Leveraging Predicate and Triplet Learning for Scene Graph Generation

Jiankai Li[1,2,3] Yunhong Wang[1] Xiefan Guo[1] Ruijie Yang[1] Weixin Li[1,2,3*]
[1] IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] State Key Laboratory of Complex & Critical Software Environment, Beihang University, Beijing, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
{lijiankai, yhwang, xfguo, rjyang, weixinli}@buaa.edu.cn

Abstract

Scene Graph Generation (SGG) aims to identify entities and predict the relationship triplets <subject, predicate, object> in visual scenes. Given the prevalence of large visual variations of subject-object pairs even in the same predicate, it can be quite challenging to model and refine predicate representations directly across such pairs, which is however a common strategy adopted by most existing SGG methods. We observe that visual variations within the identical triplet are relatively small and certain relation cues are shared in the same type of triplet, which can potentially facilitate the relation learning in SGG. Moreover, for the long-tail problem widely studied in SGG task, it is also crucial to deal with the limited types and quantity of triplets in tail predicates. Accordingly, in this paper, we propose a Dual-granularity Relation Modeling (DRM) network to leverage fine-grained triplet cues besides the coarse-grained predicate ones. DRM utilizes contexts and semantics of predicate and triplet with Dual-granularity Constraints, generating compact and balanced representations from two perspectives to facilitate relation recognition. Furthermore, a Dual-granularity Knowledge Transfer (DKT) strategy is introduced to transfer variation from head predicates/triplets to tail ones, aiming to enrich the pattern diversity of tail classes to alleviate the long-tail problem. Extensive experiments demonstrate the effectiveness of our method, which establishes new state-of-the-art performance on Visual Genome, Open Image, and GQA datasets. Our code is available at https://github.com/jkli1998/DRM

1. Introduction

Entities and their associated relationships form the cornerstones of visual contents in images [14]. Scene Graph Generation (SGG), a fundamental task in visual scene understanding, is designed to detect these entities and predict their pairwise relationships, encapsulating them into *<subject, predicate, object>* triplets [4, 26, 51]. The generated compact graph-structured image representation can be utilized in a range of applications, *e.g.* embodied navigation [7, 33], image retrieval [5, 8], visual question answering [1, 32], *etc.*, so the SGG task has received widespread attention in recent years [12, 16, 22, 45].

Figure 1. The illustration of large visual variations within the predicate "eating". Identical predicate can appear differently under distinct subject-object pairs, encompassing a different set of visual cues within each manifestation. Identifying discriminative relation cues that are shared across diverse subject-object pairs within the same predicate can be challenging. Yet, they can be easily captured when the scope is narrowed to the identical triplet.

Existing SGG methods are mostly dedicated to generating discriminative predicate representations for the detected entities, based on their appearances, relative positions, contextual cues, *etc.* [15, 19, 37, 44, 54]. However, as shown in Figure 1, large visual variations due to different subject-object combinations are inherent even in the same predicate, presenting an obstacle for SGG methods to capture robust predicate cues across distinct triplet types. To alleviate this problem, PE-Net [54] utilizes textual semantics of predicate categories as the prototype and model predicate cues by reducing the intra-class variance and inter-class similarity. Despite the improved prediction accuracy, PE-Net still follows the previous strategy of straightforwardly amalga-

*Corresponding author

mating predicate cues that probably contain extensive visual variations of diverse triplets.

Moreover, many recent efforts have been devoted to the long-tail problem in SGG task. Insufficient samples with limited triplet types lead to the reduced diversity observed in tail predicate categories, making it a challenging task to learn and adapt to their distributions. Existing methods boost model's attention towards tail predicate categories through re-sampling [19], re-weighting[28], or the utilization of mixture of experts [6, 34]. However, they mostly fall short of directly tackling the core of the long-tail issue, *i.e.*, the insufficient patterns for tail predicates, leaving space for further improvement.

Reflecting on similarities and differences among predicates, we find that despite the non-negligible or even large variations inherent in the same predicate, the visual diversity of the same triplet is relatively small (*e.g.* the two instances of <*elephant, eating, leaf*> in Figure 1). Accordingly, considering the more fine-grained triplet cues in addition to the coarse-grained predicate ones can help prevent the model from getting stuck in the refinement process of predicate features with potentially large variations, and promote it to strike a balance between different granularities during the relationship learning process. For the long-tail problem in SGG task which primarily emerges due to the insufficient tail predicate patterns and corresponding limited types and quantity of triplets, it then becomes a natural choice to enrich tail predicate patterns using head predicates and their triplets.

Based on the aforementioned insights, we propose a Dual-granularity Relation Modeling (DRM) network that models triplet cues to facilitate predicate learning and transfer knowledge from the head classes to tail classes for relation recognition. In our DRM network, as shown in Figure 2, 1) besides a predicate branch that models coarse-grained predicate cues by leveraging their contextual predicates and subject-object pairs, a triplet branch is also presented to strengthen fine-grained triplet representations via jointly exploring their visual contents and corresponding label semantics. We also devise dual-granularity constraints to prevent the degradation of predicate and triplet feature spaces during model training. Subsequently, 2) the Dual-granularity Knowledge Transfer (DKT) strategy is proposed to transfer the class variance from head classes to tail ones from both the predicate and triplet perspectives for unbiased SGG. Distributions of tail predicates/triplets are designed to be calibrated using variance from head ones that are most similar to them. New predicate/triplet samples are generated as well based on the calibrated distributions to enrich the pattern diversity of tail predicates/triplets. Extensive experiments conducted on the widely used Visual Genome [14], Open Image [13], and GQA [10] datasets for SGG demonstrate the state-of-the-art performance of our method.

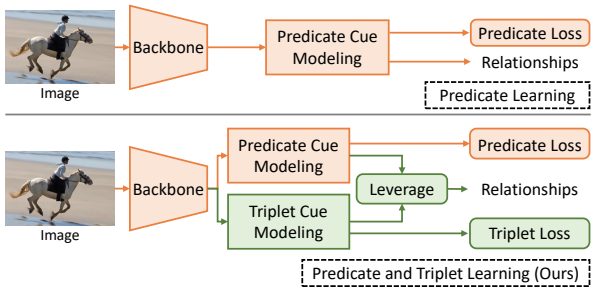


Figure 2. Comparison of different pipelines for relation recognition. Previous methods focus on learning predicate cues shared across various triplets with diverse visual appearance. Our method learns and leverages both triplet cues within the same triplet and predicate cues across triplets, to better handle the visual diversity.

