Learning to Generate an Unbiased Scene Graph by Using Attribute-Guided Predicate Features

Lei Wang, Zejian Yuan, Badong Chen*

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
leiwangmail@stu.xjtu.edu.cn, {yuan.ze.jian, chenbd}@mail.xjtu.edu.cn

Abstract

Scene Graph Generation (SGG) aims to capture the semantic information in an image and build a structured representation, which facilitates downstream tasks. The current challenge in SGG is to tackle the biased predictions caused by the long-tailed distribution of predicates. Since multiple predicates in SGG are coupled in an image, existing data re-balancing methods cannot completely balance the head and tail predicates. In this work, a decoupled learning framework is proposed for unbiased scene graph generation by using attribute-guided predicate features to construct a balanced training set. Specifically, the predicate recognition is decoupled into Predicate Feature Representation Learning (PFRL) and predicate classifier training with a class-balanced predicate feature set, which is constructed by our proposed Attribute-guided Predicate Feature Generation (A-PFG) model. In the A-PFG model, we first define the class labels of ⟨subject-predicate-object⟩ and corresponding visual feature as attributes to describe a predicate. Then the predicate feature and the attribute embedding are mapped into a shared hidden space by a dual Variational Auto-encoder (VAE), and finally the synthetic predicate features are forced to learn the contextual information in the attributes via cross reconstruction and distribution alignment. To demonstrate the effectiveness of our proposed method, our decoupled learning framework and A-PFG model are applied to various SGG models. The empirical results show that our method is substantially improved on all benchmarks and achieves new state-of-the-art performance for unbiased scene graph generation. Our code is available at https://github.com/wanglei0618/A-PFG.

Introduction

Scene Graph Generation (SGG) aims to capture the semantic information in an image by detecting visual objects and their relationships, as shown in Figure 1(a). The structured ⟨subject-predicate-object⟩ triplet of the scene graph can bridge the gap between vision and language, which is beneficial for many downstream visual understanding tasks, such as image captioning (Zhong et al. 2020), visual question answering (Hudson and Manning 2019), image retrieval (Johnson et al. 2015), and image generation(Johnson, Gupta, and Fei-Fei 2018).

*Corresponding Author.

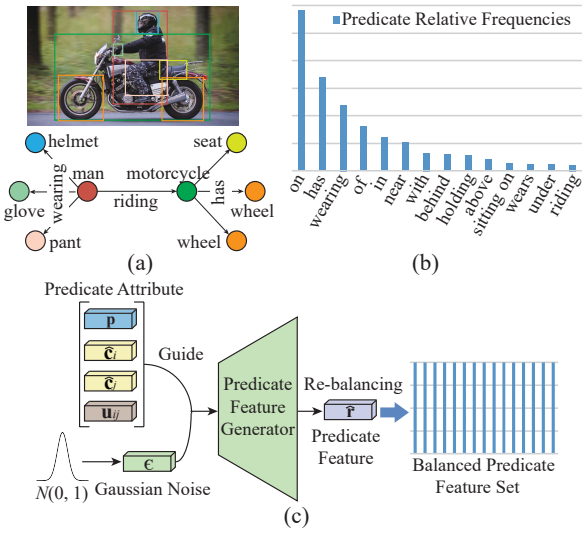


Figure 1: (a) The head ("has, wearing") and tail ("riding") predicates are coupled in the image. (b) The long-tailed distribution of predicates in the Visual Genome (VG) dataset. (c) Our approach constructs a class-balanced predicate feature training set by the attribute-guided predicate feature generation model for unbiased scene graph generation.

However, the current scene graph generation suffers from the problem of biased predictions due to the long-tailed distribution of predicates in the dataset. As shown in Figure 1(b), the annotations in the commonly used Visual Genome (VG) dataset (Krishna et al. 2017) mainly concentrate on a few head predicates and are much sparse on most tail predicates. The long-tailed distribution results in coarse-grained head classes (e.g., on/has) dominating the predictions, while tail classes (e.g., wears/riding) containing more fine-grained semantic information are difficult to be predicted.

Recently, many de-biasing strategies have been introduced for unbiased scene graph generation, such as causal analysis (Tang et al. 2020) and loss re-weighting (Yan et al. 2020; Yu et al. 2020). However, the data re-balancing methods commonly used in long-tailed recognition (Zhou et al. 2020), such as re-sampling and data augmentation, are rarely applied to SGG. This is attributed to the fact that the head

and tail predicates are coupled in the image. As shown in Figure 1(a), the tail predicate ("riding") and multiple head predicates ("has, wearing") are coupled in the image. Consequently, the image-level data re-balancing methods will introduce more head predicates when enhancing the tail predicate in that image and cannot obtain a balanced predicate distribution. Therefore, we propose a decoupled learning framework for unbiased scene graph generation to separate the different predicates in an image to achieve predicate features re-balancing. Specifically, the predicate recognition is decoupled into Predicate Feature Representation Learning (PFRL) and predicate classifier training with a balanced predicate feature set. We first learn the representation of predicate features in the original imbalanced dataset based on our PFRL model, and then construct a class-balanced set by our predicate feature generation model to fine-tune the predicate classifier. Equipped with our decoupled learning framework, data re-balancing strategies for unbiased SGG can be implemented based on the predicate features.

