Fine-Grained Predicates Learning for Scene Graph Generation

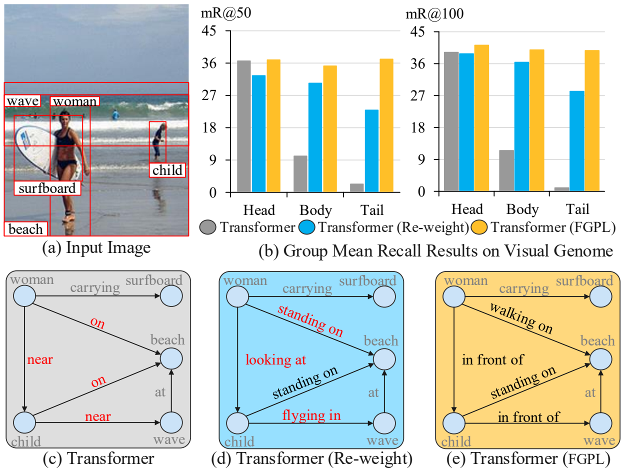
Xinyu Lyu[1] Lianli Gao[1] * Yuyu Guo[1] Zhou Zhao[2] Hao Huang[3] Heng Tao Shen[1]
Jingkuan Song[1]
[1]Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China
[2]Zhejiang University, China
[3]Kuaishou, China

Abstract

*The performance of current Scene Graph Generation models is severely hampered by some hard-to-distinguish predicates, e.g., "woman-on/standing on/walking on-beach" or "woman-near/looking at/in front of-child". While general SGG models are prone to predict head predicates and existing re-balancing strategies prefer tail categories, none of them can appropriately handle these hard-to-distinguish predicates. To tackle this issue, inspired by fine-grained image classification, which focuses on differentiating among hard-to-distinguish object classes, we propose a method named **Fine-Grained Predicates Learning (FGPL)** which aims at differentiating among hard-to-distinguish predicates for Scene Graph Generation task. Specifically, we first introduce a **Predicate Lattice** that helps SGG models to figure out fine-grained predicate pairs. Then, utilizing the Predicate Lattice, we propose a **Category Discriminating Loss** and an **Entity Discriminating Loss**, which both contribute to distinguishing fine-grained predicates while maintaining learned discriminatory power over recognizable ones. The proposed model-agnostic strategy **significantly** boosts the performances of three benchmark models (Transformer, VCTree, and Motif) by **22.8%, 24.1% and 21.7% of Mean Recall (mR@100)** on the Predicate Classification sub-task, respectively. Our model also outperforms state-of-the-art methods by a large margin (i.e., **6.1%, 4.6%, and 3.2% of Mean Recall (mR@100)**) on the Visual Genome dataset. Codes are publicly available[1].*

1. 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 [16, 33], image captioning [6, 30], and visual question answering [9, 10, 12, 17, 18, 22, 26].



(a) Input Image (b) Group Mean Recall Results on Visual Genome

(c) Transformer (d) Transformer (Re-weight) (e) Transformer (FGPL)

Figure 1. **The illustration of handling hard-to-distinguish predicates for SSG models.** (b) Transformer (FGPL) outperforms both Transformer and Transformer (Re-weight) on Group Mean Recall. (c) Transformer [20, 23] is prone to predict head predicates. (d) Transformer (Re-weight) prefers tail categories. (e) Transformer (FGPL) 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".

Prior works [8, 13, 14, 19, 22, 27, 32] have devoted great efforts to exploring representation learning for scene graph generation, but the biased prediction 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 [20, 23] shown in Fig. 1(c). Recent works [2, 7, 11, 31] have exploited re-balancing meth-

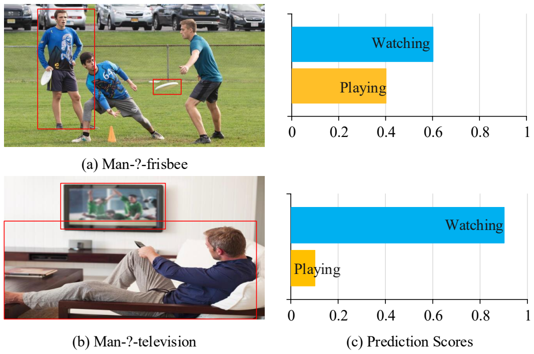
*Corresponding author.
[1] https://github.com/XinyuLyu/FGPL

ods 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, relying on the class distribution, existing re-balancing strategies prefer predicates from tail categories while being hampered by some hard-to-distinguish predicates. For instance, as shown in Fig. 1(d), Transformer (Re-weight) misclassifies *"woman-in front of-child"* as *"woman-looking at-child"* in terms of visual correlations between "in front of" and "looking at".

The origin of the issue lies in the fact that differentiating hard-to-distinguish predicates requires exploring their correlations. Underestimating correlations among predicates, existing methods [28,31] cannot choose hard-to-distinguish ones for sufficient punishment. To acquire complete predicate correlations, we consider contextual information since correlations between a pair of predicates may dramatically vary with contexts as stated in [15]. 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 in Fig. 2(a).

Inspired by the above observations, we propose a Fine-Grained Predicates Learning (FGPL) framework by thoroughly exploiting predicate correlations. We first introduce a Predicate Lattice to help understand ubiquitous predicate correlations concerning all scenarios in the SGG dataset. With the Predicate Lattice, we devise a Category Discriminating Loss (CDL) and an Entity Discriminating Loss (EDL), which both discriminate hard-to-distinguish predicates while maintaining learned discriminatory power over recognizable ones. In particular, Category Discriminating Loss (CDL) attempts to figure out and differentiate hard-to-distinguish predicates. Furthermore, as predicates' correlation varies with contexts of entities, Entity Discriminating Loss (EDL) adaptively adjusts the discriminating process according to predictions of entities. Using CDL and EDL, our method can determine whether predicate pairs are hard-to-distinguish or not during training, which guarantees a more balanced learning process among different categories than previous methods [2,7,11,28,31].

