Computational Visual Media
https://doi.org/10.26599/CVM.XXXX

Research Article

Relationship Representation Diversity Enhancement for Scene Graph Generation

Yunqing He[1], Ruichao Hou(✉)[1], Jia Bei[1], and Tongwei Ren[1]

Abstract Prevailing Scene Graph Generation (SGG) approaches primarily focus on long-tail problem from the perspective of semantic labels, such as by designing unbiased model architectures or employing balanced sampling techniques. However, concentrating solely on the semantic biases neglects images that display substantial visual dissimilarities yet bear analogous semantics, thereby inducing ambiguity of feature representations. To bridge the gap between semantic features and diverse visual content, we propose a lightweight method called Relationship Representation Diversity Enhancement (RDE) to facilitate the training process of SGG models. To simultaneously account for informative visual cues and robust semantics, RDE adopts parametric reconstruction to disentangle the relationship representation into the mean and standard deviation parameters of a Gaussian mixture model. We validated the effectiveness of RDE by integrating with several typical SGG approaches during the training phase on the Visual Genome dataset. The experimental results show that RDE significantly improves the performance of existing approaches without any additional inference cost or model structure modification.

Keywords Relationship representation diversity, scene graph generation, feature decoupling, visual-semantic representation fusion.

1 Introduction

Scene Graph Generation (SGG) focuses on the semantics of image content expression, which plays an important role in the understanding of cross-modal visual scenes [1, 2]. Due to significant advancements in object detection, object representations have achieved a high degree of efficacy and are proven to be crucial to SGG performance [3, 4]. Relationship representations, akin to object representations, are equally essential to the efficacy of SGG. Nevertheless, it is still an open challenge to produce high-quality relationship representations. Inferior relationship representations result in increased intra-class variance and ambiguous inter-class gap, complicating the classifier's ability to establish precise decision boundaries. This demonstrates the effectiveness of recent approaches that are equipped with heavy classifiers [5, 6]. Existing research predominantly emphasizes the quality of relationship representations through prototype learning [7, 8]. However, while prototype learning ensures only the consistency of semantics, it fails to foster representation diversity. As shown in Figure 1, neglecting visual distinctions neutralize the discriminability of relationship representations in various visual scenarios.

Concretely, the relationship representations without diversity constraints suggest two drawbacks. Firstly, according to information bottleneck theory, the encoder consistently eliminates a substantial amount of visual information to ensure that relationships within the same category have highly similar embeddings, which hinders understanding of varying visual scenes [9]. Secondly, the visual information retained by the encoder is predominantly representative of head classes, thus indicating the significant long-tail problem, and leading to conspicuous intra-class diversity and ambiguous inter-class boundaries at the feature level, especially when dealing with similar images with distinct semantics [10–13]. Meanwhile, incorporating both the robustness and diversity of representations presents a trade-off problem. The challenge stems from the heterogeneity of visual image content and semantic relationship labels. In other words, similar images can convey disparate semantics and the visual characteristics of the same interaction can exhibit significant variation

1 State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, China. E-mail: heyq@smail.nju.edu.cn, rchou@nju.edu.cn, beijia@nju.edu.cn, rentw@nju.edu.cn.

Manuscript received: 2025-09-29; accepted: 2025-12-21

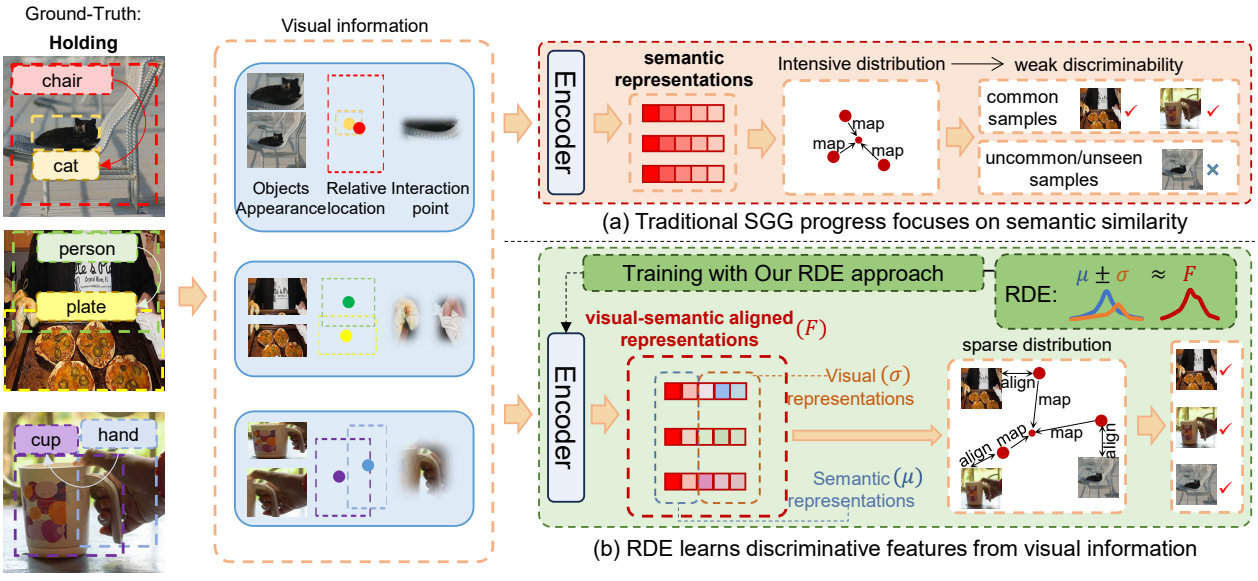


Fig. 1 Motivation of relationship representation diversity enhancement. The small red dots denote semantic prototypes, and the bigger dots represent relationship feature vectors of different instances. (a) Neglecting visual distinctions neutralizes the discriminability of relationship representations. (b) Diverse relationship representations benefit rare samples by enabling larger and sparser latent space.

between images. Representation refinement approaches based on prototypes emphasize the importance of semantic consistency [7, 14, 15]. Consequently, the decision boundary of each category tends to be excessively rigid, leading to confusion when the visual content undergoes significant changes. Alternatively, if excessive visual information is incorporated for diversity, the classifier will struggle with the gap between sparse visual information and dense semantics, adversely impeding classification performance [16, 17]. Considering that relying solely on either semantic consistency or visual diversity proves insufficient, it requires to propose an approach to balance informative visual cues and robust semantics within relationship representations.

In this paper, we propose a novel method, Relationship Representation Diversity Enhancement (RDE), that explicitly supervises the training process of relationship representation generation. To enhance relationship representations, the characteristics of ideal optimal representations are first analyzed. It is posited that high-quality representations encompass two fundamental properties: (a) *Precision*: the capacity to accurately present semantics; (b) *Diversity*: the ability to convey unique details of visual content. In other words, the optimal representation can be viewed as a visual-semantic aligned fused representation. Given that recent methodologies have adequately addressed the semantic precision problem of representations, the proposed RDE method focuses on enhancing representation diversity.

Specifically, the relationship representations are conceptualized as a Gaussian mixture model (GMM) to achieve a balance in the integration of visual and semantic representations. Firstly, a feature disentangling network, FDVAE, is developed based on variational auto-encoder (VAE) architecture, which aids in differentiating between visual-irrelevant semantic features and susceptible visual features to facilitate representation modeling. By decoupling relationship representations, we ensure a more focused representation of the underlying semantics while simultaneously allowing for the incorporation of intricate visual cues. Subsequently, a prototype learning module is introduced to enforce constraints on the semantic representations. Finally, an information maintenance module is proposed that aligns the distribution of visual representations with the comprehensive visual content.

Our contributions are summarized as: (a) We propose a novel and lightweight plug-and-play method RDE to enhance the diversity of the relationship representations in SGG without additional inference cost or baseline modification, which explores finer-grained visual-semantic aligned representations based on instance-level visual content. (b) We propose three key modules, consisting of a prototype learning module for representation robustness preservation, an FDVAE network for diverse representation disentangling, and an information maintenance module for visual information integration, to significantly improve the efficacy of SGG by modeling diversity-enhanced relationship representations

with various visual details.

2 Related Work

2.1 Scene Graph Generation

A highly discussed topic in recent SGG research is the problem of data bias, which highlights the significant challenge posed by long-tailed distributions [18]. Some recent work addresses the issue of data bias through diverse strategies, like relationship correlations finding [19–22], network architecture improvement [6, 23–25] and data augmentation [26, 27]. Prototype learning and decoupling learning are also recognized as important strategies to facilitate unbiased SGG research. The first attempt to incorporate prototype learning into SGG is documented as a memory machinism [14]. To further exploit the representation properties of scene graphs, HLB concentrates on suppressing message passing of heterogeneous nodes, emphasizing the feature distribution of relationships, and constructing implicit soft prototypes [15]. The proposal of PENET represents a substantial advancement in illustrating the efficacy of the prototype-based approach [7]. It seeks to construct semantic prototypes for both objects and relationships. Subsequently, more prototype-based approaches have been proposed. Chen *et al.* constructs three prototype centers for each relationship for a stronger prototype representation capability [8]. Zhang *et al.* introduces a more complex prototype strategy on concepts to model relation representations [28]. Decoupled learning for SGG is also a rising field. Tao *et al.* emphasize the importance of accurate object labeling and decompose object representations, allowing relationship prediction to benefit from precise object labeling [4]. Recent studies have also delved into spatial relationship modeling within 3D scenes. ScenePalette models multiplex relations among 3D objects to facilitate contextual exploration [29]. Zhang et al. utilized spatial relation priors to expedite 3D indoor scene synthesis [30]. These investigations underscore the significance of spatial and relational diversity in 3D environments, which resonates with our emphasis on representation diversity in common scene graphs.

