This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Probabilistic Modeling of Semantic Ambiguity for Scene Graph Generation

Gengcong Yang[1*†], Jingyi Zhang[2*], Yong Zhang[3‡], Baoyuan Wu[4,5‡], Yujiu Yang[1‡]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, [2]Ascend Lab, Huawei Technologies.
[3]Tencent AI Lab, [4]School of Data Science, The Chinese University of Hong Kong, Shenzhen,
[5]Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data
ygc19@mails.tsinghua.edu.cn, jgg_jingyizhang@foxmail.com, zhangyong201303@gmail.com
wubaoyuan@cuhk.edu.cn, yang.yujiu@sz.tsinghua.edu.cn

Abstract

To generate "accurate" scene graphs, almost all existing methods predict pairwise relationships in a deterministic manner. However, we argue that visual relationships are often semantically ambiguous. Specifically, inspired by linguistic knowledge, we classify the ambiguity into three types: Synonymy Ambiguity, Hyponymy Ambiguity, and Multi-view Ambiguity. The ambiguity naturally leads to the issue of implicit multi-label*, motivating the need for diverse predictions. In this work, we propose a novel plug-and-play Probabilistic Uncertainty Modeling (PUM) module. It models each union region as a Gaussian distribution, whose variance measures the uncertainty of the corresponding visual content. Compared to the conventional deterministic methods, such uncertainty modeling brings stochasticity of feature representation, which naturally enables diverse predictions. As a byproduct, PUM also manages to cover more fine-grained relationships and thus alleviates the issue of bias towards frequent relationships. Extensive experiments on the large-scale Visual Genome benchmark show that combining PUM with newly proposed ResCAGCN can achieve state-of-the-art performances, especially under the mean recall metric. Furthermore, we show the universal effectiveness of PUM by plugging it into some existing models and provide insightful analysis of its ability to generate diverse yet plausible visual relationships.*

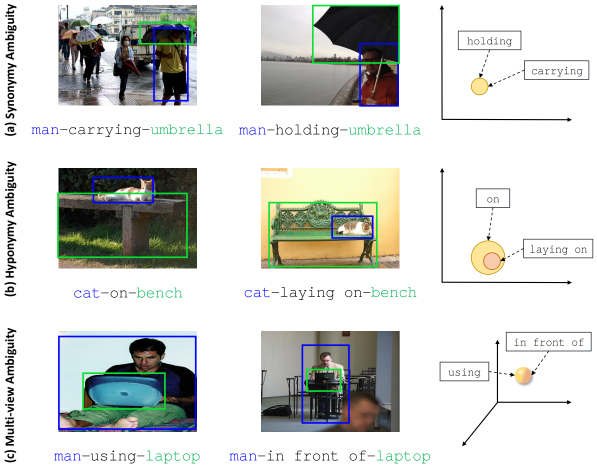

Figure 1. Examples of semantic ambiguity in Visual Genome dataset. The first two columns show the comparisons between two plausible predicates for similar visual scenes and the rightmost column illustrates the corresponding phenomenons in semantic space. (a) `carrying` and `holding` share overlapping definitions and are interchangeable to describe the relationship between a man and an umbrella. (b) Both `on` and `laying on` are reasonable to describe the scene where a cat is on top o a bench, despite their semantic specificity difference. (c) Different human annotators focus on different points of view, *i.e.* `using` (*actional*) vs. `in front of` (*spatial*), to describe a working man and a laptop.

1. Introduction

Scene graph generation (SGG) has been an important task in computer vision, serving as an intermediate task to bridge the gap between upstream object detection [21] and downstream high-level visual understanding tasks, such as image captioning [33, 42] and visual question answering [36]. Intuitively, the latter would get greater benefit from more human-like scene graphs.

Almost all existing works [30, 35, 4, 24] view SGG as an objective task and predict pairwise relationships in a de-

[*] Equal contribution.

[†] Work done in part during an internship at Tencent AI Lab.

[‡] Yong Zhang, Baoyuan Wu and Yujiu Yang are the corresponding authors. This research was partially supported by the Key Program of National Science Foundation of China under Grant No. U1903213 and the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011387). Baoyuan Wu is supported by the Natural Science Foundation of China under grant No. 62076213, the university development fund of the Chinese University of Hong Kong, Shenzhen under grant No. 01001810, and the special project fund of Shenzhen Research Institute of Big Data under grant No. T00120210003.

terministic manner. Namely, given a pair of objects, they always generate an identical predicate. However, compared with humans, such methods pursue "accurate" scene graphs but overlook the intrinsic semantic ambiguity of visual relationships. Specifically, the collaborative annotations from human annotators tend to be diverse, covering different descriptions of relationships for similar visual scenes.

We observe that there exist multiple types of semantic ambiguity in the large-scale Visual Genome dataset. Inspired by linguistic knowledge, we classify the ambiguity into three types. The first type is **Synonymy Ambiguity**, where multiple synonymous predicates that share overlapping definitions are suitable to describe similar visual scenes. For example, in Figure 1 (a), `carrying` and `holding` are interchangeable to describe the relationship between a man and an umbrella. If we visualize these two words in the semantic space, they should point to the same position where the visual relationship lies. The second one is **Hyponymy Ambiguity**, indicating that different humans tend to use predicates across adjacent abstract levels. One would simply use `on` to describe the visual scene where a cat is on top of a bench, while others may choose to use more fine-grained `laying on`, as shown in Figure 1 (b). In this case, `laying on` is a hyponym of `on`. Namely, the semantic range of the former is included by that of the latter. As for the third type, we notice that different human annotators often focus on different types of visual relationships, which originate from different points of view. Therefore, we refer to this phenomenon as **Multi-view Ambiguity**. An example is illustrated in Figure 1 (c), where both `using` (*actional*) and `in front of` (*spatial*) are plausible to describe the relationship between a working man and a laptop. If we consider the visual scene in three-dimension space, it can be a multicolor sphere that reflects different colors from different views. Although most predicates have single labels in the dataset, due to the ubiquitous semantic ambiguity mentioned above, we argue that many of them should have multiple labels, since similar visual scenes are annotated as different predicates. We refer to the issue as an *implicit multi-label* problem, which motivates the need to generate diverse predictions for visual relationships.