Contribution: Our main contributions are summarized as follows: 1). We propose a novel plug-and-play Fine-Grained Predicates Learning (FGPL) framework to differentiate hard-to-distinguish predicates for scene graph generation. 2). We devise a Predicate Lattice to obtain complete predicate correlations between each predicate pair concerning context information. Category Discriminating Loss (CDL) aims at figuring out and differentiating hard-to-distinguish predicates. Moreover, Entity Discriminating Loss (EDL) adaptively adjusts the discriminating process according to predictions of entities. 3). Our FGPL greatly boosts performances of three benchmark models (Transformer, VCTree, and Motif) by 22.8%, 24.1%, and 21.7% of Mean Recall (mR@100) on Predicate Classification subtask and achieves superior performances over state-of-the-art methods by a large margin (*i.e.*, 6.1%, 4.6% and 3.2% of Mean Recall (mR@100)) on Visual Genome dataset.



(a) Man-?-frisbee

(b) Man-?-television

(c) Prediction Scores

Figure 2. **The illustration of predicate correlations concerning contexts.** The predicate correlation between "watching" and "playing" varies with contexts. Especially, "watching/playing" is weakly correlated or distinguishable in (b), while they are strongly correlated or hard-to-distinguish in (a).

2. Related work

Scene Graph Generation: Suffering from biased prediction, today's SGG task is far from practical. To deal with the problem, some methods [1, 11, 21] are proposed to balance discriminating process in accordance with class distribution or visual clues. [28,31] explore predicate correlations with hierarchical or global structures to discriminate predicates. While correlations among predicates varies with contexts, it is neither hierarchical nor global. Thus, we focus on discriminating hard-to-distinguish predicates with pair-wise predicate correlations, constructed as a predicate graph.

Long-Tailed Distribution Classification: To solve long-tailed problem, various distribution-based re-balancing learning strategies [4,24,25] have been proposed. However, besides class distribution, correlations are crucial for differentiating hard-to-distinguish predicates in SGG. Therefore, in this work, we take advantage of both predicate distribution and predicate correlations to handle this issue.

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 from two perspectives, representation-encoding [5, 34] and local recognition [3, 29]. However, due to complex relationships among predicates, such a coarse-to-fine discriminatory manner may fail to differentiate predicates for SGG. Particularly, different predicates may share similar meanings in a specific scenario, while a predicate may have different meanings in different contexts. Instead of hierarchical structures, predicate correlations should be formed as a graph. Concretely, we construct Predicate Lattice to comprehend predicate correlations for predicate discriminating.

3. Fine-Grained Predicates Learning

3.1. Problem Formulation

Scene graph generation is typically a two-stage multi-class classification task. In the first stage, Faster R-CNN detects instance labels $O = \{o_i\}$, bounding boxes $B = \{b_i\}$, and feature maps $X = \{x_i\}$ within an input image I. In the second stage, scene graph models infer the predicate category from subject i to subject j, *i.e.*, $R = \{r_{ij}\}$, based on the detection results, *i.e.*, $Pr(R|O, B, X)$.

Within our Fine-Grained Predicates Learning (FGPL) framework, shown in Fig. 4, we first construct a Predicate Lattice concerning context information to understand ubiquitous correlations among predicates. Then, utilizing the Predicate Lattice, we develop a Category Discriminating Loss and an Entity Discriminating Loss which help SGG models differentiate hard-to-distinguish predicates.

3.2. Predicate Lattice Construction

To fully understand relationships among predicates, we build a Predicate Lattice, which includes correlations for each pair of predicates concerning contextual information. In general, predicate correlations are acquired under different contexts, since contexts (i.e., visual or semantic information of predicates' subjects and objects) determine relationships among predicates. Specifically, we extract their contextual-based correlations from biased predictions containing all possible contexts between each pair of predicates. The construction procedure is shown in Fig. 3.

Context-Predicate Association: We first establish Context-Predicate associations between predicate nodes and context nodes. As contexts determine correlations among predicates, predicate correlations are constructed as a Predicate Lattice containing predicates and related contexts (*i.e.*, visual or semantic information of predicates' subjects and objects). In Fig. 3(a), we show structures of our Predicate Lattice. There are two kinds of nodes in Predicate Lattice, namely Predicate nodes and Context nodes, which indicate predicate categories and labels of subject-object pairs, respectively. Several predicate nodes connect to the same context node, which denotes that several predicates can describe relationships in the same context. For instance in Fig. 3(a), both "holding" and "carrying" can be utilized to describe relationships for "person-racket". Specifically, we adopt Frequency model [32] to derive every subject-object pair as the context for each predicate from the SGG dataset (VG). Moreover, weights of edges between predicate nodes and context nodes, *i.e.*, $Pr(r_{ij}|o_i, o_j)$, denote occurrence frequency for each "subject(o_i)-predicate(r_{ij})-object(o_j)" triplet in dataset. In this way, we establish connections between predicate nodes and context nodes in Predicate Lattice.

