Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation

Yukuan Min
Xidian University
yukuanmin@gmail.com

Aming Wu*
Xidian University
amwu@xidian.edu.cn

Cheng Deng*
Xidian University
chdeng.xd@gmail.com

Abstract

The scene graph generation (SGG) task is designed to identify the predicates based on the subject-object pairs. However, existing datasets generally include two imbalance cases: one is the class imbalance from the predicted predicates and another is the context imbalance from the given subject-object pairs, which presents significant challenges for SGG. Most existing methods focus on the imbalance of the predicted predicate while ignoring the imbalance of the subject-object pairs, which could not achieve satisfactory results. To address the two imbalance cases, we propose a novel Environment Invariant Curriculum Relation learning (EICR) method, which can be applied in a plug-and-play fashion to existing SGG methods. Concretely, to remove the imbalance of the subject-object pairs, we first construct different distribution environments for the subject-object pairs and learn a model invariant to the environment changes. Then, we construct a class-balanced curriculum learning strategy to balance the different environments to remove the predicate imbalance. Comprehensive experiments conducted on VG and GQA datasets demonstrate that our EICR framework can be taken as a general strategy for various SGG models, and achieve significant improvements.

1. Introduction

Scene graph generation [39] (SGG) aims to predict the corresponding predicate (e.g., riding) based on the given subject-object pairs (e.g., (man, bike)). As an intermediate visual understanding task, it can serve as a fundamental tool for high-level vision and language tasks, such as visual question answering [33, 1, 20], image captioning [4, 13, 49], and cross-model retrieval [11, 30], which promotes the development of visual intelligence.

Though many advances have been achieved [45, 33], SGG is still far from satisfactory for practical applications due to the imbalance phenomenon in the given datasets [32]. To this end, most existing methods focus on addressing the class imbalance from the predicted predicates to generate accurate relation words. Particularly, some works propose resampling [10, 19] and reweighting [38] strategies to balance the head and tail predicate classes, which alleviates the imbalance and improves the performance of SGG.

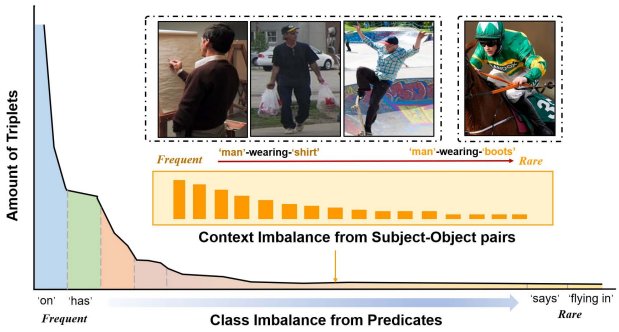


Figure 1. For SGG datasets, besides the class imbalance from predicates, there exists another imbalance phenomenon, i.e., context imbalance from subject-object pairs, which is easily ignored. To this end, this paper delves into class and context imbalance. And a method of Environment-Invariant Curriculum Relation Learning is proposed to generate fine-grained scene graphs effectively.

Besides the class imbalance from the predicted predicate, there exists another context imbalance from the given subject-object pairs, which is prone to be ignored. As shown in Fig. 1, since the predicate prediction relies on the given subject-object context, the number imbalance of the given subject-object pairs easily incorrectly predicts the relation between subjects and objects. For example, there exist a large number of relations between '(man, shirt)' and 'wearing' in the dataset. When giving a rare subject-object pair, e.g., '(man, boots)', the model is prone to generate an incorrect prediction. In Fig. 2 (a), we make an analysis of a popular SGG dataset VG [17]. We observe that the number of context subject-object types will change significantly with the number of predicate categories. Moreover, Fig. 2

*Corresponding author
Codes available at: https://github.com/myukzzz/EICR

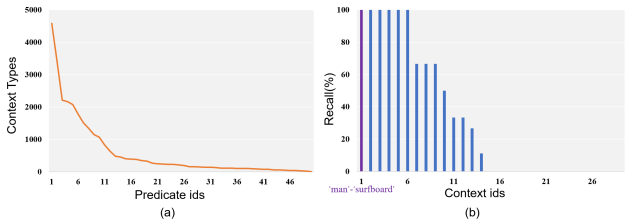


Figure 2. (a) The number of context subject-object types in all predicate classes. (b) The Recall@100 performance for different context subject-object in the predicate class 'carrying'.

(b) further shows that performance can be severely affected by the context subject-object. These phenomena show that the current SGG dataset does have the context imbalance problem, and resolving this problem will help produce fine-grained scene graphs and improve the performance.

To address the problems of the two imbalances mentioned above, we propose a novel framework named Environment Invariant Curriculum Relation learning (EICR), which can be equipped with different baseline models in a plug-and-play applied in a fashion. We construct different distribution environments for the context subject-object and propose a curriculum learning strategy to balance the environments. Specifically, to solve the context imbalance of various subject-object pairs, we construct three different distribution environments: normal, balanced, and over-balanced for the context subject-object pairs, and then apply Invariant Risk Minimization (IRM) [2] to learn a context-unbiased relation classifier that is invariant to these environments. To solve the class imbalance, we utilize a class-balanced curriculum learning strategy to first explore the general patterns from the head predicates in the normal environment and then gradually focus on learning the tail predicates in the over-balanced environment.

Our contributions can be summarized as follows:

(1) Except for the existing class imbalance, we explore and address the under-explored context imbalance problem in the current SGG dataset.

(2) We construct an environment-invariant relation classifier to solve the context imbalance of the subject-object pairs and present a new curriculum learning strategy to consolidate the relation classifier from head to tail predicates and solve the class imbalance of the predicates.

(3) Our EICR can be applied in a plug-and-play fashion for the SGG baselines and get competitive results among various SOTA methods. By applying our proposed method, a VCTree [33] model is improved over **14%** on mR@50/100 and over **12%** on the metric F@50/100.

2. Related Work

Scene Graph Generation. SGG provides an efficient way for connecting vision and language [35, 48], and has drawn widespread attention from the community. Early approaches focus on visual relation detection [26, 9, 22, 21] and are mainly dedicated to incorporating more features from various modalities. To further enhance the relations, considering that relations are highly dependent on their context, different methods [24, 45, 33, 16] are further proposed to refine the object and relation representations in the scene graph. Motifs [45] chose the Bi-LSTM framework for the object and predicate context encoding and VCTree [33] constructs a tree structure to encode the hierarchical and parallel relationships between objects. Moreover, other works also refine the message-passing strategy [24].

