machine learning & knowledge extraction

MDPI

Article

Skew Class-Balanced Re-Weighting for Unbiased Scene Graph Generation

Haeyong Kang and Chang D. Yoo *

School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST),
Daejeon 34141, Republic of Korea
* Correspondence: cd_yoo@kaist.ac.kr

Abstract: An unbiased scene graph generation (SGG) algorithm referred to as Skew Class-Balanced Re-Weighting (SCR) is proposed for considering the unbiased predicate prediction caused by the long-tailed distribution. The prior works focus mainly on alleviating the deteriorating performances of the minority predicate predictions, showing drastic dropping recall scores, i.e., losing the majority predicate performances. It has not yet correctly analyzed the trade-off between majority and minority predicate performances in the limited SGG datasets. In this paper, to alleviate the issue, the *Skew Class-Balanced Re-Weighting* (SCR) loss function is considered for the unbiased SGG models. Leveraged by the skewness of biased predicate predictions, the SCR estimates the target predicate weight coefficient and then re-weights more to the biased predicates for better trading-off between the majority predicates and the minority ones. Extensive experiments conducted on the standard Visual Genome dataset and Open Image V4 and V6 show the performances and generality of the SCR with the traditional SGG models.

Keywords: scene graph generation (SGG); skew class-balanced re-weighting (SCR); predicate sample estimates; skew class-balanced effective number

check for updates

Citation: Kang, H.; Yoo, C.D. Skew Class-Balanced Re-Weighting for Unbiased Scene Graph Generation. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 287–303. https://doi.org/10.3390/make5010018

Academic Editor: Andreas Holzinger

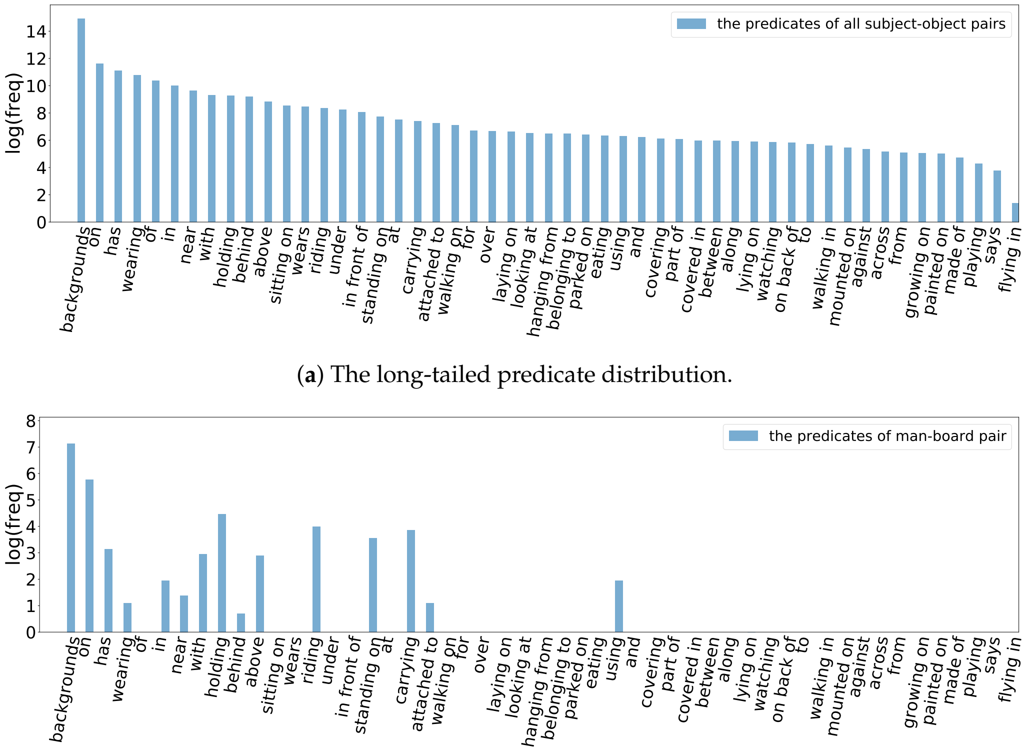
Received: 27 January 2023
Revised: 17 February 2023
Accepted: 7 March 2023
Published: 10 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Currently, computer vision [1–6] has been prevalent in various fields: material [7], chemical [8,9], and medical science [10]. Of great interest are tasks related to classifying and detecting objects by classical neural networks [11–15]. With the development of computer vision, graph neural networks [16–22] and multimodal text and image models [23–28] have also been intensely investigated. However, there is less information on Scene Graph Generation (SGG) models in the literature. For this reason, SGG has recently been receiving increased attention for improving image comprehension, divided into two approaches: one-stage [29,30] and two-stage [31–45]. The core building blocks of SGG are the objects in the image. There can be diverse relationships among the objects [46,47]. All relationships can be represented as triplets *subject*, *relation*, *object*, which can be used for generating the scene graph. The precise interpretation of an image depends on the core relations between the *subject–object* pairs. For example, given an image of *dog, woman, kite* objects, the image can be interpreted in multiple ways such as $\langle dog, has, leg \rangle$, $\langle dog, near, woman \rangle$, and $\langle woman, holding, kite \rangle$.

In general, however, the real-world large data sets often have shown long-tailed predicate distributions as depicted in Figure 1a—the predicate proportion of the Visual Genome [31], containing 51 predicates. One cause of long-tail distribution is the background predicates. The background predicates account for more than 80% in total predicates since it is not trivial to annotate all possible relationships, leading to possibly far more *subject-object* pairs without annotating ground truth than ones with ground truth labels. Another reason is the visual event frequency, as shown in Figure 1b. The majority predicates more frequently occur than the minority ones given a visual *subject–object* pair in the real-world scene. For example, given a *man-board* pair, the possible predicates—*backgrounds, on, holding,* *riding, etc.* represents a long-tail distribution. The predicate *on* is more frequently observed than *holding* and *riding*.



(**a**) The long-tailed predicate distribution.

(**b**) The predicate distribution of the *subject–object* pair.

Figure 1. The long-tailed predicate distribution: (**a**) the predicate proportion of the visual genome [31], where backgrounds are possibly far more than others and (**b**) the possible predicates of a *man-board* pair.

This long-tailed label distribution causes the trained model to tend to be biased toward the majority predicates. The SGG models [43,48–52] focus on increasing the minority predicate performances for the unbiased scene graph generation, i.e., increasing mean recall scores. However, these methods lead to deteriorating performances of the majority predicates, drastically dropping recall scores, i.e., losing the majority predicate performances. Recently, the issue has been alleviated through resampling images and object instances [49]. However, this method needs more computing power and time to train the model on the resampled samples. The current state-of-the-art models focus on re-balancing biased loss [53] or correcting noisy labels [54] to acquire an unbiased SGG model. Nevertheless, the prior works have not yet adequately described and analyzed the trade-off performances between majority and minority predicates for learning SGG models based on the given imbalanced datasets.

In this paper, the *Skew Class-Balanced Re-Weighting* (SCR) loss function is considered for alleviating the issues and acquiring the best trade-off performances in the multiple SGG tasks. Leveraged by the skewness of predicate predictions, the SCR estimates its weight coefficients and then reweights more to biased predicate samples to adaptively be unbiased SGG models. The extensive experimental results show that the SCR loss function gives more generalized performances than priors in the multiple SGG tasks on the Visual Genome dataset [55] and the Open Images datasets [56].

Contributions of our SCR learning scheme to unbiased SGG models:

- Leveraged by the skewness of biased predicate predictions, the *Skew Class-Balanced Re-Weighting* (SCR) loss function is firstly proposed for the unbiased scene graph generation (SGG) models.
- The SCR is applied to the current state-of-the-art SGG models to show its effectiveness, leading to more generalized performances: the SCR outperforms the prior reweighted methods on both mean recall and recall measurements in the multiple SGG tasks.

This paper is organized as follows. The Related Work section provides discussions on unbiased scene graph generation. The unbiased SGG section presents scene graph generation. In the Skew Class-Balanced Re-Weighting (SCR) section, the SCR loss function is depicted in detail. In the experimental section, the results of scene graph generation on the Visual Genome and Open Image V4 and V6 dataset are examined, along with an analysis and ablation study on the SCR with the current state-of-the-art SGG models. Finally, we conclude.