In this work, we focus on modeling the semantic ambiguity of visual relationships and propose a novel plug-and-play Probabilistic Uncertainty Modeling (PUM) module which can be easily deployed in any existing SGG model. Specifically, we utilize a probability distribution to represent each union region, rather than a deterministic feature vector as in previous methods. From a geometric perspective, the probabilistic representation allows us to map each visual relationship to a soft region in space, instead of merely a single point [26]. For ease of modeling, we adopt Gaussian distributions to represent them. Namely, each union region is now parametrized by a mean and variance. The former acts like the normal feature vector as in the conventional model, whereas the latter measures the feature uncertainty. To some extent, in this way, the feature instance of each union region can be viewed as a random variable drawn from a Gaussian distribution. Thanks to this uncertainty modeling, ambiguous union regions will be assigned to Gaussian distributions with large variances, which generate diverse samples and result in diverse predictions. As a byproduct, we find that PUM also manages to cover more fine-grained relationships and thus well alleviates the infamous issue of bias towards frequent relationships [4, 24].

We firstly demonstrate the effectiveness of PUM on the Visual Genome benchmark. Combining with the newly proposed Residual Cross-attention Graph Convolutional Network (ResCAGCN) in concurrent work [39], we achieve state-of-the-art performances under the existing evaluation metrics, especially the mean recall. Note that our PUM can serve as a plug-and-play component. Therefore, we plug PUM into state-of-the-art models and observe obvious universal improvement over these baselines, which mainly lies in the mean recall again. We owe the performance gain in the mean recall to the ability to generate diverse relationships, which improves the chances to hit the ground-truth with infrequent predicate labels. We further propose oracle recall as an indirect evaluation metric to measure the diversity of multiple inferences, which takes results of multiple consecutive predictions as an ensemble and computes recall. The oracle recall of the proposed model increases with the number of predictions, indicating that the model generates plausible diverse relationships and thus gradually covers the ground-truth more and more.

Overall, our contributions can be summarized as follows:

- We identify the semantic ambiguity of visual relationships and propose a novel plug-and-play Probabilistic Uncertainty Modeling (PUM) module, which utilizes a probability distribution to represent each union region instead of a deterministic feature vector.
- Combining PUM with ResCAGCN, we achieve state-of-the-art performances on the large-scale Visual Genome benchmark under the existing evaluation metrics, especially the mean recall.
- Extensive evaluations demonstrate the superiority of PUM to alleviate the bias towards frequent categories when plugged into the existing SGG models, reflected in the improvement on the mean recall.
- To the best of our knowledge, we are the first to explore diverse predictions for SGG. We conduct experiments both qualitatively and quantitatively to demonstrate that the proposed PUM module can generate diverse yet plausible relationships.

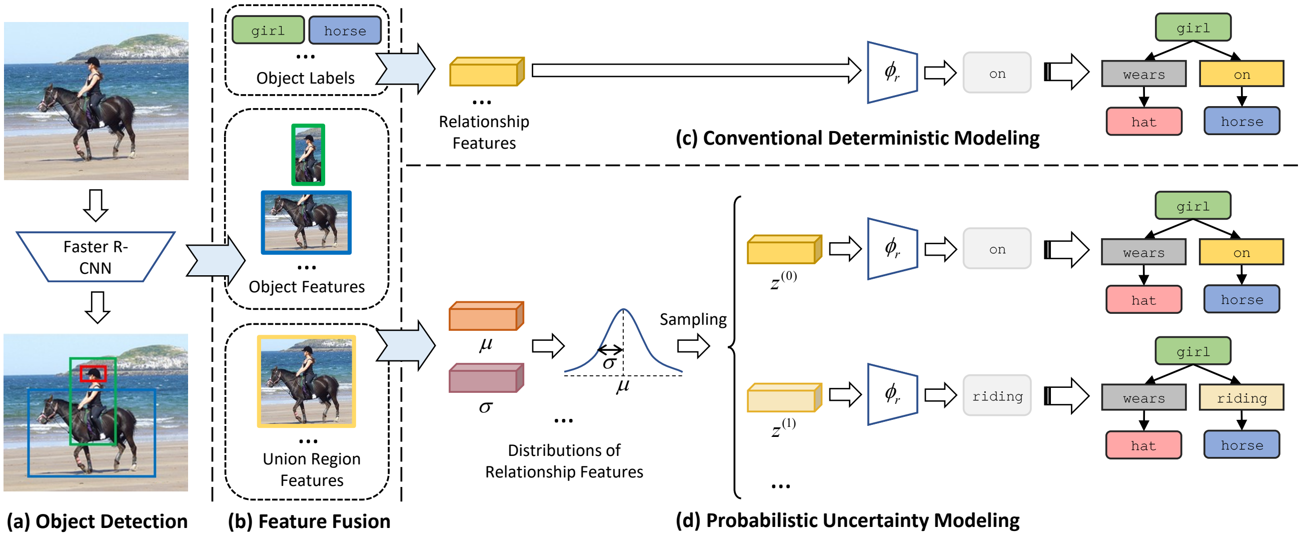

Figure 2. Existing SGG framework usually includes the following steps: (a) utilizing Faster R-CNN to obtain object proposals; (b) fusing features globally to obtain object labels, object features, and union region features; (c) conditioned on the results of the previous step, modeling each union region as a deterministic vector to predict the relationship. In this work, we replace (c) with Probabilistic Uncertainty Modeling (PUM) in (d), where each union region is represented by a probability distribution instead. In such way, diversity in scene graph generation is naturally achieved.

