CDC: Enhancing Scene Graph Generation for IoST-Driven Social Behavioral Modeling With Cooperative Dual Classifier

Zhaodi Wang, Yangyan Zeng, Biao Leng, and Xiaokang Zhou, *Member, IEEE*

***Abstract*—Scene graph generation (SGG) plays an important role in the intelligence of social things (IoST) framework by extracting structured semantic representations from social device data, thereby supporting advanced scene understanding and behavioral-cultural modeling. However, the intrinsic long-tail nature of real-world social device data, coupled with the semantic entanglement between head and tail categories (e.g., "on" versus "standing on"), presents significant challenges for fine-grained SGG. This often results in biased models and suboptimal generalization to rare but semantically informative relations. To address these issues, we propose a novel cooperative dual classifier (CDC) framework for fine-grained SGG in IoST-driven social systems. CDC introduces a cooperative learning mechanism that combines two classifiers. The frozen prototype classifier is designed with maximum interclass margins to alleviate class imbalance. In parallel, a learnable classifier dynamically adjusts decision boundaries to improve discriminative precision. To further enhance the integration between the two classifiers, we introduce a weight knowledge transfer (WKT) module and a collaborative constraint term, facilitating robust adaptation to tail categories. Extensive experiments on the Visual Genome and GQA datasets demonstrate that CDC outperforms state-of-the-art SGG methods, particularly in modeling fine-grained relations under long-tail distributions. These results highlight the capability of CDC to advance semantic understanding of complex behavioral and cultural patterns within computational social systems.**

***Index Terms*—Behavioral-cultural modeling, cooperative learning, intelligence of social things (IoST), long-tail distribution, prototype learning, scene graph generation (SGG).**

Received 31 March 2025; revised 4 June 2025 and 2 July 2025; accepted 10 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 72274058. *(Corresponding author: Yangyan Zeng.)*

Zhaodi Wang and Biao Leng are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: wzdaisy@buaa.edu.cn; lengbiao@buaa.edu.cn).

Yangyan Zeng is with Xiangjiang Laboratory, Changsha 410205, China (e-mail: yangyanz0930@163.com).

Xiaokang Zhou is with the Faculty of Business Data Science, Kansai University, Osaka 565-8585, Japan, and also with RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@kansai-u.ac.jp).

Digital Object Identifier 10.1109/TCSS.2025.3600391

I. Introduction

As intelligent devices become increasingly interconnected in modern environments, the intelligence of social things (IoST) provides a promising framework for understanding the complex interactions among humans, machines, and environments [1]. In computational social systems, understanding these interactions requires not only perception of physical scenes, but also semantic abstraction of underlying behavioral and cultural patterns, enabling interpretable results for human-centric analysis and structured inputs for downstream tasks. In particular, IoST-enabled cooperative learning leverages the collective intelligence of distributed devices to model social behaviors and cultural dynamics in a scalable manner. Scene graph generation (SGG), as a fundamental task in visual scene understanding, serves this need by extracting structured semantic representations (triplets $\langle \text{subject}, \text{predicate}, \text{object} \rangle$) from social device data. By representing visual scenes as structured graphs, SGG not only provides interpretable abstractions but also facilitates scalable integration of heterogeneous data streams from multiple social devices, enabling joint analysis of behavioral and cultural patterns across diverse environments. These semantic graphs bridge low-level visual signals and high-level reasoning, enabling deeper behavioral-cultural modeling across a variety of IoST-enabled applications, including visual question answering [2], image retrieval [3], and embodied navigation [4].

However, real-world data collected from social devices often exhibit long-tail distributions, which further intensify the challenges of behavioral-cultural modeling. Unlike other long-tail classification tasks [5], SGG is uniquely characterized by semantic entanglement among predicate categories where head predicates (e.g., "on") subsume or overlap with tail predicates (e.g., "standing on", "lying on"). Long-tail distribution combined with semantic entanglement undermines the capability to capture informative behavioral patterns, as models overfit to frequent head predicates while neglecting the rich contextual information embedded in tail predicates. To address the above problems, an increasing number of studies [6], [7], [8], [9], [10] explored various strategies to enhance the fine-grained capability of SGG models. Early efforts primarily focused on model reweighting [7] or data resampling [8], [9], [10] techniques, aiming to balance the training process. More recent studies adopt contrastive learning strategies or prototypical

networks [6], [11], [12] to model interclass relations. Among them, PE-Net [6] leverages contrastive learning constraint terms to improve the separability between classifier parameters and samples, thereby enhancing fine-grained relation recognition. However, the abundance of head samples in the repulsion term of contrastive objectives tends to overwhelm the optimization process. This often pushes the representations and classifier parameters for tail categories into unstable or random states, significantly undermining classification accuracy. Inspired by this insight and the neural collapse theory [13], we propose to fix the classifier and decouple it from the SGG optimization process.

Moreover, the feature distributions of head and tail categories exhibit significant density variations. In particular, the features of tail class tend to be sparser and exhibit higher variance due to limited training samples, making it difficult to fully capture their complex behavioral semantics. In contrast, head classes benefit from abundant data, resulting in more compact and stable representations. To address these imbalances, prototype-based approaches have attracted increasing attention, as they offer a compact representation of each class and naturally encourage larger interclass margins in the embedding space, which is essential for discriminating semantically similar predicates. In these methods, decision boundaries are usually placed equidistantly between the learned class prototypes. Such a fixed structure can be problematic under imbalanced distributions: the sparse and scattered nature of tail features makes them more likely to cross into neighboring decision regions, increasing the risk of misclassification. This motivates the need for adaptive boundary adjustment mechanisms that account for class-specific distribution characteristics and help maintain the transparency and reliability required in IoST-enabled social learning environments.

To tackle the aforementioned challenges, we propose a novel cooperative dual classifier (CDC) method for fine-grained SGG. The CDC framework introduces a cooperative learning mechanism that synergistically integrates a prototype classifier and a learnable classifier. Before SGG training, we construct an optimal prototype classifier with maximum interclass margins to alleviate the long-tail problem. During the SGG training process, this classifier guides feature refinement and produces coarse relation predictions. In parallel, we design a learnable classifier that dynamically adjust the decision boundaries of the fixed prototype classifier, enabling precise relation classification. To facilitate effective interaction between the two classifiers, we further propose a weight knowledge transfer (WKT) module and a collaborative constraint term, ensuring that the learnable classifier inherits relational structure knowledge while maintaining flexibility to adapt to class-specific feature distributions. Overall, our dual-classifier design not only mitigates the bias toward head categories but also enhances generalization and adaptability to diverse real-world social device data, enabling more accurate modeling of complex behavioral patterns in IoST applications.

Specifically, the main contributions addressed in this article can be summarized as follows.

1) We propose a novel CDC framework for fine-grained SGG, integrating a fixed prototype classifier to guide feature refinement with a learnable classifier to adjust the decision boundaries. This cooperative design aligns with IoST-enabled cooperative learning, facilitating more accurate semantic understanding of fine-grained behavioral patterns derived from social device data in computational social systems.
2) We introduce a WKT module and a collaborative constraint term to enhance the cooperation between the two classifiers. These components effectively support the optimization of decision boundaries and improve the model's adaptability under long-tail data distributions.
3) We conduct comprehensive experiments on the Visual Genome (VG) and GQA datasets, demonstrating that CDC outperforms state-of-the-art SGG methods, particularly in capturing fine-grained relations under long-tail distributions. These results highlight CDC's capability in advancing the comprehension of nuanced behavioral patterns in social device data, contributing to the development of intelligent social systems.

Beyond the immediate improvements in SGG, the structured and semantically enriched outputs of our CDC framework hold strong potential for a wide range of downstream tasks in computational social systems. In particular, the integration of CDC within IoST-enabled cooperative learning infrastructures can further support behavior-aware human–computer interaction, real-time decision-making, and fine-grained analysis of social and cultural dynamics. By improving the reliability and granularity of scene understanding, CDC lays the foundation for more context-aware and interpretable intelligent services, offering promising directions for future research and application in socially intelligent systems.

The rest of this article is organized as follows. Section II presents an overview of related works. Section III elaborates on the modeling of the CDC framework. In Section IV, we demonstrate experiment and evaluation results, providing both qualitative insights and quantitative metrics. Finally, Section V makes a conclusion of this article.

II. Related Work

In this section, various issues associated with the proposed method are discussed, with a focus on the evolution from vanilla SGG toward fine-grained approaches.

A. Vanilla SGG

In recent years, the SGG task has emerged as a central research direction in computer vision. It offers a structured representation of semantic perceptions for visual social device data, capturing entities and their relationships in the triplet format $\langle \text{subject}, \text{predicate}, \text{object} \rangle$. This structured representation is particularly valuable for IoST applications, as it enables the modeling of complex social interactions and behavioral patterns. Early methods [14], grounded in the assumption of relation independence, primarily focused on designing feature extraction modules from multimodal features, such as visual, spatial, and semantic features. However, these methods often

overlooked the critical role of contextual information in enhancing overall scene understanding and improving the accuracy of SGG. Consequently, subsequent studies shift their focus to exploring contextual features using message passing [8], [9], [15], LSTM [16], or tree structures [17]. These contextual methods leverage structured message propagation frameworks to exchange information across entity and relation nodes, thereby capturing cooccurrence patterns between objects and relations in the scene. By modeling surrounding context, they aim to disambiguate visually similar relationships and improve overall relation reasoning. However, such strategies often suffer from the introduction of redundant or irrelevant contextual cues, leading to noisy message aggregation and degraded relation classification performance.

