Refine and Redistribute: Multi-Domain Fusion and Dynamic Label Assignment for Unbiased Scene Graph Generation

Yujie Zang[1], Yaochen Li[2,*], Yuan Gao[3], Yimou Guo[4], Wenneng Tang[5], Yanxue Li[6], Meklit Atlaw[7]
School of Software Engineering, Xi'an Jiaotong University
[1,3,4,5,6]{zyj1176422374,gaoyuan419,guoym,twn29004,yanxue_li}@stu.xjtu.edu.cn
[2]yaochenli@mail.xjtu.edu.cn, [7]meklitm29@gmail.com

Abstract

Scene Graph Generation (SGG) plays an important role in enhancing visual image comprehension. However, existing approaches often struggle to represent implicit relationship features, resulting in a limited ability to distinguish predicates. Meanwhile, they are vulnerable to skewed instance distributions, which impairs effective training for fine-grained predicates. To address these problems, we propose a novel feature refinement and data redistribution framework (RAR). Specifically, a multi-domain fusion (MDF) module is designed to acquire comprehensive predicate representations, integrating global knowledge from the contextual domain and local details in the spatial-frequency domains. Then, we introduce a dynamic label assignment (DLA) strategy to tackle the long-tailed problem. Different predicate categories are adaptively grouped, accommodating varying training conditions. Guided by this strategy, we leverage a hierarchical auto-encoder to generate siamese samples, expanding the label cardinality. Furthermore, we explore the updated sample space to derive reliable samples and assign tailored labels, ultimately achieving the data rebalancing. Experiments on VG and GQA demonstrate that our model contributes to correcting prediction bias and achieves a significant improvement of approximately 10% in mean recall compared to baseline models.

1. Introduction

Scene Graph Generation (SGG) aims to detect a compact graphical structure that expresses rich semantic information from images. It organizes the visible objects and their inherent relationships as nodes and edges, converting the abstract visual features into intelligible linguistic descriptions. Additionally, SGG serves as a fundamental block for advanced visual tasks, including cross-modal retrieval [11], image captioning [4], and visual question answering [14]. However, existing SGG approaches encounter challenges such as imbalanced instance distribution and suboptimal performance, underlining the potential for further optimization.

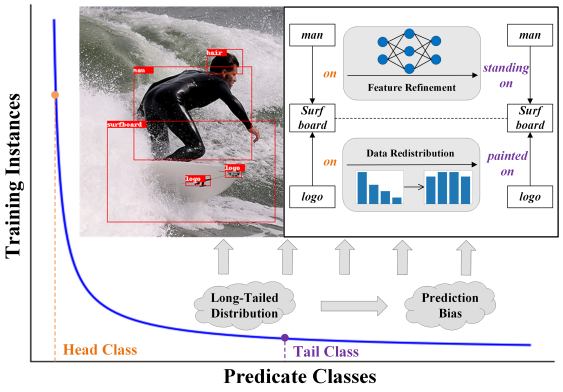

Figure 1. Two insights for eliminating bias. 1) Feature refinement involves the multi-domain fusion to enhance the predicate representation, distinguishing "on" from informative "standing on". 2) Data redistribution employs the label assignment strategy to derive balanced instances for each class, which provides sufficient training for infrequent "painting on" and facilitates accurate inference.

As research advances, the issue of long-tailed distribution [34] in training data has become increasingly serious. This is compounded by the inherent ambiguity of visual information and the scarcity of data annotations, leading to an evident bias in predicate reasoning. In response to this issue, crafting an informative scene graph has become a popular research topic, originating from the introduction of the mean recall metric [5, 26] and the release of the unbiased scene graph benchmark [25]. As shown in Fig. 1, some methods propose improvements in network design [3, 8, 9], which significantly enhance the ability of relationship generation, but they overlook the data-level processing. While techniques such as re-weighting [2, 24] and re-sampling [10] have been developed to correct the long-tailed distri-

*Corresponding author. This work is supported by Key Research and Development Program of Shaanxi Province under grant no. 2022GY-080.

bution, they may lead to overfitting in the tail classes and impair the performance of the head classes. Fig. 1 also demonstrates the impact of data balancing techniques.

Intuitively, we can integrate the above strengths into a comprehensive solution. To implement this intention, we first propose a feature refinement framework incorporating multi-domain fusion, which introduces predicate attributes from distinct perspectives. Within this, the contextual domain employs stacked Transformer to compute self-attention maps for both object and relationship sequences, thereby capturing global semantic information. Subsequently, we concentrate on the region of interest for object pairs, where the most descriptive local information about relationships is contained. We introduce a local saliency adapter: The ROIs are subdivided into smaller patches, allowing attention maps to be exchanged through sliding windows. These maps are then utilized by deformable convolutions to characterize the irregular predicate. Moving forward, we apply the adapter in spatial and frequency domains, driven by the fact that both domains can offer valuable visual insights. Especially, frequency components reflect the edge details and behavioral tendencies, making them excel in distinguishing tail classes that typically exhibit single behavioral pattern. Leveraging the adapter outputs, we feed them into feedforward networks for spatial feature encoding. Additionally, we adopt the Fourier Transform for converting them into the frequency domain, truncating high-frequency components of the adapter and encoding them through Transformer layers. Ultimately, we align both global and local information as the refined features, which exhibit both robustness and unbiasedness.