Since the predicates (e.g., "riding") in scene graphs are built within the context of subjects (e.g., "man") and objects (e.g., "motorcycle"), prevalent generative methods (Schonfeld et al. 2019; Xian et al. 2019) that use only class labels as conditions to generate predicate features are inadequate. Therefore, we propose an Attribute-guided Predicate Feature Generation (A-PFG) model, which can guide synthetic predicate features to learn the contextual information of predicates via attributes. Specifically, the class labels of ⟨subject-predicate-object⟩ and corresponding visual feature are define as attributes to describe a predicate. Then, the A-PFG model maps the predicate feature and the attribute embedding into a shared hidden space by a dual Variational Auto-encoder (VAE), and forces the predicate feature generator to learn the common information contained in predicate features and attributes via cross reconstruction and distribution alignment.

We demonstrate the efficacy of our decoupled learning framework and A-PFG model by applying them to multiple representative SGG models. The results show that the proposed method improves significantly on all benchmarks and achieves new state-of-the-art performance on our PFRL model for unbiased scene graph generation. The main contributions of this work are summarized as follows:

- A decoupled learning framework is proposed to solve the multiple predicates coupling problem. With our framework, data re-balancing can be achieved based on predicate features for unbiased scene graph generation.
- An attribute-guided predicate feature generation model is developed to synthesize the tail predicate features to alleviate the long-tail problem.

Related Works

Scene Graph Generation Early works (Xu et al. 2017; Zellers et al. 2018; Tang et al. 2019) in scene graph generation were devoted to sophisticated architecture design or contextual feature fusion strategies. However, those methods suffer from the biased predictions induced by the long-tailed distribution of predicates. Therefore, many recent works (Desai et al. 2021; Chen et al. 2022) attempt to address the long-tail problem for unbiased scene graph generation, for example, TDE (Tang et al. 2020) eliminates prediction bias through causal analysis, Cogtree (Yu et al. 2020) utilizes cognitive tree loss to achieve unbiased prediction, BGNN (Li et al. 2021) balances the head and tail predicates by bi-level re-sampling. Unlike the above works of unbiased SGG, which directly recognize the predicates of object pairs in the imbalanced dataset, we extract predicate features by predicate feature representation learning, and then fine-tune the predicate classifier with a balanced predicate feature set constructed by our generative model.

Imbalanced Learning Imbalanced learning methods can be broadly divided into two categories: **Data Re-balancing** is a prevalent strategy to balance the data by oversampling and undersampling. Oversampling can be implemented with duplicate samples (Zhou et al. 2020) or synthetic data (Xian et al. 2019). **Loss Re-weighting** forces the classifier to focus on tail classes by assigning higher costs to tail samples. The weights for each class can be determined by prior knowledge, for example, class balanced weight (Cui et al. 2019) uses the inverse effective number of samples. In this work, the proposed method lies in the data re-balancing, where a decoupled learning framework is presented to address the problem of inaccessible data re-balancing in SGG due to multiple predicates coupled in the image. Then a balanced training set can be constructed based on the predicate features to achieve accurate recognition of the tail predicates.

Generative Model The generative model aims to learn the probability distribution of data points to perform data augmentation by random sampling. Then the imbalance classification problem can be solved with a balanced dataset constructed by the synthetic samples. The commonly used generative model consists of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and Variational Auto-encoder (VAE) (Kingma and Welling 2013). Many generative models (Schonfeld et al. 2019; Xian et al. 2019) have been applied to imbalance learning by synthesizing visual features. All the above generative models use class labels or fixed class attributes as conditions to synthesize images or visual features, but since the predicate in SGG is built in contextual information, an attribute-guided predicate feature generation model is developed in this work, which synthesizes predicate features that contain contextual information by attribute guidance.

Method

Problem Formulation Given an image $\mathcal{I}$, the task of scene graph generation is to predict a scene graph $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$, where $\mathcal{O} = \{o_i\}_{i=1}^{n}$ is a set of n objects in the image and $\mathcal{R} = \{r_k\}_{k=1}^{m}$ is the set of relationships between m pairs of objects. The set of objects $\mathcal{O}$ can be further decomposed into a set of bounding boxes $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{n}$ and a set of class labels $\mathcal{C} = \{c_i\}_{i=1}^{n}$, and then the possibility of generating a scene graph $\mathcal{G}$ from an image $\mathcal{I}$ can be formulated as

$$Pr(\mathcal{G}|\mathcal{I}) = Pr(\mathcal{B}|\mathcal{I})Pr(\mathcal{C}|\mathcal{I},\mathcal{B})Pr(\mathcal{R}|\mathcal{I},\mathcal{B},\mathcal{C}) \quad (1)$$

where $Pr(\mathcal{B}|\mathcal{I})$ represents object proposals generation from input image, $Pr(\mathcal{C}|\mathcal{I},\mathcal{B})$ and $Pr(\mathcal{R}|\mathcal{I},\mathcal{B},\mathcal{C})$ denote object classification and relationship prediction, respectively. In addition, since each relationship $r_k = \langle o_i, p_k, o_j \rangle$ is a triplet $\langle$subject,predicate,object$\rangle$ format, the relationship prediction $Pr(\mathcal{R}|\mathcal{I},\mathcal{B},\mathcal{C})$ is equivalent to recognize the predicate label p_k between subject o_i and object o_j.

Method Overview We first use the Predicate Feature Representation Learning (PFRL) model to extract predicate features, and then synthesis tail predicate features based on the Attribute-guided Predicate Feature Generation (A-PFG) model to construct a balanced training set to fine-tune the predicate classifier.

Predicate Feature Representation Learning

The architecture of our PFRL model is summarized in Figure 2, which is composed of the object detector, the object encoder, the predicate encoder, and the predicate classifier. We utilize a pre-trained Faster RCNN (Ren et al. 2015) as object detector to generate a set of object proposals from image $\mathcal{I}$. For each object proposal, it provides a visual feature $\mathbf{v}_i$, the bounding box coordinates $\mathbf{b}_i$, and an initial object label prediction c_i^0.

