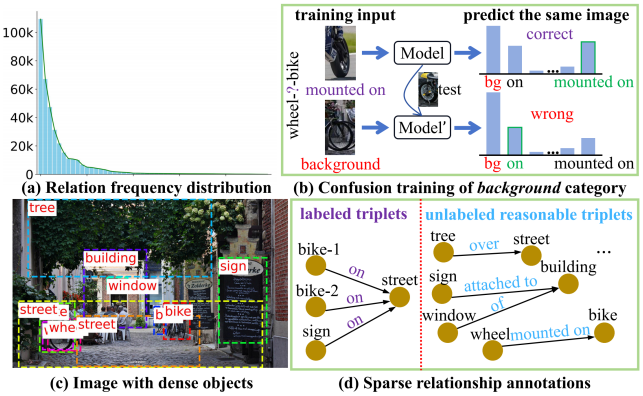

Figure 1: Confusion training caused by the long-tail distribution and the sparse annotations, thus hindering the correct prediction of low-frequency samples. (a) Relationship frequency distribution in VG (Krishna et al. 2017). (b) Due to the insufficient samples, the confidence of low-frequency relationships can be easily influenced by *background* in similar scenarios. (c) Image with dense objects. (d) Sparse annotations result in rich interaction relationships existing in unlabeled samples.

Introduction

A scene graph (SG) is a graphical structure including all of the objects and their pairwise relationships, thus providing a comprehensive and hierarchical understanding of the scene. The SGG model is characterized by interaction relationships, which are symbolically represented through triplets of the form <*Subject-Predicate-Object*>. The culmination of these relationships generates a well-defined graph structure that facilitates the solution of complex computer vision tasks, such as VQA (Zhu et al. 2020), image retrieval (Johnson et al. 2015), and image generation (Yang et al. 2022b).

Existing SGG datasets (*e.g.* VG) commonly suffer sparse and biased annotation problems. Due to the substantial human labor costs and apparent human preferences, it is impractical to develop a densely and consistently annotated SGG dataset. As shown in Fig. 1(c)(d), the image contains dense objects, but the annotator only identifies three relationships between object pairs, yet the image still contains a wealth of obvious interaction information, such as *sign-attached to-building* and *wheel-mounted on-bike*. In previous research, the common approach is to simply treat these unlabeled samples as *background* during training. In this paper, we argue that this approach has two drawbacks: 1) *Confuse the training process*. When conspicuous interaction samples are not labeled, the model is prone to predict them as *background*. This situation restricts the model from generating high-confidence predictions when faced with similar scenes, thereby affecting the final prediction results; 2) *Exacerbate biased predictions of the model*. Due to the sufficient samples, high-frequency

*Part of the work was conducted at IIE, CAS
†Corresponding author.

predicates (*e.g. on*) can easily attain high confidence, leading to a skew in the model parameters space. Conversely, with their scarcity of effective samples, low-frequency predicates (*e.g. mounted on*) are often mistakenly predicted as the related high-frequency predicates during confusion training. As shown in Fig. 1(b), *wheel-mounted on-bike* exists in the bottom image but is not labeled. Therefore, the model tends to increase its confidence in the *background* and reduce its confidence in the *mounted on* after training. Due to the long-tail distribution in SGG dataset, the high-frequency predicate *on* always maintains high confidence. When the *mounted on* is disturbed by *background* during training, the model tends to predict it as *on*, thereby exacerbating biased predictions.

The core of alleviating these two problems is to find effective interaction samples from unlabeled object pairs, especially low-frequency samples. In this paper, we propose a novel *Semi-Supervised Clustering framework for Scene Graph Generation* (SSC-SGG). The primary goal of our framework design is: *Identify and utilize unlabeled obvious low-frequency interaction samples to enhance the overall sample richness of the dataset.* Given that clustering focuses more on the underlying structure of data and is less susceptible to dataset bias, we design our framework based on clustering. Unlike previous SGG frameworks, we generate different augmentations (or "views") for each image during the training phase, which are then input into the model together for feature extraction. We introduce the prototypes, which can be trained by labeled data, to ensure the local consistency of features in multiple views and global consistency between labeled and unlabeled data. Then, we further use these prototypes to calculate the predicted logits for each view and design a dynamic pseudo-label assignment algorithm to conduct pseudo-labeling based on a mini-batch. This algorithm uses historical assignment information to enhance the detection sensitivity of low-frequency samples. Finally, we perform joint training with the soft pseudo-labels and the labeled data and employ the swap-prediction on different views to further strengthen the robustness of feature learning.

The main contributions are summarized as follows:

- We propose a semi-supervised framework for SGG, which uses labeled data to guide the model to mine the effective interaction sample from unlabeled object pairs.
- We design a method for prototype-based clustering of relationships in different views of images, which ensures the alignment of relationships across multiple views while enhancing the robustness of the model's feature learning.
- We propose a dynamic pseudo-label assignment algorithm, which uses historical assignment to enhance the model's sensitivity to the detection of low-frequency samples.
- The proposed SSC-SGG framework merely mines effective interaction samples from unlabeled object pairs, yet achieves superior overall performance on various SGG datasets, proving the importance of enhancing the overall richness of samples for SGG.