To address data imbalance, we propose a dynamic label assignment strategy that operates at the data level. Initially, we separate the training data into three exclusive groups, applying different training conditions via stepwise factors. During the training process, we continuously fine-tune group elements based on prediction performances. Subsequently, we devise an "expanding-balancing" pipeline according to the groups. 1) To alleviate the scarcity of tail group, we employ a hierarchical auto-encoder to mimic the original instances themselves and generate siamese duplicates, thereby expanding the sample cardinality and feature space. 2) Acknowledging the distribution disparities between the head group and tail group, we derive samples for the tail while removing redundant ones from the head. Specifically, we learn from the reconstructed sample space and partition the existing data into two distinct sets: the annotated samples L^+ and the unannotated ones L^-. We aggregate the limited labels from L^+ to build high-quality auxiliary-labels, which are directly utilized for optimizing the predicate classifier. Meanwhile, we reabsorb some valid instances from L^- and recombine them with pseudo-labels, which contribute to balancing the data distribution and enhancing the discriminative ability for the unbiased model.

The main contributions of our work are threefold:

- A feature refinement framework is proposed, which aggregates feature knowledge from multiple domains in parallel, forming more robust predicate representations in both global and local perspectives.
- A plug-and-play module called dynamic label assignment is introduced, which adaptively sets distinct training conditions through predicate grouping and balances data distribution via the sample derivation and label assignment.
- Our model exhibits superior capabilities in bias elimination compared to the typical baselines. Experimental results on VG and GQA validate its effectiveness and demonstrate the state-of-the-art performance.

2. Related Work

Scene Graph Generation SGG aims to achieve image understanding through structured information. Early methods adopt message passing with graphs to update themselves by combining the receptive fields of adjacent nodes [31]. Subsequent methods mostly employ recurrent neural networks to extract visual context and achieve two-stage reasoning of objects and relationships. Then, more powerful and flexible Bi-LSTM and Bi-TreeLSTM are proposed which can encode in both directions and adapt to the possible one-to-many cases of object combinations under the current scenes [26, 32]. With the improvement of logical theory, more complex models have been derived, such as hierarchical graphs [28], probabilistic graphs [30] and aware graphs [23]. However, these models still have limitations in utilizing contextual information. With the popularity of the Transformer, its feature extraction ability enables the end-to-end structure to be implemented. The Transformer framework is designed to bypass object detection and infer through entity and predicate decoders, leading to significant improvements in both speed and accuracy [8, 19].

Unbiased Scene Graph Generation The presence of long-tailed distribution in scene graph tasks has attracted significant attention, primarily focusing on eliminating performance disparities among different predicates. The mainstream approaches include re-weighting [2, 7, 13, 21], re-sampling [10, 12, 13, 20], and pseudo-label generation [18, 27, 33]. RTPB [2] proposes resistance training, which fine-tunes predicate weights based on prior biases to improve the generalizability of tail classes. DT2-ACBS [10] introduces an alternating class balance sampling strategy that better captures interactions from imbalanced entity distributions in visual relationships. Despite the potential improvement in overall model performance offered by re-weighting

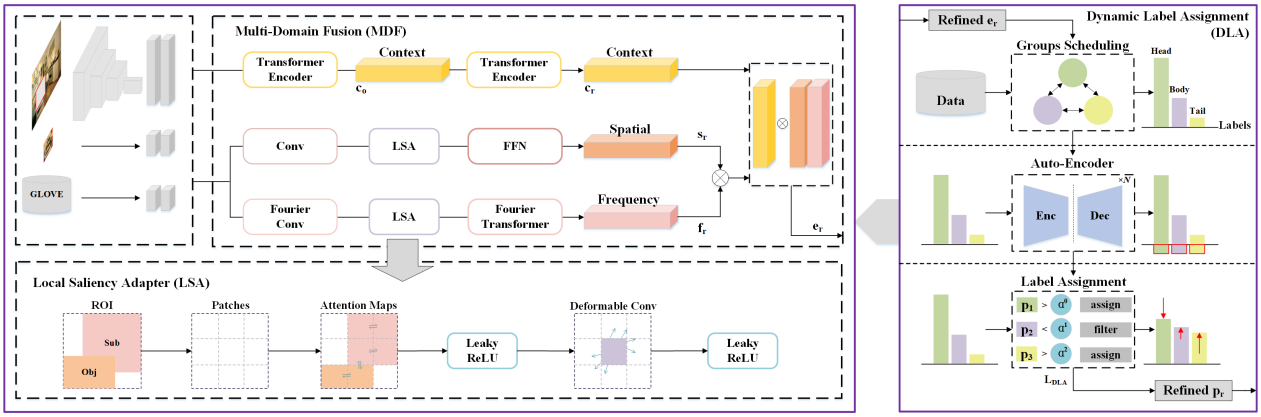

Figure 2. The overall structure of proposed method. It comprises two main components: MDF for refining predicate features, and DLA for addressing data redistribution through the provision of additional training instances.

and re-sampling, these methods often encounter the trade-off dilemma of overfitting to tail classes or underfitting to head classes. Sparse R-CNN [27], in contrast, introduces a learnable query and generates informative pseudo-labels from a siamese network. Nevertheless, this approach necessitates an additional mechanism for label generation and assignment. Distinct from the above methods, our novel approach introduces dynamic balancing to scale up valid samples and labels based on existing data, addressing the long-tailed challenges.