2. Related Work

Scene Graph Generation. Visual relationships have raised wide concern in the computer vision community since Lu *et al*. [18] formalized Visual Relationship Detection as a visual task. In early works [18, 44, 43, 38, 37], objects and pairwise relationships were detected independently. Such models overlooked the rich visual context and led to suboptimal performance. To make full use of visual context, later methods consider the whole image and utilized various kinds of message passing mechanisms [30, 15, 16, 32, 35, 24, 8, 27, 17]. For example, Xu *et al*. [30] were the first to formally define the problem of SGG and addressed it with iterative refinement via message passing. Afterward, Zellers *et al*. [35] represented the global context via LSTM, a recurrent sequential architecture. More recently, Chen *et al*. [4] incorporated statistical correlations into the graph propagation network. Meanwhile, Tang *et al*. [24] composed dynamic tree structures to allow content-specific message passing. While all these methods overlook the semantic ambiguity of visual relationships and make inferences in a deterministic manner, we propose to address the ambiguity via probabilistic modeling.

Uncertainty Modeling. Conventionally, the high-level representation of an input instance, *e.g*. an image or a word, is modeled as a fixed-length feature vector, namely, a single *point* in $\mathbb{R}^D$. However, such a point estimate is not sufficient to express uncertainty. In recent years, Gaussian embedding has been getting more attention in deep learning since [26] utilized it to represent words instead of the conventional word2vec [19], where the covariance naturally measures the ambiguity of the words. In the computer vision community, there exist prior works on modeling images as Gaussian distributions [20, 34, 3]. However, all the existing SGG methods represent each union region as a deterministic vector. In this work, we are the first to focus on the intrinsic semantic ambiguity of visual relationships and model each union region as a Gaussian distribution.

Diverse Predictions. Generally, there are two types of approaches to generate multiple diverse predictions. One is to train multiple models and aggregate their predictions. To better obtain diversity in the union of predictions, Multiple Choice Learning (MCL) [9] and other variants [14, 13, 25] were proposed to establish cooperation among all the models and train each to specialize on one particular subset of the data distribution. Another is to infer diverse predictions from a single model. Before deep learning, this type of methods mainly focused on probabilistic graphical models [2]. Afterward, existing single-model methods can be roughly categorized into two types: 1) via random noise added to Generative Adversarial Networks (GANs), applied in image captioning [5], image annotation [28], text generation [31] and so on; 2) mapping an instance to a probability distribution in latent space and sampling from it [6, 41, 7]. Our model can be regarded as the second category. While the other methods in this category usually utilize Variational Autoencoders (VAEs) in a generative way, we simply make use of Gaussian distributions to generate stochastic representations without a reconstruction loss and achieve diversity via this stochasticity in the inference stage.

3. Approach

Generally, a *scene graph* is a structured representation that describes the contents of a visual scene, which encodes object instances via nodes and relationships between objects via edges [10]. As defined by Xu *et al*. [30], the task of Scene Graph Generation is to generate an accurate visually-grounded scene graph associated with an image.

Mathematically, a scene graph can be defined as $G = \{B, O, R\}$, where B is a set of bounding boxes, O is object labels, and R is relationship labels. Conventionally, given an image I, the probability distribution of a scene graph $P(G|I)$ is decomposed into three factors [35, 4]:

$$P(G|I) = P(B|I)P(O|B,I)P(R|O,B,I). \tag{1}$$

Firstly, the widely used Faster R-CNN [21] is utilized to model $P(B|I)$ and generates a set of object proposals. Next, conditioning on the candidate bounding boxes, the object model $P(O|B,I)$ predicts the class label regarding each box. Finally, based on the result of object detection, the relationship model $P(R|O,B,I)$ infers the relationship of each object pair, leading to the whole scene graph for the current image. Existing works treat $P(R|O,B,I)$ as a deterministic model, which always generates an identical label for the same object pair. This framework is illustrated in Figure 2 (a)(b)(c). However, such a method overlooks the intrinsic semantic ambiguity of visual relationships and is likely to get stuck in the issue of biased predictions [4, 24], with the tendency to generate frequent and “safe” labels.

In this work, we propose a plug-and-play module for the relationship model, named Probabilistic Uncertainty Modeling (PUM), which addresses the semantic ambiguity mentioned above in a probabilistic manner. We replace the conventional deterministic modeling with PUM, as illustrated in Figure 2 (d). To better demonstrate the effectiveness of PUM, we adopt the newly proposed ResCAGCN [39] as our object model, which is introduced in Section 3.1. However, note that any object model from the existing SGG methods will be compatible with PUM theoretically. Then, we describe our PUM module in detail in Section 3.2.

3.1. Object Model

In our approach, we take Residual Cross-attention Graph Convolutional Network (ResCAGCN) from [39] as our object model to fuse object features and predict object labels.