2. Related Works

Unbiased Scene Graph Generation (SGG). Predicate distribution is much more long-tailed than object distribution. For N objects and R predicates, the model has to address the fundamental challenge of learning $O(N^2R)$ relations with few [57,58]. To overcome the limited training dataset, the linguistic external knowledge [38,57,59] was used by Yu et al. [60], regularizing the deep neural network; using linguistic knowledge, the probabilistic model has also alleviated the semantical ambiguity of visual relationships [61]. Furthermore, to alleviate the imbalanced relationship distribution, Yin et al. [38] reformulated the conventional one-hot classification as a n-hot multiclass hierarchical recognition via novel Intra-Hierarchical trees (IH-trees) for each label set in the triplet $\langle subject, predicate, object\rangle$. Recently, unbiased SGG [43,47–54,62–71] has drawn unprecedented interest for more generalized SGG models. Occurrence-based Node Priority Sensitive (NPS)-loss [47] was used for balancing predictions; the Total Direct Effect (TDE) method has proposed firstly for unbiased learning by Tang et al. [43], which directly separates the bias from biased predictions through the counterfactual methodologies on causal graphs; CogTree [48] addressed the debiasing issue based on the coarse-to-fine structure of the relationships from the cognition viewpoint; Li et al. [49] improved the context modeling for tail categories by designing the bipartite graph network and message propagation on resampled objects and images. Lastly, the Predicate Probability Distribution based Loss (PPDL) [53] has proposed to train the biased SGG models, which measure the semantic predicate representation to re-balance the biased training loss. In this work, the *Skew Class-Balanced Re-Weighting* (SCR) loss function is proposed for alleviating biased predicate predictions, leading to the most generalized SGG models through the novel adaptive re-weighting learning scheme.

Re-Weighting Based Unbiased SGG. Overall unbiased SGG models can be categorized into the re-balancing strategy of re-weighting [47,48,50,53] and re-sampling [49] and biased model-based strategy [43,51,52]. For unbiased SGG modes, Tang et al. [43] first investigated the re-weighting learning algorithm. However, they observed that the performances of the majority predicates were drastically dropped, resulting in low recall scores while with high mean recall scores. This shows the general tendency that there is a trade-off performance between majority predicates and minority ones. To alleviate the trade-off issue, Yu et al. [48] proposed CogTree based on the coarse-to-fine structure of the relationships from the cognition viewpoint. Recently, the Predicate Probability Distribution (PPD) [53] re-balances the biased training loss according to the similarity between the predicted probability distribution and the estimated one. However, it has not yet correctly analyzed the trade-off performances between the majority and minority classes in various SGG tasks. In this paper, we measure the sample skew score based on the sample estimates for bias toward the majority classes to assign the sample weight correctly. The sample skewness is computed as the Fisher–Pearson coefficient of skewness on its sample mean value [72]. However, since the mean value tends to be biased toward the majority predicates, we measure the sample skew score fairly on its target logit instead of its mean value. Based on the sample skew score, the SCR assigns the sample weights adaptively—if there is no bias, we assign fewer weights to the sample (majority). If it is biased to one side, we assign larger weights to the samples (minority). Such that the SGG models with SCR show superior performances and generality on the multiple SGG tasks.

3. Unbiased Scene Graph Generation

In this section, we discuss the general scene graph generation model of the object and predicate predictions and depict the predicate sample estimates for measuring its skewness of biased predictions.

3.1. Scene Graph Generations

Given an image I, a scene graph model generates a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges, respectively. Each node $\boldsymbol{o}_i \in \mathcal{V}$ is represented by a bounding box $\boldsymbol{v}^{bbox} \in \mathbb{R}^4$ and a corresponding class label $o_y \in \mathcal{C}_{obj}$. Each edge $\boldsymbol{r}_{i,j} \in \mathcal{E}$ represents the predicate between the subject node $\boldsymbol{v}_i$ and the object node $\boldsymbol{v}_j$. The corresponding predicate label is $r_y \in \mathcal{C}_{rel}$.

We depict the proposed unbiased SGG framework as shown in Figure 2. Object detector outputs its predictions (a): *man* and *hourse*. Given *man-horse* label prediction pair, the traditional SGG model predicts (d) predicate predictions ($\hat{\boldsymbol{R}}$ see Equation (3)) through a fusion of (b) non-visual FREQ prior predicates and (c) visual predicate predictions ($\hat{\boldsymbol{R}}_{vis}$). For unbiased SGG model training, our SCR estimates (e) a sample size of the possible predicate candidates through FREQ ($\boldsymbol{R}_{freq}$ or $\boldsymbol{R}_{emb}$) and measures (f) the target label skew score and then calculates (g) the training target sample weight for the adaptive re-weighted loss. In (f) and (g), red lines around the target label indicate predicate skew (S^i_{skew} see Equation (9)) and re-weight scores ($\boldsymbol{W}^i$ see Equation (7)), respectively, where they have an approximately inverse relationship.

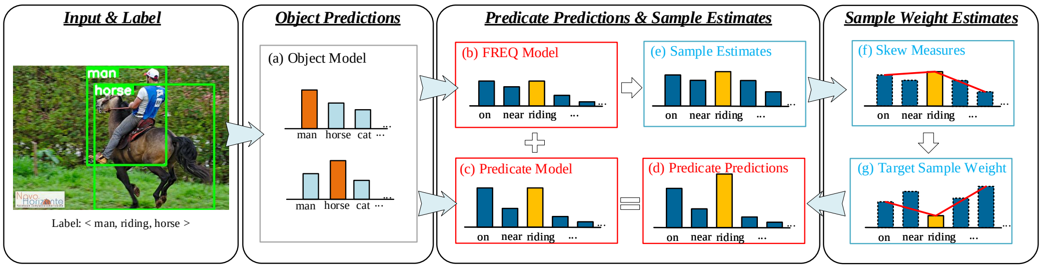


Figure 2. Skew class-balanced re-weighting (SCR) for unbiased SGG models. Based on (**a**) object detection, the SGG model predicts (**d**) predicates of (**b**) FREQ and (**c**) visual predictions; unbiased SGG is trained by (**g**) target sample weight, which is estimated by (**f**) skew measures of (**e**) possible sample estimates.

3.1.1. Object Predictions

Following [43], the node features are derived from the object detector. In particular, for each bounding box v^{bbox}_i, the detector returns a RoI-Align feature $\boldsymbol{x}^{RoI}_i$ and an object label embedding $\boldsymbol{l}_i$. In general SGG model, the $\mathcal{N}$ number of node features are constructed by vector concatenation $\boldsymbol{X} = \{[\boldsymbol{x}^{RoI}_i; \boldsymbol{l}_i; \boldsymbol{b}_i]\}^{\mathcal{N}}_{i=1}$, where $\boldsymbol{b}_i$ is the embedded box feature from the box coordinate v^{bbox}_i.

$$\tilde{\boldsymbol{X}} = fc_{obj}(\boldsymbol{X}) \in \mathbb{R}^{\mathcal{N} \times \mathcal{D}_{obj}}, \tag{1}$$

where all fc_* denote a fully connected layer for linear transformations or logits, and the object feature dimension $\mathcal{D}_{obj}$ depends on the SGG model. As depicted in Figure 2a, the predicted object label of $\hat{\boldsymbol{O}} \in \mathbb{R}^{\mathcal{N} \times |\mathcal{C}_{obj}|}$ is estimated by object logits $fc(\tilde{\boldsymbol{X}})$ as follows:

$$\hat{\boldsymbol{O}} = fc(\tilde{\boldsymbol{X}}). \tag{2}$$

3.1.2. Predicate Predictions

Inspired by Tang et al. [43], Zhu et al. [73], predicate predictions can be made by employing multiple logits from visual and non-visual features. We follow the sum over all outputs to generate the final predicate prediction. As illustrated in Figure 2d, the combined predicate logit $\hat{\boldsymbol{R}}$ is estimated based on the summation of the visual logits and the non-visual logits as follows:

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{R}}_{vis} \oplus \hat{\boldsymbol{R}}_{freq} \oplus \hat{\boldsymbol{R}}_{emb}, \tag{3}$$

where $\hat{\boldsymbol{R}} \in \mathbb{R}^{\mathcal{N}(\mathcal{N}-1)\times|\mathcal{C}_{rel}|}$; $\oplus$ is an element-wise sum; as shown in Figure 2c, $\hat{\boldsymbol{R}}_{vis}$ is the predicate logits from visual feature $\boldsymbol{F}_{vis}$ such as a $\mathcal{D}_{vis}$ dimensional union feature and a *subject–object* pair feature, which also depends on the SGG model,

$$\begin{aligned}
\hat{\boldsymbol{R}}_{vis} &= fc_{vis}(\boldsymbol{F}_{vis}), \;\; \boldsymbol{F}_{vis} \in \mathbb{R}^{\mathcal{N}(\mathcal{N}-1)\times\mathcal{D}_{vis}}, \\
\hat{\boldsymbol{R}}_{freq} &= \mathrm{Sigmoid}(\boldsymbol{R}_{freq}) \in \mathbb{R}^{\mathcal{N}(\mathcal{N}-1)\times|\mathcal{C}_{rel}|}, \\
\hat{\boldsymbol{R}}_{emb} &= fc_{emb}(\boldsymbol{L}_{emb}), \;\; \boldsymbol{L}_{emb} \in \mathbb{R}^{\mathcal{N}(\mathcal{N}-1)\times 400}.
\end{aligned} \tag{4}$$

The FREQ [74] as a non-visual feature, $\boldsymbol{R}_{freq}$ (Figure 2b) looks up the empirical distribution over relationships between subject $\hat{o}_i$ and object $\hat{o}_j$ as computed in the training set where $\hat{o}_i, \hat{o}_j \in \hat{\boldsymbol{O}}$. However, since FREQ does not consider any image representations when predicting predicates, it tends to lead to biased predicate predictions due to its imbalanced predicate distribution. To minimize the biased effects, we use the Sigmoid-activated FREQ predicate logits. In addition, for acquiring the more smoothness of the empirical distribution, the concatenated *subject–object* embedding $\hat{\boldsymbol{R}}_{emb}$ is added to the predicate predictions where $\boldsymbol{L}_{emb} = \{[\boldsymbol{l}_i; \boldsymbol{l}_j]\}_{i,j}^{\mathcal{N}}$.