3. Methodology

3.1. Methods Overview

Fig. 2 illustrates the overall process of our approach. We extract global contextual features c_r, local ROI features s_r and f_r to refine visual representations (see Section 3.2). To alleviate the long-tailed distribution, we employ a dynamic label assignment strategy, which effectively achieves data redistribution for balanced training (see Section 3.3).

3.2. Element Feature Refinement

To refine object and relationship features, we introduce a framework that integrates multi-domain fusion (MDF), significantly boosting the representation learning in SGG.

We adopt the Faster R-CNN for object detection and infer the object class $\mathcal{O} = \{o_i\}_{i=1}^{N}$. To explore the relationships within candidate o_i, we devise a visual context encoder to extract the essential global information. As the Transformer network excels in establishing self-attention for o_i sequences, we merge the FPN backbone output v_o, position information pos_o, and word embeddings Emb_o to serve as its input. This process yields the contextual representation c_r, as expressed by the following formulation:

$$\begin{aligned} c_o &= Transformer_o(W_o[v_o, pos_o, Emb_o]) \\ c_r &= Transformer_r(W_r[c_{sub}, c_{obj}]) \end{aligned} \quad (1)$$

where W_o and W_r stand for fully-connected layer. c_{sub}, c_{obj} are the context of subject and object calculated by c_o.

We further delve into the exploration of o_i combinations which exhibit higher attention coefficients. Our focus lies on the predicate's regions of interest U, namely the bounding boxes union of subject and object. This region directly contains the visual features of the predicate, where we introduce a local saliency adapter (LSA) to extract: Breaking down region U into equally sized patches, we calculate attention maps within sliding windows. This strategy encourages the exchange of channel information among shuffled patches, extracting the salient features from patches with strong interactions. Due to the irregular predicate regions occupied in images, traditional rectangular boxes may not effectively capture them. To overcome this limitation, we intersperse 3x3 deformable convolution layers among the standard convolution at alternating positions, enhancing the scope of the receptive field. The deformable convolution is equipped with a 3×3 offset field, uniquely designed to accurately locate features and adapt to the changes in shape. Finally, we use linear enhancement based on MLP and add skip connections to boost the robustness of deformable features. The outcome $u_{i \to j}$ is calculated as follows:

$$u_{i \to j} = LN([Att(U(x,y)), \Delta U(x,y)] + U(x,y)) \quad (2)$$

where Att is the attention map and LN is the LayerNorm operation, ΔU is the learnable offset between subject i and object j. Each block is processed by LeakyReLU to facilitate thorough information exchange, allowing secondary extraction of local visual features for predicates.

We then utilize LSA in both the spatial and frequency domains of U, which are essential representational spaces in image comprehension. Guided by LSA, we can precisely extract cross-domain knowledge, thereby enhancing predicate representation. Specifically, in the spatial domain, we focus on shallow visual characteristics such as color and

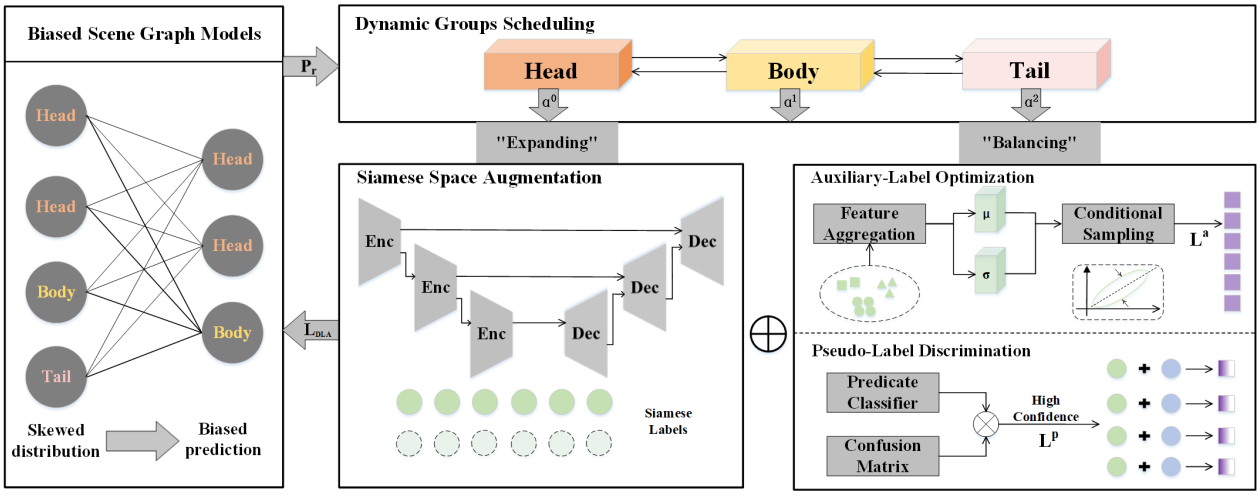

Figure 3. Illustration of the proposed dynamic label assignment strategy (DLA). We dynamically partition predicates and apply an "expanding-balancing" pipeline for data redistribution. This strategy introduces three label types: siamese labels augment the internal data, expanding instance cardinality; auxiliary-labels L^a and pseudo-labels L^p incorporate external data to achieve balanced training.