The core of ResCAGCN is the cross-attention module ($\mathcal{CA}$), which is designed to capture the semantic relevance among the object features and the pairwise union region features. The module is formulated as:

$$\begin{aligned}\mathcal{CA}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = &\left(\boldsymbol{W}_i\boldsymbol{x}_i \odot \sigma\left(\boldsymbol{W}_j^{'}\boldsymbol{x}_j\right) \oplus \boldsymbol{W}_i\boldsymbol{x}_i\right) \\ &\odot \left(\boldsymbol{W}_j\boldsymbol{x}_j \odot \sigma\left(\boldsymbol{W}_i^{'}\boldsymbol{x}_i\right) \oplus \boldsymbol{W}_j\boldsymbol{x}_j\right),\end{aligned} \tag{2}$$

where $\odot$ and $\oplus$ denote element-wise product and sum, respectively. σ is the sigmoid function to normalize the attention scores. All $\boldsymbol{W}_*$ denote linear transformations to embed features into the same dimension, both here and below.

Given two object features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and their union region feature $\boldsymbol{u}_{ij}$, to model the contextual information, ResCAGCN utilizes the cross-attention module to compute the contextual coefficient $\boldsymbol{c}_{ij}$, which is formulated as:

$$\boldsymbol{c}_{ij} = \sigma\left(\boldsymbol{W}_c^T\left(\mathcal{CA}\left(\mathcal{CA}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right), \boldsymbol{u}_{ij}\right)\right)\right). \tag{3}$$

Instead of directly using the aggregated features as the output features, ResCAGCN uses a residual connection to add them back to the original features:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \mathrm{ReLU}\left(\boldsymbol{W}_1 LN\left(\boldsymbol{W}_2 \sum_{j\in\mathcal{N}_i} \boldsymbol{c}_{ij} \otimes \boldsymbol{W}_3\boldsymbol{x}_j\right)\right), \tag{4}$$

where $\otimes$ denotes Kronecker product, $\mathcal{N}_i$ denotes the i-th node’s neighborhood, and LN denotes layer normalization [1]. The refined object features $\hat{\boldsymbol{x}}_i$ are then fed into a classifier to predict the object labels.

3.2. Probabilistic Uncertainty Modeling

Conventionally, the union of two proposals is represented as a single point in space, namely, *point embedding* [20]. As [26] observed, however, such point estimate does not naturally express the uncertainty induced by the input. In the case of visual relationships, this could be caused by ambiguous annotations, *e.g.* `holding` and `looking at` may be both plausible to describe a scene containing a man and a cell phone.

As shown in Figure 2 (d), in order to capture the intrinsic uncertainty of visual relationships, we propose to *explicitly* model the feature distribution of each union region as Gaussian. That is, we represent each union region as *stochastic embedding* instead of the conventional *point embedding*. From a stochastic perspective, the final representation of each union region is no longer a deterministic vector but randomly drawn from a Gaussian distribution. As a result, we can generate predicates diversely for the same object pair, leading to diversity in scene graph generation.

Stochastic Representation. For each object pair, following ResCAGCN, we first fuse their contextual object features $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{x}}_j$, as described in Section 3.1, and their visual union region feature $\boldsymbol{u}_{ij}$ to obtain the relationship feature:

$$\boldsymbol{e}_{ij} = \hat{\boldsymbol{x}}_i \diamond \hat{\boldsymbol{x}}_j \diamond \boldsymbol{u}_{ij}, \tag{5}$$

where $\diamond$ denotes the fusion function defined in [40, 24], $\boldsymbol{x} \diamond \boldsymbol{y} = \mathrm{ReLU}\left(\boldsymbol{W}_x\boldsymbol{x} + \boldsymbol{W}_y\boldsymbol{y}\right) - \left(\boldsymbol{W}_x\boldsymbol{x} - \boldsymbol{W}_y\boldsymbol{y}\right)^2$. Based on each fused relationship feature, we define the associated representation $\boldsymbol{z}_{ij}$ in latent space as a Gaussian distribution,

$$p(\boldsymbol{z}_{ij}|\boldsymbol{e}_{ij}) = \mathcal{N}(\boldsymbol{z}_{ij}; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}^2_{ij}), \tag{6}$$

where $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\sigma}^2_{ij}$ refer to mean vector and diagonal covariance matrix respectively. They are formulated as:

$$\boldsymbol{\mu}_{ij} = \boldsymbol{W}_\mu \boldsymbol{e}_{ij}, \tag{7}$$

$$\boldsymbol{\sigma}^2_{ij} = \boldsymbol{W}_\sigma \boldsymbol{e}_{ij}. \tag{8}$$

At test time, we sample multiple $\boldsymbol{z}_{ij}$s, feed them into the relationship classifier ϕ_r respectively and compute the average posterior probability distribution:

$$P_{ij} = \frac{1}{K}\sum_{k=1}^{K} \phi_r(\boldsymbol{z}^{(k)}_{ij}), \tag{9}$$

where $\boldsymbol{z}^{(k)}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}^2_{ij})$, and K is the number of samples drawn from the Gaussian. Then we simply take the argmax of P_{ij} as the predicted relationship label.

Uncertainty-aware Loss. $\boldsymbol{\mu}_{ij}$ can be viewed as the original deterministic representation of the union box, while the random variable $\boldsymbol{z}_{ij}$ serves as a stochastic representation sample. Here, we consider both representations and feed them into ϕ_r respectively. Then, we train the relationship model with cross-entropy loss,

$$L_{ce} = (1-\lambda)\mathbf{CE}(\phi_r(\boldsymbol{\mu}), \boldsymbol{y}) + \lambda(\mathbb{E}_{z\sim p(z|e)}\mathbf{CE}(\phi_r(\boldsymbol{z}), \boldsymbol{y})), \tag{10}$$

where λ is the weight to trade off between deterministic prediction and stochastic predictions, and **CE** means cross-entropy loss. Note that we omit the subscripts ij for clarity.