3.2. Sample Estimates

Non-visual predicate features tend to be more biased than visual features due to an imbalanced training set. If the degree of biased predictions can be measured, we can leverage its value to learn the SGG models without bias. According to Brown [72], the degree of bias prediction can generally be measured in a skew score. In general, if there is no bias, the skew score is close to 0, and if it is biased to one side, the skew score is over either -1 or $+1$.

In SCR, we need to estimate how many predicate samples are biased to measure the predicate skew scores. To approximate the biased sample numbers, we use two non-visual prior predicate distributions of FREQ, $\hat{\boldsymbol{R}}_{freq}$ and *subject–object* label embedding $\hat{\boldsymbol{R}}_{emb}$ as described in Figure 2e. Based on the two predicate distributions, at first, we define the skew logit $\hat{\boldsymbol{R}}_{skew}$ with the Sigmoid activation function as follows:

$$\hat{\boldsymbol{R}}_{skew} = \mathrm{Sigmoid}(\boldsymbol{R}_{freq} \oplus \hat{\boldsymbol{R}}_{emb}). \tag{5}$$

To estimate the predicate sample weights properly, the SCR approximates skewness through the predicate sample estimates $\hat{\boldsymbol{R}}_{skew}$ in Equation (5). We investigate the best predicate sample estimates through several experiments with a combination of non-visual predictions as follows:

- SCR of EMB: $\hat{\boldsymbol{R}}_{skew} = \mathrm{Sigmoid}(\hat{\boldsymbol{R}}_{emb})$;
- SCR of FREQ: $\hat{\boldsymbol{R}}_{skew} = \mathrm{Sigmoid}(\boldsymbol{R}_{freq})$;
- SCR of FREQ+EMB: $\hat{\boldsymbol{R}}_{skew} = \mathrm{Sigmoid}(\boldsymbol{R}_{freq} \oplus \hat{\boldsymbol{R}}_{emb})$.

Then, the predicate sample estimates are acquired as follows:

$$\mathcal{M} = \sum_{i}^{\mathcal{N}(\mathcal{N}-1)} \hat{\boldsymbol{r}}_i, \tag{6}$$

where $\hat{\boldsymbol{r}}_i \in \hat{\boldsymbol{R}}_{skew}$; the estimated $m_y \in \mathcal{M}$ is the number of yth predicate sample size and $m_y \in [0, \mathcal{N}(\mathcal{N}-1))$.

Moreover, the skew predicate $\hat{\boldsymbol{R}}_{skew}$ serves to estimate the predicate candidates that a *subject–object* pair can have fairly since the Sigmoid activation function suppresses the more extensive biased predictions. For example, a *Man-Horse* pair may have predicates such as *riding, on, with*, etc. The Sigmoid activation function amplifies the frequency of the minority predicates while squeezing that of the majority ones, as shown in Figure 2b,e, which is used to calculate Skew Class-Balanced Effective Number that we depict in the following section.

Note in Equation (5), the Sigmoid activation function was applied to estimate the predicate sample sizes (Equation (6)) and measure skewness (Equation (9)), given a subject–object pair. The main reason was that it provides smoothly activated outputs from 0 to 1. We assumed that an output value of 1 is regarded as one predicate.

4. Skew Class-Balanced Re-Weighting (SCR)

In this section, leveraged by the biased predictions deduced by the predicate sample estimates, the *Skew Class-Balanced Re-Weighting* (SCR) performs sample weight estimates for learning unbiased SGG models.

4.1. Skew Class-Balanced Effective Number

The *Skew Class-Balanced Effective Number* approximates the mini-batch class-balanced re-weighting coefficients $\boldsymbol{E}_{m_y}$ based on the predicate skew logit $\hat{\boldsymbol{R}}_{skew}$ in Equation (5). The ith predicate sample $E^i_{m_y}$ is defined as follows as [75,76]:

$$E^i_{m_y} = \frac{1-\beta_i^{m_y}}{(1-\beta_i)}, \tag{7}$$

where $\beta_i = (m_y - 1)/m_y$; the effective number satisfies the following properties that $E^i_{m_y} = 1$ if $\beta_i = 0$ $(m_y = 1)$; $E^i_{m_y} \to m_y$ as $\beta_i \to 1$ $(m_y \to \infty)$ such that β_i controls how fast $E^i_{m_y}$ grows as the target predicate sample size m_y increases.

To estimate the ith predicate effective number $E^i_{m_y}$, we adaptively estimate the $\beta_i \in [0,1)$ by using the entropy H^i_{skew} and skew score function S^i_{skew} of the $\hat{\boldsymbol{R}}_{skew}$ as shown in Algorithm 1: if $S^i_{skew} > S_{th}$ then the $\beta_i \in [0,1)$ that assigns more weights to the minority predicate samples than the majority ones; otherwise, $S^i_{skew} \leq S_{th}$ then $\beta_i = 0$ that re-weights uniformly over the entire class sample loss, i.e., conventional cross-entropy loss. The threshold S_{th} is determined as follows:

$$S_{th} = \bar{S}_{skew} - \delta, \tag{8}$$

where $\bar{S}_{skew}$ and $\delta = 0.7$ are the mean value of $\boldsymbol{S}_{skew}$ and the hyper-parameter used, respectively, in all experiments.

4.2. Skew Measures

In training, the predicate sample skewness depends on the predicate label. In other words, according to Brown [72], some majority predicate skew values tend to be greater than zero while others tend to be zero or negative. This is an unfair skew measure, leading the SCR Algorithm 1 to more weights either in the majority predicates or in the minority ones. In this paper, as portrayed in Figure 2f, to measure the skew value of all target labels fairly, we use the target skew logit instead of the mean value. Therefore, the ith predicate sample skew S^i_{skew} is firstly measured by the following equation given the target label index y as follows:

$$S^i_{skew} = \frac{\frac{1}{|\mathcal{C}_{rel}|}\sum_{\hat{r}_{i,j} \in \hat{\boldsymbol{R}}^i_{skew}} (\hat{r}_{i,j} - \hat{r}_{i,y})^3}{\left(\frac{1}{|\mathcal{C}_{rel}|}\sum_{\hat{r}_{i,j} \in \hat{\boldsymbol{R}}^i_{skew}} (\hat{r}_{i,j} - \hat{r}_{i,y})^2\right)^{3/2}}. \tag{9}$$

Then, we use the uniformness to determine the β. The uniformness provides confidence in the predicate sample estimates. To calculate the uniformness of the ith predicate sample, we estimate the entropy H^i_{skew} as follows:

$$H^i_{skew} = -\lambda_{skew} \sum_{\hat{r}_{i,j} \in \hat{\boldsymbol{R}}^i_{skew}} p(\hat{r}_{i,j}) \log_{|\mathcal{C}_{rel}|} p(\hat{r}_{i,j}), \tag{10}$$

where the number of predicates $|\mathcal{C}_{rel}|$ is used as the base of the logarithm, which $H^i_{skew} \in [0, 1]$; the $\hat{\boldsymbol{r}}_i \in \hat{\boldsymbol{R}}_{skew}$ have uniform distributions when H^i_{skew} is close to 1; otherwise, the predicate sample may have either skew distributions or correct distributions and the $\lambda_{skew} = 0.06$ is the coefficient of S_{skew}. The following section depicts the relationship between skew and entropy in detail.

Algorithm 1 Skew Class-Balanced Effective Number.