Unbiased Scene Graph Generation. Although making steady progress on improving recall on SGG tasks, further research has shown that SGG models are easy to collapse to several general predicate classes because of the long-tail effect in the SGG dataset [5, 14]. For example, from the causal view [42], TDE [32] employs a causal inference framework to eliminate predicate class bias during the inference process. BGNN [19] constructs a bi-level resampling strategy during the training process. Inspired by the application of noisy label learning [36, 37], NICE [18] formulate SGG as an out-of-distribution detection [40, 41] problem and propose a noisy label correction strategy for unbiased SGG. Different from existing SGG works, we are the first to explicitly define and address the context imbalance of the subject-object pairs on SGG datasets.

3. EICR for Class and Context Imbalances

For SGG, this paper aims to address the two different kinds of distribution imbalance, i.e., class imbalance of predicates and context imbalance of subject-object pairs.

3.1. Preliminary

The scene graph generation task aims to generate a summary graph $\mathcal{G}$ for the given image I. Specifically, a scene graph $\mathcal{G} = \{(O, E)\}$ corresponding to I contains a set of target entities $O = \{(o_i)\}_{i=1}^{N_o}$ and a set of relational triplets $E = \left\{\left(o_i, p_{(o_i,o_j)}, o_j\right)\right\}_{i,j=1}^{N_e}$, where $o_i \in O$ and $o_j \in O$, $p_{(o_i,o_j)}$ is defined as the relation between them and belongs to the predefined predicate class set $\mathcal{P}$.

Classifying a relation p as the predicate class c can be preliminarily defined as predicting $P(r = c \mid p)$ based on the dataset of the relations and its label pairs $\{(p, r)\}$ [45]. Using Bayes theorem [3], the predictive model could be decomposed as $P(r = c \mid p) = \frac{P(p|r=c)\cdot P(r=c)}{P(p)}$, where $P(r = c)$ is the class distribution, and $P(p)$ is the marginal distribution of the relations. Previous SGG meth-

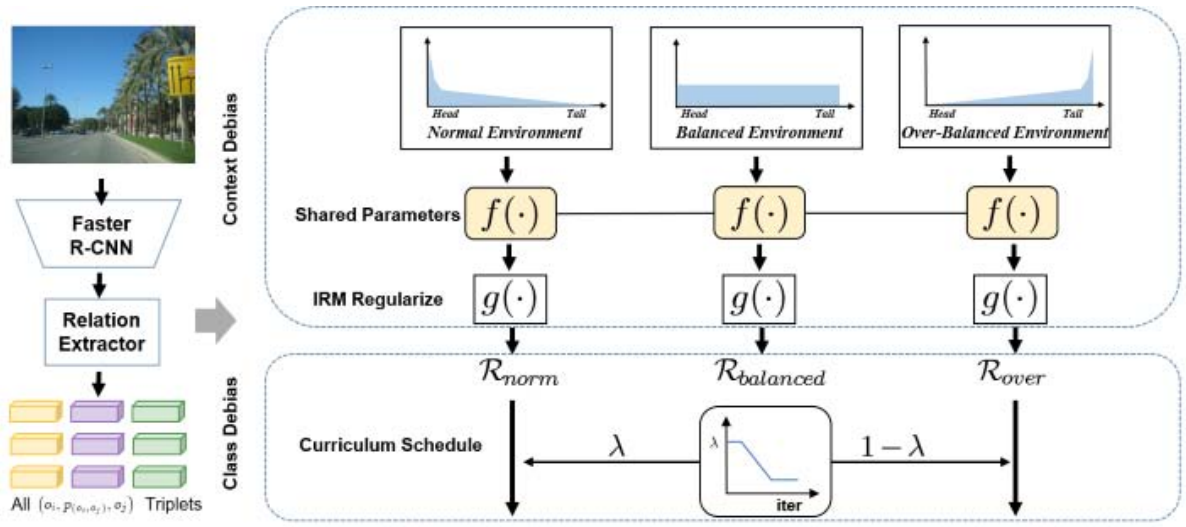


Figure 3. Illustration of our method for alleviating the context and class imbalances in SGG. Firstly, an environment-invariant learning module to build multiple different distribution environments of subject-object pairs, which is beneficial for obtaining an invariant relation classifier and alleviating the context imbalance. Next, a class-balanced curriculum learning strategy is designed to balance the built multiple environments, alleviating the class imbalance.

ods [32, 19] only consider class imbalance $P(r=c)$ while ignoring the context imbalance from the different marginal distribution $P(p)$ based on subject-object pairs.

To explicitly define the relation classification $P(r=c \mid p)$, we assume that a relation p is generated by a set of hidden attributes $z=\{z_1, z_2, z_3, \ldots\}$. Since there exists predicate class imbalance and context imbalance in the existing SGG dataset, we defined two disjoint subsets for the hidden attributes: class-specific attributes z_c (e.g., the frequency of the predicate classes [32]) and context-specific attributes z_e (e.g., the various context subject-object pairs $\{o_i, o_j\}$). Thus we can further decompose the relation prediction model $P\left(r=c \mid z_c, z_e\right)$ as follows:

$$P\left(r=c \mid z_c, z_e\right)=\frac{P\left(z_c \mid r=c\right)}{P\left(z_c, z_e\right)} \cdot \underbrace{P\left(z_e \mid r=c, z_c\right)}_{\text{context imbalance}} \cdot \underbrace{P(r=c)}_{\text{class imbalance}} . \tag{1}$$

From Eq. 1, the relation classifier $P\left(r=c \mid z_c, z_e\right)$ is affected by two imbalances:

Class Imbalance: in previous SGG works [19, 32, 38], the distribution of $P(r=c)$ is considered as the main cause of the performance degradation. As $P(r=c)$ can be explicitly calculated from the training data, the majority of previous methods directly alleviate its effect by class-wise adjustment [38] or re-sampling [19, 12].

Context Imbalance: we argue that the context imbalance suffers the relation classifier in two ways. First, as shown in Fig. 2 (a), the different diversity for the context makes imbalanced influences for the predicate classes. Second, Fig. 2 (b) shows that the predicate is high-related to certain context subject-object pairs. These phenomenons will create spurious correlations between the subject-object pairs and the predicates, which will weaken the relation classification performance, especially in the tail predicates whose context subject-object pairs are rare due to sampling scarcity. Since we have concluded that the context subject-object pair is highly related to the predicate class, we formulate the context imbalance as $P\left(z_e \mid r=c, z_c\right)$.

