State-aware Compositional Learning towards Unbiased Training for Scene Graph Generation

Tao He[1], Lianli Gao[2], Jingkuan Song[2], Yuan-Fang Li[1*]
[1]Department of Data Science and AI, Faculty of Information Technology, Monash University
[2]Center for Future Media, University of Electronic Science and Technology of China
{tao.he,yufang.li}@monash.edu, lianli.gao@uestc.edu.cn,jingkuan.song@gmail.com

***Abstract*—How to avoid biased predictions is an important and active research question in scene graph generation (SGG). Current state-of-the-art methods employ debiasing techniques such as resampling and causality analysis. However, the role of intrinsic cues in the features causing biased training has remained under-explored. In this paper, for the first time, we make the surprising observation that object identity information, in the form of object label embeddings (e.g. GLOVE), is principally responsible for biased predictions. We empirically observe that, even without any visual features, a number of recent SGG models can produce comparable or even better results solely from object label embeddings. Motivated by this insight, we propose to leverage a conditional variational auto-encoder to decouple the entangled visual features into two meaningful components: the object's intrinsic *identity* features and the extrinsic, relation-dependent *state* feature. We further develop two compositional learning strategies on the relation and object levels to mitigate the data scarcity issue of rare relations. On the two benchmark datasets Visual Genome and GQA, we conduct extensive experiments on the three scenarios, i.e., conventional, few-shot and zero-shot SGG. Results consistently demonstrate that our proposed Decomposition and Composition (DeC) method effectively alleviates the biases in the relation prediction. Moreover, DeC is model-free, and it significantly improves the performance of recent SGG models, establishing new state-of-the-art performance.**

***Index Terms*—Scene Graph Generation, Feature Decomposition, Compositional Learning, Data Augmentation.**

I. Introduction

SCENE graph generation (SGG), or visual relation detection (VRD), is a pivotal step towards scene understanding of visual contents. SGG has received enormous research interest in recent years [1]–[6] as it can provide detailed structured-representation of an image for conducting high-level visual reasoning, making it useful for many down-stream tasks such as visual captioning [7]–[9], visual question answering [10], [11], numan-object interaction detection [12] and 3D scene understanding [13], [14]. In general, SGG aims to produce an object-based relation graph, namely a *scene graph* (SG) that contains grounded visual representation of an image. Particularly, an SG can be presented as a set of relation triples, each of which can be denoted as a triple format, i.e., `<subject, PREDICATE, object>`.

One of the prominent challenging problems in SGG is the heavily skewed relation distribution in benchmark SGG datasets, e.g., Visual Genome [15]. Neural Motifs [16], which was the first to identify this skew, outperformed the then-SOTA models using solely the statistical prior knowledge of the co-occurrence of subject and object categories. This biased distribution leads to the realisation that, as long as the categories of the subject and object are known, the model can readily guess the relationship between them, even without resorting to any visual contents of an image.

A number of following works [17]–[21] have endeavoured to tackle this bias with a number of de-biasing techniques, including causality analysis [19], external knowledge [17], and energy-based training [21]. One commonality between these works is that they consistently use the word embeddings (e.g., GLOVE embeddings [22]) that contain strong object identities as the auxiliary features for relation classification. This raises a natural question that, *are visual features necessary at all if an SGG model is equipped with those category cues?*

With this question in mind, we empirically conduct a number of ablation experiments on several representative SOTA SGG methods by removing visual features from model training. Our analyses (§III) on both the standard and zero-shot SGG reveal some surprising results. In the conventional SGG, we observed that the object identity (category) clues are far more important to model performance than visual features, while visual features are almost redundant when the strong identities are available. Moreover, unexpectedly, when the GLOVE embeddings are replaced with randomly initialized object identity embeddings, i.e., removing the language prior in the GLOVE vectors, the model performance remains virtually unchanged. This observation demonstrates that the object identity information but not language priors, is crucial for relation prediction.

In zero-shot SGG, on the contrary, solely using object identify features results in a large performance degradation compared to the variant using visual features. These results confirm that although identity cues could benefit the prediction between known subject-object pairs due to the skewed relation distribution, they harm the model's generalizability to unseen pairs. In other words, SGG models would heavily rely on the object identity cues to predict predicates instead of learning visual relation patterns from images. More detailed analyses and experimental results are provided in §III.

Therefore, we deem that it is harmful for SGG models to

This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

[*]Corresponding author.

indiscriminately employ the object identity information, particularly for highly-skewed, natural datasets, e.g., VG, otherwise the ensuing over-reliance on this information would lead to biased predictions and inferiority of the learned relationship feature. Besides the explicit object identities in word embeddings, we find the visual features from object detection network (e.g., Faster-RCNN [23]) can also implicitly contain object identity information. Based on those observations, we propose to decompose an object's representation into two parallel representations: the intrinsic *identity* and the extrinsic *state*, that are responsible for classifying object labels and relation predicates respectively. More concretely, the identity feature aims to capture unique object category information, while the state feature focuses on variable relationship-related but not object-related information, such as pose or gesture. For instance, the object dog in Figure 3 can be conceptually represented in two parts: its identity that is unique to the dog category, and its state that is characterized by the pose of sitting.

In this paper, we model the object identity and state as the mean and variance of a Gaussian distribution by a conditional variational autoencoder network (CVAE) [24]. Naturally, we could assume that state features are category-agnostic. That is, object instances could share the same state feature regardless of their categories. For instance, "sitting dog" and "sitting wolf" have the same state descriptions, i.e., the action of "sitting", though dog and wolf are different classes. Thus, this separation motivates us to propose a dataset-wide compositional learning strategy to dynamically construct relation pairs beyond image-based construction techniques [1], [16], [25], which limits the subject and objects in a relation triple to be from the same image. Specifically, given a relation triple instance $<s, p, o>$, we can dynamically compose new and realistic triples by hallucinating s or o with other instances s' or o' of the same state p. The main benefit of our dynamic construction method is that we could generate more diverse relation samples so that the model can be effectively trained, especially for the data-starved relations.

In summary, our contribution are four-fold:

- For the first time, we empirically observe that directly feeding object identity cues into an SGG model could result in severely biased relation predictions, even without any statistic priors. Though, to some extend, it can improve the SGG performance, it harms a model's generalizability on few- and zero-shot learning.
- We propose to decompose the representation of an object into two task-specific components, the identity and state features, by a conditional variational autoencoder network. Compared to the conventional entangled representation, the decomposed representation mechanism can alleviate the biases in relation predictions caused by the involvement of the object identity clues.
- Based on the decomposed features, we develop a compositional learning strategy to generate additional relation samples to mitigate the problem of data-starving relations and particularly benefit the few- and zero- shot scenarios.
- We evaluate our method on two standard SGG datasets: Visual Genome and GQA. On a wide range of evaluation tasks, including in the few- and zero-shot settings, current SOTA methods enjoys substantial improvements when trained with our method.

II. Related Work

In this section, we briefly review three closely related areas: scene graph generation (**SGG**), compositional learning (**CL**) and decomposed representation (**DR**).

Scene Graph Generation (SGG) [26] is a task to detect visual relationships between objects in images, which requires an SGG model to localize and recognize objects. Most of the earlier SGG methods predominantly focused on refinement of object and relation features [1]–[3], [5], [6], [27]. For example, Xu *et al.* [1] first developed an iterative message passing (IMP) mechanism to refine object features and improve the quality of relation representations. Lu *et al.* [2] proposed to exploit language prior knowledge (e.g., GLOVE embeddings [22]) as the auxiliary information to align the visual content with its linguistic. More recently, researchers started to observe the biased relation distribution, revealing the fact that an SGG model can even predict the predicate as long as it is given the labels of the subject and the object. This insightful observation was first revealed by Neural Motifs [16], which achieves promising results solely by the statistical frequency priors. Subsequently, a suite of recent works [4], [17], [19], [21], [25], [28]–[30] invested heavily into addressing the data bias or long-tail problem in SG datasets. Although considerable performance improvements have been achieved by these models on many standard SGG evaluation tasks, particularly in terms of the unbiased metric of mean Recall@K (mR@K), all of them consistently ignore the biased impact of directly feeding subject and object identity clues into the relation classifier, as we discussed in the previous section. In this work, we first empirically find that, even without any visual cues or statistical information, SGG models can still learn the bias from the training set, resulting in severe biased relation prediction. To address this issue, we propose to decompose the representation for an object, aiming at decoupling the object's *identity* cue from its *state* cue and practically reducing biased relation predictions.

Compositional Learning has received tremendous attention in many vision tasks [31]–[34]. The core idea of compositional learning is to use limited samples to compose additional exemplars that maintain the main semantic meanings. Burgess *et al.* [31] proposed a compositional generative model, named Multi-Object Network (MONet), to decompose scenes into abstract building blocks by several Variational Auto-Encoder networks. For the multi-label classification problem, Alfassy *et al.* [35] proposed a implicit feature compositional network, named LaSO, to generate new multi-labels by simulating set operations. Concretely, they trained three parallel feature learning branches for union, subtraction and intersection operations by three objective functions. By those operations, they can synthesis novel composed features with diverse multi-labels. However, one defect of LaSO is that the synthesized features cannot be guaranteed with the consistent distribution with the original, because no any knowledge priors is used to ensure generated multi-labels are feasible. Latter on, researchers started to study composition learning in human-object interaction (HOI), defined as a human-object pair. Hence, an intuitive idea is to use extra human or object instances to compose versatile human-object pairs, but not restricted to the annotated HOI pairs. To this end, Kato *et al.* [32] devised a composition learning technique for zero-shot HOI detection by external knowledge graph and graph convolutional networks. A main difference from LaSO is that Kato *et al.* leverages external knowledge to ensure the composed sample are reasonable. Subsequently, Hou *et al.* [36] based on this idea proposed a a simple yet efficient visual composition learning (VCL) for HOI detection to generate new samples in the feature space. Besides, Peyre *et al.* [37] used analogies to learn the compositional representations for subjects, predicates, and objects, and during testing stage, use the nearest neighbor search to retrieve similar triples. In scene graph generation, Knyazev *et. al* [29] devised a generative compositional data augmentation strategy to generate hallucinated samples by a generative adversarial network (GAN). Unfortunately, both these works [29] ignore the biased predictions caused by an object's identify clue. In this work, we first decompose a object visual feature into two parts and compose new samples based on the decomposed features. One advantage of our method is that our method could eliminate the biased prediction caused by the object identity and largely increase the samples of tail relationships.

