Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation

Arushi Goel[1], Basura Fernando[2], Frank Keller[1], and Hakan Bilen[1]
[1]School of Informatics, University of Edinburgh, UK
[2]CFAR, IHPC, A*STAR, Singapore.

Abstract

Scene graph generation (SGG) aims to capture a wide variety of interactions between pairs of objects, which is essential for full scene understanding. Existing SGG methods trained on the entire set of relations fail to acquire complex reasoning about visual and textual correlations due to various biases in training data. Learning on trivial relations that indicate generic spatial configuration like 'on' instead of informative relations such as 'parked on' does not enforce this complex reasoning, harming generalization. To address this problem, we propose a novel framework for SGG training that exploits relation labels based on their informativeness. Our model-agnostic training procedure imputes missing informative relations for less informative samples in the training data and trains a SGG model on the imputed labels along with existing annotations. We show that this approach can successfully be used in conjunction with state-of-the-art SGG methods and improves their performance significantly in multiple metrics on the standard Visual Genome benchmark. Furthermore, we obtain considerable improvements for unseen triplets in a more challenging zero-shot setting.

1. Introduction

In this paper, we look at a structured vision-language problem, scene graph generation [25, 55], which aims to capture a wide variety of interactions between pairs of objects in images. SGG can be seen as a step towards comprehensive scene understanding and benefits several high-level visual tasks such as object detection/segmentation [16, 19, 40], image captioning [1, 17, 21, 29], image/video retrieval [15], and visual question answering [2, 33]. In the literature, SGG is typically formulated as predicting a triplet of a localized subject-object pair connected by a relation (*e.g. person **wear** shirt*). Broadly, recent advances in SGG have been obtained by extracting local and global visual features in convolutional neural networks [22,44,67] or graph neural networks [30, 55, 60] combined with language embeddings [35, 38] or statistical priors [64] for predicting relations between objects.

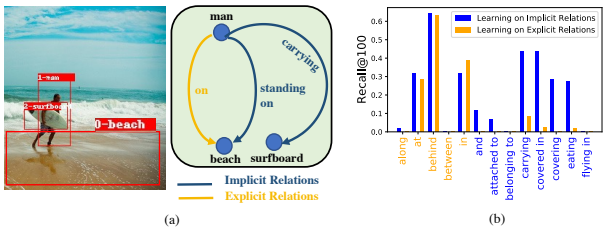

Figure 1. (a) An example of a scene graph with *implicit* and *explicit* relations. (b) Per-class Recall@100 for an SGG model trained either on only explicit (orange) or implicit (blue) relations.

Despite the remarkable progress in this task, various factors (long-tail data distribution, language or reporting bias [37]) in the established SGG benchmarks (*e.g.* Visual Genome [25]) have been shown to drive existing methods towards biased and inaccurate relation predictions [48, 49]. One major cause is that each subject-object pair is annotated with *only one positive* relation, which typically depends on the annotator's preference and is *subjective*, while other plausible relations are treated as negative.[1] For instance, a subject-object pair *man–beach* in Fig. 1a is only annotated with one relation as *man **on** beach*, even though other plausible relations are available such as *standing on*. Hence, the models [22, 28, 49, 64] trained on this data become biased towards more frequently occurring labels, as reported in [48]. To alleviate the biased training, Tang *et al.* [48] employ counterfactual causality to force the SGG model to base its predictions only on visual evidence rather than the data bias. Wang *et al.* [53] propose a semi-supervised technique that jointly imputes the missing labels of subject–object pairs with no annotated relations to obtain more balanced triplet distributions. Suhail *et al.* [46] propose an en-

[1]Note that both training and test sets of the standard benchmarks are subject to the similar biases.

ergy based method that can learn to model the joint probability of triplets from few samples and thereby avoids generating biased graphs with inconsistent structure.

While the recent methods [46,48,53] successfully tackle the bias towards more frequently occurring labels, this paper studies another type of bias related to *label informativeness*. It also manifests itself in missing annotations and has not been addressed in SGG before. In particular, we hypothesize that certain relation labels (implicit labels) are more informative than others (explicit labels) and training on implicit labels improves a model's ability to reason over complex correlations in visual and textual data.

Our key intuition comes from prior computational and cognitive models [31] and recent work [9] that categorize relations into *explicit (or spatial)* and *implicit* based on whether the relation defines the relative spatial configuration between the two objects implicitly or explicitly (*e.g.* *man* ***standing on*** *beach vs. man* ***on*** *beach* in Fig. 1a). Explicit relations are often *easy to learn*, *e.g.* from the spatial coordinates of subject–object pairs, thanks to their highly deterministic spatial arrangements, while implicit ones are often challenging due to the relative spatial variation and require deliberate reasoning. To test our hypothesis, we conduct experiments where we train a SGG model either only on explicit, or only implicit ones, and evaluate them on a test set including both types (*i.e.* zero-shot implicit or explicit relation classification), see Fig. 1b. Surprisingly, training only on implicit relations obtains good performance not only over implicit ones but also unseen explicit ones (only 2% lower in average training on explicit relations and 4% lower when trained on all labels), while training only on explicit relations performs poorly on implicit relations (where the performance drops to 0.1% from 24.3%).[2] In other words, training on implicit labels enables the model to better generalize to unseen explicit labels. However, due to partially annotated training data, many subject-object pairs are only labelled by explicit relations and their implicit relations are missing and obscured by the explicit ones.