Related Work

Scene Graph Generation. SGG aims to use a comprehensive graph structure to represent a scene image. Early works pay more attention to exploring different networks, such as GNN (Li et al. 2017), CRF (Cong, Wang, and Lee 2018), and RNN/LSTM (Tang et al. 2019), to model the message-passing mechanisms between the entities and predicates. Follow-up works (Suhail et al. 2021; Lin et al. 2022a; Zheng et al. 2023) focus on extracting more powerful global contextual information. Recently, more research has shifted the attention to the severe long-tail problem of SGG datasets, like Visual Genome (Krishna et al. 2017) and Open Image (Kuznetsova et al. 2020). (Tang et al. 2020) proposes the first solution for unbiased SGG model prediction. Most works mainly utilize re-sample (Guo et al. 2021) or re-weight (Yan et al. 2020), and their variants (Min, Wu, and Deng 2023; Jiao et al. 2024, 2022, 2021, 2023) to alleviate biased prediction. (Zhang et al. 2022) proposes the first method for assigning fine-grained labels to unlabeled data. However, this label assignment method is based on t,he pre-trained model's prediction confidence for individual relationship samples, which is easily influenced by dataset bias. (Kim et al. 2024) performs pseudo-labeling in a self-training manner; however, this approach requires complex manual design, yet results in slight performance enhancements. In this work, we adopt a prototype-based clustering assignment for pseudo-label generation. , We propose a dynamic label assignment algorithm to simultaneously assign pseudo-labels to all unlabeled samples in a mini-batch. This strategy bypasses the influence of dataset bias and enhances the model's detection sensitivity to low-frequency samples, thus increasing the sample size of low-frequency predicates.

Semi-Supervised Learning. Semi-supervised methods aim to exploit a limited amount of annotations and large volumes of unlabeled data. The most intuitive approach is Pseudo-Labels (Lee et al. 2013), which continually assigns pseudo-labels to unlabeled data using a model trained on labeled data in a self-training manner, but this can lead to confirmation bias (Arazo et al. 2020). To overcome this issue, researchers have shown benefits from soft labels and confidence thresholding (Arazo et al. 2020) and designed different training strategies like model distillation (Xie et al. 2020) and consistency regularization (Sohn et al. 2020; Yang et al. 2022a) to generate more balanced and effective pseudo-labels. In this paper, we conduct prototype-based clustering to ensure the local consistency of multi-view features and global consistency between labeled and unlabeled data and assign pseudo-labels to samples through prototypes similarity.

Problem Setting and Overview

Problem Setting. The task of scene graph generation is to parse an image I into a scene graph $\mathcal{G} = \{\mathcal{E}_{sub}, \mathcal{P}, \mathcal{E}_{obj}\}$, where $\mathcal{E}_{sub}$ and $\mathcal{E}_{obj}$ represent the set of subject and object entities, respectively. $\mathcal{P} \in \mathbb{R}^{\mathcal{C}_p+1}$ denotes the set of predicates for all entity-pairs and $\mathcal{C}_p$ is the number of predicate categories with one category reserved for *background*. (For simplicity, we use $C = \mathcal{C}_p + 1$ in the following.) The mainstream SGG framework is still two-stage: first detecting objects, then predicting relationships between them. In the first stage, a pre-trained object detector with fixed parameters is typically used for proposal extraction from the image, which contains spatial, semantic, and visual information. In the next stage,

the SGG model utilizes these features to generate relationship candidates and make predictions. A common operation is to calculate the cross-entropy loss between the predicted logits and the ground truth labels:

$$\mathcal{L}_{rel} = -\frac{1}{C}\sum_{k=0}^{\mathcal{C}_p} y_k \log(\text{softmax}(f_p(\mathbf{p}))_k), \quad (1)$$

where $\mathbf{p}$ represents the predicate feature and f_p is a classifier.

In the SGG dataset, we generally default object pairs without relationship annotations to the *background*, i.e., y_0, which occupies the majority of the supervised categories. Therefore, when conducting actual predictions, we should exclude this category to predict effective predicate labels:

Overview. In this paper, our objective is to discover valuable interaction information from unlabeled object pairs, especially for low-frequency samples, thus enhancing the overall sample richness of the dataset. Due to the abundance of labeled samples, the model can train a robust feature space for high-frequency labels. In contrast, low-frequency samples, with their scarcity of annotation, are highly susceptible to interference from other predicates during the feature training process. Therefore, by providing new low-frequency samples from unlabeled samples, we enable the model to train a relatively robust sample space for them.

We draw inspiration from clustering-based self-supervised learning methods. We use prototypes trained with labeled data to cluster different relationship augmented features of all object pairs, thus enhancing the model's feature learning capability. Then a dynamic pseudo-label assignment algorithm is designed to assign pseudo-labels for the entire mini-batch from the predicted logits based on historical assignment information. Finally, we conduct joint training between pseudo-labels and labeled data. An overview of the framework is shown in Fig. 2.

Multi-View Prototype-based Clustering Framework

Scene images always contain a large number of unlabeled object pairs, brimming with effective interaction information. To distinguish them and construct connections to labeled data, one intuitive idea is to gather the same type of interaction information together. Directly using the label propagation mechanism largely relies on predicate labels, presenting over-confidence in high-frequency predicates. Based on this consideration, we start with feature clustering that mainly focuses on the inherent structure of data representation. We construct a multi-view prototype clustering framework, aiming to learn connections between labeled and unlabeled data in the feature space.

Relationship Augmentation. To fully leverage the information within the scene, data augmentation techniques have been widely employed at the input level. For scene graph generation, since the union boxes of all object pairs in a scene image contain an abundance of overlapping areas, directly augmenting each relationship separately is computationally redundant. To reduce these redundant computations, we instead augment the input image as a whole. We generate different views for an image and predict the relationships for each of them. In this way, the model only needs to extract relationship features on the corresponding feature maps, thus greatly reducing the computation. However, this brings another problem: *how do we align the multi-view relationship prediction results?* Considering the randomness of image augmentation, we only augment the image based on the pixel level, like *ColorJitter* and *GaussianBlur*. In addition, for different augmented images, a detector with fixed parameters might predict different object proposals, which influences the alignment of multi-view relationships. To address this issue, as shown in Fig. 2, we only predict the proposals of one augmented view (*e.g.* the first view). Then, we use these proposals to extract relationships on different augmented feature maps through RoI module, thereby automatically aligning the multi-view relationships. Through relationship augmentation, we obtain consistent multi-view initial relationship features. We then input these features into some common-used message-passing network (MPN), like Motifs (Zellers et al. 2018), to conduct feature refinement.