texture. Leveraging LSA and feedforward network to adaptively extract representations denoted as s_r. In contrast, we take measures to mitigate the impact of visual attributes like color in the frequency domain. This involves normalizing the region U to grayscale and resizing it to the predefined size through linear interpolation and average pooling. Next, we apply a two-dimensional Fast Fourier Transform (FFT) to shift region U into the spectral space U_F. The transformation formulation is provided below:

$$U_F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} U(x,y) W_M^{ux} W_N^{vy}, \qquad W_K^k = e^{\frac{-j2\pi k}{K}} \tag{3}$$

where (x, y) represents spatial domain pixel and (u, v) represents frequency domain pixel. Building upon the ability of high-frequency components to capture unique behavioral patterns for informative predicates, we devise extracting these components using a Gaussian filter, concurrently eliminating low-frequency components located at the center of the spectrum. Subsequently, we further adopt LSA to mine the local information via the truncated feature maps. To establish long-time dependencies in high-frequency components across different patches, we employ the Transformer layers to process the outcome of LSA, ultimately yielding the representation f_r.

The integration of knowledge from diverse domains enriches the representation space, resulting in more precise and robust features. Furthermore, through feature fusion, we align the global information c_r, local information s_r and f_r using self-attention matching, ultimately producing a refined predicate representation e_r.

3.3. Data Balance Enhancement

In this section, we alleviate the long-tailed distribution from the perspective of data redistribution. Illustrated in Fig. 3, we introduce a novel architecture named dynamic label assignment (DLA).

Dynamic Groups Scheduling To ensure balanced training, it's necessary to handle predicates differently based on their prior and posterior states. Toward this end, we initialize three mutually exclusive groups, denoted as $G_{rel} = G_{head} \cup G_{body} \cup G_{tail}$, determined by the volume of training instances. Intuitively, we expect a decline in performance from G_{head} to G_{tail}, which requires a corresponding increase in training intensity. Within this perspective, we employ the true positive rate (TP) from the confusion matrix as the metric and compare it against confidence scores in the training process, dynamically fine-tuning the members within each group: if the TP of a certain head class decreases, it will be transferred from G_{head} to G_{body} with a higher emphasis on training; in case a tail class exhibits overfitting, it will be shifted from G_{tail} to G_{body} while moderating the training intensity. Additionally, as the constraints applied on distinct groups should display diversity, we adopt a stepwise factor $\alpha \in (0, 1)$ to emulate this design. For instance, rather than using the initial confidence scores T_c equally across all groups, we iteratively modify the conditions, leading to an asymmetric set of scores $\{T_c * \alpha^i, i = 0, 1, 2\}$. The merit of this approach lies in classifying each predicate into its respective group and subsequently offering customized training.

Siamese Space Augmentation Undoubtedly, samples from tail classes are extremely rare in terms of their absolute count, making it challenging to train robust classifiers.

Hence, we consider reconstructing samples from the original data, which share identical distribution and augment the siamese sample space (SSA).

We design an auto-encoder architecture to simulate two-dimensional feature reconstruction, ensuring precise alignment between the generated siamese duplicates and the original samples. The encoder $q_\phi(z|f)$ projects the input feature f into a latent space and induces the basis vectors of the sample distribution. This block is composed of a couple of fully-connected layers, gradually reducing the channel dimensions of the feature map. Meanwhile, the decoder $q_\psi(f|z)$ assists the basis vectors in reconstructing samples from a random normal distribution, adopting dual layers.

Furthermore, we employ a hierarchical auto-encoder to better capture intricate predicate representations. We design L stacked encoder blocks, $q_\phi(z_{l+1}|z_l)$, which sequentially yield a series of latent variables $\{z_0, z_1..., z_n, f = z_0\}$. The output of z_l is passed as the input to z_{l+1} in an autoregressive manner. During the decoding stage, we introduce relation-aware skip connections between corresponding layers of the encoder and decoder to inject the inherent characteristics of predicates. These ensure the consistency of generated samples in visual content. Then, we restore the latent variables and employ a progressive optimization approach to approximate the source data. The outputs of the last layer serve as the reconstructed siamese samples, assigned labels consistent with the inputs. The overall loss is the cumulative sum of individual decoders. Each block is composed of both the Mean Squared Error for the reconstructed features and the Kullback-Leibler Divergence loss for the distinct distribution. The formulation is as follows:

$$\begin{aligned} L_{SSA}(f,\phi,\psi) &= \mathbb{E}_{q_\psi(z|f)}[log(p_\psi(f|z))] \\ &- \sum_{l=1}^{L} \mathbb{E}_{q_\phi(z_{<l}|f)}[KL(q_\phi(z_l|z_{<l},f)||p_\psi(z_l|z_{<l}))] \end{aligned} \tag{4}$$

where ϕ and ψ denote the model parameters.