Contributions of this paper can be summarized as:

- We propose to learn scene graphs in a dual-granularity manner, integrating both coarse-grained predicate cues and fine-grained triplet cues.
- We introduce the DRM network to model predicate and triplet cues, and propose the DKT strategy to propagate knowledge from head predicates/triplets to tail ones to mitigate the long-tail problem in SGG.
- We comprehensively evaluate the proposed method and demonstrate its superior performance.

2. Related Work

2.1. Scene Graph Generation

Scene graph is a structured representation of the image content, where its essential constituents are the relationships (or triplets <*subject, predicate, object*>). The direct prediction of triplet categories presents a significant challenge given the extensive combinations of subjects, predicates, and objects. Existing Scene Graph Generation (SGG) methods [11, 16, 25, 42, 52, 54] decompose this prediction target into two components: entities and predicates. Early SGG approaches [26, 49, 51] explore to integrate multiple modalities like positional information and linguistic features into relationships. Later methodologies [11, 27, 35, 36, 39] identify the value of visual context in SGG. Some of them encode contextual information utilizing techniques *e.g.* message passing [43], LSTM [37, 41, 49], graph neural networks [19, 44], and self-attention modules [3, 6, 20]. Others [27, 39] refine the detected scene graph and optimize the features of refined predicates based on high-confidence predictions. PE-Net [54] proposes to utilize text embeddings as the centroid of predicates, aiming to minimize the intra-class variance and the inter-class similarity. Despite the improved prediction accuracy, it is still hard to extract discriminative relation cues across the various subject-object pairs in the same predicate. The core of this issue lies in the fact that the identical predicates can manifest differently under distinct subject-object pairs, encompassing a unique array

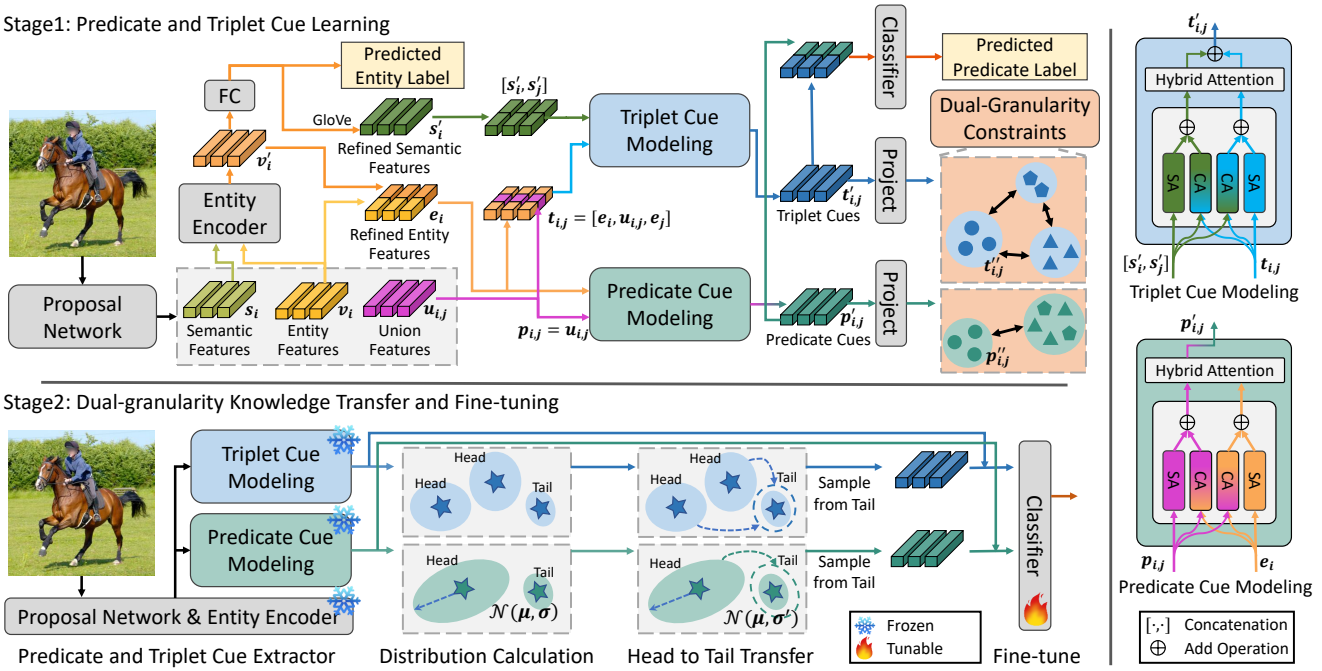


Figure 3. Illustration of the proposed Dual-granularity Relation Modeling (DRM) network. The learning procedure of DRM is composed of two stages. In the first stage, we capture the coarse-grained predicate cues shared across different subject-object pairs and learn the fine-grained triplet cues under specific subject-object pairs. In the second stage, the Dual-granularity Knowledge Transfer (DKT) strategy transfers the variation from head predicates with their associate triplets to the tail. Then DRM exploits the real instances along with synthetic samples from the calibrated tail distribution to fine-tune the relation classifier, which alleviates the long-tail problem in SGG.

of visual cues within each manifestation. The direct aggregation of predicate features tends to overlook the inherent triplet cues present under the various subject-object pairs. We propose to explicitly model both coarse-grained predicate cues and fine-grained triplet cues for biased and unbiased SGG.

2.2. Unbiased Scene Graph Generation

Real-world data tends to obey a long-tailed distribution, and imbalanced samples in scene graph present the challenge to learn and adapt to the distribution in the tail predicates [37, 38, 48, 53]. Recently, many methods have been proposed to deal with the biased prediction problem in scene graphs, which can be roughly divided into three categories. The first category of methods use re-balancing strategies that alleviate the long-tail problem by re-sampling images and triplet samples to enhance the performance of tail predicates [19] or by enhancing the loss weights of tail predicates and easily misclassified categories through the pre-defined Predicate Lattice [28]. The second category of methods utilize noisy label learning to explicitly re-label relational triplets missed by annotators and correct the mislabeled predicates [17, 50]. The last category exploit the mixture of experts to let different experts separately deal with a subpart of predicates and merge their outputs [6, 34]. Different from these approaches, we mitigate the long-tail problem in SGG by exploiting knowledge from the head predicates and triplets. Our method involves transferring the abundant variations from the head predicates and their associate triplets to the tail, thus enriching patterns of the tail predicates.