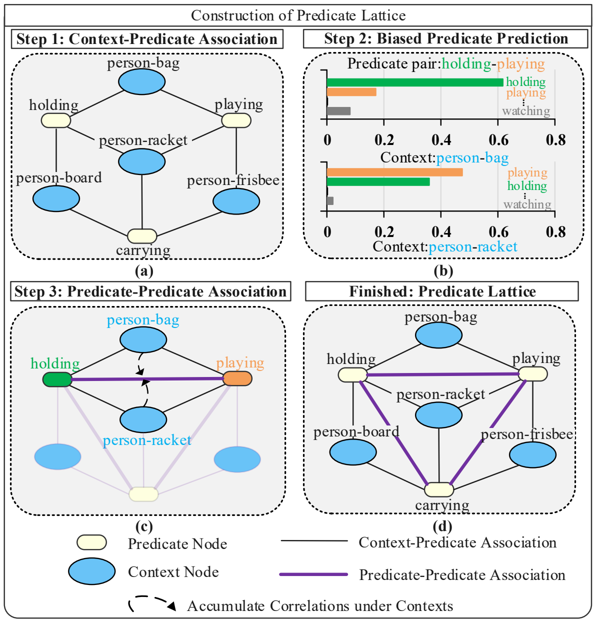


Figure 3. **Construction of Predicate Lattice.** The whole process is divided into three steps: (1) Context-Predicate Association; (2) Biased Predicate Prediction; (3) Predicate-Predicate Association.

Biased Predicate Prediction: To associate predicate pairs with predicate correlations in the next step, we acquire Biased Predicate Prediction from SGG models. Firstly, we incorporate Context-Predicate Association, constructed in step one, into SGG models. Particularly, we extract the Context-Predicate Association for each "subject-predicate-object" triplet as semantic information. Then, to acquire complete contextual information, we combine semantic information with visual features, *i.e.*, $B = b_i$ and X, of subjects o_i and objects o_j to predict predicates $Pr(r_{ij}|o_i, o_j, b_i, b_j, x_i, x_j)$. With the contextual information, we derive the Biased Predicate Prediction of pre-trained SGG models by inferring on the training set of the SGG dataset concerning all scenarios. In this way, the Biased Predicate Prediction contains predicate predictions under all possible scenarios for each predicate pair. For instance, as shown in Fig. 3(b), we infer the pre-trained SGG model under all possible scenarios for predicate "playing" or "holding", such as "person-racket" and "person-bag".

Predicate-Predicate Association: At last, we establish Predicate-Predicate Association among predicates with context-based correlations obtained from the Biased Predicate Prediction. The Biased Predicate Prediction implies the context-based correlations between each pair of predicates. For instance, if most samples are predicted as j but labeled as i in ground truth, predicate i is correlated to pred-

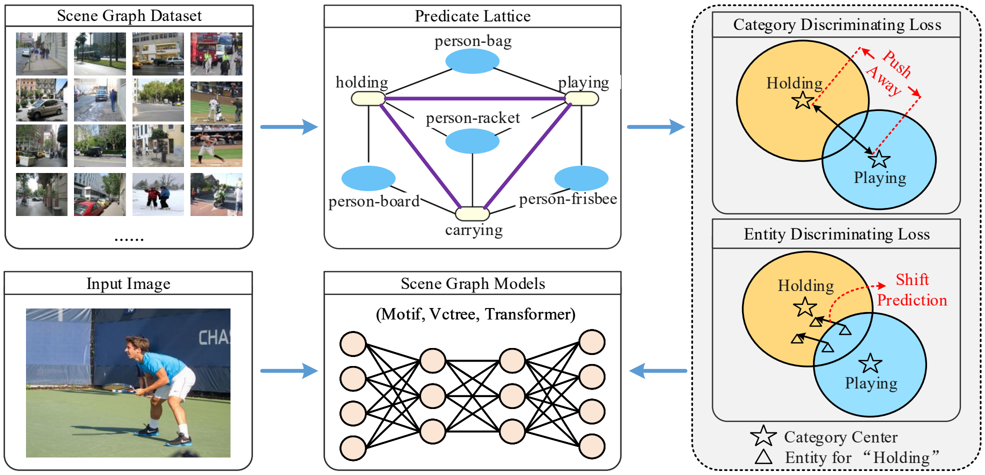


Figure 4. **The Overview of our Fine-Grained Predicates Learning (FGPL) framework.** It includes three parts: Predicate Lattice, Category Discriminating Loss, Entity Discriminating Loss. Fine-Grained Predicates Learning is incorporated into several state-of-the-art SGG models. Predicate Lattice is constructed from the SGG dataset (Visual Genome) to help understand predicates correlations. With Predicate Lattice, the architecture is optimized with two terms: Category Discriminating Loss and Entity Discriminating Loss.

icate j in most contexts. Based on the above observation, we accumulate prediction results from each possible context to obtain the holistic predicate correlations between each pair of predicates, shown in Fig. 3(c). For instance, given predicate pair "playing-holding", we gather their correlations under all contexts/scenarios, such as "person-racket" and "person-bag". Moreover, if predicate i is correlated to predicate j in most contexts, they are prone to be strongly correlated. Therefore, we normalize the gathered predicate correlations as $S = \{s_{ij}\}$ with $s_{ij} \in [0, 1]$, which indicates the proportion of samples labeled as i but predicted as j. In particular, higher s_{ij} means a stronger correlation between predicate pair i and j. Then, we associate predicate pairs with predicate correlations s_{ij}. Finally, predicate correlations are formed as a Predicate Lattice, shown in Fig. 3(d).

3.3. Category Discriminating Loss

In this section, we first analyze the limitations of re-weighting methods. Then, we introduce our Category Discriminating Loss (CDL) in detail.