Decomposed (Decoupled) Representation (DR) [38]–[41] aims at decoupling high-dimensional features as several meaningful components for different subtasks and then aggregate them for the global task. The main challenge lies in how to develop a decomposition function so that decoupled features have consistent semantics with the corresponding labels. Sordoni *et al.* [38] first proposed to decouple the problem of mutual information estimation into a number of sub-estimation problems by leveraging the chain rule to aggregate them. Bai *et al.* [39] focused on the out-of-distribution (OOD) generalization and addressed this issue by a decomposed representation and semantic augmentation approach, named DecAug, by disentangling the category- and context-related features. The core idea of DecAug is to disentangle two important features: category-related features that are more related to causal information of an object and context-related features depicting side information, such as attributes, styles, backgrounds, or scenes, etc. Since both features are independent, DecAug can construct new samples by crosswise combining them. Jing *et al.* [40] proposed to learn various decoupled linguistic representations to resist language priors for visual question answering, since the language priors usually causes biased prediction in VQA. For graphical models, Wang *et al.* [41] developed a tree decomposition paradigm to decouple neighborhoods in a graph neural network (GNN) to overcome the feature smoothing problem among different layers of the GNN. To some extent, distribution embedding (DE) [42], [43] that decomposes a point feature into the mean and variance vectors can be viewed as a special case of decomposed representation, though DE focuses more on addressing the uncertainty problem in representation learning.

III. Object Identity Biases: a Motivating Study

Preliminary: without loss of generality, given a subject s and an object o, we could characterize the scene graph generation problem [1], [16], [19], [21], [25] as follows:

$$[\mathbf{u}_1, \mathbf{u}_2 \cdots \mathbf{u}_n] = \mathcal{F}(\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots \boldsymbol{v}_n) \tag{1}$$

$$\boldsymbol{p}_i^o = \boldsymbol{W}_o \otimes \mathbf{u}_i \tag{2}$$

$$\boldsymbol{p}_{so}^r = \boldsymbol{W}_r \otimes (\boldsymbol{W} \otimes [\mathbf{u}_s; \mathbf{u}_o]) \tag{3}$$

where $\mathcal{F}$ can be any context learning module (e.g., Bi-LSTM [16] or GCN [44]); $[;]$ and $\otimes$ are the concatenation operation and Kronecker Product, respectively; $\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots \boldsymbol{v}_n$ are comprehensive object features, e.g., the combination of the visual, spatial, and language priors from word embeddings while $\mathbf{u}$ denotes the refined object features; $\boldsymbol{W}_o$ is an object classifier and $\boldsymbol{p}_i^o$ is the prediction score of object labels; $\boldsymbol{W}$ is a linear project matrix and $\boldsymbol{p}_{so}^r$ is the relation score of a pair (s, o) classified by the linear classifier $\boldsymbol{W}_r$.

As noted in the literature recently [16], the biased relation distribution intrinsically exists in benchmark SGG datasets and it hampers the performance of SGG models. The prior work Neural Motifs [16] empirically demonstrated that simply using the statistical frequency of subject-object pairs without training any parameters can achieve competitive results, alluding to the important roles of the subject and object labels in relation prediction. In this work, we take it a step further and experiment with the idea of training SGG models simply with object identity clues. Thus, we pose the question that, *can a model generate scene graphs solely by the given the object identities (i.e. their labels) without visual content?* It is worth noting that this attempt is different from Freq [16] solely using the statistic co-occurrence probabilities between subjects and objects, because we need to train the full SGG models to predict relations based on the given features.

To answer this question, we first ablate visual representations from raw images in the VG dataset, and denote the ablated images as *scene layout images* as shown in Figure 1. Note that the relative spatial information of objects in the scene layout images remains untouched with the corresponding raw images. In practice, the features of region of interest (ROI) in scene layout images are presented by two types of object identity features: (1) word embeddings of object labels derived from GLOVE [22] vectors; and (2) randomly generated word embeddings aiming to ablate language priors from the GLOVE embeddings.

Experimental setups: we conduct the experiments on several representative SGG methods, such as Freq [16], Neural Motifs [16], TDE [19] and EBM [21] in terms of five features: (1) BASELINE using both visual features and GLOVE embeddings of object labels; (2)BASELINE* means replacing the GLOVE embeddings in the BASELINE with the random word embeddings; (3) VISUAL using the visual features; (4) GLOVE using the GLOVE word embeddings of object classes; and (5) RANDOM using the random word embeddings. We report the results of the metrics mR@K and R@K on the tasks of predicate classification and scene graph classification of SGG as well as the results of Zs-SGG on the task of

TABLE I

Results of the metrics mR@K and R@K on the tasks of predicate classification and scene graph classification with different features. Note that no statistical prior frequency knowledge [16] is applied for all models except for Freq [16]. ⋆ means replacing the Glove embeddings in the with the random word embeddings.

Models	Features	Predicate Classification				Scene Graph Classification			
		mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100
Freq [16]	-	13.0	16.0	60.6	62.2	7.2	8.5	32.3	32.9
Motifs [16]	Baseline	17.8	19.3	**64.3**	**65.8**	11.0	11.8	34.3	35.1
	Baseline⋆	18.0	19.4	64.0	65.5	11.2	11.5	34.5	35.0
	Visual	14.3	15.8	59.3	61.3	8.1	8.6	31.5	33.5
	Glove	19.2	**21.5**	64.2	65.4	**12.1**	**13.7**	34.2	**35.2**
	Random	**19.4**	21.2	63.9	65.2	11.7	13.5	**34.4**	34.9
TDE [19]	Baseline	25.4	28.7	47.2	51.6	12.2	14.1	25.4	27.9
	Baseline⋆	24.3	26.5	47.0	51.4	12.1	14.4	25.2	28.0
	Visual	20.4	23.7	45.7	48.9	11.6	12.8	22.6	25.2
	Glove	**25.3**	29.0	**49.1**	**53.3**	**13.7**	15.3	**27.5**	**29.0**
	Random	25.1	**29.3**	48.4	52.8	13.3	**15.4**	26.2	28.6
EBM [21]	Baseline	18.3	19.9	63.6	64.7	12.5	13.4	33.8	34.2
	Baseline⋆	18.2	19.8	**63.7**	**64.8**	12.2	13.7	33.5	34.0
	Visual	15.1	16.7	59.0	60.9	10.4	11.5	30.2	32.5
	Glove	**20.4**	**22.6**	63.1	63.8	13.0	14.4	**33.7**	**34.4**
	Random	20.2	22.4	63.0	63.6	**13.2**	**14.5**	33.3	34.2

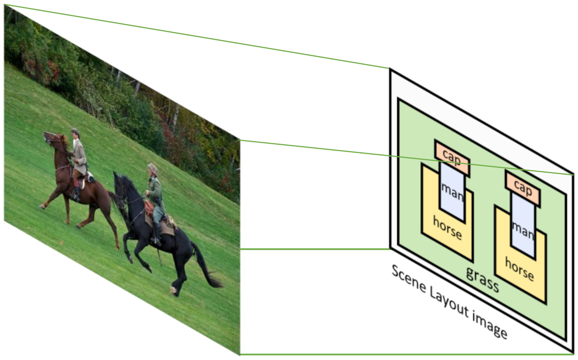


Fig. 1. An example of our de-visualization for an raw image (The image is *2353896.jpg* in Visual Genome [15] dataset.), while the left is a non-visual image, namely scene layout image, simply with the object identity and localization information.

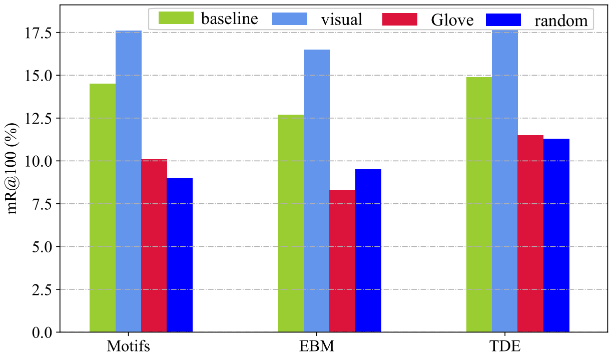


Fig. 2. ZS-SGG results of predicate classification on VG in terms of R@100.

predicate classification in terms of R@100. Note that we discard the third task scene graph detection on the both SGG scenarios, because it is infeasible to conduct an object detection network on scene graph layout images. Since the models can effortlessly classify the identity features into object categories on the task of scene graph classification, we use object classification results on raw images as the predicted object labels of the corresponding scene graph layout images for a fair comparison. We also apply the resampling strategy to all models, as [19] has demonstrated it is effective to alleviate the bias.