3. Method

An overview of our Dual-granularity Relation Modeling (DRM) network is illustrated in Figure 3. DRM aims to balance and integrate coarse-grained predicate cues and fine-grained triplet cues for relation recognition. To this end, a predicate cue modeling module and a triplet cue modeling module are designed to extract predicate cues shared across diverse subject and object pairs, and triplet cues shared by specific subject-object pairs, respectively. Moreover, the Dual-granularity Knowledge Transfer (DKT) strategy is proposed for unbiased SGG. This strategy transfers the knowledge of both predicates and triplets from head categories to tail ones, with the objective of enriching the pattern of tail predicates and their associated triplets.

3.1. The DRM Network Backbone

The backbone of our DRM network is composed of a proposal network and an entity encoder. Features generated by the backbone are further fed into subsequent predicate and triplet cue modeling modules.

Proposal Network. Given an image $\mathcal{I}$, the proposal network generates N entities along with their corresponding visual features, label predictions, and spatial features from their bounding boxes. A pre-trained Faster RCNN [31] is adopted as the proposal network in this paper. And following previous works [6, 19, 37, 49], we initialize the entity representations $\{\mathbf{v}_i\}_{i=1}^N$ with their visual and spatial features, encode the union feature $\mathbf{u}_{i,j}$ between the i-th entity and the j-th entity with their relative spatial representation and the ROI feature of their union box, and obtain the semantic features $\{\mathbf{s}_i\}_{i=1}^N$ of entities using word embeddings of their class labels.

Entity Encoder. The entity encoder is designed to refine features of entities with their contexts for further predictions. Inspired by Dong *et al.* [6], we utilize the Hybrid Attention (HA) to incorporate semantic cues $\{\mathbf{s}_i\}_{i=1}^N$ into entities $\{\mathbf{v}_i\}_{i=1}^N$ while modeling the scene context. Each layer of Hybrid Attention is composed of two Self-Attention (SA) units and two Cross-Attention (CA) units, which is built upon the multi-head attention module [40]. The Hybrid Attention at the l-th layer can be formulated as:

$$\begin{cases} \mathbf{X}^{(l)} = SA(\mathbf{X}^{(l-1)}) + CA(\mathbf{X}^{(l-1)}, \mathbf{Y}^{(l-1)}), \\ \mathbf{Y}^{(l)} = SA(\mathbf{Y}^{(l-1)}) + CA(\mathbf{Y}^{(l-1)}, \mathbf{X}^{(l-1)}), \end{cases} \tag{1}$$

where $\mathbf{X}^{(l-1)}$ and $\mathbf{Y}^{(l-1)}$ are inputs of the l-th HA layer. We directly fuse outputs of SA and CA units with an addition operation.

Our entity encoder Enc_{ent} is consisted of a stacked 4-layer Hybrid Attention, and we have $\mathbf{X}^{(0)} = \{\mathbf{v}_i\}_{i=1}^N$, $\mathbf{Y}^{(0)} = \{\mathbf{s}_i\}_{i=1}^N$, and:

$$\{\mathbf{v}'_i\}_{i=1}^N = Enc_{ent}(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{s}_i\}_{i=1}^N), \tag{2}$$

where the refined entity representations $\{\mathbf{v}'_i\}_{i=1}^N$ are obtained by summing up of outputs from the last HA layer.

3.2. Predicate and Triplet Cue Modeling

Previous approaches tend to categorize predicates in a coarse manner, most of which only focus on predicate cues shared among various subject-object pairs, and thus cannot effectively deal with the potentially large visual variations inherent in the identical predicate. In contrast to these approaches, our method considers both coarse-grained predicate cues and fine-grained triplet cues, leveraging and striking a balance between the dual-grained cues for accurate predicate categorization. Our DRM network involves a Predicate Cue Modeling module and a Triplet Cue Modeling module for extracting features at these two granularities respectively. Additionally, we introduce Dual-Granularity Constraints to decrease the intra-class variance and increase inter-class distinguishability, explicitly enforcing the predicate and triplet branches to concentrate on cues from their corresponding granularities and preventing the degradation of dual-granularity space.

Predicate Cue Modeling. Our predicate cue modeling module Enc_{prd}, comprising a 2-layer Hybrid Attention, aims to capture predicate cues across different subject-object pairs. In each HA layer of Enc_{prd}, the two Self-Attention units are designed to model predicate and entity contextual information. And the two Cross-Attention units are designed to capture the dependency between entities and predicates, where the predicate $\mathbf{p}_{i,j}$ solely queries the contexts of its corresponding subject $\mathbf{e}_i$ and object $\mathbf{e}_j$, and the entity $\mathbf{e}_i$ only queries predicates related to it. Our Predicate Cue Modeling process can thus be formulated as:

$$\{\mathbf{p}'_{i,j}\}_{i\neq j}^M = Enc_{prd}(\{\mathbf{p}_{i,j}\}_{i\neq j}^M, \{\mathbf{e}_i\}_{i=1}^N), \tag{3}$$

where the predicate representation $\mathbf{p}_{i,j}$ is initialized with the union feature $\mathbf{u}_{i,j}$, the entity representation $\mathbf{e}_i$ is obtained with the concatenation of $\mathbf{v}'_i$ and $\mathbf{v}_i$, $M = N \times (N-1)$ is the number of subject-object pairs, and $\mathbf{p}'_{i,j}$ is the corresponding output of $\mathbf{p}_{i,j}$ at the last HA layer.