Limitations of Re-weighting Methods: 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. To be specific, the state-of-the-art re-weighting method [24] adjusts weights for each class in Cross-Entropy Loss on the basis of the proportion of training samples as follows:

$$\mathcal{L}_{CD}(\eta) = -\textstyle\sum_{i=1}^{C} y_i log(\hat{\phi}_i) \ ,$$
$$\hat{\phi}_i = \frac{e^{\eta_i}}{\sum_{j=1}^{C} w_{ij} e^{\eta_j}} \ , w_{ij} = \begin{cases} (\frac{n_j}{n_i})^{\alpha}, & \text{if } n_j > n_i \\ 1, & \text{if } n_j \le n_i \end{cases} \ , \qquad (1)$$

where $\eta = [\eta_1, \eta_2, ..., \eta_C]$ and $\hat{\phi} = [\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_C]$ denote predicted logits and re-weighted probabilities for each class. The label $Y = [y_1, y_2, ..., y_C]$ is a one-hot vector. Additionally, w_{ij} denotes the re-weighting factor concerning distribution between positive class i and negative class j. Explicitly, w_{ij} is calculated based on the proportion of distribution between class i and j, as shown in Eq. 1, where $\alpha > 0$.

$$\frac{\partial \mathcal{L}_{CD}(\eta)}{\partial \eta_j} = \frac{w_{ij} e^{\eta_j}}{\sum_{k=1}^{C} w_{ik} e^{\eta_k}} \ . \qquad (2)$$

Eq .2 shows negative gradients for category j. If positive category i is less frequent than negative category j, *i.e.*, $n_j > n_i$ with $w_{ij} > 1$, it will strengthen the punishment to negative class j. On the contrary, if $n_j \le n_i$ with $w_{ij} = 1$, 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 correlation reveals 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 head classes to tail classes, such discriminating process is prone to over-suppress weakly correlated predicate pairs and degrades the learned discriminatory ability of recognizable predicates as maintained in [4,25]. 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 hard-to-distinguish predicates. Thus, based on the re-weighting method in Eq. 1, 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_{ij}, defined in Sec. 3.2, 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_{ij}, in Eq. 1, 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:

$$
\begin{aligned}
w_{ij} &= \begin{cases} \mu_{ij}^{\beta}\ (\geq 1), & \text{if } \mu_{ij} \geq 1 \textit{ and } \varphi_{ij} > \xi \\ 1, & \text{if } \mu_{ij} \geq 1 \textit{ and } \varphi_{ij} \leq \xi \\ 1, & \text{if } \mu_{ij} < 1 \textit{ and } \varphi_{ij} > \xi \\ \mu_{ij}^{\alpha}\ (< 1), & \text{if } \mu_{ij} < 1 \textit{ and } \varphi_{ij} \leq \xi \end{cases}, \\
\mu_{ij} &= \frac{n_j}{n_i},\ \varphi_{ij} = \frac{s_{ij}}{s_{ii}},
\end{aligned}
\tag{3}
$$

where φ_{ij} is calculated by the proportion between s_{ij} and s_{ii}, revealing correlations between predicate i and j. In addition, α and β are hyper-parameters larger than 0. For instance, when $n_j \geq n_i$ ($\mu_{ij} \geq 1$), if $\varphi_{ij} > \xi$ of strongly correlated predicate pair i and j, w_{ij} is larger than 1 to strengthen the punishment on negative predicate j. In contrast, if $\varphi_{ij} \leq \xi$ of weakly correlated predicate pair i and j, w_{ij} 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 $n_j < n_i$ ($\mu_{ij} < 1$), we set $w_{ij} \leq 1$ (including $\varphi_{ij} > \xi$ and $\varphi_{ij} \leq \xi$) to relieve the over-suppression from head predicate i to tail one j. Moreover, if $\varphi_{ij} \leq \xi$, we set $w_{ij} = \mu_{ij}^{\alpha}$ (< 1) to mitigate the magnitude of the penalty on negative predicate j.

3.4. Entity Discriminating Loss

Although CDL can effectively differentiate hard-to-distinguish predicates, it still has a limitation: weights assigned to predicates are stable during training, which can neither adapt to the gradually obtained discriminatory power during training nor contexts varied with training samples. Hence, we individually treat prediction results of each sample as signals to adjust the decision boundary. Based on the observations, we propose Entity Discriminating Loss (EDL), which adapts the discriminating process to the learning status and contexts, shown as below:

$$
\mathcal{L}_{ED}(\eta) = \frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} \max(0, \phi_j - \phi_i + \delta) \frac{n_j}{n_i}\ , \tag{4}
$$

where $\mathcal{V}_i$ is defined as a set of strongly correlated predicates selected in reference to predicate correlations s_{ij} in Predicate Lattice. For each predicate category i, M predicates with the highest s_{ij} in the Predicate Lattice are chosen to construct $\mathcal{V}_i$. Given the input sample η, ϕ_i and ϕ_j are the predicted probabilities for predicates i and j, and $\phi_j - \phi_i$ implies the learned discriminatory ability between them during training. The δ is a hyper-parameter, which denotes prediction margins for predicates. Furthermore, EDL is reduced to zero if predicate pairs are distinguishable enough *i.e.*, $\phi_i - \phi_j \geq \delta$. Moreover, we also adopt the balancing factor $\frac{n_j}{n_i}$ to alleviate imbalanced gradients between classes with fewer or more observations.