To solve these two imbalances and obtain unbiased relations for the scene graphs, we propose the following Environment Invariant Curriculum Relation learning (EICR) framework to learn an unbiased relation classifier invariant to the change of the various predicate classes and context subject-object pairs.

3.2. Environment Invariant Learning

To alleviate the context imbalance, we need to eliminate the impact of the contexts for the relation classification $P\left(z_e \mid r=c, z_c\right)$. To this end, based on the theory of Invariant Risk Minimization (IRM) [2], we can first construct a set of environments $\mathcal{E}=\{e_1, e_2, \ldots\}$ with diverse context distribution. Then, by regularizing the relation classifier $g(\cdot)$ to be equally optimal across the environments with different context distributions, we can alleviate the influence of the context. Thus objective function can be defined as follows:

$$\begin{aligned} &\min_{g} \sum_{e \in \mathcal{E}} R^e(I, r; f(\cdot), g(\cdot)), \\ &\text{subject to } g \in \arg\min_{g} R^e \text{ for all } e \in \mathcal{E}, \end{aligned} \tag{2}$$

where $R^e(I, r; f(\cdot), g(\cdot))$ is the risk under environment e (i.e., the loss for training), $f(\cdot)$ is the relation feature extractor, $g \in \arg\min_g R^e(I, r; f(\cdot), g(\cdot))$ for all $e \in \mathcal{E}$ means that the invariant identifier g should minimize the risk under all environments simultaneously. Following IRM [2], we use a gradient norm penalty term to minimize g at each environment, i.e., $\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}} R^e(\Phi) + \lambda \cdot \left\|\nabla_{w|w=1.0} R^e(w \cdot \Phi)\right\|^2$,where Φ is the invariant model, $R(\cdot)$ denotes the training loss under different environments $e \in \mathcal{E}$ and we set $\lambda = 1$. The detailed process of the environment construction is introduced below:

The set of diverse environments should ensure the variance of the context influence and ideally are orthogonal distributions [34, 43]. However, considering the computation consumption and feasibility of the strategy, to construct different $P(z_e \mid r = c, z_c)$, it is hard to change the $r = c$ since the relation labels are predefined. Thus we construct three learning environments with different z_c, i.e., the frequency of the predicate classes. As illustrated in Fig. 3, each learning environment constructs different frequencies of the predicate classes:

• The normal environment maintains the raw distribution of the predicate classes in the dataset.

• The class-balanced environment constructs the resampling strategy [12] to sampling in all predicate categories at balanced frequencies. Specifically, we first calculate the median amount of the samples over all predicate classes $Median(r)$. Then, for each predicate class r_i with n_i samples, we calculate the sampling rate s_i as follows:

$$s_i = \begin{cases} \frac{Median(r)}{n_i} & \text{if } Median(r) \le n_i, \\ 1 & \text{if } Median(r) > n_i. \end{cases} \tag{3}$$

• The over-balanced environment is constructed to overcorrect the imbalanced predicate class distribution $P(r)$. Thus we first construct a resampling strategy for the balanced sampling as in the class-balanced environment. Then we adopt an extra reweighting strategy for over-balanced weighting, the loss can be formulated as:

$$L_{over} = -\sum_{i=1}^{C} w_i r_i \log(g(f(I))), \tag{4}$$

where $w_i = 1/n_i$ and C denotes the number of the predicate categories. This environment deliberately picks relation triplets with the probability negatively correlated with predicate class size.

3.3. Class-Balanced Curriculum Learning

After obtaining a context-unbiased relation representation from the environment learning module, we assume the network has already modeled the $P(z_e \mid r = c, z_c)$. Therefore, we only need to tackle the class imbalance $P(r = c)$ in the context-balanced SGG data. We devise a curriculum schedule for environment learning to make the relation prediction model successfully explore general patterns from head predicates and then gradually focus on the tail predicates. Specifically, we adjust the learning weights between the over-balanced environment and the normal environment by a trade-off factor λ which is defined as:

$$\lambda = \begin{cases} \lambda_{max} & \text{if } t \le T, \\ \max(H(t), \lambda_{min}) & \text{if } T < t \le 2T, \\ \lambda_{min} & \text{if } t > 2T, \end{cases} \tag{5}$$

where t is the current training iteration, T is predefined intermediate training iterations for different stages of curriculum learning. $\lambda_{min}, \lambda_{max} \in [0, 1]$ are hyper-parameters. In order to ensure the scale invariance, $\lambda_{min} + \lambda_{max} = 1$. $H(t)$ is a curriculum schedule function decreasing from 1 to 0 with the input iteration t, which can be defined as:

$$H(t) = \frac{2T - t}{T}(\lambda_{max} - \lambda_{min}), \tag{6}$$

thus the joint loss function for the three environments can be formulated as:

$$\mathcal{R}_{\text{hybird}} = \lambda \cdot \mathcal{R}_{norm} + (1 - \lambda) \cdot \mathcal{R}_{over} + \mathcal{R}_{balanced}, \tag{7}$$

where $\mathcal{R}_{norm}, \mathcal{R}_{over}, \mathcal{R}_{balanced}$ are the risks under normal, class-balanced, and over-balanced environments. The class-balanced curriculum learning strategy thus can be divided into three phases. In the first training phase ($t \le T$), the model is mainly focused on the normal environment to learn the general patterns from head predicates. In the second phase ($T < t \le 2T$), λ gradually decreases during the training. The model's learning focus shifts from the normal environment to the over-balanced environment to incrementally learn the fine-grained tail predicates while retaining the general patterns. In the third phase ($t > 2T$),

Algorithm 1 EICR Framework

Input: SGG Dataset $\{(I, r)\}$, # Iteration T.