Triplet Cue Modeling. Our triplet cue modeling module Enc_{tpt} is also constructed using a 2-layer Hybrid Attention. This module is responsible for obtaining fine-grained triplet cues that are shared by specific subject-object pairs. The two Self-Attention units in each HA layer of Enc_{tpt} are designed to model visual and semantic contextual information of triplets, respectively, and the two Cross-Attention units aim at fusing semantic cues into the visual information of triplets. We initialize the triplet representation $\mathbf{t}_{i,j}$ with the concatenation of the subject representation $\mathbf{e}_i$, predicate representation $\mathbf{p}_{i,j}$, and object representation $\mathbf{e}_j$. Our Triplet Cue Modeling process is formulated as:

$$\{\mathbf{t}'_{i,j}\}_{i\neq j}^M = Enc_{tpt}(\{\mathbf{t}_{i,j}\}_{i\neq j}^M, \{[\mathbf{s}'_i, \mathbf{s}'_j]\}_{i\neq j}^M), \tag{4}$$

where $\mathbf{s}'_i$ is the word embedding of predicted entity label, and $[\cdot,\cdot]$ denotes the concatenate operation. $\mathbf{t}'_{i,j}$ is the contextually and semantically aware triplet feature, and is derived from the addition of outputs from the last HA layer.

Dual-granularity Constraints. Although we explicitly model predicate and triplet cues with Enc_{prd} and Enc_{tpt}, they may degrade beyond our desires with a single predicate cross-entropy loss. To prevent this degradation, we propose the dual-granularity constraints to guide the predicate and triplet cue modeling modules to refine representations in their desired granularities. Specifically, we generate two views of an input relation and impose a predicate category-aware supervised contrastive learning loss on predicate representations to capture the coarse-grained predicate cues as:

$$\mathcal{L}_p = -\log \frac{\exp(\langle \mathbf{p}''_{i,j}, \mathbf{p}''_{pos_p} \rangle / \tau_p)}{\sum_{b\in\mathcal{B}(i,j)} \exp(\langle \mathbf{p}''_{i,j}, \mathbf{p}''_b \rangle / \tau_p)}, \tag{5}$$

where $\mathbf{p}''_{i,j}$ is obtained by passing $\mathbf{p}'_{i,j}$ through a projection layer comprising two fully connected layers, τ_p is the tem-

perature, pos_p denotes the subscripts of positive samples belonging to the same predicate category as $\mathbf{p}''_{i,j}$, $\langle \cdot, \cdot \rangle$ is the cosine similarity function, and $\mathcal{B}(i,j)$ denotes the subscript set of samples within the same batch.

Similarly, a triplet category-aware supervised contrastive learning loss is applied on triplet representations, aiming at capturing the fine-grained triplet cues as:

$$\mathcal{L}_t = -\log \frac{\exp(\langle \mathbf{t}''_{i,j}, \mathbf{t}''_{pos_t} \rangle / \tau_t)}{\sum_{b \in \mathcal{B}(i,j)} \exp(\langle \mathbf{t}''_{i,j}, \mathbf{t}''_{b} \rangle / \tau_t)}, \tag{6}$$

where τ_t is the temperature, pos_t denotes subscripts of positive samples from the same triplet category as $\mathbf{t}''_{i,j}$, and $\mathbf{t}''_{i,j}$ is obtained by passing $\mathbf{t}'_{i,j}$ through a projection layer.

Scene Graph Prediction. For each relationship proposal, our predicate classifier utilizes two fully connected layers to integrate the coarse-grained and fine-grained cues, *i.e.* $\mathbf{p}'_{i,j}$ and $\mathbf{t}'_{i,j}$, to obtain the final relation label prediction. For each entity, we also introduce a fully connected layer with softmax function to get its refined label prediction.

Training Loss. During the training in the first stage of DRM, the overall loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_e \mathcal{L}_e + \lambda_r \mathcal{L}_r + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t, \tag{7}$$

where $\mathcal{L}_e$ and $\mathcal{L}_r$ are cross-entropy losses of entities and relationships respectively. λ_e, λ_r, λ_p, and λ_t are pre-defined weight hyper-parameters for corresponding loss terms.

3.3. Dual-granularity Knowledge Transfer

The SGG task typically suffers from the long-tail distribution problem. This problem primarily originates from the tail predicate classes, which possesses limited types and quantities of triplets. The head predicate classes contain a relatively larger number of samples and an abundance of triplet types. Thus to alleviate the long-tail problem, we propose the Dual-granularity Knowledge Transfer (DKT) strategy to transfer knowledge in predicate and triplet feature spaces from head predicate class to the tail one. DKT generates samples belonging to the tail predicates with their associated predicate and triplet features, thereby enriching and diversifying the patterns of tail predicates.

Specifically, DKT first calculates the distributions of predicate and triplet features. We assume that the distribution of each category Ω_c follows a multidimensional Gaussian distribution. Formally, it can be expressed as $\{\Omega_c = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c) | c \in \mathcal{C}\}$, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ denote the mean and covariance of Ω_c, and c denotes the predicate or triplet category. After the first-stage pre-training of DRM network, we freeze the proposal network, entity encoder, and the predicate and triplet cue modeling modules. Subsequently, the compact predicate and triplet features, *i.e.* $\mathbf{p}'$ and $\mathbf{t}'$, are extracted to calculate the predicate and triplet feature distributions, respectively. The mean, denoted as $\boldsymbol{\mu}_c$, is calculated as $\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_k^{N_c} \mathbf{x}_k^c$. The covariance, symbolized as $\boldsymbol{\sigma}_c$, is calculated as $\boldsymbol{\sigma}_c = \frac{1}{N_c - 1} \sum_k^{N_c} (\mathbf{x}_k^c - \boldsymbol{\mu}_c)(\mathbf{x}_k^c - \boldsymbol{\mu}_c)^T$. Here, $\mathbf{x}_i^c$ denotes the feature of category c, and N_c is the number of $\mathbf{x}_k^c$. The feature $\mathbf{x}$ can be either $\mathbf{p}'$ or $\mathbf{t}'$.

We then transfer knowledge of feature distributions of head predicate classes to the tail ones. Specifically, we arrange the predicate classes in descending order based on their sample numbers, choosing half of the predicate classes as head predicates and the remaining ones as tail predicates. We further select triplets that appear more than certain times in the head predicates to be the head triplets (we use 64 times as the threshold in this paper), and those in the tail predicates to be the tail triplets. For each tail category $i \in \mathcal{C}$, we compute the euclidean distance $d_{i,j}$ between its center $\boldsymbol{\mu}_i$ and the center $\boldsymbol{\mu}_j$ of head category $j \in \mathcal{C}$. The closer the centers of two categories are to each other in either predicate or triplet space, the more similar they are, which also increases the likelihood of knowledge sharing between them. Based on $d_{i,j}$, we achieve the knowledge transfer as:

$$\boldsymbol{\sigma}'_i = \frac{N_i}{Q_i} \boldsymbol{\sigma}_i + (1 - \frac{N_i}{Q_i}) \sum_j \alpha_{i,j} \boldsymbol{\sigma}_j, \tag{8}$$

where $\alpha_{i,j}$ denotes the softmax normalized form of $d_{i,j}$, Q_i denotes the desired number of predicate/triplet instances of tail class i and it is identical for every predicates in the tail. It suggests that the tail category with fewer samples requires more knowledge from the head for calibration.