Table I shows the results of the metrics mR@K and R@K on the tasks of predicate classification and scene graph classification of SGG on the VG. Stunningly, however, based on the extensive experiments, we observe that, even though without any visual cues, those models on scene graph layout images (solely using the both identity features) can still show competitive or even better performance compared to their results on the full features (denoted Baseline in Table I). The independent two identity presentations, Glove and Random, show very close performance but consistently surpass Freq [16] merely employing statistical priors. This is possibly because Freq only considers the regional co-occurrence of subject-object pairs, while the variant models, e.g., Glove and Random, consider learning not only the regional co-occurrence but the global co-occurrence of all objects in an image. Hence, to some extent, they can learn more accurate bias priors. Moreover, we test the impact of the language priors by replacing the Glove embeddings used in the baselines with our randomly generated word embeddings, denoted as Baseline⋆ in Table I. Likewise, we could see that the modified baselines hardly suffer performance declines, e.g., the fluctuation is less than 0.2 point. These results further confirm that the benefit of the language priors is trivial but the object identity cues are vital in the Glove embeddings.

In addition, we investigate the more challenging task of zero-shot SGG. The results can be seen in Figure 2, where we test the three models on the four different feature sets: Baseline using both visual features and Glove embeddings, and the other single features as in Table I. Different from the observation on the standard SGG tasks, we could find that removing the visual features has a negative impact, e.g., a performance decrease of more than 30% from **Baseline** (green bar) to **Glove** (red bar). In contrast and surprisingly, solely using visual features (light blue bar) even outperforms **Baseline** (green bar).

These results confirm that although the learned bias knowledge from the object identity cues could benefit known subject-object pairs, it damages the model's generalizability for unseen pairs and eventually causes SGG models to heavily rely on object identity cues to predict predicates instead of essentially learning visual relation patterns from images.

In summary, we could make the following conclusions:

1) The addition of object identities could result in biased prediction which improves the performance of the seen relation pairs but inevitably harms the model's generalizability to unseen object pairs.
2) An SGG model can learn more accurate bias from the given object identity features than the statistical frequency bias [16].
3) The essential knowledge in the word embeddings (e.g. GLOVE) benefiting relation prediction is the object identity cues but not the language priors.

Based on the above conclusions, we believe that object identities should not be added indiscriminatingly for relation classification, especially on the heavily skewed datasets e.g. VG. That is, we should learn object and relationship features separately instead of using the entangled feature for both object and relationship classification. Therefore, we propose a representation decomposition mechanism to decouple the object identity cues from object representations to achieve reliable unbiased relation prediction.

IV. Methodology

Figure 3 shows the overview of our proposed framework, where the top half depicts the workflow of conventional SGG models, and the bottom half shows our decomposed unbiased SGG. More concretely, our framework consists of two modules: *feature decomposition* aiming at decomposing the object identity cues from comprehensive object features by a variational autoencoder network; and *relation and object composition* to construct additional relation triple instances to mitigate the highly skewed relation distribution.

A. Learning to Decompose Object Representations

As can be seen in our analysis above, the object identity is an essential cause of the biased relation prediction. In this subsection, we introduce our feature decomposition strategy based on a conditional variational autoencoder (CVAE) network.

1) Visual feature extraction: following the most SGG models [16], [47], we use the pre-trained Faster-RCNN [48] as the backbone object detection network. Subsequently, a region proposal network (RPN) is deployed to generate a set of n bounding boxes $\mathcal{B}_i = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ for each image i, where $\mathbf{b}_j = [x_t^j, y_t^j, x_b^j, y_b^j]$, (x_t^j, y_t^j) denotes the top-left coordinate and (x_b^j, y_b^j) is the bottom-right coordinate. The corresponding object visual features are represented as $\mathcal{O}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\}$, aligned by a RoIAlign module [23], where $\boldsymbol{v}_j \in \mathbb{R}^{4096}$ for each j. Note that, for simplicity, we omit the image subscript i in the following notations. Additionally, we extract the localisation feature of the object proposals and transform each bounding box $\mathbf{b}_j \in \mathbb{R}^4$ into a 128-dimensional informative spatial feature vector [16], denoted as $\mathbf{l}_j \in \mathbb{R}^{128}$.

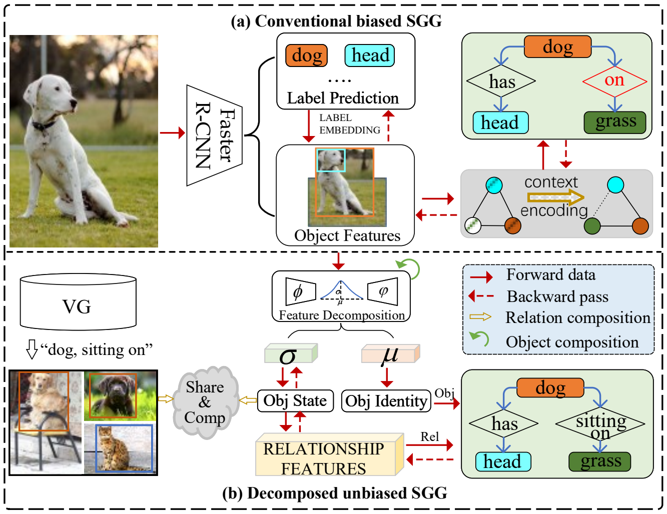


Fig. 3. The overall architecture of our framework, where the top block denotes general Scene Graph Generation models, such as Motifs [16], TransE [45] and GPSN [46], while the bottom pipeline denotes our proposed decomposed unbiased SGG. Specifically, our method first decomposes a visual feature from an object detection network (e.g., Faster-RCNN) into two parts, identity and state features, by a conditional variational autoencoder network (CVAE). To generate scene graphs, we leverage the identity feature for object classification and the state feature for relation classification. Moreover, we further use the state and identity feature to synthesize novel relation triples by our relation and object composition strategies to alleviate the data-starving issue in SGG datasets (e.g., VG).

2) Decomposing object identities: we could naturally assume that the object identities could be derived from two aspects: the explicit, extrinsic word embedding of object labels (e.g., GLOVE) [16], [21], [29] and the implicit, intrinsic object identity cues embedded in the object visual feature from the object detection network (e.g., Faster-RCNN). In this work, we focus on decomposing the latter, implicit identities, because for the former, we could directly discard the word embeddings. To this end, we attempt to decompose an entangled visual feature $\boldsymbol{v}_j$ into two separate components representing the object's *identity* and *state* as shown in Figure 3. More Concretely, the identity should contain the object category-unique information while the state feature excludes it, i.e., being category-agnostic, and captures more relationship-related but not object-related information.

Inspired by Gaussian embeddings [42], [49], they represent a point as a Gaussian distribution consisting of a pivotal mean and a diagonal variance vector. Concretely, the pivot mean can be viewed as the center point of a distribution, that is, the cluster center of a suite of data points, and is intuitively equipped with rich object identity clues. The other variance depicts the distribution's density or uncertainty [49]. However, different distributions could share the same uncertainty (i.e., with a similar distribution density), which makes it possible that the variance term can be shared between inter- or intra-object instances. Hence, we model an object's visual representation $\boldsymbol{v}_i$ as a Gaussian distribution $\boldsymbol{z}_i$, that is:

$$p\left(\boldsymbol{z}_j \mid \boldsymbol{v}_j, y_j\right)=\mathcal{N}\left(\boldsymbol{z}_j ; \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2\right) \tag{4}$$

where y_j is the object label; $\boldsymbol{\mu}_j$ denotes the identity, i.e., the mean vector of the distribution while $\boldsymbol{\sigma}_j$ is the state, i.e., the diagonal variance. In practice, we manipulate a conditional variational autoencoder network [24] to simulate the decomposition. Specifically, we first employ an encoder network ϕ to embed $\boldsymbol{v}_j$ into the mean and variance by:

$$\boldsymbol{\mu}_j=\mathrm{LN}_1(\phi(\boldsymbol{v}_j)),\ \boldsymbol{\sigma}_j=\mathrm{LN}_2(\phi(\boldsymbol{v}_j)) \tag{5}$$

where $\mathrm{LN}_1(.)$ and $\mathrm{LN}_2(.)$ are two projection functions consisting of multiple non-linear transformation layers for the identity and state. Subsequently, we utilize a decoder network φ to reconstruct the object feature $\boldsymbol{v}_j$.

To optimize Eq.(4), we employ a reconstruction and regularization loss as [24]:

$$\mathcal{L}_v=\underbrace{\mathbb{E}_{q(\boldsymbol{z} \mid \boldsymbol{v}, \boldsymbol{y})}[\log p(\boldsymbol{v} \mid \boldsymbol{y}, \boldsymbol{z})]}_{\text{reconstruction}}+\underbrace{\mathrm{KL}\left(q(\boldsymbol{z} \mid \boldsymbol{v}, \boldsymbol{y}) \| p(\boldsymbol{z} \mid \boldsymbol{y})\right)}_{\text{regularization}} \tag{6}$$

where $\boldsymbol{y}$ is the object label and in practice, we use the word embedding (e.g. GLOVE) to represent it; and $q(\boldsymbol{z}|\boldsymbol{v},\boldsymbol{y})$ and $p(\boldsymbol{v}|\boldsymbol{y},\boldsymbol{z})$ can be viewed as the encoder ϕ and the decoder φ, respectively. Following [24], we use a reparameterization trick to sample L points from the distribution, i.e., $\tilde{\boldsymbol{v}}_{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2\right)$, to model the Gaussian distribution.

B. Compositional learning for scene graph generation

In this section, we will elaborate our proposed category-agnostic relation classification based on our decomposed features. Specifically, we use the identity feature $\boldsymbol{\mu}$ for object label classification while the state feature $\boldsymbol{\sigma}$ is for relation classification. Then, to alleviate the data-starving problem, we further leverage the decomposed features and develop two level composition strategies: relation and object composition.