Motivated by our analysis, we design a novel model-agnostic training procedure for SGG that jointly extracts more information from partially labeled data by mining the missing implicit labels, trains a SGG model on them and boosts its performance. In particular, our method involves a two stage training pipeline. The first stage trains a SGG model on a subset of training data including only annotated implicit relations, which allows the model to learn rich correlations in the data and encourages it to predict more informative implicit labels in the next stage. The second stage includes an alternating procedure that imputes missing implicit labels on the subset of samples annotated with explicit relations, followed by training on both the annotated and imputed labels, called *label refinement*. In this stage, a model is prone to confirm to its own (wrong) predictions to achieve a lower loss as observed in semi-supervised learning (*e.g.* [3, 50]). To prevent such overfitting, we regularize the model by a latent space augmentation strategy. We demonstrate that our method yields significant performance gains in the SGG task for the standard and zero-shot settings on the Visual Genome [25] when applied to several existing scene graph generation models.

In short, our contributions are as following. We identify a previously unexplored issue, *missing informative labels* in the standard SGG benchmark and address this through a model agnostic training procedure based on alternating label imputation and model training with effective regularization strategies. This method can be incorporated into state-of-the-art SGG models and boosts their performance by a significant margin.

2. Related work

Scene Graph Generation. SGG has been extensively studied in the past few years with the goal of better understanding the object relations in an image by either focusing on architecture designs [7, 28, 49, 55, 60, 64] or feature fusion methods [13, 18, 22, 27, 57, 62, 63, 67]. Recently, Tang *et al.* [48, 49] reported that the performance gains from these methods largely come from improved performance only on the head classes (frequently occurring relations) while the performance on most other relations is poor. They propose replacing the biased evaluation metric Recall@k with mean-Recall@k to assign equal importance to all labels.

The same authors [48] report that the bias in the data often drives SGG models to predict frequent labels and propose to use counterfactual causality *i.e.* measure the difference in predictions between the original scene and a counterfactual one to remove the effect of context bias and focus on the main visual effects of the relation. Suhail *et al.* [46] propose an energy based loss that learns the joint likelihood of object and relations instead of learning them individually. This helps to incorporate commonsense structure (*man* ***riding*** *horse* and *man* ***feeding*** *horse* occurring together are highly improbable) and in better context aggregation. Unlike [46, 48], our focus is on extracting more information from the training data through mining informative labels. In fact, we show that our model is orthogonal to theirs and boosts performance when incorporated to theirs. The most similar to ours, Wang *et al.* [53] propose a semi-supervised method that employs two deep networks, where the auxiliary one imputes missing labels of unlabeled pairs and self-trains on them and transfers its knowledge to the main network. Unlike [53], who treat all the labels equally, our method only imputes informative implicit labels. This is crucial, as shown in Tab. 3, because, without such consideration, imputing labels cannot extract any substantial

[2] We provide the full analysis in the results section and supplementary.

information from unlabeled samples leading to only minor gains. In addition, our framework is more efficient as it involves only a single network that jointly infers labels and trains on them, outperforming [53] significantly.

Label Completion. There is a rich body of work in the literature that focuses on learning from partial/missing labels in a multi-label learning setting where each image is labelled for multiple categories with some missing labels [5,6,8,14,20,34]. Common strategies to address this can be divided into two categories: 1) graph based methods [24,54] that exploit similarity between samples to predict missing labels, and 2) low rank matrix completion which extracts label correlations [6, 14, 56, 59] to complete missing labels. There is another setting in which some instances miss all the labels, also called semi-supervised learning in multi-label classification [47,68]. In this setting, the classifier is trained for unseen data. While related to our setting, we look at the classification of relations conditioned on subject–object pairs (rather than on the image level), with each pair already labeled with one relation (rather than unlabeled images). Finally, we group the label set in two groups and treat them asymmetrically in our training.

Semi-supervised learning. Semi-supervised learning methods exploit unlabeled data via either pseudo-labeling or imputation with small amounts of labeled data [41, 43] or by enforcing consistency regularization on the unlabeled data to produce consistent predictions over various perturbations of the same input [50, 51] by applying several augmentation strategies such as Mixup [65], RandAugment [11], AutoAugment [10] or combine both pseudo-labeling and consistency regularization [4, 45]. Inspired from pseudo-labeling in semi-supervised learning, the main motivation of our method is to impute informative missing labels to improve generalization and learn complex features by relying on partially labeled data and still predict more accurate labels on the biased test set.

3. Methodology

3.1. Revisiting SGG Pipeline

In SGG, we seek to localize and classify the objects in an image followed by labeling the visual relations between each pair of objects (or subject and object). Concretely, let C and P denote the object and relation classes respectively. Each subject or object $e = (e^b, e^c) \in \mathcal{E}$ consists of a bounding box $e^b \in \mathbb{R}^4$ and a class label $e^c \in C$. A relation tuple is a triplet of the form $r = (s, p, o)$ where the subject s and the object o ($s, o \in \mathcal{E}$) are joined by the relation $p \in P$, *e.g.* *man **wearing** shirt*. Given an image I, we can then use a set of objects $E = \{e_i\}_{i=1}^m$ and a set of relations $R = \{r_j\}_{j=1}^n$, where m and n are the number of subject/objects and relation triplets in an image respectively, to define a scene graph $S = (E, R)$. A scene graph can also be written as a combination of a set of bounding boxes $B = \{e_i^b\}_{i=1}^m$, a set of class labels $Y = \{e_i^c\}_{i=1}^m$ and a set of relations R.