Object Encoder (OE) Our object encoder used to predict object labels consists of a stack of transformer encoders (Vaswani et al. 2017). Each transformer encoder accepts input encodings and weighs their relevance to each other to generate output encodings by self-attention mechanism. With the object proposals, the refined object feature $\mathbf{f}_i$ used to predict the object labels is formulated as

$$\mathbf{f}_i = \text{OE}([\mathbf{v}_i, pos(\mathbf{b}_i), emb(c_i^0)]) \tag{2}$$

where $[\cdot,\cdot]$ denotes the concatenation operation, pos is a fully-connected layer for object position encoding, emb is a pre-trained Glove language model to acquire the word embedding. Subsequently, the final object label prediction $\hat{c}_i$ can be calculated by the refined object feature $\mathbf{f}_i$.

Predicate Encoder (PE) In order to encode the predicate features, we first use the predicate encoder, which is also composed of a stack of transformer encoders, to obtain the context-aware object feature $\mathbf{e}_i$ as follows:

$$\mathbf{e}_i = \text{PE}([\mathbf{v}_i, \mathbf{f}_i, emb(\hat{c}_i)]) \tag{3}$$

Then the predicate feature $\mathbf{r}_k$ between subject feature $\mathbf{e}_i$ and object features $\mathbf{e}_j$ are given as

$$\mathbf{r}_k = \text{LeakyReLu}(W_r[W_u[\mathbf{e}_i, \mathbf{e}_j]\mathbf{u}_{ij}, \mathbf{e}_i, \mathbf{e}_j]) \tag{4}$$

where W_r is a linear transformation layer used to compress the predicate features, $\mathbf{u}_{ij}$ is the union visual feature of subject o_i and object o_j, and W_u is a fully-connected layer to weight the union visual feature.

Predicate Classifier (PC) Finally, the predicate feature $\mathbf{r}_k$ is used to recognize predicate label $\hat{p}_k$ as follows:

$$\hat{\mathbf{p}}_k = \text{Softmax}(W_{cls}\mathbf{r}_k) \tag{5}$$

where W_{cls} is the parameter of predicate classifier.

Since the prediction of the predicate classifier trained on the original imbalanced dataset is dominated by the head predicates, a balanced set of predicate features needs to be built to fine-tune the classifier for unbiased prediction. Therefore, we extract the set of all predicate features $\mathcal{R}^f = \{\mathbf{r}_k\}_{k=1}^{N_f}$ from the original dataset based on the PFRL model to train the A-PFG model and construct the balanced set.

Attribute-Guided Predicate Features Generation

To provide contextual guidance for predicate feature generation, the labels of $\langle$subject-predicate-object$\rangle$ and corresponding visual feature are defined as attributes to synthesize predicate features. Then, the attributes are compressed as predicate attribute embedding input to the A-PFG model as follows:

$$\mathbf{a}_k = W_a[\varphi(p_k), \varphi(\hat{c}_i), \varphi(\hat{c}_j), \mathbf{u}_{ij}] \tag{6}$$

where p_k is the label of predicate, $\hat{c}_i$ and $\hat{c}_j$ are the predicted labels of subject and object, respectively, $\mathbf{u}_{ij}$ is the union visual features, W_a is a fully-connected layer to encode predicate attributes, and φ is a one-hot function.

A-PFG Model As shown in Figure 3, the A-PFG model contains two Variational Auto-encoder (VAE) (Kingma and Welling 2013). The attribute VAE uses the Predicate Attribute Encoder (PAE) to receive the attribute embedding $\mathbf{a}$ to map into an attribute latent variable $\mathbf{z}_a = \text{PAE}(\mathbf{a})$ and reconstruct attribute embedding $\hat{\mathbf{a}}_a = \text{PAD}(\mathbf{z}_a)$ by the Predicate Attribute Decoder (PAD). The feature VAE uses Predicate Feature Encoder (PFE) to receive the predicate feature $\mathbf{r}$ to map into a feature latent variable $\mathbf{z}_r = \text{PFE}(\mathbf{r})$ and reconstruct predicate feature $\hat{\mathbf{r}}_r = \text{PFD}(\mathbf{z}_r, \mathbf{a})$ by the Predicate Feature Decoder (PFD) conditioned on $\mathbf{a}$.

To capture the common information between predicate features and attributes, the A-PFG model uses cross-reconstruction and distribution alignment to learn the representation of cross-modal data in the shared latent space. Therefore, each decoder in A-PFG decodes the latent variable from another modal encoder to obtain the cross reconstruction. The cross-reconstructions of predicate attribute embedding and predicate feature are denoted as $\hat{\mathbf{a}}_r = \text{PAD}(\mathbf{z}_r)$ and $\hat{\mathbf{r}}_a = \text{PFD}(\mathbf{z}_a, \mathbf{a})$, respectively.