In practice, we approximate the expectation term via Monte-Carlo sampling from $\boldsymbol{z}^{(k)} \sim p(\boldsymbol{z}|\boldsymbol{e})$:

$$L_{ce} \approx (1-\lambda)\mathbf{CE}(\phi_r(\boldsymbol{\mu}), \boldsymbol{y}) + \lambda(\frac{1}{N}\sum_{k=1}^{N}\mathbf{CE}(\phi_r(\boldsymbol{z}^{(k)}), \boldsymbol{y})), \tag{11}$$

where N is the number of samples drawn from the Gaussian. It is clear that conventional deterministic training can be seen as a special case of Eq. 11 where λ is set to 0.

Inspired by [34], as training progresses, the variance $\boldsymbol{\sigma}^2$ always decreases with L_{ce} alone and reverts our stochastic representation back to deterministic model. This problem could be alleviated by the following regularization term:

$$L_{reg} = \max(0, \gamma - h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2))), \tag{12}$$

where γ is a margin to bound the uncertainty level, and $h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2))$ is the differential entropy of a multivariate Gaussian distribution which is actually only related to $\boldsymbol{\sigma}$:

$$h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)) = \frac{1}{2}\log(\det(2\pi e\boldsymbol{\sigma}^2)). \tag{13}$$

It is obvious that L_{reg} will maintain the uncertainty level of the learned stochastic representations.

In conclusion, our final uncertainty-aware loss for the relationship model is expressed as:

$$L_{rel} = L_{ce} + \alpha L_{reg}, \tag{14}$$

where α is the weight of regularization term.

Reparameterization Trick. Sampling $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^{\mathbf{2}})$ directly will prevent gradients from propagating back to the preceding layers. Thus, we use the reparameterization trick [11] to bypass the problem. Specifically, we first sample a random noise ϵ from the standard Gaussian and generate $\boldsymbol{z}$ as the equivalent sampling representation,

$$\boldsymbol{z} = \boldsymbol{\mu} + \epsilon\boldsymbol{\sigma}, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{15}$$

4. Experiments

4.1. Experiment Setting

Dataset. We evaluate the proposed method on the popular large-scale Visual Genome (VG) benchmark [12], which originally contains 108,077 images with average annotations of 38 objects and 22 relationships per image. Since the majority of the annotations are noisy, following previous works [35, 4, 24], we adopt the most popular split from [30], which selects top-150 object categories and top-50 predicate categories by frequency.

Evaluation. We follow three conventional tasks to evaluate the proposed SGG model: (1) **Predicate Classification (PredCls)**: given the bounding boxes and their object labels in an image, predict the predicates of all pairwise relationships. (2) **Scene Graph Classification (SGCls)**: given the ground-truth bounding boxes in an image, predict the predicate as well as the object labels in every pairwise relationship. (3) **Scene Graph Detection (SGDet)**: given merely an image, simultaneously detect a set of objects and predict the predicate between each pair of the detected objects.

Since the distribution of relationships in the VG dataset is highly imbalanced, we follow [4, 24] to utilize mean Recall@K (short as mR@K) to evaluate each relationship in a balanced way. For reference, all the methods are also evaluated by the conventional Recall@K (short as R@K) metric.

Implementation Details. Following the prior works [35, 4, 24], we adopt the same Faster-RCNN [21] to detect object bounding boxes and extract RoI features. For the hyperparameters in PUM, we set K to 8, N to 8, λ to 0.1, and γ to 200. We optimize the proposed model by the Adam optimizer with a batch size of 8, and momentums of 0.9 and 0.999. Our method is implemented by Pytorch and MindSpore. Intuitively, our uncertainty modeling would cause variance of performances. However, in practice, under the

Table 1. Comparisons among various methods on mR@K (%). † denotes the re-implemented version from [35]. ↑ and ↓ indicate the performance change before and after plugging in PUM.

	SGDet		SGCls		PredCls		
Methods	mR@50	mR@100	mR@50	mR@100	mR@50	mR@100	Mean
IMP† [30, 35]	3.8	4.8	5.8	6.0	9.8	10.5	6.8
FREQ [35]	4.3	5.6	6.8	7.8	13.3	15.8	8.9
SMN [35]	5.3	6.1	7.1	7.6	13.3	14.4	9.0
KERN [4]	6.4	7.3	9.4	10.0	17.7	19.2	11.7
VCTREE-SL [24]	6.7	7.7	9.8	10.5	17.0	18.5	11.7
VCTREE-HL [24]	6.9	8.0	10.1	10.8	17.9	19.4	12.2
IMP† + PUM	4.5 ↑ **0.7**	5.5 ↑ **0.7**	6.4 ↑ **0.6**	6.8 ↑ **0.8**	11.3 ↑ **1.5**	12.3 ↑ **1.8**	7.8 ↑ **1.0**
SMN + PUM	7.5 ↑ **2.2**	8.6 ↑ **2.5**	9.4 ↑ **2.3**	10.1 ↑ **2.5**	16.4 ↑ **3.1**	18.1 ↑ **3.7**	11.7 ↑ **2.7**
KERN + PUM	6.5 ↑ **0.1**	7.4 ↑ **0.1**	9.9 ↑ **0.5**	10.6 ↑ **0.6**	18.7 ↑ **1.0**	20.4 ↑ **1.2**	12.3 ↑ **0.6**
VCTREE-SL + PUM	7.1 ↑ **0.4**	8.2 ↑ **0.5**	11.0 ↑ **1.2**	11.9 ↑ **1.4**	19.0 ↑ **2.0**	20.9 ↑ **2.4**	13.0 ↑ **1.3**
ResCAGCN [39]	**7.9**	8.8	10.2	11.1	18.3	19.9	12.7
ResCAGCN + PUM	7.7 ↓ 0.2	**8.9** ↑ **0.1**	**11.9** ↑ **1.7**	**12.8** ↑ **1.7**	**20.2** ↑ **1.9**	**22.0** ↑ **2.1**	**13.9** ↑ **1.2**