In addition to contextual modeling, existing methods [6], [8], [15], [18] commonly exploit multiple modalities, such as visual appearance, spatial features, and semantic embeddings, to represent relation instances. These modalities provide complementary cues: visual features capture object textures and poses, spatial features describe geometric interactions, and semantic embeddings offer prior knowledge derived from language. However, inappropriate fusion strategies, such as concatenation [8], [15], often fail to capture the intrinsic interaction patterns between subjects and objects. This leads to scattered intraclass distributions and interclass overlapping in the relation feature space, thereby hindering discriminative relation learning. The problem is further exacerbated by the diverse subject-object compositions under the same predicate. Thus, recent works [6] attempt to realize feature refinement. For example, PE-Net [6] introduces a prototype-based embedding network to explore an intrinsic and compact feature space. However, the learning of feature representation modules remains influenced by the long-tail distribution of SGG data, resulting in overfitting to abundant head classes while under-representing rare tail classes. As a result, the learned relation features are biased toward frequent relation patterns, failing to capture the subtle and diverse semantics of infrequent relations, which ultimately results in suboptimal generalization of SGG models.

B. Fine-Grained SGG

To address the challenges posed by long-tail data distributions, a growing number of studies [6], [9], [10], [19] have explored debiasing techniques for SGG, making fine-grained SGG an emerging research hotspot. Fine-grained SGG aims to accurately identify semantically subtle and less frequent predicates, which is particularly vital for real-world visual understanding and downstream reasoning tasks in IoST scenarios. To this end, VCTree [17] introduced the mean recall metric, which averages recall values across all predicate categories, providing a fairer evaluation of fine-grained SGG by emphasizing tail predicate performance. Early fine-grained SGG methods primarily adopted debiasing techniques, including data resampling [8], [9], [10], data augmentation [20], [21], and loss reweighting [7]. For instance, Tang et al. [10] proposed TDE, which employs causal inference to reweight training samples and increase the focus on tail categories. While these methods mitigate sample imbalance to some extent, they often struggle with overfitting to synthetic or reweighted data, leading to limited generalization in diverse real-world scenarios. Moreover, some recent methods [6] utilize contrastive learning strategies to enhance the separation between head and tail categories by pulling features of the same class closer together while pushing apart those of different classes. The long-tailed distribution challenge is not unique to SGG but is widely observed in image classification tasks [22], [23], where various strategies, including class-balanced loss functions [23] and decoupled training [22], have been proposed to mitigate class imbalance. Inspired by these approaches, recent SGG models [6], [24] have attempted to adapt these principles to the relational domain. The aforementioned methods generate scene graphs using a single classifier. In contrast, a series of recent works [25] have shifted toward multiple experts, where each expert is responsible for a subset of predicates, and their results are integrated into the final scene graph [26].

Despite their effectiveness, existing debiasing strategies largely rely on manipulating training distributions (e.g., resampling), which can introduce synthetic biases or unstable gradients. Even though contrastive learning methods enhance feature separation, they rely on a single global classifier, which lacks adaptability to the fine-grained variations across head and tail predicates. Moreover, expert-based methods [25], [26] segment the classification space but fail to dynamically coordinate decision boundaries across the entire feature space. Unlike the above methods, we address the long-tail problem in SGG through the cooperative learning of two classifiers, where the first guides feature refinement and the second dynamically calibrates the decision boundary based on the unbiased feature space.

III. CDC

In this section, we propose a novel CDC framework for fine-grained SGG within the context of the IoST. The CDC framework is composed of three core modules: the Proposal Network for entity detection, the Feature Refinement module for multimodal feature extraction and refinement, and the collaborative predicate classifier (CPC) module. As our primary contribution, the CPC module infers predicates through a cooperative classification strategy that leverages both prototype-based and learnable classifiers. These modules work synergistically to generate structured scene graphs with rich and informative relations, facilitating accurate modeling of social behaviors in IoST-enabled applications.

A. Preliminaries and Problem Definition

Given a visual input $\mathcal{I}$ from a social device, the objective of SGG is to construct a directed scene graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. This graph serves as a critical tool for behavior and relationship modeling within the IoST framework (as illustrated in Fig. 1). Here, each node $N_i \in \mathcal{N}$ represents an object (e.g., person), characterized by its bounding box and object category. Each edge $E_i \in \mathcal{E}$ denotes a predicate category C_i^P that defines the relationship (e.g., eating) between a pair of objects. The generated scene graph $\mathcal{G}$ captures visual relation triplets in the

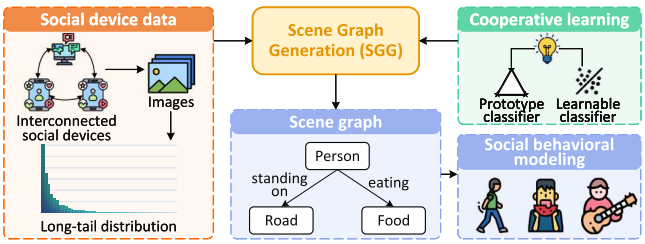


Fig. 1. SGG for social behavioral modeling within the IoST framework. SGG processes image data collected from multiple social devices with long-tail distributions to generate structured scene graphs. To enhance SGG, cooperative learning incorporates both prototype-based and learnable classifiers, refining feature representations and improving the expression of tail predicates. The generated graphs are further leveraged for social behavioral modeling, enabling intelligent perception in IoST-driven environments.

form of ⟨subject, predicate, object⟩, providing a comprehensive semantic representation of social device data that often exhibit long-tail distributions. This capability is essential for IoST-driven applications, such as autonomous vehicles and smart cities.

A major challenge in SGG arises from the long-tail distribution inherent in predicate categories, where a few frequent predicates dominate the dataset, while many rare and fine-grained predicates remain underrepresented. This imbalance hinders accurate prediction of tail predicates, often leading to biased models that prioritize head categories and overlook nuanced relationships essential for comprehensive scene understanding. Addressing this issue is especially crucial in IoST-driven applications, where rare but meaningful predicates, such as specific social interactions or contextual behaviors, are key to robust perception and behavioral modeling.

To achieve fine-grained SGG in the IoST framework, we propose a SGG method CDC, which leverages cooperative learning to enhance predicate classification. By integrating prototype-based and learnable classifiers, CDC refines feature representations and improves the expression of tail predicates, thereby enabling intelligent perception in IoST-driven scenarios. As illustrated in Fig. 2, the CDC framework is primarily composed of three components: the proposal network, the feature refinement module, and the CPC module. Following previous works [6], [16], [25], we employ Faster R-CNN [27] as the proposal network for entity detection on social device data. It generates entity proposals along with visual features, categories, and bounding boxes. In the feature refinement module, feature refinement of entities and relations is conducted based on multimodal features, as detailed in Section III-B. Additionally, we use GloVe [28] to embed categories into semantic space, incorporating contextual word relationships and cooccurrence statistics derived from large-scale text corpora. We also encode bounding boxes as spatial features. We derive the union feature $\mathbf{u}$ for each entity pair by encoding the spatial and visual features of their union region, which represents the region of interest for the relation. We design the CPC module to further guide feature refinement and adjust decision boundaries through cooperative learning between the prototype classifier and the learnable classifier, as elaborated in Section III-C.

B. Feature Refinement

Feature refinement is a crucial step for improving entity and relation representations in SGG, enabling precise modeling of social behaviors. It involves two key components: the entity encoder, which refines individual object features, and the relation encoder, which extracts discriminative relation-centric features. In this section, we describe these encoders in detail. We explain their roles in capturing intrinsic characteristics of entities and their interactions, which are essential for fine-grained relation detection and intelligent perception in IoST-driven applications.

1) Entity Encoder: To refine the feature representations of subjects $\mathbf{s}$ and objects $\mathbf{o}$, we employ an entity encoder based on the prototype-based embedding network [6]

$$\mathbf{s} = \mathbf{W_s}\mathbf{t_s} + \mathbf{v_s} \tag{1}$$

$$\mathbf{o} = \mathbf{W_o}\mathbf{t_o} + \mathbf{v_o} \tag{2}$$

where $\mathbf{W_s}$ and $\mathbf{W_o}$ are learnable parameters. $\mathbf{t_s}$ and $\mathbf{t_o}$ respectively represent the semantic features of the subject and object, derived from the GloVe embedding. Moreover, we leverage a gating mechanism to capture the distinctive contents of each subject and object as $\mathbf{v_s}$ and $\mathbf{v_o}$

$$\mathbf{v_s} = \sigma(\mathrm{FC}((\mathbf{W_s}\mathbf{t_s}) \oplus \mathrm{h}(\mathbf{x_s})) \odot \mathrm{h}(\mathbf{x_s}) \tag{3}$$

$$\mathbf{v_o} = \sigma(\mathrm{FC}((\mathbf{W_o}\mathbf{t_o}) \oplus \mathrm{h}(\mathbf{x_o})) \odot \mathrm{h}(\mathbf{x_o}) \tag{4}$$

where $\mathrm{FC}(\cdot)$ is the fully connected layer, $\mathrm{h}(\cdot)$ is the visual-to-semantic function, and $\sigma(\cdot)$ is the sigmoid activation function. $\oplus$ denotes concatenation operation and $\odot$ denotes element-wise product. $\mathbf{x_s}$ and $\mathbf{x_o}$ represent visual features of the subject and object.

2) Relation Encoder: After extracting the subject and object features, the relation encoder uses these representations to derive discriminative relation-centric features. This is a critical step in feature refinement for enhancing predicate classification in SGG. We adopt a two-step process to achieve this goal, addressing challenges such as background noise and the need for interaction-specific representations, ultimately preparing robust features for the CPC module.