The SGG models can be decomposed as:

$$P(S|I) = P(B|I)\, P(Y|B, I)\, P(R|B, Y, I) \qquad (1)$$

where $P(B|I)$ is the object detector or bounding box prediction model, $P(Y|B, I)$ is an object class model and $P(R|B, Y, I)$ is a relation prediction model.

Existing methods [23, 26, 32, 49, 55, 64, 66, 69] often employ a two-step process for the scene graph generation task. First, bounding-box proposals ($P(B|I)$) with class predictions and confidence scores ($P(Y|B, I)$) are extracted using off-the-shelf object detectors [16, 40]. Then, a multimodal feature fusion model combines visual, language and spatial features to predict the relation for a given subject-object pair ($P(R|B, Y, I)$). Several methods adopt BiLSTMs [64], Bi-TreeLSTMs [49] or fully connected layers [23, 66] to encode the co-occurrence between object pairs for relation prediction.

3.2. Missing Relation Labels

Many visual relations are hypernyms, hyponyms, or synonyms [39, 58] and hence are non-mutually exclusive. The standard SGG datasets (*e.g.* Visual Genome [25]) ignore this fact and only assume *one* annotated label per subject–object pair. Which one is assumed strongly depends on the annotator (manifesting as labeling or language/reporting bias [37]).

One way to circumvent this problem is to collect multiple labels for each triplet, which is however expensive and time consuming. Another potential solution is to use linguistic sources such as WordNet [36] or VerbNet [42] to automatically obtain the missing labels by exploiting the linguistic dependencies between relations. However, this is not trivial, as some of the relation and spatial vocabulary in the SGG datasets are not included in WordNet. Moreover, the context of relations in these language resources does not provide always allow the right inferences. For instance, in WordNet, *person **riding** horse* does not imply *person **on** horse* (no hyponymy relation), but this is the visual implication in the SGG datasets.

While one can use existing methods [53] to infer the missing labels, the estimated labels can be noisy and uninformative such that re-training a model on them may not improve the generalization performance. Here, inspired by previous work [9], we propose to group visual relations into two sets: explicit and implicit.[3] Explicit relations encode spatial information between two objects such as 'on', 'in front of' or 'under' and are typically easy to learn, even only based on subject–object locations [9]. The implicit ones are

[3] Further details about the explicit and implicit relations is in the experiments and the supplementary.

normally verbs such as 'riding', 'walking' and for learning them the model has to find complex correlations in visual and textual data. In existing SGG datasets, some object pairs are annotated with implicit labels, while other pairs are labeled only with explicit ones and their implicit labels are missing. We propose to divide the set of predicate labels P into two sets, *explicit* and *implicit* and denoted them by $\mathbb{E}$ and $\mathbb{I}$ respectively.

3.3. Proposed Method

In this section, we explain our proposed method for training the relation classifier f_θ to implement $P(R|B, Y, I)$. For each image I, we assume that an object detector provides a set of candidate subject–object pairs $\{(s-o)\}$ and each pair is represented by a d-dimensional joint embedding $\boldsymbol{x} \in \mathbb{R}^d$ including its visual, semantic and spatial features. Note that we apply our method to various existing SGG models which uses different object detector and joint embedding functions, and we provide these details in Sec. 4. In particular, f_θ is instantiated as a deep neural network parameterized by θ, takes in a joint feature embedding $\boldsymbol{x}$ for a subject–object pair $s-o$ and outputs a softmax probability over $|P|$ relations, *i.e.* $f_\theta(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{|P|}$. Our goal is to learn a relation classifier f_θ that can correctly estimate the relation label of a subject-object pair in an unseen image.

Given a training set $\mathcal{D}$ with $|\mathcal{D}|$ samples, each including tuples of subject-object pairs $s-o$ along with the relation label p and the joint feature embedding $\boldsymbol{x}$, which we denote with $X = \{(s, p, o, \boldsymbol{x})\}_{i=1}^{|X|}$ with $|X|$ tuples. We formulate the learning problem as minimization of two loss terms:

$$\min_\theta \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left(\mathcal{L}^{\mathbb{I}}(X_i; \theta) + \mathcal{L}^{\mathbb{E}}(X_i; \theta) \right) \tag{2}$$

where $\mathcal{L}^{\mathbb{I}}(X_i; \theta)$ and $\mathcal{L}^{\mathbb{E}}(X_i; \theta)$ are the loss terms defined over implicit and explicit relations respectively.

For a given image I and its tuple X, we pick the tuples whose relation is annotated only with implicit relation label (*i.e.* $X^{\mathbb{I}} = \{(s, p, o, \boldsymbol{x}) \mid p \in \mathbb{I}\}$) and define the implicit loss term as:

$$\mathcal{L}^{\mathbb{I}}(X^{\mathbb{I}}; \theta) = \frac{1}{|X^{\mathbb{I}}|} \sum_{(s,p,o,\boldsymbol{x}) \in X^{\mathbb{I}}} \mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), p) \tag{3}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss. In other words, for the implicit relations, we follow the standard practice and compute its loss by using its ground-truth implicit relation label p which is an one-hot vector, as each subject–object pair is annotated with only one label.

