Adaptive Fine-Grained Predicates Learning for Scene Graph Generation

Xinyu Lyu, Member, IEEE, Lianli Gao, Member, IEEE, Pengpeng Zeng, Heng Tao Shen, Fellow, IEEE, and Jingkuan Song, Senior Member, IEEE

Abstract—The performance of current Scene Graph Generation (SGG) models is severely hampered by hard-to-distinguish predicates, e.g., “woman-on/standing on/walking on-beach”. As general SGG models tend to predict head predicates and re-balancing strategies prefer tail categories, none of them can appropriately handle hard-to-distinguish predicates. To tackle this issue, inspired by fine-grained image classification, which focuses on differentiating hard-to-distinguish objects, we propose an Adaptive Fine-Grained Predicates Learning (FGPL-A) which aims at differentiating hard-to-distinguish predicates for SGG. First, we introduce an Adaptive Predicate Lattice (PL-A) to figure out hard-to-distinguish predicates, which adaptively explores predicate correlations in keeping with model’s dynamic learning pace. Practically, PL-A is initialized from SGG dataset, and gets refined by exploring model’s predictions of current mini-batch. Utilizing PL-A, we propose an Adaptive Category Discriminating Loss (CDL-A) and an Adaptive Entity Discriminating Loss (EDL-A), which progressively regularize model’s discriminating process with fine-grained supervision concerning model’s dynamic learning status, ensuring balanced and efficient learning process. Extensive experimental results show that our proposed model-agnostic strategy significantly boosts performance of benchmark models on VG-SGG and GQA-SGG datasets by up to 175% and 76% on Mean Recall@100, achieving new state-of-the-art performance. Moreover, experiments on Sentence-to-Graph Retrieval and Image Captioning tasks further demonstrate practicability of our method.

Index Terms—Scene graph generation, visual relationship, fine-grained learning, adaptive learning.

I. INTRODUCTION

Scene graph generation plays a vital role in visual understanding, which intends to detect instances together with their relationships. By ultimately representing image contents in a graph structure, scene graph generation serves as a powerful means to bridge the gap between visual scenes and human languages, benefiting several visual-understanding tasks, such as image retrieval [1], image captioning [2], [3], and visual question answering [4], [5], [6], [7].

Prior works [4], [10], [11], [12], [13], [14], [15] have devoted great efforts to exploring representation learning for scene graph generation, but the biased prediction issue is still challenging because of the long-tailed distribution of predicates in SGG datasets. Trained with severely skewed class distributions, general SGG models are prone to predict head predicates, as results of Transformer [8], [9] shown in Fig. 1(c). Recent works [16], [17], [18], [19] have exploited re-balancing methods to solve the biased prediction problem for scene graph generation, making predicates distribution balanced or the learning process smooth. As demonstrated in Fig. 1(b), Transformer (Re-weight) achieves a more balanced performance than Transformer. However, re-balancing on the class distribution, existing re-balancing strategies prefer predicates from tail categories while being hampered by some hard-to-distinguish predicates. In fine-grained image classification task, hard-to-distinguish object classes include

Fig. 1. Illustration of handling hard-to-distinguish predicates for SGG models. (a) Transformer (Re-weight) prefers tail categories. (b) Transformer (FGPL-A) outperforms both Transformer and Transformer (Re-weight) on Group Mean Recall with a balanced discrimination among predicates with different frequencies. (c) Transformer [8], [9] is prone to predict head predicates. (d) Transformer (Re-weight) prefers tail categories. (e) Transformer (FGPL-A) can appropriately handle hard-to-distinguish predicates, e.g., “woman-on/standing on/walking on-beach” or “woman-near/looking at/in front of child”.

This work was supported in part by National Key R&D Program of China under Grant 2022YFC2009903/2022YFC2009900, in part by the National Natural Science Foundation of China under Grants 62120218, 62020106008, 61772116, and 61872064, in part by Fok Ying-Tong Education Foundation under Grant 171106, and in part by SongShan Laboratory under Grant YYJCD2022019. Recommended for acceptance by E. Kalogerakis. (Corresponding author: Lianli Gao.)

The authors are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China (e-mail: xinyuluyu68@gmail.com; lianli.gao@uestc.edu.cn; is.pengpengzeng@gmail.com; shenhengtao@hotmail.com; jingkuan.song@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3298356, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3298356
species of birds, flowers, or animals; and the makes or models of vehicles. Similarly, typical hard-to-distinguish predicates in SGG include types of human-object-interactions (e.g., “on”, “standing on” or “walking on”), spatial (e.g., “in front of”, “behind” or “near”) or possessive relationships (e.g., “part of”, “attached to” or “covering”). And they cannot be well addressed by either general SGG models or re-balancing methods. For instance, as shown in Fig. 1(d), Transformer (Re-weight) misclassifies “woman-in front of-beach” as “woman-looking at-child” in terms of visual correlations between “in front of” and “watching” in this scenario.

The origin of the issue lies in the fact that differentiating among hard-to-distinguish predicates requires exploring their correlations first. As an inherent characteristic of predicates, the predicate correlations reflect the difficulty of distinction between predicate pairs, i.e., if two predicates are hard-to-distinguish or not for SGG models. Different from fine-grained image classification task where hard-to-distinguish object classes correlation are predefined, the hard-to-distinguish predicates correlation in SGG task is unknown. Hence, without exploration of predicate correlations, the existing re-balancing SGG methods [18], [20], [21], [22] cannot adaptively adjust discriminating process in accordance with the difficulty of distinction, resulting in an inefficient learning process. Particularly, roughly adjusting discriminating process based on the distribution prior, the dataset-based weights used in re-balancing SGG methods (i.e., re-sampling [17], [18], [21], re-weighting [23], [24], [25], logit adjustment [22], [26], [27]) may over-emphasize the tail predicates while over-suppressing the head classes. Consequently, the over-adjustment of optimization tends to make model’s learning process less effective or even misleading. Ultimately, as claimed in [28], [29], the inefficient learning process may cause over-confidence on tail predicates with under-represented head classes, which degrades model’s discriminatory power among hard-to-distinguish predicates and thereby deteriorates SGG model’s performance on generating fine-grained scene graphs. For example, in Fig. 1(d), “in front of” is misclassified to “flying in” simply because “flying in” is a tail predicate with less observations in SGG datasets.

To acquire comprehensive predicate correlations, we consider the contextual information for predicate pairs, since correlations between a pair of predicates may dramatically vary with contexts as stated in [30]. Particularly, contexts are regarded as visual or semantic information of predicates’ objects and subjects in scene graph generation. Take predicate correlations analysis between “watching” and “playing” as an example. “Watching/playing” is weakly correlated or distinguishable in Fig. 2(b), while they are strongly correlated or hard-to-distinguish for SGG models in Fig. 2(a). Except for the contextual information, the predicate correlations may gradually change during model’s learning process. Revealing whether the predicates are hard-to-distinguish or not, the predicate correlations reflects model’s learning status over predicate classes, which varies with its dynamic learning pace in training iterations. For instance, the predicates pair “on” and “standing on” is hard-to-distinguish for SGG models at first, and then gradually becomes recognizable during training. Hence, the predicate correlations need progressive refinement in keeping with model’s dynamic learning status for the comprehensive understanding.

Inspired by the above observations, we propose an Adaptive Fine-Grained Predicates Learning (FGPL-A) framework for SGG, which dynamically figures out and discriminates among hard-to-distinguish predicates in keeping with model’s dynamic learning pace. We first introduce an Adaptive Predicate Lattice (PL-A) to help SGG models understand ubiquitous predicate correlations with fine-grained supervisions. Utilizing the predicate correlations of PL-A, we devise an Adaptive Discriminating Loss (EDL-A), which both differentiate among hard-to-distinguish predicates while maintaining learned discriminatory power over recognizable ones. In particular, CDL-A progressively regularizes model’s learning procedure concerning category-wise predicate correlations. However, the category-wise regularization adopted in CDL-A can only fit a portion of samples/entities since predicate correlations vary with contexts in different entities. Hence, EDL-A is proposed to adaptively adjust model’s discriminating process considering entity-wise predicate correlations with fine-grained supervisions. Utilizing CDL-A and EDL-A, our method can determine whether predicate pairs are hard-to-distinguish or not during training, which guarantees a more balanced and efficient learning process than previous methods [16], [17], [23], [18]. Finally, as the current evaluation metrics cannot comprehensively reflect model’s capability of discriminating among hard-to-distinguish predicates, we introduce the Discriminatory Power (DP) to evaluate SGG models’ discriminatory ability among hard-to-distinguish predicates.

This article reinforces the preliminary version of our work [31] with an effective optimization mechanism and insight.
analysis of each key component. The extensive contributions within this article are fourfold. First, we enrich the related works with additional details on more literature from three relevant research areas, i.e., scene graph generation, long-tailed distribution classification and fine-grained image classification. Second, we propose an adaptive version of FGPL (i.e., FGPL-A) which can better differentiate among hard-to-distinguish predicates based on more comprehensive understanding of predicate correlations. In practice, we design a Batch-Refinement (BR) regime which iteratively refines predicate correlations of PL-A in keeping with model’s dynamic learning status, providing faithful guidance for predicates discrimination. Third, by exploring the refined predicate correlations, the CDL-A progressively optimizes model’s learning process with fine-grained supervision, while the EDL-A adaptively selects hard-to-distinguish predicates for sufficient regularization concerning model’s current learning status. Taking advantage of both CDL-A and EDL-A, our proposed FGPL-A ensures a more balanced and efficient learning procedure compared to its original version, i.e., FGPL. Finally, we present a more comprehensive experimental evaluation of FGPL and FGPL-A on two SGG datasets (i.e., VG-SGG and GQA-SGG) and validate the practicability of our methods on Sentence-to-Graph Retrieval and Image Captioning tasks.

Our main contributions are summarized as follows:

- Beyond the long-tailed problem in SGG tasks, we are the first to propose the fine-grained predicates problem, a new perspective of cases that hamper current SGG models. We further propose a novel plug-and-play Adaptive Fine-Grained Predicates Learning (FGPL-A) framework to well address this problem.
- We devise an Adaptive Predicate Lattice (PL-A) to adaptively explore pair-wise predicate correlations based on model’s ongoing predictions of the current mini-batch, which provides faithful guidance for model’s discriminating process. The Adaptive Category Discriminating Loss (CDL-A) helps SGG models learn to figure out and differentiate among hard-to-distinguish predicates with fine-grained supervision. Moreover, the Adaptive Entity Discriminating Loss (EDL-A) adaptively adjusts the discriminating process under predictions of entities, ensuring a balanced and efficient learning procedure.
- Extensive experimental results demonstrate that our FGPL-A dramatically boosts the performance of benchmark models on the VG-SGG and GQA-SGG datasets (e.g., Transformer, VCTree, and Motifs improved by 142.3%, 175.2%, 157.6% of Mean Recall (mR@100) under the Predicate Classification sub-task on VG-SGG dataset), reaching a new state-of-the-art performance. Furthermore, we conduct experiments on Sentence-to-Graph Retrieval and Image Captioning tasks, which demonstrate the practicability of our FGPL-A.
- Finally, we introduce the Discriminatory Power (DP) and a new metric (DP@K) to reflect SGG models’ discriminatory ability against hard-to-distinguish predicates for scene graph generation. Our codes are available at https://github.com/XinyuLyu/FGPL.