In the first step, we focus on suppressing predicate-irrelevant background noise in the union feature $\mathbf{u}$, which encapsulates the subject status, object status, and their interaction patterns. To this end, we apply the gating mechanism to filter out background noise from $\mathbf{u}$

$$\mathbf{u_{wb}} = \sigma(\mathrm{FC}(\mathrm{F}(\mathbf{s}, \mathbf{o}) \oplus \mathrm{h}(\mathbf{u}))) \odot \mathrm{h}(\mathbf{u}) \tag{5}$$

where $\mathbf{u_{wb}}$ represents the background-suppressed union feature. $\mathrm{F}(\mathbf{s}, \mathbf{o}) = \mathrm{ReLU}(\mathbf{s} + \mathbf{o}) - (\mathbf{s} - \mathbf{o})^2$ integrates contents of the subject and object.

In the second step, we further refine $\mathbf{u_{wb}}$ to isolate interaction-specific content by subtracting the subject-object status $\mathrm{F}(\mathbf{s}, \mathbf{o})$, thereby ensuring that the resulting representation focuses on relational dynamics. The final relation representation $\mathbf{r}$ is derived as

$$\mathbf{r} = \mathbf{u_{wb}} - \mathrm{F}(\mathbf{s}, \mathbf{o}) \tag{6}$$

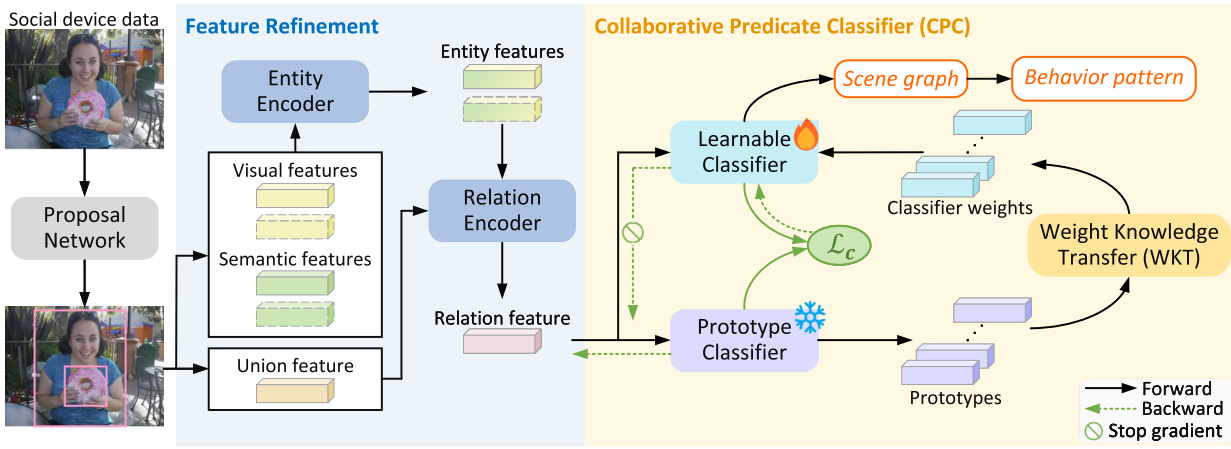


Fig. 2. Overview of the CDC framework. The CDC framework is mainly composed of three modules: 1) the proposal network perceives and detects visual data from social devices, generating a set of entity proposals; 2) the feature refinement module refines entity and relation representations by leveraging multimodal features; and 3) the CPC module constructs an optimal prototype classifier to guide feature refinement. It simultaneously devises a learnable classifier to adjust decision boundaries. To boost the recalibration process, a WKT module and a collaborative constraint term are introduced to support effective cooperative learning.

C. CPC

In this section, we elaborate on the collaboration mechanism of dual classifiers and their design for fine-grained relation recognition. The CPC module leverages the complementary strengths of a prototype classifier and a learnable classifier to address the long-tail challenge in SGG, which is a critical barrier to modeling nuanced social behaviors in IoST-driven applications.

Unlike conventional methods that rely on joint learning of feature representations and classifiers, our CPC module introduces a cooperative learning mechanism to address the long-tail challenge. As shown in Fig. 3, traditional SGG pipelines [Fig. 3(a)] jointly optimize the relation encoder and classifier, often leading to overfitting on head predicates and poor performance on tail predicates. In contrast, our CPC module [Fig. 3(b)] leverages the complementary strengths of a prototype classifier and a learnable classifier. The prototype classifier establishes a balanced geometric configuration to guide relation encoding, mitigating bias toward head predicates. Meanwhile, the learnable classifier dynamically adjusts decision boundaries to enhance recognition of tail predicates. The cooperative mechanism, facilitated by the WKT module and a collaborative constraint term, ensures that the two classifiers work synergistically to improve fine-grained relation modeling.

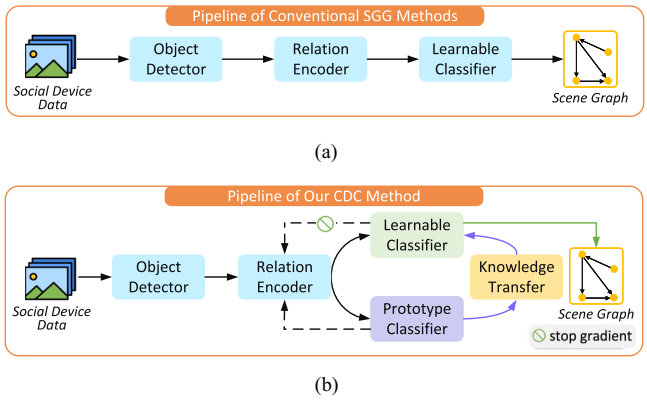


Fig. 3. Comparison of different pipelines for SGG. (a) Conventional methods jointly learn the relation encoder and the classifier, which often leads to bias toward head predicates. (b) Our CDC method introduces cooperative learning between a prototype classifier and a learnable classifier to address the long-tail challenge, thereby enhancing fine-grained relation modeling.

1) Prototype Classifier: The design of the prototype classifier is crucial for addressing the long-tail challenge in SGG datasets. Such imbalance distorts the feature space, causing class centers of tail predicates to cluster closely together and leading to interclass overlap, which impairs the classifier's relation recognition capability. The neural collapse theory [13] demonstrates that on balanced datasets, the within-class means of features and classifier vectors align with the vertices of a simplex equiangular tight frame at the final training stage, maximizing interclass angular separation. Inspired by this, we employ a prototype classifier with maximum interclass margins to counteract this imbalance. Specifically, we construct an optimal prototype classifier by learning a balanced geometric configuration in a task-independent manner, as illustrated in Fig. 4. This classifier is preestablished before SGG training. The establishment solution ensures that tail classes with sparse data are allocated sufficient representation space, enhancing their distinguishability. During SGG training, the frozen prototype classifier guides the feature refinement process by aligning features with this balanced structure, thereby reducing head category dominance and establishing a robust foundation for fine-grained relation modeling.

To realize the balanced geometric configuration of the prototype classifier, we construct the prototype matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n] \in \mathbb{R}^{d \times n}$. Each prototype vector is initialized by sampling values from a uniform distribution $\mathcal{U}(-1, 1)$. Here, d denotes the feature dimension, and n represents the number of relation classes. We then employ the prototype regularization loss $\mathcal{L}_{pr}$ to supervise prototype learning, randomly selecting n'

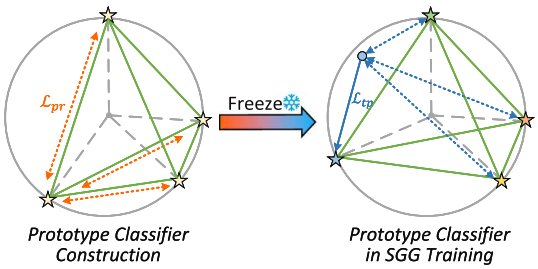


Fig. 4. Pipeline for constructing the prototype classifier and its application in SGG training to mitigate long-tail bias.

prototypes for optimization in every training iteration

$$\mathcal{L}_{pr} = \log\left(\frac{1}{n'}\sum_{i=1}^{n'}(g_+^i + g_-^i)\right) \tag{7}$$

where g_+^i and g_-^i are calculated using the Gaussian potential kernel G between the selected prototype $\mathbf{p}_i$ and itself, as well as between $\mathbf{p}_i$ and all other prototypes

$$g_+^i = \mathrm{G}(\mathbf{p}_i, \mathbf{p}_i), \quad g_-^i = \sum_{j=1, j\neq i}^{n'} \mathrm{G}(\mathbf{p}_i, \mathbf{p}_j) \tag{8}$$

$$\mathrm{G}(\mathbf{p}_i, \mathbf{p}_j) = e^{-\alpha\|\mathbf{p}_i - \mathbf{p}_j\|_2^2}, \alpha > 0 \tag{9}$$

here $\|\cdot\|_2$ represents the L2 norm, and α is a hyper-parameter.

After establishing the optimal prototype classifier $\mathbf{P}$, we freeze its parameters during the SGG training phase and apply the class-balanced loss [23] along with the triplet loss to optimize the feature refinement module. To compute the class-balanced loss $\mathcal{L}_{cb1}$, we first calculate the classification score y_p as the cosine similarity between the relation representation $\mathbf{r}$ and the prototypes $\mathbf{P}$

$$y_p = \frac{\mathbf{r}\cdot\mathbf{P}}{\|\mathbf{r}\|_2\|\mathbf{P}\|_2}. \tag{10}$$

The classification score guides the computation of the class-balanced loss $\mathcal{L}_{cb1}$, which addresses the long-tail distribution by reweighting classes based on their frequency, as detailed in [23]. Additionally, we introduce a triplet loss $\mathcal{L}_{tp}$ to enhance the feature refinement by ensuring that the relation representation $\mathbf{r}$ is positioned closer to its ground truth prototype while being sufficiently distant from others in the Euclidean space. To achieve this, we first compute the Euclidean distances D between $\mathbf{r}$ and all prototypes in $\mathbf{P}$

$$D = \{\|\mathbf{r} - \mathbf{p}_i\|_2^2 \mid \mathbf{p}_i \in \mathbf{P}, 0 < i \leq n\} \tag{11}$$

where n is the number of relation classes.