Table 2. Comparisons among various methods on R@100 (%).

	SGDet	SGCls	PredCls
Methods	R@100	R@100	R@100
IMP† [30, 35]	24.5	35.4	61.3
FREQ [35]	30.1	32.9	62.2
SMN [35]	30.3	36.5	67.1
KERN [4]	29.8	37.4	67.6
VCTREE-SL [24]	31.1	38.6	67.9
VCTREE-HL [24]	31.3	38.8	68.1
IMP† + PUM	25.0 ↑ **0.5**	35.7 ↑ **0.3**	61.8 ↑ **0.5**
SMN + PUM	30.6 ↑ **0.3**	37.4 ↑ **0.9**	67.5 ↑ **0.4**
KERN + PUM	29.8	37.1 ↓ 0.3	67.5 ↓ 0.1
VCTREE-SL + PUM	30.9 ↓ 0.2	38.1 ↓ 0.5	67.6 ↓ 0.3
ResCAGCN [39]	30.9	38.7	67.9
ResCAGCN + PUM	**31.3** ↑ **0.4**	**39.0** ↑ **0.3**	**68.3** ↑ **0.4**

hyperparameter setting mentioned above, we observe that the variance is always negligible enough[1] to be ignored.

4.2. Comparisons with State-of-the-Art Methods

Comparing Methods. In this part, we compare our model with existing state-of-the-art methods: (1) designed to improve the recall, including Iterative Message Passing (**IMP**) [30], frequency baseline without using visual contexts (**FREQ**) [35] and Stacked Motif Network (**SMN**) [35]; (2) intended for more balanced prediction on relationships, including Knowledge-Embedded Routing Network (**KERN**) [4] and Visual Context Tree model (**VCTREE-SL**, trained by supervised learning, and **VCTREE-HL**, trained by hybrid learning) [24]. Although [17] also reported new state-of-the-art performances recently, we argue that their results are not directly comparable to ours. Please refer to the supplementary material for details.

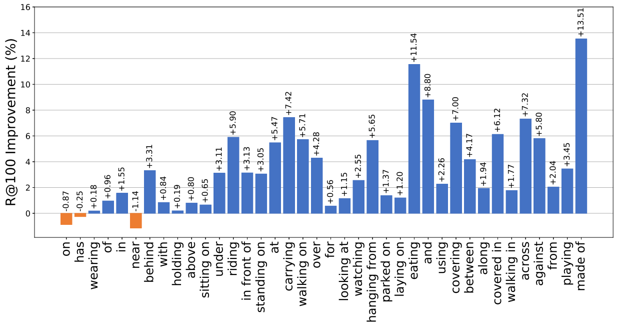

Figure 3. The R@100 improvement (%) of different predicate categories over VCTREE-HL in the PredCls setting. The x-axis labels are in descending order according to their number of samples in the dataset. Categories with the same performances are filtered out. Blue (orange) indicates increase (decrease) in performance.

Quantitative Results. From Table 1, compared with the previous state-of-the-art methods, the proposed model (**ResCAGCN + PUM**) shows the best performances on the mR@K metric, with a relative improvement of 13.9% compared with VCTREE-HL according to the mean. This indicates that the proposed model achieves notable improvement on infrequent categories. Meanwhile, it does not sacrifice frequent categories a lot, since its performances on R@100 also reach state-of-the-art, as shown in Table 2.

To gain a more comprehensive understanding of this

[1] 0.03% at most, with respect to mR/R@K.

phenomenon, as depicted in Figure 3, we further present the R@100 improvement of each predicate category over VCTREE-HL in the PredCls setting. Note that the x-axis labels are in descending order according to their number of samples in the VG dataset and categories with the same performances are filtered out. It is obvious that the proposed model achieves significant improvement in most categories. Importantly, the improvement is much larger for those infrequent categories in the long tail. We mainly owe this phenomenon to the byproduct of PUM, which endows the model with more chances to cover infrequent categories and thus alleviates the issue of biased predictions.

4.3. Ablation Study

To better prove the effectiveness of PUM, in this part, we explore the gain of PUM as a plug-and-play module. Firstly, as illustrated in the third part of Table 1, PUM brings about a significant increase over the vanilla ResCAGCN [39] by a margin of 1.2% on mR@K, according to the mean. Meanwhile, from Table 2, PUM also improves the model moderately on R@K. The results suggest that PUM indeed plays a critical role in the proposed model, which especially lies in more balanced predictions.