II. RELATED WORK

A. Scene Graph Generation

Suffering from the biased prediction, today’s SGG task is far from practical. Prior works [32], [33], [34] independently detect entities and relationships, ignoring the rich visual context information. To deal with the problem, some methods [4], [10], [11], [12], [35], [36], [37], [38], [39], [40] are proposed to refine the relation representation by exploring the contextual information via message passing mechanism. Motifs [11] acquires the rich context representation among objects and relationships by encoding their features utilizing Bi-LSTMs. For VCTree [4], it adaptively adjusts the structure of entities within the images to obtain the hierarchical information with a dynamic tree structure. Taking advantage of self-attention mechanism, Transformer [8] explores the interactions among instances and their relationships, which contributes to refining the insufficient context information within each entity. Moreover, [18] proposes a bipartite graph neural network to capture the dependency between entities and predicates. Other methods [17], [18], [23], [41], [42] are proposed to balance the discriminating process in accordance with class distribution or visual clues. [18] proposes a bi-level re-sampling scheme to achieve a balanced data distribution for training. Besides, TDE [41] attempts to disentangle predicates’ unbiased representation from biased predictions. Concerning the consistency in scene graphs, [14] makes SGG models aware of the structural information in output space to constrain the predicates prediction. [23] regularizes model’s learning process with the global predicate correlation, defined as the independence degree among other predicates. However, the independence degree only provides a coarse-grained understanding of predicate correlations which cannot achieve fine-grained relation recognition in SGG. Additionally, [16] explores predicate correlations with a hierarchical cognitive tree, in which model’s learning process is regularized in a coarse-to-fine manner. However, the predicate correlation dynamically varies with contexts, which can not be formed as such hierarchical structure. Different from them, we focus on discriminating among hard-to-distinguish predicates with pair-wise predicate correlation which is constructed as a predicate graph.

B. Long-Tailed Distribution Classification

The long-tailed distribution reveals the nature of the real-world that a small number of classes dominates the majority of our observations. Recently, the long-tailed problems received growing attentions. To solve the long-tailed problem, various distribution-based re-balancing learning strategies [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53] have been proposed. With respect to the imbalanced distribution, [51] has been proposed to re-balance the optimizing process according to the class-wise frequency. Besides, [53] introduces the LDAM to deal with the class-imbalanced problem by minimizing a margin-based generalization bound. Similarity, [52] introduces the Balanced Meta-Softmax to address label distribution differences between training and testing. Considering the over-suppression problem, [45], [49] reduces overwhelming
punishment from head classes. As a re-weighting method, Focal Loss [54] is proposed to adaptively adjust the loss weights to model’s learning quality. To keep a balanced learning process, [49] regularizes the learning process in reference to both model’s prediction results and class distribution, which copes with the over-confidence problem on tail classes. Similarly, [43] adaptively adjusts the training process for different classes on the basis of probabilities concerning the discriminatory difficulty between different classes. Furthermore, [44] adapts the training procedure to model’s learning status, which ensures a robust optimization process. However, due to the considerable semantic overlap among predicates, the predicate correlations are crucial for differentiating among hard-to-distinguish predicates in SGG. Without thoroughly exploring predicate correlations, the re-balancing methods (i.e., re-sampling [17], [18], [21], re-weighting [23], [24], [25], logit adjustment [22], [26], [27]) would roughly adjust the discriminating process based on the distribution prior obtained from statistics of SGG datasets. As a result, they may over-emphasize the tail predicates and over-suppress the head classes, which degrade model’s discriminatory power among hard-to-distinguish predicates and thereby deteriorate SGG model’s performance on generating fine-grained scene graphs. Therefore, in this work, we take advantage of both predicate distributions and predicate correlations to handle this issue.

C. Fine-Grained Image Classification

Fine-Grained Image Classification aims to recognize hard-to-distinguish objects in a coarse-to-fine manner. Existing methods tackle the problem roughly from two perspectives, localization-recognition [55], [56], [57] and representation-encoding [58], [59], [60]. Specifically, the former [55], [56], [57], [61], [62] localize recognizable parts and take advantage of higher-order structural relationships for fine-grained image classification. Moreover, the latter [58], [59], [60], [63] differentiate among hard-to-distinguish objects by enriching representation with more discriminate features. In [62], the self-attention module is extended to the global-local cross-attention mechanism, which enables the model to learn subtle feature embeddings for recognizing fine-grained object categories. Additionally, [63] proposes the MC-Loss, a loss function that prompts the model to identify and concentrate on subtle details of discriminative regions in terms of channel-wise feature maps for fine-grained image classification. However, since there is no constant hierarchical semantic structure among predicate categories, the Fine-Grained Image Classification methods are prone to fail in fine-grained predicate classification. Specifically, the hard-to-distinguish objects are predefined and related to each other in a constant hierarchical structure (i.e., coarse-to-fine) in fine-grained image classification. For example, “African hunting dog”, “Afghan bound”, and “Bedlington terrier” are always sub-breeds of “Dog”. However, since the predicate correlation dynamically varies with contexts, there isn’t such constant hierarchical structure among predicates in scene graph generation as shown in Fig. 2. Thus, the context-varied predicate correlations make existing fine-grained classification methods (i.e., adopting the coarse-to-fine recognition manner) fail to differentiate hard-to-distinguish predicates in fine-grained relation recognition. Instead of a hierarchical structure in a coarse-to-fine manner, we claim that predicate correlation should be formed as a graph structure. Concretely, we construct a Predicate Lattice to comprehend predicate correlations for predicate discriminating.

III. ADAPTIVE FINE-GRAINED PREDICATES LEARNING

A. Problem Formulation and Overview

Problem Formulation: Scene graph generation is typically a two-stage multi-class classification task. In the first stage, Faster R-CNN [64], [65] detects instance labels $T = \{t_s, t_o\}$, bounding boxes $B = \{b_s, b_o\}$, and feature maps $X = \{x_s, x_o\}$ within an input image $I$. In the second stage, scene graph models infer predicates from subject $s$ to object $o$, i.e., $R = \{r_{s \to o}\}$, based on detection results, i.e., $Pr_t(R|O, B, X)$.

Overview of FGPL framework: Within our Fine-Grained Predicates Learning framework (FGPL) [31], we first construct a Predicate Lattice (PL) concerning context information to understand ubiquitous correlations among predicates. Then, utilizing the PL, we develop a Category Discriminating Loss (CDL) and an Entity Discriminating Loss (EDL) which help SGG models differentiate among hard-to-distinguish predicates.

Limitations of FGPL: Obtained from models’ biased predictions of SGG dataset before training, the predicate correlations of PL are pre-determined and static during training. Regularized by the PL-based CDL and EDL, FGPL fails to cope with model’s dynamic learning pace, rendering over-confidence [28], [29] on predicate discrimination.

Overview of FGPL-A Framework: Based on the above observations, we further devise an adaptive version for FGPL as Adaptive Fine-Grained Predicate Learning (FGPL-A) by introducing an Adaptive Predicate Lattice (PL-A), which progressively updates predicate correlations in keeping with model’s dynamic learning status. Practically, we propose a Batch-Refinement (BR) regime for the PL-A, which iteratively refines pair-wise predicate correlations via exploring model’s ongoing predictions of the current mini-batch in training iterations. Then, by exploring the refined predicate correlations, we propose the adaptive version of CDL and EDL as Adaptive Entity Discriminating Loss (EDL-A) and Adaptive Entity Discriminating Loss (CDL-A) for the adaptive discrimination among hard-to-distinguish predicates, resulting in a balanced and efficient learning process. The overview of FGPL-A framework is shown in Fig. 3.

B. Adaptive Predicate Lattice Construction

In this section, we describe the construction process of Adaptive Predicate Lattice (PL-A), which provides comprehensive understanding on predicate correlations, concerning contexts and model’s dynamic learning status. First, we build the Predicate Lattice (PL) by exploring model’s biased predictions of SGG dataset in Section III-B1. Since the PL fails to cope with model’s dynamic learning status, we further devise an Adaptive Predicate Lattice (PL-A) by introducing a Batch-Refinement (BR) regime in Section III-B2.
Fig. 3. Overview of the Adaptive Fine-Grained Predicates Learning (FGPL-A) framework. It includes three parts: Adaptive Predicate Lattice (PL-A), Adaptive Category Discriminating Loss (CDL-A), and Adaptive Entity Discriminating Loss (EDL-A). FGPL-A can be incorporated into several benchmark SGG models. The Adaptive Predicate Lattice (PL-A) is initialized as the PL, and then gets iteratively refines by the Batch-Refinement (BR) based on model’s predictions of the current mini-batch. By exploring the refined predicate correlations of PL-A, CDL-A, and EDL-A adaptively optimize SGG models, ensuring a balanced and efficient learning process.

1) Predicate Lattice: To gain a comprehensive understanding of correlations between predicates, we construct a Predicate Lattice (PL) (i.e., $S' = \{s'_{r_i \rightarrow o, t_s, t_o}, r_j \rightarrow o, t_s, t_o\} \in \mathbb{R}^{C \times E \times E \times C}$), that incorporates correlations between pairs of predicates based on contextual information. In general, predicate correlations are established by considering various contexts (i.e., visual or semantic information of predicates’ subjects and objects), for contexts determine correlations among predicates. Specifically, we extract predicates’ contextual-based correlations by analyzing biased predictions encompassing all potential contexts between each pair of predicates. The construction procedure, illustrated in Fig. 4, can be divided into three steps:

Context-Predicate Association: The primary objective of constructing the Predicate Lattice (PL) is to derive predicate correlations. As contexts determine correlations among predicates, the PL consists of two types of nodes: Predicate nodes and Context nodes. Predicate nodes represent predicate categories, while Context nodes represent labels of subject-object pairs. An illustration of the Predicate Lattice structure can be seen in Fig. 4(a). Within the PL, multiple predicate nodes can be connected to the same context node, indicating that several predicates can describe relationships within the same context. For instance, in Fig. 4(a), both “holding” and “carrying” can be used to describe relationships for “person-racket”. Hence, to establish the Context-Predicate Associations between predicate nodes and context nodes, we begin by deriving the contexts for each predicate. Specifically, predicates’ contexts are derived from the Frequency model [11], which includes the occurrence frequency for each “subject-predicate-object” triplet (i.e., $F = \{f_{r_i \rightarrow o, t_s, t_o}\} \in \mathbb{R}^{C \times E \times E}$) in the dataset. Furthermore, predicate nodes and context nodes are associated with edges weighted as $f_{r_i \rightarrow o, t_s, t_o}$. In this way, we establish connections between predicate nodes and context nodes in Predicate Lattice.

Biased Predicate Prediction: To associate predicate pairs with predicate correlations in the next step, we acquire Biased Predicate Prediction, i.e., $Q = \{q_{r_i \rightarrow o, t_s, t_o}\} \in \mathbb{R}^{C \times E \times E}$, from SGG models. First, we incorporate Context-Predicate Association, constructed in the previous step, into SGG models. This involves extracting the Context-Predicate Association as semantic information from each “subject-predicate-object”
Predicate Lattice Construction.

**Input:** Training set $D = \{d^h\}_{h=1}^H$, pretrained biased SGG model $SGG$, number of object class $E$, number of predicate class $C$, biased prediction matrix $Q = \{q_{r_s-o,t_o}\} \in \mathbb{R}^{C \times E \times E}$, frequency-bias matrix $F = \{f_{r_s-o,t_o}\} \in \mathbb{R}^{C \times E \times E}$.

**Output:** Predicate Lattice $S' = \{s'_{i,j}^r\} \in \mathbb{R}^{C \times E \times E \times C}$, predicate correlation matrix $S = \{s_{i,j}\} \in \mathbb{R}^{C \times C}$.