After dual-granularity knowledge transfer, we generate synthetic samples in tail predicate using corresponding predicate and triplet features from calibrated distributions:

$$\{\tilde{\boldsymbol{x}} | \tilde{\boldsymbol{x}} \sim \Omega'_c = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}'_c)\}. \tag{9}$$

Finally, we under-sample head predicates to form a balanced dataset and input the real instances together with synthetic ones into the relation classifier for fine-tuning. Benefiting from the dual-granularity knowledge transfer, the patterns of tail predicates and their associate triplets can be enriched, thus alleviating the long-tail problem.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our method on three commonly used SGG datasets, namely Visual Genome [14], Open Image [13], and GQA [10]. For the Visual Genome dataset, we adopt the VG150 split following previous approaches [23, 29, 37, 43, 49], which contains the most frequent 150 object categories and 50 predicate categories. We use 70% of images for training, 30% images for testing and 5k images from the training set for validation. As for Open Image, we apply the Open Image V6 protocol, which has 301 object categories and 31 predicate categories. It contains

Models	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
IMP [43]$_{CVPR'17}$	61.1 / 63.1	11.0 / 11.8	37.5 / 38.5	6.2 / 6.5	25.9 / 31.2	4.2 / 5.3
VTransE [51]$_{CVPR'17}$	65.7 / 67.6	14.7 / 15.8	38.6 / 39.4	8.2 / 8.7	29.7 / 34.3	5.0 / 6.1
MOTIFS [49]$_{CVPR'18}$	66.0 / 67.9	14.6 / 15.8	39.1 / 39.9	8.0 / 8.5	32.1 / 36.9	5.5 / 6.8
G-RCNN [44]$_{ECCV'18}$	65.4 / 67.2	16.4 / 17.2	37.0 / 38.5	9.0 / 9.5	29.7 / 32.8	5.8 / 6.6
VCTREE [37]$_{CVPR'19}$	65.5 / 67.4	16.7 / 17.9	40.3 / 41.6	7.9 / 8.3	31.9 / 36.0	6.4 / 7.3
GPS-Net [22]$_{CVPR'20}$	65.2 / 67.1	15.2 / 16.6	37.8 / 39.2	8.5 / 9.1	31.3 / 35.9	6.7 / 8.6
RU-Net [24]$_{CVPR'22}$	67.7 / 69.6	- / 24.2	42.4 / 43.3	- / 14.6	32.9 / 37.5	- / 10.8
HL-Net [23]$_{CVPR'22}$	67.0 / 68.9	- / 22.8	**42.6** / **43.5**	- / 13.5	**33.7** / **38.1**	- / 9.2
PE-Net(P) [54]$_{CVPR'23}$	**68.2** / **70.1**	23.1 / 25.4	41.3 / 42.3	13.1 / 14.8	32.4 / 36.9	8.9 / 11.0
VETO [34]$_{ICCV'23}$	64.2 / 66.3	22.8 / 24.7	35.7 / 36.9	11.1 / 11.9	27.5 / 31.5	8.1 / 9.5
TDE$^\diamond$ [38]$_{CVPR'20}$	46.2 / 51.4	25.5 / 29.1	27.7 / 29.9	13.1 / 14.9	16.9 / 20.3	8.2 / 9.8
CogTree$^\diamond$ [47]$_{IJCAI'21}$	35.6 / 36.8	26.4 / 29.0	21.6 / 22.2	14.9 / 16.1	20.0 / 22.1	10.4 / 11.8
BPL-SA$^\diamond$ [9]$_{ICCV'21}$	50.7 / 52.5	29.7 / 31.7	30.1 / 31.0	16.5 / 17.5	23.0 / 26.9	13.5 / 15.6
VisualDS$^\diamond$ [46]$_{ICCV'21}$	- / -	16.1 / 17.5	- / -	9.3 / 9.9	- / -	7.0 / 8.3
NICE$^\diamond$ [17]$_{CVPR'22}$	55.1 / 57.2	29.9 / 32.3	33.1 / 34.0	16.6 / 17.9	27.8 / 31.8	12.2 / 14.4
PPDL$^\diamond$ [21]$_{CVPR'22}$	47.2 / 47.6	32.2 / 33.3	28.4 / 29.3	17.5 / 18.2	21.2 / 23.9	11.4 / 13.5
GCL$^\diamond$ [6]$_{CVPR'22}$	42.7 / 44.4	36.1 / 38.2	26.1 / 27.1	20.8 / 21.8	18.4 / 22.0	16.8 / 19.3
IETrans$^\diamond$ [50]$_{ECCV'22}$	- / -	35.8 / 39.1	- / -	21.5 / 22.8	- / -	15.5 / 18.0
INF$^\diamond$ [2]$_{CVPR'23}$	51.5 / 55.1	24.7 / 30.7	32.2 / 33.8	14.5 / 17.4	23.9 / 27.1	9.4 / 11.7
CFA$^\diamond$ [18]$_{ICCV'23}$	54.1 / 56.6	35.7 / 38.2	34.9 / 36.1	17.0 / 18.4	27.4 / 31.8	13.2 / 15.5
EICR$^\diamond$ [29]$_{ICCV'23}$	55.3 / 57.4	34.9 / 37.0	34.5 / 35.4	20.8 / 21.8	27.9 / 32.2	15.5 / 18.2
BGNN [19]$_{CVPR'21}$	59.2 / 61.3	30.4 / 32.9	37.4 / 38.5	14.3 / 16.5	31.0 / 35.8	10.7 / 12.6
SHA+GCL [6]$_{CVPR'22}$	35.1 / 37.2	41.6 / **44.1**	22.8 / 23.9	23.0 / 24.3	14.9 / 18.2	17.9 / 20.9
PE-Net [54]$_{CVPR'23}$	64.9 / 67.2	31.5 / 33.8	39.4 / 40.7	17.8 / 18.9	30.7 / 35.2	12.4 / 14.5
SQUAT [12]$_{ICCV'23}$	55.7 / 57.9	30.9 / 33.4	33.1 / 34.4	17.5 / 18.8	24.5 / 28.9	14.1 / 16.5
CaCao [48]$_{ICCV'23}$	- / -	**41.7** / 43.7	- / -	**24.0** / **25.0**	- / -	**18.3** / **22.1**
DRM w/o **DKT**	**70.2** / **72.1**	23.3 / 25.6	**44.3** / **45.2**	13.5 / 14.6	**34.0** / **38.9**	9.0 / 11.2
DRM	43.9 / 45.8	**47.1** / **49.6**	27.5 / 28.4	**27.8** / **29.2**	19.0 / 22.9	**20.4** / **24.1**