The positive distance d_+ is defined as the Euclidean distance between $\mathbf{r}$ and its ground truth prototype $\mathbf{p}_{gt}$, i.e., $d_+ = \|\mathbf{r} - \mathbf{p}_{gt}\|_2^2$. For the negative distance d_-, we select a subset of k negative prototypes $\mathbf{P}^-$ by identifying the k nearest prototypes based on Euclidean distances, excluding the ground truth prototype. The negative distance d_- is then computed as the average distance to these selected prototypes

$$d_- = \frac{1}{k}\sum_{\mathbf{p}_i \in \mathbf{P}^-} \|\mathbf{r} - \mathbf{p}_i\|_2^2. \tag{12}$$

Algorithm 1: Prototype Classifier Construction.

1: Initialize prototype matrix $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_n] \in \mathbb{R}^{d\times n}$ with values sampled from uniform distribution $\mathcal{U}(-1, 1)$
2: **for** epoch $t = 1...T$ **do**
3: Randomly select n' prototypes from $\mathbf{P}$
4: Compute Gaussian potential kernel for each selected prototype $\mathbf{p}_i$ using (9)
5: Compute the prototype regularization loss $\mathcal{L}_{pr}$ using (7)
6: Update $\mathbf{P}$ by minimizing $\mathcal{L}_{pr}$
7: **end for**
8: **return** $\mathbf{P}$

The triplet loss $\mathcal{L}_{tp}$ is then formulated to enforce a margin between the positive and negative distances

$$\mathcal{L}_{tp} = \max(0, d_+ - d_- + \delta) \tag{13}$$

where δ is a margin hyperparameter that controls the separation threshold.

In summary, we present Algorithm 1 to describe the establishment process of the prototype classifier.

2) WKT: To enhance the precision of relation classification, we propose the WKT module. This module synergistically leverages the balanced geometric configuration prior of the prototype classifier, while empowering the learnable classifier to refine the decision boundaries. Specifically, this module transforms the prototypes $\mathbf{P}$ into the weights $\mathbf{W_l}$ of the learnable classifier through a MLP composed of two linear layers, each followed by batch normalization and ReLU activation, formalized as

$$\mathbf{W_l} = \mathrm{WKT}(\mathbf{P}). \tag{14}$$

The WKT module is of paramount importance to the entire collaborative mechanism, facilitating a dynamic interplay between the prototype and learnable classifiers. While the prototype classifier, pre-trained with an optimal balanced geometric configuration (e.g., a simplex equiangular tight frame), maximizes inter-class separation, its fixed decision boundaries may not effectively adapt to the varying feature distributions across classes. The WKT module addresses this limitation by enabling the learnable classifier to dynamically adjust these boundaries. It inherits the prototype's prior knowledge of relation class distributions while optimizing the decision surface to better fit the complex feature distributions of both head and tail predicates. Moreover, this process enhances the learning efficiency of the learnable classifier.

3) Learnable Classifier: To address the limitations of the prototype classifier's rigid decision boundary, we design the learnable classifier as a complementary component. Leveraging the knowledge transferred via the WKT module, the learnable classifier dynamically refines the decision boundary, formulated as follows:

$$y_l = \mathbf{W_l}\mathbf{r} + \mathbf{b} \tag{15}$$

where $\mathbf{b}$ is a learnable bias parameter, and y_l denotes the classification score of the learnable classifier to predict the relationship class.

Importantly, the predicted relationship, obtained from y_l, serves as the semantic basis for behavioral understanding, enabling our framework to extract socially meaningful patterns from visual content. This makes y_l not only a decision output, but also a bridge connecting low-level visual representations with high-level behavioral-cultural interpretation.

To optimize the learnable classifier independently, we employ the class-balanced loss $\mathcal{L}_{cb2}$, which focuses on refining its parameters without propagating gradients to the feature refinement module, as illustrated in Fig. 2. This design ensures that the feature refinement module remains unaffected by the optimization of the learnable classifier, allowing for a more stable training process. Tailored to mitigate the long-tail bias, the class-balanced loss $\mathcal{L}_{cb2}$ assigns higher weights to tail predicates, encouraging the learnable classifier to pay greater attention to underrepresented classes.

To further enhance the adaptability of the learnable classifier's decision boundaries, we introduce a collaborative constraint term $\mathcal{L}_c$, formulated as follows:

$$\mathcal{L}_c = \max(0, \mathcal{L}_{cb2} - \mathcal{L}_{cb1} + \gamma) \tag{16}$$

where γ is a hyper-parameter.

The collaborative constraint term $\mathcal{L}_c$ plays a pivotal role in ensuring that the learnable classifier benefits from the prototype classifier's guidance while dynamically adjusting its decision boundary to better fit the data distribution. By enforcing a margin between the performance of the learnable classifier (measured by $\mathcal{L}_{cb2}$) and that of the prototype classifier (measured by $\mathcal{L}_{cb1}$), this term encourages the learnable classifier to outperform the prototype classifier, particularly in scenarios requiring fine-grained distinctions. This collaborative mechanism not only mitigates the rigidity of the prototype classifier's decision boundary but also enhances the model's ability to generalize across diverse relation categories.

During the SGG training stage, the overall training loss $\mathcal{L}$ is defined as

$$\mathcal{L} = \beta_1 \mathcal{L}_{cb1} + \beta_2 \mathcal{L}_{tp} + \beta_3 \mathcal{L}_{cb2} + \beta_4 \mathcal{L}_c \tag{17}$$

where the hyper-parameters β_1, β_2, β_3, and β_4 are tuned to balance the relative contributions of each loss term, ensuring that the CDC framework achieves both robustness and accuracy in fine-grained relation prediction. The class-balanced loss $\mathcal{L}_{cb1}$ and triplet loss $\mathcal{L}_{tp}$ focus on optimizing the feature refinement module guided by the prototype classifier, while $\mathcal{L}_{cb2}$ and $\mathcal{L}_c$ refine the learnable classifier's performance. The balanced loss configuration enhances the CDC framework's generalization across diverse IoST-driven scenarios, such as behavioral analysis in smart environments, where precise predicate classification is vital for downstream tasks.

In summary, we design Algorithm 2 to describe the CPC module.

Algorithm 2: The Collaborative Predicate Classifier (CPC) Module.

Input: Pre-trained prototype matrix $\mathbf{P} \in \mathbb{R}^{d \times n}$, relation representation $\mathbf{r}$
Output: Classification score y_l
Step 1: Prototype Classification
Compute prototype classifier score $y_p \leftarrow \frac{\mathbf{r} \cdot \mathbf{P}}{\|\mathbf{r}\|_2 \|\mathbf{P}\|_2}$
Step 2: Weight Knowledge Transfer
Compute learnable classifier weights $\mathbf{W_l} \leftarrow \mathrm{WKT}(\mathbf{P})$
Step 3: Learnable Classification with Dynamic Boundary Adjustment
Compute learnable classifier score $y_l \leftarrow \mathbf{W_l}\mathbf{r} + \mathbf{b}$
if training **then**
    Compute class-balanced loss $\mathcal{L}_{cb1}$, $\mathcal{L}_{cb2}$
    Compute triplet loss $\mathcal{L}_{tp}$
    Compute collaborative constraint term $\mathcal{L}_c$
    Total loss $\mathcal{L} \leftarrow \beta_1 \mathcal{L}_{cb1} + \beta_2 \mathcal{L}_{tp} + \beta_3 \mathcal{L}_{cb2} + \beta_4 \mathcal{L}_c$
    Parameter update
else
    Inference: Use learnable classifier to get prediction y_l
end if
return y_l

IV. Experiment and Analysis

A. Experimental Settings

We evaluate our method on two commonly used SGG datasets, namely VG [29] and GQA [30]. The VG dataset is the most prevalent benchmark, containing 108 000 images. Each image is annotated with an average of 38 objects and 22 relationships. In this article, we adopt the widely used VG150 split for VG. This split collects images from the social platform Flickr and retains the 150 most frequent object categories and 50 predicate categories. We use 70% of the images for training, 30% for testing, and 5000 images from the training set for validation. The GQA dataset, derived from VG, refines annotations by removing inaccurate predicates and enriching object and relation labels. For GQA, we use the GQA200 split, which includes 200 object categories and 100 predicate categories.

We evaluate our method on three tasks: 1) predicate classification (PredCls) infers the predicates of entity pairs with ground-truth bounding boxes and class labels; 2) scene graph classification (SGCls) predicts both the entity class labels and their pairwise relationships with ground-truth bounding boxes; and 3) scene graph detection (SGDet) jointly detects entities and their pairwise relationships from raw images without access to ground-truth bounding boxes and labels, making it a more challenging and realistic SGG setting.

Consistent with recent works [6], [21], [25], [31], we evaluate the performance of SGG methods on the VG150 and GQA200 datasets using two standard metrics: Recall@K (R@K) and mean Recall@K (mR@K). While R@K measures the proportion of correct triplets among the top-K predictions, it is inherently biased toward head predicates due to their higher frequency. In contrast, mR@K computes the average R@K across all predicate classes, thus offering a more balanced evaluation that highlights the model's performance on rare tail predicates. As such, mR@K serves as a more informative and critical metric for assessing fine-grained relation understanding under long-tail distributions.