Similarly, we formulate the explicit term $\mathcal{L}^{\mathbb{E}}(X^{\mathbb{E}}; \theta)$ for the tuples with explicit labels:

$$\mathcal{L}^{\mathbb{E}}(X^{\mathbb{E}}; \theta) = \frac{1}{|X^{\mathbb{E}}|} \sum_{(s,p,o,\boldsymbol{x}) \in X^{\mathbb{E}}} \mathcal{L}_{KL}(f_\theta(\boldsymbol{x}), \hat{p}) \tag{4}$$

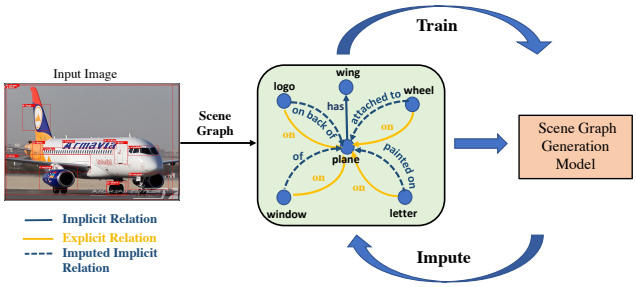

Figure 2. Pipeline of our proposed framework for mining informative labels for scene graph generation. An input image can be represented as a scene graph (in green). The yellow and blue solid arrows in the scene graph denote ground-truth explicit and implicit relations, respectively. The scene graph generation model is first trained on a subset of the implicit relations. In an iterative fashion, we then impute implicit relations (dashed blue arrows) for the triplets annotated with explicit relations and train the model with all relations (implicit, explicit, and imputed).

where $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, $X^{\mathbb{E}} = \{(s, p, o, \boldsymbol{x}) \mid p \in \mathbb{E}\}$ and $\hat{p}$ is the *imputed* relation label for the subject-object pair, which is a vector with soft probabilities. Next, we discuss our method to obtain $\hat{p}$.

Label imputation. The subject–object pairs that are annotated with explicit relations only can often be labelled with more informative implicit relation labels. For instance, the ground truth label may be *person* ***beside*** *table*, but *person* ***eating at*** *table* is also correct in this case, and more informative. To impute the missing implicit relation for a subject–object pair, which is originally annotated with an explicit relation label, we follow a two-step procedure.

First, we take each subject–object pair annotated with an explicit label from $X^{\mathbb{E}}$ along with its joint embedding and impute its implicit label through the relation classifier f_θ as:

$$\bar{p} = \arg\max_{i \in \mathbb{I}} \left[\frac{\exp(f_\theta^i(\boldsymbol{x}))}{\sum_{j \in \mathbb{I}} \exp(f_\theta^j(\boldsymbol{x}))} \right] \tag{5}$$

where f^i denotes its logit for the i-th relation class. In words, we compute softmax probabilities only over implicit relation labels and pick the highest scoring implicit label to obtain an one-hot vector $\bar{p}$. Note that one can also obtain a soft probability vector over all implicit label classes, however, we empirically show that the former works better.

Second, as the subject–object pair is originally labelled with an explicit label p, we also use this information and average the imputed label $\bar{p}$ with its original label p as:

$$\hat{p} = (p + \bar{p})/2. \tag{6}$$

We call this step as ***Label Refinement***. As $\hat{p}$ includes equal probabilities (*i.e.* 0.5) for one explicit and implicit labels,

it is not an one-hot vector. Thus, we use KL divergence Eq. (4) to encourage the model to predict both the labels. Compared to standard cross-entropy loss, the KL divergence loss increases the model entropy by reducing overconfidence, resulting in smoother predictions. Most of the traditional methods for SGG trained with cross-entropy may get confused by inconsistent annotations, where same relation is labeled with less informative spatial relation in some images while more informative labels are used in some other images. Our loss function formulation and multi-label nature of the targets addresses this inconsistency, unlike previous work [46, 49].

Latent space augmentation. Our training pipeline is illustrated in Fig. 2 which follows an alternating optimization and consists of two alternating steps where we employ the relation classifier f_θ to impute the implicit labels and simultaneously optimize the classifier parameters θ. The main challenge here is that the model parameters can overfit to its own imputed labels quickly, resulting in a local optimum solution. This problem is notoriously known as confirmation bias that also occurs in many semi-supervised problems [3, 45]. To prevent overfitting to the wrong imputed labels, existing solutions include applying various kinds of data augmentation including standard geometric and color transformations [12], their combinations [10, 11, 61] and also generating augmented version of samples [52, 65].

As many SGG methods build on the feature space of an object detector, many augmentation strategies that are applied on raw pixels are not applicable to our case. Hence, we use Manifold Mixup [52] that generates augmented embeddings (in the maninfold space rather than in the pixel space) by taking a convex combination of different pairs of embeddings ($\boldsymbol{x}$ and $\boldsymbol{x}'$) and also their labels (p and p'):

$$\begin{aligned} \tilde{\boldsymbol{x}} &= \lambda.\boldsymbol{x} + (1-\lambda).\boldsymbol{x}' \\ \tilde{p} &= \lambda.p + (1-\lambda).p' \end{aligned} \tag{7}$$

where λ is sampled from a beta distribution, *i.e.* $\lambda \sim \text{Beta}(\alpha, \alpha)$ with α as a hyperparameter. Note that we apply this augmentation to the whole training set and allow mixing embeddings across samples from both the implicit and explicit set of relations. This augmentation acts as a regularizer and accounts for overfitting to the incorrect imputed labels while training.