Table 1. Comparison results with state-of-the-art SGG methods on the VG150 dataset. "$\diamond$" denotes the combination of MOTIFS with a model-agnostic unbiasing strategy. The best and second best results under each setting are respectively marked in **red** and **underline blue**.

Model	R@50	WmAP		score_{wtd}
		rel	phr	
MOTIFS [49]$_{CVPR'18}$	71.6	29.9	31.6	38.9
G-RCNN [44]$_{ECCV'18}$	74.5	33.2	34.2	41.8
VCTREE [37]$_{CVPR'19}$	74.1	34.2	33.1	40.2
GPS-Net [22]$_{CVPR'20}$	74.8	32.9	34.0	41.7
BGNN [19]$_{CVPR'21}$	75.0	33.5	34.2	42.1
RU-Net [24]$_{CVPR'22}$	**76.9**	35.4	34.9	43.5
HL-Net [23]$_{CVPR'22}$	**76.5**	35.1	34.7	43.2
PE-Net [54]$_{CVPR'23}$	**76.5**	**36.6**	**37.4**	**44.9**
SQUAT [12]$_{ICCV'23}$	75.8	34.9	35.9	43.5
DRM w/o **DKT**	75.9	**40.5**	**41.4**	**47.9**

Table 2. Comparison results with state-of-the-art SGG methods on Open Image V6. The best and second best results under each metric are respectively marked in **red** and **underline blue**.

126,368, 1,183, and 5,322 images for training, validation, and testing, respectively. For the GQA dataset, we follow previous works [6, 34] and utilize the GQA200 split, which includes 200 object categories and 100 predicate categories.

Tasks. We adopt three SGG tasks for evaluation: 1) Predicate Classification (Predcls) infers the predicates of entity pairs with ground-truth bounding boxes and categories. 2) Scene Graph Classification (SGCls) aims to predict the triplet categories with ground-truth bounding boxes. 3) Scene Graph Detection (SGDet) detects bounding boxes of entity pairs and infers their predicate categories.

Evaluation Metrics. We use Recall@K (R@K) and mean Recall@K (mR@K) as evaluation metrics on VG150 and GQA200 datasets, following recent works [6, 34, 54]. R@K tends to prioritize frequent predicates, while mR@K exhibits a preference for less frequent predicates. Results on Open Image dataset are evaluated using Recall@50 (R@50), weighted mean AP of relations (wmAP_{rel}), weighted mean AP of phrase (wmAP_{phr}), and a weighted score of them $score_{wtd} = 0.2 \times R@50 + 0.4 \times \text{wmAP}_{rel} + 0.4 \times \text{wmAP}_{phr}$, following previous works [19, 23, 52].

Implementation Details. Following previous works [6, 34, 54], we adopt the pre-trained Faster RCNN with ResNeXt-101-RPN in the proposal network to detect entities in the image. GloVe [30] is applied to embed the semantic features. We set the loss weight parameters λ_r, λ_e, λ_t, and λ_p as 3, 0.5, 0.1, and 0.1, respectively. Temperatures τ_p and τ_t are set as 0.2, and 0.1 considering that the predicate feature space exhibits greater variety than its asso-

Models	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
VTransE [51]$_{CVPR'17}$	55.7 / 57.9	14.0 / 15.0	33.4 / 34.2	8.1 / 8.7	27.2 / 30.7	5.8 / 6.6
MOTIFS [49]$_{CVPR'18}$	**65.3** / **66.8**	16.4 / 17.1	**34.2** / **34.9**	8.2 / 8.6	**28.9** / **33.1**	6.4 / 7.7
VCTREE [37]$_{CVPR'19}$	63.8 / 65.7	16.6 / 17.4	34.1 / 34.8	7.9 / 8.3	28.3 / 31.9	6.5 / 7.4
SHA [6]$_{CVPR'22}$	63.3 / 65.2	19.5 / 21.1	32.7 / 33.6	8.5 / 9.0	25.5 / 29.1	6.6 / 7.8
VETO [34]$_{ICCV'23}$	64.5 / 66.0	21.2 / 22.1	30.4 / 31.5	8.6 / 9.1	26.1 / 29.0	7.0 / 8.1
VTransE+GCL [6]$_{CVPR'22}$	35.5 / 37.4	30.4 / 32.3	22.9 / 23.6	16.6 / 17.4	15.3 / 18.0	14.7 / 16.4
MOTIFS+GCL [6]$_{CVPR'22}$	44.5 / 46.2	36.7 / 38.1	23.2 / 24.0	17.3 / 18.1	18.5 / 21.8	16.8 / 18.8
VCTREE+GCL [6]$_{CVPR'22}$	44.8 / 46.6	35.4 / 36.7	23.7 / 24.5	17.3 / 18.0	17.6 / 20.7	15.6 / 17.8
SHA+GCL [6]$_{CVPR'22}$	42.7 / 44.5	**41.0** / **42.7**	21.4 / 22.2	**20.6** / **21.3**	14.8 / 17.9	**17.8** / **20.1**
DRM w/o **DKT**	**66.9** / **68.4**	18.1 / 19.0	**36.4** / **37.2**	7.1 / 7.4	**30.6** / **34.6**	6.9 / 8.4
DRM	43.2 / 44.4	**41.9** / **43.5**	23.3 / 23.9	**19.9** / **20.7**	18.6 / 21.7	**18.9** / **21.0**

Table 3. Comparison results with state-of-the-art SGG methods on the GQA200 dataset. The best and second best results under each setting are respectively marked in **red** and **underline blue**.