Finally, we combine CDL and EDL as Eq. 5, which distinguishes hard-to-distinguish predicates while maintaining the performance between distinguishable ones.

$$
\mathcal{L}(\eta) = \mathcal{L}_{CD}(\eta) + \lambda \mathcal{L}_{ED}(\eta)\ , \tag{5}
$$

where $\mathcal{L}_{CD}$ and $\mathcal{L}_{ED}$ denote Category Discriminating Loss and Entity Discriminating Loss. Futhermore, λ is a hyper-parameter balancing CDL and EDL.

4. Experiments

4.1. Experiment Setting

Dataset: Following previous works [1, 21, 32], we adopt widely used Visual Genome split for scene graph generation. Under the setting, the Visual Genome dataset has 150 object categories and 50 relationship categories. Then, we divide the dataset into 70% training set, 30% testing set, and $5k$ images from the training set for validation.

Model Configuration: For our Fine-Grained Predicates Learning (FGPL) is model-agnostic, following recent works [7], we incorporate it into VCTree [22], Motif [32], and Transformer [23] in the SGG benchmark [20].

Evaluation Metrics: We evaluate our methods on three sub-tasks in scene graph generation, including PredCls, SGCls, and SGDet. Following recent works [1, 11, 22], we evaluate the performance of prior methods on mR@K and Group Mean Recall, *i.e.*, head, body, and tail. 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 ($j \in \mathcal{V}_i'$). Furthermore, $\mathcal{V}_i'$ 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

Method	Predicate Classification (PredCls)			Scene Graph Classification (SGCls)			Scene Graph Detection (SGDet)		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
BGNN [11]	-	30.4	32.9	-	14.3	16.5	-	10.7	12.6
PCPL [28]	-	35.2	37.8	-	18.6	19.6	-	9.5	11.7
TDE-VCTree [21,22]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
CogTree-Motif [31,32]	20.9	26.4	29.0	12.1	14.9	16.1	7.9	10.4	11.8
CogTree-VCTree [22,31]	22.0	27.6	29.7	15.4	18.8	19.9	7.8	10.4	12.1
CogTree-Transformer [23,31]	22.9	28.4	31.0	13.0	15.7	16.7	7.9	11.1	12.7
Reweight*-Motif [24,32]	18.8	28.1	33.7	10.7	15.6	18.3	7.2	10.5	13.2
Reweight*-VCTree [22,24]	19.4	29.6	35.3	13.7	19.9	23.5	7.0	10.5	13.1
Reweight*-Transformer [23,24]	19.5	28.6	34.4	11.9	17.2	20.7	8.1	11.5	14.9
FGPL-Motif	**24.3**	**33.0**	**37.5**	**17.1**	**21.3**	**22.5**	**11.1**	**15.4**	**18.2**
FGPL-VCTree	**30.8**	**37.5**	**40.2**	**21.9**	**26.2**	**27.6**	**11.9**	**16.2**	**19.1**
FGPL-Transformer	**27.5**	**36.4**	**40.3**	**19.2**	**22.6**	**24.0**	**13.2**	**17.4**	**20.3**

Table 1. **Comparison between existing methods and FGPL**. * denotes state-of-the-art re-weighting method proposed in [24].

matrix $S' \in \mathbb{R}^{C \times C}$ from the model's prediction results, with $s'_{ij} \in [0, 1]$, which denotes the degree of confusion between the predicate pair i and j. For each predicate category i, k predicates with the highest s'_{ij} are chosen to construct $\mathcal{V}'_i$. In a word, a higher score of DP@K means stronger discriminatory power against hard-to-distinguish predicates.

4.2. Implementation Details

Detector: For object detectors, we utilize the pre-trained Faster R-CNN by [21] to detect objects in images. Moreover, weights of the object detectors are frozen during training of scene graph generation for all three sub-tasks.

Scene Graph Generation Model: Following [20], baselines are trained with Cross-Entropy Loss and SGD optimizer with an initial learning rate of 0.01, batch size as 16.

Fine-Grained Predicates Learning: We incorporate our FGPL into baselines in Model Zoo [20] with the same hyper-parameters for CDL and EDL. In particular, we set α, β, and ξ as 1.5, 2.0, and 0.9 for CDL. Additionally, we set the number of hard-to-distinguish predicates (*i.e.*, $|\mathcal{V}_i|$) as 5 for EDL. Furthermore, the boundary margin δ, and the hyper-parameter λ are set as 0.5 and 0.1, respectively.

4.3. Comparison with State of the Arts

We evaluate our FGPL by incorporating them into three SGG baselines, namely Transformer [23], Motif [32], and VCTree [22]. Quantitative results compared with state-of-the-art methods on Visual Genome are shown in Tab. 1. Specifically, FGPL-Motif, FGPL-VCTree, and FGPL-Transformer outperform CogTree-Motif, CogTree-VCTree, and CogTree-Transformer with consistent improvements as 8.5%, 10.5%, and 9.3% on mR@100 for PredCls, respectively, demonstrating the effectiveness of the Lattice-Structured Predicate Correlation against the Tree-Structured one, *i.e.*, CogTree. It is worth noting that, although Reweight*-Motif, Reweight*-VCTree, and Reweight*-Transformer exceed most of the prior works on all metrics, FGPL-Motif, FGPL-VCTree, and FGPL-Transformer still achieve a large margin of improvements by 3.8%, 4.9%, and 5.9% on mR@100 for PredCls, verifying the significant efficacy of FGPL for improving discriminatory power over predicates. Intuitively, fully understanding relationships over predicates, our method can adjust the re-weighting process based on predicate correlations, enhancing the discriminatory ability over predicates.