Although relationship enhancement produces doubling views, adding an extra 33% training time, this method further enhances the model robustness of feature training, thereby improving the overall performance. Additionally, it does not require multiple views during the inference phase, thus not affecting the inference speed, making it effective.

Prototype-based Clustering. Given the refined representations of relationships and partially labeled data, our goal is to exploit the interaction information for unlabeled object pairs. To achieve this, we build a set of category prototypes, which serve as cluster centroid within the feature space. These prototypes can be linked to all object pairs by considering their feature similarity and constraining the local consistency of multi-view relationship features. To construct prototypes, we simply allocate a prototype for each category, which is initialized as a word vector from the pre-trained GloVe (Pennington, Socher, and Manning 2014) with a 300-dimensional feature vector and undergoes fine-tuning with labeled data during the pre-training phase of the model:

$$\mathbf{q}_k = MLP(Embed(l_k)), \quad (2)$$

where l_k denotes the k^{th} relationship label and MLP is used to map from word vector space to relationship feature space.

In our design, the prototype essentially provides a cluster centroid for each category. To ensure the effectiveness of clustering, we adopt prototype regularization (Zheng et al. 2023) to make the cluster centroids far apart from each other in the feature space, which use $\ell_{2,1}$-norm to minimize the cosine similarity between different prototypes:

$$\mathcal{L}_p = \|\mathbf{Q}\cdot\mathbf{Q}^T\|_{2,1}, \quad (3)$$

where $\mathbf{Q} \in \mathbb{R}^{C\times d_p}$ is a ℓ_2 normalized prototype matrix.

To link samples with prototypes, we cluster the multi-view relationship features by calculating their similarity to the prototype matrix $\mathbf{Q}$. Specifically, for each view, by calculating the cosine similarity with each prototype and then conducting normalization, we obtain the predicted logits:

$$r_i^k = \frac{\exp(\|\mathbf{p}_i^T\|_2\|\mathbf{q}_k\|_2/\tau)}{\sum_{k'}\exp(\|\mathbf{p}_i^T\|_2\|\mathbf{q}_{k'}\|_2/\tau)}, \quad (4)$$

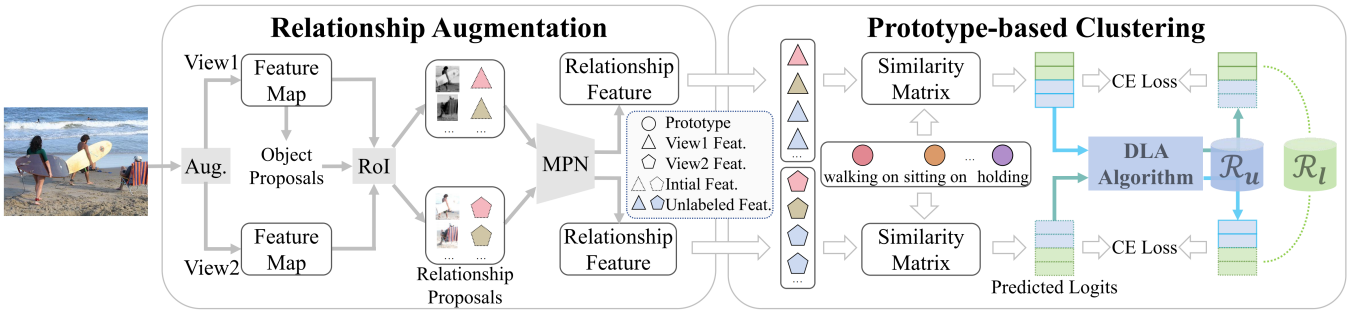

Figure 2: Overview of the proposed SSC-SGG framework.

where r_i^k denotes the logits output of the i^{th} sample on the k^{th} category, $\|\cdot\|_2$ represents ℓ_2-norm and τ is a scalar temperature parameter. Using Eq. 4, we obtain the prototype-based probability distributions of the two views:

$$x_{v1} = [r_{v1}^0, r_{v1}^1, ..., r_{v1}^{\mathbf{C}_p}], \\ x_{v2} = [r_{v2}^0, r_{v2}^1, ..., r_{v2}^{\mathbf{C}_p}]. \tag{5}$$

These multi-view predicted logits supervise the clustering process using a swap-prediction mode (Caron et al. 2020), keeping prediction probability distribution consistent of the model across different views.

Our vanilla multi-view prototype-based clustering framework improves the robustness of feature learning, and links labeled and unlabeled data in the feature space. In the next, we further utilize these robust features to mine effective interaction information from a large number of unlabeled object pairs, enhancing the diversity of samples across different categories, especially for low-frequency samples.

Dynamic Pseudo-Label Assignment

Given the predicted logits of relationships for each view from Eq. 5, a direct strategy to conduct pseudo-label assignment is to choose the category with the highest confidence. However, under the long-tail data distribution of SGG datasets, this approach inevitably exacerbates the model's preference for high-frequency predicates. A common way to alleviate this phenomenon is threshold controlling, but it usually requires tedious manual design. To automatically and reasonably select effective and balanced unlabeled interaction samples, we consider a pseudo-label assignment in a holistic way and conduct it across the entire samples in a mini-batch instead of individual ones. To guarantee a balanced assignment across samples, we impose regularization constraints.