Module				PredCls		SGCls	
P	T	A	C	R@50/100	mR@50/100	R@50/100	mR@50/100
				56.5/60.4	15.5/17.4	37.5/39.3	10.1/11.3
✓				67.6/69.5	18.1/19.9	41.8/42.7	10.8/11.8
	✓			67.6/69.8	20.5/22.5	42.3/43.2	12.6/13.6
✓	✓			69.3/71.3	20.8/22.8	42.8/43.8	12.8/13.9
✓	✓	✓		69.8/71.6	20.4/22.3	43.4/44.4	12.8/13.8
✓		✓	✓	69.8/71.6	22.0/24.0	43.4/44.3	12.3/13.3
	✓	✓	✓	69.7/71.5	21.9/24.3	43.8/44.8	12.7/13.9
✓	✓	✓	✓	**70.2/72.1**	**23.3/25.6**	**44.3/45.2**	**13.5/14.6**

Table 4. Ablation studies on predicate and triplet cues learning. "P", "T", "A", and "C" denote the predicate cue modeling module, triplet cue modeling module, augmentation in dual-granularity constraints, and the dual-granularity constraints loss, respectively.

Module	PredCls		SGCls	
	R@50/100	mR@50/100	R@50/100	mR@50/100
None	**70.2/72.1**	23.3/25.6	**44.3/45.2**	13.5/14.6
DKT-P	42.4/44.2	45.0/47.3	26.8/27.8	26.9/28.1
DKT-T	40.4/42.2	46.1/48.7	24.3/25.3	26.4/28.0
DKT	43.9/45.8	**47.1/49.6**	27.5/28.4	**27.8/29.2**

Table 5. Ablation studies on dual-granularity knowledge transfer. "DKT-P" and "DKT-T" denote the predicate knowledge transfer and triplet knowledge transfer.

ciated triplet one. We optimize our method via SGD, using an initial leaning late of 10^{-4} with a batch size of 16.

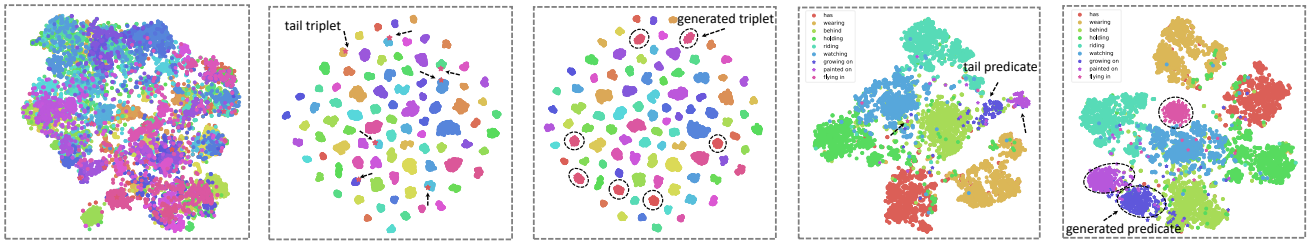
4.2. Comparison with State-of-the-art Methods

To evaluate the performance our model, we compare it with several state-of-the-art SGG approaches on Visual Genome, Open Image, and GQA datasets. The comparison methods include IMP [43], MOTIFS [49], VCTREE [37], RU-Net [24], HL-Net [23], PE-Net [54], and VETO [34], which focus on the prediction of every relationship in an image. We also compare with methods for unbiased scene graph generation, including TDE [38], CogTree [47], NICE [17], INF [2], CFA [18], EICR [29], BGNN [19], SHA+GCL [6], SQUAT [12], and CaCao[48]. In addition, we compare our method with VETO+MEET [34] in the setting without graph constraint in the Supplementary Material.

Visual Genome. Table 1 shows the comparison results of different approaches on VG150. From these results, we have the following observations: 1) our proposed method significantly outperforms all baselines on all three tasks. More specifically, our DRM w/o DKT outperforms the recent PE-Net [54] by 2.0%, 2.9%, and 2.0% at R@100 on PredCls, SGCls, and SGDet, respectively. Unlike approaches such as MOTIFS [49] and VCTREE [37], which only utilize a predicate classifier to predict predicates and overlook the triplet cues, our method leverages coarse-grained predicate cues and fine-grained triplet cues for relation recognition and thus achieves superior results. 2) Our DRM method also has considerably better performance compared to the baseline unbiased SGG methods. Notably, based on the proposed DKT strategy in DRM, our method outperforms the recent multi-expert method SHA+GCL [6] by 5.5%, 4.9%, and 3.2% at mR@100 on three tasks. This demonstrates that transferring knowledge from head predicates to tail predicates at dual granularities and enriching the patterns in tail predicates can effectively mitigate the long-tail problem.

Open Image. Compared to the VG dataset, Open Image provides a relatively complete labeling of relationships in the images. Consequently, the model's capability to generate scene graphs can be evaluated at a fine-grained level utilizing the AP metric. To evaluate the generalizability of our method across various datasets, we conduct experiments on Open Image V6. Since the metrics in Open Image V6 tend to prioritize frequent predicates, we just compare our DRM w/o DKT with state-of-the-art approaches. The comparison results shown in Table 2 indicate that our DRM w/o DKT significantly outperforms the recent approaches, *i.e.*, SQUAT [12] and PE-Net [54].

GQA. Compared to Visual Genome and Open Image datasets, GQA200 contains a broader range of predicates. So we further confirm the generality of our model on the



(a) MOTIFS, Triplet (b) DRM w/o DKT, Triplet (c) DRM, Triplet (d) DRM w/o DKT, Predicate (e) DRM, Predicate

Figure 4. The comparison of t-SNE visualization results on predicate and triplet feature distributions within the VG dataset. "MOTIFS, Triplet" and "DRM w/o DKT, Triplet" visualize the same set of samples, where each unique color represents a different type of triplet.

GQA200 dataset. As shown in Table 3, our method significantly outperforms the recent VETO [34] at R@100 on three tasks. When equipped with DKT, our DRM outperforms SHA+GCL [6] by an average of 0.4% at mR@100 across three tasks. These results demonstrate the consistent effectiveness of our method in handling relation recognition under different data distributions.

4.3. Ablation Study

To verify the contributions of different components in our DRM network, we conduct the following ablation studies.