4.4. Generalization on SGG Models

To verify that both CDL and EDL of FGPL are plug-and-play, we incorporate them into different benchmark models, including Transformer, VCTree, and Motif. Quantitative results on Visual Genome are shown in Tab. 2. From Tab. 2, compared with baselines, we observe considerate improvements on Transformer-FGPL (CDL) (17.5% *vs.* 35.4%), VCTree-FGPL (CDL) (16.1% *vs.* 35.3%), Motif-FGPL (CDL) (15.8% *vs.* 34.4%) on mR@100 of PredCls task, showing notable generalizability for FGPL (CDL). The reason lies in the fact that CDL helps to figure out and differentiate hard-to-distinguish predicates. Furthermore, after being integrated with FGPL (EDL), our Transformer-FGPL (CDL+EDL), VCTree-FGPL (CDL+EDL), and Motif-FGPL (CDL+EDL) achieve further progress as 4.9%, 4.9%, and 3.1% on mR@100 of PredCls task, which manifests the great compatibility of our FGPL (EDL). The possible reason is that EDL adjusts the learning process according to the discriminatory ability and contexts varied with learning process and training samples, respectively.

4.5. Predicate Discrimination of FGPL

We observe that FGPL helps SGG models differentiate hard-to-distinguish predicates, and hence give quantitative and qualitative studies to obtain deep insights into FGPL.

Quantitative Analysis: As hypothesized, our FGPL improves discriminatory power among hard-to-distinguish predicates while preserving distinguishable ones compared with re-weighting methods. Accordingly, we conduct experiments among three settings to testify our hypothesis: 1) Baselines with traditional Cross-Entropy Loss. 2) Baselines with the state-of-the-art re-weighting method in [24]. 3) Baselines with our FGPL. To focus on predictions of predicates, we only conduct experiments on PredCls task. Tab. 3 presents comparisons among three settings on Transformer, VCTree, and Motif. Besides mR@50, we also evaluate

Method	Predicate Classification (PredCls)			Scene Graph Classification (SGCls)			Scene Graph Detection (SGDet)		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
Transformer	12.4	16.0	17.5	7.7	9.6	10.2	5.3	7.3	8.8
Transformer-FGPL(CDL)	23.0 ↑ **10.6**	31.4 ↑ **15.4**	35.4 ↑ **17.9**	14.3 ↑ **6.6**	18.9 ↑ **9.3**	21.2 ↑ **11.0**	9.4 ↑ **4.1**	13.3 ↑ **6.0**	16.5 ↑ **7.7**
Transformer-FGPL(CDL+EDL)	27.5 ↑ **15.1**	36.4 ↑ **20.4**	40.3 ↑ **22.8**	19.2 ↑ **11.5**	22.6 ↑ **13.0**	24.0 ↑ **13.8**	13.2 ↑ **7.9**	17.4 ↑ **10.1**	20.3 ↑ **11.5**
VCTree	11.7	14.9	16.1	6.2	7.5	7.9	4.2	5.7	6.9
VCTree-FGPL(CDL)	23.0 ↑ **11.3**	31.6 ↑ **16.7**	35.3 ↑ **19.2**	15.7 ↑ **9.5**	21.1 ↑ **13.6**	23.3 ↑ **15.4**	11.0 ↑ **6.8**	14.7 ↑ **9.0**	17.5 ↑ **10.6**
VCTree-FGPL(CDL+EDL)	30.8 ↑ **19.1**	37.5 ↑ **22.6**	40.2 ↑ **24.1**	21.9 ↑ **15.7**	26.2 ↑ **18.7**	27.6 ↑ **19.7**	11.9 ↑ **7.7**	16.2 ↑ **10.5**	19.1 ↑ **12.2**
Motif	11.5	14.6	15.8	6.5	8.0	8.5	4.1	5.5	6.8
Motif-FGPL(CDL)	22.2 ↑ **10.7**	30.3 ↑ **15.7**	34.4 ↑ **18.6**	12.6 ↑ **6.1**	16.7 ↑ **8.7**	18.5 ↑ **10.0**	8.2 ↑ **4.1**	11.6 ↑ **6.1**	14.3 ↑ **7.5**
Motif-FGPL(CDL+EDL)	24.3 ↑ **12.8**	33.0 ↑ **18.4**	37.5 ↑ **21.7**	17.1 ↑ **10.6**	21.3 ↑ **13.3**	22.5 ↑ **14.0**	11.1 ↑ **7.0**	15.4 ↑ **9.9**	18.2 ↑ **11.4**

Table 2. **Quantitative results on generalizability of CDL and EDL in FGPL.** We validate generalization capability of our proposed components, *i.e.*, Entity Discriminating Loss (EDL) and Category Discriminating Loss (CDL), in comparison with baselines.

them with DP@K to show discriminatory ability among hard-to-distinguish predicates. After being integrated with FGPL, Transformer (FGPL), VCTree (FGPL), and Motif (FGPL) greatly surpass baselines on DP@10 with a large margin as 22.9% , 22.1%, and 22.1%. It provides direct evidence that our FGPL considerably improves discriminatory power against hard-to-distinguish predicates. It is also important to note that Transformer (FGPL), VCTree (FGPL), and Motif (FGPL) achieve consistent progress on DP@10 compared with Transformer (Re-weight), VCTree (Re-weight), and Motif (Re-weight). It reflects that our FGPL improves discriminatory ability over the re-weighting method [24] to generate fine-grained predicates. One possible reason is that FGPL makes the learning process both adapt to correlations of predicates and inherent contextual information of each sample, strengthening discriminatory power against hard-to-distinguish predicates.