1: // Step 1: Context-Predicate Association. */
2: for $i = 0, j = 0$ to $C$ do
3: if $f_{r_s-o,t_o} = 0$ or $f_{r_s-o,t_o} = 0$ then
4: $s'_{i,j} = 0$
5: else
6: $s'_{i,j} = \frac{q_{r_s-o,t_o} \times q_{r_s-o,t_o}}{f_{r_s-o,t_o} \times f_{r_s-o,t_o}}$
7: end if
8: end for
9: // Step 2: Biased Predicate Prediction. */
10: for $d^h$ in training set $D$ do
11: $Pr(r_s-o\to t_o) \leftarrow SGG(d^h)$
12: $q_{r_s-o,t_o} \leftarrow q_{r_s-o,t_o} + Pr(r_s-o\to t_o) \leftarrow q_{r_s-o,t_o}$
end for
14: // Step 3: Predicate-Predicate Association. */
15: for $i = 0, j = 0$ to $C$ do
16: for $s = 0, o = 0$ to $E$ do
17: $s'_{i,j} = s'_{i,j} \leftarrow s'_{i,j} + s'_{i,j}$
18: end for
19: end for
20: // gather predicates correlations under all contexts. */
21: for $s = 0, o = 0$ to $E$ do
22: $s_{i,j} = s_{i,j} \leftarrow l_2\text{norm}(s'_{i,j})$
23: end for
24: // Abbreviate $s_{r_s-o,t_o}$ as $s_{i,j}$ in manuscripts. */

Algorithm 1: Predicate Lattice Construction.

As a result, the Biased Predicate Prediction encompasses predicate predictions for each predicate pair under all possible scenarios. For instance, as depicted in Fig. 4(b), we infer the pre-trained SGG model under all possible scenarios for predicates “playing” or “holding”, such as scenarios involving “person-racket” and “person-bag”.

**Predicate-Predicate Association:** Finally, we establish Predicate-Predicate Association among predicates with context-based correlations obtained from the Biased Predicate Prediction. The Biased Predicate Prediction provides insights into the context-based correlations between each pair of predicates. For instance, if the majority of samples are predicted as predicate $j$ but labeled as predicate $i$ in the ground truth, it indicates that predicate $i$ correlates to predicate $j$ in most contexts.

Based on the above observation, we accumulate prediction results from each possible context to obtain comprehensive predicate correlations between each pair of predicates, i.e., $s_{r_s-o,t_o} = q_{r_s-o,t_o} \times q_{r_s-o,t_o}$, as illustrated in Fig. 4(c). For instance, given the predicate pair “playing-holding”, we gather their correlations under all contexts/scenarios, such as “person-racket” and “person-bag”. Moreover, if predicate $i$ is consistently correlated with predicate $j$ across most contexts, it suggests a strong correlation between them. Hence, to quantitatively capture predicates’ correlations, we normalize the accumulated predicate correlations as $S = \{s_{i,j}\} \in \mathbb{R}^{C \times C}$ with $s_{i,j}$, where $s_{i,j}$ denotes the proportion of samples labeled as $i$ but predicted as $j$. In particular, higher $s_{i,j}$ indicates a stronger correlation between predicate pair $i$ and $j$. Finally, we associate predicate pairs with their corresponding predicate correlations $s_{i,j}$, forming a Predicate Lattice, as depicted in Fig. 4(d).

2) Adaptive Predicate Lattice With Batch-Refinement: Limitations of PL: Ignoring model’s dynamic learning status, the PL (developed in Section III-B1) fails to accurately monitor model’s learning quality among predicates throughout training. Specifically, derived from the biased predictions before training, the predicate correlations of PL are pre-determined and remain static after that. However, as model’s discriminatory power is progressively acquired, the static predicate correlations of PL cannot catch the variations of model’s learning status (i.e., if the predicates are hard-to-distinguish or recognizable) during the learning procedure. Hence, regularized by the static predicate correlations, model’s learning process tends to be inconsistent with its gradually gained discriminatory power, rendering the over-confident issue on predicate discrimination. Thus, explorations on the refinement mechanism of predicate correlations need to be further carried out for a balanced and efficient learning process.

**Formulation of PL-A:** To address the limitation mentioned above, we further extend the Predicate Lattice (PL) as Adaptive Predicate Lattice (PL-A). Practically, to make predicate correlations in line with model’s gradually obtained discriminatory power, we devise a Batch-Refinement (BR) regime, which iteratively refines predicate correlations of PL-A via exploring model’s ongoing predictions during training. For Batch-Refinement (BR) regime, it first comes to mind to iteratively refine predicate correlations by acquiring model’s biased predictions of the whole training set at each training step. However, obtaining evaluation results for the whole training set at each training step is infeasible, which brings a tremendous computational cost. Instead, we iteratively update the predicate correlations with the accumulated predictive scores for each mini-batch, since the batch-based predictions approximately reflect model’s learning status at each learning pace.

To achieve that, PL-A is first initialized with the predicate correlations of PL (i.e., $s_{i,j} = s_{i,j}$), since PL provides fundamental insights on comprehending the predicate correlations. Then, we propose the Batch-Refinement (BR) regime to refine the predicate correlations of PL-A $s_{i,j}$ with Refining Momentum
in the $t$-th iteration. Furthermore, the Refining Momentum $S^{t}_{i,j}$ is calculated by exploring model’s biased predictions within the current mini-batch. The corresponding Batch-Refinement (BR) regime can be written as follows:

$$s^{t}_{i,j} = \begin{cases} \tau s^{t-1}_{i,j} + (1 - \tau) S^{t-1}_{i,j}, & \text{if } t > 0, \\ s^{0}_{i,j}, & \text{if } t = 0 \end{cases}$$

$$S^{t}_{i,j} = \tilde{s}^{t}_{i,j} + \hat{s}^{t}_{i,j},$$

(1)

where $s^{t}_{i,j}$ and $s^{t-1}_{i,j}$ indicates the predicate correlations between class $i$ and $j$ at the $t$-th and the $t-1$ iteration. Moreover, $\tilde{s}^{t}_{i,j}$ is refined by iteratively adding the Refining Momentum $S^{t-1}_{i,j}$ to $s^{t-1}_{i,j}$ (i.e., $s^{t}_{i,j} = \tau s^{t-1}_{i,j} + (1 - \tau) S^{t-1}_{i,j}$). Intuitively, $s^{t-1}_{i,j}$ and $s^{t}_{i,j}$ reflects the historical and coming-batch statistics of predicate correlations. Besides, $\tau$ denotes the hyper-parameter for a trade-off between $s^{t-1}_{i,j}$ and $S^{t-1}_{i,j}$. Furthermore, the Refining Momentum $S^{t}_{i,j}$ is composed of both Entity Refining Momentum ($EM^{t}_{i,j}$) and Category Refining Momentum ($CM^{t}_{i,j}$) $\tilde{s}^{t}_{i,j}$, i.e., $S^{t}_{i,j} = \tilde{s}^{t}_{i,j} + \hat{s}^{t}_{i,j}$.

Particularly, CRM updates the predicate correlations by exploring model’s category-wise biased predictions, while ERM refines the predicate correlations concerning the entity-specific context within each sample. The illustrations for CRM and ERM are shown in Fig. 5.

The Category Refining Momentum (CRM) acquires category-wise predicate correlations by jointly exploring biased predictions from all samples in each mini-batch. Specifically, the CRM is formulated as follows:

$$\hat{\phi} = [\hat{\phi}^c_1, \hat{\phi}^c_2, \ldots, \hat{\phi}^c_C]$$

where $\hat{\phi}$ denotes the weighted logit predictions for each class. Practically, CRM first accumulates predictive scores of negative class $j$ and positive class $i$ from $M$ samples in the $t$-th mini-batch, as $\sum_{m=1}^{M} \hat{\phi}^{l,m}_j$ and $\sum_{m=1}^{M} \hat{\phi}^{l,m}_i$, respectively.

Then, the CRM obtains distributions of biased predictions between negative class $j$ and positive class $i$, as $\hat{s}^{t}_{i,j} = \frac{\sum_{m=1}^{M} \hat{\phi}^{l,m}_j}{\sum_{m=1}^{M} \hat{\phi}^{l,m}_i}$. Intuitively, the CRM $\hat{s}^{t}_{i,j}$ indicates the difficulty of distinction between the positive category $i$ and the negative category $j$ for SGG models at current stage. It is worth noting that the categories that do not appear in current mini-batch are skipped and not updated in practice.

Additionally, the Entity Refining Momentum (ERM) investigates the predicate correlations within each sample to deal with the intra-class variance caused by entity-specific context, which is computed as:

$$\hat{s}^{t,m}_{i,j} = \frac{\phi^{t,m}_j}{\phi^{t,m}_i},$$

(3)

First, ERM individually calculates the predictive scores of negative class $j$ and positive class $i$ within in sample $m$ at iteration $t$, i.e., $\hat{\phi}^{t,m}_j$ and $\hat{\phi}^{t,m}_i$. Then, it calculates the ratio between $\hat{\phi}^{t,m}_j$ and $\hat{\phi}^{t,m}_i$, as $\hat{s}^{t,m}_{i,j}$, which reveals how likely the given sample $m$ would be misclassified as a negative class $j$ at $t$-th iteration.

Refined in terms of the CRM (category-wise) and the ERM (entity-wise), the predicate correlations in the PL-A provide a faithful guidance (i.e., if predicate pairs are hard-to-distinguish or recognizable) for the Category Discriminating Loss (CDL) and the Entity Discriminating Loss (EDL), which are elaborated in the following sections.

C. Adaptive Category Discriminating Loss

To compensate for the limitations of the re-weighting method, we introduce our Category Discriminating Loss (CDL) in Section III-C1, which attempts to differentiate among hard-to-distinguish predicates while maintaining the discriminatory power on distinguishable ones. Since the CDL fails to cope with model’s dynamic learning status, we further devise an Adaptive Category Discriminating Loss (CDL-A) in Section III-C2, which contributes to achieving more substantial discriminatory power among predicates concerning model’s dynamic learning pace.

1) Category Discriminating Loss: Limitations of Re-weighting Method:

Overall, recent re-weighting methods re-balance the learning process by strengthening the penalty to head classes while scaling down the overwhelming punishment to tail classes. Specifically, the state-of-the-art re-weighting method [49] adjusts weights for each class in Cross-Entropy Loss on the basis of the proportion of training samples as follows:

$$L_{CE}(\eta) = - \sum_{i=1}^{C} y^i \log(\hat{\phi}^i),$$

$$\hat{\phi}^i = \frac{e^{\eta y^i}}{\sum_{j=1}^{C} w_{i,j} e^{\eta y^j}}, \quad w_{i,j} = \begin{cases} \left( \frac{n_j}{n_i} \right)^{\alpha}, & \text{if } n_j > n_i, \\ 1, & \text{if } n_j \leq n_i \end{cases},$$

(4)

where $\eta = [\eta_1, \eta_2, \ldots, \eta_C]$ and $\hat{\phi} = [\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_C]$ denote predicted logits and re-weighted probabilities for each class. The label $Y = [y_1, y_2, \ldots, y_C]$ is a one-hot vector. Additionally, $w_{i,j}$ denotes the re-weighting factor concerning distribution between...
positive class \(i\) and negative class \(j\). Explicitly, \(w_{i,j}\) is calculated based on the proportion of distribution between class \(i\) and \(j\), as shown in (4), where \(\alpha > 0\).

\[
\frac{\partial \mathcal{L}_{CE}(\eta)}{\partial \eta_{ij}} = \frac{w_{i,j} \eta_{ij}}{\sum_{k=1}^{C} w_{i,k} \eta_{ik}}. \tag{5}
\]

Equation (5) shows negative gradients for category \(j\). If positive category \(i\) is less frequent than negative category \(j\), i.e., \(n_j > n_i\) with \(w_{i,j} > 1\), it will strengthen the punishment to negative category \(j\). On the contrary, if \(n_j \leq n_i\) with \(w_{i,j} = 1\), it will degrade the penalty to negative class \(j\). Finally, it results in a balanced learning process.