3.4. Algorithm

In Algorithm 1, we detail our training pipeline. To obtain the initial parameters θ_0, we first train our model on the tuples with only implicit labels as following (Line 2):

$$\theta_0 = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}^{\mathbb{I}}(X_i; \theta). \tag{8}$$

The key intuition behind this is that model learned on only the implicit relations are more likely to produce confident predictions over implicit labels and hence not 'distracted' by explicit relation labels.

After the training on implicit relations, we iteratively impute implicit relation labels for the subject-object pairs annotated with the explicit relations (Line 5) and update the model parameters using Eq. (2) (Line 7). The model parameters optimized in Eq. (2) take as input the augmented versions of the sample and label pair (as in Eq. (7)) for both the implicit and explicit set of relation labels (Line 6).

Algorithm 1 Our proposed optimization of the SGG model

1: **Input:** Training set $\mathcal{D}$ with $|\mathcal{D}|$ samples, each including a set of tuples X with $(s, p, o, \boldsymbol{x})$ with subject, object and relation label, joint embedding resp., Explicit and implicit relation sets $\mathbb{E}$ and $\mathbb{I}$ resp., relation classifier f_θ, T number of iterations, η learning rate.
2: Initialize θ as in Eq. (8)
3: **for** $t = 0, \ldots, T$ **do**
4: Sample a minibatch $B = (X_1, \ldots, X_{|B|}) \sim \mathcal{D}$,
5: **Impute** $\hat{p}$: Impute implicit labels $\hat{p}_t$ for $B^{\mathbb{E}}$ by using Eq. (5) and Eq. (6),
6: Augment B by applying manifold mixup in Eq. (7),
7: **Update** θ: $\theta_{t+1} \leftarrow \theta_t + \eta\Delta_\theta$ where Δ_θ is the update for θ obtained from Eq. (2),
8: **end for**
9: **return** θ

4. Experiments

Dataset and Evaluation Settings. We evaluate our proposed method for scene graph generation on the Visual Genome (VG) [25] dataset. We use the pre-processed version of the VG dataset as proposed in [55]. The datatset consists of 108k images with 150 object categories and 50 relation categories. The training, test and validation split used in the experiments also follow previous work [46, 48, 55].

For evaluation on the **Visual Genome dataset**, we follow [55] and report performance on three settings: **(1) Predicate Classification (PredCls).** This task measures the accuracy of relation (also termed as predicate in literature) prediction when the ground truth object classes and boxes are given. It is not affected by the object detector accuracy. **(2) Scene Graph Classification (SGCls).** In this setting, we know the ground truth boxes and we have to predict the object classes and the relations between them. **(3) Scene Graph Detection (SGDet).** This is the most challenging setting and the models are used to predict object bounding boxes, object classes and the relations between them. We measure Mean Recall@K (mR@K) [48, 49] to evaluate scene graph generation models. More recent work has

Models	Method	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP [7, 55]	-	-	9.8	10.5	-	5.8	6.0	-	3.8	4.8
FREQ [49, 64]	-	8.3	13.0	16.0	5.1	7.2	8.5	4.5	6.1	7.1
Motifs [64]	-	10.8	14.0	15.3	6.3	7.7	8.2	4.2	5.7	6.6
KERN [7]	-	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
VCTree [49]	-	14.0	17.9	19.4	8.2	10.1	10.8	5.2	6.9	8.0
VCTree-L2+cKD [53]	-	14.4	18.4	20.0	9.7	12.1	13.1	5.7	7.7	9.0
IMP [55]	Baseline	8.9	11.0	11.8	5.4	6.4	6.7	2.2	3.3	4.1
	Ours	**12.3**	**14.6**	**15.3**	**7.1**	**8.0**	**8.3**	**6.9**	**7.8**	**8.1**
Motif-TDE-Sum [48, 64]	Baseline	17.9	24.8	28.7	9.8	13.2	15.1	6.6	8.9	11.0
	Ours	**21.3**	**27.1**	**29.7**	**11.3**	**14.3**	**15.7**	**8.4**	**10.4**	**11.8**
VCTree [49]	Baseline	13.1	16.5	17.8	8.5	10.5	11.2	5.3	7.2	8.4
	Ours	**18.0**	**21.7**	**23.1**	**11.9**	**14.1**	**15.2**	**7.1**	**8.2**	**8.7**
VCTree-EBM [46]	Baseline	14.2	18.2	19.7	10.4	12.5	13.4	5.7	7.7	9.1
	Ours	**21.0**	**24.9**	**26.5**	**14.0**	**16.2**	**17.1**	**7.8**	**10.1**	**11.8**
VCTree-TDE [48]	Baseline	16.3	22.9	26.3	11.9	15.8	18.0	6.6	9.0	10.8
	Ours	**22.2**	**28.1**	**30.6**	**17.8**	**22.0**	**23.6**	**8.4**	**10.3**	**11.5**

Table 1. Scene Graph Generation performance comparison on mean Recall@K [48] under all three experimental settings. We compare the results of our proposed framework (Ours) with the original model (Baseline) using different SGG architectures.

preferred mR@K over regular Recall@K [55] due to data imbalance [48]. Mean Recall@K treats each relation separately and then averages Recall@K over all relations. We also measure the **zero-Shot Recall**, zsR@K, which helps to evaluate the generalization ability of the model in predicting subject–relation–object triplets that are not seen during training. We measure zsR@K for all three settings of PredCls, SGCls and SGDet.