The pseudo-label assignment is formulated as the following optimal transport problem (Cuturi 2013):

$$\underset{\mathbf{Y}\in\mathcal{U}}{\arg\max}\ Tr(\mathbf{Y}^T\mathbf{X}) + \lambda H(\mathbf{Y}), \tag{6}$$

where $\mathbf{X} \in \mathbb{R}^{N\times C}$ is a predicted logits matrix, $\mathbf{Y} \in \mathbb{R}^{N\times C}$ is exactly the optimal assignment matrix what we need to solve for. $H(\mathbf{Y}) = -\sum_{ij} \mathbf{Y}_{ij} \log \mathbf{Y}_{ij}$ is an entropy regularization term used to constrain smooth category assignment and avoid trivial solutions. $\mathcal{U}$ is the constraint conditions defined as:

$$\mathcal{U} = \{\mathbf{Y} \in \mathbb{R}_+^{N\times C} \mid \mathbf{Y}\mathbf{1}_C = \mathbf{1}_N, \mathbf{Y}^T\mathbf{1}_N = \frac{N}{C}\mathbf{1}_C\}, \tag{7}$$

Algorithm 1: Dynamic Label Assignment Algorithm

Input: $\mathbf{X} \in \mathbb{R}^{N\times C}$: predicted logits; t: iteration times;
Output: $\mathbf{Y} \in \mathbb{R}^{N\times C}$: assignment soft pseudo-labels;

$\mathbf{Y} = \exp(\frac{\mathbf{X}^\mathbf{T}}{\lambda})$
$\mathbf{Y} = \mathbf{Y}/\text{sum}(\mathbf{Y})$
set $r = \text{ones}(N)$, $c = \text{ones}(C) * \frac{N}{C}$
calculate w from Eq. 8
for i from 1 to t **do**
  $u = \mathbf{Y}.\text{sum}(1) * w$
  $\mathbf{Y} = \mathbf{Y} * c/u$
  $\mathbf{Y} = \mathbf{Y} * r/\mathbf{Y}.\text{sum}(0)$
end for
return $\mathbf{Y}^T$

where $\mathbf{1}_N$ denotes the ones vector in dimension N. Eq. 7 enforces that on average each category is selected N/C times in a mini-batch while the assignment for single sample satisfies the probability distribution.

Under long-tail data distribution, the assignment rule satisfying Eq. 7 improves the detection sensitivity of low-frequency samples. After pre-training the SGG model with biased data, the model is prone to generating higher confidence for high-frequency predicates. For high-frequency predicates, the constraint of Eq. 7 forces the algorithm to assign an average lower value to these samples. For low-frequency predicates, this constraint amplifies interaction samples that probably contain this type of relationship. Since other samples hardly have this relationship and the predicted confidence is extremely low, the algorithm needs to increase its assignment weight to satisfy $\mathbf{Y}^T\mathbf{1}_N = \frac{N}{C}\mathbf{1}_C$.

Although Eq. 6 ensures the detection sensitivity to low-frequency samples, we find that this static assignment method still harbors some biases. In Fig. 4, we intuitively demonstrate this bias: *it lacks the ability to correct itself and will only assign an increasing number of samples to the preferred categories, leading to another form of data bias.* To overcome this issue, we further propose a dynamic assignment mechanism. We consider dynamically adjusting the weight constraint of Eq. 7 based on historical assignment statistics:

$$w_k = (s_k^h * f_k + 1)^\sigma, \quad w = \frac{w}{\overline{w}}, \tag{8}$$

s_k^h represents the average assignment count for the k^{th}

Baseline	Methods	PredCls			SGCls			SGDet		
		R@50 / 100	mR@50 / 100	M@50 / 100	R@50 / 100	mR@50 / 100	M@50 / 100	R@50 / 100	mR@50 / 100	M@50 / 100
Motifs	Baseline	65.2 / 67.5	15.3 / 17.0	40.3 / 42.3	39.1 / 39.9	8.0 / 8.5	23.6 / 24.2	32.1 / 36.9	5.5 / 6.8	18.8 / 21.9
	TDE (Tang et al. 2020)	46.2 / 51.4	25.5 / 29.1	35.9 / 40.3	27.7 / 29.9	13.1 / 14.9	20.4 / 22.4	16.9 / 20.3	8.2 / 9.8	12.6 / 15.1
	NICE (Li et al. 2022)	55.1 / 57.2	29.9 / 32.3	42.5 / 44.8	33.1 / 34.0	16.6 / 17.9	24.9 / 26.0	27.8 / 31.8	12.2 / 14.4	20.0 / 23.1
	IETrans* (Zhang et al. 2022)	54.7 / 56.7	30.9 / 33.6	42.8 / 45.2	32.5 / 33.4	16.8 / 17.9	24.7 / 25.7	26.4 / 30.6	12.4 / 14.9	19.4 / 22.8
	CFA (Li et al. 2023)	54.1 / 56.6	35.7 / 38.2	44.9 / 47.4	34.9 / 36.1	17.0 / 18.4	26.0 / 27.3	27.4 / 31.8	13.2 / 15.5	20.3 / 23.7
	ST-SGG* (Kim et al. 2024)	63.4 / 65.4	22.4 / 24.1	42.9 / 44.8	36.8 / 37.8	12.1 / 12.8	24.5 / 25.3	29.7 / 34.8	8.5 / 10.1	19.1 / 22.5
	SSC-SGG*	59.7 / 62.0	31.5 / 34.0	**45.6 / 48.0**	37.4 / 38.5	17.8 / 18.9	**27.6 / 28.7**	28.6 / 33.0	12.3 / 14.4	**20.5 / 23.7**
VCTree	Baseline	66.2 / 68.1	14.9 / 16.1	40.6 / 42.1	40.5 / 41.4	7.5 / 7.9	24.0 / 24.7	31.5 / 36.2	5.7 / 6.9	18.6 / 21.6
	TDE (Tang et al. 2020)	47.2 / 51.6	25.4 / 28.7	36.3 / 40.2	25.4 / 27.9	12.2 / 14.0	18.8 / 21.0	19.4 / 23.2	9.3 / 11.1	14.4 / 17.2
	NICE (Li et al. 2022)	55.0 / 56.9	30.7 / 33.0	42.9 / 45.0	37.8 / 39.0	19.9 / 21.3	28.9 / 30.2	27.0 / 30.8	11.9 / 14.1	19.5 / 22.5
	IETrans* (Zhang et al. 2022)	53.0 / 55.0	30.3 / 33.9	41.7 / 45.0	32.9 / 33.8	16.5 / 18.1	24.7 / 26.0	25.4 / 29.3	11.5 / 14.0	18.5 / 21.7
	CFA (Li et al. 2023)	54.7 / 57.5	34.5 / 37.2	44.6 / 47.4	42.4 / 43.5	19.1 / 20.8	30.8 / 32.2	27.1 / 31.2	13.1 / 15.5	20.1 / 23.4
	ST-SGG* (Kim et al. 2024)	64.2 / 66.2	21.5 / 22.9	42.9 / 44.6	37.5 / 38.4	12.0 / 12.5	24.8 / 25.5	30.4 / 34.7	8.7 / 10.1	19.6 / 22.4
	SSC-SGG*	59.5 / 61.7	31.1 / 33.4	**45.3 / 47.6**	42.9 / 44.3	21.7 / 23.2	**32.3 / 33.8**	28.2 / 32.5	12.7 / 14.6	**20.5 / 23.6**
Transformer	Baseline	65.1 / 66.8	16.1 / 17.7	40.6 / 42.3	38.4 / 39.1	9.2 / 10.0	23.8 / 24.6	31.2 / 35.6	7.2 / 8.4	19.2 / 22.0
	CogTree (Yu et al. 2020)	38.4 / 39.7	28.4 / 31.0	33.4 / 35.4	22.9 / 23.4	15.7 / 16.7	19.3 / 20.1	19.5 / 21.7	11.1 / 12.7	15.3 / 17.2
	IETrans* (Zhang et al. 2022)	51.8 / 53.8	30.8 / 34.7	41.3 / 44.3	32.6 / 33.5	14.7 / 19.1	23.7 / 26.3	25.5 / 29.6	12.5 / 15.0	19.0 / 22.3
	CFA (Li et al. 2023)	59.2 / 61.5	30.1 / 33.7	44.7 / 47.6	36.3 / 37.3	15.7 / 17.2	26.0 / 27.3	27.7 / 32.1	12.3 / 14.6	20.0 / 23.4
	SSC-SGG*	60.1 / 62.3	31.5 / 33.8	**45.8 / 48.1**	37.3 / 38.4	18.6 / 19.9	**28.0 / 29.2**	28.9 / 33.0	12.3 / 14.4	**20.6 / 23.7**

Table 1: Comprehensive comparison on VG150 with state-of-the-art methods based on ResNeXt-101-FPN backbone. The best and second-best is highlighted in bold and underlined, respectively. * denotes pseudo-labeling-like method.

category in the most recent h mini-batches. $f_k = -\log(n_k/n_{total})$ is a frequency factor used to give lower-frequency categories larger weight and σ is a smooth factor. In order to make the two constraints in Eq. 7 compatible after re-weighting, we need to normalize the weights to have an average of 1.0.

After replacing $\frac{N}{C}$ with $w\frac{N}{C}$ in Eq. 7, Eq. 6 will adjust the assignment weight based on the historical assignment information of the current mini-batch. A portion of low-frequency samples will gradually reduce their weight as the number of assignments increases, thus allowing other low-frequency samples to be assigned. The solution $\mathbf{Y}^*$ of Eq. 6 can be solved by the iterative Sinkhorn-Knopp algorithm (Cuturi 2013) shown in Algorithm 1. $\mathbf{Y}^* \in \mathbb{R}^{N\times C}$ is exactly the normalized soft pseudo-labels to all samples in a mini-batch after assignment. We take the top-n samples with the highest confidence for pseudo-labeling and keep the remaining samples as *background*.

Swap-Joint Training

Finally, we conduct multi-task joint training with the pseudo-labels and the labeled data:

$$\mathcal{L}_u(x) = -(\sum_k y_{v2}^{(k)} \log x_{v1}^{(k)} + \sum_k y_{v1}^{(k)} \log x_{v2}^{(k)}), \quad (9)$$

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_l + \mathcal{L}_p, \quad (10)$$

where $\mathcal{L}_u$ and $\mathcal{L}_l$ represent the loss of unlabeled and labeled data respectively, which optimize with the same cross-entropy loss. In Eq. 9, we leverage the swap-prediction (Caron et al. 2020) to ensure the local consistency of multi-view relationship features during the clustering process.

Experiments

Experimetal Setup

Dataset. We evaluate our proposed SSC-SGG framework on two common-used SGG datasets, *i.e.*, Visual Genome (VG) (Krishna et al. 2017) and Open Image (OI) V6 (Kuznetsova et al. 2020). Following previous work (Zellers et al. 2018; Tang et al. 2019), We adopt the most popular pre-processed VG150 including 108k images, the most frequent 150 object classes, and 50 predicate categories. For OI V6, we follow the same data pre-processing and evaluation protocols utilized in (Li et al. 2021; Lin et al. 2022a), including 602 object classes and 30 predicate categories.