Different from the existing approaches, we focus primarily on the balance of stable semantics and diverse visual information, rather than merely the consistency of semantic representations.

2.2 Feature Disentangling Learning

The primary objective of feature disentangling is to identify component of task-specific characteristics that are robust to varying inputs. In several fields, feature disentangling learning represents significant performance, including human pose reconstruction, object classification, and action recognition tasks [31–33]. In the context of disentangling-based SGG, a notable challenge arises from the substantial visual differences, resulting in significant intra-class variations and inter-class ambiguity. Existing work addresses the issue of visual variance by isolating the semantics of objects and relations [4]. The central concept of this disentanglement paradigm is that various visual contents can convey unique semantic meanings, thereby pushing the feature representations towards a finer-grained fashion. However, it overlooks the possibility that identical semantics can manifest itself in a variety of visual contexts.

Consequently, we propose a disentangling strategy, FDVAE, that emphasizes the disentanglement of relationship representations. The proposed FDVAE not only suggests the disentangling of visual and semantic representations, but also examines the implications for the quality of relationship representations when decoupled representations are recombined to varying degrees.

3 Our Method

3.1 Problem Formulation

A scene graph is composed of a set of relation triplets $\langle \text{subject}, \text{relationship}, \text{object} \rangle$. In our method, we focus on improving the quality of the relationship representations $F = \{f_1, \ldots, f_K\}$ derived from relationships in relation triplets. Here, K equals the total number of different types of relation triplets, *i.e.*, the relationship in each type of relation triplet is represented separately. Assuming that well-learned representations consist of stable semantic facts and informative visual cues, we formulate the relationship representation distribution as a GMM, *i.e.*, $f_k \sim GMM(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$. Here, $\boldsymbol{\mu}_k$ denotes the vector of the mean value of parameterized f_k; $\boldsymbol{\sigma}_k$ denotes the vector of the variation of parameterized f_k. $\boldsymbol{\sigma}_k$ is further divided into two components, $\boldsymbol{\sigma}_k^s$ and $\boldsymbol{\sigma}_k^o$, to depict the variation in relationship features caused by the semantic category of the subject and the semantic category of the object, respectively. For example, the visual representation of the "holding" relationship may differ when the subject is "person" versus "dog", which is captured by $\boldsymbol{\sigma}_k^s$. We refining f_k to ensure that $\boldsymbol{\mu}_k$ remains consistent with the underlying semantic information, while $\boldsymbol{\sigma}_k$ represents diverse visual content.

3.2 Feature Disentangling Module

We design a VAE-based network, FDVAE, for the parametric disentangling of relationship representations. The training

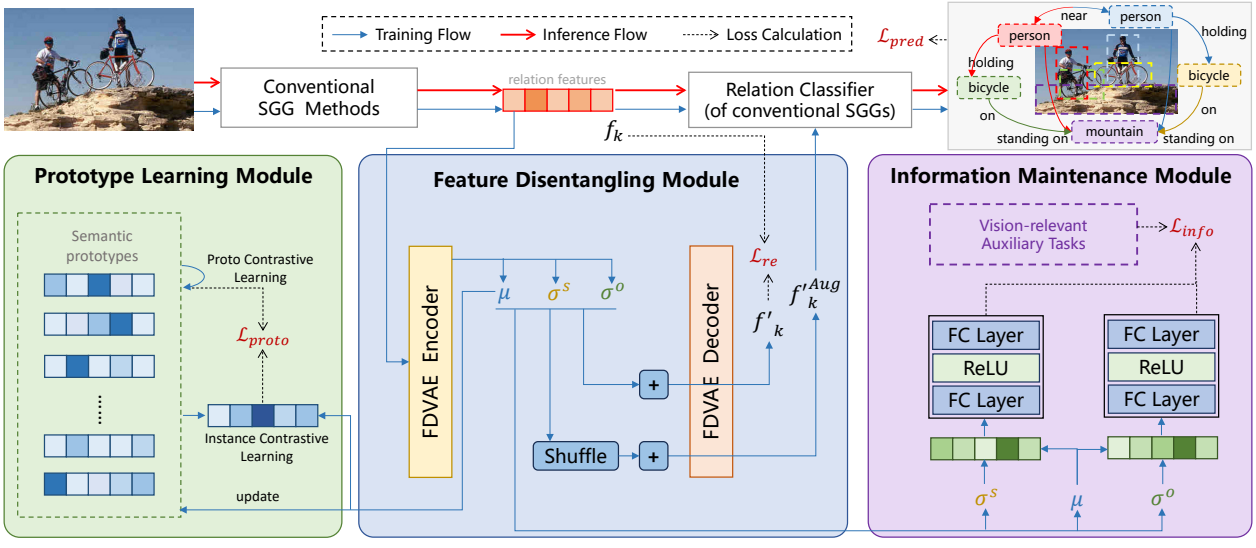


Fig. 2 The overview of RDE. Feature disentangling module decomposes relation features generated by conventional SGG approaches into parameters of GMM. The prototype learning module supervises the generation of mean values. The information maintenance module ensures that the variance values keep rich visual cues. Relation Classifier is inherited from conventional SGG approaches with shared parameters. $\mathcal{L}$ in red indicates loss function.

target of FDVAE is to learn a reversible progress for modeling f_k, *i.e.*, FDVAE can translate f_k into a combination of $\boldsymbol{\mu}_k, \boldsymbol{\sigma}^s_k, \boldsymbol{\sigma}^o_k$, and vice versa:

$$f_k \leftrightarrow \boldsymbol{\mu}_k + \boldsymbol{\sigma}^s_k + \boldsymbol{\sigma}^o_k. \tag{1}$$

We then present the network structure of FDVAE and illustrate the rationale of the underlying design, specifically focusing on the aspect of information transmission. The iterative encoder network consists of six stacked encoder blocks. This fixed number of blocks is selected based on preliminary experiments to balance representation quality and computational cost. As depicted in Figure 3, we illustrate the information transmission route in FDVAE of a given relationship sample. The encoder component of FDVAE is responsible for decomposing the representations of the input relationships f_k, which encode various forms of information $\mathcal{I}$, into various parameters of the GMM. These parameters represent distinct types of information $\mathcal{I}$, presented in the red dotted box in Figure 3. The original representation f_k comprises the information $\mathcal{I}(S_r, V_s, V_o, V_\delta)$, where S_r presents the semantic label of the relationship, V_s and V_o stand for the visual cues of the subject and the object, respectively. In addition, some negligible visual details, denoted as V_δ, which make minimal contributions to the quality of feature representations will be discarded. Based on the principles of information bottleneck theory, deep neural networks strive to concentrate on the most informative attributes while discarding extraneous or duplicate ones [17]. Thus, we can model the forward propagation of the encoder as a process of information attenuation. As shown in Figure 4, we design the encoder network as an information attenuation network. Considering that the removal of information, *e.g.*, sparsification, is not robust for training and could lead to the collapse of model parameters, we employ an equivalent process to weaken unimportant information by repeatedly emphasizing important ones. The progress of the encoder network can be formulated as follows:

$$\begin{aligned} \mathcal{X}^{(0)} &= f_k, \\ \mathcal{X}^{(l)} &= \mathcal{X}^{(l-1)} + \Delta\mathcal{I}, \end{aligned} \tag{2}$$

where $\mathcal{X}^{(l)}$ is the output of the l-th encoder block, and also the input of the $(l+1)$-th encoder block; $\Delta\mathcal{I}$ is an updater of $\mathcal{X}$ by extracting relevant information from the reference vectors. Specifically, the encoding process of $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ can be formulated as follows:

$$\begin{aligned} \mathcal{X}^{(l)}_{\boldsymbol{\mu}_k} &= \mathcal{X}^{(l-1)}_{\boldsymbol{\mu}_k} + g(Attn(\mathcal{X}^{(l-1)}_{\boldsymbol{\mu}_k}, \mathcal{P}_k) \cdot \mathcal{P}_k), \\ \mathcal{X}^{(l)}_{\boldsymbol{\sigma}_k} &= \mathcal{X}^{(l-1)}_{\boldsymbol{\sigma}_k} + g((1 - Attn(\mathcal{X}^{(l-1)}_{\boldsymbol{\sigma}_k}, \boldsymbol{\mu}_k)) \cdot \boldsymbol{\mu}_k), \end{aligned} \tag{3}$$

where $\mathcal{P}_k$ is the prototype of $\boldsymbol{\mu}_k$, and $g(\cdot)$ is a basic non-linear unit, including a Linear layer and a ReLU layer. The function $Attn(\cdot,\cdot)$ denotes the computation of the cross-attention map between the input vector $\mathcal{X}$ and the reference vector ref, which indicates the semantic similarity between these vectors. For $\mathcal{X}_{\boldsymbol{\mu}_k}$ and $\mathcal{P}_k$, the outcome of $Attn(\mathcal{X}_{\boldsymbol{\mu}_k}, \mathcal{P}_k)$ is simply employed as a standard cross-attention layer. Conversely, when addressing $\mathcal{X}_{\boldsymbol{\sigma}_k}$ and $\boldsymbol{\mu}_k$, the objective is to extract