Auxiliary-Label Optimization While siamese samples do contribute to augmenting in-distribution data, their category-agnostic nature prevents achieving instances balance across all predicates. To tackle this problem, we collect the annotated samples L^+ from the original data, learning category-specific knowledge to derive out-of-distribution samples and thus achieve data rebalancing.

When considering a sample, we recognize that each component of its feature vector holds individual meanings. This leads to the existence of some specific components for which variations in the attributes they represent have minimal impact on the ultimate semantic label [29]. Leveraging these components enables us to assign extra labels denoted as auxiliary-labels L^a. Particularly, we first aggregate all samples within each category. Taking category p_i as an example, we normalize the collected feature matrixes, estimating the mean μ_i as the centroid of p_i. This implies the rotation stability among categories, guaranteeing that reconstructed samples must remain close to their original center μ_i. To identify variable semantic directions for modification, we compute the intra-class covariance matrix σ_i and pinpoint the component with the largest value as the adjustable direction $\Delta\sigma_i$. σ_i provides translation variance for augmenting samples, which indicates that moderate variations will bring acceptable supplementary samples. Then, we employ conditional sampling guided by the acquired μ_i and σ_i. To establish equilibrium spanning from G_{head} to G_{tail}, we iteratively formulate sampling rates T_s regulated by the stepwise factor α and grouping results. This yields an array of values $\{T_s/\alpha^i, i = 0, 1, 2\}$, maximizing the likelihood of choosing tail samples.

The auxiliary-labels are introduced from the ground-truth distribution of L^+, closely aligning with the inherent samples and demonstrating high quality and confidence. As a result, they can be directly utilized for supervised training as a strong regularization constraint, significantly promoting the performance of classifiers. To better distinguish predicates, we introduce the notion of category margin $\delta_i = \frac{n_i^{0.5}}{\sum_{j=1}^{C} n_j^{0.5}}$, where n_i means the number of training instances in the prior statistics. By incorporating δ_i into the cross-entropy loss, the decision boundary shifts towards the more general classes. In other words, it can partially lift up the confidence scores of tail classes and alleviate the impact of imbalanced data again. We define margin-CE as follows:

$$L_{ALO} = -\sum_i L_i^a log(\frac{e^{P_i-\delta_i}}{\sum_j e^{P_j-\delta_j}}) \tag{5}$$

where L_i^a refers to auxiliary-label. P_i and P_j are the predicted probabilities.

Pseudo-Label Discrimination Owing to the issue of sparse labeling within the data, the quantity of unannotated samples in L^- far exceeds the annotated samples in L^+. However, we discover that a portion of L^- contains potential semantic relationships and can be reabsorbed as valid training samples. These samples are constructed as pseudo-labels L^p, establishing a more balanced distribution.

Specifically, our process begins with the selection of suitable samples. The probabilities of the samples are estimated by a pre-trained classifier and the corresponding thresholds $\{T_c * \alpha^i, i = 0, 1, 2\}$ are adjusted in a step-wise manner across three groups. When the predicted probability of a sample matches the category p_i and its confidence score surpasses the threshold of the group where p_i belongs, the condition is satisfied. Next, this sample is collected and assigned a preprocessed label $\widetilde{L}^p$. It can be observed that higher threshold filters out the redundant samples in G_{head}. In contrast, G_{tail} employs more lenient criteria, aiming to accommodate candidate labels as much as possible.

Model	PredCls			SGCls			SGDet		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
KERN [5]	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
GPS-Net [22]	**17.4**	**21.3**	**22.8**	10.0	11.8	12.6	6.9	8.7	9.8
DTrans [2]	15.1	19.3	21.0	9.9	12.1	13.0	6.6	9.0	10.8
MDF	17.0	20.4	22.1	**10.5**	**12.4**	**13.3**	**7.5**	**9.7**	**11.3**
Motifs [32]	11.7	14.8	16.1	6.7	8.3	8.8	5.0	6.8	7.9
+TDE [25]	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
+BPL [13]	22.6	27.1	29.1	13.0	15.3	16.2	9.7	12.4	14.4
+NICE [18]	25.6	29.9	32.3	14.3	16.6	17.9	9.1	12.2	14.4
+RTPB [2]	28.8	35.3	37.7	16.3	20.0	21.0	9.7	13.1	15.5
+Inf [1]	30.9	35.3	**38.2**	18.0	19.8	20.7	10.9	14.1	16.8
+DLA	**31.2**	**35.9**	**38.2**	**18.6**	**20.8**	**21.9**	**11.2**	**14.7**	**17.1**
VCTree [26]	13.1	16.7	18.1	9.6	11.8	12.5	5.4	7.4	8.7
+TDE [25]	18.4	25.4	28.7	8.9	12.2	14.0	8.9	9.3	11.1
+BPL [13]	23.8	28.4	30.4	15.6	18.4	19.5	9.9	12.5	14.4
+NICE [18]	24.9	30.7	33.0	16.8	19.9	21.3	9.0	11.9	14.1
+RTPB [2]	27.3	33.4	35.6	20.6	24.5	25.8	9.6	12.8	15.1
+Inf [1]	28.9	35.3	35.7	20.3	23.6	25.4	10.1	13.4	15.5
+DLA	**29.3**	**34.5**	**36.4**	**21.2**	**24.7**	**26.3**	**10.8**	**14.0**	**16.0**
RelTR [8]	22.6	28.2	31.1	14.1	18.2	19.7	9.7	14.2	16.3
EOA [6]	30.8	36.7	39.2	14.9	17.3	18.3	10.5	14.2	16.1
SQUAT [16]	25.6	30.9	33.4	14.4	17.5	18.8	10.6	14.1	16.5
RAR(ours)	**32.8**	**37.0**	**39.4**	18.9	21.7	23.1	**12.4**	**16.1**	**18.6**