Models		mR@50	R@50	wmAP		$score_{wtd}$
				rel	phr	
RelDN (Zhang et al. 2019)		33.98	73.08	32.16	33.39	40.84
RU-Net (Lin et al. 2022b)		-	76.90	35.40	34.90	43.50
PE-Net (Zheng et al. 2023)		-	76.50	36.60	37.40	44.90
SGTR+ (Li, Zhang, and He 2023)		42.70	72.16	39.54	41.53	45.59
Motifs	Baseline	32.68	71.63	29.91	31.59	38.93
	+TDE	35.50	69.30	30.70	32.80	39.30
	+ST-SGG	34.10	71.80	31.10	32.50	39.80
	+MPC	39.83	76.62	38.13	39.53	46.21
	+SSC-SGG	42.48	75.11	36.97	38.20	44.69
Transformer	Baseline	32.12	74.78	33.71	34.59	42.19
	+MPC	38.51	76.03	39.27	40.38	**47.02**
	+SSC-SGG	**44.26**	74.90	38.07	39.08	45.61

Table 2: SGDet comparison on Open Image V6 dataset.

Tasks and Evaluation Metrics. The scene graph generation is divided into three sub-tasks with the acquisition of objects: Predicate Classification (**PredCls**) that takes ground truth object detection as inputs; Scene Graph Classification (**SGCls**) gives ground truth bounding boxes; and Scene Graph Detection (**SGDet**) detect the whole scene graph from the scratch. To estimate the performance, we use **Recall@K (R@K)**, **mean Recall@K (mR@K)**, and **M@K** as evaluation metrics. The **M@K** is defined as the mean of **R@K** and **mR@K**, which reflect the comprehensive performance of the model.

Implementation Details: For the object detector selection, we utilize the pre-trained Faster R-CNN (Ren et al. 2015) with ResNeXt-101-FPN (Xie et al. 2017) and freeze the model parameters during training. In dynamic pseudo-label assignment algorithm, we set the regularization coefficient $\lambda = 0.05$, the number of iterations $t = 3$, and the smooth factor $\sigma = 0.5$. Then, we pseudo-label the top-5 samples ($n = 5$) with the highest assignment confidence scores for each image on average. The model is trained by an SGD optimizer with

Module		PredCls			SGCls			SGDet		
MPC	DLA	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100
✗	✗	65.2 / 67.5	15.3 / 17.0	40.3 / 42.3	39.1 / 39.9	8.0 / 8.5	23.6 / 24.2	32.1 / 36.9	5.5 / 6.8	18.8 / 21.9
✓	✗	65.7 / 67.8	20.2 / 22.1	43.0 / 45.0	40.4 / 41.4	12.1 / 13.2	26.3 / 27.3	31.8 / 36.2	7.9 / 9.3	19.9 / 22.8
✓	✓	59.7 / 62.0	31.5 / 34.0	**45.6 / 48.0**	37.4 / 38.5	17.8 / 18.9	**27.6 / 28.7**	28.6 / 33.0	12.3 / 14.4	**20.5 / 23.7**

Table 3: Ablation study of the framework components. None of them denotes the vanilla Motifs (Zellers et al. 2018).

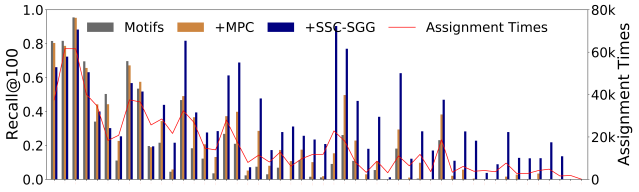

Figure 3: The performance of predicate classes on Recall@100 (bar graph) and the assignment times in the whole semi-supervised training process (line graph).

60k iterations. The initial learning rate is 1.0×10^{-3} with being decayed by a factor of 10 at 28k and 48k iterations. The pre-training process is executed for the 50k iterations using proposed multi-view prototype-based clustering framework without pseudo-labels, followed by semi-supervised training for the subsequent 10k iterations. The batch size is set to 8. All our experiments are conducted using a RTX A5000 GPU.

Performance Comparisons

Visual Genome. We apply our semi-supervised framework to various popular baseline models including Motifs (Zellers et al. 2018), VCTree (Tang et al. 2019), and Transformer (Vaswani et al. 2017). As shown in Table 1, compared with vanilla baseline models, the SSC-SGG framework average outperforms with 13.4%, 24.7%, and 8.4% increment on M@100 across PredCls, SGCls, and SGDet tasks in three baseline models. SSC-SGG achieves significant improvements in mR@K while maintaining relatively high R@K performance, which indicates that SSC-SGG provides a beneficial boost to the prediction results, *i.e.*, achieving model de-biasing without sacrificing the performance of high-frequency predicates.

Furthermore, our proposed SSC-SGG framework doesn't apply any de-biasing operations to the labeled data, but merely mines effective interaction samples from unlabeled object pairs, yet it achieves state-of-the-art comprehensive performance, *i.e.*, M@K, in most sub-tasks. In particular, compared with the previous SOTA method CFA (Li et al. 2023), the SSC-SGG average outperforms with 2.7%, 5.7%, and 0.7% increment on M@100 across all sub-tasks. In addition, compared with SOTA pseudo-labeling methods, *i.e.*, IETrans (Zhang et al. 2022) and ST-SGG (Kim et al. 2024), SSC-SGG shows a huge performance gain with average 10.2% and 11.8% on M@100, which powerfully proves that SSC-SGG possesses a strong ability to mine effective interaction samples from unlabeled object pairs.

Open Image. We further verify the generalization of our proposed method on Open Image V6 (OI V6) (Kuznetsova et al. 2020). Following (Li et al. 2021), we use the mean Recall@50 (mR@50), Recall@50 (R@50), weighted mean AP of relationships (wmAPrel), and weighted mean AP of phrase (wmAPphr) as evaluation metrics. Following standard evaluation metrics of Open Images, the weight metric $score_{wtd}$ is computed as $score_{wtd} = 0.2 \times R@50 + 0.4 \times wmAP_{rel} + 0.4 \times wmAP_{phr}$. As shown in Table 2, our Multi-view Prototype-based Clustering (MPC) framework shows superior performance compared with baseline models, Motifs and Transformer, demonstrating good generalization of MPC framework. After applying the dynamic pseudo-label assignment algorithm, the framework achieves significant improvements on mR@50, which indicates that our proposed pseudo-labeling algorithm can effectively exploit useful low-frequency interaction samples.