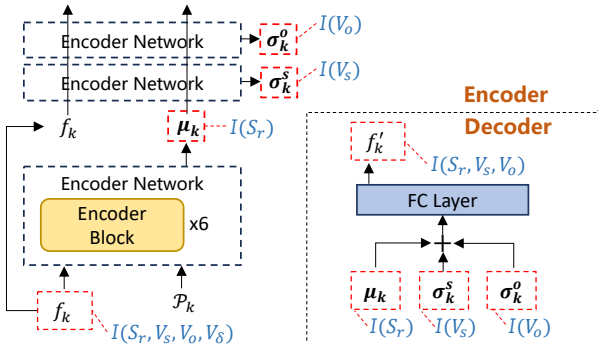


Fig. 3 Design of our proposed FDVAE network.

information from $\boldsymbol{\sigma}_k$ that is unrelated to $\boldsymbol{\mu}_k$. Consequently, a negation operation is applied to the attention matrix, as illustrated in Eq(3).

In addition, the role of the decoder is to recombine GMM parameters to reconstruct the original input relation features, effectively restoring the decoupled information. The decoder is designed with three key considerations. First, reversibility is ensured by directly incorporating the parameterized components, and a simple linear layer is introduced as sampling generation. Second, random noise is avoided. In general, a conventional VAE network incorporates a random sampler, and the decoding process can be denoted as $f'_k = \epsilon * (\boldsymbol{\mu}_k + \boldsymbol{\sigma}_k)$, to enhance the diversity of the reconstructed features. Unlike conventional VAEs that rely on random samplers to introduce noise for enhancing diversity, our $\boldsymbol{\sigma}_k$ encodes meaningful visual variations derived from individual instances. Consequently, we omit the random sampler to prevent the dilution of useful visual cues, and our decoding progress can be formulated as $f'_k = \boldsymbol{\mu}_k + \epsilon * \boldsymbol{\sigma}_k$. Third, to improve the robustness of the reconstructed relationship representations, we further adopt a shuffle operation, as shown in Figure 2, to combine $\boldsymbol{\mu}_k$ with random $\boldsymbol{\sigma}$ as f''_k. For each input data pair ($\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k$) in a training batch, we construct two training samples for decoder input, including the original data pair ($\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k$) and the shuffled data pair ($\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_j$). Specifically, $\boldsymbol{\sigma}_j$ in the shuffled pair is randomly selected from the $\boldsymbol{\sigma}$ parameters of other samples within the same batch.

3.3 Prototype Learning Module

The prototype learning module primarily employs a contrastive learning strategy to ensure the discrimination of prototypes. For the relationship representation optimization, we aim to minimize the distance between each parameterized feature $\boldsymbol{\mu}_k$ and its corresponding prototype $\mathcal{P}_k$. The semantic prototypes $\mathcal{P}_k$ are learnable vectors initialized with random values. During training, $\mathcal{P}_k$ are updated iteratively along

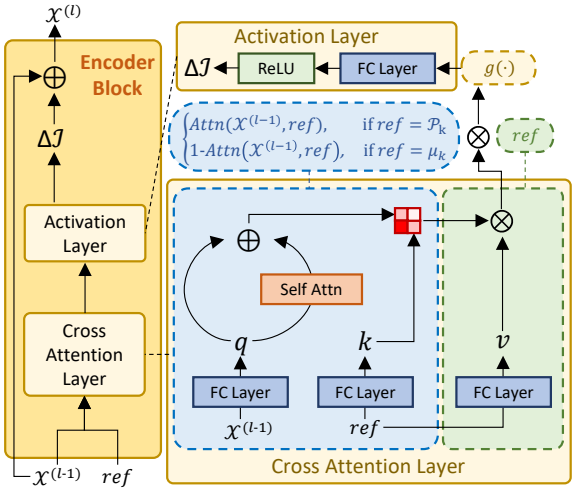


Fig. 4 Detailed design of the encoder block in FDVAE.

with other model parameters to minimize the prototype contrastive loss. This joint training ensures that the prototypes are dynamically aligned with the semantic characteristics of the relationship categories. In practice, the distance $\mathcal{D}$ between $\boldsymbol{\mu}_k$ and $\mathcal{P}_k$ is defined as follows:

$$\mathcal{D}(\boldsymbol{\mu}_k, \mathcal{P}_k) = L^2(\boldsymbol{\mu}_k) \cdot L^2(\mathcal{P}_k) + ||\boldsymbol{\mu}_k - \mathcal{P}_k||^2, \quad (4)$$

where L^2 is a standard L2 normalization function. Following L2 normalization, the dot product can be utilized to measure the cosine similarity between $\boldsymbol{\mu}_k$ and $\mathcal{P}_k$. Meanwhile, the magnitude of $\boldsymbol{\mu}_k - \mathcal{P}_k$ serves as a metric to evaluate the Euclidean distance. Furthermore, due to the highly imbalanced distribution of real samples, alignment between instances and prototypes leads to significant bias. In other words, the data bias tends to push a large number of head samples away from tail samples, while the distance between tail samples is challenging to increase. To address this data bias problem, we also employ a similar contrastive learning process between prototypes to avoid mode collapse. The process of contrastive learning is guided by a prototype learning loss, denoted as $\mathcal{L}_{proto}$. This loss function is a composite of $\mathcal{D}(\boldsymbol{\mu}_k, \mathcal{P}_k)$ and $\mathcal{D}(\mathcal{P}_k, \mathcal{P}_{\sim k})$. Here, $\mathcal{D}(\cdot, \cdot)$ represents the distance metric as defined in Eq.(4), which incorporates both cosine similarity and Euclidean distance. The term $\mathcal{P}_{\sim k}$ refers to the set of all prototypes excluding $\mathcal{P}_k$. Finally, the initialization strategy for prototypes is implemented using random vectors. Sometimes, word vectors from a pre-learned vocabulary are employed for initialization [34]. However, while initialization remarkably influences non-learning prototype strategies, such as the exponential moving average, it has minimal impact on learning-based prototype updating strategies [4].

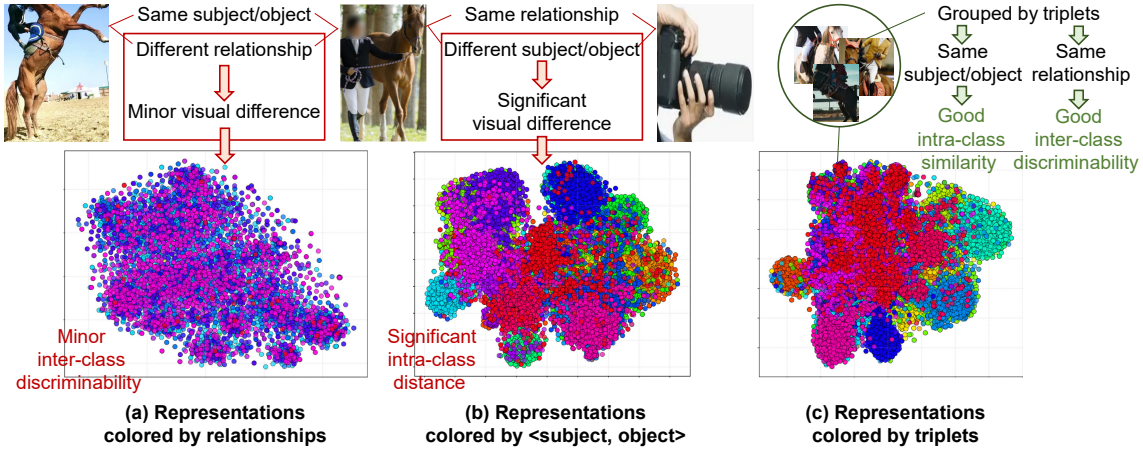


Fig. 5 Motivation of the Information Maintenance Module. (a) Different relationships with same subject/object are hard to distinguish. (b) Same relationship with varying subject/object have significant visual difference. (c) IMM introduces triplets to model relationship representations to ensure similar semantics share similar visual representations.