Without considering predicate correlations, re-weighting methods cannot adaptively adjust discriminating process in accordance with difficulty of discrimination, resulting in an inefficient learning process. As an inherent characteristic of predicates, predicate correlations reflect difficulty of discrimination for different pairs of predicates. However, ignoring predicate correlations in learning process, the re-weighting method roughly reduces negative gradients for all negative predicates with fewer samples than the positive predicate. As a process to push away the decision boundary from tail classes to head classes, such discriminating process is prone to over-suppress weakly correlated predicate pairs and degrades the learned discriminatory capability of recognizable predicates as maintained in [43], [44]. Take an example among “on/has/standing on”, where “on/standing on” are strongly correlated and “has/standing on” are weakly correlated. To prevent the tail class “standing on” from being over-suppressed, the re-weighting method roughly degrades negative gradients from both “on” and “has”. Although it strengthens discriminatory power between “on” and “standing on”, it is prone to reduce that between “has” and “standing on” simultaneously.

**Formulation of CDL.** Based on the above observations, we should both consider the class distribution and predicate correlations to differentiate among hard-to-distinguish predicates. Thus, based on the re-weighting method in (4), we devise Category Discriminating Loss (CDL), which adjusts the re-weighting process according to predicate correlations obtained from Predicate Lattice. Overall, we utilize predicate correlations \(s_{i,j}\), defined in Section III-B1, as a signal to adjust the degree of re-weighting between predicates \(i\) and \(j\). Especially, we mitigate the magnitude of re-weighting for weakly correlated predicates while strengthening that for strongly correlated ones by setting \(w_{i,j}\), in (4), with different values. In this way, we maintain gained discriminatory power among recognizable predicates and further enhance that among hard-to-distinguish ones, shown as below:

\[
w_{i,j} = \begin{cases} 
\mu^\beta_{i,j} (> 1), & \text{if } \mu_{i,j} \geq 1 \text{ and } \varphi_{i,j} > \xi \\
1, & \text{if } \mu_{i,j} \geq 1 \text{ and } \varphi_{i,j} \leq \xi \\
1, & \text{if } \mu_{i,j} < 1 \text{ and } \varphi_{i,j} > \xi \\
\mu^\alpha_{i,j} (< 1), & \text{if } \mu_{i,j} < 1 \text{ and } \varphi_{i,j} \leq \xi 
\end{cases}, \tag{6}
\]

\[
\mu_{i,j} = \frac{n_j}{n_i} \varphi_{i,j} = \frac{s_{i,j}}{s_{i,i}}, \quad \varphi_{i,j} = \frac{s_{i,j}}{s_{i,i}}.
\]

where \(\varphi_{i,j}\) is calculated by the proportion between \(s_{i,j}\) and \(s_{i,i}\), revealing correlations between predicate \(i\) and \(j\). In addition, \(\alpha\) and \(\beta\) are hyper-parameters larger than 0. For instance, when \(n_j \geq n_i (\mu_{i,j} \geq 1)\), if \(\varphi_{i,j} > \xi\) of strongly correlated predicate pair \(i\) and \(j\), \(w_{i,j}\) is larger than 1 to strengthen the punishment on negative predicate \(j\). In contrast, if \(\varphi_{i,j} \leq \xi\) of weakly correlated predicate pair \(i\) and \(j\), \(w_{i,j}\) is set as 1 to mitigate the magnitude of penalty on negative predicate \(j\). That is because the excessive punishment is unnecessary for the weakly correlated predicate \(j\), which is easy to distinguish from predicate \(i\) for models. When \(n_j < n_i (\mu_{i,j} < 1)\), we set \(w_{i,j} \leq 1\) (including \(\varphi_{i,j} > \xi\) and \(\varphi_{i,j} \leq \xi\)) to relieve the over-suppression from head predicate \(i\) to tail one \(j\). Moreover, if \(\varphi_{i,j} \leq \xi\), we set \(w_{i,j} = \mu^\alpha_{i,j} (< 1)\) to mitigate the magnitude of the penalty on negative predicate \(j\).

2) **Adaptive Category Discriminating Loss: Limitations of CDL.** Founded on the Cross-Entropy Loss, the CDL proposed in Section III-C1 has two limitations. First, the hard targets (one-hot encoded) in CDL roughly regularize model’s learning process in a coarse-grained manner. The coarse-grained supervision inevitably introduces huge ambiguity into model’s learning procedure, which is inadequate for the fine-grained predicates learning. Thus, we argue that fine-grained supervisions ought to be employed for model’s fine-grained learning process. Second, ignoring model’s dynamic learning status, the predetermined and static label distribution has been found to incur an over-confident problem [28], [29]. Consequently, it may lead to degradation in model’s discriminatory power and dramatically limits its learning efficacy. Hence, we claim that different label distributions should be progressively assigned for each category in keeping with model’s learning pace over the training course.

**Formulation of CDL-A:** To address the limitations mentioned above, we adaptively set learning targets utilizing the fine-grained predicate correlations for SGG models. To achieve that, we introduce an Adaptive Label Softening (ALS) scheme. Practically, the ALS dynamically softens hard targets with predicate correlations of PL-A. Since predicate correlations get progressively refined during training, the ALS adaptively adjusts model’s learning objectives according to its dynamic learning status. Thus, it makes CDL-A consistent with model’s learning status and results in a balanced and effective discriminating process. Besides, in contrast to hard targets (one-hot encoded) in CDL, the ALS provides an insightful understanding of predicate correlations during training, which makes CDL-A progressively optimize model’s learning process with fine-grained supervisions. Under the guidance of fine-grained objectives, CDL-A enhances models’ capability of fine-grained relation recognition against hard-to-distinguish predicates.

The corresponding Adaptive Category Discriminating Loss (CDL-A) can be written as follows:

\[
\mathcal{L}_{CDA}(\eta) = \mathcal{L}_{CD}(\eta) + \theta \mathcal{L}_{ALS}(\eta),
\]

\[
\mathcal{L}_{ALS}(\eta) = - \sum_{i=1}^{C} s_{i,j} \log(\phi_{i,j}), \tag{7}
\]

where \(\mathcal{L}_{ALS}\) and \(\mathcal{L}_{CD}\) denotes the Adaptive Label Softening (ALS) regularization and the Category Discriminating Loss (CDL). It is worth noting that, since the CDL provides
fundamental effectiveness in differentiating among hard-to-distinguish predicates, we treat the ALS (L_{ALS}) as the regularization for CDL, with \( \theta \) as the trade-off coefficient. Additionally, \( s_{i,j}^{t} \) and \( \bar{\phi}_{i} \) denotes the refined predicate correlations at the \( t \)-th iteration mentioned in Section III-B2, and the re-weighted logits for class \( i \) in Section III-C1. Intuitively, the \( s_{i,j}^{t} \) of ALS can be regarded as the softened label for sample \( \eta \) at \( t \)-th iteration.

D. Adaptive Entity Discriminating Loss

In this section, we introduce our Entity Discriminating Loss (EDL) in Section III-D1, which adapts the discriminating process to different contexts within entities. Furthermore, to compensate for the constraints of EDL, we propose the Adaptive Entity Discriminating Loss (EDL-A) in Section III-D2, which adaptively adjusts the discriminating process to model’s ongoing predictions of the current mini-batch.

1) Entity Discriminating Loss: Limitations of CDL: Although CDL effectively differentiates hard-to-distinguish predicates, it still has a limitation: it only considers the category-wise difficulty of distinction between predicates but ignores entity-wise difficulty varied with contexts within different entities. As model’s predictive scores reflects discrimination difficulty within the specific context, we individually treat prediction results of each sample as signals to adjust the decision boundary.

Formulation of EDL: Based on the observations, we propose an Entity Discriminating Loss (EDL), which adapts the discriminating process to the contexts within entities, shown as below:

\[
L_{ED}(\eta) = \frac{1}{|V_i|} \sum_{j \in V_i} \max(0, \hat{\phi}_j - \hat{\phi}_i + \delta) \frac{n_j}{n_i},
\]

where \( V_i \) is defined as a set of strongly correlated predicates selected in reference to predicate correlations \( s_{i,j} \) in Predicate Lattice (PL). For each predicate category \( i \), \( M \) predicates with the highest \( s_{i,j} \) in the Predicate Lattice are chosen to construct \( V_i \). Given the input sample \( \eta \), \( \hat{\phi}_i \) and \( \hat{\phi}_j \) are the predicted probabilities for predicates \( i \) and \( j \). Intuitively, \( \hat{\phi}_j - \hat{\phi}_i \) implies the discrimination difficulty between predicates \( i \) and \( j \) of the specific context within sample \( \eta \). The \( \delta \) is a hyper-parameter, which denotes prediction margins for predicates. Furthermore, EDL is reduced to zero if predicate pairs are recognizable enough i.e., \( \hat{\phi}_i - \hat{\phi}_j \geq \delta \). Moreover, we also adopt the balancing factor \( \frac{n_j}{n_i} \) to alleviate imbalanced gradients between classes with fewer or more observations.

2) Adaptive Entity Discriminating Loss: Limitations of EDL: While the EDL proposed in Section III-D1 takes the intra-class variance of contexts into account, its effectiveness is still limited by the pre-determined predicate correlations within the Predicate Lattice (PL). Based on the static predicate correlations of PL, EDL only concentrates on differentiating a small set of strongly correlated predicates, which are pre-defined before training. Furthermore, it ignores the fact that predicate correlations may gradually change during model’s learning process. Hence, the episode-fixed regularization of EDL is prone to be inconsistent with model’s gradually obtained discriminatory power, resulting in the degradation on the model’s performance.

Formulation of EDL-A: To remedy such phenomenon, we further devise an adaptive version of EDL, namely the Adaptive Entity Discriminating Loss (EDL-A), which adaptively picks strongly correlated predicates in accordance with model’s gradually obtained discriminatory power, concerning inherent contexts within entities. As EDL-A adaptively chooses which category to suppress for each sample in training iterations, it adapts the discriminating process to both model’s learning status and contexts of entities, resulting in a robust and efficient learning process. The EDL-A is written as follows:

\[
L_{EDA}(\eta) = \frac{1}{|V_i|} \sum_{j \in V_i} \max(0, \hat{\phi}_j - \hat{\phi}_i + \delta) \frac{n_j}{n_i},
\]

where \( s_{i,j}^{t} \) and \( \bar{\phi}_{i} \) denote the refined predicate correlations at current stage \( t \) mentioned in Section III-B2 and the re-weighted logits for class \( i \) in Section III-C1. Moreover, \( V_i \) indicates a set of hard-to-distinguish (strongly correlated) predicates, which are dynamically chosen in reference to the refined predicate correlations \( s_{i,j}^{t} \) of the PL-A. In practice, \( V_i \) is built by figuring out the predicates \( j \) with the top-k strongest correlations \( s_{i,j}^{t} \) to predicate \( i \) at current stage \( t \).

E. Adaptive Discriminating Loss for FGPL-A

Including both CDL-A in (7) and EDL-A in (9), the Adaptive Discriminating Loss for FGPL-A can be expressed as below:

\[
L_{DA}(\eta) = L_{CDA}(\eta) + \gamma L_{EDA}(\eta),
\]

where \( \gamma \) balances the regularization between CDL-A and EDL-A. By exploring the progressively refined predicate correlation, the Adaptive Category/Entity Discriminating Loss (CDL-A/EDL-A) adaptively collaborates category-level and entity-level optimization, which guarantees an efficient learning procedure and enhances models discriminatory power among hard-to-distinguish predicates.