Model Generalization. Our proposed framework has the flexibility to be trained with any scene graph generation model. Hence, we train with different model architectures to demonstrate the generalizability of our approach: Iterative Message Passing (IMP) [55], Neural-Motifs [64] and VCTree [49]. We also train with two other debiasing methods that build upon these models, Energy-based Modeling (EBM) [46], where the authors propose to train with an additional energy-based loss and Total Direct Effect (TDE). where counterfactual reasoning is used during inference.

Explicit and Implicit Relation Labels. For the Visual Genome dataset [25], Xu *et al*. [55] released a version of the dataset with 50 relations and 150 object categories. These 50 relations are: *above, across, against, along, and, at, attached to, behind, belonging to, between, carrying, covered in, covering, eating, flying in, for, from, growing on, hanging from, has, holding, in, in front of, laying on, looking at, lying on, made of, mounted on, near, of, on, on back of, over, painted on, parked on, part of, playing, riding, says, sitting on, standing on, to, under, using, walking in, walking on, watching, wearing, wears, with.*

Inspired by [9], we divide the relation label space into a set of explicit and implicit relations. We define explicit relations when the spatial arrangement of objects are implied by the label itself *e.g*., "on", "below", "next to" and so on. For implicit relations, the spatial arrangement of objects is only indirectly implied, "riding", "walking", "holding" etc. More specifically, the explicit relations are *above, across, against, along, at, behind, between, in, in front of, near, on, over, under* and the rest are treated as implicit relations.

Implementation Details. Following previous work [48, 49], we train the scene graph generation models on top of the pre-trained Faster R-CNN object detector with ResNetXt-101-FPN backbone [40]. The weights of the SGG models' object detector are are frozen during training in all the three settings – PredCls, SGCls and SGDet. The mAP of the object detector on the Visual Genome dataset is 28% using 0.5 IoU. For training each scene graph generation model using our proposed method, we use the default training settings as in [46, 48] for fair comparisons. The models are trained with the SGD optimizer with a batch size of 12, an initial learning rate of 10^{-2} and 0.9 momentum.

The models are trained for the first 30,000 batch iterations on the implicit label subset with the standard cross-entropy loss. After label imputation, the model is trained on the rest of the imputed data and the implicit subset for another 20,000 batch iterations. The value of α is set to 4 from which λ is sampled for the mixing function in the latter half of training. Our code will be made publicly available upon acceptance.

5. Results

Quantitative Results. Table 1 compares the performance of the state-of-the-art methods when trained with our proposed training framework on the Visual Genome dataset [25]. Our method consistently improves performance on all the three evaluation settings (PredCls, SGCls, SGDet) when trained with existing methods. With IMP [55] and

Models	Method	PredCls zsR@20/50	SGCls zsR@20/50	SGDet zsR@20/50
IMP [55]	Baseline	**12.17/17.66**	**2.09/3.3**	0.14/0.39
	Ours	7.12/10.50	1.57/2.32	**1.52/2.48**
Motif-TDE-Sum [48]	Baseline	8.28/14.31	**1.91**/2.95	1.54/2.33
	Ours	**9.33/14.43**	1.87/**2.99**	**2.06/3.05**
VCTree [49]	Baseline	1.43/**4.0**	**0.39/1.2**	0.19/0.46
	Ours	**1.51**/3.7	0.36/1.0	**0.43/0.95**
VCTree-TDE [48]	Baseline	8.98/**14.52**	3.16/4.97	1.47/2.3
	Ours	**9.11**/13.52	**4.26/6.20**	**2.24/3.25**

Table 2. Zero shot recall performance for our proposed method compared with the original model (baseline).

Motif-TDE-Sum [48, 64], we obtain an absolute improvement of 3.4% and 1.7% in PredCls and SGCls, respectively. For the most challenging setting of SGDet, there is an improvement of 4.7% with the IMP model, showing the generalization ability of our approach on a visual-only model (no language/textual features). When debiasing approaches such as TDE [48] and EBM [46] are incorporated, we obtain consistent improvements for VCTree. In both the cases, we gain significantly in all the settings and achieve a new state-of-the-art performance in scene graph generation by combining our proposed method with VCTree-TDE.

In Table 2, we report the zero-shot recall performance and compare it with baselines. Our proposed method achieves improvements in most of the settings with different SGG backbones, except for IMP in PredCls and SGCls. IMP being a visual-only model fails to learn correlations via textual features for different relation classes and hence performs poorly in zero-shot, due to low recall on explicit relations. However, the multi-modal nature of Motif and VCTree brings out the strength of our method in generalizing to unseen triplets during test time.

Method	Predicate Classification			
	Train Label	Imputed with	Imputed on	mR@20/50
Baseline	-	-	-	17.85/24.75
Ours	Random	Top1	Random	17.11/23.56
	Explicit	Top1-Explicit	Implicit	14.23/20.51
	Implicit	Top1-Implicit	Explicit	**21.26/27.14**

Table 3. Experimental results on the Predicate Classification setting with different ways of label imputation.

Ablative Analysis. In this section, we study different components of our method separately to validate their effectiveness. Table 3 evaluates the contribution of our proposed label imputation strategy. All the models are trained using the Motif-Sum [64] backbone with TDE [48] at inference time, Motif-TDE-Sum. *Random* implies that we randomly divide the relation labels into two sets without the knowledge of *explicit* and *implicit* relations. We compare the results under different settings when the model is trained on a particular subset of *Train Label*. The relation labels are imputed on the hold-out relation set with the top-1 (best) in the set of labels that it was trained on. Training and imputing with the random or explicit set of labels decreases performance compared to the baseline, with a significant drop when only explicit relations are considered. This shows the importance of learning with and imputing implicit relations: they provide useful information about interactions between object pairs not captured by explicit relations.