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART SGG METHODS ON VG DATASET

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
VTransE [14]	65.7/67.6	14.7/15.8	38.6/39.4	8.2/8.7	29.7/34.3	5.0/6.1
MOTIFS [16]	65.2/67.0	14.8/16.1	38.9/39.8	8.3/8.8	32.8/37.2	6.8/7.9
VCTree [17]	65.4/67.2	16.7/18.2	**46.7**/**47.6**	11.8/12.5	31.9/36.2	7.4/8.7
GPS-Net [8]	65.2/67.1	15.2/16.6	37.8/39.2	8.5/9.1	31.3/35.9	6.7/8.6
BGNN [9]	59.2/61.3	30.4/32.9	37.4/38.5	14.3/16.5	31.0/35.8	10.7/12.6
BPL-SA$^\diamond$ [35]	50.7/52.5	29.7/31.7	30.1/31.0	16.5/17.5	23.0/26.9	13.5/15.6
HL-Net [15]	**67.0**/**68.9**	-/22.8	**42.6**/**43.5**	-/13.5	**33.7**/**38.1**	-/9.2
RU-Net [19]	**67.7**/**69.6**	-/24.2	42.4/43.3	-/14.6	**32.9**/**37.5**	-/10.8
GCL$^\diamond$ [36]	42.7/44.4	36.1/38.2	26.1/27.1	20.8/21.8	18.4/22.0	**16.8**/**19.3**
NICE [37]	55.1/57.2	29.9/32.3	33.1/34.0	16.6/17.9	27.8/31.8	12.2/14.4
IETrans$^\diamond$ [38]	48.6/50.5	35.8/39.1	29.4/30.2	**21.5**/**22.8**	23.5/27.2	15.5/18.0
HetSGG [39]	57.8/59.1	31.6/33.5	37.6/38.7	17.2/18.7	30.0/34.6	12.2/14.4
DKBL$^\diamond$ [40]	57.2/58.8	29.7/32.2	32.7/33.4	18.2/19.4	27.0/30.7	12.6/15.1
CV-SGG [21]	58.2/62.4	32.6/36.2	-/-	-/-	27.8/32.0	14.6/17.0
INF$^\diamond$ [41]	51.5/55.1	24.7/30.7	32.2/33.8	14.5/17.4	23.9/27.1	9.4/11.7
PE-Net [6]	64.9/67.2	31.5/33.8	39.4/40.7	17.8/18.9	30.7/35.2	12.4/14.5
SQUAT [42]	55.7/57.9	30.9/33.4	32.9/34.3	17.5/18.9	24.5/28.9	14.1/16.5
CFA$^\diamond$ [20]	54.1/56.6	35.7/38.2	34.9/36.1	17.0/18.4	27.4/31.8	13.2/15.5
EICR$^\diamond$ [43]	55.3/57.4	34.9/37.0	34.5/35.4	20.8/21.8	27.9/32.2	15.5/18.2
VETO [25]	64.2/66.3	22.8/24.7	35.7/36.9	11.1/11.9	27.5/31.5	8.1/9.5
HiKER-SGG [44]	-/-	**39.3**/**41.2**	-/-	20.3/21.4	-/-	-/-
DPL$^\diamond$ [31]	54.4/56.3	33.7/37.4	32.6/33.8	18.5/20.1	24.5/28.7	13.0/15.6
SBG$^\diamond$ [45]	55.4/57.3	32.1/34.4	34.9/35.7	17.5/18.6	27.0/31.3	13.8/16.1
CooK+TF-*l*-IDF [46]	60.4/62.3	35.4/37.2	36.4/37.6	19.1/20.3	27.7/32.7	14.2/16.3
CDC (Ours)	37.9/42.3	**43.6**/**47.4**	23.6/26.0	**24.4**/**26.8**	23.3/27.6	**17.3**/**20.3**

Note: "$\diamond$" denotes MOTIFS with a model-agnostic method. The best and second best results are respectively marked in red and underline blue.

In this article, the pretrained Faster R-CNN [27] with ResNeXt-101-RPN [32], [33], [34] is utilized to detect objects in the images, following previous works [6], [25]. We optimize our method via SGD optimizer for 60 000 iterations, starting with a learning rate of 10^{-3} and a batch size of 8. We set the loss weight parameters β_1, β_2, β_3 and β_4 as 10, 1, 10 and 1, respectively. For other hyper-parameters, we set $\gamma = 0.01$, $\alpha = 2$, and $\delta = 0.95$. Moreover, k is set to 30 for VG, and 90 for GQA. All experiments are carried out using PyTorch and trained with an NVIDIA GeForce RTX 3090 GPU.

B. Overall Performance Comparison

Aiming to evaluate the effectiveness of our proposed CDC, we compare its performance against several state-of-the-art SGG methods on VG and GQA datasets in this section.

1) VG: In Table I, we present a comprehensive performance comparison of our CDC method against state-of-the-art SGG methods on the VG150 dataset, emphasizing fine-grained relation modeling under long-tail distributions. Among all evaluated methods, CDC achieves the highest mR@K scores across all three SGG tasks. Specifically, CDC attains mR@50/mR@100 of 43.6/47.4 in PredCls, 24.4/26.8 in SGCls, and 17.3/20.3 in SGDet, consistently outperforming all competing methods. This consistent superiority across tasks underscores CDC's robust capability in capturing fine-grained and long-tail relationships, an area where most prior approaches struggle due to their bias toward frequent head categories. Moreover, CDC maintains this advantage even in the more challenging SGDet setting, where object detection and relation prediction are jointly performed, highlighting its end-to-end effectiveness under noisy visual conditions.

Compared with the baseline PE-Net [6], CDC demonstrates significant improvements in addressing long-tail problems. In PredCls, CDC's mR@50 of 43.6 and mR@100 of 47.4 outperform PE-Net's 31.5 and 33.8 by relative increases of 38.4% and 40.2%, respectively. Similarly, in SGCls, CDC achieves mR@50 of 24.4 and mR@100 of 26.8, surpassing PE-Net's 17.8 and 18.9 by 37.1% and 41.8%. In SGDet, CDC further improves upon PE-Net's mR@50 and mR@100 of 12.4 and 14.5, reaching 17.3 and 20.3, which constitute relative gains of 39.5% and 40.0%. This marked enhancement stems from CDC's cooperative dual-classifier architecture, which integrates feature refinement and dynamic decision boundary adjustment to effectively distinguish subtle differences in predicate semantics. By leveraging fine-grained representations and adjusting decision boundaries, CDC mitigates the overfitting to dominant categories observed in PE-Net, enabling a more balanced and accurate representation of long-tail distributions. These results highlight CDC's potential for applications requiring nuanced scene understanding, such as social interaction analysis, where rare relationships are critical.

2) GQA: In Table II, we present a comprehensive performance comparison of our CDC method against state-of-the-art SGG methods on the GQA200 dataset, which poses a more fine-grained and semantically diverse challenge compared with VG150. Across all three SGG tasks, CDC achieves the highest mR@K scores, demonstrating its superior generalization ability across a broader range of relations. Specifically, CDC obtains mR@50/100 of 40.3/42.7 in PredCls, 18.6/19.8 in SGCls, and

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART SGG METHODS ON GQA200 DATASET

Methods	PredCls		SGCls		SGDet	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
VTransE [14]	55.7/57.9	14.0/15.0	33.4/34.2	8.1/8.7	27.2/30.7	5.8/6.6
MOTIFS [16]	**65.2**/**66.8**	16.4/17.1	**34.2**/**34.9**	8.2/8.6	**28.9**/**33.1**	6.4/7.7
VCTree [17]	63.8/65.7	16.6/17.4	**34.1**/**34.8**	7.9/8.3	**28.3**/**31.9**	6.5/7.4
CFA$^{\diamond}$ [20]	-/-	31.7/33.8	-/-	14.2/15.2	-/-	11.6/13.2
GCL$^{\diamond}$ [36]	44.5/46.2	**36.7**/**38.1**	23.2/24.0	**17.3**/18.1	18.5/21.8	**16.8**/**18.8**
EICR$^{\diamond}$ [43]	56.4/58.1	36.3/38.0	28.8/29.4	17.2/**18.2**	24.6/28.4	16.0/18.0
VETO [25]	**64.5**/**66.0**	21.2/22.1	30.4/31.5	8.6/9.1	26.1/29.0	7.0/8.1
DPL$^{\diamond}$ [31]	50.3/52.3	31.6/33.9	25.0/25.9	13.3/14.4	15.0/19.0	11.1/13.1
CDC (Ours)	36.9/40.1	**40.3**/**42.7**	19.0/20.5	**18.6**/**19.8**	16.2/19.8	**17.1**/**19.2**

Note: "$\diamond$" denotes MOTIFS with a model-agnostic method. The best and second best results are respectively marked in red and underline blue.

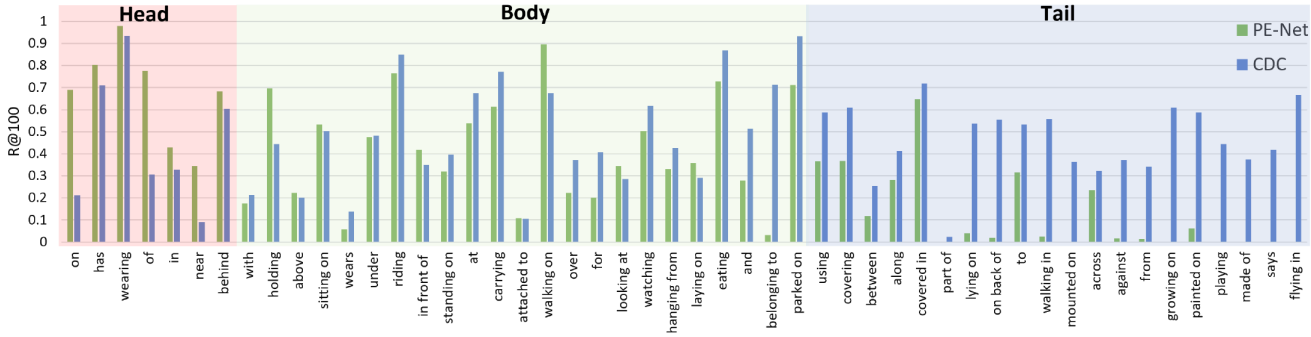


Fig. 5. Comparison of R@100 of all predicate classes under the PredCls task on the VG dataset. Predicates are sorted in decreasing order of sample frequency, and are divided into three groups: Head (red), body (green), and tail (blue).