IV. EXPERIMENTS

A. Experiment Setting

Dataset: Following previous works [11], [42], we evaluate our methods on the Visual Genome (VG-SGG) dataset. It contains 108k images with 75k and 37k objects and predicates classes. Due to the severely skewed class distributions within the VG-SGG, following [11], [41], we adopt the widely used split for SGG. Under the setting, the VG-SGG dataset has 150 object categories and 50 relationship categories. Then, we further divide it into 70% training set, 30% testing set, and 5 images, selected from the training set, for validation. To further validate the generalizability of our methods under different distributions, we also conduct experiments on the other two challenging SGG datasets, i.e., GQA-SGG and OI-SGG. In contrast to VG-SGG, GQA-SGG has more object categories (1704 classes) and predicate categories (311 classes), which contains more complex scenario information. Similarly, for GQA-SGG, we adopt a 70-30 split for training (75k) and testing set (10k),
and further sample a small set (5k) for validation. In addition, for OI-SGG (OpenImages V6), we follow previous works [18], [37] and divide the dataset into 126k, 2k and 5k images for training, validation and testing, containing 301 object categories and 31 predicate classes. For Sentence-to-Graph Retrieval task, we follow [41] to sample the overlapped 41k images between VG and MS-COCO [66]. Then, they are divided into train (36k), test (1k), and validation (5k) sets. For the image captioning task, we follow the split of [67], which contains 113k images in training set, 5k images for testing, and 5k for validation, collected from MS-COCO [66].

**Evaluation Tasks:** Following recent works [4], [18], we evaluate our methods on three sub-tasks of SGG, including PredCls (Predicate Classification), SGCls (Scene Graph Classification), and SGDet (Scene Graph Detection). In the PredCls, model needs to predict predicates given ground-truth object bounding boxes together with their labels. For the SGCls task, taking ground truth bounding boxes as input, the model predicts both object labels and predicates (relationships) between them. The SGDet require the SGG model to generate the object labels with relationships from scratch, i.e., without ground-truth bounding boxes and object labels. Moreover, following [41], we conduct Sentence-to-Graph retrieval as a downstream task to verify the effectiveness of fine-grained scene graphs generated by our Adaptive Fine-Grained Predicate Learning (FGPL-A), reporting Recall@K on the 1k/5k gallery of image captions. We also perform Image Captioning task to prove that the generated scene graphs precisely describes scenarios. Specifically, object visual representations and scene graphs are combined to generate image captions. Moreover, we evaluate the generated image captions with Bleu-4 [68], Meteor [69], Cider [70], and Spice [71].

**Evaluation Metrics:** Following recent works [4], [18], we evaluate model’s performance on R@K and mR@K. However, different trade-offs between R@K and mR@K are made in different methods, which makes it hard to make a direct comparison. Therefore, following [72], to jointly evaluate R@K and mR@K, we further evaluate them with an overall metric F@K, which is the harmonic average of R@K and mR@K. Furthermore, we also evaluate them on DP@K and Group Mean Recall metrics to comprehensively validate their capability of fine-grained relation recognition. For Group Mean Recall, 50 categories of predicates are sorted and divided into Head Group (17), Body Group (17) and Tail Group (16) according to their frequency in VG-SGG dataset. Besides, we introduce DP@K (%) to indicate models’ Discriminatory Power among top-k hard-to-distinguish predicates. Generally, DP@K is calculated by averaging the difference between the proportion of samples correctly predicted as i and the proportion of samples misclassified as hard-to-distinguish predicates j (j ∈ V′ i). Furthermore, V′ i is defined as a set of top-k hard-to-distinguish predicates for predicate i. Especially, to figure out hard-to-distinguish predicates, we collect a normalized confusion matrix S′ ∈ [R^C×C] from the model’s prediction results, with s′ i,j, which denotes the degree of confusion between the predicate pair i and j. For each predicate category i, k predicates with the highest s′ i,j are chosen to construct V′ i. In a word, a higher score of DP@K means more substantial discriminatory power among hard-to-distinguish predicates. Moreover, the intuitive explanation of DP@K with Peruse-code is shown in Algorithm 2. For OI-SGG dataset, following previous works [18], [37], we employ the Recall@K, weighted mean AP of relationships (wmAP_{rel}) and weighted mean AP of phrase (wmAP_{phr}) for evaluation. Furthermore, to uniformly reflect model’s overall performance, we report the score_{wtd}, which is calculated as score_{wtd} = 0.2 × R@50 + 0.4 × wmAP_{rel} + 0.4 × wmAP_{phr}.

**B. Implementation Details**

**Detector:** For object detectors, we utilize the pre-trained Faster R-CNN by [41], [64] to detect objects in images. Moreover, object detectors’ weights are frozen during scene graph generation training for all three sub-tasks.

**Scene Graph Generation Model:** The benchmark models in experiments are all implemented in [9]. For Motifs, it divides the procedure of scene graph generation into two stages, i.e., the object context encoding and the relation context encoding, which are implemented as Bi-LSTMs for capturing global contextual information in images. For Transformer, it is implemented by replacing the Bi-LSTMs in Motifs’ object context encoder and relation context encoder with self-attention-based Transformer encoders. For VCTree, it is implemented by constructing a dynamic tree structure which explores the hierarchical contextual information among instances within the image. Next, following [9], these benchmark models in Model Zoo are all trained with Cross-Entropy Loss and SGD optimizer with an initial learning rate of 0.01, batch-size M as 16. For the GQA-SGG dataset, we follow the standard Faster R-CNN settings to pre-train our object detector.

**Fine-Grained Predicates Learning:** We incorporate our Fined-Grained Predicate learning framework (FGPL) into benchmark models in Model Zoo [9]. Moreover, they share the same hyper-parameters for Category Discriminating Loss (CDL) and Entity Discriminating Loss (EDL). In particular, we set α, β, and ξ as 1.5, 2.0, and 0.9 for CDL. Moreover, when constructing the Predicate Lattice, the Biased Predicate Prediction is derived from the corresponding benchmark model, i.e., Transformer, Motifs and VCTree under the Predcls task. Additionally, based on the correlations derived from Predicate Lattice, we set the number of hard-to-distinguish predicates (i.e.,

---

**Algorithm 2:** Discriminatory Power (DP@K).

**Input:** Confusion Matrix S′ ∈ [R^C×C], with s′ i,j ∈ [0, 1].

**Output:** Models’ Discriminatory Power among top-K hard-to-distinguish predicates, y^K.

for i = 0 to C do
  2: create a Set V′_i = topK(s′_i,j,j∉i)
  3: for s′_i,j in V′_i do
  4: y^K ← y^K + (s′_i,j - s′_i,j)/K
  end for
  6: end for

```latex
y^K ← y^K/C
```
Finally, we empirically set $\gamma$ as 0.1.

Adaptive Fine-Grained Predicate Learning: The coefficient of the Batch-Refinement (BR) regime, $\tau$ is set as 0.99. Moreover, the hyper-parameter $\theta$ in CDL-A is set as 0.1. Besides, all the hyper-parameters (i.e., $\alpha$, $\beta$, $\xi$, $\delta$, $\tau$ and $\theta$) are set the same values in the experiments on the VG-SGG, GQA-SGG, and OI-SGG datasets. Due to the limitation of GPU’s memory, we only conduct experiments on two benchmark models, i.e., Motifs and Transformer, for the GQA-SGG dataset.

Experimental Devices: We perform all the experiments on the server with Ubuntu 20.04.4 LTS and 1 NVIDIA GeForce RTX 3090 GPU. Our codes are implemented with PyTorch 1.9.0.

C. Comparison With State of the Arts

To evaluate our methods’ capability on scene graph generation, we compare them with several state-of-the-art SGG methods under two SGG datasets, i.e., VG-SGG and OI-SGG.

Quantitative Analysis on VG-SGG: We compare our methods with state-of-the-art SGG models via incorporating FGPL and FGPL-A into three SGG benchmark models, namely Transformer [8], Motifs [11], and VCTree [4]. Quantitative results compared with state-of-the-art methods on the VG-SGG dataset is shown in Table 15, available online. First, VCTree (FGPL-A), Transformer (FGPL-A), and Motifs (FGPL-A) outperform all the state-of-the-art methods on all SGG tasks of all the metrics, achieving 44.3%, 42.4% and 40.7% of VCTree (FGPL-A), Transformer (FGPL-A) and Motifs (FGPL-A) on mR@100 under the PredCls task. Specifically, compared with the state-of-the-art SGG method ITrans [72], FGPL-A achieves superior performance on three benchmark models of all the metrics, which demonstrates the effectiveness of FGPL-A on predicates discrimination. To test the hypothesis that fine-grained and long-tail classification methods are ineffective for fine-grained predicate classification, we compare our FGPL-A with state-of-the-art fine-grained/long-tail classification methods. The experiments are conducted by integrating these methods onto an SGG benchmark model, i.e., Transformer. As illustrated in Table I, our FGPL-A outperforms fine-grained (i.e., MC [63] and DCAL [62]) and long-tail recognition methods (i.e., BALMS [52] and LDAM [53]) on all metrics across the three sub-tasks. This highlights the weakness of fine-grained/long-tail classification methods in the fine-grained predicate classification task. For more detailed comparisons, please refer to Table 23 of Appendix F, available online. Moreover, to verify the significant efficacy of predicate correlations for improving discriminatory power among predicates, we make comparisons between benchmark models trained with the Re-weight and FGPL-A learning strategies. We observe that compared with Reweight*-Motifs, Reweight*-VCTree, and Reweight*-Transformer, FGPL-A-Motifs, FGPL-A-VCTree, and FGPL-A-Transformer achieve large margins of improvements by 7.0%, 9.0%, and 8.0% on mR@100 for PredCls, verifying the effectiveness of FGPL-A. Intuitively, fully understanding relationships among predicates, our method can adjust the re-weighting process based on predicate correlations, which boosts and sustains the discriminatory capability over hard-to-distinguish and recognizable predicates, respectively. Finally, it is worth noting that our FGPL-A outperforms LA (i.e., the logit adjustment method adopted in [27]) by 8.4%, 5.9%, and 3.3% on mR@100 (PredCls) of all three benchmark models. We owe the effectiveness to FGPL-A’s adaptive adjustment, which makes the model’s learning process more efficient than the prior distribution-based method (i.e., LA used in [27]). More comparative experiments between FGPL-A and LA are provided in Section IV-E.

Quantitative Analysis on OI-SGG: To verify if our method can work well under different data distributions, we conduct experiments on OI-SGG dataset. Consistent with the performance achieved on VG-SGG and QGA-SGG, we observe that our FGPL-A also achieves state-of-the-art performance on OI-SGG dataset. Specifically, FGPL-A significantly outperforms RU-Net [79] on wmAP$_{rel}$ and wmAP$_{phr}$ by large margins of 0.5% and 1.9%, achieving a new state-of-the-art performance (i.e., score$_{wmL}$ = 43.8%) on the OI-SGG dataset. Since VG-SGG and OI-SGG have different data distributions, the experiments in Tables I and II powerfully demonstrate the practicability of our methods under different data distributions.

D. Generalization on SGG Models

In this section, we verify the generalizability of FGPL and FGPL-A on different SGG benchmark models under two SGG datasets, i.e., VG-SGG and GQA-SGG.