3.4 Information Maintenance Module

The objective of the information maintenance module is to model the distribution of relationship representations by learning specific distribution parameters $\boldsymbol{\sigma}_k$. Even within the same category of relationships, there are notable differences in visual content. These substantial visual variations are primarily influenced by relation triplets. As illustrated in Fig. 5, we performed a statistical analysis on approximately 30k samples and employed T-SNE to visualize their visual representations, thereby further validating the significance of relation triplets. When representations are grouped by different predicates, the clusters exhibit ambiguous separability. A comparison between Fig. 5(a) and Fig. 5(b) indicates that substantial visual variations are primarily introduced by the subject/object rather than the predicate. We aim for IMM to model the distribution of relationships such that the same predicate semantics can share similar representations. To this end, we introduce triplets encompassing the subject, object, and predicate to simultaneously account for both semantic and visual differences. In this context, the information maintenance module enables a diverse understanding of visual content by explicitly modeling the latent space of relationships corresponding to the relation triplets.

Specifically, for a given relation triplet k, we propose that the relationship representation f_k should follow a GMM distribution GMM_k. With N relationship classes and M object classes, we can construct M^2N distinct relation triplets, resulting in M^2N distinct GMM distributions. We define GMM sparsity as the number of distinct GMM distributions, indicating that the sparsity of relationship representation constructed by this module is M^2N. Furthermore, we observe that, in most cases, merely using the categories of objects can determine the predominant visual content, while other visual information contributes minimally to understanding relationships beyond the scope of a relation triplet region. Finally, we can use the subject/object classification task to regulate the $\boldsymbol{\sigma}_k$ as follows:

$$\mathcal{L}_{info} = \text{argmin}(w_2 \cdot ReLU(w_1 \cdot \boldsymbol{\sigma}_k), m), \tag{5}$$

where m is the one-hot distribution of subject/object classes, and $\mathcal{L}_{info}$ is the cross entropy loss of object classification. We also explore alternative GMM modeling strategies with varying sparsity, which are further elaborated in the ablation study section.

3.5 Training Strategy

The training loss in RDE is simply designed as a combination of the losses in each branch. As shown in Figure 2, the overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{proto} + \mathcal{L}_{re} + \mathcal{L}_{info} + \mathcal{L}_{pred}, \tag{6}$$

where $\mathcal{L}_{proto}$ pulls in the distances between μ_k and $\mathcal{P}_k$, and pushes out the distances between prototypes simultaneously; $\mathcal{L}_{re}$ is the reconstruction loss, which is a combination of MSE and cosine similarity loss functions, to ensure that the parameterization process of the FDVAE is reversible; Finally, $\mathcal{L}_{pred}$ is the conventional relationship classification loss.

4 Experiments

4.1 Datasets and Evaluation Metrics

We leverage the Visual Genome-150 (VG-150) as our primary benchmark dataset [46]. It comprises annotations for

Table 1 Comparison results of **mR@K** (K=20, 50 and 100) on VG-150 dataset. The **bold** values stand for the best results based on the same baseline approach, and the underlined values represent the sub-optimal results.

	PredCls			SGCls			SGDet		
	mR@20	**mR@50**	**mR@100**	**mR@20**	**mR@50**	**mR@100**	**mR@20**	**mR@50**	**mR@100**
ReIDN [35]	-	15.8	17.2	-	9.3	9.6	-	6.0	7.3
GBNet-β [36]	-	22.1	24.0	-	12.7	13.4	-	7.1	8.5
EBM [37]	19.9	26.7	30.0	13.9	18.2	20.5	7.1	9.7	11.6
NARE [27]	22.2	28.1	30.6	17.8	22.0	23.6	8.4	10.3	11.5
PPDL [19]	-	33.0	36.2	-	20.2	22.0	-	12.2	14.4
DeC- [4]	24.1	32.6	35.2	15.0	18.3	19.1	9.5	12.8	15.3
SQUAT [38]	25.6	30.9	33.4	14.4	17.5	18.8	10.6	14.1	16.5
TsCM [39]	-	37.8	40.9	-	22.4	23.8	-	17.4	19.7
FGPL-A [40]	-	36.3	40.7	-	23.2	24.5	-	17.0	19.8
CFA [41]	-	35.7	38.2	-	17.0	18.4	-	13.2	15.5
PENet [7]	-	38.8	40.7	-	22.2	23.5	-	16.7	18.8
Cook [42]	-	35.4	37.2	-	19.2	20.3	-	14.2	16.3
Motifs [11]	10.8	14.0	15.3	6.3	7.7	8.2	4.2	5.7	6.6
+HLB [15]	11.99	15.39	16.74	7.20	8.90	9.44	5.37	7.19	8.43
+TDE [18]	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
+Reweight [7]	-	33.7	36.1	-	17.7	19.1	-	13.3	15.4
+GCL [5]	30.5	36.1	38.2	18.0	20.8	21.8	**12.9**	16.8	19.3
+FGPL [43]	24.3	33.0	37.5	17.1	21.3	22.5	11.1	15.4	18.2
+ADTrans [44]	29.0	36.2	38.8	14.8	17.0	17.8	10.6	15.5	18.1
+BiC [45]	-	37.4	40.2	-	19.0	21.0	-	**17.2**	19.9
+RDE	**32.01**	**39.47**	**42.28**	**19.09**	**23.26**	**24.47**	12.62	16.82	**20.19**
BGNN [23]	-	30.4	32.9	-	14.3	16.5	-	10.7	12.6
+HLB [15]	23.35	28.20	30.43	13.91	16.72	18.09	9.16	12.57	15.03
+RDE	**26.28**	**33.26**	**36.10**	**15.37**	**18.96**	**20.59**	**9.34**	**13.29**	**16.04**
Transformer [43]	12.4	16.0	17.5	7.7	9.6	10.2	5.3	7.3	8.8
+Reweight [43]	19.5	28.6	34.4	11.9	17.2	20.7	8.1	11.5	14.9
+FGPL [43]	27.5	36.4	40.3	19.2	22.6	24.0	**13.2**	**17.4**	**20.3**
+BiC [45]	-	34.6	37.2	-	19.7	21.0	-	16.7	19.1
+RDE	**31.86**	**39.26**	**42.25**	**19.82**	**23.72**	**24.86**	12.80	16.97	19.55

108,077 images, encompassing 1,366,673 object instances and 1,531,448 pairs of relations. These annotations are associated with 108,249 scene graphs. SGG models are commonly assessed using three distinct tasks [47]: Predicate Classification (PredCls), Scene Graph Classification (SGCls), and Scene Graph Detection (SGDet). SGDet limits that only raw images are available. The PredCls task facilitates a more accurate assessment of relationship classification by allowing manually labeled object locations and categories as input. SGCls requires the model to predict both object and relationship categories. We adopt mR@K as the evaluation metrics, which is widely used in SGG research.

Moreover, we also introduce Open Image V6 (OI-V6) dataset for more comprehensive evaluation. OI-V6 dataset has 602 object classes and 30 predicate categories. Following previous work, we use the Recall@50 (R@50), weighted mean AP of relationships ($wmAP_{rel}$), and weighted mean AP of phrase ($wmAP_{phr}$) as evaluation metrics. The weight metric $score_{wtd}$ is computed as $score_{wtd} = 0.2 \times R@50 + 0.4 \times wmAP_{rel} + 0.4 \times wmAP_{phr}$. Similar to previous studies, the OI V6 dataset is predominantly utilized for assessing the performance of the SGDet task.

Table 2 More integration results with latest (DRM) or other plug-and-play approaches (GCL) on VG dataset for the PredCls task. **Bold** indicates the best results.

	R@50	R@100	mR@50	mR@100
GCL [5]	42.7	44.4	36.1	**38.2**
+RDE	**46.53**	**48.59**	**36.98**	38.00
DRM [48]	43.9	45.8	47.1	49.6
+RDE	**44.92**	**46.96**	**47.48**	**50.21**

4.2 Comparison with State-of-the-Arts

Firstly, as shown in Table 1, we follow previous work to integrate the RDE with some commonly-used baselines for fair comparison. Motifs+RDE achieves the best overall performance among the compared plug-and-play approaches. With the enhancement of RDE, Motifs [11] and Transformer [43] improve by 184.9% and 147.9%, respectively. RDE also shows an average improvement of 17.1%, 9.6%, and 41.2% compared to the baseline approaches with unbiased training, *i.e.*, Motifs+Reweight [7], BGNN [23],