Initialize the pretrained relation feature extractor f and relation classifier g
while $t \le T$ **do**
  // *Context-Debias*
  Generate multiple environments $\{e_1, e_2, e_3\}$
  Learn parameters of g through IRM with Eq. 2
  // *Class-Debias*
  Reweight environments by schedule in Eq. 5
  Update the model by balanced risks from Eq. 7.
end while

Output: The debiased relation feature extractor f and relation classifier g

Method	PredCls			SGCls			SGDet		
	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100
IMP [35]	61.1 / 63.1	11.0 / 11.8	18.6 / 19.9	37.4 / 38.3	6.4 / 6.7	10.9 / 11.4	23.6 / 28.7	3.3 / 4.1	5.8 / 7.2
GPS-Net [25]	65.2 / 67.1	15.2 / 16.6	24.7 / 26.6	37.8 / 39.2	8.5 / 9.1	13.9 / 14.8	31.1 / 35.9	6.7 / 8.6	18.9 / 22.3
BGNN [19]	59.2 / 61.3	30.4 / 32.9	40.2 / 42.8	37.4 / 38.5	14.3 / 16.5	20.7 / 23.1	31.0 / 35.8	10.7 / 12.6	15.9 / 18.6
DT2-ACBS [10]	23.3 / 25.6	35.9 / 39.7	28.3 / 31.1	16.2 / 17.6	24.8 / 27.5	19.6 / 21.5	15.0 / 16.3	22.0 / 24.0	17.8 / 19.4
SHA-GCL [12]	35.1 / 37.2	41.6 / 44.1	38.1 / 40.4	22.8 / 23.9	23.0 / 24.3	22.9 / 24.1	14.9 / 18.2	17.9 / 20.9	16.3 / 19.5
Motifs [45]	65.2 / 67.0	14.8 / 16.1	24.1 / 26.0	38.9 / 39.8	8.3 / 8.8	13.7 / 14.8	31.1 / 35.9	6.7 / 8.6	11.0 / 13.9
+ TDE [32]	46.2 / 51.4	25.5 / 29.1	32.9 / 37.2	27.7 / 29.9	13.1 / 14.9	17.8 / 19.9	16.9 / 20.3	8.2 / 9.8	11.0 / 13.2
+ PCPL [38]	54.7 / 56.5	24.3 / 26.1	33.7 / 35.7	35.3 / 36.1	12.0 / 12.7	17.9 / 18.8	27.8 / 31.7	10.7 / 12.6	15.5 / 18.0
+ EBM [31]	65.2 / 67.3	18.0 / 19.5	28.2 / 30.2	39.2 / 40.0	10.2 / 11.0	16.2 / 17.3	31.7 / 36.3	7.7 / 9.3	12.4 / 14.8
+ NICE [18]	55.1 / 57.1	29.9 / 32.3	38.8 / 41.3	33.1 / 34.0	16.6 / 17.9	22.1 / 23.5	27.8 / 31.8	12.2 / 14.4	17.0 / 19.8
+ IETrans [46]	48.6 / 50.5	35.8 / 39.1	41.2 / 44.1	29.4 / 30.2	21.5 / 22.8	24.8 / 26.0	23.5 / 27.2	15.5 / 18.0	18.7 / 21.7
+ EICR	55.3 / 57.4	34.9 / 37.0	**42.8 / 45.0**	34.5 / 35.4	20.8 / 21.8	**25.9 / 27.0**	27.9 / 32.2	**15.5 / 18.2**	**19.9 / 23.3**
VCTree [33]	65.4 / 67.2	16.7 / 18.2	26.6 / 28.6	46.7 / 47.6	11.8 / 12.5	18.8 / 19.8	31.9 / 36.2	7.4 / 8.7	12.0 / 14.0
+ TDE [32]	47.2 / 51.6	25.4 / 28.7	33.0 / 36.9	25.4 / 27.9	12.2 / 14.0	16.5 / 18.6	19.4 / 23.2	9.3 / 11.1	12.6 / 15.1
+ PCPL [38]	56.9 / 58.7	22.8 / 24.5	32.6 / 34.6	40.6 / 41.7	15.2 / 16.1	22.1 / 23.2	19.4 / 23.2	9.3 / 11.1	12.6 / 15.0
+ EBM [31]	64.0 / 65.8	18.2 / 19.7	28.3 / 30.3	44.7 / 45.8	12.5 / 13.5	19.5 / 20.9	31.4 / 35.9	7.7 / 9.1	12.4 / 14.5
+ NICE [18]	55.0 / 56.9	30.7 / 33.0	39.4 / 41.8	37.8 / 39.0	19.9 / 21.3	26.1 / 27.6	27.0 / 30.8	11.9 / 14.1	16.5 / 19.3
+ IETrans [46]	48.0 / 49.9	37.0 / 39.7	41.8 / 44.2	30.0 / 30.9	19.9 / 21.8	23.9 / 25.6	23.6 / 27.8	12.0 / 14.9	15.9 / 19.4
+ EICR	56.0 / 57.9	35.6 / 37.9	**43.6/ 45.8**	39.4 / 40.5	**26.2 / 27.4**	**32.8 / 33.9**	26.0 / 30.1	**15.2 / 17.5**	**19.2 / 22.1**
Transformer [32]	63.6 / 65.7	19.7 / 19.6	27.9 / 30.2	38.1 / 39.2	9.9 / 10.5	15.7 / 16.6	30.0 / 34.3	7.4 / 8.8	11.9 / 14.0
+ CogTree [44]	38.4 / 39.7	28.4 / 31.0	32.7 / 34.8	22.9 / 23.4	15.7 / 16.7	18.6 / 19.5	19.5 / 21.7	11.1 / 12.7	14.1 / 16.0
+ IETrans [46]	49.0 / 50.8	35.0 / 38.0	40.8/ 43.5	29.6 / 30.5	20.8 / 22.3	24.4 / 25.8	23.1 / 27.1	15.0 / 18.1	18.2 / 21.7
+ EICR	52.8 / 54.7	**36.9 / 39.1**	**43.5/ 45.6**	31.4 / 32.4	**21.6 / 22.4**	**25.6 / 26.5**	23.7 / 27.7	**17.3 / 19.7**	**20.0 / 23.0**

Table 1. Performance (%) of our method and other baselines on VG dataset. + EICR denotes different models equipped with our EICR.

Method	PredCls			SGCls			SGDet		
	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100
Motifs [45]	65.3 / 66.8	16.4 / 17.1	26.2 / 27.2	34.2 / 34.9	8.2 / 8.6	13.2 / 13.8	28.9 / 33.1	6.4 / 7.7	10.5 / 12.5
+ GCL [12]	44.5 / 46.2	36.7 / 38.1	40.2 / 41.8	23.2 / 24.0	17.3 / 18.1	19.8 / 20.6	18.5 / 21.8	16.8 / 18.8	17.6 / 20.2
+ EICR	56.4 / 58.1	36.3 / 38.0	**44.2 / 46.0**	28.8 / 29.4	17.2 / **18.2**	**21.5 / 22.5**	24.6 / 28.4	16.0 / 18.0	**19.4 / 22.0**
VCTree [33]	63.8 / 65.7	16.6 / 17.4	26.3 / 27.5	34.1 / 34.8	7.9 / 8.3	12.8 / 13.4	28.3 / 31.9	6.5 / 7.4	10.6 / 13.2
+ GCL [12]	44.8 / 46.6	35.4 / 36.7	39.5 / 41.1	23.7 / 24.5	17.3 / 18.0	20.0 / 20.8	17.6 / 20.7	15.6 / 17.8	16.5 / 19.1
+ EICR	55.3 / 57.0	**35.9 / 37.4**	**43.5 / 45.2**	28.4 / 29.1	**17.8 / 18.6**	**21.9 / 22.7**	24.0 / 27.6	14.7 / 16.3	**18.2 / 20.5**

Table 2. Performance comparison of different methods on three tasks of GQA dataset

the model avoids overfitting the general patterns from the normal environment when focusing on the tail predicates at later training periods. Algorithm. 1 shows details of EICR.