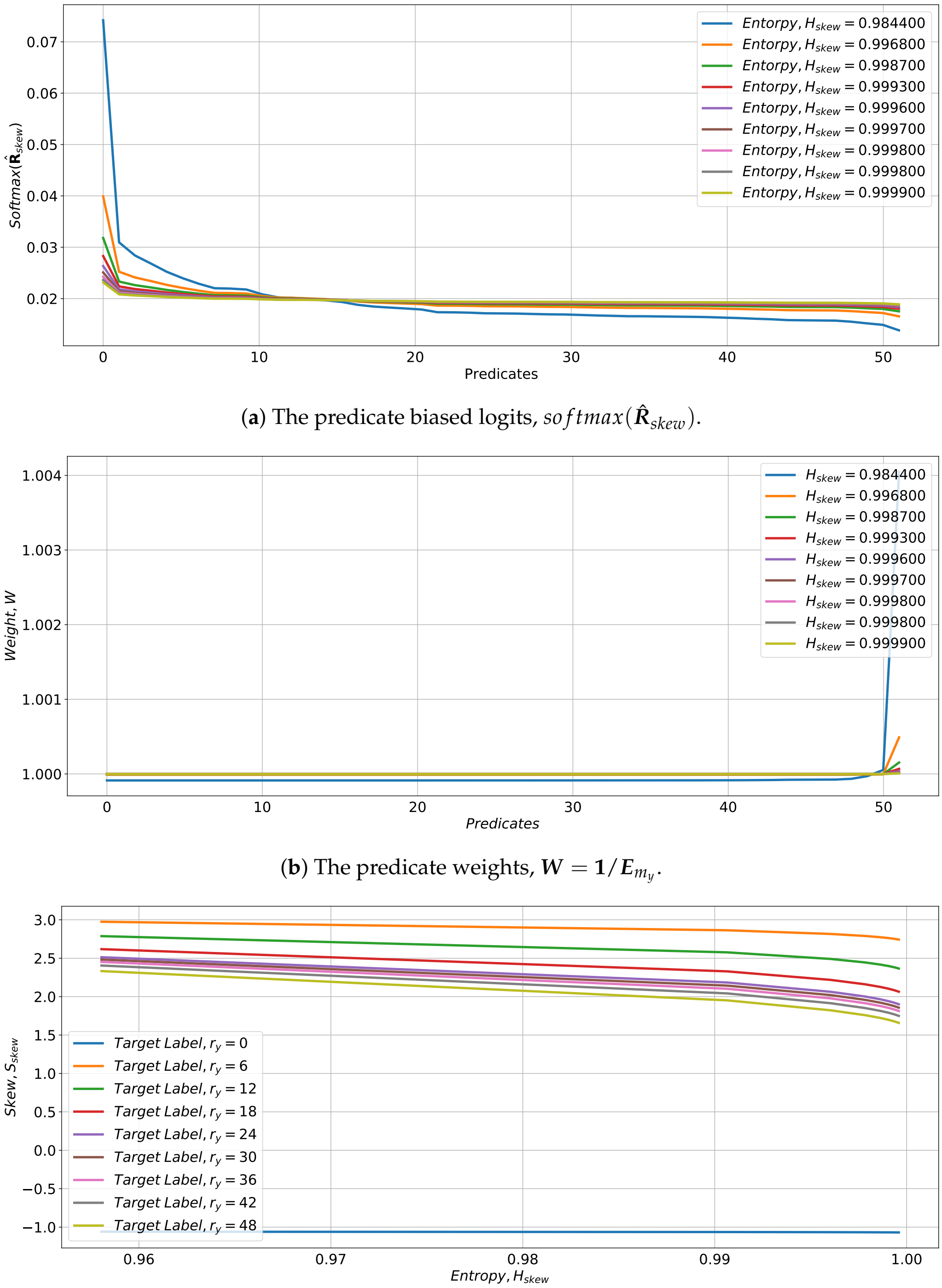
```
Require: Dataset $\mathcal{D}$, SGG Model $f_\theta$
for $t = 0, 1, 2, \ldots, T$ do
    $\mathcal{B} \leftarrow$ Minibatch($\mathcal{D}$)
    $\hat{\boldsymbol{R}}_{skew} \leftarrow Skew\_Logits(\hat{\boldsymbol{R}}_{freq}, \hat{\boldsymbol{R}}_{emb}; \mathcal{B}, f_\theta)$    (Equation (5))
    $\mathcal{M} \leftarrow Sample\_Estimates(\hat{\boldsymbol{R}}_{skew})$    (Equation (6))
    $\boldsymbol{S}_{skew} \leftarrow Skew(\hat{\boldsymbol{R}}_{skew}, \boldsymbol{R})$    (Equation (9))
    $\boldsymbol{H}_{skew} \leftarrow Entropy(\hat{\boldsymbol{R}}_{skew})$    (Equation (10))
    if $\boldsymbol{S}_{skew} > S_{th}$ then
        $\boldsymbol{\beta} = \mathbf{1.0} - \boldsymbol{H}_{skew}$
    else
        $\boldsymbol{\beta} = \mathbf{0.0}$
    end if
    $E^i_{m_y} = \frac{1-\beta_i^{m_y}}{1-\beta_i}; m_y \in \mathcal{M}, i \in [0, \mathcal{N}(\mathcal{N}-1))$    (Equation (7))
end for
```

4.3. Target Sample Weights

The interpretation of the relationship between skew and weight is depicted based on biased predicate prediction as shown in Figure 3. In Figure 3a, to understand the sample weight estimate from the predicate sample estimates, we assume that the predicate biased predictions $\hat{\boldsymbol{R}}_{skew}$ are given by the predicate sample frequency $\mathcal{M}^{1/5}$ (see Figure 1a) and the predicate entropy $\boldsymbol{H}_{skew}$, i.e., the more frequent sample is the more biased prediction is; the more biased the prediction is the less entropy is. The skew $\boldsymbol{S}_{skew}$ measures not only the degree of true/false prediction but also biased predictions as shown in Figure 3c. The following simple equation summarizes the degree of true/false predictions:

$$\begin{cases} S^i_{skew} > 0 & \text{if } \arg\max_j \hat{\boldsymbol{r}}_{i,j} \neq r_y \\ S^i_{skew} < 0 & \text{otherwise.} \end{cases} \tag{11}$$

where $S^i_{skew} \approx -1$ for the true prediction; $S^i_{skew} = 0$ for the uniform prediction and $S^i_{skew} \gg 0$ for the false and biased prediction. Moreover, the predicate target label-wise skew has more discrete at high entropy. The final ith sample weight $w^i_y = 1/E^i_{m_y}$ (Figure 2g) is acquired by the criteria of skew and entropy as shown in Figure 3b. The minority predicates have larger weights than the majority when $S^i_{skew} \gg 0$.



(**a**) The predicate biased logits, $softmax(\hat{\mathbf{R}}_{skew})$.

(**b**) The predicate weights, $\boldsymbol{W} = \mathbf{1}/\boldsymbol{E}_{m_y}$.

(**c**) The predicate target label-wise skew, $\boldsymbol{S}_{skew}$

Figure 3. The biased predicate weights: (**a**) the predicate biased logits $\hat{\mathbf{R}}_{skew}$ of label frequency $\mathcal{M}^{1/5}$, depending on entropy H_{skew}, (**b**) the sensitivity of sample weights $\boldsymbol{W}$, according to entropy H_{skew}, and (**c**) the skew measures $\boldsymbol{S}_{skew}$ of a target label r_y.

4.4. Learning with SCR

Except for object loss, all traditional SGG models are learned by the skew class-balanced re-weighting cross-entropy loss functions. The conventional cross-entropy loss for the objects is computed, given object predictions $\hat{\boldsymbol{O}}$ and the object ground-truth $o_y^i \in \boldsymbol{O}$ as follows:

$$\mathcal{L}_{obj}(\hat{\boldsymbol{O}}, \boldsymbol{O}) = \sum_{i}^{\mathcal{N}} -\gamma_{obj} \cdot w_y^i \log\left(\frac{\exp(\hat{o}_y)}{\sum_{\hat{o}_j \in \hat{\boldsymbol{O}}_i} \exp(\hat{o}_j)}\right), \tag{12}$$

where, given $\mathcal{N}$ of objects, ith object sample weight is $w_y^i = 1$ s.t. $|\mathcal{C}_{obj}| = \sum_j^{|\mathcal{C}_{obj}|} w_j^i$; here, the sample weight denominator $\gamma_{obj} = 1/\sum_i^{\mathcal{N}} w_y^i$ is used for the mean object cross-entropy loss. The skew class-balanced cross-entropy loss computes the skew predicate-balanced cross-entropy, based on the predicate predictions $\hat{\boldsymbol{R}}$ and the predicate ground-truth $r_y^i \in \boldsymbol{R}$:

$$\mathcal{L}_{rel}(\hat{\boldsymbol{R}}, \boldsymbol{R}) = \sum_{i}^{\mathcal{N}(\mathcal{N}-1)} -\gamma_{rel} \cdot w_y^i \log\left(\frac{\exp(\hat{r}_y)}{\sum_{\hat{r}_j \in \hat{\boldsymbol{R}}_i} \exp(\hat{r}_j)}\right), \tag{13}$$

where, given $\mathcal{N}(\mathcal{N}-1)$ of object-subject pairs, the ith predicate sample weight is $w_y^i = \frac{1}{E_{m_y}^i} = \frac{1-\beta_i}{1-\beta_i^{m_y}}$ s.t. $|\mathcal{C}_{rel}| = \sum_j^{|\mathcal{C}_{rel}|} w_j^i$; here, we set the predicate sample normalizer as $\gamma_{rel} = 1/\sum_i^{\mathcal{N}(\mathcal{N}-1)} w_y^i$ for mean predicate loss. In summary, the total objective loss function $\mathcal{L}_{total}$ for unbiased SGG learning can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{obj} + \mathcal{L}_{rel}. \tag{14}$$

5. Experiments

The proposed SCR is evaluated with the traditional SGG models on the Visual Genome benchmark datasets [55], and the performances of the SCR are compared with others in the multiple SGG tasks. The source code is available at https://github.com/ihaeyong/Unbiased-SGG (accessed on 26 January 2023).