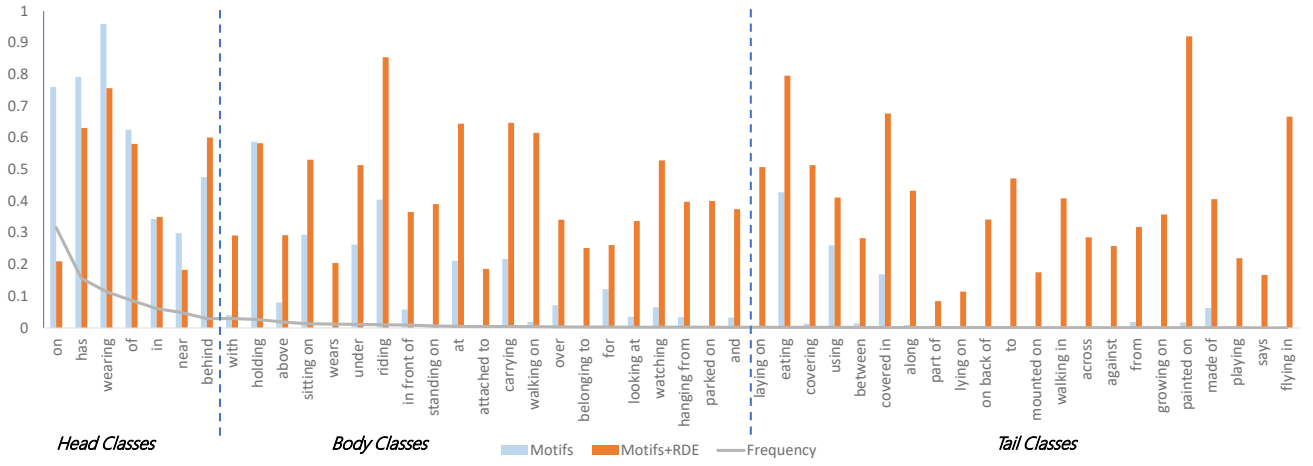


Fig. 6 Comparison results of **R@100** of each predicate class on VG-150 dataset.

Table 3 SGDet comparison on Open Image V6 dataset.

	R@50	wmAP rel	wmAP phr	$score_{wtd}$
Motifs [11]	71.6	29.9	31.6	38.9
+RDE	**73.3**	**31.5**	**32.6**	**40.3**
Unbiased [18]	69.3	30.7	32.8	39.3
BGNN [23]	75.0	33.5	34.2	42.1
RUNet [24]	76.9	35.4	34.9	43.5
PENet [7]	76.5	36.6	37.4	44.9
SGTR+ [49]	72.2	39.5	41.5	45.6
MPC [50]	76.0	39.3	40.4	47.0
DRM [48]	75.9	40.5	41.4	47.9
+RDE	**76.8**	**40.7**	**41.5**	**48.2**

and Transformer+Reweight [43]. The SGCls task introduces inaccurate object labels. Despite this, RDE outperforms other plug-and-play methods. Although RDE incorporates object information to aid in relationship representation modeling, it does not depend on the accuracy of object classification, and exhibits significant improvement compared to the baselines on the SGCls and SGDet task. In comparison to existing approaches based on prototypes or feature disentanglement, *i.e.*, PENet [7] and DeC- [4], the RDE demonstrates superior overall performance. As illustrated in Figure 6, we also present the performance of RDE across each predicate class. The average R@100 metrics for the head, body, and tail classes are 47.32, 42.91, and 40.07, respectively. This indicates that the RDE-enhanced SGG results exhibit a more balanced performance on long-tail data. Secondly, we integrate RDE with a more recent baseline, DRM [48], as well as another plug-and-play method, GCL [5]. The comparison results in Table 2 demonstrate that RDE is compatible with other plug-and-play methods and remains effective on the latest baseline. Finally, as shown in Table 3, RDE shows significant improvements in both the earliest and latest baselines on the OI V6 dataset.

Table 4 Comparison with VLMs on VG dataset for the SGDet task. **Bold** indicates the best results.

	mR@50	mR@100
InternVL2.5-4B (w/o finetuning) [51]	0.13	0.13
QWen2-VL-7B (w/o finetuning) [52]	0.89	1.06
LLM4SGG [53]	6.26	7.60
PGSG [54]	10.5	12.7
Motifs+RDE (Ours)	**16.82**	**20.19**

As shown in Table 4, we also perform comparison with some recent Vision-Language Models (VLMs) on the SGDet task using VG dataset for more comprehensive evaluations. There are two approaches for integrating VLMs into SGG models. The first involves directly training a VLM-based SGG model, while the second involves incorporating a pretrained VLM. For SGG with pretrained VLM, we evaluate InternVL2.5-4B [51] and Qwen2-VL-7B [52]. For fair comparison, the object detection phase is pre-processed, and the VLMs are merely tasked with selecting relationships from a close-set of relationships, thereby bridging the gap between open-set and closed-set conditions. For VLM-based SGG model, we choose LLM4SGG [53] and PGSG [54] for comparison, as they are also specifically designed for the SGG task. While VLMs excel in generalizability and open-domain capabilities, they still exhibit a performance gap compared to task-specific small models in conventional SGG scenarios. VLMs adopt general-purpose frameworks trained on scene graph-specific datasets rather than being tailored for SGG tasks. As high data noise remains a prominent challenge in the SGG field, these general architectures are more susceptible to interference from noisy annotations during training. In contrast, RDE is specifically designed to address the core demands of SGG by balancing semantic robustness and visual

Table 5 Component analysis under both biased and unbiased setting on PredCls task.

Components	Bias	mR@20	mR@50	mR@100
Motifs	Yes	10.8	14.0	15.3
+Proto		12.04	15.33	16.77
+Visual		**13.19**	**16.76**	**18.09**
+Full		12.46	15.97	17.53
Motifs+RW*	No	28.64	34.34	36.50
+Proto		29.88	36.46	39.05
+Visual		30.29	36.80	39.23
+Full		**32.01**	**39.47**	**42.28**

diversity in relationship representations, enabling it to better mitigate the impact of data noise and capture fine-grained visual-semantic alignments.

4.3 Ablation Study

We conduct ablation studies from three perspectives. Considering that RDE mainly concentrates on the quality of relationship representations, ablation studies are performed on the PredCls task to diminish bias from object detection. Firstly, to understand the importance of semantic and visual information, respectively, we perform experiments that separately focus on semantic prototypes and visual information. In addition to evaluating the feasibility of incorporating semantic and visual cues, we identify how visual information is preserved by GMM modeling. Finally, we investigate how the FDVAE filters and retains information during the process of decomposition and reconstruction.

Component analysis. We first analyze the components of the network, and the results are presented in Table 5. It should be noted that due to the absence of specified parameters for previous Reweight approaches [5, 7], for a fair comparison, we utilize our customized implementation, referred to as RW* in Table 5. In our experiments, all hyperparameters and codes for the reweight operation are kept consistent. In the analysis process, biased and unbiased baselines are constructed using Motifs and Motifs+RW* approaches, respectively. It is interesting to find that incorporating only visual cue enhancement yields the best results in the case of biased training. This phenomenon is attributed to the rich and detailed information contained in the visual cues, which provides more discriminative evidence for the classifier. On the contrary, although the construction of prototypes is beneficial to the learning of tail classes to a certain extent, the imbalanced learning process limits the upper bound of prototype learning, making the generated samples easier to approach the representations of head classes.

GMM modeling for visual cues. As illustrated in Table 6, we investigate various strategies for GMM modeling. Based on GMM sparsity, we devise five strategies to analyze the granularity of vision-semantic alignment. We adopt the prototype branch as the baseline, which means relationship representations adhere to N GMM distributions defined by N prototype centers. We subsequently implement the information maintenance module as an auxiliary classification branch for those hard-to-distinguish representations, where the GMM sparsity is N^2. This auxiliary branch is trained using samples grouped by a statistical confusion matrix. Additionally, we associate the representation variances to the appearance of subjects and objects to expand the auxiliary branch, extending the GMM sparsity to N^3. In the proposed RDE method, each type of relation triplet is designed to learn a distinct GMM distribution, leading to a GMM sparsity of M^2N. Furthermore, by integrating the RDE with the aforementioned auxiliary classification branch, the GMM sparsity increases to M^2N^3. Finally, we investigate the extremely sparse distribution by assigning each individual relation triplet instance a variance through self-supervised contrastive learning. However, sparser GMM distributions do not necessarily result in better performance, as shown in the last row in Table 6. This suggests that the granularity of vision-semantic alignment in the SGG task corresponds to the granularity of relation triplets, which is also intuitive.