A-PFG Training All predicate feature set and corresponding attributes are used to train the A-PFG model. The training loss $\mathcal{L}_{A\text{-}PFG}$ is composed of the basic VAE loss $\mathcal{L}_{VAE}$, the cross-reconstruction loss $\mathcal{L}_{C\text{-}Re}$ and the distribution alignment loss $\mathcal{L}_{DA}$ as follows:

$$\mathcal{L}_{A\text{-}PFG} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{C\text{-}Re} + \delta\mathcal{L}_{DA} \tag{7}$$

where γ and δ are hyper-parameters weighting $\mathcal{L}_{C\text{-}Re}$ and $\mathcal{L}_{DA}$, respectively. The basic VAE loss is the sum of attribute VAE loss and feature VAE loss as follows:

$$\begin{aligned}\mathcal{L}_{VAE} =& \|\mathbf{a} - \hat{\mathbf{a}}_a\| + \|\mathbf{r} - \hat{\mathbf{r}}_r\| \\ &- \beta\Big(D_{KL}(\text{PAE}(\mathbf{z}_a|\mathbf{a})\|p_\theta(\mathbf{z}_a)) \\ &+ D_{KL}(\text{PFE}(\mathbf{z}_r|\mathbf{r})\|p_\theta(\mathbf{z}_r))\Big)\end{aligned} \tag{8}$$

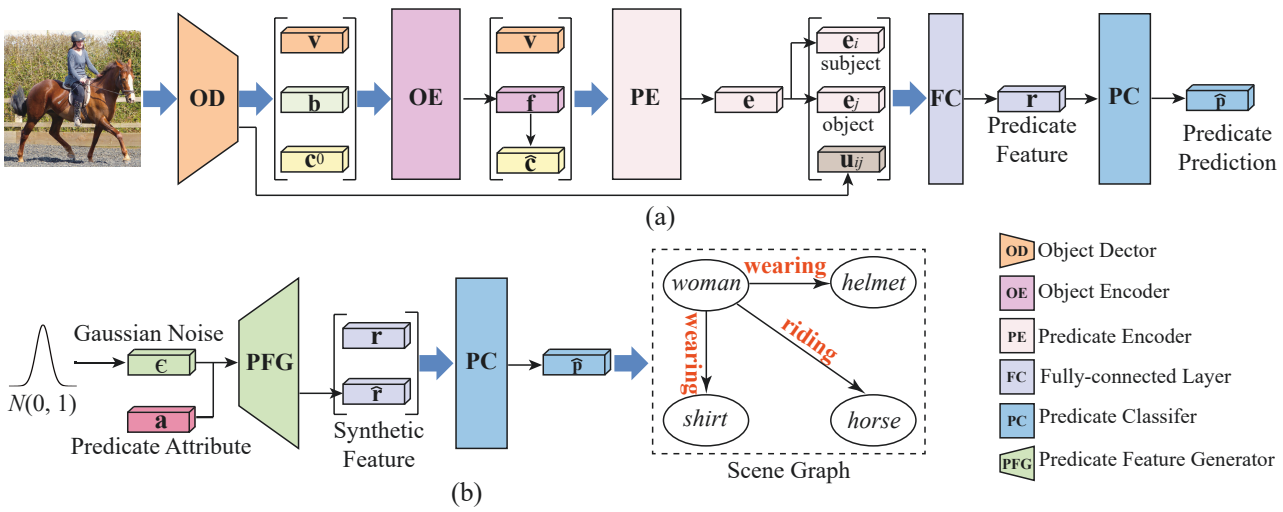


Figure 2: Overall pipeline of our decoupled learning framework. (a) The Predicate Feature Representation Learning (PFRL) model; (b) The predicate classifier fine-tuning for unbiased scene graph generation by using a balanced predicate feature training set. The Predicate Feature Generator (PFG) corresponds to the Predicate Feature Decoder (PFD) in Figure 3.

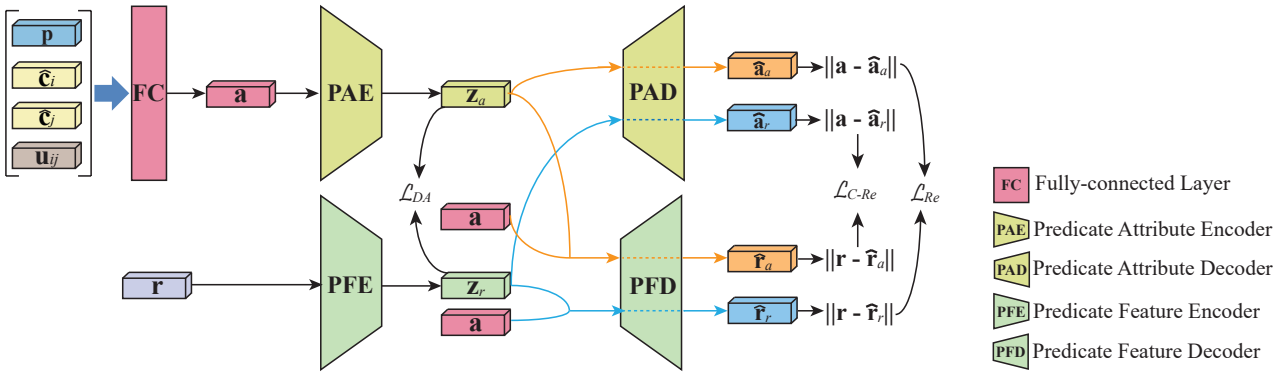


Figure 3: Overview of the Attribute-guide Predicate Feature Generation (A-PFG) model. The colors orange and blue indicate the feature generation based on the latent variables of attribute embedding **a** and predicate feature **r**, respectively.

where the first two terms is the reconstruction error $\mathcal{L}_{Re}$ and the last term indicates the unpacked Kullback–Leibler (KL) divergence, the prior $p_\theta(\mathbf{z}_a)$ and $p_\theta(\mathbf{z}_r)$ are assumed to be $\mathcal{N}(0,1)$, and the hyper-parameter β provides a trade-off between reconstruction loss and the KL-divergence. The cross-reconstruction loss $\mathcal{L}_{C\text{-}Re}$ is calculated as:

$$\mathcal{L}_{C\text{-}Re} = \|\mathbf{a} - \hat{\mathbf{a}}_r\| + \|\mathbf{r} - \hat{\mathbf{r}}_a\| \tag{9}$$