Quantitative Analysis on VG-SGG (Two-stage): To verify the CDL, EDL of FGPL and the CDL-A, EDL-A within FGPL-A are all plug-and-play, we gradually incorporate them into different benchmark models (i.e., Transformer, VCTree and Motifs) on the VG-SGG dataset. Quantitative results are shown in Table III. Integrated with CDL, we observe considerable improvements as at least 17.9% on three benchmark models of mR@100 under the PredCls task, showing the notable generalizability for CDL on figuring out and differentiating among hard-to-distinguish predicates. Besides, after EDL is applied to benchmark models, the performance further boosts about 5.0% on mR@100 under the PredCls task, which manifests the remarkable compatibility of EDL. Moreover, after being extended from CDL to CDL-A, Transformer (CDL-A), VCTree (CDL-A), and Motifs (CDL-A) achieve more considerable gains as 3.5%, 2.5%, and 3.5% on mR@100 under the PredCls task than Transformer (CDL), VCTree (CDL) and Motifs (CDL). It demonstrates the effectiveness in two aspects: 1) The progressively assigned targets increase model’s discriminatory power. 2) The softened targets provide SGG models with comprehensive understanding for fine-grained predicate discrimination. Finally, Transformer (CDL-A+EDL-A), VCTree (CDL-A+EDL-A), and Motifs (CDL-A+EDL-A) achieve the best performance against benchmark models, indicating the generalizability of both CDL-A and EDL-A. We conjecture that by exploring the refined predicate correlations during training, both CDL-A and EDL-A make model’s learning procedure consistent with its gradually gained discriminatory power, ensuring a robust and efficient learning process.
TABLE I
COMPARISON BETWEEN EXISTING METHODS AND OUR METHODS (I.E., FGPL-A AND FGPL) UNDER THREE SUB-TASKS OF mR@K(%) ON THE VG-SGG DATASET

| Method       | Predicate Classification (PredCls) | Scene Graph Classification (SGCls) | Scene Graph Detection (SGDet) |
|--------------|------------------------------------|-----------------------------------|-------------------------------|
| Transformer  |                                    |                                   |                               |
| Transformer  [8] | 12.4 16.0 17.5 7.7 9.6 10.2 5.3 7.3 8.8 |
| -LDA1 [53] | 8.8 11.8 13.9 3.5 5.2 6.6 3.9 5.6 6.9 |
| -DCL [65] | 10.5 14.1 15.8 4.9 7.9 8.9 4.2 5.9 7.0 |
| -LDA2 [55] | 8.5 13.3 16.5 4.9 7.4 8.7 3.2 4.7 5.9 |
| -MC [63] | 12.5 16.4 18.5 7.5 9.6 10.8 5.4 7.3 8.7 |
| -CocTree [16] | 22.9 28.4 31.0 13.0 15.7 16.7 7.9 11.1 12.7 |
| -FGP [45] | 24.3 28.5 31.6 12.3 14.4 18.2 8.1 11.8 14.5 |
| -LA [27] | 25.5 31.7 34.0 17.3 20.1 21.3 11.9 15.7 18.4 |
| -BASOG [17] | 26.7 31.9 34.2 15.7 18.5 19.4 11.4 14.8 17.1 |
| -Reweight* [49] | 19.5 28.6 34.4 11.9 17.2 20.7 8.1 11.5 14.9 |
| -IAFL [73] | 27.5 33.3 35.9 15.7 19.1 20.4 11.4 14.9 17.7 |
| -IAFL [72] | - 35.0 38.0 - 20.8 22.3 - 15.0 18.1 - |
| -FGPL [31] | 27.5 36.4 40.3 19.2 22.6 24.0 13.2 17.4 20.3 |
| -FGPL-A | 28.4 38.0 42.4 20.5 24.0 25.4 13.4 18.0 21.0 |

TABLE II
COMPARISON BETWEEN EXISTING METHODS AND OUR FGPL-A ON THE OI-SGG DATASET

| Model       | R@50 | wmAP_rel | wmAP_phr | score / std |
|-------------|------|----------|----------|-------------|
| Motifs [11] | 11.5 | 14.6 | 15.8 | 6.3 | 8.0 | 8.5 | 4.1 | 3.5 | 6.8 |
| RelDN [37]  | 10.9 | 13.9 | 15.9 | 6.3 | 7.7 | 8.3 | 3.9 | 5.3 | 6.6 |
| -EBM [14]  | 14.2 | 18.0 | 19.5 | 8.2 | 10.2 | 11.0 | 5.7 | 7.7 | 9.3 |
| -Resample [41] | 14.7 | 18.5 | 20.0 | 9.1 | 11.0 | 11.8 | 5.9 | 8.2 | 9.7 |
| -SO [74] | 14.5 | 18.5 | 20.3 | 8.9 | 11.2 | 12.1 | 6.4 | 8.3 | 9.2 |
| -CocTree [97] | 16.6 | 20.4 | 21.9 | 8.5 | 9.9 | 10.4 | 3.9 | 5.5 | 6.9 |
| -PCF [33] | 24.3 | 26.4 | 29.0 | 12.1 | 14.9 | 16.1 | 7.9 | 10.4 | 11.8 |
| -TDE [41] | 18.5 | 24.9 | 28.3 | 11.1 | 13.9 | 15.2 | 6.6 | 8.5 | 9.9 |
| -FGPL [20] | 21.7 | 26.9 | 28.8 | 12.8 | 15.2 | 15.9 | 8.6 | 11.7 | 13.8 |
| -FGPL-A | 20.9 | 26.4 | 29.0 | 12.1 | 14.9 | 16.1 | 7.9 | 10.4 | 11.8 |
| -Reweight* [49] | 24.8 | 29.7 | 31.7 | 14.0 | 16.5 | 17.5 | 10.7 | 13.5 | 15.6 |
| -BASOG [17] | 29.5 | 35.4 | 37.4 | 16.3 | 19.4 | 20.6 | 9.7 | 13.1 | 15.5 |
| -OCL [20] | 29.7 | 35.3 | 37.2 | 16.3 | 19.4 | 20.6 | 9.7 | 13.1 | 15.5 |
| -IAFL [72] | - | 36.6 | 37.2 | - | 20.8 | 21.8 | - | 16.8 | 19.3 |
| -IAFL [73] | - | 35.8 | 37.1 | - | 21.5 | 23.2 | - | 15.5 | 18.0 |
| -FGPL [31] | 24.3 | 33.0 | 35.7 | 17.1 | 21.3 | 22.5 | 11.1 | 15.4 | 18.2 |
| -FGPL-A | 27.2 | 36.3 | 40.7 | 19.9 | 23.2 | 24.5 | 12.5 | 17.0 | 19.8 |

Table adapted from [79].

Quantitative Analysis on VG-SGG (One-stage): Obtained from biased predictions of the pre-trained SGG model, the proposed Adaptive Predicate Lattice (PL-A) of our FGPL-A can work as an inductive bias to regularize the optimization process of existing one-stage models. Hence, to evaluate the generalizability of the proposed FGPL-A, we conduct experiments on the VG-SGG dataset, considering the following baselines:

- RelTR [83] employs the DETR backbone with a triplet decoder taking subject and object queries as input tokens.
- SGTR [84] utilizes an entity-aware predicate representation with a graph assembling module to infer the connectivity of the bipartite scene graph.
- SSR-CNN [27] introduces a cascaded architecture that progressively refines scene graphs through knowledge-distillation from a Siamese Sparse R-CNN.

The comparison results are shown in Table V. The proposed FGPL-A approach achieves significant improvements on mR@K while maintaining comparable R@K. For instance, SGTR (FGPL-A) outperforms SGTR by 4.5%, 2.4% on R@100, 6.6% on R@50, and 8.8% on mR@K. This is attributed to the inductive bias provided by the proposed Adaptive Predicate Lattice (PL-A) and the re-implemented weight-adjustment method proposed in [49].
mR@100. This highlights the generalizability of FGPL-A in assisting one-stage SGG models in generating fine-grained scene graphs.

**Quantitative Analysis on GQA-SGG:** We further verify the generalizability of FGPL and FGPL-A on a more challenging dataset, namely GQA-SGG. Quantitative results are shown in Table IV. From Table IV, integrated with FGPL-A and FGPL benchmark models achieve superior performances on mR@100, and meanwhile attain comparable performances on R@100 of three SGG sub-tasks. It verifies FGPL and FGPL-A’s generalizability on boosting model’s discriminatory power among hard-to-distinguish predicates while maintaining that on recognizable categories.

| Method | Predicates Classification (PredCls) | Scene Graph Classification (SGCls) | Scene Graph Detection (SGDet) |
|--------|------------------------------------|-----------------------------------|-------------------------------|
|        | mR@20 | mR@50 | mR@100 | mR@50 | mR@100 | mR@50 | mR@100 | mR@50 | mR@100 |
| Motifs | 4.6/58.6 | 2.1/24.1 | 1.9/23.3 |
| -FGPL | 6.1/58.8 | 2.7/23.3 | 2.3/23.3 |
| -FGPL-A | 7.9/57.7 | 3.1/23.0 | 2.9/23.3 |
| Transformer | 4.5/57.8 | 2.3/24.1 | 1.7/23.3 |
| -FGPL | 5.2/57.3 | 3.0/23.3 | 3.1/23.4 |
| -FGPL-A | 7.9/57.8 | 3.5/23.5 | 3.2/23.5 |

We validate generalization capability of our FGPL and FGPL-A on GQA-SGG in comparison to benchmark SGG models.

**TABLE VI**

| Method | Predicate Classification (PredCls) | DP@5 | DP@10 | DP@20 | Mean |
|--------|------------------------------------|-------|-------|-------|------|
| Transformer | 15.6 | 17.4 | 18.5 | 17.2 |
| -LA | 27.3 | 30.3 | 32.4 | 30.0 |
| -Re-weight | 33.3 | 36.1 | 38.1 | 35.8 |
| -FGPL | 37.9 | 40.3 | 42.1 | 40.1 |
| -FGPL-A | 38.6 | 41.1 | 42.9 | 40.3 |
| VCTree | 14.1 | 15.7 | 17.3 | 15.7 |
| -LA | 33.8 | 36.2 | 37.9 | 35.9 |
| -Re-weight | 33.9 | 36.5 | 38.4 | 36.3 |
| -FGPL | 35.4 | 37.8 | 39.6 | 37.6 |
| -FGPL-A | 36.6 | 39.3 | 41.0 | 39.0 |
| Motifs | 13.4 | 17.1 | 18.2 | 16.9 |
| -LA | 32.7 | 34.8 | 36.9 | 34.8 |
| -Re-weight | 33.1 | 35.7 | 37.5 | 35.4 |
| -FGPL | 36.1 | 38.7 | 40.6 | 38.5 |
| -FGPL-A | 37.1 | 39.6 | 41.5 | 39.4 |

Re-weight denotes the re-implemented state-of-the-art re-weighting method proposed in [49].
To verify that FGPL and FGPL-A guarantee for an intuitive illustration of FGPL and logit adjustment used in for an intuitive illustration of FGPL S

M (RA

VG-SGG D

B 6

FGPL-A) I

G L M R

D R OF

13934 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 45, NO. 11, NOVEMBER 2023

in Fig. 6. Effectiveness of FGPL and FGPL-A among hard-to-distinguish predicates. Inner-ring (Transformer (FGPL-A)), first-middle-ring (Transformer (FGPL)), second-middle-ring (Transformer (Re-weight)), and outer-ring (Transformer) represent prediction distribution of hard-to-distinguish predicates acquired from different models, for samples with ground truth as “looking at” above, “walking on” below on the VG-SGG dataset.

learning status, our FGPL-A considerably enhances the model’s discriminatory power among hard-to-distinguish predicates. It is also important to note that Transformer (FGPL-A), VCTree (FGPL-A), and Motifs (FGPL-A) achieve consistent progress on DP@10 compared with Transformer (Re-weight/LA), VCTree (Re-weight/LA), and Motifs (Re-weight/LA). It highlights the limitation of re-weighting [49] and logit adjustment used in [27] methods to generate fine-grained predicates. The possible reason is that the fixed bias of re-weighting/logit adjustment, which relies on database statistics, fails to catch up with the model’s dynamic learning status. This leads to excessive adjustment on the model’s training process, thus decreasing the models’ discriminatory power among hard-to-distinguish predicates.