5.1. Visual Genome

We used Visual Genome (VG) [55] dataset to train and evaluate our models, which is composed of 108k images across 75k object categories and 37k predicate categories. We followed the widely adopted VG split [31,42,74] containing the most frequent 150 object categories and 50 predicate categories. The original split only has a training set (70%) and a test set (30%). We followed Zellers et al. [74] to sample a 5k validation set from the training set for parameter tuning.

5.2. Open Images

The Open Images dataset [56] is a large-scale dataset proposed by Google recently. Compared with the Visual Genome dataset, it has a superior annotation quality for the scene graph generation. In this work, we conduct experiments on Open Images $V4\&V6$, following similar data processing and evaluation protocols in [47,56,77]. The Open Images V4 is introduced as a benchmark for scene graph generation by Zhang et al. [77] and Lin et al. [47], which has 53,953 and 3234 images for the train and validation sets, 57 objects categories, and 9 predicate categories in total. The Open Images V6 has 126,368 images used for training, 1813, and 5322 images for validation and testing, respectively, with 301 object categories and 31 predicate categories. This dataset has a comparable amount of semantics categories with the VG.

5.3. Experiments Configurations

State-of-the-Art Comparisons. For fair comparisons, all the compared SGG models should use the FREQ [74], which looks up the empirical distribution over relationships between subject prediction $\hat{o}_i$ and object ones $\hat{o}_j$. To evaluate the effectiveness of the SCR learning algorithm, we follow the same experimental settings as the CogTree [48]. We also set the current state-of-the-art SGG models as the baseline: MOTIFS [74], VCTree [41] and SG-Transformer [48] which contains 3 *O2O* blocks and 2 *R2O* blocks with 12 attention heads, and Bipartite-Graph [49] without resampling layers and compare the performance with the state-of-the-art debiasing approach TDE [43], CogTree [48], PCPL [50], DLFE [51], BPL-SA [52], PPDL [53], and NICE [54].

Implementation. Following the previous works [43,48,49], the object detector is the pre-trained Faster R-CNN [78] with ResNeXt-101-FPN [79]. In bi-level resampling [49], we also set the repeat factor $t = 0.07$, instances drop rate $\gamma_d = 0.7$, and weight of fusion the entities features $\rho = -5$. The α, β are initialized as 2.2 and 0.025, respectively.

5.4. Evaluations

Our SCR has the following two evaluations:

Relationship Retrieval (RR) contains three sub-task: (1) Predicate Classification (PredCls): taking ground truth bounding boxes and labels as inputs, (2) Scene Graph Classification (SGCls): using ground truth bounding boxes without labels, (3) Scene Graph Detection (SGDet): detecting SGs from scratch. The conventional metric of RR is **Recall@K (R@K)**, included in this paper even though the biased prediction is reported by Misra et al. [80] for the performance of the SCR. Moreover, to evaluate the general performances, we adopted **mean Recall@ K (mR@K)** that retrieves each predicate separately and then averages R@K for all predicates.

Zero-Shot Relationship Retrieval (ZSRR). The **Zero-Shot Recall@K** was firstly evaluated on the VG dataset in [43], which reports the R@K of those subject-predicate-object triplets that have never been observed in the training set. ZSRR also has three sub-tasks as RR.

5.5. Quantitative Results

Visual Genome. The SCR is compared with others on the two evaluation tasks: RR and ZSRR, which are the same as shown in Tables 1 and 2. The SCR achieves the best and second best performances over the previous methods: TDE, PCPL, Cogtree, DLFE, BPL-SA PPDL, and NICE, demonstrating its generality and effectiveness on the two measures of RR task. Moreover, the SCR shows the best trade-off performances on the ZSRR task as shown in Figure 2.

The Best Trade-off performances. To analyze the trade-off between majority and minority predicate performances, firstly, we need to understand the two measurements, mRR and RR. The mRR was introduced to calculate the recall on each predicate category independently and then to average the results by Tang et al. [41]. This is because of RR's bias, which was reported by Misra et al. [80]. The higher the recall scores (RR) are, the more biased the majority categories are. Therefore, a higher mRR is a more unbiased SGG model; a higher RR is a more biased SGG model. Traditional SGG models show the tendency that if mRR is higher than RR, RR performances are low, i.e., MOTIFS, VCTREE, and SG-Transformer show the best Recall but low mRR@100 in Table 1. However, our SCR shows the best trade-off in terms of mRR and RR measurements; MOTIFS with SCR shows the best performances in mRR and the second-best performances in RR except for MOTIFS. So do the other models' results. Figure 4 supports our analysis of the best trade-off since overall predicate performances were better than the traditional SGG model (SG-Transformer).

Open Image V4 and V6. To show the effectiveness of SCR, we set BGNN as the baseline, as shown in Table 3. On Open Images Dataset V4, SCR outperformed BGNN except for $score_{wtd}$ measurement. Mainly, SCR shows outstanding performance in terms of phrase evaluation. Moreover, on Open Images Dataset V6, SCR outperformed all baselines such as BGNN, GPS-Net, Etc., showing a good trade-off between mean recall@50 and recall@50. These results proved that we could have a good trade-off performance in long-tailed predicated distributions if properly assigning weights in training.

Table 1. The SGG performances of relationship retrieval on mean Recall@K and Recall@K. SCR† denotes SCR of FREQ+EMB. Note the best and second best methods under each setting are marked according to format.

	PredCls		**SGCls**		**SGDet**	
Model	mR@20/50/100	R@20/50/100	mR@20/50/100	R@20/50/100	mR@20/50/100	R@20/50/100
IMP+ [31]	-.-/ 9.8/10.5	52.7/59.3/61.3	-.-/ 5.8/ 6.0	31.7/34.6/35.4	-.-/ 3.8/ 4.8	14.6/20.7/24.5
FREQ [74]	8.3/13.0/16.0	53.6/60.6/62.2	5.1/ 7.2/ 8.5	29.3/32.3/32.9	4.5/ 6.1/ 7.1	20.1/26.2/30.1
KERN [46]	-.-/17.7/19.2	-.-/65.8/67.6	-.-/ 9.4/10.0	-.-/36.7/37.4	-.-/ 6.4/ 7.3	-.-/27.1/29.8
MOTIFS [74]	10.8/14.0/15.3	58.5/65.2/67.1	6.3/ 7.7/ 8.2	32.9/35.8/36.5	4.2/ 5.7/ 6.6	21.4/27.2/30.3
VCTree [41]	14.0/17.9/19.4	60.1/66.4/68.1	8.2/10.1/10.8	35.2/38.1/38.8	5.2/ 6.9/ 8.0	22.0/27.9/31.3
MSDN [33]	-.-/15.9/17.5	-.-/64.6/66.6	-.-/ 9.3/ 9.7	-.-/38.4/39.8	-.-/ 6.1/ 7.2	-.-/31.9/36.6
G-RCNN [37]	-.-/16.4/17.2	-.-/64.8/66.7	-.-/ 9.0/ 9.5	-.-/38.5/37.0	-.-/ 5.8/ 6.6	-.-/29.7/32.8
BGNN [49]	-.-/30.4/32.9	-.-/59.2/61.3	-.-/14.3/16.5	-.-/37.4/38.5	-.-/10.7/12.6	-.-/31.0/35.8
DT2-ACBS [44]	-.-/35.9/39.7	-.-/23.3/25.6	-.-/24.8/27.5	-.-/16.2/17.6	-.-/22.0/24.4	-.-/15.0/16.3
MOTIFS [74]	11.5/14.6/15.8	59.5/66.0/67.9	6.5/ 8.0/ 8.5	35.8/39.1/39.9	4.1/ 5.5/ 6.8	25.1/32.1/36.9
+ TDE [43]	18.5/25.5/29.1	33.6/46.2/51.4	9.8/13.1/14.9	21.7/27.7/29.9	5.8/ 8.2/ 9.8	12.4/16.9/20.3
+ PCPL [50]	-.-/24.3/26.1	-.-/54.7/56.5	-.-/12.0/12.7	-.-/35.3/36.1	-.-/10.7/12.6	-.-/27.8/31.7
+ CogTree [48]	20.9/26.4/29.0	31.1/35.6/36.8	12.1/14.9/16.1	19.4/21.6/22.2	7.9/10.4/11.8	15.7/20.0/22.1
+ DLFE [51]	22.1/26.9/28.8	-.-/52.5/54.2	12.8/15.2/15.9	-.-/32.3/33.1	8.6/11.7/13.8	-.-/25.4/29.4
+ BPL-SA [52]	24.8/29.7/31.7	-.-/50.7/52.5	14.0/16.5/17.5	-.-/30.1/31.0	10.7/13.5/15.6	-.-/23.0/26.9
+ PPDL [53]	-.-/32.2/33.3	-.-/47.2/47.6	-.-/17.5/18.2	-.-/28.4/29.3	-.-/11.4/13.5	-.-/21.2/23.9
+ NICE [54]	-.-/29.9/32.3	-.-/55.1/57.2	-.-/16.6/17.9	-.-/33.1/34.0	-.-/12.2/14.4	-.-/27.8/31.8
+ **SCR† (ours)**	25.9/31.5/33.6	51.0/57.9/60.1	14.2/17.1/18.2	27.1/31.0/32.3	9.6/13.5/15.9	18.1/25.1/29.5
VCTree [41]	11.7/14.9/16.1	59.8/66.2/68.1	6.2/ 7.5/ 7.9	37.0/40.5/41.4	4.2/ 5.7/ 6.9	24.7/31.5/36.2
+ TDE [43]	18.4/25.4/28.7	36.2/47.2/51.6	8.9/12.2/14.0	19.9/25.4/27.9	6.9/ 9.3/11.1	14.0/19.4/23.2
+ PCPL [50]	-.-/22.8/24.5	-.-/56.9/58.7	-.-/15.2/16.1	-.-/40.6/41.7	-.-/10.8/12.6	-.-/26.6/30.3
+ CogTree [48]	22.0/27.6/29.7	39.0/44.0/45.4	15.4/18.8/19.9	27.8/30.9/31.7	7.8/10.4/12.1	14.0/18.2/20.4
+ DLFE [51]	20.8/25.3/27.1	-.-/51.8/53.5	15.8/18.9/20.0	-.-/33.5/34.6	8.6/11.8/13.8	-.-/22.7/26.3
+ BPL-SA [52]	26.2/30.6/32.6	-.-/50.0/51.8	17.2/20.1/21.2	-.-/34.0/35.0	10.6/13.5/15.7	-.-/21.7/25.5
+ PPDL [53]	-.-/33.3/33.8	-.-/47.6/48.0	-.-/14.3/15.7	-.-/32.1/33.0	-.-/11.3/13.3	-.-/20.1/22.9
+ NICE [54]	-.-/30.7/33.0	-.-/55.0/56.9	-.-/19.9/21.3	-.-/37.8/39.0	-.-/11.9/14.1	-.-/27.0/30.8
+ **SCR† (ours)**	27.7/33.5/35.5	49.7/56.4/58.3	15.4/18.9/20.1	26.7/30.6/31.9	10.3/13.8/16.3	18.1/25.0/29.4
SG-Transformer [48]	14.8/19.2/20.5	58.5/65.0/66.7	8.9/11.6/12.6	35.6/38.9/39.8	5.6/ 7.7/ 9.0	24.0/30.3/33.3
+ CogTree[48]	22.9/28.4/31.0	34.1/38.4/39.7	13.0/15.7/16.7	20.8/22.9/23.4	7.9/11.1/12.7	15.1/19.5/21.7
+ SCR† (ours)	27.0/32.2/34.5	45.3/52.7/55.0	14.9/17.7/18.7	25.1/28.9/30.2	10.4 /13.4/15.0	17.7/23.2/26.2