Ablation Study and Model Analysis

We further conduct a detailed ablation study over components in our framework and show the performance gain in Table 3. MPC and DLA are the abbreviation of *Multi-view Prototype-based Clustering framework*, and *Dynamic pseudo-Label Assignment*, respectively. None of them denotes the vanilla Motifs (Zellers et al. 2018).

Analysis on Components. As shown in Table 3, we quantitatively verify the effectiveness of each component. Compared to the vanilla Motifs, utilizing the proposed MPC framework achieves an average improvement of 40.7% and 7.8% on mR@K and M@K respectively, and maintains equal or even better performance on R@K. It convinces that the proposed MPC framework constrains the local feature consistency of multi-view samples and the distinctiveness of features among samples of different categories through the prototypes, enhancing the robustness of the feature learning. We then apply the proposed dynamic pseudo-label assignment algorithm to this framework. As shown in Fig. 3, the performance of low-frequency predicates has significantly improved, which further improves the overall performance on M@K by 5.2% on average. This proves that our method successfully mines effective low-frequency samples from unlabeled object pairs, enhancing the overall sample richness while hardly interfering with the training of high-frequency samples.

Discussion on Label Assignment Methods. We discuss the impact of different pseudo-label assignment methods on the overall performance in Table 4. We design three strategies. *Static hard label* indicates that we convert the soft pseudo-labels into one-hot encoded hard labels, while *static soft label* retains the obtained soft labels. *Dynamic soft label* is our proposed DLA algorithm, adjusting weights based on historical assignment information. Compared to the use of hard pseudo-labels, soft labels show a huge advantage in performance. This may be because soft labels provide more

Method	PredCls		
	R@50/100	mR@50/100	M@50/100
no assignment	65.7 / 67.8	20.2 / 22.1	43.0 / 45.0
static hard label	57.6 / 60.9	27.2 / 29.6	42.4 / 45.3
static soft label	60.8 / 63.5	28.2 / 30.8	44.5 / 47.2
dynamic soft label	59.7 / 62.0	31.5 / 34.0	**45.6 / 48.0**

Table 4: Performance on different label assignment methods.

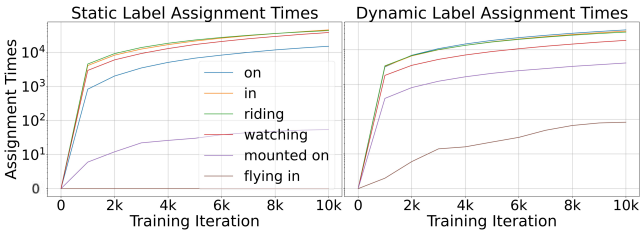

Figure 4: Statistics on the assignment times of the selected typical high (*on, in*), medium (*riding, watching*), and low-frequency (*mounted on, flying in*) predicates using static (left) and dynamic (right) label assignment methods.

probabilistic information, which is more in line with the relationship prediction environment. In addition, our proposed DLA algorithm significantly improves the mR@K while retaining the R@K. Fig. 4 shows the statistics of the label assignment times for each category in the semi-supervised training process both in static and dynamic methods. The static assignment method, as training proceeds, assigns more frequently to some predicates with medium-frequency, like *riding* and *watching*, but rarely caters to low-frequency samples. This results in another form of data bias, where the medium-frequency predicates upgrade to high-frequency, but the low-frequency samples remain unchanged. However, the dynamic method effectively avoids this problem. As shown in Fig. 4 (right), it can cater to low-frequency predicates, while reducing the assignment times for high-frequency predicates.

Fair comparison with other pseudo-labeling methods in SGG. To fairly compare with other pseudo-labeling methods in SGG, we also test the proposed dynamic label assignment (DLA) algorithm on vanilla Motifs framework. As shown in Table 5, **compared with existed SGG pseudo-labeling methods, IETrans (Zhang et al. 2022) and ST-SGG (Kim et al. 2024), DLA shows the best overall performance in all sub-tasks and exhibits significant marginal gains on PredCls and SGCls tasks.** We attribute this to our designed holistic assignment scheme, which not only shows more flexible and balanced assignment but also reduces the manual design of algorithmic strategies. We conduct detailed ablation studies on DLA algorithm in the supplementary materials.

Discussion on Multi-View Setting. In the SSC-SGG framework, we enhance the features of the relationships by introducing multi-view clustering, thereby improving the robustness of model training. To validate the effectiveness of this scheme, we maintain the contrastive learning of relationship features and prototypes but only enhance the features

Method	PredCls	SGCls	SGDet
	M@50/100	M@50/100	M@50/100
vanilla	40.3 / 42.3	23.6 / 24.2	18.8 / 21.9
IETrans	42.8 / 45.2	24.7 / 25.7	19.4 / 22.8
ST-SGG	42.9 / 44.8	24.5 / 25.3	19.1 / 22.5
DLA (Ours)	**44.9 / 47.6**	**26.8 / 27.9**	**19.6 / 22.9**

Table 5: Comparison with other pseudo-labeling methods in SGG under vanilla Motifs.