Table 1. Comparisons of different approaches on VG in terms of mR@20/50/100. Three tasks of PredCls, SGCls and SGDet are evaluated in model-based, model-agnostic and complete-model modes. The sign "+" denotes the combination with the universal de-biased module.

However, given that $\widetilde{L}^p$ originate from background instances, there could be semantic ambiguity associated with the labels. To avoid this, we adopt a mixed augmentation approach. This involves choosing a sample from L^+ and another from $\widetilde{L}^p$, which share a certain degree of semantic similarity. These selected samples are then matched in proportion of β, aiming for recombining hybrid samples in both feature and label spaces. Operations are shown as follows:

$$\begin{aligned} X^p &= \beta X_s^+ + (1-\beta)\widetilde{X}_t^p \\ L^p &= \beta L_s^+ + (1-\beta)\widetilde{L}_t^p \end{aligned} \tag{6}$$

where s and t represent arbitrary pairs, with $s \in L^+$and $t \in \widetilde{L}^p$. Here, X^p denotes the generated sample while L^p is defined as the pseudo-label updated from hard label to soft label. The benefit is that even if the information of $\widetilde{L}^p$ may not be entirely reliable, a portion of L^+ is still incorporated, preventing the model from learning blindly. Lastly, we train the model by recycling pseudo-labels as a weak regularization constraint, which provide robust discriminative ability. The soft-CE loss is defined as:

$$L_{PLD} = -\sum_i \sum_k \mathbb{I}(L_{ik}^p > 0) log(\frac{e^{P_k}}{\sum_j e^{P_j}}) \tag{7}$$

where $\mathbb{I}$ is indicator function and L_{ik}^p shows whether pseudo-label L_i^p contains component p_k. P_k and P_j are the predicted probabilities.

To better coordinate the auxiliary-label optimization (ALO) and pseudo-label discrimination (PLD), we devise a positive feedback loop. The ALO functions as a pre-classifier to provide prediction proposals for the PLD. Meanwhile, the PLD supplements huge out-of-distribution samples for extensive training, which in turn contributes to the improvement of the ALO. The overall loss is formulated as $L_{ALO} + \lambda L_{PLD}$, with λ denoting a hyper-parameter.

4. Experiments

4.1. Experimental Settings

Dataset We conduct experiments on two datasets. Visual Genome (VG) [17] is the benchmark for SGG task, consisting of 108K images and 2.3M relationship annotations. Following prior works [12, 25], we adopt the most commonly used VG150 split, which contains the most frequent 150 object classes and 50 predicate classes. GQA [15] is a scene graph based visual question answering dataset consisting of over 200K images and 113K questions. We follow prior

Model	PredCls	SGCls	SGDet
	mR@50/100	mR@50/100	mR@50/100
Motifs [32]	16.4 / 17.1	8.2 / 8.6	6.4 / 7.7
VCTree [26]	16.6 / 17.4	7.9 / 8.3	6.5 / 7.4
SHA [12]	19.5 / 21.1	8.5 / 9.0	6.6 / 7.8
MDF	20.2 / 21.7	10.7 / 11.4	7.7 / 9.0
+TED [25]	21.0 / 22.5	11.3 / 12.5	8.7 / 9.9
+RTPB [2]	27.2 / 28.7	17.8 / 18.5	12.7 / 14.5
+Inf [1]	28.3 / 30.1	19.3 / 20.5	14.5 / 15.9
+DLA	**32.1 / 33.6**	**20.3 / 21.2**	**15.3 / 17.4**

Table 2. Comparisons of different approaches on GQA in terms of mR@50/100. Three tasks of PredCls, SGCls and SGDet are evaluated in model-based and model-agnostic modes.

work [12] in using the GQA200 split, selecting the most frequent 200 object classes and 100 predicate classes.

Tasks We evaluate our model on the widely used tasks: 1) Predicate Classification (PredCls). 2) Scene Graph Classification (SGCls). 3) Scene Graph Detection (SGDet).

Metric Following previous works [12, 25], we use mean Recall@K (mR@K) for unbiased SGG, which is defined as the average of each predicate in Recall@K (R@K).