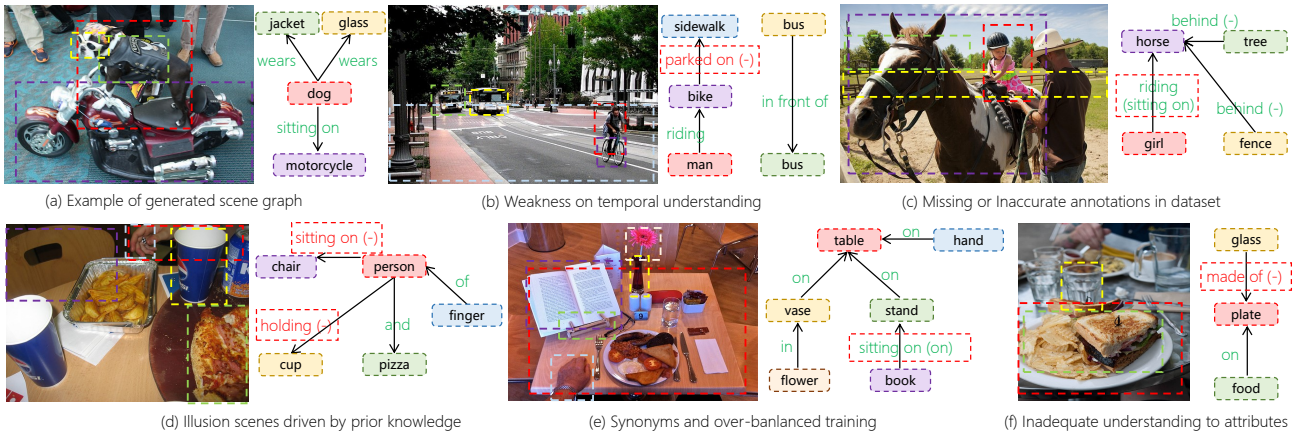
Table 6 Evaluation of distinct GMM modeling on PredCls task. underlined indicates the visual cues used by RDE.

GMM sparsity	Visual cues	mR@20	mR@50	mR@100
N	-	29.88	36.46	39.05
N^2	p	29.57	36.13	38.59
N^3	$p\|s, p\|o$	30.06	36.73	39.28
M^2N	s, p, o	**32.01**	**39.47**	**42.28**
M^2N^3	$s, p\|s, p, p\|o, o$	31.07	38.18	40.86
$+inf$	$(s, p, o)_i$	31.05	38.28	40.94

Table 7 Ablation study on Head/Body/Tail classes.

Models	All Classes	Head	Body	Tail
Motifs	15.3	60.77	12.13	4.53
Motifs+RDE	42.28	37.00	47.33	37.87

Long-tail Performance Analysis. We subsequently perform a comprehensive analysis of performance across Head, Body, and Tail data categories. The mR@100 metric is employed to assess performance on the PredCls task. The classification of Head, Body, and Tail classes adheres to the standard definitions established in prior research. The RDE method achieves a balanced performance across data partitions, with mR@100 scores of 37.00 for head classes and 37.87 for tail classes. This indicates that our approach effectively mitigates the long-tail bias, avoiding excessive reliance on high-frequency head classes while



(a) Example of generated scene graph (b) Weakness on temporal understanding (c) Missing or Inaccurate annotations in dataset

(d) Illusion scenes driven by prior knowledge (e) Synonyms and over-banlanced training (f) Inadequate understanding to attributes

Fig. 7 Qualitative results and analysis of our RDE method on VG dataset. The relationships marked with green denote the correctly predicted relations, and the red words represent the wrongly predicted ones with notated labels in brackets. Several challenges or limitations observed during the testing phase are highlighted in red dashed boxes.

Table 8 Different VAEs in Motifs+RDE framework.

Network Structure	mR@20	mR@50	mR@100
Standard VAE	31.13	38.76	41.72
Transformer VAE	29.34	36.07	38.98
FDVAE (ours)	**32.01**	**39.47**	**42.28**

significantly enhancing the performance of tail classes. The slight performance reduction in head classes is primarily due to synonym confusion. For example, the model may predict "on" as semantically similar relationships such as "sitting on", which reflects a more fine-grained understanding of relationships rather than an actual performance decline.

Network Structure of the FDVAE. Finally, the network design of the FDVAE is validated, as shown in Table 8. Our primary focus lies on the network design of the encoder section, since the decoder section comprises only a singular fully connected layer. We first conduct a standard VAE network for comparison, which is a multi-layer perceptron that includes multiple fully connected layers and ReLU activation units. Subsequently, we incorporate a transformer encoder to maximize the non-linear capability of the VAE network. Both the standard VAE and transformer VAE used in our experiments are built following the general framework of VAE, designed to ensure fair comparison with our proposed FDVAE [55]. All three models share the same overall architecture with 6 encoder layers, differing only in the implementation of the encoder block. While FDVAE adopts the iterative information enhancement block, the two baselines use alternative block designs with consistent activation layers. For standard VAE, the cross-attention layer is replaced with two fully connected layers. For transformer VAE, the cross-attention layer is substituted with a standard cross-attention mechanism configured with 8 attention heads,

Table 9 Training cost of our method. All experiments are conducted in a single RTX 4090 GPU.

Baselines	Feature Dim.	Params. (M) Orig.	Params. (M) +RDE	Speed (FPS) Orig.	Speed (FPS) +RDE
Motifs	4096	453.6	567.0 +25.0%	18.9	16.4 +13.2%
Transformer	1024	394.3	472.9 +19.9%	16.3	15.4 +5.5%

retaining the same feature propagation workflow as FDVAE. The experimental results indicate that our proposed FDVAE network, which is based on iterative information enhancement, significantly outperforms the above networks.

4.4 Analysis of Training Overhead

Given that RDE is a plug-and-play training approach that incurs no additional inference cost, only the training cost and the size of trainable parameters are reported. The training cost is detailed in Table 9. As the objective of RDE is to refine the original relationship feature f_k into an enhanced relationship representation, the number of network parameters in the RDE branch varies with the dimension of f_k. In particular, since few methods surpass a feature size of 4096 in Motifs [11], the RDE can be considered to introduce fewer than 113.4M additional parameters. For a more typical feature size of 1024 in Transformer [43], the RDE introduces only 78.6M parameters. Furthermore, the reduction in training speed is negligible.

4.5 Visualization and Qualitative Analysis

We provide some real cases produced by Motifs+RDE, as shown in Figure 7. Although RDE can effectively improve the performance of the baseline model, we notice that there are

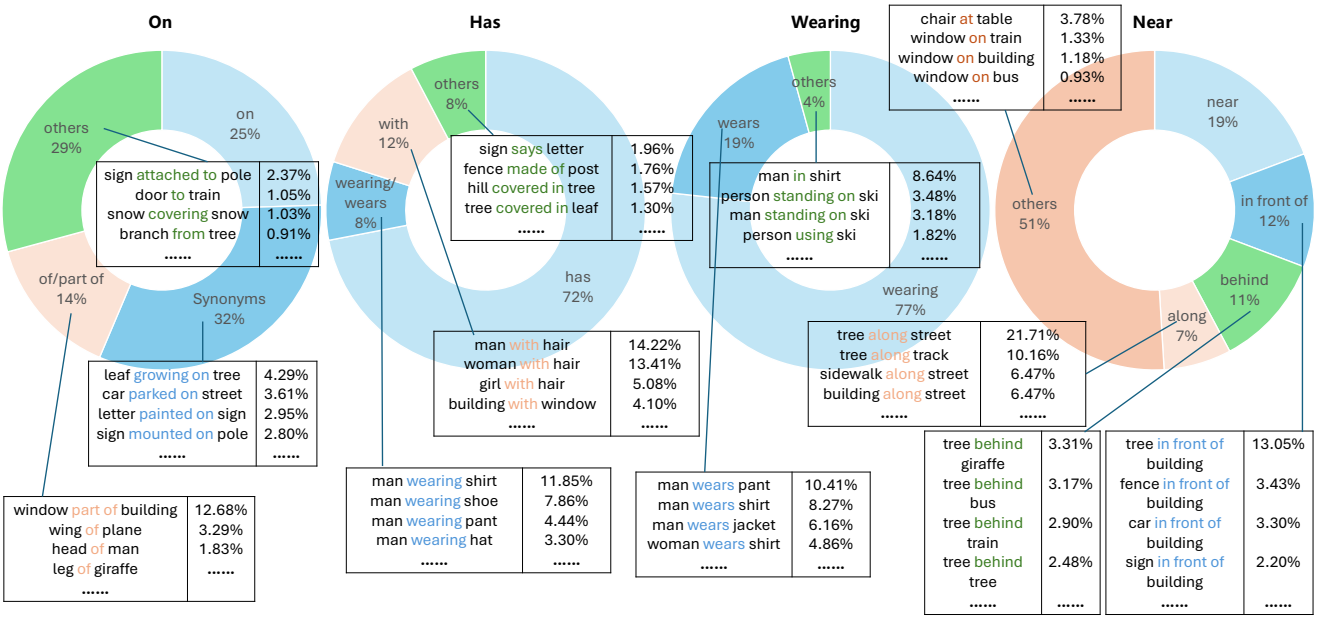


Fig. 8 Detailed statistics of mismatched predictions on four head classes. The majority of the discrepancies arise from the confusion of interchangeable (near-)synonyms.