Table 2. The SGG Performances of zero-shot relationship retrieval on Recall@K. SCR† denotes SCR of FREQ+EMB. The SGG models re-implemented under our codebase are denoted by the superscript *.

Zero-Shot Relationship Retrieval		**PredCls**		**SGCls**		**SGDet**	
Model	Method	R@50	R@100	R@50	R@100	R@50	R@100
MOTIFS [43]	baseline [43]	10.9	14.5	2.2	3.0	0.1	0.2
	Reweight [43]	0.7	0.9	0.1	0.1	0.0	0.0
	TDE [43]	14.4	18.2	3.4	4.5	2.3	2.9
	CogTree [48] *	2.4	4.0	0.9	1.5	0.3	0.6
	SCR† (ours)	**18.0**	**21.1**	**5.1**	**5.9**	**2.4**	**3.8**
VCTree [43]	Baseline [43]	10.8	14.3	1.9	2.6	0.2	0.7
	TDE [43]	14.3	17.6	3.2	4.0	2.6	3.2
	CogTree [48] *	3.3	5.0	2.1	2.6	0.4	0.6
	SCR† (ours)	**17.6**	**20.4**	**4.5**	**5.2**	**2.5**	**3.5**
SG-Transformer [48] *	Baseline *	4.1	6.3	1.6	2.3	0.2	0.5
	CogTree [48] *	5.2	7.3	2.3	3.0	0.3	0.5
	SCR† (ours)	**16.4**	**19.6**	**4.6**	**5.3**	**2.0**	**3.2**
BGNN [49]	BGNN *	15.0	18.0	4.5	5.4	4.5	5.3
	CogTree [48] *	13.4	16.1	5.0	5.7	0.5	0.8
	SCR† (ours)	**16.3**	**19.5**	**4.9**	**5.9**	**1.9**	**3.0**

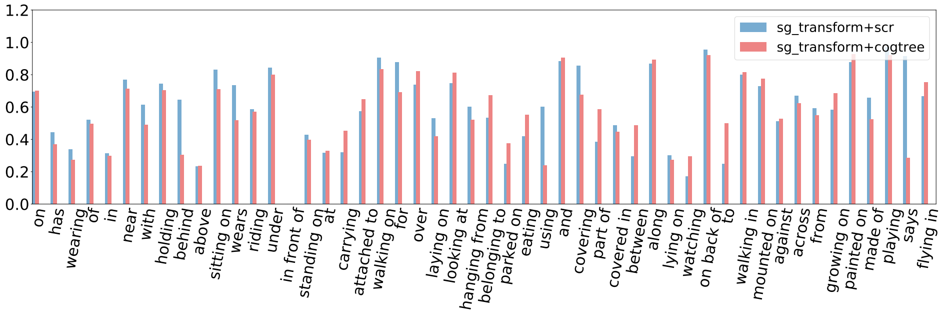


Figure 4. The Recall@100 on PredCls: we compare the SCR of FREQ+EMB with CogTree using the SG-Transformer [48].

Table 3. The performances of open images dataset. $*$ denotes results reproduced by Li et al. [49]. SCR† denotes SCR of FREQ+EMB.

Dataset	Models	mR@50	R@50	wmAP		score$_{\text{wtd}}$
				rel	phr	
V4	RelDN [77] *	70.40	75.66	36.13	39.91	45.21
	GPS-Net [47] *	69.50	74.65	35.02	39.40	44.70
	BGNN [49]	72.11	75.46	37.76	41.70	**46.87**
	BGNN+SCR† (ours)	**72.20**	**75.48**	**38.64**	**45.01**	45.01
V6	RelDN [77] *	33.98	73.08	32.16	33.39	40.84
	VCTree [41] *	33.91	74.08	34.16	33.11	40.21
	MOTIFS [74] *	32.68	71.63	29.91	31.59	38.93
	TDE [43] *	35.47	69.30	30.74	32.80	39.27
	GPS-Net [47] *	35.26	74.81	32.85	33.98	41.69
	BGNN [49]	40.45	74.98	33.51	34.15	42.06
	BGNN+SCR† (ours)	**42.43**	**75.21**	**33.98**	**35.13**	**42.66**

5.6. Ablation Study

We investigate the predicate-biased prediction and the best hyper-parameter settings of the SCR loss function for the better generalized SGG models.

Predicate Bias. To estimate the predicate bias and assign the proper sample weights, we define the predicate sample estimates $\hat{\boldsymbol{R}}_{skew}$ of $\hat{\boldsymbol{R}}_{freq}$ and $\hat{\boldsymbol{R}}_{emb}$ in Equation (5). To investigate the proper predicate sample estimates, we examined the effectiveness of the predicate sample estimates-SCR of EMB, SCR of FREQ, and SCR of FREQ+EMB with the fixed predicate predictions (Equation (3)) as shown in Table 4. In the experiments, the SCR of FREQ+EMB leads to more generalized performances over others. In summary, the previous Re-Weight methods worsen the recall performances, while the SCR does not.

Table 4. The ablation study for the predicate sample estimates on Recall@100. The underbar represents the predicate sample estimates for better generality.

Relationship Retrieval		PredCls			SGCls			SGDet		
Model	Method	mRR	ZSRR	RR	mRR	ZSRR	RR	mRR	ZSRR	RR
VCTree	SCR of FREQ	31.6	22.2	60.0	16.9	6.2	35.4	13.3	3.5	31.6
	SCR of EMB	33.3	21.1	60.1	17.9	5.9	35.2	15.4	3.5	30.0
	SCR of FREQ+EMB	$\underline{33.7}$	$\underline{20.1}$	$\underline{60.0}$	$\underline{18.0}$	$\underline{6.1}$	$\underline{33.3}$	$\underline{14.4}$	$\underline{3.5}$	$\underline{31.9}$

Hyper-parameters. The hyper-parameters of SCR control the weight of the SGG loss. We investigate the best hyper-parameter settings as shown in Table 5. The best hyper-parameter settings $\delta = 0.7$ and $\lambda_{skew} = 0.06$ show the generalized performances of the SGG tasks. The smaller λ_{skew} is, the higher mean Recall scores are, while the higher λ_{skew} is, the higher Recall scores are, i.e., the λ_{skew} controls the trade-off between the majority predicates and the minority ones since the smaller H_{skew} tend to assign more weights to the minority predicates in the SCR. The $\delta = 0.7$ shows the best proportion of Re-weighting the predicate samples.