The A-PFG also aligns the latent variable distributions of predicate features and attributes by minimizing the Wasserstein distance. Since the latent variables $\mathbf{z}_r$ and $\mathbf{z}_a$ are Gaussian distribution, the distribution alignment loss $\mathcal{L}_{DA}$ can be simplified (Schonfeld et al. 2019) and written as:

$$\mathcal{L}_{DA} = (\|\boldsymbol{\mu}_{\mathbf{z}_r} - \boldsymbol{\mu}_{\mathbf{z}_a}\|^2 + \|\boldsymbol{\sigma}_{\mathbf{z}_r}^{\frac{1}{2}} - \boldsymbol{\sigma}_{\mathbf{z}_a}^{\frac{1}{2}}\|_{\mathrm{F}}^2)^{\frac{1}{2}} \tag{10}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are, respectively, the mean and variance of the latent variables $\mathbf{z}_r$ and $\mathbf{z}_a$.

Predicate Feature Generation Since predicate and attribute are one-to-one correspondence, as shown in Figure 2 (b), we first randomly sample a predicate attribute embedding $\mathbf{a}^{(p_k)}$ from the class p_k as condition, and then input it into the Predicate Feature Generator (PFG) along with random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$ to synthesize a predicate feature of p_k as follows:

$$\hat{\mathbf{r}}^{(p_k)} = \mathrm{PFG}(\boldsymbol{\epsilon}, \mathbf{a}^{(p_k)}) \tag{11}$$

Decoupled Training Strategy

Our proposed method decouples the SGG training into a two-stage process:

PFRL Training The PFRL model is trained with the original imbalanced dataset, and the training loss $\mathcal{L}_{rep}$ is formulated as:

$$\mathcal{L}_{rep} = \mathcal{L}_{CE}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{L}_{CE}(\mathbf{p}, \hat{\mathbf{p}}) \tag{12}$$

where $\mathcal{L}_{CE}(\mathbf{c}, \hat{\mathbf{c}})$ and $\mathcal{L}_{CE}(\mathbf{p}, \hat{\mathbf{p}})$ are the cross-entropy loss functions of object classification and predicate classification, respectively.

Predicate Classifier Fine-tuning The predicate classifier is fine-tuned using the balanced predicate feature training set $\mathcal{D}^{bal}$ as follows:

$$\mathcal{D}^{bal} = \{\mathcal{R}^{syn}, \bar{\mathcal{R}}^{f}, \mathcal{R}^{b}\} \tag{13}$$

where $\mathcal{R}^{syn}$ is a synthetic predicate feature set generated by the A-PFG model, $\bar{\mathcal{R}}^{f}$ is the pruned predicates feature set, which removes the instances beyond threshold N_p of each class in the set $\mathcal{R}^{f}$. Since background relations (i.e., the classification of a predicate is irrelevant) need to be recognized in SGG, the background feature set $\mathcal{R}^{b} = \{\mathbf{r}_k^b\}_{k=1}^{N_b}$, which is extracted by the PFRL model based on the original dataset, is used to balance the predicates and background, wherein N_b is the number of background features. Finally, the predicate classifier is fine-tuned in $\mathcal{D}^{bal}$ with the cross-entropy loss function $\mathcal{L}_{ft} = \mathcal{L}_{CE}(\mathbf{p}, \hat{\mathbf{p}})$.

Experiments

Experimental Settings

Dataset Following previous works (Zellers et al. 2018; Tang et al. 2019; Yu et al. 2020; Li et al. 2021), the proposed method and recent methods are evaluated on the widely used subset of Visual Genome dataset (i.e., VG150) (Krishna et al. 2017), which includes the most frequent 150 object classes and 50 predicate classes. Then, we divide it into 70% training set, 30% testing set, and 5k images selected from the training set for validation.

Task & Evaluation Metrics The SGG contains three tasks (Xu et al. 2017): 1) Predicate Classification (**PredCls**) predicts the pairwise predicates with the ground-truth object labels and bounding boxes. 2) Scene Graph Classification (**SGCls**) predicts the object labels and their pairwise predicates with the ground-truth bounding boxes. 3) Scene Graph Detection (**SGDet**) detects all the objects in an image and predicts their bounding boxes, labels, and predicates.

Due to the imbalanced distribution of predicates in the VG dataset, the early evaluation metric Recall@K (R@K) is easily dominated by the head classes. Therefore, we follow (Tang et al. 2020) to adopt mean Recall@K (mR@K) to evaluate the unbiased SGG, mR@K calculates R@K for each class independently, and then average the results. The results of Recall@K are reported in the Suppl.

Implementation Details We employ the widely used pretrained Faster RCNN with ResNeXt-101-FPN provided by (Tang et al. 2020) as object detector. For the PFRL model, the number of object encoder layers and predicate encoder layers are 4 and 2, respectively, the dimension of the predicate feature is 1024. For classifier fine-tuning, the number of instances for each predicate class N_p is 5000, and the number of background features N_b is 5×10^6. For the A-PFG model, the encoders and decoders are 3-layers fully-connected network, with each layer followed by the LeakyReLu activation function, the dimension of the predicate attribute embedding is 1024, and the dimensions of the latent variables $\mathbf{z}_r$ and $\mathbf{z}_a$ are 256. The hyperparameter γ is 1, β and δ are increased 0.5 per epoch. The PFRL is implemented on two NVIDIA 3090 GPUs with batch size 16 and learning rate 0.001, and the classifier is fine-tuned with batch size 16 and learning rate 2×10^{-6}. The A-PFG model is trained for 200 epochs with batch size 64 and learning rate 2×10^{-4}.