Qualitative Analysis: For an intuitive illustration of FGPL's discriminatory power among hard-to-distinguish predicates, we visualize discrimination among hard-to-distinguish predicates of Transformer, Transformer (Re-weight), and Transformer (FGPL), shown in Fig. 5. 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 "standing on" in Fig. 5, Transformer struggles to distinguish it from its correlated predicates, *e.g.*, "in" or "on". Besides, Transformer (Re-weight) fails to distinct among hard-to-distinguish predicates, *e.g.*, "standing on", "walking on", and "sitting on". For Transformer (FGPL), the proportion of correctly classified samples rises from 6% to 39% compared with Transformer. Meanwhile, hard-to-distinguish predicates are more recognizable than Transformer (Re-weight), *i.e.*, "walking on" dropping from 16% to 14% and "sitting on" from 5% to 4%. Consequently, the results validate our FGPL's efficiency of discriminatory ability against hard-to-distinguish predicates.

4.6. Ablation Study

To deeply investigate our FGPL, we further study different ablation variants of CDL and EDL on PredCls task.

Method	Predicate Classification (PredCls)			
	mR@50	DP@1	DP@5	DP@10
Transformer	16.0	9.9	15.6	17.4
Transformer (Re-weight)	28.6	25.3	33.3	36.1
Transformer (FGPL)	**36.4**	**30.1**	**37.9**	**40.3**
VCTree	14.9	10.5	14.1	15.7
VCTree (Re-weight)	29.6	26.2	33.9	36.5
VCTree (FGPL)	**37.5**	**27.1**	**35.4**	**37.8**
Motif	14.6	10.0	15.1	16.6
Motif (Re-weight)	29.6	25.6	33.0	35.6
Motif (FGPL)	**33.0**	**28.6**	**36.1**	**38.7**

Table 3. **Quantitative results on discriminatory power** of top-k hard-to-distinguish Predicates (DP@K(%)) on PredCls.

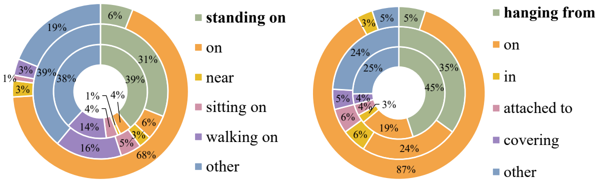


Figure 5. **The effectiveness of FGPL among hard-to-distinguish predicates.** The inner-ring, middle-ring, and outer-ring represent prediction distribution of hard-to-distinguish predicates acquired from Transformer (FGPL), Transformer (Re-weight), and Transformer respectively, for samples with ground truth as **"standing on"** on the left, **"hanging from"** on the right.

Entity Discriminating Loss: To validate the superiority for each component of Entity Discriminating Loss, *i.e.*, Predicate Correlation (PC) and Balancing Factor (BF), we experiment with the following four settings: 1) Transformer with EDL (without PC and BF). 2) Transformer with EDL (without PC), *i.e.*, setting $\mathcal{V}_i$ in Eq. 4 as a set containing all predicate categories. 3) Transformer with EDL (without BF), *i.e.*, removing the balancing factor $\frac{n_j}{n_i}$ in Eq. 4. 4) Transformer with EDL (with PC and BF). The experimental results are shown in Tab. 4. Without Predicate Correlation (PC), we observe a steep decrease on mR@50 (22.0% *vs.* 17.0%) and Group Mean Recall (head:39.2% *vs.* 37.2%, body:19.7% *vs.* 11.4%, tail:7.4% *vs.* 3.7%). It verifies the usefulness of PC for improving discriminatory capability for SGG models. The possible reason is that EDL (PC) explores the underlying context information within each entity and adjusts the discriminating process based on the gradually obtained discriminatory capability to alleviate the issue of imbalanced learning. Additionally, it can be observed that trained without BF, there is a substantial reduc-

EDL		Predicate Classification (PredCls)			
PC	BF	mR@50	head (16)	body (17)	tail (17)
×	×	16.2	36.5	10.5	2.5
×	✓	17.0	37.2	11.4	3.7
✓	×	18.9	38.4	16.3	5.7
✓	✓	**22.0**	**39.2**	**19.7**	**7.4**

Table 4. **Ablation study on each component of EDL**. PC and BF denote Predicate Correlation and Balancing Factor, respectively. The results are obtained with Transformer as the baseline.

CDL		Predicate Classification (PredCls)			
PC	RF	mR@50	head (16)	body (17)	tail (17)
×	×	16.0	36.6	10.1	2.3
×	✓	28.6	32.5	30.4	22.9
✓	✓	**31.4**	**37.7**	**33.5**	**23.3**

Table 5. **Ablation study on PC and RF of CDL**. PC denotes Predicate Correlation. RF denotes the Re-weighting Factor. The results are obtained with Transformer as the baseline.

tion on mR@50 (22.0% *vs.* 18.9%) and Group Mean Recall (head:39.2% *vs.* 38.4%, body:19.7% *vs.* 16.3%, tail:7.4% *vs.* 5.7%), demonstrating the efficacy of BF for a more efficient learning process. We think this may be caused by alleviating over-suppression to tail classes, which leads to a balanced discriminating process among classes with different frequency. At last, when both discarding PC and BF, we observe a larger margin of reduction on mR@50 and Group Mean Recall, demonstrating effectiveness of PC and BF.