Table 5. The ablation study for the scr hyper-parameters on Recall@100. The underbar represents the best trade-off performances between mRecall and Recall. SCR† denotes SCR of FREQ+EMB.

Relationship Retrieval			**PredCls**		**SGCls**		**SGDet**	
Model	λ_{skew}	δ	mRR	RR	mRR	RR	mRR	RR
VCTree +**SCR**†	0.03	0.7	**35.5**	58.3	**20.1**	31.9	**16.3**	29.4
	$\underline{0.06}$	0.7	$\underline{33.7}$	$\underline{60.0}$	$\underline{18.0}$	$\underline{33.3}$	$\underline{14.4}$	$\underline{31.9}$
	0.08	0.7	32.8	60.4	17.2	34.2	13.5	32.6
	0.06	0.6	33.5	60.6	17.8	34.3	14.2	32.0
	0.06	$\underline{0.7}$	$\underline{33.7}$	$\underline{60.0}$	$\underline{18.0}$	$\underline{33.3}$	$\underline{14.4}$	$\underline{31.9}$
	0.06	0.8	33.3	59.3	17.6	33.0	14.0	30.6

5.7. Qualitative Examples

To demonstrate the effectiveness of the sample-wise SCR Re-weighting, we show the comparison of Recall @100 on PredCls of all predicates based on the SG-Transformer [48] as shown in Figure 4. The SCR of FREQ+EMB achieves a significant performance gain on the overall predicate categories. Moreover, the skew predicate $\hat{\boldsymbol{R}}_{skew}$ serves to estimate not only the predicate candidates that a *subject–object* pair can have but also the target skew S_{skew} which determines the sample weight. Figure 5 includes a *subject–object* pair which have possible predicates, its predicate prediction $\hat{\boldsymbol{R}}_i$, and the target skew S^i_{skew} deduced by the predicate sample estimate $\hat{\boldsymbol{R}}^i_{skew}$. The higher the sample estimate is, the lower the target skew is. To be specific, when one *woman*- $S^{\text{wearing}}_{skew}$ -coat pair have possible predicates such as *backgrounds, has, wearing, in,* etc., we show predicate predictions $\hat{\boldsymbol{R}}_i$ with its target skew S^i_{skew} deduced by its sample estimates $\hat{\boldsymbol{R}}^i_{skew}$. The larger $\hat{R}^i_{skew} \in \hat{\boldsymbol{R}}^i_{skew}$ tends to be a smaller S^i_{skew}, when compared to another *hill*- $S^{\text{covered in}}_{skew}$ -snow pair in a SCR-trained SGG model.

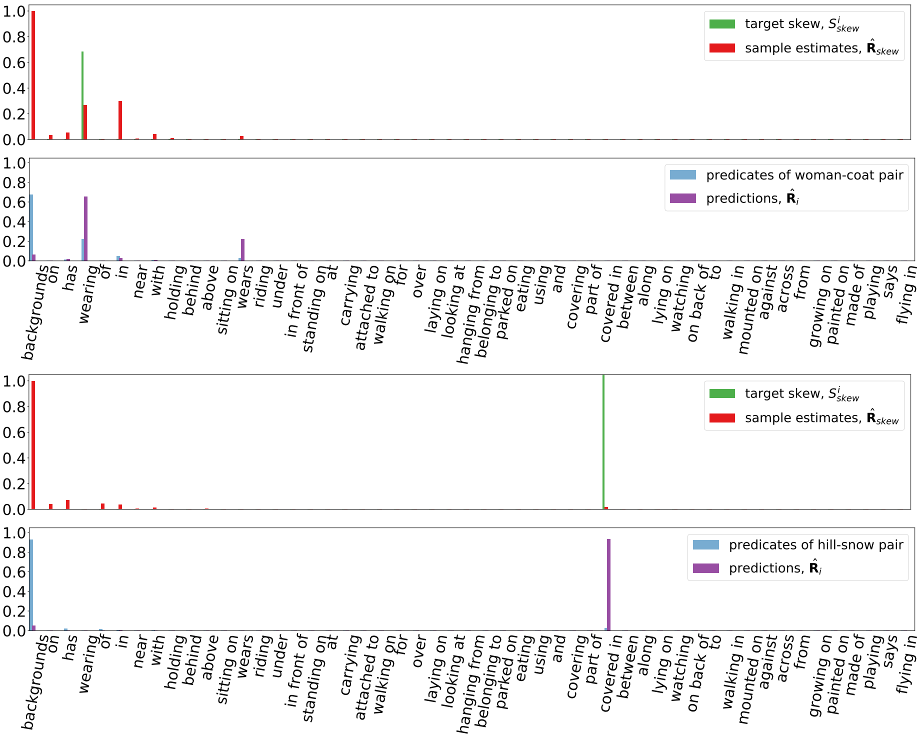


Figure 5. The predicate predictions and sample estimates: in a SCR-trained SGG model, the larger $\hat{R}^i_{skew}$ tends to be a smaller S^i_{skew}, i.e., *woman*- $S^{\text{wearing}}_{skew}$ -coat (**upper**) < *hill*- $S^{\text{covered in}}_{skew}$ -snow (**lower**).

6. Conclusions

In this paper, the unbiased Scene Graph Generation (SGG) algorithm, referred to as Skew Class-Balanced Re-Weighting (SCR), was proposed for considering the unbiased predicate prediction caused by the long-tailed distribution. The prior works focus mainly on alleviating the deteriorating performances of the minority predicate predictions, showing drastic dropping recall scores, i.e., forgetting the majority predicate class. It has not yet properly analyzed the trade-off performances between majority and minority predicates in the given SGG datasets. In this paper, to address the issues leveraged by the skewness of biased predicate predictions, firstly, the SCR estimated the predicate re-weighting coefficient and then re-weighted more to the biased predicates for the better trading-off performances between the majority and the minority predicates. Extensive experiments conducted on the standard Visual Genome dataset and Open Image V4 and V6 showed the SCR's effectiveness and generality with the traditional SGG models.

Author Contributions: Writing—original draft, H.K.; Writing—review & editing, C.D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions), and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2012706).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jarvis, R.A. A perspective on range finding techniques for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **1983**, *PAMI-5*, 122–139. [CrossRef] [PubMed]
2. Forsyth, D.A.; Ponce, J. *Computer Vision: A Modern Approach*; Prentice Hall Professional Technical Reference: Hoboken, NJ, USA, 2002.
3. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
4. Kang, J.-S.; Kang, J.; Kim, J.J.; Jeon, K.W.; Chung, H.J.; Park, B.H. Neural Architecture Search Survey: A Computer Vision Perspective. *Sensors* **2023**, *23*, 1713. [CrossRef] [PubMed]
5. Høye, T.T.; Ärje, J.; Bjerge, K.; Hansen, O.L.; Iosifidis, A.; Leese, F.; Mann, H.M.; Meissner, K.; Melvad, C.; Raitoharju, J. Deep learning and computer vision will transform entomology. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2002545117. [CrossRef] [PubMed]
6. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
7. Sethian, J.A. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*; Cambridge University Press: Cambridge, UK, 1999; Volume 3.
8. Scheuerman, M.K.; Hanna, A.; Denton, E. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. Acm Hum. Comput. Interact.* **2021**, *5*, 1–37. [CrossRef]
9. Verma, T.; De, A.; Agrawal, Y.; Vinay, V.; Chakrabarti, S. Varscene: A deep generative model for realistic scene graph synthesis. In Proceedings of the International Conference on Machine Learning, Guangzhou, China, 18–21 February 2022; PMLR; pp. 22168–22183.
10. Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Mottaghi, A.; Liu, Y.; Topol, E.; Dean, J.; Socher, R. Deep learning-enabled medical computer vision. *NPJ Digit. Med.* **2021**, *4*, 5. [CrossRef]
11. Andriyanov, N.A.; Dementiev, V.E.; Tashlinskiy, A. Detection of objects in the images: From likelihood relationships towards scalable and efficient neural networks. *Comput. Opt.* **2022**, *46*, 139–159. [CrossRef]
12. Dutordoir, V.; van der Wilk, M.; Artemev, A.; Hensman, J. Bayesian Image Classification with Deep Convolutional Gaussian Processes. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Online, 26–28 August 2020; pp. 1529–1539.
13. Papakostas, G.; Diamantaras, K.; Papadimitriou, T. Parallel pattern classification utilizing GPU-based kernelized Slackmin algorithm. *J. Parallel Distrib. Comput.* **2017**, *99*, 90–99. [CrossRef]
14. Joseph, K.J.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5830–5840.

15. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
16. Zhang, Q.; Wipf, D.; Gan, Q.; Song, L. A Biased Graph Neural Network Sampler with Near-Optimal Regret. *arXiv* **2021**, arXiv:2103.01089.
17. Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C.D.; Leskovec, J. GreaseLM: Graph REASoning Enhanced Language Models. In Proceedings of the International Conference on Learning Representations, Online, 25–29 April 2022.
18. Wu, Y.X.; Wang, X.; Zhang, A.; He, X.; Chua, T.S. Discovering Invariant Rationales for Graph Neural Networks. In Proceedings of the International Conference on Learning Representations, Online, 25–29 April 2022.
19. Gao, J.; Ribeiro, B. On the Equivalence Between Temporal and Static Equivariant Graph Representations. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 7052–7076.
20. Yoon, M.; Wu, Y.; Palowitch, J.; Perozzi, B.; Salakhutdinov, R. Scalable and Privacy-enhanced Graph Generative Model for Graph Neural Networks. *arXiv* **2022**, arXiv:2207.04396
21. Andriyanov, N. Application of Graph Structures in Computer Vision Tasks. *Mathematics* **2022**, *10*, 4021. [CrossRef]
22. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
23. Chen, S.; Li, B. Multi-Modal Dynamic Graph Transformer for Visual Grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15534–15543.
24. Ding, Y.; Yu, J.; Liu, Y.; Hu, Y.; Cui, M.; Wu, Q. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-Based Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5089–5098.
25. Lou, C.; Han, W.; Lin, Y.; Zheng, Z. Unsupervised Vision-Language Parsing: Seamlessly Bridging Visual Scene Graphs with Language Structures via Dependency Relationships. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15607–15616.
26. Walmer, M.; Sikka, K.; Sur, I.; Shrivastava, A.; Jha, S. Dual-Key Multimodal Backdoors for Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15375–15385.
27. Koh, J.Y.; Salakhutdinov, R.; Fried, D. Grounding Language Models to Images for Multimodal Generation. *arXiv* **2023**, arXiv:2301.13823.
28. Iwamura, K.; Kasahara, J.L.; Moro, A.; Yamashita, A.; Asama, H. Image Captioning Using Motion-CNN with Object Detection. *Sensors* **2021**, *21*, 1270. [CrossRef]
29. Liu, H.; Yan, N.; Mortazavi, M.; Bhanu, B. Fully convolutional scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11546–11556.
30. Cong, Y.; Yang, M.Y.; Rosenhahn, B. RelTR: Relation Transformer for Scene Graph Generation. *arXiv* **2022**, arXiv:2201.11460.
31. Xu, D.; Zhu, Y.; Choy, C.B.; Li, F.-F. Scene graph generation by iterative message passing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5410–5419.
32. Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3076–3086.
33. Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; Wang, X. Scene graph generation from objects, phrases and region captions. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1261–1270.
34. Li, Y.; Ouyang, W.; Wang, X. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *arXiv* **2017**, arXiv:1702.07191.
35. Hwang, S.J.; Ravi, S.N.; Tao, Z.; Kim, H.J. Collins, M.D.; Singh, V. Tensorize, factorize and regularize: Robust visual relationship learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1014–1023.
36. Li, Y.; Ouyang, W.; Zhou, B.; Shi, J.; Zhang, C.; Wang, X. Factorizable net: An efficient subgraph-based framework for scene graph generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 335–351.
37. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Graph r-cnn for scene graph generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–685.
38. Yin, G.; Sheng, L.; Liu, B.; Yu, N.; Wang, X.; Shao, J.; Change Loy, C. Zoom-net: Mining deep feature interactions for visual relationship recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 322–338.
39. Woo, S.; Kim, D.; Cho, D.; Kweon, I.S. LinkNet: Relational Embedding for Scene Graph. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 558–568.
40. Wang, W.; Wang, R.; Shan, S.; Chen, X. Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8188–8197.
41. Tang, K.; Zhang, H.; Wu, B.; Luo, W.; Liu, W. Learning to compose dynamic tree structures for visual contexts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 155–20 June 2019; pp. 6619–6628.

42. Chen, L.; Zhang, H.; Xiao, J.; He, X.; Pu, S.; Chang, S.F. Counterfactual Critic Multi-Agent Training for Scene Graph Generation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4613–4623.
43. Tang, K.; Niu, Y.; Huang, J.; Shi, J.; Zhang, H. Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3716–3725.
44. Desai, A.; Wu, T.Y.; Tripathi, S.; Vasconcelos, N. Learning of visual relations: The devil is in the tails. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15404–15413.
45. Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eledath, J.; Medioni, G.; Sigal, L. Energy-Based Learning for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13936–13945.
46. Chen, T.; Yu, W.; Chen, R.; Lin, L. Knowledge-Embedded Routing Network for Scene Graph Generation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
47. Lin, X.; Ding, C.; Zeng, J.; Tao, D. Gps-net: Graph property sensing network for scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3746–3753.
48. Yu, J.; Chai, Y.; Hu, Y.; Wu, Q. CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation. *arXiv* **2020**, arXiv:2009.07526.
49. Li, R.; Zhang, S.; Wan, B.; He, X. Bipartite Graph Network With Adaptive Message Passing for Unbiased Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11109–11119.
50. Yan, S.; Shen, C.; Jin, Z.; Huang, J.; Jiang, R.; Chen, Y.; Hua, X.S. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In Proceedings of the 28th ACM International Conference on Multimedia, Online, 12–16 October 2020; pp. 265–273.
51. Chiou, M.J.; Ding, H.; Yan, H.; Wang, C.; Zimmermann, R.; Feng, J. Recovering the unbiased scene graphs from the biased ones. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 1581–1590.
52. Guo, Y.; Gao, L.; Wang, X.; Hu, Y.; Xu, X.; Lu, X.; Shen, H.T.; Song, J. From general to specific: Informative scene graph generation via balance adjustment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16383–16392.
53. Li, W.; Zhang, H.; Bai, Q.; Zhao, G.; Jiang, N.; Yuan, X. PPDL: Predicate Probability Distribution Based Loss for Unbiased Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19447–19456.
54. Li, L.; Chen, L.; Huang, Y.; Zhang, Z.; Zhang, S.; Xiao, J. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 18869–18878.
55. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
56. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The open images dataset v4. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [CrossRef]
57. Zhan, Y.; Yu, J.; Yu, T.; Tao, D. On Exploring Undetermined Relationships for Visual Relationship Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5128–5137.
58. Sadeghi, M.A.; Farhadi, A. Recognition using visual phrases. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1745–1752.
59. Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; Ling, M. Scene graph generation with external knowledge and image reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1969–1978.
60. Yu, R.; Li, A.; Morariu, V.I.; Davis, L.S. Visual relationship detection with internal and external linguistic knowledge distillation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1974–1982.
61. Yang, G.; Zhang, J.; Zhang, Y.; Wu, B.; Yang, Y. Probabilistic Modeling of Semantic Ambiguity for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12527–12536.
62. Lyu, X.; Gao, L.; Guo, Y.; Zhao, Z.; Huang, H.; Shen, H.T.; Song, J. Fine-Grained Predicates Learning for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19467–19475.
63. Dong, X.; Gan, T.; Song, X.; Wu, J.; Cheng, Y.; Nie, L. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 19427–19436.
64. Goel, A.; Fernando, B.; Keller, F.; Bilen, H. Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15596–15606.

65. Li, Y.; Yang, X.; Xu, C. Dynamic Scene Graph Generation via Anticipatory Pre-Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13874–13883.
66. Teng, Y.; Wang, L. Structured sparse r-cnn for direct scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19437–19446.
67. Zhang, A.; Yao, Y.; Chen, Q.; Ji, W.; Liu, Z.; Sun, M.; Chua, T.S. Fine-Grained Scene Graph Generation with Data Transfer. *arXiv* **2022**, arXiv:2203.11654
68. Lin, X.; Ding, C.; Zhang, J.; Zhan, Y.; Tao, D. RU-Net: Regularized Unrolling Network for Scene Graph Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19457–19466.
69. Yang, J.; Ang, Y.Z.; Guo, Z.; Zhou, K.; Zhang, W.; Liu, Z. Panoptic Scene Graph Generation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–196.
70. Deng, Y.; Li, Y.; Zhang, Y.; Xiang, X.; Wang, J.; Chen, J.; Ma, J. Hierarchical Memory Learning for Fine-Grained Scene Graph Generation. *arXiv* **2022**, arXiv:2203.06907.
71. He, T.; Gao, L.; Song, J.; Li, Y.F. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 56–73.
72. Brown, S. Measures of shape: Skewness and kurtosis. *Retrieved August* **2011**, *20*, 2012.
73. Zhu, Z.; He, X.; Qi, G.; Li, Y.; Cong, B.; Liu, Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf. Fusion* **2023**, *91*, 376–387. [CrossRef]
74. Zellers, R.; Yatskar, M.; Thomson, S.; Choi, Y. Neural motifs: Scene graph parsing with global context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5831–5840.
75. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277.
76. Kang, H.; Vu, T.; Yoo, C.D. Learning Imbalanced Datasets With Maximum Margin Loss. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1269–1273.
77. Zhang, J.; Shih, K.J.; Elgammal, A.; Tao, A.; Catanzaro, B. Graphical contrastive losses for scene graph parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 155–20 June 2019; pp. 11535–11543.
78. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in neural information processing systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
79. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
80. Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; Girshick, R. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2930–2939.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.