Category-agnostic relation classification: as analyzed before, $\boldsymbol{v}$ has implicit object identity clues, potentially leading to biased relation predictions. Thus, we use our decomposed state representations $\boldsymbol{\sigma}$ in Eq. (5) for relation classification, that is,

$$[\mathbf{u}_1, \mathbf{u}_2 \cdots \mathbf{u}_n]=\mathcal{F}(\boldsymbol{\sigma}'_1, \boldsymbol{\sigma}'_2 \cdots \boldsymbol{\sigma}'_n) \tag{7}$$

where $\boldsymbol{\sigma}'_i$ is a concatenation feature of $\boldsymbol{\sigma}_n$ and the spatial feature $\mathbf{l}_i$ and then use Eq. (3) to predict relation scores. At the same time, the identity feature $\boldsymbol{\mu}$ is employed for object classification:

$$\boldsymbol{p}_i^o=\boldsymbol{W}_o \otimes \boldsymbol{\mu}_i \tag{8}$$

Thus far, we address the biased prediction issue caused by the object identity. However, the extremely skewed relation distribution in the dataset still poses a challenge for the data-starved relations. To further debiase the predictions for tail relations, we develop two composition mechanisms to augment the relations and objects by synthesizing novel relation pairs and objects, respectively.

1) Relation composition: Recall in Eq. (1), for context learning, the object features $\{\boldsymbol{v}_i\}_{i=1}^n$ are obtained from the same image [1], [16], [19], [21], [25], hence, the number of all relation pair combinations is limited to the Cartesian product of the set of $\{\boldsymbol{v}_i\}_{i=1}^n$, namely $\{\boldsymbol{v}_i\}_{i=1}^n \times \{\boldsymbol{v}_i\}_{i=1}^n$ where × is the Cartesian product. In practice, although $|\{\boldsymbol{v}_i\}_{i=1}^n \times \{\boldsymbol{v}_i\}_{i=1}^n|$ is large, the majority are negative pairs, i.e., the non-relation, and the positive pairs are extremely scarce. Unfortunately, nevertheless, the positive pairs play a much more essential role in training an unbiased SGG model. To address this issue, Yang *et al.* [44] proposed a relation network to filter out redundant negative pairs, implicitly increasing the ratio of positive pairs during relation classification. However, the number of positive pairs does not increase, limiting the diversity of training data.

In this work, we propose a alternate relation composition method that doest not limit relation pairs to being constructed by the object instances only from the sampled image. For example, given a positive relation pair (s, o) with the annotated predicate r, we seek to replace s or o with another candidate object s' or o' from a different image to compose a new relation pair, e.g., (s', o) or (s, o') also with the same predicate r. Toward this goal, we need to address two questions: (1) how to guarantee the synthesized novel pairs have the same relation semantics with the original one; and (2) how to avoid involving the object identity clues in the new pair.

As mentioned before, the state feature $\boldsymbol{\sigma}$ can be naturally shared for different $\boldsymbol{\mu}$, i.e., different object categories. In other words, if s and s' have the same or close state $\boldsymbol{\sigma}$, we view that they are mutually replaceable. However, during training, we can not pointwise calculate s' state with all other s, which is extremely time-consuming. Hence, the remaining question is how to devise a simple and effective way to identify whether or not any two object instances have the same state.

To this end, we treat the combination of relation label and the object label as the *state label* for the subject/object in a relation triple. For example, given a relation annotation <dog, sitting on, grass>, we designate the dog's state label as "dog-standing on". However, due to the directionality of the predicate, we use the position of the predicate to differentiate the subject and object, namely "sitting on-grass" for the grass. In this way, we could know "dog-sitting on" and "sitting on-dog" are different state labels. Note that since an object could have different relationships with other objects in an image, an object instance possibly has several different state labels. Therefore, the state label of an object instance is a set instead of a single one. If the intersection of two instances' state labels is not empty, we deem that they have the same state.

Based on the object labels of s' and s, we could divide the composition into two types: *intra-category relation composition*, which requires both of them to have the same label, and *inter-category relation composition*, where they belong to different object categories. For the intra-category relation composition, we can only consider the candidate object with the same state label as s'. Figure 4 shows an example of the intra-category relation composition. For the latter, we first use the language prior, i.e., the word embedding $\boldsymbol{\mu}_i$ of the object category, to retrieve semantically similar object categories by a similarity threshold. For example, for object category "car",

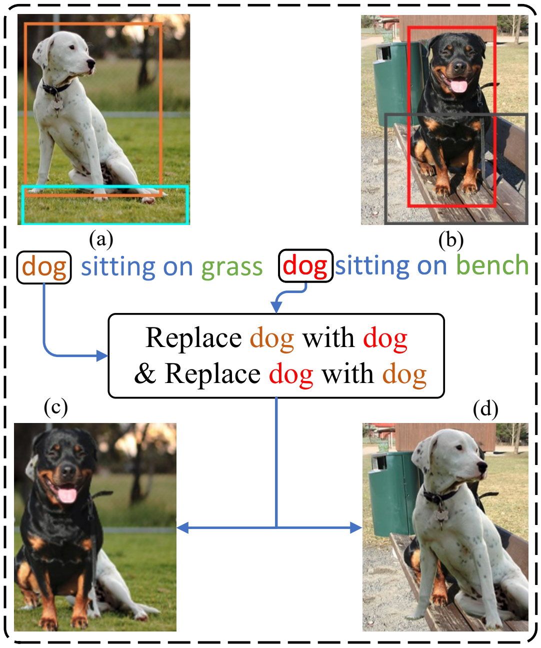


Fig. 4. An illustration of our proposed intra-category relation composition. Since both dogs in the two images have the state label,i.e., ”dog-sitting on”, we can interchangeably replace the dogs and synthesis other two novel relation triples, i.e., Figure (c) and (d).

the retrieved object categories are likely to be “train”, “plane” or “bike”, as all of them are vehicles having similar language prior. Once we obtain the feasible object categories, we select the candidate s' with the same state label as s ignoring the object part. It is worth noting the inter-category composition method could create novel subject-object combinations, which could benefit the zero-shot scenario in SGG.

2) Object composition: Intuitively, we could reuse the reconstructed $\tilde{\boldsymbol{v}}$ by the conditional variational auto-encoder to augment object samples. Specifically, since we represent an object's feature as a Gaussian distribution, we could sample extra points from the distribution. Given $\tilde{\boldsymbol{v}}_{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2\right)$, we first reconstruct the visual feature based on $\tilde{\boldsymbol{v}}_{(k)}$ using the decoder φ, and then decompose it by:

$$\boldsymbol{\mu}'_k = \mathrm{LN}_1(\phi(\varphi(\tilde{\boldsymbol{v}}_{(k)}))),\ \boldsymbol{\sigma}'_k = \mathrm{LN}_2(\phi(\varphi(\tilde{\boldsymbol{v}}_{(k)}))) \quad (9)$$

Where $\mathrm{LN}_1(.)$ and $\mathrm{LN}_2(.)$ are the same as Eq. (5). Then, we treat $\boldsymbol{\sigma}'_k$ as the candidate object to replace its original state feature $\boldsymbol{\sigma}_k$ and further augment the relation triples.

3) Training: There are in total two main components: representation decomposition and scene graph generation. We jointly optimize them as follows.

$$\mathcal{L} = \mathcal{L}_o + \mathcal{L}_r + \lambda \mathcal{L}_v \quad (10)$$

where $\mathcal{L}_o = \mathrm{CE}(\boldsymbol{p}_i^o, y^o)$ and $\mathcal{L}_r = \mathrm{CE}(\boldsymbol{p}_{so}^r, y^r)$ are the object and predicate classification losses by minimizing the cross-entropy (calculated by the function of $\mathrm{CE}(.)$) of Eq.(8 and 3), respectively; and $\mathcal{L}_v$ denotes the decomposition loss in Eq.(6) balanced by the hyper-parameter λ.

Finally, it is worth noting that since our method is model-free, a major advantage is that it can be readily plugged into existing SGG models to improve their performance.

V. Experiments

In this section, we evaluate our decomposition and composition strategy (denoted **DeC**) on the task of standard SGG as well as two more challenging settings, namely few-shot and zero-shot SGG (FS-SGG and ZS-SGG respectively). We further discuss the effectiveness of each component in an ablation study. The experiments are conducted on the widely-used SGG dataset Visual Genome [15] as well as the more challenging dataset GQA [50].

A. Datasets

1) Visual Genome (VG) [15]: a prevailing benchmark dataset for SGG. Following the widely adopted split [1], [16], [19], [51], we choose the most frequent 150 object classes and 50 relations to generate scene graphs, with 57,723 images for training and 26,443 images for testing. Additionally, 5,000 images make up the validation set to select the best model and to finetune model parameters. As revealed by many previous works [16], [19], [47], the relation distribution in VG dataset is extremely skewed. Specifically, the top 10 most frequent head predicates have almost 90% samples while the remaining 40 predicates simply account for ∼10%.

2) GQA [50]: a much more challenging and complete dataset derived from VG but has richer object category information and predicate words. More specifically, it contains 1,704 object categories and 311 predicate words. Following the split of [21], we have 72,580 training images, 2, 573 validation images, and 7, 722 test images.

B. Baselines

To demonstrate the effectiveness of our method, we select the following representative state-of-the-art SGG methods as our comparison baselines.

- IMP [1] first leverages an iterative message passing mechanism to improve the quality of object and relation representations by a graph neural network.
- Motifs [16] points out the relation distribution bias in the VG dataset and proposes a sequence-to-sequence to model the relation prediction.
- Freq [16] directly uses the statistical frequency prior as the relation prediction score.
- VCTree [25] proposes dynamic tree structures to model the objects in an image into a visual context.
- TDE [19] uses a suit of causality analysis techniques to remove the prediction bias in the training stage.
- GCA [29] develops a generative compositional augmentation method to hallucinate extra relation triple samples to mitigate the long-tail problem.
- BGNN [52] developed a confidence-aware bipartite graph neural network to adaptively propagate messages for unbiased SGG.
- EBM [21] uses a energy-based loss function to constrain the structure in the output space and enables the model to learn from a small number of labels.