4. Experiments

In this section, we first show the generalizability of our method with different baseline models and the expansibility to different SGG datasets. Ablation studies are also constructed to explore the influence of different modules and hyperparameters. Finally, we conduct several analyses to show the effectiveness of our method in solving both the context imbalance and the class imbalance.

4.1. Experimental Settings

Dataset. In the SGG task, we choose Visual Genome (VG) [17] dataset which comprises 75k object categories and 40k predicate categories. We applied the widely accepted benchmark [45, 33, 32, 47, 28], using the 150 highest frequency objects categories and 50 predicate categories. GQA [15] is another dataset for vision-language tasks with more than 3.8M relation annotations. Following previous work [12], we select Top-200 object classes as well as Top-100 predicate classes by their frequency for the GQA200 benchmark. For both datasets, the training set is set to be 70%, and the testing set is 30%, with 5k images from the training set for validation [45].

Tasks. Following previous works [45, 33, 12, 6], we evaluate our model on three widely used SGG tasks: (1) Predicate Classification (PredCls): given images, object bounding boxes, and object labels, models are required to recognize predicate classes. (2) Scene Graph Classification (SGCls): gives images and object bounding boxes and asks models to predict object labels and relationship labels between objects. (3) Scene Graph Detection (SGDet): models are required to localize objects, recognize objects, and predict their relationships directly from images.

Metrics. Following previous works [19, 38], we use Recall@K (R@K) and mean Recall@K (mR@K) as our metrics. Moreover, inspired by previous work [46], we use the overall metric F@K to jointly evaluate R@K and mR@K, which is the harmonic average of R@K and mR@K.

Method	PredCls			SGCls			SGDet		
	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100
Motifs [45]	65.2 / 67.0	14.8 / 16.1	24.1 / 26.0	38.9 / 39.8	8.3 / 8.8	13.7 / 14.8	31.1 / 35.9	6.7 / 8.6	11.0 / 13.9
+ TDE [32]	46.2 / 51.4	25.5 / 29.1	32.9 / 37.2	27.7 / 29.9	13.1 / 14.9	17.8 / 19.9	16.9 / 20.3	8.2 / 9.8	11.0 / 13.2
+ EIL	64.1 / 65.8	24.5 / 26.5	**35.5 / 37.8**	39.3 / 40.1	**15.4 / 16.1**	**22.1 / 23.0**	32.2 / 36.8	**10.6 / 12.6**	**15.9 / 18.7**
VCTree [33]	65.4 / 67.2	16.7 / 18.2	26.6 / 28.6	46.7 / 47.6	11.8 / 12.5	18.8 / 19.8	31.9 / 36.2	7.4 / 8.7	12.0 / 14.0
+ TDE [32]	47.2 / 51.6	25.4 / 28.7	33.0 / 36.9	25.4 / 27.9	12.2 / 14.0	16.5 / 18.6	19.4 / 23.2	9.3 / 11.1	12.6 / 15.1
+ EIL	64.5 / 66.5	22.8 / 24.3	33.7/ 35.6	45.9 / 46.9	**17.8 / 18.9**	**25.6 / 26.9**	31.2 / 35.5	**10.6 / 12.4**	**15.9 / 18.3**
Transformer [32]	63.6 / 65.7	19.7 / 19.6	27.9 / 30.2	38.1 / 39.2	9.9 / 10.5	15.7 / 16.6	30.0 / 34.3	7.4 / 8.8	11.9 / 14.0
+ CogTree [44]	38.4 / 39.7	28.4 / 31.0	32.7 / 34.8	22.9 / 23.4	15.7 / 16.7	18.6 / 19.5	19.5 / 21.7	11.1 / 12.7	14.1 / 16.0
+ EIL	63.3 / 65.0	27.7 / 29.8	**38.6/ 40.9**	38.2 / 40.0	**15.7** / 16.5	**22.3 / 23.2**	31.7 / 36.1	**12.7 / 14.9**	**18.1 / 21.0**

Table 3. Ablation study of the Environment-Invariant Learning (EIL) module on VG dataset.

Environments			SGCls		
Normal	Balanced	Over-Balanced	R@50 / 100	mR@50 / 100	F@50 / 100
			38.9 / 39.8	8.3 / 8.8	13.7 / 14.8
	✓	✓	21.5 / 22.6	22.1 / 23.4	21.8 / 23.0
✓		✓	39.8 / 40.6	13.7 / 14.8	20.3 / 21.6
✓	✓		39.7 / 40.5	13.3 / 14.0	19.9 / 20.8
✓	✓	✓	39.3 / 40.1	15.4 / 16.1	**22.1 / 23.0**

Table 4. Ablation study of constructing different environments on VG dataset.

Model	SGCls		
	R@50 / 100	mR@50 / 100	F@50 / 100
w/o-Curriculum Schedule	39.3 / 40.1	15.4 / 16.1	22.1 / 23.0
w/o-Norm Schedule	39.2 / 40.0	14.8 / 15.8	21.5 / 22.7
w/o-Over Schedule	35.6 / 36.4	18.1 / 19.1	24.0 / 25.1
w-Curriculum Schedule	34.5 / 35.4	**20.8 / 21.8**	**25.9 / 27.0**

Table 5. Ablation study for curriculum learning strategy on VG dataset.

Setting	PredCls		
	R@50 / 100	mR@50 / 100	F@50 / 100
w/o IRM term	52.4 / 54.4	35.4 / 37.4	42.2 / 44.3
w/ IRM term	**55.3 / 57.4**	34.9 / 37.0	**42.8 / 45.0**

Table 6. Ablation study of the IRM term.