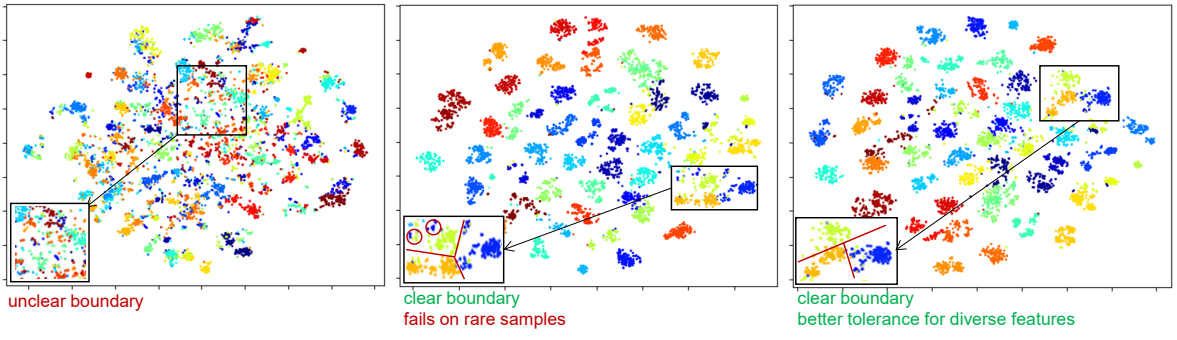


Fig. 9 T-SNE visualization of latent representations.

still some potential limitations to be discussed. Specifically, Figure 7(a) represents a successful predicted result, where RDE proves to accurately predict complex relationships between objects, even for unusual classes of relation triples. Figure 7(b) shows a moving bicycle is incorrectly predicted as "parked on", which indicates that the current SGG model for still images struggles to predict actions involving temporal information. Figure 7(c) highlights inherent problems in the current SGG dataset, where a mass of inconspicuous relationships is not labeled, and some of the labeled ones are inaccurate, *e.g.*, "riding" labeled "sitting on" and "sitting on" labeled as "on". Figure 7(d) demonstrates the limited ability to capture spatial details. Many predicted relationships are "illusions" at the semantic level, which means that they seem reasonable but do not correspond to the actual visual scenes. Figure 7(e) presents the difficulty in handling synonyms and the challenges posed by the long-tail phenomenon. Figure 7(f) reveals the limitations in representing relationships involving object parts or attributes, such as "made of".

Meanwhile, as shown in Figure 8, we perform a detailed statistical analysis focusing on four head categories. It is observed that the majority of errors arise from ambiguity in semantic expressions. For instance, 32% of "on" predictions are identified as synonyms, including terms such as "lying on" and "sitting on", and 14% of "on" are classified as subordinate relationships, such as "of" and "part of". A large number of bad cases are attributed to the ambiguity of human annotation, and the current SGG benchmarks fail to evaluate when models

attempt to provide more detailed expressions.

Figure 9 illustrates the T-SNE visualization of latent representations. As depicted in Figure 9(b), the prototypes leads to better classification boundary. In contrast, the incorporation of visual information in Figure 9(c) results in a sparser intra-class distribution to enable better tolerance for diverse features. This enhances the model's ability to generate relationship representations within an expanded latent space, while maintaining the distinguishability of these representations.

4.6 Limitations and Future Work

Limitations. While the proposed method enhances the diversity of relationship representations for scene graph generation, it still has significant limitations that warrant attention. Firstly, it underutilizes external knowledge, relying solely on single-dataset learning and lacking effective integration of prior semantic information, which could further enrich the model's understanding of complex relationships. Secondly, although current VLMs still have room for improvement in specific scene graph tasks, they have shown considerable potential, particularly in zero-shot reasoning tasks, and the method has yet to explore how to harness this potential. Thirdly, due to the difficulty and complexity of relationship annotation, high dataset noise, especially semantic ambiguity that leads to confusion between similar relationships, is almost inevitable, and the method lacks targeted strategies to utilize such noisy data for learning fine-grained relationship representations.

Future Work. To overcome the aforementioned limitations, our future work will concentrate on three key areas, with plans for further exploration and validation. First, we aim to enhance the integration of external prior knowledge, which could help us move beyond the constraints of single datasets and enrich the model's semantic understanding of relationships within the open-vocabulary domain. Additionally, we plan to investigate potential methods for combining the proposed approach with VLMs, leveraging their strengths in complex reasoning to improve the accuracy of relationship inference. Lastly, we will develop dedicated strategies to effectively utilize noisy dataset resources, aiming to reduce the impact of semantic ambiguity while enabling the model to learn more precise, fine-grained relationship representations.

5 Conclusion

In this paper, we proposed a novel RDE method by exploring the diversity of relationship representations in SGG. The RDE method can be adopted in various SGG approaches without any additional inference cost or modification of the model architecture. Inside the RDE method, an FDVAE network is designed to decompose relationship representations into semantic and visual parameters. The semantic features are constrained by prototype learning, while the visual features are supervised by relation triplets. We applied RDE in several baselines and conducted comprehensive experiments to demonstrate the effectiveness of RDE in the VG-150 and OI V6 dataset. The results show that the learned relationship representations can facilitate finer-grained relationship representations for SGG.

Availability of data and materials

The data that support the findings of this study can be downloaded from publicly available storage: VG-150 and Open Image V6.

Author contributions

Yunqing He is responsible for investigation, methodology, software and original draft writing. Ruichao Hou is responsible for project administration and validation. Jia Bei is responsible for supervision. Tongwei Ren is responsible for funding acquisition, resources, supervision. All the authors have approved the final manuscript.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62072232), the Fundamental Research Funds for the Central Universities (021714380026) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Liu J, Wei K, Liu C. Multimodal Event Causality Reasoning with Scene Graph Enhanced Interaction Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 8778–8786, doi:10.1609/AAAI.V38I8.28724.

[2] Wu Z, Li H, Chen G, Yu Z, Gu X, Wang Y. 3D Question Answering with Scene Graph Reasoning. In *Proceedings of the ACM International Conference on Multimedia*, 2024, 1370–1378, doi:10.1145/3664647.3681517.

[3] Zhang H, Zhang P, Hu X, Chen Y, Li LH, Dai X, Wang L, Yuan L, Hwang J, Gao J. GLIPv2: Unifying Localization and Vision-Language Understanding. In *Advances in Neural Information Processing Systems*, 2022, 36067–36080.

[4] Tao H, Lianli G, Jingkuan S, Yuan-Fang L. State-Aware Compositional Learning Toward Unbiased Training for Scene Graph Generation. *IEEE Transactions on Image Processing*, 2023, 32: 43–56.

[5] Xingning D, Tian G, Xuemeng S, Jianlong W, Yuan C, Liqiang N. 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*, 2022, 19405–19414, doi: 10.1109/CVPR52688.2022.01882.

[6] Yao T, Limin W. Structured Sparse R-CNN for Direct Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 19415–19424, doi:10.1109/CVPR52688.2022.01883.

[7] Chaofan Z, Xinyu L, Lianli G, Bo D, Jingkuan S. Prototype-based Embedding Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 22783–22792, doi:10.1109/CVPR52729.2023.02182.

[8] Chen L, Song Y, Cai Y, Lu J, Li Y, Xie Y, Wang C, He G. Multi-Prototype Space Learning for Commonsense-Based Scene Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 1129–1137, doi:10.1609/AAAI.V38I2.27874.

[9] Naftali T, Noga Z. Deep learning and the information bottleneck principle. In *Proceedings of the IEEE Information Theory Workshop*, 2015, 1–5, doi:10.1109/ITW.2015.7133169.

[10] Xu D, Zhu Y, Choy CB, Fei-Fei L. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 3097–3106, doi:10.1109/CVPR.2017.330.

[11] Zellers R, Yatskar M, Thomson S, Choi Y. Neural Motifs: Scene Graph Parsing with Global Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 5831–5840, doi:10.1109/CVPR.2018.00611.

[12] Chen T, Yu W, Chen R, Lin L. Knowledge-Embedded Routing Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 6163–6171, doi:10.1109/CVPR.2019.00632.

[13] Zhang H, Kyaw Z, Chang S, Chua T. Visual Translation Embedding Network for Visual Relation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 3107–3115, doi: 10.1109/CVPR.2017.331.

[14] Wang W, Liu R, Wang M, Wang S, Chang X, Chen Y. Memory-Based Network for Scene Graph with Unbalanced Relations. In *Proceedings of the ACM International Conference on Multimedia*, 2020, 2400–2408, doi:10.1145/3394171.3413507.

[15] Yunqing H, Tongwei R, Jinhui T, Gangshan W. Heterogeneous Learning for Scene Graph Generation. In *Proceedings of the ACM International Conference on Multimedia*, 2022, 4704–4713, doi:10.1145/3503161.3548356.

[16] Jiankang D, Jia G, Jing Y, Alexandros L, Stefanos Z. Variational Prototype Learning for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 11906–11915, doi:10.1109/CVPR46437.2021.01173.

[17] Yuxuan C, Yizhuang Z, Qi H, Jianjian S, Xiangwen K, Jun Li XZ. Reversible Column Networks. In *Proceedings of the International Conference on Learning Representations*, 2023, 7650–7673.

[18] Tang K, Niu Y, Huang J, Shi J, Zhang H. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 3713–3722, doi:10.1109/CVPR42600.2020.00377.