Noticeably, GCA, TDE and EBM are three model-free methods, and for a fair comparison, we choose the VCTree as the base model for all of them.

C. Evaluation Metrics

We evaluate our technique on the three common task of SGG and its few-shot (FR-SGG) and zero-shot (ZR-SGG) variants, which are increasingly difficult due to the reduction in training samples. Specifically, FR-SGG tests a model's ability to learn from a few training examples. ZR-SGG [2] aims to evaluate a model's generalisability on unseen relation triples. At the testing stage, we select new relation triples, i.e., the combinations of subject-object pairs that are not present in the training set, to evaluate the SGG models following TDE [19].

Each task includes three sub-tasks: (a) Predicate Classification (**PredCls**) with ground-truth object labels and boxes provided; (b) Scene Graph Classification (**SGCls**) with only the labels provided; and (c) Scene Graph Detection (**SGDet**) with neither of them.

In this paper, we mainly report the unbiased metric mean Recall@K (**mR@K**) instead of the conventional metric **R@K**, as it has been shown in the literature that **R@K** is biased and does not reflect a model's true performance on tail relations [17], [19], [51]. Note that all experiments are under the 'constraint' scheme [17].

D. Implementation Details

Following recent works [16], [19], we use the pre-trained Faster-RCNN [48] with the backbone of ResNeXt101-FPN [53] as the object detection network and freeze its parameters. The dimension of GLOVE word embedding is set to 300 as in Motifs [16], and the object identity feature $\boldsymbol{\mu}$ and the state feature $\boldsymbol{\sigma}$ are also 300-D. The hyper-parameter λ is set to 1. The counterparts of non-linguistic embeddings are randomly initialized vectors with the same dimension. Each of $\mathrm{LN}_1(.)$ and $\mathrm{LN}_2(.)$ consist of three non-linear transformation layers. For the reparameterization trick, we empirically set L as 64.

In relation composition, we choose the relation pairs (s, o) where s and o have a small overlap to compose, as they are much more decomposable because they contain less background information. We use IoU [23] score to measure the overlap of two objects. Besides, for the candidate instance selection, we do not generate all object's state features, but explore an object state feature memory bank to dynamically store the candidates during training. When the bank is full, we randomly replace some old candidates with the new ones. Additionally, as mentioned before, we select the candidate with the same state label as the original object. However, in practice, there are a numerous candidates meeting this requirements. Therefore, we further design another measurement, i.e., bounding box shape similarity, to choose the best-match candidate. More concretely, we first align two bounding boxes' center coordinates and calculate their IoU score [48] as their box similarity. In inter-category composition, we first use word embeddings from GLOVE to retrieve the semantically similar object category based on the object label of the replaced object and set the category similarity threshold as 0.4, i.e., when two categories' word embedding similarity is greater than 0.4, we think the object category is valid, and then follow the techniques in intra-relation composition to synthesize inter-relation samples. Besides, since the top 10 frequent relationships have sufficient samples, we do not synthesize additional examples for them, but only augment samples of the remaining other relationships. Based on our statistics, during training, we construct extra 800K novel relation triple instances.

All experiments are conducted on four 2080 Ti GPUs and we implement our experiments based on the released code [19].

E. Main results

As discussed above, object identity is the key leading to biased predictions. In this section, thus, we report the results of the state-of-the-art methods with object identity cues removed. Results can be seen in Table II, where we report results under two metrics: R@K and mR@K. We apply our decomposition and composition strategy (**DeC**) to three representative SOTA SGG models: Motifs [16], EBM [21], and TDE [19]. It is worth noting that for EBM, we provide another competitive baseline (denoted EBM*), in which we apply the TDE technique to EBM. Also, the resampling strategy are applied to all models, but the statistical prior information is discarded for a fair comparison.

In terms of the unbiased metric mR@K, it can be clearly observed in Table II that TDE + **DeC** achieves the best performance in all terms. Equally importantly, when equipped with our proposed **DeC** strategy, the three baseline models, Motifs, EBM and TDE, all significantly surpass their respective performance without **DeC**. For instance, TDE + **DeC** obtains approx. 22% improvements on average compared with its baseline across all three subtasks, and surpasses the competitive method EBM*. On the conventional biased metric R@K, we could see that EBM + **DeC** gains the best results over five terms, except for R@100 of the prediction classification. All baseline models with the object identities suffer from significant performance degradation compared to their original results. Moreover, we could observe that many baselines are even worse than Freq [16] solely using the statistic information from the training set. This again confirms the critical role of the object identity in relation prediction. However, when applied with our DeC, the three methods consistently acquire relative improvements, especially TDE, with about $\sim$3 points on average. On both metrics, all results can demonstrate the effectiveness of our proposed **DeC**.

For a more comprehensive comparison, we further test specific improvement on each predicate with object identity information removed. As shown in Figure 5, we report the results of mR@100 on the task of PredCls over two representative methods, Motifs and TDE. The predicates on the x-axis are sorted by the numbers of instances in the training set, with the most frequent on the left.

From this analysis, we can observe that while our model achieves competitive on the head relations, it shows significantly better performance on the tails relations. It can be seen that both Motifs and TDE achieve a 0 or near 0 R@100 value for all of the 15 least frequent relations (`part of` onwards). In contrast, TDE + **DeC** achieves substantially better results, with R@100 scores of at least 15% for 10 of

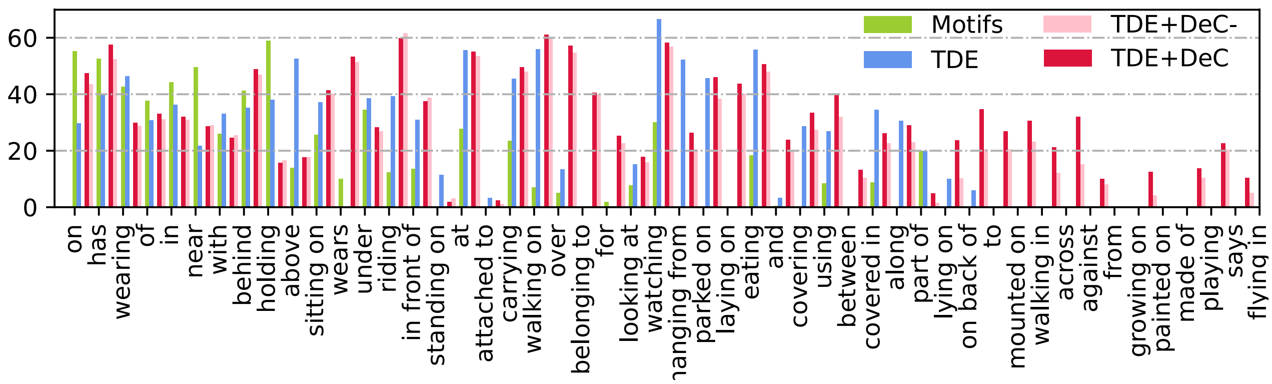


Fig. 5. R@100 results of PredCls on VG. Note that the frequencies of the relations on the x-axis descend from left to right.

the 15 relations. Interestingly, we can observe that although both Motifs and TDE have discarded the GLOVE embeddings, they are still unable to predict the majority of tail predicates. This demonstrates that even if we remove the explicit object identities, the model can still suffer from biased predictions. As we analysed in Section III, this is mainly because the visual features contain many implicit object identity cues from the object classifier in an object detection network (e.g. Faster-RCNN).

In addition, we also apply **DeC** to the state-of-the-art methods in the biased scenario, i.e., adding the full object identity clues (including both explicit and implicit). Note that we remove the decomposition module for DeC in this setting as we need to maintain the implicit object identities in visual features. The results are shown in Table III. It is evident to see when the object identity cues are added into the relation features, a significant performance improvements over all the metrics can be observed. As discussed in Section III, this is mainly because the object identity information provides strong biased cues for relation prediction and the models can readily recognize relationships. Nonetheless, this identity information hampers a model's generalizability to unseen relation pairs (also see the results on the few- and zero-shot SGG in the following Sec. V-F) due to the over-reliance on it during training, also as discussed in Sec. III. To some extent, although Table III shows better performance, Table II can fairly reflect the model's performance. Particularly, in Table III TDE, when applied with **DeC**, gains the best performance in the majority of terms except for the task of SGCLS in terms of mR@100. This demonstrates our proposed composition learning strategy is also effective for biased SGG.

On the other challenging dataset GQA, we follow the settings in EBM [21], and report the results on the metric of mR@K only on the tasks of PREDCLS and SGCLS, as shown in Table IV. Additionally, we further change Motifs' encoder network Bi-LSTM into the Transformer [54] as done in EBM [21]. We denote this model as Trans. From the results, we could observe that the Transformer-based encoder gains the best results over all metrics, and when plugging our **DeC** strategy into the baselines, noticeable improvements are achieved, e.g., about 25% on average with respect to Motifs. Hence, **DeC** is not only effective on VG, but also on GQA, demonstrating its wide applicability.

F. Results on FS-SGG and ZS-SGG

Few-shot learning for SGG is an important and realistic task, as the tail relations have very few samples. To evaluate the performance of FS-SGG, we randomly sample S images for each predicate to train the models. More concretely, we select $S = [5, 10]$ as the number of images per predicate for training. Table V shows the results of FS-SGG on the three subtasks, where we also apply our **DeC** strategy to Motifs, TDE and EBM. The statistical bias knowledge proposed in Motifs [16] and GLOVE embeddings are discarded for all methods. Besides, we report the results of baseline models + **Dec-**, i.e., adding object identities and our proposed compositional learning strategy.