Then, we apply PUM to existing state-of-the-art methods, which conventionally utilize deterministic representations for visual relationships. Specifically, by regarding the original relationship features in each model as the $\boldsymbol{e}_{ij}$ in Eq. 5, we adopt the subsequent operations in Section 3.2 to model the uncertainty of relationships in a probabilistic manner. We present comparisons on mR@K between the existing state-of-the-art methods (IMP [30], SMN [35], KERN [4] and VCTREE-SL[2] [24]) and the counterparts plugged in PUM in the second part of Table 1. We find that PUM improves the performances of all models by a significant margin, with a relative improvement of 14.7%, 30.0%, 5.1% and 11.1% compared with the baselines respectively. The results show the universal superiority of PUM to the deterministic modeling, which mainly lies in the effectiveness to alleviate the issue of biased predictions towards frequent relationships. We also present comparisons on R@K in Table 2. Note that PUM does not necessarily improve all baselines under this metric. However, according to the discussions by Tang *et al.* [23], R@K is not a "qualified" metric for SGG, since simply catering to frequent categories while neglecting the infrequent ones would unfairly obtain a good performance. Meanwhile, similar to [23], we also observe that the performance drops caused by PUM mainly originate from classifying trivial "head" predicates into more fine-grained "tail" classes, *e.g.* from `on` to `parked on`.

4.4. Understand Uncertainty Modeling

We observe that the proposed model can generate diverse relationships, which helps to address the *implicit multi-label* issue caused by the semantic ambiguity. In this part, we qualitatively and quantitatively analyze this characteristic to gain more insights about our uncertainty modeling.

Qualitative Results. From Eq. 9, the relationship features fed into the classifier are randomly drawn from Gaussian distributions, resulting in varied predicted confidences, even for the same union region. Therefore, given a pair of objects, the proposed model can produce different plausible predicate at each inference. In other words, it is able to describe the same visual scene in different ways, leading to more human-like diverse predictions. This diversity well matches the three types of ambiguity illustrated in Figure 1. We show qualitative examples from two consecutive predictions of the proposed model in the PredCls setting in Figure 4. From the first row, the proposed model generates semantically-similar predicates consecutively, *i.e.* `at` vs. `near` and `holding` vs. `carrying`. Although the ground-truth only considers a single label, we argue that there exists such Synonymy Ambiguity, where multiple synonyms are plausible at the same time. Hyponymy Ambiguity is also a common phenomenon, where predicates across adjacent abstract levels are interchangeable. In the second row, the ground truth can be fine-grained (`walking on`) or coarse-grained (`on`). Thanks to our uncertainty modeling, the proposed model covers both levels of granularity and thus increases the chance of hitting the ground-truth. In Figure 1 (c), different human annotators tend to describe similar visual scenes from different points of view, resulting in Multi-view Ambiguity. We observe that the proposed model also simulates this phenomenon well. From Figure 4 (c), for the relationship between `person` and `sheep`, the proposed model focuses on either the spatial position (`behind`) or the person's action (`looking at`). On the other hand, for the scene on the right, the predictions could be a possessive verb (`has`) or an actional verb (`holding`). In a word, these qualitative examples suggest that semantic ambiguity is quite common when describing visual relationships, while the proposed model manages to generate *diverse* yet *plausible* predictions.

Oracle Evaluation. Inspired by the *oracle error rate* in Multiple Choice Learning [9, 14, 13, 25], we propose to use oracle Recall (short as oR) to measure the diversity of predictions indirectly, which counts a hit if one of the multiple consecutive predictions matches the ground-truth:

$$oR = \frac{\sum_{i=1}^{R} \mathbf{1}(\sum_{m=1}^{M} \mathbf{1}(\hat{y}_{m,i} = y_i) > 0)}{R}, \tag{16}$$

$$\mathbf{1}(x) = \begin{cases} 1 & x = \text{True}, \\ 0 & x = \text{False}. \end{cases} \tag{17}$$

[2] Since the reinforcement learning of VCTREE is independent of our method, we only conduct experiments on its one-stage supervised version for simplicity.