Implementation Details. We employ a pre-trained Faster-RCNN [29] with ResNeXt-101-FPN [23] provided by [32] as the object detector. We use Glove [27] to obtain the semantic embedding. In the training process, the parameters of the detector are fixed to reduce the computation cost. The hyper-parameter lambda which balances the different environments is set to 0.9. We optimize all models with an Adam optimizer with a momentum of 0.9. The batch size is set to 4, and the total training stage lasts for 120,000 steps with $T = 30000$ and $\lambda_{max} = 0.9$. The initial learning rate is 0.001, and we adopt the same warm-up and decayed strategy as [12]. One RTX2080 Ti is used to conduct all the experiments.

4.2. Compared Methods

We demonstrate the effectiveness of our method by comparing the results with current SOTA methods and validate its generalizability with different baseline models. On the one hand, to prove its performance, we select some dedicated designed SGG models with state-of-the-art performance, including re-produced IMP [35], GPS-Net [25], DT2-ACBS [10], SHA-GCL [12], and BGNN [19]. On the other hand, to demonstrate the generalizability of our EICR, we compare our method with the model-agnostic baselines which can be applied in a plug-and-play fashion, including TDE [32], CogTree [44], PCPL [38], EBM [31], DLFE [7], GCL [12], NICE [18] and IETrans [46].

λ_{max}	SGCls		
	R@50 / 100	mR@50 / 100	F@50 / 100
w/o-λ_{max}	39.3 / 40.1	15.4 / 16.1	22.1 / 23.0
0.7	37.1 / 38.0	18.0 / 19.0	24.3 / 25.2
0.8	36.2 / 37.1	19.1 / 20.0	25.0 / 26.0
0.87	34.8 / 35.6	18.9 / 19.7	24.5 / 25.4
0.9	34.5 / 35.4	20.8 / 21.8	25.9 / 27.0
0.92	32.9 / 33.8	21.1 / 22.1	25.7 / 26.7
0.95	33.3 / 34.2	20.0 / 21.1	25.0 / 26.1
0.99	27.6 / 28.7	20.9 / 21.9	23.8 / 24.8

Table 7. Parameter analysis towards λ_{max} on VG dataset.

4.3. Main Results

We report the results of the proposed EICR framework and baselines on the VG and GQA datasets in Table 1 and Table 2. From the results of various tasks and baselines, we have several observations as follows:

On the one hand, our EICR is adaptive to different baselines. We adapt our method to 3 popular baselines for SGG, including Motifs [45], VCTree [33], and Transformer [32]. These baselines include various architectures such as conventional LSTM (Motifs), tree structure (VCTree), and self-attention layers (Transformer). Various training algorithms are also contained such as supervised training and reinforcement learning (VCTree). Specifically, our method can boost all models' mR@50/100 metric and the overall F@50/100 metric. With our method, the results for VCTree are improved over 14% on mR@50/100 and improved over 12% across all 3 tasks on the metric F@50/100. Moreover, com-

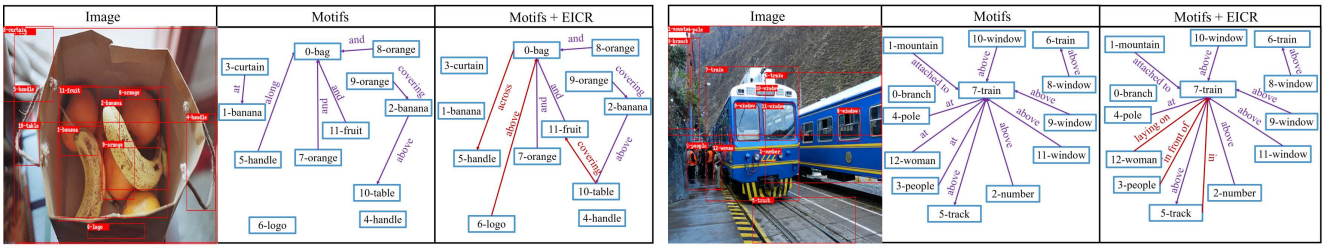
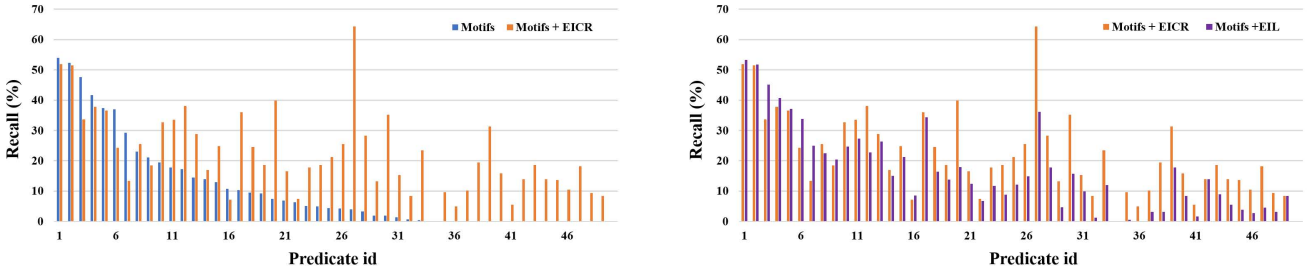


Figure 4. Visualization scene graphs between Motifs and Motifs + EICR with regard to R@20 on PredCls setting. Purple edges represent the reasonable relationships predicted by Motifs. Red edges represent the refined reasonable relationships which are predicted by Motifs + EICR but failed to be detected by Motifs.



(a) R@100 of all the predicate classes of Motifs and Motifs + EICR on VG

(b) R@100 of all the predicate classes of Motifs + EIL and Motifs + EICR on VG

Figure 5. R@100 of 50 predicate classes on SGCls on the VG dataset.

pared with other model-agnostic methods, our method outperforms all of them on the F@50/100 and gains competitive results on mR@50/100. By applying our methods to VCTree and Transformer on SGCls and SGDet, our model can achieve the highest R@50/100 and mR@50/100 among all model-agnostic baselines.

On the other hand, compared with strong specific baselines, our method can also achieve competitive performance on mR@50/100 and the best overall performance on F@50/100. Our method with VCTree is close to the SOTA results in DT2-ACBS on SGCls and SGDet tasks on mR@50/100 while outperforming much better than them on R@50/100. For an overall comparison of the F@50/100 metrics, our method with VCTree can achieve the best F@50/100 on PredCls and SGCls and our method with Motif achieves the best F@50/100 in the SGDet task.