From the results it is easy to see the challenging nature of the FS-SGG task. Though GCA achieves competitive results on the three tasks, it can be clearly observed that when equipped with **DeC**, Motifs, TDE and EBM gain results comparable to and better than GCA, except for the metric R@100 on PREDCLS. Moreover, we could find that with the object identity removed, the models gain modest improvements. For example, Motifs† is on average better than Motifs by 1.43 at $S = 5$, and the same observations can be made on TDE. These results confirm that the addition of object identity could cause a model's over-reliance on the identities, and deteriorates the model's generalizability on the few-shot scenario. Besides, when applying our composition learning technique to the original baseline models i.e., with the object identities, we could see there are noticeable improvements compared to the original baselines. However, their performance is lower than our full model without the object identity and with our composition learning technique. This is possibly because although our composition strategy can generate more examples during training, the intrinsic bias still exists in the constructed exemplars due to the addition of the object identities.

Table VI shows the comparison results on the task of ZS-SGG. Following the settings in TDE [19], we select unseen relation triples of the test set as the evaluated samples. Likewise, the statistical prior knowledge is removed during training stage. Note that we provide two baselines for EBM, Motifs and TDE, that is whether or not to remove the explicit object

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART SGG MODELS ON THE METRIC OF MR@K IN THE VG DATASET. † DENOTES THE RESULTS WITHOUT THE EXPLICIT OBJECT IDENTITY, I.E., THE GLOVE EMBEDDINGS, AND **DEC** DENOTES OUR PROPOSED DECOMPOSITION AND COMPOSITION METHOD. ⋆ DENOTES THE METHOD IS APPLIED TDE [19].

Method	Predicate Classification				Scene Graph Classification				Scene Graph Detection			
	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100
Freq [17]	13.0	16.2	60.6	62.2	7.2	8.5	32.3	32.9	6.1	7.1	26.2	30.1
IMP [1]	9.8	10.5	58.3	60.3	5.8	6.0	33.0	34.2	3.8	4.8	20.7	24.5
VCTree† [25]	12.9	15.4	59.2	61.5	9.5	10.2	35.2	36.4	6.5	7.7	26.3	30.3
GCA† [29]	17.8	18.3	58.2	60.4	11.2	12.6	34.2	35.3	9.0	10.2	26.0	29.8
BGNN† [52]	23.5	26.4	57.0	58.9	13.0	15.2	33.4	34.5	10.3	11.8	27.8	31.5
Motifs† [16]	14.3	15.8	59.3	61.3	8.1	8.6	31.5	33.5	5.6	6.8	26.3	29.5
Motifs + **DeC**	18.3	20.3	59.2	60.6	11.8	12.3	34.6	35.9	9.0	10.4	27.7	30.8
EBM† [21]	15.1	16.7	59.0	60.9	10.3	11.4	30.2	32.6	7.2	9.0	25.7	29.6
EBM† + **DeC**	17.4	19.3	**60.7**	61.8	10.7	12.2	**35.8**	**36.7**	8.7	10.0	**28.6**	**32.4**
EBM†*	21.7	24.3	51.7	54.4	12.6	13.8	25.7	28.1	9.6	11.6	19.0	23.7
EBM†* + **DeC**	23.3	25.8	54.2	56.1	13.4	14.3	28.3	29.5	10.5	12.0	22.5	25.1
TDE† [19]	20.4	23.7	45.7	48.9	11.6	13.0	22.6	25.2	9.3	11.1	18.9	22.7
TDE† + **DeC**	**25.1**	**28.9**	49.5	51.3	**14.2**	**16.1**	25.3	28.7	**12.0**	**13.6**	21.4	25.2

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART SGG MODELS ON THE METRIC OF MR@K IN THE VG DATASET WITH THE BIASED FULL OBJECT IDENTITY CUES. **DEC**- REPRESENTS OUR METHOD WITHOUT THE DECOMPOSITION COMPONENT.

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 [1]	7.1	9.8	10.5	5.0	5.8	6.0	3.1	3.8	4.8
Freq [17]	8.3	13.0	16.2	5.1	7.2	8.5	4.5	6.1	7.1
VCTree [25]	14.5	18.2	20.4	8.2	10.4	11.2	5.2	6.9	8.3
GCA [29]	16.4	20.1	22.3	9.6	11.2	12.6	8.0	9.3	11.1
BGNN [52]	22.7	30.4	32.9	11.7	14.3	16.2	8.5	10.7	12.6
Motifs [16]	14.7	18.5	20.0	9.1	11.0	11.8	5.9	8.2	9.7
Motifs + **DeC**-	20.3	23.6	26.1	12.5	14.8	16.6	8.2	10.3	12.4
EBM [21]	14.2	18.2	19.7	10.4	12.5	13.5	5.7	7.7	9.1
EBM + **DeC**-	18.1	22.4	24.1	13.6	15.2	16.8	7.2	9.3	10.3
EBM*	19.9	26.7	30.0	13.9	18.2	**20.5**	7.1	9.7	11.6
EBM* + **DeC**-	22.6	29.4	32.7	14.0	17.4	18.6	8.1	11.2	13.4
TDE [19]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
TDE + **DeC**-	**24.1**	**32.6**	**35.2**	**15.0**	**18.3**	19.1	**9.5**	**12.8**	**15.3**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART SGG MODELS WITHOUT OBJECT IDENTITY CUES ON THE METRIC OF MR@K IN THE GQA DATASET. ° MEANS THE RESULTS REPORTED IN [21].

Models	PredCls		SGCls	
	mR@50	mR@100	mR@50	mR@100
IMP$^{\circ}$ [1]	.94	1.32	.50	.65
VCTree† [25]	1.20	1.48	.62	.79
EBM† [21]	1.35	1.58	.75	.88
EBM + **DeC**	1.76	1.94	.93	1.13
Motifs† [16]	1.81	2.75	.81	1.18
Motifs + **DeC**	2.63	3.54	1.17	1.52
Trans† [54]	2.48	3.69	.97	1.29
Trans† + EBM	2.94	4.71	1.32	**1.77**
Trans† + **DeC**	**3.13**	**4.92**	**1.46**	1.73

identity cues, differentiating by the symbol †. Based on the results, when we apply our **DeC** strategy to the three methods, we can observe considerable improvements, especially on TDE, which enjoys the largest performance boost. Moreover, the models with the explicit identities could obviously suffer from a relative performance degradation, e.g. TDE on average gains about ∼20% improvements when removing the identities. This further demonstrates the object identity could pose the degradation of the model's generalizability on the unseen pairs.

G. Ablation Study

We further study the effectiveness of each module in our framework for the three tasks: SGG, FS-SGG and ZS-SGG on the VG dataset. Specifically, we divide our framework into six variants: (1) -COMP, only with the decomposed representation, i.e., without any relation and object composition strategies; (2) -INTER, removing the inter-category relation composition; (3) -INTRA, removing intra-category relation composition; (4) -OBJC, removing object composition; (5) DEC-S, only synthesizing novel relation triples based on the same image, i.e., the candidate object or subject must be from the sampled

TABLE V
The results of few-shot scene graph generation (FS-SGG) on VG. † denotes the results without the explicit object identity, i.e., the Glove embeddings.

	Method	SGDet	SGCls	PredCls
		mR@50/100	mR@50/100	mR@50/100
S=5	IMP [1]	1.8 / 3.5	3.2 / 4.5	6.2 / 7.4
	GCA† [29]	3.1 / 4.0	5.1 / 6.3	8.9 / 10.2
	EBM [21]	2.0 / 3.1	3.3 / 4.5	6.4 / 7.9
	EBM†	2.7 / 3.8	4.0 / 5.3	7.1 / 8.7
	EBM+**DeC**-	3.2 / 4.0	4.5 / 5.8	8.4 / 9.4
	EBM†+**DeC**	3.9 / 4.3	5.2 / 6.1	9.7 / 10.6
	TDE [19]	1.7 / 2.2	3.1 / 4.2	6.6 / 8.1
	TDE†	2.5 / 3.1	3.5 / 4.6	8.5 / 9.4
	TDE+**DeC**-	2.9 / 3.8	5.6 / 6.4	10.3 / 11.6
	TDE†+**DeC**	**3.7** / 4.5	6.3 / **7.7**	**12.3 / 13.8**
	Motifs [16]	1.8 / 2.9	3.7 / 5.0	6.9 / 7.8
	Motifs†	2.2 / 3.0	4.2 / 5.6	7.6 / 8.4
	Motifs+**Dec**-	3.0 / 4.1	5.3 / 6.8	8.2 / 9.3
	Motifs†+**Dec**	3.5 / **4.7**	**6.6** / 7.3	9.4 / 10.3
S=10	IMP	3.2 / 4.4	3.9 / 5.3	8.1 / 9.3
	GCA†	4.1 / 5.1	5.7 / 6.6	11.5 / 12.2
	EBM	3.0 / 3.9	4.2 / 5.5	6.5 / 7.6
	EBM†	3.9 / 4.7	5.3 / 6.1	7.2 / 8.0
	EBM +**DeC**-	4.1 / 5.0	5.6 / 6.3	9.0 / 10.9
	EBM† +**DeC**	**4.5** / 5.2	6.0 / 6.8	10.2 / 13.4
	TDE	2.0 / 2.7	3.7 / 4.6	8.4 / 9.7
	TDE†	2.9 / 3.2	4.3 / 5.0	9.6 / 10.2
	TDE+**DeC**-	3.6 / 4.7	5.2 / 6.4	10.4 / 12.0
	TDE†+DeC	4.4 / **5.3**	**6.5 / 7.2**	**12.7 / 14.1**
	Motifs	2.4 / 3.3	4.0 / 5.1	7.4 / 8.6
	Motifs†	2.8 / 3.6	4.8 / 5.3	8.4 / 9.5
	Motifs+**DeC**-	3.4 / 4.2	5.3 / 6.4	9.5 / 10.7
	Motifs†+**DeC**	4.1 / 4.8	6.1 / 7.0	11.2 / 12.3

image; and (6) DeC-d, only constructing triples based on different images, i.e., the candidate object or subject must derive from different images. We choose TDE as the baseline model and remove identity clues for all models, and the results are shown in Table VII. Note that on the task of FR, we set S=10 for each predicate.