Comparison with State of the Arts

In this subsection, the biased scene graph generation methods that focus on feature representation learning is first compared, including KERN, Motif, VCTree, SG-Trans and our PFRL. However, the performance of them are unsatisfactory due to the imbalanced distribution of predicates, and their mean recall is relatively low as shown in Table 1.

Then, an in-depth comparison between proposed method and several debiasing methods is performed to further demonstrate the effectiveness. The proposed decoupled learning framework and A-PFG model are applied to four baseline models: Motif, VCTree, SG-Trans, and our PFRL. Compare to other debiasing approaches (TDE, EBM, BPL+SA, GCL, and RTPB) in Table 1, our method can achieve consistent improvement in all benchmarks. For example, the improvement on the Motifs for mR@100 across three tasks are 143% (from 17.3 to 42.0), 154% (from 9.3 to 23.6) and 126% (from 8.1 to 18.3), respectively. The A-PFG outperforms on most of the metrics with the same representation model. Compared to specific benchmark methods, such as PCPL, GAN+Gra-N, BGNN, Trans+BPL+SA, PCL, SHA+GCL, and DTrans+RTPB, our approach can also achieve outperforming performance. In addition, GAN+Gra-N also uses a generative model to synthesize visual features for zero-/few-shot SGG, but it does not achieve satisfactory performance on the mean Recall metric. Importantly, our full model PFRL+A-PFG achieves new state-of-the-art performance with mR@100 of 44.3, 25.8, and 21.7 in PredCls, SGCls, and SGDet, respectively.

Class-Wise Performance Improvement To analyze the performance improvement of different frequency classes, the 50 predicate classes are divided into three parts based on their counts: Head (16), Body (17), and Tail (17). Figure 4 shows the mR@100 performance comparison between PFRL and PFRL+A-PFG on these partitions. Comparing the performance improvements across all tasks, it can be observed that PFRL+A-PFG gains a huge boost in body and tail classes with only a slight sacrifice of the head classes. For example, in task SGDet, Body and Tail boost 16.5 (from 6.3 to 22.8) and 21.5 (from 0.1 to 21.6), respectively, while Head only drops 1.6 (from 22.1 to 20.5). This demonstrates that the proposed method can achieve a balance among all the classes for unbiased SGG.

Ablation Studies

Decoupled Learning Framework Comparing the data generation methods PFG- and the re-sampling method Rs-bi in Table 2, PFG- is significantly better than Rs-bi. For example, the mR@100 of PFG-p on task SGDet is 21.4, while Rs-bi is only 14.9. This is because Rs-bi is an image-based bi-

Method	PredCls			SGCls			SGDet		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
KERN(Chen et al. 2019)	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
PCPL(Yan et al. 2020)	-	35.2	37.8	-	18.6	19.6	-	9.5	11.7
GAN+Gra-N(Knyazev et al. 2021)	-	28.6	-	-	-	18.3	-	-	9.9
BGNN(Li et al. 2021)	-	30.4	32.9	-	14.3	16.5	-	10.7	12.6
Trans+BPL+SA(Guo et al. 2021)	26.7	31.9	34.2	15.7	18.5	19.4	11.4	14.8	17.1
PCL(Tao et al. 2022)	33.2	36.3	39.2	17.9	20.7	21.8	11.1	15.2	18.3
SHA+GCL(Dong et al. 2022)	35.6	41.6	44.0	19.6	23.0	24.3	14.2	17.9	20.9
DTrans+RTPB(Chen et al. 2022)	30.3	36.2	38.1	19.1	21.8	22.8	12.7	16.5	19.0
MOTIFS†(Zellers et al. 2018)	12.7	16.0	17.3	7.0	8.7	9.3	5.0	6.8	8.1
+TDE(Tang et al. 2020)	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
+EBM(Suhail et al. 2021)	14.2	18.0	19.5	8.2	10.2	11.0	5.7	7.7	9.3
+BPL+SA(Guo et al. 2021)	24.8	29.7	31.7	14.0	16.5	17.5	10.7	13.5	15.6
+GCL(Dong et al. 2022)	30.5	36.1	38.2	18.0	20.8	21.8	**12.9**	**16.8**	**19.3**
+RTPB(Chen et al. 2022)	28.8	35.3	37.7	16.3	19.4	20.6	9.7	13.1	15.5
+A-PFG (ours)	**33.5**	**39.9**	**42.0**	**19.8**	**22.7**	**23.6**	11.9	15.8	18.3
VCTree†(Tang et al. 2019)	13.2	16.6	17.9	6.6	8.1	8.6	5.2	7.0	8.2
+TDE(Tang et al. 2020)	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
+EBM(Suhail et al. 2021)	19.9	26.7	30.0	13.9	18.2	20.5	7.1	9.7	11.6
+BPL+SA(Guo et al. 2021)	26.2	30.6	32.6	17.2	20.1	21.2	10.6	13.5	15.7
+GCL(Dong et al. 2022)	31.4	37.1	39.1	19.5	22.5	23.5	11.9	15.2	17.5
+RTPB(Chen et al. 2022)	27.3	33.4	35.6	**20.6**	**24.5**	**25.8**	9.6	12.8	15.1
+A-PFG (ours)	**34.6**	**41.7**	**43.8**	17.3	20.4	21.8	**12.3**	**16.6**	**19.1**
SG-Trans(Yu et al. 2020)	14.8	19.2	20.5	8.9	11.6	12.6	5.6	7.7	9.0
+CogTree(Yu et al. 2020)	22.9	28.4	31.0	13.0	15.7	16.7	7.9	11.1	12.7
+A-PFG (ours)	**32.2**	**38.2**	**40.7**	**19.5**	**22.9**	**24.5**	**13.9**	**17.9**	**20.2**
PFRL (ours)	14.5	18.4	20.0	7.9	9.8	10.4	5.8	8.0	9.3
+A-PFG (ours)	**35.4**	**41.9**	**44.3**	**20.9**	**24.5**	**25.8**	**14.6**	**18.9**	**21.7**

Table 1: Performance comparison of different methods on VG150 with respect to mean Recall. †denotes our reproduced model.