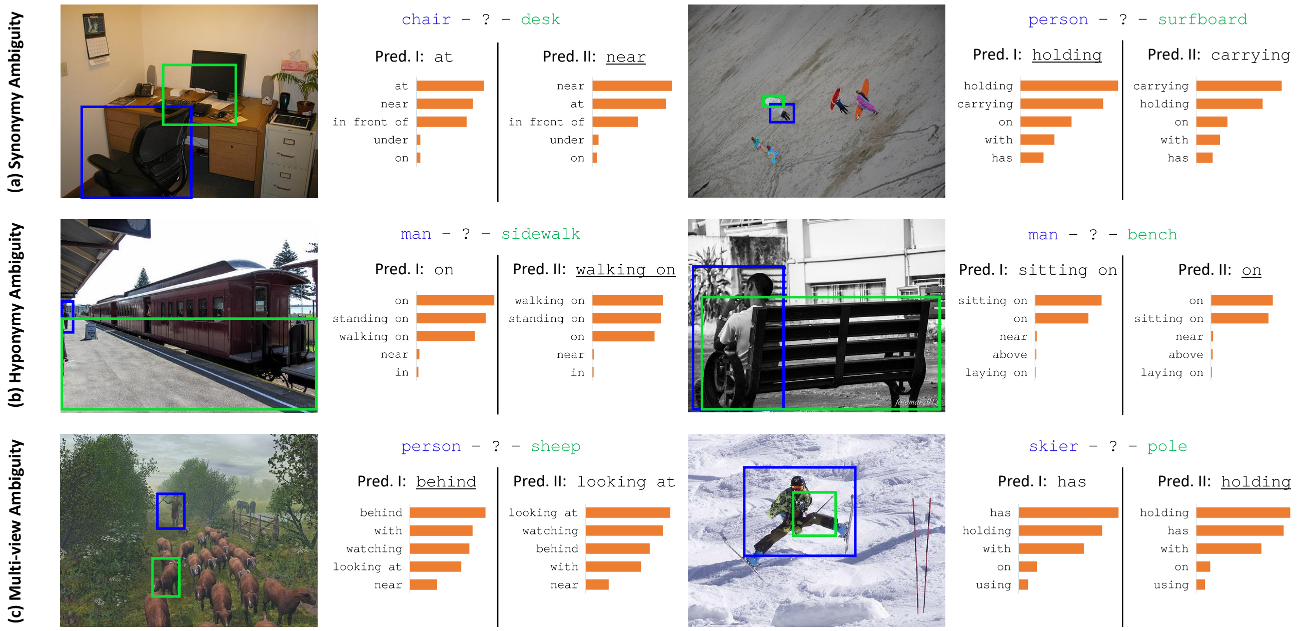

Figure 4. Qualitative examples from two consecutive predictions of the proposed model in the PredCls setting. Blue color indicates subjects; green color indicates objects; underlining indicates predictions that hit the ground-truth. For each prediction, we also show the top-5 classes with the highest confidences. Note that the confidences vary between two consecutive predictions, because of the stochasticity of the relationship features.

Above, R is the number of the ground-truth relationships in an image, M is the number of consecutive predictions, $\mathbf{1}(\cdot)$ is an indicator function, and $\hat{y}_{m,i} = y_i$ means the m-th prediction on the i-th relationship hits the ground-truth. if M is set to 1, oR is reduced to the normal recall. Note that we omit the averaging over all images in Eq. 16 for clarity.

In order to focus on the prediction of predicates, we conduct experiments in the PredCls setting. It is obvious that a model with the ability to make diverse inferences will achieve better performance under this metric. As illustrated in Figure 5, we evaluate the oR of the proposed model with and without PUM, respectively. As M increases, while the performance of ResCAGCN remains unchanged due to the lack of diversity, ResCAGCN + PUM gets improved steadily. The result suggests that our PUM not only boosts the coverage of predicted relationships in a single inference (as indicated when $M = 1$), but generates fresh relationships diversely in the next consecutive new predictions, which improves the opportunities to hit the ground-truth.

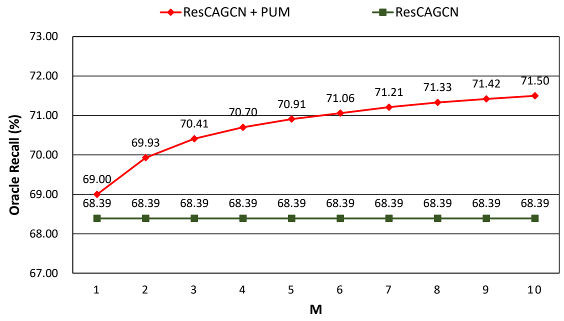

Figure 5. The effects of different number of consecutive predictions (M) on oR (%). The performance of ResCAGCN + PUM is higher than that of ResCAGCN and increases with M.

5. Conclusion

In this work, we considered the semantic ambiguity of visual relationships, which could be classified into Synonymy Ambiguity, Hyponymy Ambiguity and Multi-view Ambiguity. To address the implicit multi-label issue caused by the ambiguity, we proposed a novel plug-and-play module dubbed PUM. Although we aimed for diverse predictions, thanks to the byproduct of PUM, we achieved state-of-the-art performances under the existing evaluation metrics when combining it with ResCAGCN. Furthermore, we showed the universal effectiveness of PUM and explored its ability to generate diverse yet plausible relationships both qualitatively and quantitatively. A possible future direction would be to apply this kind of uncertainty modeling in down-stream tasks that also emphasize diversity, such as diverse visual captioning [5, 22, 29].

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse m-best solutions in markov random fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–16. Springer, 2012.

[3] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020.

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

[5] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[6] Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054, 2018.

[7] Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1229–1238, 2019.

[8] 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 Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2019.

[9] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2012.

[10] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] 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.

[13] Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. *arXiv preprint arXiv:1706.03475*, 2017.

[14] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016.

[15] 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.

[16] 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.

[17] 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.

[18] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 852–869. Springer, 2016.

[19] 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.

[20] Seong Joon Oh, Kevin P Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C Gallagher. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2018.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017.

[23] 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.

[24] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019.

[25] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6349–6357, 2019.

[26] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *International Conference on Learning Representations*, 2015.

[27] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2019.

[28] Baoyuan Wu, Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, and Siwei Lyu. Tagging like humans: Diverse and distinct image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7967–7975, 2018.

[29] Huanhou Xiao and Jinglun Shi. Diverse video captioning through latent variable expansion with conditional gan. *arXiv preprint arXiv:1910.12019*, 2019.

[30] 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.

[31] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, 2018.

[32] 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.

[33] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[34] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 552–561, 2019.

[35] 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.

[36] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. In *British Machine Vision Conference (BMVC)*, 2019.

[37] 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.

[38] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5678–5686, 2017.

[39] Jingyi Zhang, Yong Zhang, Baoyuan Wu, Yanbo Fan, Fumin Shen, and Heng Tao Shen. Dual resgcn for balanced scene graphgeneration. *arXiv preprint arXiv:2011.04234*, 2020.

[40] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018.

[41] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, 2017.

[42] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. *arXiv preprint arXiv:2007.11731*, 2020.

[43] Yaohui Zhu and Shuqiang Jiang. Deep structured learning for visual relationship detection. In *AAAI*, pages 7623–7630, 2018.

[44] 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.