[19] Wei L, Haiwei Z, Qijie B, Guoqing Z, Ning J, Xiaojie Y. PPDL: Predicate Probability Distribution based Loss for Unbiased Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 19425–19434, doi:10.1109/CVPR52688.2022.01884.

[20] Yan S, Shen C, Jin Z, Huang J, Jiang R, Chen Y, Hua X. PCPL: Predicate-Correlation Perception Learning for Unbiased Scene Graph Generation. In *Proceedings of the ACM International Conference on Multimedia*, 2020, 265–273, doi:10.1145/3394171.3413722.

[21] Leitian T, Li M, Nannan L, Xianhang C, Yaosi H, Zhenzhong C. Predicate Correlation Learning for Scene Graph Generation. *IEEE Transactions on Image Processing*, 2022, 31: 4173–4185.

[22] Chen Z, Luo Y, Shao J, Yang Y, Wang C, Chen L, Xiao J. Dark Knowledge Balance Learning for Unbiased Scene Graph Generation. In *Proceedings of the ACM International Conference on Multimedia*, 2023, 4838–4847, doi:10.1145/3581783.3612031.

[23] Li R, Zhang S, Wan B, He X. 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*, 2021, 11109–11119, doi:10.1109/CVPR46437.2021.01096.

[24] Xin L, Changxing D, Jing Z, Yibing Z, Dacheng T. RU-Net: Regularized Unrolling Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 19435–19444, doi: 10.1109/CVPR52688.2022.01885.

[25] Xin L, Changxing D, Yibing Z, Zijian L, Dacheng T. HL-Net: Heterophily Learning Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 19454–19463, doi: 10.1109/CVPR52688.2022.01887.

[26] Lin L, Long C, Yifeng H, Zhimeng Z, Songyang Z, Jun X. 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, 2022, 18847–18856, doi:10.1109/CVPR52688.2022.01830.

[27] Arushi G, Basura F, Frank K, Hakan B. Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 15575–15585, doi: 10.1109/CVPR52688.2022.01515.

[28] Zhang R, Shang Z, Wang F, Yang Z, Cao S, Cen Y, An G. Synergetic Prototype Learning Network for Unbiased Scene Graph Generation. In *Proceedings of the ACM International Conference on Multimedia*, 2024, 945–954, doi:10.1145/3664647.3680973.

[29] Zhang SK, Xie WY, Wang C, Zhang SH. ScenePalette: Contextually Exploring Object Collections Through Multiplex Relations in 3D Scenes. *Journal of Computer Science and Technology*, 2024, 39(5): 1180–1192.

[30] Zhang SH, Zhang SK, Xie WY, Luo CY, Yang YL, Fu H. Fast 3D indoor scene synthesis by learning spatial relation priors of objects. *IEEE Transactions on Visualization and Computer Graphics*, 2021, 28(9): 3082–3092.

[31] Zhang D, Yu J, Zhang C, Cai W. PaRot: Patch-Wise Rotation-Invariant Network via Feature Disentanglement and Pose Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 3418–3426, doi:10.1609/AAAI.V37I3.25450.

[32] Li X, Xu Z, Wei K, Deng C. Generalized Zero-Shot Learning via Disentangled Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 1966–1974, doi: 10.1609/AAAI.V35I3.16292.

[33] Zhu Z, Wang L, Tang W, Liu Z, Zheng N, Hua G. Learning Disentangled Classification and Localization Representations for Temporal Action Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 3644–3652, doi: 10.1609/AAAI.V36I3.20277.

[34] Jeffrey P, Richard S, Christopher D M. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, 1532–1543, doi:10.3115/V1/D14-1162.

[35] Zhang J, Shih KJ, Elgammal A, Tao A, Catanzaro B. Graphical Contrastive Losses for Scene Graph Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 11535–11543, doi:10.1109/CVPR.2019.01180.

[36] Zareian A, Karaman S, Chang SF. Bridging Knowledge Graphs to Generate Scene Graphs. In *Proceedings of the European Conference on Computer Vision*, 2020, 606–623, doi:10.1007/978-3-030-58592-1_36.

[37] Suhail M, Mittal A, Siddiquie B, Broaddus C, Eledath J, Medioni G, Sigal L. Energy-Based Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 13936–13945, doi:10.1109/CVPR46437.2021.01372.

[38] Deunsol J, Sanghyun K, Won K Hwa, Minsu C. 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*, 2023, 18664–18674, doi: 10.1109/CVPR52729.2023.01790.

[39] Sun S, Zhi S, Liao Q, Heikkilä J, Liu L. Unbiased Scene Graph Generation via Two-Stage Causal Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(10): 12562–12580.

[40] Lyu X, Gao L, Zeng P, Shen HT, Song J. Adaptive Fine-Grained Predicates Learning for Scene Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(11): 13921–13940.

[41] Li L, Chen G, Xiao J, Yang Y, Wang C, Chen L. Compositional Feature Augmentation for Unbiased Scene Graph Generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023, 21628–21638, doi:10.1109/ICCV51070.2023.01982.

[42] Kim H, Kim S, Ahn D, Lee JT, Ko BC. Scene Graph Generation Strategy with Co-occurrence Knowledge and Learnable Term Frequency. In *Proceedings of the International Conference on Machine Learning*, 2024, 24094–24109.

[43] Xinyu L, Lianli G, Yuyu G, Zhou Z, Hao H, Heng Tao S, Jingkuan S. Fine-Grained Predicates Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 19445–19453, doi:10.1109/CVPR52688.2022.01886.

[44] Li L, Ji W, Wu Y, Li M, Qin Y, Wei L, Zimmermann R. Panoptic Scene Graph Generation with Semantics-Prototype Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 3145–3153, doi:10.1609/AAAI.V38I4.28098.

[45] Yang J, Wang C, Yang L, Jiang Y, Cao A. Adaptive Feature Learning for Unbiased Scene Graph Generation. *IEEE Transactions on Image Processing*, 2024, 33: 2252–2265.

[46] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma D, Bernstein M, Li FF. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 2017, 123(1): 32–73.

[47] Johnson J, Krishna R, Stark M, Li L, Shamma DA, Bernstein MS, Fei-Fei L. Image Retrieval Using Scene Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, 3668–3678, doi: 10.1109/CVPR.2015.7298990.

[48] Li J, Wang Y, Guo X, Yang R, Li W. Leveraging Predicate and Triplet Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 28369–28379, doi:10.1109/CVPR52733.2024.02680.

[49] Li R, Zhang S, He X. SGTR+: End-to-End Scene Graph Generation With Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(4): 2191–2205.

[50] Yang J, Wang C, Zhang J, Wu S, Zhao J, Liu Z, Yang L. Semi-

Supervised Clustering Framework for Fine-grained Scene Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 9220–9228, doi:10.1609/AAAI.V39I9.32998.

[51] Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, et al.. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 24185–24198.

[52] Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, et al.. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, 2024.

[53] Kim K, Yoon K, Jeon J, In Y, Moon J, Kim D, Park C. Llm4sgg: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 28306–28316.

[54] Li R, Zhang S, Lin D, Chen K, He X. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 28076–28086, doi:10.1109/CVPR52733.2024.02652.

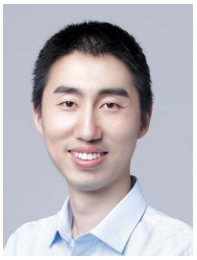
[55] Diederik PK, Max W. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations*, 2014, 158–171.

Author biography

Yunqing He received the B.S. and the M.S. degree from Nanjing University, Nanjing, China, in 2021, where he is currently pursuing the Ph.D. degree. He has authored several papers in ACM MM, ICMR, and ICME. His research interest mainly includes visual relation understanding, representation learning and image generation.

Ruichao Hou (IEEE Member) received his Ph.D degree from the Department of Computer Science and Technology, Nanjing University in 2023. He is currently an Assistant Researcher at the Software Institute of Nanjing University. His research mainly focuses on multi-modal object detection and tracking.

Jia Bei received his B.S and Ph.D. degrees from Nanjing University, Nanjing, China, in 2001 and 2006, respectively. He joined Nanjing University in 2006, and at present he is an associate professor. His research interest mainly includes multimodal information fusion and understanding. He has published more than 30 papers in top-tier journals and conferences.



Tongwei Ren (IEEE Member) received the bachelor's, master's, and doctoral degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He joined the Software Institute of Nanjing University as an Assistant Professor in 2010, and at present he is an Associate Professor with Nanjing University. He visited the Hong Kong Polytechnic University in 2008 and the National University of Singapore from 2016 to 2017. He has authored more than 50 papers in international journals/conferences, such as TIP, TOMM, TNNLS, MM, ICCV, and AAAI, and won the best paper honorable mention of ICIMCS 2014, the best paper runner-up of PCM 2015, the champions of ECCV 2018 PIC challenge and MM 2019 VRU challenge, and the second places of ICME 2019 SVU challenge and MM 2019 CBVRP challenge. His research interest mainly includes visual multimedia computing and its application.