4.4. Ablation Studies

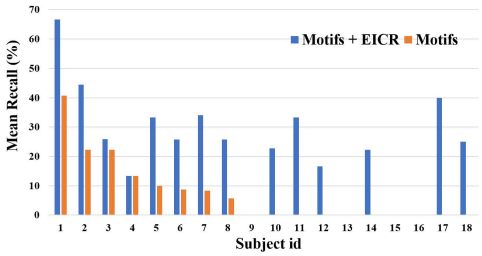
In this part, we analyze the influence of the environment-invariant learning, curriculum learning strategy, and corresponding parameter λ_{max}.

Influence of Environment-Invariant Learning (EIL). Table 3 and Table 4 present the results of all the ablation models. As shown in Table 3, only using environment-invariant learning is hard to boost the mR@50/100 and F@50/100 performance as much as EICR. The reason is that the training procedure is still led by the normal environments and overfitting the corresponding general patterns. However, though the performance on the mR@50/100 and F@50/100 are not boosted so much, the EIL retains the performance on the R@50/100 compared with the original baselines (Motifs, VCTree, Transformer). We can conclude that by introducing EIL to cope with the context imbalance problem, the model learns the context-unbiased relation representation and make a gain on the mR@50/100 metric while retaining the general patterns without the decrease on the R@50/100 metric. As shown in Table 4, different settings of learning environments all achieve improvements on mR@50/100 and F@50/100 compared with Motifs. However, its performance is poor compared with EIL, which shows the importance of combing multiple environments by EIL. Integration of multiple learning environments can alleviate the context imbalance and improve the performance on SGG benchmarks.

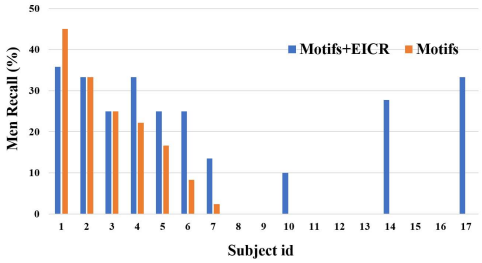
Influence of Curriculum Learning Strategy. As aforementioned, we propose the class-balanced curriculum learning strategy to alleviate the class imbalance. In order to prove the effectiveness of the above components, we test various ablation models on the VG dataset as follows:

(1) w/o-Curriculum Schedule: To evaluate the effectiveness of the curriculum schedule, we do not use curriculum schedule, i.e., $\mathcal{R}_{\text{hybird}} = \mathcal{R}_{norm} + \mathcal{R}_{over} + \mathcal{R}_{balanced}$.

(2) w/o-Norm Schedule: To evaluate the effectiveness of the changing weight of the normal environment, we remove the curriculum schedule for the normal environment risk and only employ the curriculum schedule for the over-



(a) mR@100 of the subject categories of Motifs and Motifs + EICR inside class 'looking at' on VG



(b) mR@100 of the subject categories of Motifs and Motifs + EICR inside class 'over' on VG

Figure 6. mR@100 of various subject categories inside the predicate classes on the VG dataset.

balanced environment, i.e., $\mathcal{R}_{\text{hybird}} = \mathcal{R}_{norm} + (1-\lambda) \cdot \mathcal{R}_{over} + \mathcal{R}_{balanced}$.

(3) w/o-Over Schedule: To evaluate the effectiveness of the curriculum schedule for the over-balanced environment, we remove the curriculum schedule for the over-balanced environment, i.e., $\mathcal{R}_{\text{hybird}} = \lambda \cdot \mathcal{R}_{norm} + \mathcal{R}_{over} + \mathcal{R}_{balanced}$.

Table 5 presents the results of all the ablation models. First, the curriculum schedule can achieve a huge improvement on the mR@50/100 and F@50/100 metrics. Compared with w/o-curriculum schedule, w-curriculum schedule boosts the mR@50/100 metric by over 5 points and improves by nearly 4 points on the F@50/100. Second, we witness an obvious performance decay when removing the curriculum schedule either for the normal environment or the over-balanced environment. It verifies that constructing curriculum learning schedules for multiple environments would effectively alleviate the class imbalance in the SGG dataset, thus leading to class-unbiased relation predictions.

Influence of IRM regularization. We take Motifs [45] as the baseline model. As shown in Table 3, we can see that adding the IRM regularization term improves the performance of R@50/100 and F@50/100, demonstrating that the IRM regularization enhances the representation ability of the predicate predictor.

Influence of λ_{max}. As shown in Table 7, the mR@50/100 metric significantly increases with the increase of the λ_{max} while the R@50/100 metric decreases at the same time. Since the over-balanced environment is highly related to the samples from the tail predicates, the increase of the λ_{max} can somewhat be considered as increasing the model's attention to the tail samples. Thus, the phenomenon indicates that the conventional structures of the SGG model (Motifs, VCTree, Transformer) may easily classify tail classes as negative samples and lead to low results on mR@50/100, while this part of the data is of vital significance for improving the model's ability to make class-unbiased predictions.

Method	PredCls		
	R@50 / 100	mR@50 / 100	F@50 / 100
BBN [50]	56.0 / 57.7	19.4 / 21.3	28.8 / 31.1
Reweight [8]	54.7 / 56.5	17.3 / 18.6	26.3 / 28.0
EICR	55.3 / 57.4	**34.9 / 37.0**	**42.8 / 45.0**

Table 8. Related class-balancing strategies on VG dataset.

Model	PredCls	SGCls	SGDet
	mT@50 / 100	mT@50 / 100	mT@50 / 100
Motifs	7.9 / 8.8	3.1 / 3.4	2.0 / 2.4
+ EICR	17.8 / 19.2	8.3 / 8.9	5.8 / 6.6
VCTree	8.4 / 9.3	4.3 / 4.8	1.7 / 2.1
+ EICR	18.3 / 19.7	11.6 / 12.4	5.8 / 6.7
Transformer	9.6 / 10.6	3.3 / 3.7	2.4 / 2.9
+ EICR	18.8 / 20.3	8.9 / 9.4	6.7 / 7.6

Table 9. Performance of balancing the contexts of our EICR method on VG dataset.

4.5. Qualitative Studies

To get an intuitive perception of the superior performance on the SGG tasks of our proposed method, we make quantitative studies.