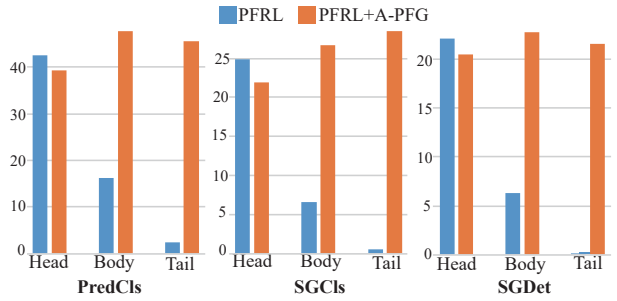


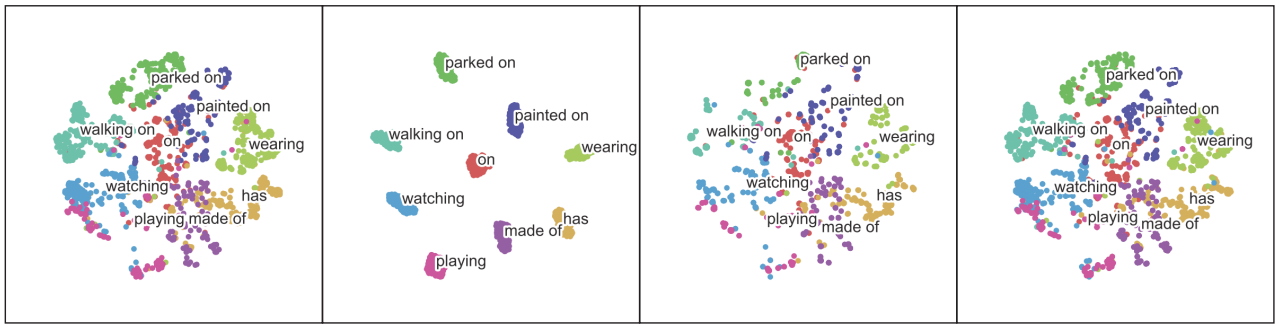
Figure 4: Performance comparison between PFRL and PFRL+A-PFG on mR@100 for head, body, tail classes.

level re-sampling method, which firstly oversamples the images containing tail predicates and then randomly discards a portion of the head predicates. The distribution of predicates in Rs-bi is still imbalanced due to the multiple predicates coupling problem. However, the PFG- models are all based on predicate features to construct a balanced training set. Hence the excellent performance of PFG- are mainly contributed to the decoupled learning framework that separates the coupled predicates in an image into individual predicate features.

Attribute-Guided Predicate Feature Generation To investigate the influence of attributes, ablation tests of different components contained in the attributes are conducted in Table 2, where PFG-p, PFG-⟨s,p,o⟩, and PFG-⟨p,u⟩ are the label of predicate, labels of ⟨subject-predicate-object⟩, and the label of the predicate combined with union visual feature as attributes to guide the generative model, respectively, and A-PFG is the full model contain all components.

For predicates in scene graph generation, even the same predicate ("riding") in different contexts, e.g., "man riding motorcycle" and "woman riding horse", features may exhibit significant variations. Therefore, PFG-p has a lower performance than others because the predicate label contains too little guidance information. The results of PFG-⟨p,u⟩ are better than PFG-⟨s,p,o⟩, for example, the mR@100 of PFG-⟨p,u⟩ are improved by 0.2 and 0.3 over PFG-⟨s,p,o⟩ in the PredCls and SGCls, respectively. That's because the specific union visual feature includes more information exclusive to the predicate feature than broad class labels. Since the attributes in A-PFG contain all the details of the predicate, A-PFG achieves optimal performance, where A-PFG improves 0.6, 0.4, 0.3 on mR@100 across three tasks compared to PFG-p. In summary, both the class labels and union visual feature provide detailed descriptions of the predicate, thus the attributes in the A-PFG can guide the synthetic predicate features to learn the contextual information.

Component Analysis of A-PFG Loss To better analyze the A-PFG model, the ablation experiments of A-PFG loss are also performed in Table 2, where $\mathcal{L}_{VAE}$ is the ba-



(a) PFRL (b) PFG-p (c) PFG-<s,p,o> (b) A-PFG

Figure 5: T-SNE visualization of predicate features on SGCls task.

Model	PredCls mR@50/100	SGCls mR@50/100	SGDet mR@50/100
PFRL	18.4/20.0	10.1/10.8	8.0/9.3
+RS-bi	30.7/32.9	19.4/20.4	12.6/14.9
+PFG-p	40.4/42.7	24.2/25.4	18.7/21.4
+PFG-⟨s,p,o⟩	41.6/43.7	24.2/25.4	18.8/21.5
+PFG-⟨p,u⟩	41.5/43.9	24.4/25.7	18.5/21.5
+A-PFG	**41.9/44.3**	**24.5/25.8**	**18.9/21.7**
$\mathcal{L}_{VAE}$	41.1/43.3	23.9/25.4	18.5/21.0
$\mathcal{L}_{VAE+C\text{-}Re}$	41.5/43.9	24.1/25.4	18.8/21.2
$\mathcal{L}_{VAE+DA}$	41.7/44.0	24.2/25.5	18.7/21.3
$\boldsymbol{\mathcal{L}_{A\text{-}PFG}}$	**41.9/44.3**	**24.5/25.8**	**18.9/21.7**

Table 2: Ablation study. PFRL is the baseline model that combines different re-sampling and data generation strategies, Rs-bi is a bi-level re-sampling strategy in (Li et al. 2021), and PFG- is the generative model guided by different information. $\mathcal{L}$ is the A-PFG loss with different components.