Category Discriminating Loss: We explore the effectiveness of the Predicate Correlation (PC) and the Re-weighting Factor (RF) of Category Discriminating Loss. To be specific, we discard PC by ignoring $\varphi_{ij} > \xi$ and $\varphi_{ij} \leq \xi$ in Eq. 3. Besides, we discard RF by setting Re-weighting Factor w_{ij} as 1 for all predicate pairs i and j in Eq. 1. The results are shown in Tab. 5. It is worth noting that CDL (RF) leads to notable progress on mR@50 and Group Mean Recall, which proves the efficacy of RF for keeping a balanced learning process. Furthermore, CDL outperforms the baseline with a considerable margin after being integrated with PC. We believe that adjusting the re-weighting process according to PC, CDL improves the discriminatory power among hard-to-distinguish predicates while maintaining the original discriminating ability among recognizable ones.

4.7. Visualization Results

To intuitively illustrate the effectiveness of our proposed FGPL, we make comparisons among scene graphs generated by Transformer, Transformer (Re-weight), and Transformer (FGPL) with the same input images in Fig. 6. We observe that Transformer (FPL) is capable of generating more fine-grained relationships between objects than Transformer and Transformer (Re-weight), such as "man-walking in-snow" rather than "man-on-snow", "tree-across-street" instead of "tree-near-street", and "sidewalk-along-street" as opposed "sidewalk-near-street".

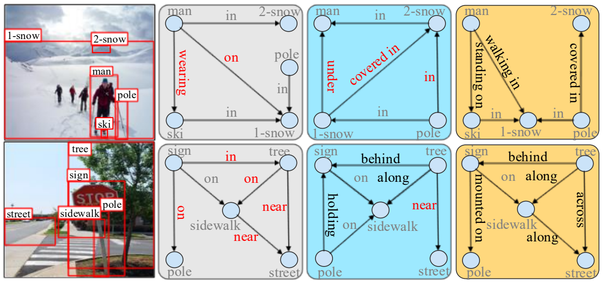


Figure 6. **Visualization results** of Transformer, Transformer (Re-weight), and Transformer (FGPL) on PredCls.

5. Conclusion

In this work, we propose a plug-and-play Fine-Grained Predicates Learning (FGPL) framework for scene graph generation. We devise a Predicate Lattice to help understand predicates correlation concerning all scenarios in the SGG dataset. Based on the Predicate Lattice, we develop a Category Discriminating Loss (CDL) and an Entity Discriminating Loss (EDL), which help differentiate hard-to-distinguish predicates while maintaining learned discriminatory power over recognizable ones. Experiments show that our FGPL can differentiate hard-to-distinguish predicates. When being integrated with our FGPL, several benchmark models achieve superior performance than existing methods, showing the great generability of our FGPL.

Broader Impact. Our research helps reduce the cost of collecting annotations for real-world scenes in applications of scenario understanding. The positive effect of our method on society is in making scenario understanding more efficient for organizations and people. The negative effect of our method on society is that in a deep learning manner, our method is susceptible to adversarial attacks. Therefore, there is a challenge to make people get misunderstood with tampered scenario information. Since training samples come from real-world scenes, it may cause the invasion of personal privacy. Thus, the dataset should be carefully used concerning copyrights and privacy problems. Moreover, the model should be distributed with limitations and regularization in specific scenes. For researchers, we should obey the ethical rules to avoid ethical risks.

Acknowledgement. This work is supported by National Key Research and Development Program of China (No. 2018AAA0102200), the National Natural Science Foundation of China (Grant No. 62122018, No. 61772116, No. 61872064), Sichuan Science and Technology Program (Grant No.2019JDTD0005).

References

[1] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019. 2, 5

[2] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *ICCV*, 2021. 1, 2

[3] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019. 2

[4] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *ICCV*, 2021. 2, 4

[5] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *CVPR*, 2019. 2

[6] Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *ICCV*, 2019. 1

[7] Yuyu Guo et al. From general to specific: Informative scene graph generation via balance adjustment. In *ICCV*, 2021. 1, 2, 5

[8] Yuyu Guo, Jingkuan Song, Lianli Gao, and Heng Tao Shen. One-shot scene graph generation. In *ACMMM*, 2020. 1

[9] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene graph reasoning for visual question answering. *CoRR*, 2020. 1

[10] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 1

[11] Rongjie Li et al. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, 2021. 1, 2, 5, 6

[12] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, 2019. 1

[13] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *ICCV*, 2019. 1

[14] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 2020. 1

[15] L. John Old. An analysis of semantic overlap among english prepositions in roget's thesaurus. In *ACL-SIGSEM*, 2003. 2

[16] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *CVPR*, 2020. 1

[17] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, 2019. 1

[18] Jingkuan Song, Pengpeng Zeng, Lianli Gao, and Heng Tao Shen. From pixels to objects: Cubic visual attention for visual question answering. In *IJCAI*, 2018. 1

[19] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gérard G. Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, 2021. 1

[20] Kaihua Tang. A scene graph generation codebase in pytorch, 2020. https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch. 1, 5, 6

[21] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 2, 5, 6

[22] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 1, 5, 6

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 5, 6

[24] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *CVPR*, 2021. 2, 4, 6, 7

[25] Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, and Ming Tang. Adaptive class suppression loss for long-tail object detection. In *CVPR*, 2021. 2, 4

[26] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021. 1

[27] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 1

[28] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. PCPL: predicate-correlation perception learning for unbiased scene graph generation. In *ACMMM*, 2020. 2, 6

[29] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *CoRR*, 2021. 2

[30] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 1

[31] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *IJCAI*, 2021. 1, 2, 6

[32] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1, 3, 5, 6

[33] Pengpeng Zeng, Lianli Gao, Xinyu Lyu, Shuaiqi Jing, and Jingkuan Song. Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, editors, *ACMMM*, 2021. 1

[34] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In *NeurIPS*, 2019. 2