Visualization. To show the potential of our method for real-world application, we visualize several PredCls examples generated from the biased Motifs and the unbiased Motifs + EICR. As shown in Fig. 4, we can observe that our method can help to generate more various relation predicates while keeping faithful to the image content. The model prefers to provide more specific relationship predictions (e.g., 'covering' and 'in front of') rather than common and trivial ones (e.g., 'at' and 'along'). Moreover, our method could also help capture potential reasonable relationships. For example, in Fig. 4, our method captures 'logo-on-bag' in the left example and 'track-in-train' in the right example. In a nutshell, the proposed method could enhance the unbiased scene graph generation and generate more informative relation triplets to support various downstream tasks.

Detailed Results. In Fig. 5 (a), we show the detailed results of comparing Motifs and Motifs + EICR with respect to R@100 of all the predicate classes on the SGCls task. We can observe that by applying our methods to the Motifs baseline, though there exists an acceptable decay on the minority of several head predicate classes, the performance on most of the predicate classes is obviously improved. More-

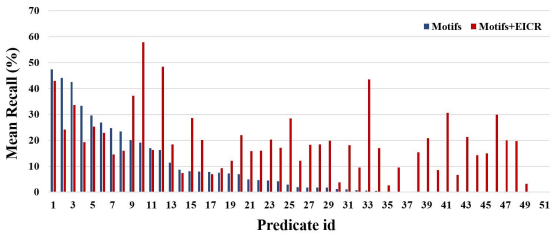


Figure 7. mR@100 of various triplets with different contexts inside all predicate classes of Motifs and Motifs + EICR on VG.

T	SGCls		
	R@50 / 100	mR@50 / 100	F@50 / 100
10000	34.3 / 35.1	19.9 / 20.9	25.2 / 26.2
20000	34.3 / 35.1	20.8 / 21.6	25.9 / 26.8
30000	34.5 / 35.4	20.8 / 21.8	25.9 / 27.0
40000	34.9 / 35.8	19.0 / 19.9	24.6 / 25.6

Table 10. Parameter analysis towards T on VG dataset.

over, we also compare Motifs + EIL (i.e., w/o curriculum schedule) and Motifs + EICR on the detailed performance towards every predicate class on VG. As shown in Fig. 5 (b), the curriculum schedule effectively prevents the model from overfitting the general pattern on the head predicate classes and achieves a better performance towards the tail predicate predictions. It demonstrates that the curriculum schedule could achieve a reasonable trade-off between the environments, and effectively alleviate the class imbalance of the predicates to get a class-unbiased relation classifier.

4.6. Further Analysis

Verification for Alleviating the Context Imbalance. To make a further analysis, we verify the effectiveness of our EICR for alleviating the context imbalance between the various subject-object pairs. Specifically, we calculate the mR@100 of all the different subject categories inside the same predicate class. i.e., the mR@100 for the subject 'man' is calculated by the mean R@100 of the relation triplets with the subject 'man' such as 'man-wearing-shirt' and 'man-wearing-boots'. Two examples on the VG dataset are shown in Fig. 6. We can see that compared with the Motifs, the EICR help to make a more balanced distribution for the various subject-object pairs thus gaining context-unbiased results.

Discussions with Relevant Long-Tailed Approaches. To demonstrate the effectiveness of our EICR strategy in alleviating the class imbalance, we compare our method with two other relevant typical class-balancing strategies on the Motifs baseline, i.e., the resampling strategy BBN [50] and the reweighting strategy [8] following SHA [12]. As shown in Table 8, we can see that our method achieves the best performance. We can see that compared with the Motifs, our EICR model could help to make a more balanced distribution for the various predicate classes thus gaining better results on the various SGG datasets.

Verification for Balancing Contexts. To provide a more detailed analysis of our method's effectiveness in alleviating the context imbalance, we report the metric mT@50/100 on the VG dataset. mT@50/100 denotes the average of the mean Recall for various triplets (i.e., the same predicate with different subject-object context) inside each predicate class. As shown in Table 9, our EICR can be applied in a plug-and-play fashion for solving the context imbalance. By adding our EICR to the three baselines, the results are significantly improved across all 3 tasks on the metric mT@50/100. Moreover, in Fig. 7, we show the detailed results of comparing Motifs and Motifs + EICR with respect to mR@100 of the triplets inside all predicate classes on the PredCls task. With our method, the performance of the triplets inside most of the predicate classes is obviously improved. It demonstrates that our method could achieve a reasonable trade-off for the existing imbalance contexts between the predicate classes, and effectively alleviate the context imbalance.

Influence of T. To provide a more detailed analysis of the influence of the curriculum learning module, we report the performance with different T. As shown in Table 10, with the increase of the intermediate training iterations T, the mR@50/100 metric and the overall metric F@50/100 first increases and then decreases. The phenomenon shows that blindly focusing on the tail predicates does not necessarily mean higher performance on the various SGG datasets.

5. Conclusions

In this paper, we design a method named EICR for fine-grained scene graph generation. We were motivated by the observation that there not only exists the class imbalance between predicate classes, but also the context imbalance for various subject-object pairs. The proposed EICR consists of two debias modules to learn a robust relation classifier unbiased to the various class and contexts. Comprehensive experiments show the effectiveness of our method. In the future, we will further analyze more effective methods to alleviate the context imbalance and explore our theory in other visual recognition problems (e.g., image classification) with similar challenges.

Acknowledgement. Our work was supported by Joint Fund of Ministry of Education of China (8091B022149). Key Research and Development Program of Shanxi (2021ZDLGY01-03). National Natural Science Foundation of China (62102293, 6213201662171343, and 62071361) and Fundamental Research Funds for the Central Universities (ZDRC2102).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

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

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

[6] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021.

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

[8] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021.

[9] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, pages 3076–3086, 2017.

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

[11] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5213–5222, 2020.

[12] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022.

[13] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332, 2019.

[14] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16383–16392, 2021.

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[16] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3820–3832, 2020.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

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

[19] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021.

[20] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937, 2022.

[21] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020.

[25] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020.

[26] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[30] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 24:2914–2923, 2021.

[31] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[32] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[33] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[34] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021.

[35] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[36] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE Transactions on Image Processing*, 32:1245–1256, 2023.

[37] Jiexi Yan, Lei Luo, Chenghao Xu, Cheng Deng, and Heng Huang. Noise is also useful: Negative correlation-steered latent contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–40, 2022.

[38] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 265–273, 2020.

[39] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[40] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10256, 2020.

[41] Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3459–3468, June 2023.

[42] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 2022.

[43] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *European Conference on Computer Vision*, pages 739–756. Springer, 2022.

[44] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020.

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

[46] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 409–424. Springer, 2022.

[47] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

[48] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834, 2021.

[49] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 211–229. Springer, 2020.

[50] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 9719–9728, 2020.