17.1/19.2 in SGDet, consistently outperforming both traditional pipelines such as MOTIFS and VCTree, which favor high-frequency patterns, and model-agnostic designs such as CFA, GCL, and EICR. This consistent superiority highlights CDC's capacity to recover rare and diverse relations, even under the compounded noise of joint object detection and relation inference in SGDet.

While CDC significantly improves the performance on tail predicates, as reflected by substantial gains in mR@K, we observe a slight drop in head class performance. This trade-off emerges not from an explicit suppression of head classes, but rather from CDC's capacity to allocate greater representational focus to underrepresented relations through its dual-classifier design. In practice, this trade-off is often beneficial in downstream tasks that require nuanced understanding of less frequent but semantically meaningful relationships, such as safety-aware perception, long-tail human–object interaction understanding, and behavioral-cultural modeling in IoST environments.

C. Detailed Performance Comparison

In this section, we conduct an in-depth evaluation of the proposed CDC framework by comparing its performance against the baseline PE-Net [6] on the VG150 dataset, focusing on per-class predicate recognition and classification consistency.

First, we analyze the performance across different frequency classes compared with PE-Net in Fig. 5. Specifically, we

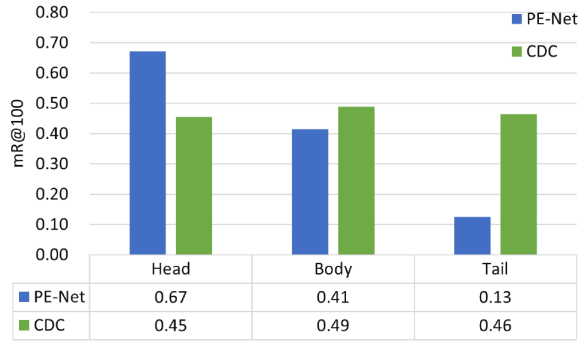


Fig. 6. Comparison of mR@100 for head, body, and tail predicate groups under the PredCls task on the VG dataset.

divide the 50 predicate classes into three groups based on instance counts in the training set: head (7), body (23), tail (20), and provide comparison of groups in Fig. 6. In comparison with PE-Net, our CDC successfully identifies all 50 predicates, while PE-Net fails to classify 6 of them. Additionally, CDC demonstrates performance improvements in R@100 for 35 relations, predominantly those fine-grained body and tail relations. Notably, CDC successfully identifies the relation "flying in", which has only four training samples but achieves an R@100 of 66.67%. As shown in Fig. 6, CDC outperforms PE-Net in mR@100 for both the body and tail groups, achieving 0.49 versus 0.41 (a 19.5% relative increase) for the body group

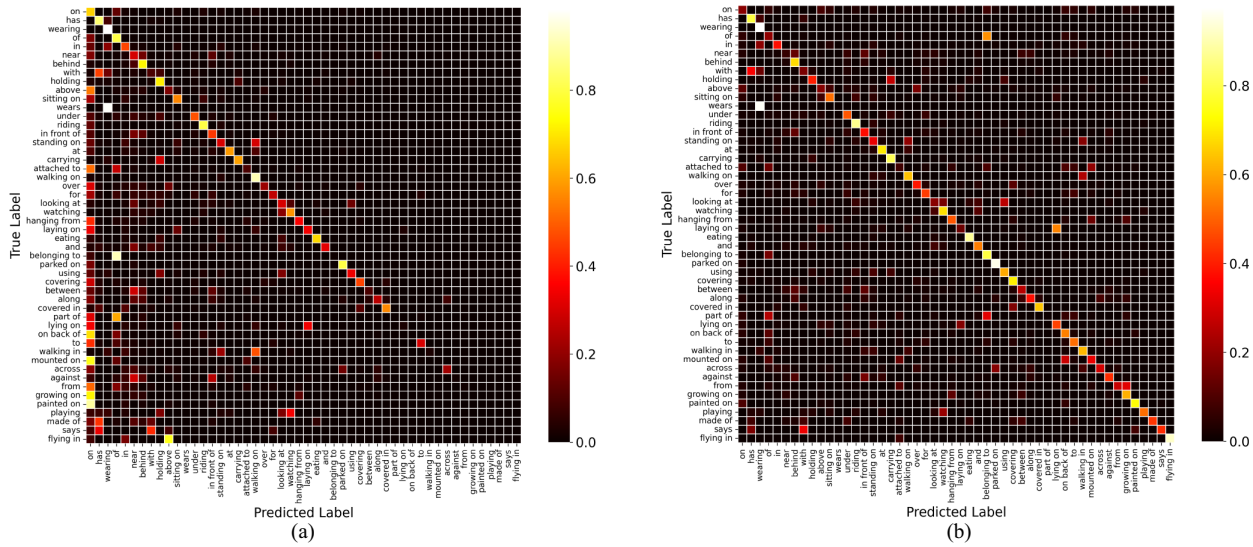


Fig. 7. Confusion matrices of predicates under the VG150 dataset. Predicates are sorted in decreasing order of sample frequency. (a) PE-Net. (b) CDC.

TABLE III
ABLATION STUDIES ON EACH COMPONENT OF CDC

Exp	Module				PredCls				SGCls			
	P	L	W	C	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100
0	✗	✗	✗	✗	39.8	41.8	36.1	38.5	27.3	28.5	21.4	22.4
1	✓	✗	✗	✗	57.8	62.2	37.2	41.1	35.0	37.8	20.6	22.6
2	✓	✓	✗	✗	52.0	56.7	7.4	9.2	28.3	31.7	4.0	5.2
3	✓	✓	✓	✗	36.0	40.2	43.4	47.0	21.8	24.2	24.0	26.4
4	✓	✓	✓	✓	37.9	42.3	43.6	47.4	23.6	26.0	24.4	26.8

Note: "P", "L", "W", and "C" denote the fixed prototype classifier, the learnable classifier, the weight knowledge transfer module, and the collaborative constraint term, respectively.

and 0.46 versus 0.13 (a 253.8% increase) for the tail group. Although PE-Net achieves higher mR@100 in the head group (0.67 versus CDC's 0.45), CDC's substantial gains in the body and tail groups highlight its superior ability to handle fine-grained and long-tail predicates, addressing the key challenge in SGG.

Second, we visualize the confusion matrices on the VG150 dataset to further investigate the classification consistency of the CDC framework in comparison with the baseline PE-Net [6]. As shown in Fig. 7(a), PE-Net's confusion matrix reveals a significant bias toward head predicates, such as "on", which dominate the predictions. Many tail predicates, including "painted on", "on back of", and "growing on", are frequently misclassified into head categories, resulting in low-recall or even absent predictions for numerous tail classes. This indicates that PE-Net struggles to capture the diversity of fine-grained relations, particularly those underrepresented tail predicates, leading to an imbalanced scene graph representation.

In contrast, the confusion matrix of CDC [Fig. 7(b)] exhibits a prominently highlighted diagonal, demonstrating that all 50 predicates are successfully recognized with high recall across categories. Notably, CDC not only achieves more balanced predictions across head, body, and tail predicates but also reassigns many instances that are previously predicted as head predicates by PE-Net (e.g., "on," "has") to more informative tail predicates (e.g., "sitting on", "belonging to"). This reassignment enhances the informativeness of the generated scene graphs by uncovering nuanced relationships that PE-Net tends to overlook. For instance, predicates such as "flying in", which are nearly absent in PE-Net's predictions, show significant recall in CDC's confusion matrix, reflecting its superior capability in addressing long-tail challenges. These improvements stem from CDC's cooperative dual-classifier architecture, which dynamically adjusts decision boundaries to better accommodate tail predicates, thereby producing more comprehensive and equitable scene graphs.

D. Ablation Studies

1) Ablation on Model Components: To assess the contribution of each component in the CDC framework, we conduct ablation studies on the VG150 dataset, with results summarized in Table III. In Exp0, we use a learnable prototype classifier optimized with the class-balanced loss and triplet loss to recognize relations. In Exp1, we evaluate the baseline performance using only the fixed prototype classifier (P) for predicate classification, establishing a foundation for coarse relation prediction. Exp2 introduces the learnable classifier (L) to refine decision

TABLE IV
HYPERPARAMETER ANALYSIS OF THE MARGIN δ

δ	PredCls			
	R@50	R@100	mR@50	mR@100
0.90	36.2	40.7	43.0	46.8
0.95	37.9	42.3	**43.6**	**47.4**
1.00	**38.6**	**43.7**	41.3	46.3

Note: The best results are shown in boldface.

boundaries, aiming for more precise classification. Exp3 incorporates the WKT module (W) to enhance the adjustment of decision boundaries, and Exp4 adds the collaborative constraint term (C) to ensure robust and efficient decision boundary adjustments, completing the full CDC framework.