Method		PredCls	SGCls	SGDet
		M@50/100	M@50/100	M@50/100
multi-view	pre-trained	43.0 / 45.0	26.3 / 27.3	19.9 / 22.8
	pseudo-label	**45.6 / 48.0**	**27.6 / 28.7**	**20.5 / 23.7**
single-view	pre-trained	43.0 / 45.1	26.2 / 27.3	19.7 / 22.6
	pseudo-label	44.5 / 47.0	27.1 / 28.5	19.6 / 22.8

Table 6: Performance of no relationship augmentation (single-view) compared with multi-view scheme.

in a single view. In Table 6, we present the performance of pre-trained model and after pseudo-labeling. Firstly, only conducting multi-view feature enhancement on the framework almost has no performance gain. We argue the reason as two folds: 1) All unlabeled samples are assumed to come from the *background* in the pure framework, consequently, there are no distinctive features that need to be learned. 2) Since the *background* has a very large number of samples, multi-view clustering does not bring much value. Secondly, when we apply the multi-view framework in a semi-supervised environment, we obtain an obvious performance gain compared to the single-view framework. This is mainly because the objects of clustering are not just the *background*, but also a certain proportion of pseudo-samples. These samples contain a lot of low-frequency interaction information, which helps the model train a robust feature space for low-frequency categories, thus improving the final performance.

Conclusion

In this paper, considering the biased and sparse annotation problems existing in the SGG dataset, we construct a clustering-based semi-supervised framework for SGG, which utilizes the labeled data to mine effective interaction information from a large number of unlabeled object pairs. These pseudo-labels enrich the original sparse labeled sample space, especially for low-frequency samples. We first design a multi-view prototype-based clustering framework to enhance the robustness of feature learning and ensure the feature consistency between unlabeled and labeled data through prototypes. Then, we propose a dynamic pseudo-label assignment algorithm, which adjusts the detection sensitivity to low-frequency interaction samples based on historical assignment information, thereby discovering more balanced and effective low-frequency samples. Finally, we use a swap-joint training strategy to guide the clustering process with labeled data. The results show that our proposed framework significantly improves the comprehensive performance of SGG.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62102413, 62376088, 62132006, 62276019), in part by the Hebei Natural Science Foundation No. F2024202047, in part by the Fundamental Research Funds for the Central Universities under Grant 2023JBZY037, 2022JBQY007, in part by the Beijing Natural Science Foundation under Grant L231019, in part by the Hebei Yanzhao Golden Platform Talent Gathering Programme Core Talent Project (Education Platform) (HJZD202509), and in part by Guangdong Provincial Key Laboratory (Grant 2023B1212060076).

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Cong, W.; Wang, W.; and Lee, W.-C. 2018. Scene graph generation via conditional random fields. *arXiv preprint arXiv:1811.08075*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Guo, Y.; Gao, L.; Wang, X.; Hu, Y.; Xu, X.; Lu, X.; Shen, H. T.; and Song, J. 2021. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16383–16392.

Jiao, Y.; Chen, S.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, 528–545. Springer.

Jiao, Y.; Chen, S.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2024. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*.

Jiao, Y.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2023. Suspected Objects Matter: Rethinking Model's Prediction for One-stage Visual Grounding. In *Proceedings of the 31st ACM International Conference on Multimedia*, 17–26.

Jiao, Y.; Jie, Z.; Luo, W.; Chen, J.; Jiang, Y.-G.; Wei, X.; and Ma, L. 2021. Two-stage visual cues enhancement network for referring image segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, 1331–1340.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3668–3678.

Kim, K.; Yoon, K.; In, Y.; Moon, J.; Kim, D.; and Park, C. 2024. Adaptive Self-training Framework for Fine-grained Scene Graph Generation. *arXiv preprint arXiv:2401.09786*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7): 1956–1981.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Li, L.; Chen, G.; Xiao, J.; Yang, Y.; Wang, C.; and Chen, L. 2023. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21685–21695.

Li, L.; Chen, L.; Huang, Y.; Zhang, Z.; Zhang, S.; and Xiao, J. 2022. The Devil is in the Labels: Noisy Label Correction for Robust Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18869–18878.

Li, R.; Zhang, S.; and He, X. 2023. SGTR+: End-to-end Scene Graph Generation with Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.

Li, R.; Zhang, S.; Wan, B.; and He, X. 2021. Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11109–11119.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, 1261–1270.

Lin, X.; Ding, C.; Zhan, Y.; Li, Z.; and Tao, D. 2022a. HL-Net: Heterophily Learning Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19476–19485.

Lin, X.; Ding, C.; Zhang, J.; Zhan, Y.; and Tao, D. 2022b. RU-Net: Regularized Unrolling Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19457–19466.

Min, Y.; Wu, A.; and Deng, C. 2023. Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13296–13307.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eledath, J.; Medioni, G.; and Sigal, L. 2021. Energy-Based Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13936–13945.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6619–6628.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Yan, S.; Shen, C.; Jin, Z.; Huang, J.; Jiang, R.; Chen, Y.; and Hua, X.-S. 2020. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 265–273.

Yang, F.; Wu, K.; Zhang, S.; Jiang, G.; Liu, Y.; Zheng, F.; Zhang, W.; Wang, C.; and Zeng, L. 2022a. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14421–14430.

Yang, L.; Huang, Z.; Song, Y.; Hong, S.; Li, G.; Zhang, W.; Cui, B.; Ghanem, B.; and Yang, M.-H. 2022b. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*.

Yu, J.; Chai, Y.; Wang, Y.; Hu, Y.; and Wu, Q. 2020. CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation. In *arXiv preprint arXiv:2009.07526*.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5831–5840.

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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, 409–424. Springer.

Zhang, J.; Shih, K. J.; Elgammal, A.; Tao, A.; and Catanzaro, B. 2019. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11535–11543.

Zheng, C.; Lyu, X.; Gao, L.; Dai, B.; and Song, J. 2023. Prototype-based Embedding Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22783–22792.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1097–1103.