Predicate and Triplet Cue Learning. We first perform an ablation study on the leverage of predicate and triplet cues. As shown in Table 4, we incrementally incorporate one component into the baseline to verify their effectiveness. Compared with the application of predicate or triplet cue modeling in isolation, leveraging both predicate and triplet together improves the performance. Our dual-granularity constraints include a constraint loss component and a two-view augmentation component to generate positive pairs in each training batch. The constraints compel predicate and triplet cue modeling to concentrate on corresponding granularities. We observe from the results that both the loss component and augmentation component contribute to the performance improvements.

Dual-Granularity Knowledge Transfer. As shown in Table 5, we conduct an ablation study for the DKT. We observe an obvious performance gain on mR@K when transferring knowledge at either predicate or triplet granularity. The mR@K is further improved when predicate and triplet granularities are simultaneously employed. In comparison to the scenario without the application of DKT, we observe a performance decrease in R@K when using it. This is attributed to the capability of our model to reasonably classify ambiguous head predicates, *e.g.* "on", into more specific tail predicates, *e.g.* "sitting on". Consequently, decreases at the R@K for these head predicates are inevitable [17, 54].

4.4. Visualization Analysis

To illustrate the ability our method to learn triplet cues and impact of the DKT strategy, we visualize the predicate and triplet feature distributions using t-SNE. The visualization results are shown in Figure 4. Comparing Figure 4a with Figure 4b, we observe that our DRM generates compact and distinguishable triplet representations, while MOTIFS [49] appears to overlook the triplet cues, leading to a challenge in distinguishing various triplet types. By comparing Figure 4b with Figure 4c, it can be observed that DKT can generate synthetic samples with diverse distributions for the tail triplets. Figures 4d and 4e demonstrate that DKT also transfers the knowledge of head predicates to tail predicates, increasing the tail predicate patterns. The visualization of predicate and triplet feature distributions demonstrates the interpretability of our method in leveraging predicate and triplet cues and transferring the dual-granularity knowledge.

5. Conclusion

In this paper, we propose a Dual-granularity Relation Modeling (DRM) network to address two issues in SGG, *i.e.* the diverse visual appearance within the same predicate and the lack of patterns in tail predicates. Our DRM network captures the coarse-grained predicate cues shared across different subject-object pairs and fine-grained triplet cues under specific subject-object pairs for relationship recognition. The Dual-granularity Knowledge Transfer (DKT) is further proposed to transfer the variation from head predicates to the tail to enrich the tail predicate patterns. Quantitative and qualitative experiments demonstrate that our method establishes new state-of-the art performances on Visual Genome, Open Image and GQA datasets.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0161901, the National Natural Science Foundation of China under Grants 62276018 and U20B2069, the Beijing Nova Program under Grant 20230484297, the Fundamental Research Funds for the Central Universities, and Research Program of State Key Laboratory of Complex & Critical Software Environment.

References

[1] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8102–8109, 2019. 1

[2] Bashirul Azam Biswas and Qiang Ji. Probabilistic debiasing of scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10429–10438, 2023. 6, 7

[3] Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, and Mohamed Elhoseiny. Reltransformer: A transformer-based long-tail visual relationship recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19507–19517, 2022. 2

[4] 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. 1

[5] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5213–5222, 2020. 1

[6] 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, 3, 4, 6, 7, 8

[7] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *European Conference on Computer Vision*, pages 19–34. Springer, 2020. 1

[8] Yutian Guo, Jingjing Chen, Hao Zhang, and Yu-Gang Jiang. Visual relations augmented cross-modal retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 9–15, 2020. 1

[9] 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. 6

[10] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 2, 5

[11] Tianlei Jin, Fangtai Guo, Qiwei Meng, Shiqiang Zhu, Xiangming Xi, Wen Wang, Zonghao Mu, and Wei Song. Fast contextual scene graph generation with unbiased context augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6302–6311, 2023. 2

[12] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18664–18674, 2023. 1, 6, 7

[13] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 2, 5

[14] 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. 1, 2, 5

[15] Sanjoy Kundu and Sathyanarayanan N Aakur. Is-ggt: Iterative scene graph generation with generative transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6292–6301, 2023. 1

[16] Jiankai Li, Yunhong Wang, and Weixin Li. Zero-shot scene graph generation via triplet calibration and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 1, 2

[17] 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. 3, 6, 7, 8

[18] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21685–21695, 2023. 6, 7

[19] 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, 3, 4, 6, 7

[20] 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

[21] 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. 6

[22] 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. 1, 6

[23] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19476–19485, 2022. 5, 6, 7

[24] Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. Ru-net: Regularized unrolling network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19466, 2022. 6, 7

[25] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11556, 2021. 2

[26] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 1, 2

[27] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15931–15941, 2021. 2

[28] Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, and Jingkuan Song. Fine-grained predicates learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19467–19475, 2022. 2, 3

[29] Yukuan Min, Aming Wu, and Cheng Deng. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13296–13307, 2023. 5, 6, 7

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 6

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 4

[32] 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

[33] Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. Scene graph contrastive learning for embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10884–10894, 2023. 1

[34] Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21882–21893, 2023. 2, 3, 6, 7, 8

[35] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13945, 2021. 2

[36] Shuzhou Sun, Shuaifeng Zhi, Qing Liao, Janne Heikkilä, and Li Liu. Unbiased scene graph generation via two-stage causal modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[37] 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. 1, 2, 3, 4, 5, 6, 7

[38] 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. 3, 6, 7

[39] Hongshuo Tian, Ning Xu, An-An Liu, Chenggang Yan, Zhendong Mao, Quan Zhang, and Yongdong Zhang. Mask and predict: Multi-step reasoning for scene graph generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 4128–4136, 2021. 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[41] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2020. 2

[42] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[43] 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, 5, 6, 7

[44] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pages 670–685, 2018. 1, 2, 6

[45] 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

[46] Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Wermter, and Maosong Sun. Visual distant supervision for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15816–15826, 2021. 6

[47] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *International Joint Conference on Artificial Intelligence*, pages 1274–1280, 2021. 6, 7

[48] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21560–21571, 2023. 3, 6, 7

[49] 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, 4, 5, 6, 7, 8

[50] 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 *European Conference on Computer Vision*, pages 409–424. Springer, 2022. 3, 6

[51] 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. 1, 2, 6, 7

[52] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 2, 6

[53] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[54] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792, 2023. 1, 2, 6, 7, 8