From the results, we could obtain the following observations. For the overall performance, we could see that the improvements from decomposed representation (i.e., CVAE module) are lower than the counterpart from composition learning. First, the CVAE eliminates the reliance on identity cues to predict relation and leads to more effective learned relationship features. However, the decomposition representation only solved the feature learning, but the severe imbalanced relation distribution still exists in the dataset. Thus, we could see CVAE slightly outperforms the baseline models (-Comp vs. TDE), but not significantly. As for our compositional learning, we construct extra ~ 800K relation triple instances, but the total ground truth training relation triples have ~ 400K, and the top 10 frequent predicates account for about 90%. This means we almost enlarged the samples of the tail relationships by 20 times. Therefore, we could see -intra and -inter strategies can bring more improvements.

Besides, our compositions strategies have different task-specific effectiveness on the various SGG tasks. More concretely, on the main task of SGG, the intra-category relation composition shows superiority to the inter-category, because the majority of relations in the test set consist of common relations, and only a small subset of them are unseen by the training model. Thus, the intra-category one plays a more essential role on the SGG task. For FS-SGG, since all relations have only few samples, intra-category can compose extra relation pairs with the same object labels. In contrast, for the ZS-SGG task that simply evaluates the unseen relation pairs, the inter-category relation composition exactly creates novel samples with different subject-object combinations, and shows

TABLE VI
The results of zero-shot scene graph generation (ZS-SGG) on VG. † denotes the results without the explicit object identity, i.e., the Glove embeddings.

Method	SGDet	SGCls	PredCls
	R@50/100	R@50/100	R@50/100
IMP [1]	.73 / 1.2	2.5 / 3.2	14.5 / 16.2
GCA† [29]	1.7 / 2.5	3.8 / 4.4	18.5 / 20.4
EBM [21]	1.0 / 1.8	2.3 / 3.1	8.0 / 12.4
EBM†	1.5 / 2.2	3.1 / 4.2	12.4 / 16.2
EBM+**DeC**-	1.9 / 2.9	3.6 / 4.5	14.6 / 17.3
EBM†+**DeC**	1.8 / **3.6**	4.3 / 5.0	17.3 / 20.5
TDE [19]	.22 / .70	1.9 / 2.6	10.8 / 14.3
TDE†	1.6 / 2.3	2.6 / 3.4	12.5 / 15.6
TDE+**DeC**-	1.9 / 2.6	3.2 / 4.1	15.2 / 17.4
TDE†+**DeC**	**2.4** / 3.1	**4.9 / 5.3**	**18.8 / 22.1**
Motifs [16]	.14 / .27	2.2 / 3.0	10.3 / 14.7
Motifs†	.32 / .61	2.8 / 3.6	13.5 / 17.2
Motifs+**Dec-**	1.5 / 2.4	3.0 / 3.9	14.2 / 18.3
Motifs+**Dec**	2.3 / 3.1	3.2 / 4.1	15.8 / 19.5

TABLE VII
The ablation study on the three tasks: **SGG**, **FS-SGG** and **ZS-SGG**. All model remove the identity clues.

Tasks	Models	SGDet mR@50/100	SGCls mR@50/100	PredCls mR@50/100
SGG	Full	**12.0 / 13.6**	**14.2 / 16.1**	**25.1 / 28.9**
	-Comp	9.7 / 11.3	12.1 / 14.3	22.7 / 25.6
	-inter	11.4 / 13.0	13.2 / 15.5	24.1 / 26.8
	-intra	11.8 / 12.6	12.8 / 13.7	23.4 / 26.2
	-ObjC	11.6 / 12.8	13.7 / 15.6	24.7 / 28.1
	DeC-s	10.2 / 11.8	12.5 / 14.6	22.5 / 25.7
	DeC-d	11.9 / 13.3	13.6 / 15.8	24.7 / 28.5
	TDE	9.3 / 11.1	11.6 / 13.0	20.4 / 23.7
FS-SGG	Full	**4.4 / 5.1**	**6.5 / 7.2**	**12.7 / 14.1**
	-Comp	2.4 / 3.3	3.9 / 5.7	9.5 / 11.4
	-inter	3.6 / 4.7	5.3 / 6.3	11.6 / 13.1
	-intra	3.4 / 4.2	4.3 / 5.9	10.3 / 12.2
	-ObjC	4.0 / 4.8	6.0 / 6.6	12.0 / 13.5
	DeC-s	2.3 / 3.6	4.1 / 5.7	10.0 / 11.8
	DeC-d	4.1 / 4.8	6.3 / 6.5	12.6 / 13.9
	TDE	2.9 / 3.2	4.3 / 5.0	9.6 / 10.2
ZS-SGG	Full	**2.4 / 3.5**	**4.9 / 5.3**	**18.8 / 19.5**
	-Comp	1.7 / 2.8	3.0 / 4.1	15.2 / 16.4
	-inter	2.0 / 3.2	3.3 / 4.5	15.9 / 17.3
	-intra	2.4 / 3.0	4.5 / 5.0	16.4 / 17.8
	-ObjC	2.2 / 3.3	4.1 / 5.3	17.9 / 18.5
	DeC-s	1.8 / 2.8	3.1 / 4.3	15.5 / 16.7
	DeC-d	2.4 / 3.4	4.6 / 5.0	18.6 / 19.1
	TDE	1.6 / 2.3	2.6 / 3.4	12.5 / 15.6

a better effectiveness than intra-category relation composition. In addition, the object composition strategy can enlarge the set of object candidates and further enhance both relation composition strategies.

To evaluate the influence of different relation composition strategies, we further report the results of DEC-D and DEC-S on the three tasks. From the results, we could find that DEC-D relatively outperforms the DEC-S. We deem the main reason is that the number of synthesized relation pairs from the same image is highly less than the counterpart from the different image. Based on our statistics, the number of novel relation triples synthesized from DEC-D are $\sim$20 times more than the counterpart from DEC-S. Thus, DEC-D can more sufficiently train the model than DEC-S. Secondly, two object instances in an image rarely have the same state label, especially rare predicates. This limits DEC-S to construct relation triples for the tail relation and thus hardly contributes to the performance of tail predicate prediction.

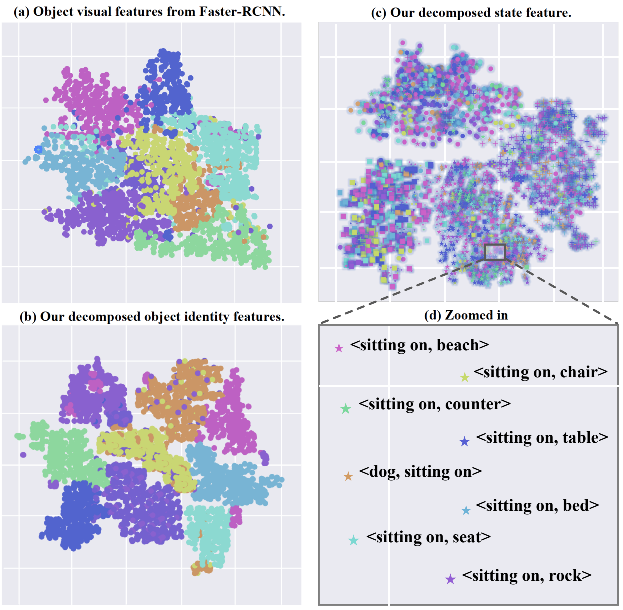


Fig. 6. The visualisation of the learned object features based on TSNE tools.

H. *Qualitative Results*

We further visualize the quality of our decomposed features by t-SNE [55] and the generated scene graphs on the VG.