Implementation Details We adopt a similar approach to the SGG benchmark [25], freezing the pre-trained Faster R-CNN and fine-tuning the multi-domain fusion network. For dynamic label assignment, we recommend a two-stage learning to prevent biased models from affecting unbiased inference. The initial partition for $G_{head}, G_{body}, G_{tail}$ is according to their instance quantity, two thresholds 20000 and 5000 are set for dividing. We initialize $T_s = 1$ and $T_c = 0.8$, signifying the sampling rate and confidence score. We use a stepwise factor $\alpha = 0.5$ to guide the group training. The SGD optimizer with learning rate and batch size of 0.0005 and 8 is employed. All experiments are conducted in Pytorch 1.9.0 and a single RTX3090.

4.2. Comparisons with State-of-the-Arts

We evaluate our proposed model against state-of-the-art approaches, categorized into three modes: model-based, model-agnostic and complete-model. From the experimental data in Tab. 1, our model achieves the best performance across the three tasks. We compare our baseline MDF with several representative approaches, including MOTIFS, VCTree and DTrans. MDF offers more precise semantic features and leads to significant metric improvements in both SGCls and SGDet. Meanwhile, in the context of our model-agnostic DLA, we contrast it with TED, BPL, NICE, RTPB, etc. DLA achieves label redistribution by creating a balanced dataset. Unlike re-weighting methods such as RTPB that utilize fixed weights for distinct categories,

Model			SGDet			
MDF†	MDF	DLA	mR@20	mR@50	mR@100	FPS
			6.0	8.1	9.6	**3.20**
✓			6.6	8.7	10.0	2.89
	✓		7.5	9.7	11.3	2.86
✓		✓	10.9	13.6	16.8	2.69
	✓	✓	**12.4**	**16.1**	**18.6**	2.67

Table 3. Ablation studies on VG, which validate the effectiveness of our proposed frameworks: MDF and DLA, where MDF† means the absence of the local saliency adapter.

Model		SGDet		
λ	SSA	mR@20	mR@50	mR@100
$\lambda = 0.5$	w/o	11.0	13.8	16.9
	w/	11.3	14.2	17.3
$\lambda = 1$	w/o	12.0	15.6	18.1
	w/	**12.4**	**16.1**	**18.6**
$\lambda = 2$	w/o	10.1	13.2	15.9
	w/	10.4	13.4	16.3

Table 4. Analysis on VG for the impact of expanding module SSA and two balancing branches, ALO and PLD, where "w/" and "w/o" respectively indicate the presence or absence of the SSA.

our dynamic strategy allows flexibly weighting. This prevents overfitting in the tail and underfitting in the head. When compared to re-sampling approaches like BPL, we introduce an extensive dataset of newly generated out-of-distribution labels. This approach surpasses the effectiveness of recurrently sampling from the original data, resulting in enhanced unbiasedness. The experiments conducted on MOTIFS+DLA and VCTree+DLA reveal our significant performance. Moreover, our holistic framework RAR incorporates the MDF and DLA. In contrast to the one-stage Transformer architecture RelTR, RAR leverages two-stage debiasing and more extensive data adjustments, leading to superior performance in SGDet. However, we have also observed a minor decrease in SGCls results. This can be attributed to the static object features provided by SGCls, which replace the refined fusion features from MDF, resulting in the loss of crucial information.

To validate the generalization performance of our model, we also conduct experiments on GQA, which is considered more challenging with a severe long-tailed distribution. The results are presented in Tab. 2. Our MDF and DLA continue to outperform the state-of-the-art approaches.

4.3. Ablation Studies

Analysis for Model Components As aforementioned, we mainly incorporate two modules in our approach. To evaluate the effectiveness of each part, we conduct ablation experiments on VG and report the model performance in Tab. 3. "✓" denotes that the module is employed. Compared to the simple baseline, our modified MDF extracts features from three distinct domains, guaranteeing complementary representations. This contributes to a 0.5% performance enhancement. Then, local saliency adapter finetunes the ROIs, focusing on essential pixels and generating an almost 1% performance boost. Moreover, the plug-and-play DLA significantly improves performance by over 2% through data-level operations. We also analyze the algorithmic complexity and computational load of the pipeline, revealing that our implementations effectively double performance with a relatively moderate time investment.

We further delve into the principles of DLA. Firstly, we notice that the intention of SSA is the augmentation of sample cardinality. Compared to solely relying on existing limited data, siamese samples enhance the feature space, providing stronger support for classifier training. This enhancement is evident in Tab. 4, confirming a 0.5% improvement. Additionally, the nucleus of data redistribution lies in the ALO and PLD branches. ALO optimizes the margin-CE loss via feature aggregation, while PLD employs label recombination to drive discriminative learning based on soft-CE loss. To ensure effective fusion, we iteratively adjust their proportion λ, finding that $\lambda = 1$ delivers optimal outcomes. Namely, regardless of which branch takes the dominant position, it would render the regularization constraints ineffective, leading to inaccurate label assignment. Furthermore, we compute the mR@100 for the head, body, and tail groups, resulting in values of 19.2, 18.2, and 18.1, respectively. This emphasizes the collaboration of the DLA again, which prevents the head classes from experiencing underfitting due to prolonged inadequate training. Conversely, it ensures timely adjustments for tail classes, mitigating the potential occurrence of overfitting.