The results in Table III reveal the impact of each component on both the PredCls and SGCls tasks. In Exp1, relying solely on the prototype classifier yields strong R@K scores (57.8 R@50 and 62.2 R@100 in PredCls; 35.0 R@50 and 37.8 R@100 in SGCls), indicating its effectiveness in capturing common relations. However, its mR@K scores (37.2 mR@50 and 41.1 mR@100 in PredCls) suggest limited capability in handling long-tail predicates. In contrast, Exp0, using a learnable prototype classifier, exhibits poorer unbiasedness between head and tail categories, sacrificing recall on head categories. This comparison reinforces our decision to adopt the fixed prototype classifier in subsequent experiments, as it provides a more balanced foundation for relation prediction. Additionally, our motivation to enable the learnable classifier to inherit the equidistributed structure of the fixed prototype further supports this choice. Incorporating the learnable classifier in Exp2 (P + L) significantly reduces R@K (e.g., 52.0 R@50 in PredCls) and mR@K (7.4 mR@50 and 9.2 mR@100 in PredCls), highlighting that the learnable classifier alone disrupts the balance achieved by the prototype classifier, likely due to overfitting to head categories without proper guidance. The introduction of the WKT module in Exp3 (P + L + W) markedly improves mR@K (43.4 mR@50 and 47.0 mR@100 in PredCls; 24.0 mR@50 and 26.4 mR@100 in SGCls), demonstrating its critical role in guiding the learnable classifier to recalibrate decision boundaries effectively, thus enhancing the model's ability to capture fine-grained and long-tail relations. However, the R@K scores remain lower (36.0 R@50 in PredCls), reflecting a trade-off for improved graph diversity. Finally, Exp4 (P + L + W + C) incorporates the collaborative constraint term, achieving the best mR@K performance (43.6 mR@50 and 47.4 mR@100 in PredCls; 24.4 mR@50 and 26.8 mR@100 in SGCls) while slightly boosting R@K (37.9 R@50 and 42.3 R@100 in PredCls), indicating that the constraint term stabilizes the recalibration process and improves overall robustness.

These findings underscore the synergistic effect of CDC's components. The prototype classifier provides a robust foundation for coarse classification, while the learnable classifier, guided by the WKT module and the collaborative constraint term, refines decision boundaries to excel in fine-grained relation modeling, particularly for long-tail predicates. This cooperative mechanism ensures a balanced trade-off between recall on common relations and diversity in rare relations, thereby validating the effectiveness of the full CDC framework in addressing the long-tail challenge in SGG.

2) Geometric Properties of the Prototype Classifier: The effectiveness of the fixed prototype classifier fundamentally relies on its ability to maintain large inter-class separation in the embedding space, thereby alleviating the feature overlap issues prevalent in long-tail relation distributions. To validate this property, we visualize and compare the prototypes of our method and PE-Net in Fig. 8. Specifically, we analyze two key geometric metrics, the pairwise cosine similarity among all prototypes and their pairwise Euclidean distances.

Fig. 8(a) and (b) presents the cosine similarity heatmaps among prototypes. Compared with PE-Net, whose certain prototype pairs show high cosine similarity, our method yields a nearly orthogonal configuration, with similarity scores close to 0. This indicates that the learned prototypes are well-separated in the angular space.

Fig. 8(c) and (d) further shows the Euclidean distance heatmaps. It can be observed that PE-Net tends to produce prototypes with relatively small Euclidean distances for some class pairs, which reflects a potential tendency toward class overlap in the embedding space. In contrast, our approach achieves uniformly large Euclidean distances, with most distances approaching 1.43, which confirms that the prototypes are not only angularly decorrelated but also spatially dispersed.

Overall, these results demonstrate that the proposed fixed prototype classifier effectively enforces a discriminative structure where interclass prototypes are evenly distributed in the embedding space, reducing the risk of feature confusion especially among tail classes.

3) Analysis on Margin Hyperparameter: To assess the influence of the triplet loss margin δ on model performance, we conduct the experiment by varying δ in the range $\{0.90, 0.95, 1.00\}$. Table IV reports the results under the PredCls setting.

We observe that $\delta = 1.00$ achieves the highest Recall@K (R@50 and R@100), indicating a slight improvement in overall retrieval capability. However, $\delta = 0.95$ yields the best mean Recall (mR@50 and mR@100), reflecting more balanced performance across head and tail predicates, which is critical for fine-grained relation recognition under long-tail distributions. Given that the main objective of this work is to enhance fine-grained and unbiased predicate classification, we adopt $\delta = 0.95$ as the final setting in all experiments.

E. Complexity Analysis

To assess model complexity, we report the inference time and network size of our CDC method and several baselines in Table V. Concretely, for the inference time, we record time (s) for inferring a single image of VG dataset (i.e., batchsize = 1). For the network size, we count the total parameters of the whole models (i.e., the object detector is also considered). We employ one NVIDIA GeForce RTX 3090 device for the experiment, and all results are based on the same pretrained Faster-RCNN (backbone: ResNeXt-101-FPN).

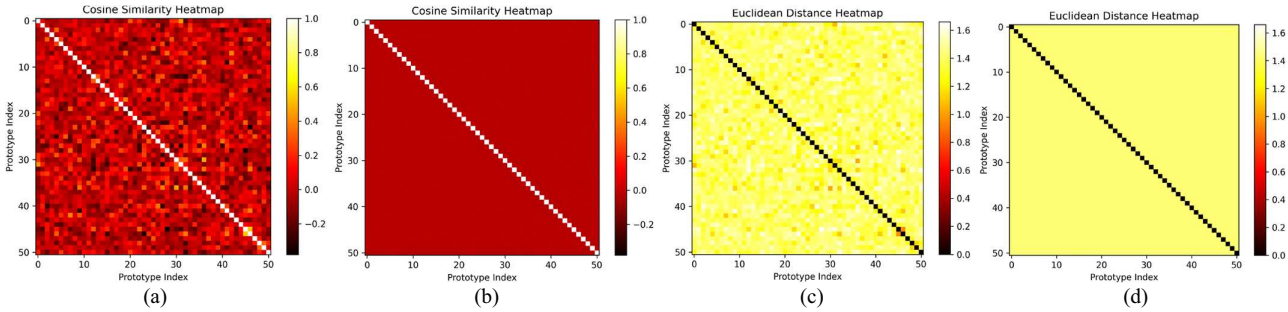


Fig. 8. Visualization of prototype geometric structure. (a) and (b) Cosine similarity between prototypes learned by PE-Net and our method CDC, respectively. (c) and (d) Euclidean distance heatmaps showing interprototype separation. Our method exhibits greater angular diversity and spatial dispersion.

TABLE V
COMPARISON OF INFERENCE TIME AND NETWORK SIZE

Methods	Inference Time (s)	Params (M)	SGDet mR@100
MOTIFS [16]	0.09	367	7.9
VCTree [17]	0.07	358	8.7
PE-Net [6]	0.31	411	14.5
DPL$^{\diamond}$ [31]	0.06	368	15.6
CDC (Ours)	0.29	459	20.3

Note: "$\diamond$" denotes MOTIFS with a model-agnostic method.

As demonstrated in Table V, the proposed CDC has a moderate increase in network parameters. Specifically, CDC incurs a 48M increase in parameters compared with PE-Net due to the incorporation of the cooperative dual-classifier and knowledge transfer modules. Nevertheless, it achieves a substantial improvement of 5.8% in mR@100, demonstrating a favorable trade-off between model complexity and performance. In terms of inference efficiency, CDC maintains a competitive inference time of 0.29 s per image, slightly faster than PE-Net (0.31 s). Although traditional approaches such as MOTIFS and VCTree, along with the recent method DPL, involve fewer parameters and are expected to offer faster inference speed, their SGG performance remains considerably lower. This demonstrates that the marginal increase in complexity brought by CDC remains practical and is well justified by its substantial improvements in fine-grained and unbiased SGG performance.

F. Visualization Results

In this section, we compare the generated scene graphs of the baseline PE-Net and CDC. As illustrated in Fig. 9, the proposed CDC generates more semantically informative triplets, such as $\langle$woman, playing, racket$\rangle$ versus $\langle$woman, holding, racket$\rangle$, $\langle$man, lying on, surfboard$\rangle$ versus $\langle$man, on, surfboard$\rangle$, and $\langle$clock, mounted on, building$\rangle$ versus $\langle$clock, on, building$\rangle$. These examples highlight CDC's ability to discern subtle behavioral and spatial dynamics, elevating the granularity of scene understanding. Additionally, the results underscore the effectiveness of our method in leveraging social device data to capture intricate behavioral patterns across "man–man", "man–machine" and "machine–machine" interactions. This improvement is evident in its capacity to model dynamic actions (e.g., "playing" versus "holding") and spatial relationships (e.g., "mounted on" versus "on"), offering a more comprehensive representation of real-world scenes.

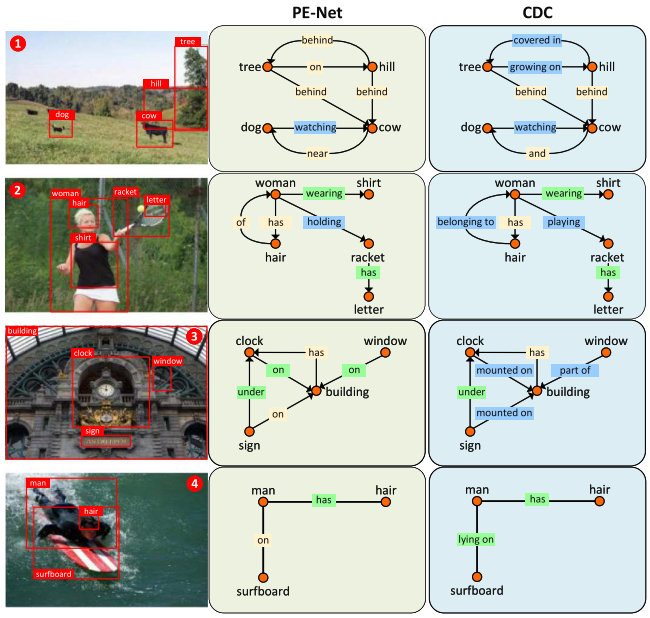


Fig. 9. Visualization results of PE-Net (in green) and CDC (in blue) under PredCls task on the VG dataset. Predicates matching the ground truth are highlighted in green. Yellow color represents predicates in head class, while blue color represents predicates in body/tail class.