Method	Predicate Classification			
	mR@20	mR@50	mR@100	zsR@20
Baseline (Learning on All Relations)	17.85	24.75	28.70	8.28
Learning on Explicit Relations	14.06	20.34	23.81	8.12
Learning on Random Relations	16.99	23.33	26.57	8.18
Learning on Implicit Relations	18.24	24.93	28.71	7.83
Ours w/o Manifold Mixup	18.26	24.23	27.48	7.35
Ours w/o Label Refinement	19.90	26.35	29.26	**9.70**
Ours w/ MSE-loss (soft imputed labels)	17.97	24.54	28.19	7.72
Ours w/ KL-loss (soft imputed labels)	20.76	27.10	**30.33**	8.25
Ours (final)	**21.26**	**27.14**	29.68	9.33

Table 4. Ablation study on our proposed training framework with Motif-TDE-Sum [48, 64] as the SGG backbone.

In Table 4, we show the effect of training with *implicit* relations on scene graph generation performance. If the model is trained with the *explicit* relations only, mean recall drops. There is also a marginal drop in performance compared to the baseline when trained on a random subset of the relations. Even training on a subset of the data with only implicit relations, we can achieve a performance which is on a par with, or better than, the baseline. We also show the importance of training with *Latent Space Augmentation* (Manifold Mixup) and *Label Refinement* in our method. They help to prevent model overfitting to incorrect imputed labels and generalize better. We replace the loss on the hard imputed labels in Equation (4) by computing the Mean Squared Error (MSE) loss or KL-loss over the *soft* imputed labels, where we consider probabilities over all classes as the target. KL-loss with soft imputed labels performs similar to our final method with hard imputed and refined labels except for the zero-shot case. This shows that the noise in the soft targets can be detrimental for generalization to unseen triplets.

In Table 5, we show the results for training on relation subsets and our final method on the explicit and implicit set of relations separately. As discussed in Section 1, when the model is only trained on explicit relations, it fails to generalize to implicit relations. This is in contrast to training only on the implicit set of relations. This indicates that implicit relations are rich in information and perhaps learn complex and generalizable features. Our final method outperforms training on all relations (original model) and the subset of relations by a significant margin, showing the strength of mining more informative labels for less informative samples.

Qualitative Results. Figure 3 visualizes the scene graphs predicted from the baseline VCTree-EBM [46] model (in

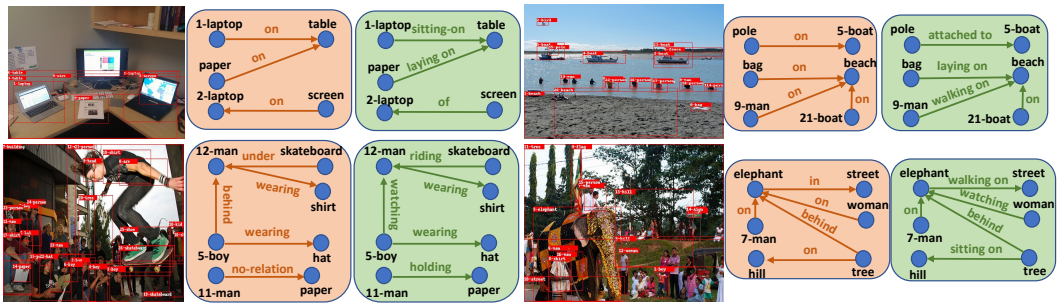

Figure 3. Visualization of scene graphs generated by the VCTree-EBM [46] based learning framework (in orange) and our proposed method using VCTree-EBM as backbone (in green).

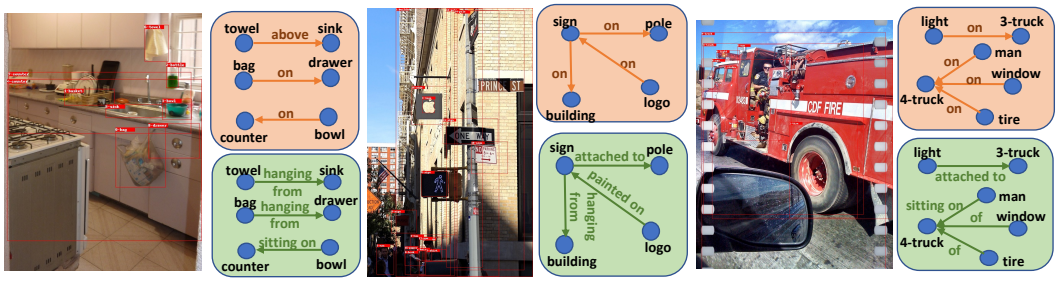

Figure 4. Scene Graph Visualization results for the missing *implicit* relations imputed for the triplets annotated with *explicit* relations. The scene graphs in orange are the ground-truth triplets with their corresponding imputed *implicit* relations in green.

