Fine-Grained is Too Coarse: A Novel Data-Centric Approach for Efficient Scene Graph Generation

Neau Maëlic, Paulo E. Santos, Anne-Gwenn Bosser, Cédric Buche

Fine-Grained is Too Coarse: A Novel Data-Centric Approach for Efficient Scene Graph Generation

Maëlic Neau[1,2] Paulo E. Santos[1] Anne-Gwenn Bosser[2] Cédric Buche[2]
[1]College of Science and Engineering, Flinders University, Australia
[2]Ecole Nationale d'Ingénieurs de Brest, France
{neau, buche, bosser}@enib.fr, paulo.santos@flinders.edu.au

Abstract

Learning to compose visual relationships from raw images in the form of scene graphs is a highly challenging task due to contextual dependencies, but it is essential in computer vision applications that depend on scene understanding. However, no current approaches in Scene Graph Generation (SGG) aim at providing useful graphs for downstream tasks. Instead, the main focus has primarily been on the task of unbiasing the data distribution for predicting more fine-grained relations. That being said, all fine-grained relations are not equally relevant and at least a part of them are of no use for real-world applications. In this work, we introduce the task of Efficient SGG that prioritizes the generation of relevant relations, facilitating the use of Scene Graphs in downstream tasks such as Image Generation. To support further approaches, we present a new dataset, VG150-curated, based on the annotations of the popular Visual Genome dataset. We show through a set of experiments that this dataset contains more high-quality and diverse annotations than the one usually use in SGG. Finally, we show the efficiency of this dataset in the task of Image Generation from Scene Graphs[1].

1. Introduction

The task of Scene Graph Generation (SGG) aims at creating a symbolic representation of a scene by inferring relations between entities as a graph structure. Typically, approaches in SGG rely on detecting object features from an image and then inferring relation predicates between object pairs as $< subject, predicate, object >$ triplets. Connections between pairs of triplets form a directed acyclic graph in which each vertex refers to an object and its associated image region. Due to its efficient representation capacity, this task holds strong promises for other downstream tasks such as Image Captioning [38, 33, 40] or Visual Question Answering [7, 16]. Recent contributions to the field highlight an opportunity for SGG to support the reasoning of an embodied agent by leveraging both the spatial and semantic latent context of a scene in a single representation [4, 21, 3]. However, despite a vast amount of work in the last few years, the performance of the best approaches is far from optimal, and the usage in downstream tasks is limited [50]. A set of problems have been raised by the community to explain this slow pace, the main one is the long-tail distribution of predicates [31, 43, 19]. In fact, due to annotation biases, datasets used in SGG tend to have more annotated samples with vague predicates (e.g. `on`, `has` or `near`) rather than with fine-grained ones (e.g. `riding`, `under` or `eating`). While this issue has been largely investigated under the name of **Unbiased SGG** [41, 44, 9, 49, 42, 10, 30, 19], other aspects of the task have been left aside, such as the amount of actual useful information conveyed by a scene graph structure. Inspired by recent approaches in this direction [35], we introduce the task of **Efficient SGG** that aims at extracting the maximum quantity of *valuable information* from an input scene, in contrast to current approaches that focus on extracting *fine-grained information* first. This new approach is highly beneficial to downstream tasks where predicting major events from the

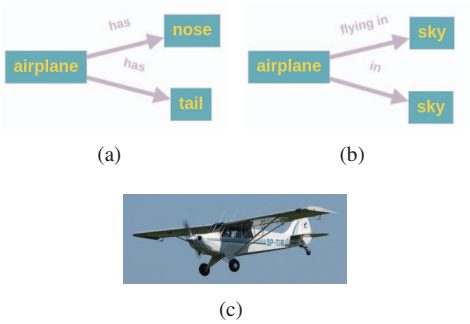


Figure 1. Top-2 relations predictions for Figure 1(c): 1(a) is using original dataset and 1(b) is using our curated dataset. Note that more relevant relations are obtained in the latter.

[1]Source code: `https://github.com/Maelic/VG_curated`

scene is more important than predicting detailed but minor ones (see a comparison in Figure 1).

To support efficient learning for this task, we focus on providing a novel high-quality dataset by leveraging the existing but noisy annotations from the largest and main used dataset in SGG, Visual Genome (VG) [15]. In contrast to other curation approaches of VG [22, 35, 26], we focus on preserving the semantics conveyed by Scene Graph structures during pruning while annotations that are irrelevant for downstream tasks are removed, thus creating an optimised dataset for the task of Efficient SGG. To demonstrate the necessity of our approach, we show that SGG models trained on the current version of Visual Genome are inefficient in downstream tasks: first, they are biased toward predicting irrelevant relations with overconfidence. Second, the poor connectivity in annotated samples penalises the learning process resulting in low-quality graphs. Our approach tackles these two problems, resulting in a new high-quality dataset for the task that improves the performance of baseline models by a strong margin. We further evaluate the use of this new dataset in the task of Image Generation.