V. CONCLUSION

In this article, we propose the CDC framework to address the challenges of fine-grained SGG within the IoST framework, particularly under long-tail distributions prevalent in real-world social device data. By integrating a cooperative learning mechanism, CDC synergistically combines a prototype classifier, which enforces equidistributed interclass separation to alleviate long-tail bias, and a learnable classifier, which dynamically adjusts decision boundaries for accurate relation prediction. To further enhance this synergy, we introduce a WKT module and a collaborative constraint term, ensuring robust performance in capturing tail predicates. Comprehensive experiments on the VG and GQA datasets validate the effectiveness of CDC, demonstrating superior performance in fine-grained relation

classification and improved generalization to underrepresented predicates. Beyond advancing semantic scene understanding, CDC offers a scalable and interpretable framework for modeling complex behavioral and cultural dynamics embedded in IoST environments, thereby offering new insights for downstream applications such as social behavior analysis, cultural context reasoning, and safety-critical perception in interconnected systems.

REFERENCES

[1] X. Zhou et al., "Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1742–1751, Aug. 2023.

[2] T. Qian, J. Chen, S. Chen, B. Wu, and Y.-G. Jiang, "Scene graph refinement network for visual question answering," *IEEE Trans. Multimedia*, vol. 25, pp. 3950–3961, 2023.

[3] S. Wang, F. Zhou, M. Yang, L. Shi, and C. Tan, "SGG-MVAR: Cross-modal retrieval with scene graph generation and multiview attribute relationship guidance," *IEEE Trans. Comput. Soc. Syst*,, early access, 2025.

[4] K. P. Singh, J. Salvador, L. Weihs, and A. Kembhavi, "Scene graph contrastive learning for embodied navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10884–10894.

[5] X. Zhou, W. Liang, I. Kevin, K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 171–178, Feb. 2020.

[6] C. Zheng, X. Lyu, L. Gao, B. Dai, and J. Song, "Prototype-based embedding network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis., Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 22783–22792.

[7] W. Li, H. Zhang, Q. Bai, G. Zhao, N. Jiang, and X. Yuan, "PPDL: Predicate probability distribution based loss for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., New Orleans, LA, USA, Jun. 2022, pp. 19447–19456.

[8] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph property sensing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Seattle, WA, USA, Jun. 2020, pp. 3746–3753.

[9] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Virtual Event*, Jun. 2021, pp. 11109–11119.

[10] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Seattle, WA, USA, Jun. 2020, pp. 3716–3725.

[11] X. Zhou et al., "Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 4, pp. 817–831, Apr. 2024.

[12] H. Gao, J. Huang, Y. Tao, W. Hussain, and Y. Huang, "The joint method of triple attention and novel loss function for entity relation extraction in small data-driven computational social systems," *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 6, pp. 1725–1735, Dec. 2022.

[13] V. Papyan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proc. Nat. Acad. Sci.*, vol. 117, no. 40, pp. 24652–24663, Oct. 2020.

[14] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5532–5540.

[15] X. Lin, C. Ding, Y. Zhan, Z. Li, and D. Tao, "HL-Net: Heterophily learning network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., New Orleans, LA, USA, Jun. 2022, pp. 19476–19485.

[16] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Salt Lake City, UT, USA, Jun. 2018, pp. 5831–5840.

[17] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Long Beach, CA, USA, Jun. 2019, pp. 6619–6628.

[18] X. Zhou, W. Liang, I. Kevin, K. Wang, and S. Shimizu, "Multi-modality behavioral influence analysis for personalized recommendations in health social media environment," *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 5, pp. 888–897, Oct. 2019.

[19] X. Lin, C. Ding, J. Zhang, Y. Zhan, and D. Tao, "RU-Net: Regularized unrolling network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., New Orleans, LA, USA, Jun. 2022, pp. 19457–19466.

[20] L. Li, G. Chen, J. Xiao, Y. Yang, C. Wang, and L. Chen, "Compositional feature augmentation for unbiased scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 21685–21695.

[21] T. Jin et al., "Fast contextual scene graph generation with unbiased context augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Vancouver, BC, Canada, Jun. 2023, pp. 6302–6311.

[22] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR),* Addis Ababa, Ethiopia, Apr. 2020.

[23] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9268–9277.

[24] J. Yu, Y. Chai, Y. Wang, Y. Hu, and Q. Wu, "CogTree: Cognition tree loss for unbiased scene graph generation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI), Virtual Event*, Montreal, Canada, Aug. 2021, pp. 1274–1280.

[25] G. Sudhakaran, D. S. Dhami, K. Kersting, and S. Roth, "Vision relation transformer for unbiased scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 21882–21893.

[26] X. Zhou et al., "Reconstructed graph neural network with knowledge distillation for lightweight anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543.

[29] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[30] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Long Beach, CA, USA, Jun. 2019, pp. 6700–6709.

[31] J. Jeon, K. Kim, K. Yoon, and C. Park, "Semantic diversity-aware prototype-based learning for unbiased scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Milan, Italy, Sep. 2024, pp. 379–395.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1492–1500.

[35] Y. Guo et al., "From general to specific: Informative scene graph generation via balance adjustment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Virtual Event*, Oct. 2021, pp. 16383–16392.

[36] X. Dong, T. Gan, X. Song, J. Wu, Y. Cheng, and L. Nie, "Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., New Orleans, LA, USA, Jun. 2022, pp. 19427–19436.

[37] L. Li, L. Chen, Y. Huang, Z. Zhang, S. Zhang, and J. Xiao, "The devil is in the labels: Noisy label correction for robust scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., New Orleans, LA, USA, Jun. 2022, pp. 18869–18878.

[38] A. Zhang et al., "Fine-grained scene graph generation with data transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 409–424.

[39] K. Yoon, K. Kim, J. Moon, and C. Park, "Unbiased heterogeneous scene graph generation with relation-aware message passing neural network," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, vol. 37, no. 3, pp. 3285–3294, Feb. 2023.

[40] Z. Chen et al., "Dark knowledge balance learning for unbiased scene graph generation," in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, ON, Canada, Oct. 2023, pp. 4838–4847.
[41] B. A. Biswas and Q. Ji, "Probabilistic debiasing of scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 10429–10438.
[42] D. Jung, S. Kim, W. H. Kim, and M. Cho, "Devil's on the edges: Selective quad attention for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 18664–18674.
[43] Y. Min, A. Wu, and C. Deng, "Environment-invariant curriculum relation learning for fine-grained scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 13296–13307.
[44] C. Zhang, S. Stepputtis, J. Campbell, K. Sycara, and Y. Xie, "Hiker-SGG: Hierarchical knowledge enhanced robust scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 28233–28243.
[45] Y. Li, T. Wang, K. Wu, L. Wang, X. Guo, and W. Wang, "Fine-grained scene graph generation via sample-level bias prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Milan, Italy, Sep. 2024, pp. 18–35.
[46] H. Kim, S. Kim, D. Ahn, J. T. Lee, and B. C. Ko, "Scene graph generation strategy with co-occurrence knowledge and learnable term frequency," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jul. 2024, pp. 24094–24109.

Zhaodi Wang received the M.S. degree in aerospace engineering from Beijing Institute of Technology, Beijing, China, in 2021. She is currently working toward the Ph.D. degree in computer science and engineering with the School of Computer Science and Engineering, Beihang University, Beijing.

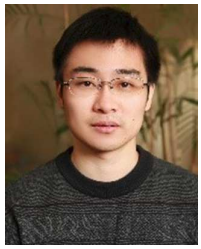
Her research interests include visual understanding and knowledge graph.

Yangyan Zeng received the B.S. degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 2004, the M.S. degree in software engineering from Hunan University, Changsha, in 2010, and the Ph.D. degree in management science and engineering from Central South University, Changsha, in 2023.

She is currently an Associate Professor with Xiangjiang Laboratory, Changsha, China. Her research interests include artificial intelligence and smart healthcare.

Biao Leng received the B.S. degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2009.

He visited Pennsylvania State University, University Park, PA, USA, for the period 2011–2012, and The Hong Kong University of Science and Technology, Hong Kong, in 2012. He is currently a Professor with the School of Computer Science and Engineering, Beihang University, Beihang, China. His research interests include artificial intelligent and industrial internet.



Xiaokang Zhou (Member, IEEE) received the Ph.D. degree in human sciences from Waseda University, Shinjuku, Japan, in 2014.

He is currently an Associate Professor with the Faculty of Business Data Science, Kansai University, Suita, Japan. From 2012 to 2015, he was a Research Associate with the Faculty of Human Sciences, Waseda University. From 2016 to 2024, he was a Lecturer/Associate Professor with the Faculty of Data Science, Shiga University, Shiga, Japan. Since 2017, he also works as a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Japan. He has been engaged in interdisciplinary research works in the fields of computer science and engineering, data science, information systems, and social and human informatics. His research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber-physical-social systems, and cyber intelligence and security.

Dr. Zhou is a member of the *IEEE Computer Society*, and *Association for Computing Machinery*, United States of America, *International Conferences-Information Processing Society*, and *Japanese Society for Artificial Intelligence*, Japan, and China Computer Federation, China.