**Qualitative Analysis:** For an intuitive illustration of FGPL and FGPL-A’s discriminatory power, we visualize model’s prediction distribution of hard-to-distinguish predicates from Transformer, Transformer (Re-weight), Transformer (FGPL), and Transformer (FGPL-A) on the VG-SGG dataset, shown in Fig. 6. The proportion of rings indicates the distribution of prediction results, including hard-to-distinguish predicates j and ground truth predicates i, for all samples with ground truth i. For predicate “looking at” in Fig. 6, Transformer struggles to distinguish it from its correlated predicates, e.g., “on” or “holding”. Besides, Transformer (Re-weight) fails to distinct “looking at” from hard-to-distinguish predicates, e.g., “holding”, “carrying”, and “riding”. For Transformer (FGPL), the proportion of correctly classified samples rises from 7.4% to 27.5% compared with Transformer. Meanwhile, hard-to-distinguish predicates are more recognizable than Transformer (Re-weight), i.e., “holding” dropping from 8.6% to 7.4% and “watching” from 12.0% to 10.0%. Consequently, the results validate our FGPL’s efficiency of discriminatory capability among hard-to-distinguish predicates. Furthermore, after being upgraded from FGPL to FGPL-A, it yields a remarkable enhancement on mode’s discriminatory power. The results substantiate that dynamically selecting hard-to-distinguish predicates and progressively setting fine-grained learning targets improve model’s discriminatory power in differentiating among hard-to-distinguish predicates.

2) Analysis of Balanced Predicate Discrimination: Quantitative Analysis: To verify that FGPL and FGPL-A guarantee the balanced predicate discrimination among predicates with different frequencies, we make comparisons among benchmark models (i.e., Transformer, VCTree, and Motifs) incorporated with different learning strategies (i.e., Re-weight, LA, FGPL and FGPL-A) on Group Mean Recall metrics under the PredCls task. The experimental results are shown in Table VII. Generally, after being integrated with FGPL or FGPL-A, we observe significant enhancements on Group Mean Recall metrics. Specifically, compared with Transformer, Transformer (Re-weight), and Transformer (LA), both Transformer (FGPL-A) and Transformer (FGPL) achieve more balanced performance among different predicate groups (e.g., Transformer (FGPL-A): Head-42.2%, Body-42.7%, Tail-42.4%, Mean-42.4%. Transformer (FGPL): Head-42.2%, Body-37.8%, Tail-40.7%, Mean-40.2%), which validates FGPL-A and FGPL’s ability to alleviate imbalanced learning issues of the general and Re-weighting SGG methods.

**Qualitative Analysis:** For an intuitive illustration of FGPL and FGPL-A’s balanced predicate discrimination, we provide Recall@100 results on predicates from Transformer, Transformer (Re-weight), Transformer (FGPL), and Transformer (FGPL-A), as shown in Fig. 7. For clarity, we choose 30 predicates with highest frequency from each predicate group. We observe that Transformer (FGPL) outperforms Transformer and Transformer (Re-weight) on almost all predicates, demonstrating its significant efficacy of ensuring a more effective and balanced learning process among different categories. After being incorporated with FGPL-A, Transformer (FGPL-A) gains further progress on

| Method          | Predicate Classification (PredCls) | Head (10) | Body (10) | Tail (10) | Mean |
|-----------------|-----------------------------------|-----------|-----------|-----------|------|
| Transformer     |                                   | 38.8      | 9.6       | 3.1       | 17.2 |
| -LA             |                                   | 37.1      | 29.0      | 20.4      | 28.9 |
| -Re-weight      |                                   | 39.8      | 34.2      | 28.8      | 34.3 |
| -FGPL           |                                   | 42.2      | 37.8      | 40.7      | 40.2 |
| -FGPL-A         |                                   | 42.2      | 42.7      | 42.4      | 42.4 |
| VCTree          |                                   | 39.5      | 9.1       | 2.9       | 17.2 |
| -LA             |                                   | 39.4      | 34.3      | 28.6      | 33.1 |
| -Re-weight      |                                   | 40.9      | 36.1      | 28.6      | 35.2 |
| -FGPL           |                                   | 38.4      | 43.7      | 38.3      | 40.1 |
| -FGPL-A         |                                   | 43.2      | 39.8      | 39.7      | 40.9 |
| Motifs          |                                   |           |           |           |      |
| -LA             |                                   | 39.8      | 33.4      | 24.2      | 32.5 |
| -Re-weight      |                                   | 40.7      | 35.1      | 24.7      | 33.5 |
| -FGPL           |                                   | 43.4      | 35.8      | 33.5      | 37.6 |
| -FGPL-A         |                                   | 41.9      | 40.5      | 39.8      | 40.7 |

Experiments all conducted under PredCls task.
most of predicate classes, which manifests the effectiveness of FGPL-A in guaranteeing an efficient learning process. Although Transformer (FGPL) and Transformer (FGPL-A) achieve superior performance compared with Transformer, we observe a reduction of performance on some predicates from the Head Group, e.g., “on”, “wearing” and “holding”. The drops of Recall on those head predicates are inevitable in fine-grained classification as observed in long-tailed tasks [85]. As claimed in [17], general SGG models are over-confident on head classes with a high Recall. For discriminating among hard-to-distinguish predicates, FGPL and FGPL-A classify some head classes (e.g., “on”) into their fine-grained ones of tail classes (e.g., “standing on”). It is prone to cause a degradation in head classes’ performance, but meanwhile improves models’ discriminatory power on tail ones with higher performance on DP@K and mR@K, as shown in Table VI and Table 15, available online.

F. Ablation Study

To deeply investigate our FGPL and FGPL-A, we further study different ablation variants of CDL/CDL-A, EDL/EDL-A, and PL-A under the PredCls task on the VG-SGG dataset.

Predicate Correlation (PC) and Re-Weighting Factor (RF) in CDL: We explore the effectiveness of the Predicate Correlation (PC) and the Re-weighting Factor (RF) of CDL. To be specific, we discard PC by ignoring $\varphi_{i,j} > \xi$ and $\varphi_{i,j} \leq \xi$ in (6). Besides, we discard RF by setting Re-weighting Factor $w_{i,j}$ as 1 for all predicate pairs $i$ and $j$ in (4). The results are shown in Table VIII. It is worth noting that CDL (RF) leads to notable progress on mR@100 and DP@K, which demonstrates the effectiveness of RF. Furthermore, CDL outperforms the baseline with a considerable margin after being integrated with PC. We believe adjusting the re-weighting process based on PC, CDL improves the discriminatory power among hard-to-distinguish predicates while maintaining the original discriminating capability among recognizable ones.

Predicate Correlation (PC) and Balancing Factor (BF) in EDL: To validate the superiority for each component of EDL, i.e., Predicate Correlation (PC) and Balancing Factor (BF), we gradually deploy BF and PC onto Transformer (EDL). The experimental results are shown in Table IX. Without Predicate Correlation (PC), we observe a decrease on mR@100 (24.4% versus 21.0%) and DP@K (e.g., DP@10: 24.8% versus 20.6%). It verifies the usefulness of PC for improving discriminatory capability. Additionally, it can be observed that trained without BF, there is a reduction on mR@100 (24.4% versus 18.2%) and DP@K (e.g., DP@10: 24.8% versus 28.2%), demonstrating the efficacy of BF for the enhanced discriminatory ability for hard-to-distinguish predicates. Finally, discarding both PC and BF, we observe a more considerable margin of reduction on mR@50 and DP@K, demonstrating the strengths of both PC and BF.

Batch-Refinement (BR) Regime in PL-A: To prove that our proposed Batch-Refinement (BR) regime (i.e., Category Refining Momentum (CRM) and Entity Refining Momentum (ERM)) strengthens model’s discriminatory power, we gradually incorporate them into Transformer (FGPL-A). Compared with results in the first line (BR without ERM and CRM) of Table X, Transformer (FGPL-A) with ERM or CRM in the second/third line achieves enhancements on both DP@K ($K = 5, 10, 20$) and mR@100 metrics. Then, it performs the best after being integrated with both ERM and CRM.
TABLE X
ABLATION STUDY ON CRM AND ERM OF BR

| BR | Predicate Classification (PredCls) | mR@100/R@100 | DP@5 | DP@10 | DP@20 |
|----|----------------------------------|---------------|------|-------|-------|
| CRM | × | 41.0/50.1 | 38.3 | 40.7 | 42.5 |
| CRM | × | 42.0/50.9 | 38.5 | 41.1 | 42.8 |
| CRM | ✓ | 41.9/53.3 | 38.4 | 41.0 | 42.7 |
| CRM | ✓ | 42.4/53.4 | 38.6 | 41.1 | 42.9 |

CRM and ERM denote category refining momentum and entity refining momentum, respectively. The results are acquired under transformer (FGPL-A).

TABLE XI
ABLATION STUDY ON HARD-TO-DISTRINGUISH THRESHOLD ξ OF CDL

| ξ | Predicate Classification (PredCls) | mR@50/R@50 | mR@100/R@100 | F@50/F@100 |
|---|----------------------------------|------------|--------------|------------|
| 0.7 | 30.8/34.0 | 33.5/37.7 | 39.1/43.7 |
| 0.8 | 30.7/34.1 | 34.4/37.8 | 39.2/43.1 |
| 0.9 | 31.4/34.9 | 35.4/38.5 | 40.0/44.1 |
| 1.0 | 30.3/34.7 | 33.8/38.2 | 39.0/42.8 |
| 1.1 | 30.4/34.7 | 33.7/38.2 | 39.1/42.7 |

All results are obtained under the transformer (CDL) model.

TABLE XII
ABLATION STUDY ON HYPER-PARAMETER τ OF CRM

| τ | Predicate Classification (PredCls) | mR@100/R@100 | DP@10/DP@20 | Head-Body-Tail |
|---|----------------------------------|---------------|------------|---------------|
| 0.9 | 42.0/53.4 | 41.0/42.7 | 41.0/40.54/1.8 |
| 0.99 | 42.4/53.4 | 41.1/42.9 | 42.2/42.7/42.4 |
| 0.999 | 42.0/53.4 | 40.4/42.3 | 42.8/40.0/42.1 |

All results are obtained under the transformer (FGPL-A) model.

regularizes model’s learning procedure in keeping with its dynamic learning pace.

Hard-to-Distinguish Threshold ξ in CDL: To investigate how the hard-to-distinguish threshold ξ influences the discriminating process among predicates, we conduct experiments with different ξ. The quantitative results are shown in Table XI. The results illustrates that increasing ξ makes the model gradually focus on discriminating among hard-to-distinguish predicates while preserving the learned discriminatory power of distinguishable ones. Particularly, it achieves the best on both mR@100 and R@100 with ξ = 0.9 (i.e., mR@100/R@100: 35.4%/58.5%). After that, the continuous increase on ξ (i.e., larger than 0.9) weakens the classifier’s overall discriminatory power among predicates. Ultimately, we empirically figure out the best value for ξ as 0.9.

Hyper-Parameter τ in CRM: In this section, we attempt to explore the impacts of different τ (i.e., the hyper-parameter in CRM) on model’s performance. Generally, as a trade-off coefficient, neither a small nor a large τ can balance the refining process of predicate correlations between the historical and the coming-batch statistics. From Table XII, we observed that model’s Predicate Discriminatory Power (represented by DP@10/DP@20), as well as its Balanced Predicate Discrimination Capability (represented by Head-Body-Tail) get progressively improved with the increase of τ and peaked at τ = 0.99 before decreasing with larger τ. Consequently, we selected τ = 0.99 empirically for the CRM.