Analysis for Hyper-parameters The stepwise factor α plays a pivotal role in DLA, determining the execution conditions of distinct groups. As shown in Tab. 5, we employ two strategies for applying α: iterative and linear. In the linear approach, the initial T_c uniformly varies with α, while the iterative method follows an exponential change. For instance, $\{T_c, T_c - \alpha/2, T_c - \alpha\}$ or $\{T_c, T_c * \alpha, T_c * \alpha^2\}$. It's crucial to control both the step size and its variation rate. Larger step size is more adaptable to the long-tailed distribution, displaying significant results when $\alpha = 0.5$. As for the variation rate, constraints should be more relaxed when moving towards tail classes. Thus, utilizing a non-uniform approach ensures distinctiveness across different groups.

Additionally, we design pseudo-labels with a mixed ratio β between real and preprocessed labels. Given the uncertainty linked to the latter, we prioritize real labels. Experiments validate our assumption for $\beta = 0.7$. Lastly, we set the auto-encoder layers $L = 3$. This choice balances against the risk of overfitting in deeper networks and the challenge of reconstructing high-quality labels with shallow layers. Details are reported in supplementary material.

α	Strategy	SGDet		
		mR@20	mR@50	mR@100
$\alpha = 0.3$	Linaer	10.8	13.2	16.3
	Iterative	10.3	12.8	15.3
$\alpha = 0.5$	Linaer	11.3	14.5	17.2
	Iterative	**12.4**	**16.1**	**18.6**
$\alpha = 0.7$	Linaer	11.5	14.6	17.2
	Iterative	11.1	14.0	16.7

Table 5. Analysis on VG for the impact of the stepwise factor α. Two strategies are employed for its implementation.

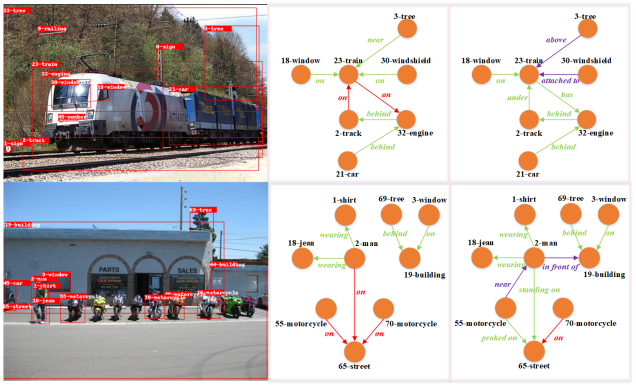

Figure 4. The visualization results for MDF and MDF+DLA. Green predicates represent correct matches with the ground-truth, while red ones are incorrect. Purple predicates represent acceptable predictions generated by our model (not in ground-truth).

Qualitative Results We visualize the results on VG. As shown in Fig. 4, our model ensures mostly accurate predictions through MDF and DLA.

5. Conclusion

In this paper, we tackle two critical challenges in generating unbiased scene graphs. We propose a domain fusion network that leverages the deformable adapter for spatial-frequency feature refinement, improving the representation performance. Additionally, we introduce a dynamic label assignment strategy that generates balanced instances and assigns reliable labels to alleviate the long-tailed distribution. Our model outperforms the state-of-the-art methods.

References

[1] 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

[2] Chao Chen, Yibing Zhan, Baosheng Yu, Liu Liu, Yong Luo, and Bo Du. Resistance training using prior bias: toward unbiased scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 212–220, 2022. 1, 2, 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. 1

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

[5] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 1, 6

[6] Zhanwen Chen, Saed Rezayi, and Sheng Li. More knowledge, less bias: Unbiasing scene graph generation with explicit ontological adjustment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4023–4032, 2023. 6

[7] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, and othrs. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 2

[8] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 6

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

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

[11] Helisa Dhamo, Azade Farshad, Iro Laina, et al. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5213–5222, 2020. 1

[12] 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, 6, 7

[13] Yuyu Guo, Lianli Gao, Xuanhan Wang, et al. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16383–16392, 2021. 2, 6

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

[15] 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. 6

[16] 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. 6

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[18] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18878, 2022. 2, 6

[19] 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

[20] 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. 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. 2

[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. 6

[23] Yichao Lu, Himanshu Rai, Jason Chang, et al. Context-aware scene graph generation with seq2seq transformers. In *IEEE/CVF international conference on computer vision*, pages 15931–15941, 2021. 2

[24] 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. 1

[25] 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. 1, 6, 7

[26] 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, 6, 7

[27] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19437–19446, 2022. 2, 3

[28] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *ECCV 2020: 16th European Conference, Proceedings, Part XIII 16*, pages 222–239. Springer, 2020. 2

[29] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3733–3748, 2021. 5

[30] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021. 2

[31] 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 (ECCV)*, pages 670–685, 2018. 2

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

[33] Ao Zhang, Yuan Yao, Qianyu Chen, et al. Fine-grained scene graph generation with data transfer. In *ECCV 2022: 17th European Conference, Proceedings, Part XXVII*, pages 409–424. Springer, 2022. 2

[34] 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. 1