Model	PredCls mR@50/100	SGCls mR@50/100	SGDet mR@50/100
L-1	41.8/43.9	24.1/25.4	18.4/21.1
L-2	41.4/44.0	24.2/25.6	18.8/21.5
L-3	**41.9/44.3**	**24.5/25.8**	**18.9/21.7**
L-4	41.8/44.0	24.2/25.6	18.8/21.4
L-5	41.3/43.7	24.2/25.5	18.6/21.2

Table 3: Parameter analysis of the number of VAE layers in A-PFG. L- is the number of layers.

sic VAE loss, $\mathcal{L}_{C\text{-}Re}$ is the cross reconstruction loss, and $\mathcal{L}_{DA}$ is the distribution alignment loss. The results of $\mathcal{L}_{VAE+C\text{-}Re}$ and $\mathcal{L}_{VAE+DA}$ indicate that both cross reconstruction and distribution alignment contribute to the model performance. The improvements of full model $\mathcal{L}_{A\text{-}PFG}$ compared to $\mathcal{L}_{VAE}$ on mR@50/100 are 0.8/1.0 (PredCls), 0.6/0.4(SGCLs), 0.4/0.7(SGDet). Thus, the dual VAE in A-PFG contributes to the model's performance improvement.

Parameter Analysis of the Number of VAE Layers Table 3 provides the parameter analysis for the choice of VAE layers in A-PFG. Comparing the results for different number of layers, L-3 achieves the optimal performance. Therefore, this work sets the number of VAE layers to 3 for all experiments. More hyper-parameter analysis and discussion about the design choice are shown in the Suppl.

Visualization of Predicate Features

We selecte nine predicate classes in the head ("on", "has", "wearing"), body ("parked on", "walking on", "watching"), and tail ("painted on", "made of", "playing"), respectively, and then t-SNE visualized the predicate features in Figure 5. The visualization results demonstrate that the PFRL model learns good feature representations, as shown in Figure 5(a), "on", "parked on", "painted on" and "walking on" containing similar semantics are distributed relatively close together in the feature space, while "wearing" and "walking on" with different semantics are distributed far away.

From Figure 5(b), it can be observed that the features of PFG-p are all concentrate in the class centers because the guidance information is only independent predicate classes, resulting in the PFG-p losing semantic information of predicates. However, the feature distribution of PFG-⟨s,p,o⟩ is close to the semantic distribution as shown in Figure 5(c), but it lacks diversity due to the broad class information of the subject and object. Further, the A-PFG introduces the union visual feature containing more predicate details, which can make the synthesized features not only obey the semantic distribution but also have diversity, as shown in Figure 5(d). Therefore, our proposed A-PFG model synthesizes predicate features based on attributes, which can provide the classifier with diverse predicate features while preserving contextual information.

Conclusion

In this work, a decoupled learning framework is proposed for unbiased scene graph generation using attribute-guided predicate features to construct a balanced training set. Our proposed decoupled learning framework enables SGG to solve the long-tail problem by implementing data rebalancing based on predicate features. Our attribute-guided predicate features generation model can construct a balanced training set by synthesizing features containing contextual information via attributes. Extensive experiments confirm the effectiveness of our proposed method and our full model achieves new state-of-the-art performance for unbiased scene graph generation.

Acknowledgments

This study was funded by the National Natural Science Foundation of China with grant numbers (U21A20485, 61976175, 61976170, 62088102).

References

Chen, C.; Zhan, Y.; Yu, B.; Liu, L.; Luo, Y.; and Du, B. 2022. Resistance Training Using Prior Bias: Toward Unbiased Scene Graph Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 212–220.

Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6163–6171.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Desai, A.; Wu, T.-Y.; Tripathi, S.; and Vasconcelos, N. 2021. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15404–15413.

Dong, X.; Gan, T.; Song, X.; Wu, J.; Cheng, Y.; and Nie, L. 2022. 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*, 19427–19436.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Gao, L.; Wang, X.; Hu, Y.; Xu, X.; Lu, X.; Shen, H. T.; and Song, J. 2021. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16383–16392.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Knyazev, B.; de Vries, H.; Cangea, C.; Taylor, G. W.; Courville, A.; and Belilovsky, E. 2021. Generative compositional augmentations for scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15827–15837.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Li, R.; Zhang, S.; Wan, B.; and He, X. 2021. 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*, 11109–11119.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eledath, J.; Medioni, G.; and Sigal, L. 2021. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13936–13945.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6619–6628.

Tao, L.; Mi, L.; Li, N.; Cheng, X.; Hu, Y.; and Chen, Z. 2022. Predicate correlation learning for scene graph generation. *IEEE Transactions on Image Processing*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10275–10284.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yan, S.; Shen, C.; Jin, Z.; Huang, J.; Jiang, R.; Chen, Y.; and Hua, X.-S. 2020. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 265–273.

Yu, J.; Chai, Y.; Wang, Y.; Hu, Y.; and Wu, Q. 2020. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5831–5840.

Zhong, Y.; Wang, L.; Chen, J.; Yu, D.; and Li, Y. 2020. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, 211–229. Springer.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. 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*, 9719–9728.