G. Practicability Analysis of FGPL-A

To verify the practicability of fine-grained predicates within scene graphs generated by our FGPL-A, we conduct experiments on both Sentence-to-Graph Retrieval and Image Captioning tasks on the Visual Genome (VG) dataset.

Sentence-to-Graph Retrieval: To affirm that scene graphs generated by our FGPL-A precisely describe the contents of images, we conduct the Sentence-to-Graph Retrieval task on the VG dataset, shown in Table XIII. Compared with benchmark models, we observe improvements on R@100 under Gallery 1000 from benchmarks trained with our FGPL-A learning strategies (e.g., Transformer (FGPL-A) (51.8% versus 35.9%), VCTree (FGPL-A) (52.1% versus 45.9%), Motifs (FGPL-A) (54.9% versus 39.3%)). It validates FGPL-A’s practicability in generating fine-grained scene graphs, which dramatically enriches the Sentence-to-Graph Retrieval task with precise scenario information.

Image Captioning: Additionally, we evaluate the practicability of fine-grained scene graphs generated by FGPL-A utilizing the Image Captioning task on the VG dataset in Table XIV. Utilizing the features of scene graphs generated by Transformer, Baseline-Transformer slightly outperforms the Baseline on the Cider and Spice metrics, which demonstrates the effectiveness of scene graphs on the scene-understanding task. Furthermore, when taking advantage of scene graphs generated by our SGG model, i.e., Transformer (FGPL-A) outperform the Baseline by a larger enhancement on almost all the evaluation metrics. It powerfully validates the practicability of fine-grained predicates generated by our FGPL-A. Intuitively, since the FGPL-A regularizes SGG models with the comprehensive understanding on predicate correlations concerning contextual information, the generated scene graphs are prone to faithfully describe the contents of scenarios, benefiting scene-understanding tasks with rich semantics information.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
H. Visualization Results

Finally, we testify the hypothesis that our proposed FGPL-A is capable of generating fine-grained predicates. To the end, we make comparisons among scene graphs generated by Transformer, Transformer (Re-weight), and Transformer (FGPL-A) with the same input images from the VG-SGG dataset in Fig. 8.

We observe that Transformer (FGPL-A) is able to generate more fine-grained relationships between objects than Transformer and Transformer (Re-weight), such as “letter-painted on-boat” rather than “letter-on-boat” in Fig. 8 (1) and “rooftop-over-bench” instead of “roof-in-bench” in Fig. 8 (2). From Fig. 8 (3), both Transformer and Transformer (Re-weight) fail to handle hard-to-distinguish predicates, i.e., “near” and “along”, for the relationship between “tree” and “street”. In contrast, our Transformer (FGPL-A) successfully figures out the fine-grained predicate as “along”, which is faithful to the scenario contexts. It powerfully demonstrates FGPL-A’s effectiveness on generating fine-grained predicates for scene graphs. Intuitively, comprehensively exploring the context information, FGPL-A can differentiate among hard-to-distinguish predicates.

Additionally, to further verify the effectiveness of FGPL-A in various scenarios, we visualize more scene graphs generated from Transformer (FGPL-A) shown in Figs. 8 (4) and 8 (5). We observe that, in Fig. 8 (4), Transformer (FGPL-A) can precisely describe the interaction between “man” and “phone” as “holding”. Similarly, in Fig. 8 (5), Transformer (FGPL-A) accurately infers the spatial relationship as “behind” between “1-elephant” and “2-elephant”. However, there are some unreasonable inference, e.g., “2-elephant-has-head” and “2-elephant-has-truck” shown in Fig. 8 (5), which are structural-inconsistent with “1-elephant-has-head” and “1-elephant-has-truck”. Only considering the triplet-level context information, our method independently treats each “subject-predicate-object” during the
inference process, while ignoring the structural information in the output space of generated scene graphs. Thus, how to generate structural-consistent scene graphs is still a challenging problem to be discussed in the future.

V. CONCLUSION

In this work, we first propose the fine-grained predicates problem, a new perspective of cases that hamper current SGG models. The address the problem, we propose a plug-and-play Adaptive Fine-Grained Predicates Learning (FGPL-A) framework for scene graph generation, which contributes to discriminating among hard-to-distinguish predicates with fine-grained ones. In practice, we devise an Adaptive Predicate Lattice (PL-A) to help understand ubiquitous predicates correlation involving scenarios in SGG datasets and model’s dynamic learning status at each learning pace. Based on the PL-A, we further develop an Adaptive Category Discriminating Loss (CDL-A) and an Adaptive Entity Discriminating Loss (EDL-A), which both help SGG models differentiate among hard-to-distinguish predicates while maintaining learned discriminatory power over recognizable ones throughout training. Finally, comprehensive experiments show that our FGPL-A can differentiate among hard-to-distinguish predicates, benefits SGG models with a balanced discriminating learning procedure, and enriches downstream tasks with precise semantic information.

REFERENCES

[1] B. Schroeder and S. Tripathi, “Structured query-based image retrieval using scene graphs,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2020, pp. 680–688.

[2] J. Gu, S. R. Jooy, J. Cai, H. Zhao, X. Yang, and G. Wang, “Unpaired image captioning via scene graph alignments,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 10322–10331.

[3] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10677–10686.

[4] 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., 2019, pp. 6612–6621.

[5] D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6693–6702.

[6] J. Shi, H. Zhang, and L. Li, “Explainable and explicit visual reasoning over scene graphs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8368–8376.

[7] M. Hildebrandt, H. Li, R. Koner, V. Tresp, and S. Gnnemann, “Scene reasoning for visual question answering,” 2020, arXiv:2007.01072.

[8] A. Vaswani et al., “Attention is all you need,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.

[9] K. Tang, “A scene graph generation codebase in PyTorch,” 2020. [Online]. Available: https://github.com/KaihuaTang/Scene-Graph-Bench mark.pytorch

[10] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3097–3106.

[11] R. Zellers, M. Yasakar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5831–5840.

[12] X. Lin, C. Ding, J. Zeng, and D. Tao, “GPS-Net: Graph property sensing network for scene graph generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 10402–10411.

[13] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, “VRt-VG: Refocusing visually-relevant relationships,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 10402–10411.

[14] M. Suhail et al., “Energy-based learning for scene graph generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13931–13940.

[15] M. Chen, X. Lyu, Y. Guo, J. Liu, L. Gao, and J. Song, “Multi-scale graph attention network for scene graph generation,” in Proc. IEEE Int. Conf. Multimedia Expo, 2022, pp. 1–6.

[16] 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., 2021, pp. 1274–1280.

[17] Y. Gao et al., “From general to specific: Informative scene graph generation via balance adjustment,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 16363–16372.

[18] 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., 2021, pp. 11104–11114.

[19] C. Zheng, X. Lyu, Y. Guo, P. Zeng, J. Song, and L. Gao, “ROUGE: A package for automatic evaluation of summaries,” in Proc. IEEE Int. Conf. Multimedia Expo, 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859711.

[20] 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 Conf. Comput. Vis. Pattern Recognit., 2022, pp. 19405–19414.

[21] A. Desai, T.-Y. Wu, S. Tripathi, and N. Vasconcelos, “Learning of visual relations: The devil is in the tails,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1538–1547.

[22] C. Chen, Y. Zhan, B. Yu, L. Liu, Y. Luo, and B. Du, “Resistance training using prior bias: Toward unbiased scene graph generation,” in Proc. Conf. Assoc. Adv. Artif. Intell., 2022, pp. 212–220.

[23] S. Yan et al., “PCPL: predicate-correlation perception learning for unbiased scene graph generation,” in Proc. ACM Int. Conf. Multimedia, 2020, pp. 265–273.

[24] L. Tao, L. Mi, N. Li, X. Cheng, Y. Hu, and Z. Chen, “Predicate correlation learning for scene graph generation,” in IEEE Trans. Image Process., vol. 31, pp. 4173–4185, 2022.

[25] 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 Conf. Comput. Vis. Pattern Recognit., 2022, pp. 19425–19434.

[26] M. Chiou, H. Ding, H. Yan, C. Wang, R. Zimmermann, and J. Feng, “Recovering the unbiased scene graphs from the biased ones,” in Proc. ACM Int. Conf. Multimedia, 2021, pp. 1581–1590.

[27] Y. Teng and L. Wang, “Structured sparse R-CNN for direct scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 19415–19424.

[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 1321–1330.

[29] Z. Zhong, J. Cui, S. Liu, and J. Jia, “Improving calibration for long-tailed recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3743–3752.

[30] L. J. Old, “An analysis of semantic overlap among English prepositions in Roget’s Thesaurus,” in Proc. Assoc. Comput. Linguistics SIG Semantics Conf., 2003, pp. 13–19.

[31] X. Lyu et al., “Fine-grained predicates learning for scene graph generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 265–273.

[32] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, “Visual relationship detection with internal and external linguistic knowledge distillation,” in Proc. Int. Conf. Comput. Vis., 2017, pp. 1974–1982.

[33] X. Liang, L. Lee, and E. P. Xing, “Deep variation-structured reinforcement learning for visual relationship and attribute detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 848–857.

[34] Y. Zhan, J. Yu, T. Yu, and D. Tao, “On exploring underdetermined relationships for visual relationship detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5128–5137.

[35] Z.-S. Hung, A. Mallya, and S. Lazebnik, “Contextual translation embedding for visual relationship detection and scene graph generation,” in IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 3820–3832, Nov. 2021.

[36] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relationship detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5532–5540.

[37] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanaro, “Graphical contrastive losses for scene graph parsing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11527–11535.

[38] Y. Li, W. Ouyang, X. Wang, and X. Tang, “VIP-CNN: Visual phrase guided convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7244–7253.
Lianli Gao (Member, IEEE) received the PhD degree in information technology from The University of Queensland (UQ), Brisbane, QLD, Australia, in 2015. She is currently a professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. She is focusing on integrating natural language for visual content understanding. She was the winner of the *IEEE Transactions on Multimedia* 2020 Prize Paper Award, the Best Student Paper Award in the Australian Database Conference, Australia, in 2017, the IEEE TC/MC Rising Star Award in 2020, and the ALIBABA Academic Young fellow.

Pengpeng Zeng received the BE degree from the Xi'an University of Technology in 2016, and the ME and PhD degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2019 and 2023, respectively. His current research interests include visual understanding, machine learning, and reinforcement learning.

Heng Tao Shen (Fellow, IEEE) received the BSc with 1st class Honours and PhD degrees from the Department of Computer Science, National University of Singapore in 2000 and 2004 respectively. He is the dean of School of Computer Science and Engineering, the executive dean of AI Research Institute with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests mainly include multimedia search, computer vision, artificial intelligence, and Big Data management. He is/was an associate editor of *ACM Transactions of Data Science, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Knowledge and Data Engineering, and Pattern Recognition*. He is a member of Academia Europaea, fellow of ACM and OSA.

Jingkuan Song (Senior Member, IEEE) is currently a professor with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include large-scale multimedia retrieval, image/video segmentation and image/video understanding using hashing, graph learning, and deep learning techniques. He has been an AC/SPC/PC member of IEEE Conference on Computer Vision and Pattern Recognition for the term 2018–2021, and so on. He was the winner of the Best Paper Award in International Conference on Pattern Recognition, Mexico, in 2016, the Best Student Paper Award in Australian Database Conference, Australia, in 2017, and the Best Paper Honorable Mention Award, Japan, in 2017.