Method	PredCls - mR@50/100		
	Train Relations ↓/Test Relations →	Explicit	Implicit
MOTIF-TDE-Sum	All Relations	24.47/28.79	24.96/28.74
	Explicit only	22.89/25.34	0.08/0.09
	Implicit only	20.10/22.89	24.34/26.03
	Ours (final)	**24.83/27.80**	**27.99/30.38**

Table 5. Performance Comparison on the Explicit and Implicit set of relations separately with different subsets of training labels.

orange) and compares it to the scene graphs obtained via our proposed training framework (in green). Our method consistently predicts more informative relations such as ***laying on, walking on, holding*** instead of simple prepositional relations such as ***on, in***. Moreover, our method also effectively identifies triplets with relations that were missed in the baseline. For instance, in the bottom-left image, our method localizes *man* ***holding*** *paper* correctly. Our method also corrects relations which are incorrectly predicted in the baseline, *woman* ***watching*** *elephant* as opposed to *woman* ***on*** *elephant* in the bottom-right image.

In Figure 4, we visualize the imputed *implicit* relations for the triplets annotated with explicit relations. In orange, we show the ground-truth scene graphs and the corresponding imputed scene graphs in green. It can be clearly observed from the ground-truth scene graphs that there is annotator bias towards spatial relations. Our label imputation strategy is able to find alternate and missing implicit relations for these triplets and exploit them during training. For instance, our method imputes important relations such as ***attached to, hanging from, sitting on*** which are more descriptive than their explicit counterpart ***on***. This shows the importance of label imputation and generating descriptive scene graphs for comprehensive scene understanding.

6. Conclusion

We proposed a novel model-agnostic training framework for scene graph generation. We introduced the concept of label informativeness, which had not been explored in SGG before. A model trained on informative relations is able to model the visual and textual context better compared to training on simple spatial relations. We showed how to impute informative relations from the partially labeled data and jointly train with imputed and ground truth relations. We improved the performance across models and tasks, including in a zero-shot setting. One limitation of our approach is its limited ability in predicting relations with very few samples, which should be investigated in future work.

Acknowledgement: AG is supported by the Armeane Choksi Scholarship and HB is supported by the EPSRC programme grant Visual AI EP/T028572/1. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1

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

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2, 5

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 3

[5] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR 2011*, pages 2801–2808. IEEE, 2011. 3

[6] Ricardo S Cabral, Fernando Torre, Joao P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In *Advances in neural information processing systems*, pages 190–198, 2011. 3

[7] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 2, 6

[8] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. 3

[9] Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 3, 6

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 3, 5

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 3, 5

[12] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017. 5

[13] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Fei-Fei Li. Visual relationships as functions: Enabling few-shot scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[14] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019. 3

[15] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 1

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 3

[17] Arushi Goel, Basura Fernando, Thanh-Son Nguyen, and Hakan Bilen. Injecting prior knowledge into image caption generation. In *European Conference on Computer Vision*, pages 369–385. Springer, 2020. 1

[18] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2019. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[20] Jun Huang, Linchuan Xu, Kun Qian, Jing Wang, and Kenji Yamanishi. Multi-label learning with missing and completely unobserved labels. *Data Mining and Knowledge Discovery*, 35(3):1061–1086, 2021. 3

[21] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019. 1

[22] Zih Siou Hung, Arun Mallya, and Svetlana Lazebnik. Union visual translation embedding for visual relationship detection and scene graph generation. *Unknown Journal*, 2019. 1, 2

[23] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[24] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020. 3

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1, 2, 3, 5, 6

[26] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1356, 2017. 3

[27] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient

subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018. 2

[28] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270, 2017. 1, 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[30] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 1

[31] Gordon D Logan and Daniel D Sadler. A computational analysis of the apprehension of spatial relations. 1996. 2

[32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 3

[33] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297, 2016. 1

[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 3

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 1

[36] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3

[37] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 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*, pages 2930–2939, 2016. 1, 3

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[39] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1100–1109, 2015. 3

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1, 3, 6

[41] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 3

[42] Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005. 3

[43] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018. 3

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[45] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3, 5

[46] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13945, 2021. 1, 2, 5, 6, 7, 8

[47] Qiaoyu Tan, Yanming Yu, Guoxian Yu, and Jun Wang. Semi-supervised multi-label classification using incomplete label information. *Neurocomputing*, 260:192–202, 2017. 3

[48] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 1, 2, 5, 6, 7

[49] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 5, 6, 7

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2, 3

[51] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 2021. 3

[52] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 5

[53] Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. Tackling the unannotated: Scene graph generation with bias-reduced models. *arXiv preprint arXiv:2008.07832*, 2020. 1, 2, 3, 6

[54] Baoyuan Wu, Fan Jia, Wei Liu, Bernard Ghanem, and Siwei Lyu. Multi-label learning with missing labels using mixed dependency graphs. *International Journal of Computer Vision*, 126(8):875–896, 2018. 3

[55] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 1, 2, 3, 5, 6, 7

[56] Linli Xu, Zhen Wang, Zefan Shen, Yubo Wang, and Enhong Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *2014 IEEE international conference on data mining*, pages 1067–1072. IEEE, 2014. 3

[57] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 265–273, 2020. 2

[58] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021. 3

[59] Hao Yang, Joey Tianyi Zhou, and Jianfei Cai. Improving multi-label learning with missing labels by structured semantic correlations. In *European conference on computer vision*, pages 835–851. Springer, 2016. 3

[60] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 1, 2

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 5

[62] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020. 2

[63] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3736–3745, 2020. 2

[64] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 1, 2, 3, 6, 7

[65] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3, 5

[66] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 3

[67] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 1, 2

[68] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 3

[69] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE international conference on computer vision*, pages 589–598, 2017. 3