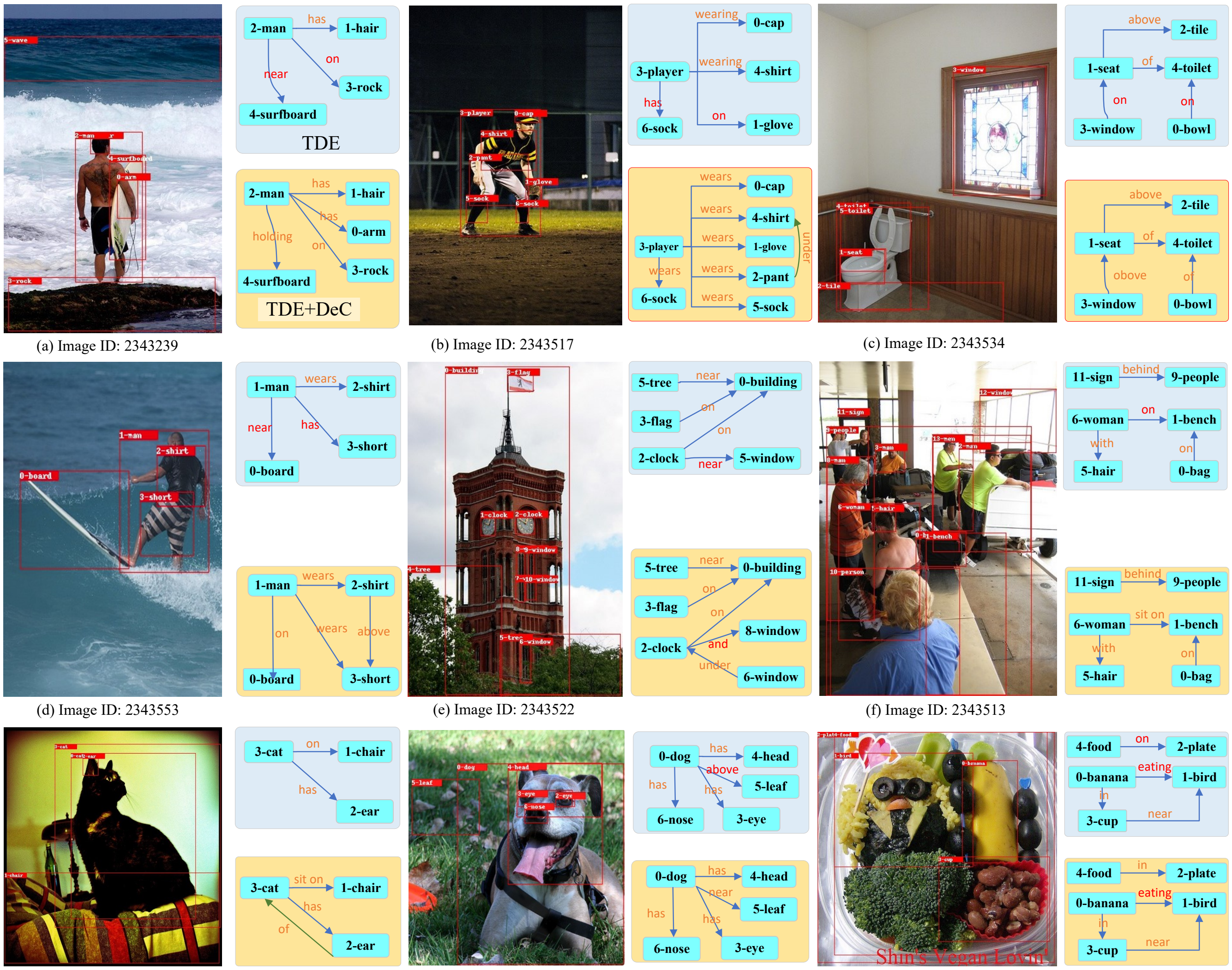
We random select nine object classes with 1,800 data points (each class with 200 data points) to visualize. Figure 6(a) shows the visual feature derived from the object detection network (e.g., Faster-RCNN). Figure 6(b) and Figure 6(c) present our decomposed object identity features and state features, i.e., $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ respectively. Figure 6(d) visualizes four predicate's state features differentiated by four markers. Note that the different colors in Figure 6 denote different object categories. From Figure 6(a), we could observe that the object features from Faster-RCNN still contain object identity clues, because the same-category objects are relatively clustered together. In contrast, our decomposed object identity features (Figure 6(b)) are much more tightly clustered and better separated, possibly because we use the GLOVE word embeddings as the supervision signal in the CVAE. In Figure 6(c), we could see that the learned state features are category-agnostic, that is, object categories do not affect state features. For example, when zoomed in on Figure 6(d), although "beach" and "chair" are very different categories, their state features are close to each other, since they have the same state "sitting on".

Some examples of scene graphs generated by TDE together with our method are shown in Figure 7, confirming that our method can generate more relations compared to TDE alone. This is shown prominently in Figure 7(b), in which our model generates a large number of correct and appropriate triples but TDE is prone to predicting biased predicates. For example, in Figure 7(a,d), TDE prefers to predict a predicate as "near". In comparison, our model can well address this, e.g, in Figure 7(a), our model correctly predicts the relationship between 2-man and 4-surfboard as "holding". Although ''near" also seems to be reasonable from the localization aspect, it does not informatively reflect the semantic meaning as well as "holding". Also, in Figure 7(f,g), TDE consistently predicts the relationship "sit on" as "on". Although, to some extent, the relationship "on" seems correct, it does not capture more high-level semantics. However, our model can accurately predict the fine-grained predicate "sit on", which is more informative and rational.

VI. CONCLUSION

Unbiased scene graph generation (SGG) techniques aim at addressing the long-tail in real-world benchmark datasets such as Visual Genome. In this paper, we first reveal the effect of *object identity information* on causing biased scene graph generation. Counter-intuitively, we observe that even without any visual features, conventional SGG models can still produce competitive results. We deem that this is mainly because of the heavily skewed relation distribution in data. Based on this insight, we propose to learn disentangled representations of visual features, aiming to decompose the raw visual features into two components: *identity* features and *state* features. In particular, state features are category-agnostic and capture relation-specific information in visual features. We further develop two compositional learning strategies on the relation and object levels to alleviate the data-starving issue for rare relations. Putting it all together, our decomposition and composition method (**DeC**) is a model-free technique that can be readily plugged into existing SGG models to improve their performance for unbiased SGG. In extensive experiments that include conventional SGG and two challenging subtasks: FS-SGG and ZS-SGG, **DeC** consistently improves performance of a number of recent strong SGG models, setting new state-of-the-art performance for unbiased SGG.

Although the object identity information could provide strong cues for relation prediction and eventually lead to biased predictions, to some extent, it can also provide essential benefits for SGG. Hence, as a future work, we will consider incorporating it into relation prediction in an unbiased manner, e.g., through an attention fusion mechanism.



(a) Image ID: 2343239

(b) Image ID: 2343517

(c) Image ID: 2343534

(d) Image ID: 2343553

(e) Image ID: 2343522

(f) Image ID: 2343513

(g) Image ID: 2343512

(h) Image ID: 2343539

(i) Image ID: 2343554

Fig. 7. Four example images from the test set on the task of predicate classification. For each image (left), the results of TDE are shown top-right, whilst the bottom-right scene graph is generated by TDE+DeC. The red predicates denote the wrong prediction.

Acknowledgement

This work was partially supported by the DARPA CCU program (HR001121S0024).

References

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

[2] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *ECCV*. Springer, 2016, pp. 852–869.

[3] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *NIPS*, 2017, pp. 2171–2180.

[4] T. He, L. Gao, J. Song, J. Cai, and Y.-F. Li, "Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation," in *IJCAI*, 2020, pp. 587—-593.

[5] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *CVPR*, 2017, pp. 3076–3086.

[6] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in *NIPS*, 2018, pp. 7211–7221.

[7] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[8] K. Ye and A. Kovashka, "Linguistic structures as weak supervision for visual scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8289–8299.

[9] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019, pp. 10 685–10 694.

[10] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *CVPR*, 2017, pp. 1–9.

[11] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *CVPR*, 2015, pp. 2425–2433.

[12] T. He, L. Gao, J. Song, and Y.-F. Li, "Exploiting scene graphs for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 984–15 993.

[13] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5664–5673.

[14] C. Zhang, J. Yu, Y. Song, and W. Cai, "Exploiting edge-oriented reasoning for 3d point-based scene graph analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

2021, pp. 9705–9715.

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[16] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018, pp. 5831–5840.

[17] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *CVPR*, 2019, pp. 6163–6171.

[18] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *CVPR*, 2019, pp. 1969–1978.

[19] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *CVPR*, 2020.

[20] A. Zareian, H. You, Z. Wang, and S.-F. Chang, "Learning visual commonsense for robust scene graph generation," in *ECCV*, 2020.

[21] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, and L. Sigal, "Energy-based learning for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 936–13 945.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[24] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, pp. 3483–3491, 2015.

[25] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6619–6628.

[26] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. G. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] V. S. Chen, P. Varma, R. Krishna, M. Bernstein, C. Re, and L. Fei-Fei, "Scene graph prediction with limited labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2580–2590.

[28] A. Zareian and S. Karaman, "Bridging knowledge graphs to generate scene graphs," in *ECCV*, 2020.

[29] B. Knyazev, H. de Vries, C. Cangea, G. W. Taylor, A. Courville, and E. Belilovsky, "Generative compositional augmentations for scene graph prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 827–15 837.

[30] T. He, L. Gao, J. Song, and Y.-F. Li, "Towards open-vocabulary scene graph generation with prompt-based finetuning," in *ECCV*, 2022.

[31] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "Monet: Unsupervised scene decomposition and representation," *arXiv preprint arXiv:1901.11390*, 2019.

[32] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–251.

[33] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417.

[34] W. Zheng, C. Wang, J. Lu, and J. Zhou, "Deep compositional metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9320–9329.

[35] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, and A. M. Bronstein, "Laso: Label-set operations networks for multi-label few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6548–6557.

[36] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 584–600.

[37] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1981–1990.

[38] A. Sordoni, N. Dziri, H. Schulz, G. Gordon, P. Bachman, and R. T. Des Combes, "Decomposed mutual information estimation for contrastive representation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9859–9869.

[39] H. Bai, R. Sun, L. Hong, F. Zhou, N. Ye, H.-J. Ye, S.-H. G. Chan, and Z. Li, "Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation," in *arXiv preprint arXiv:2012.09382*, 2021.

[40] C. Jing, Y. Wu, X. Zhang, Y. Jia, and Q. Wu, "Overcoming language priors in vqa via decomposed linguistic representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 181–11 188.

[41] Y. Wang and T. Derr, "Tree decomposed graph neural network," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2040–2049.

[42] L. Vilnis and A. McCallum, "Word representations via gaussian embedding," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[43] C. Qian, F. Feng, L. Wen, and T.-S. Chua, "Conceptualized and contextualized gaussian embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 683–13 691.

[44] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *ECCV*, 2018, pp. 670–685.

[45] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5532–5540.

[46] X. Lin, C. Ding, J. Zeng, and D. Tao, "Gps-net: Graph property sensing network for scene graph generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[47] Z. Chen, Y. Fu, Y.-X. Wang, L. Ma, W. Liu, and M. Hebert, "Image deformation meta-networks for one-shot learning," in *CVPR*, June 2019.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[49] A. Bojchevski and S. Günnemann, "Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking," in *ICLR*, 2018.

[50] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[51] T.-J. J. Wang, S. Pehlivan, and J. Laaksonen, "Tackling the unannotated: Scene graph generation with bias-reduced models," in *BMVC*, 2020.

[52] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 109–11 119.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[55] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.