The main contributions of this paper are threefold: **(i)** a study on the impact of sample connectivity and irrelevant relations in the learning process of baseline models in SGG ; **(ii)** a new curation process of the Visual Genome dataset that results in VG150-curated, a new high-quality dataset for SGG ; **(iii)** a study on the benefits of VG150-curated in SGG and downstream tasks.

2. Related Work

Since the first description of the task [37], SGG has drawn widespread attention in the computer vision and natural language processing communities. Popular approaches combine object detection backbones such as the popular Faster-RCNN [28] with a graph generation model in a two-stage pipeline [37, 47, 32, 19, 23, 10, 34]. However, concerns about biases in large-scale datasets such as Visual Genome [15] have been rapidly raised, resulting in the task of Unbiased SGG [36, 31, 42, 39, 43, 9, 8]. The idea is to improve predicate prediction using different model-agnostic techniques and training strategies such as the Total Direct Effect [31] or probability distribution [20] and evaluate them on a set of baseline models [45, 32, 37]. On the other hand, new approaches [5, 18, 25, 24] are considering the task in a one-stage fashion, learning relationships from image features directly. Still, these approaches are assuming that all relations are equal and share the same amount of information on the scene in the learning process. This creates models that extensively predict meaningless relations with high confidence, hindering the performance of downstream tasks that depend on relevant predictions. A first step towards solving this issue is to enforce *relevant* annotations in data-centric approaches [17, 46].

VG Split	#Images	#Objects	#Predicates
VG80K [48]	104,832	53,304	29,086
VG150 [37]	105,414	150	50
VrR-VG [22]	58,983	1,321	117

Table 1. Comparison of the main splits of Visual Genome used in SGG. #Images is the total number of annotated images, #Objects and #Predicates count the number of classes.

2.1. Data-Centric Approaches in SGG

To the best of our knowledge, only a few current approaches are considering the Visual Genome dataset biases from a data-centric perspective. In VrR-VG [22], the authors based their assumption on the fact that relations that can be easily inferred with only spatial information from an object pair (i.e. bounding box coordinates) are not visually relevant. This results in the removal of common relations and leads to sparse annotations where only rare and very specific relations are annotated for which the use in downstream tasks is very limited. Other approaches are focusing on balancing the predicate distribution [26] or filtering similar or vague predicates [2] to improve the relevance of the annotations. However, these methods assume a consistent use of the same predicate across the annotations which is not true due to the inherent *semantic ambiguity* of natural language [46]. Thus, we believe that curating the annotations based on the predicate distribution alone is not a viable strategy. Intuitively, taking into account the triplets' distribution seems to alleviate this semantic ambiguity and could be a more beneficial strategy.

Recent work has focused on re-sampling and de-noising the Visual Genome dataset. Different techniques were used to improve the quality of annotations such as internal and external transfer [46] or clustering strategies [17]. The reported results showed a non-negligible impact on the training of baseline models for SGG (by up to 25.2% in certain cases [17]). In fact, looking at those results [46], it is possible to conclude that at this stage cleaning the dataset is more beneficial for the task than implementing new models. Finally, research reported in [47], [6], [1], and [37] are splitting the original annotations based on different object and predicate classes frequency. These approaches were not filtering relationships to explicitly address inherent biases, here the data split addressed only minor annotation issues such as object class de-noising [47]. A comparison of the different splits of Visual Genome reported in SGG papers is available in Table 1. In contrast to prior work, the present paper focuses on building a data split that encompasses only visually-relevant annotations to support usage in downstream tasks. To do so, we introduce a new definition of visually-relevant relation based on the following assumption: *a relation is not relevant if it describes a composition between parts of an entity that is true in a general*

sense and that could be inferred using external knowledge (e.g. $< man, has, arm >$*).*

This definition of relevant relations is not related to the nature of the dataset but models instead the semantic information conveyed by a scene graph structure. Moreover, we also consider the issue of connectivity of the annotated samples which has never been addressed before. The next section details these issues, while also arguing why their solution causes a positive impact on the downstream tasks.

3. Problem Definition

The Visual Genome dataset [15] is the largest and most widely adopted dataset for SGG. Its annotations have been collected by annotators in the form of region captions. Then, different parsing techniques have been applied to retrieve $< subject, predicate, object >$ triplets for each region. Because annotators were not constrained to use any particular vocabulary, this process resulted in more than 53K object classes and 27K predicate classes, where more than 50% of them only have one sample. This split of the data is usually referred to as VG80K [48]. To support efficient learning for SGG, the common practice is to prune annotations from VG80K, keeping only a selection of the top-k predicate and object classes. However, when doing so, no current approaches are aiming at preserving the graph structure as well as keeping the relevant information about the scene. In this work, we present an approach that is able to extract annotations for k object and predicate classes while ensuring to preserve most of the information from the original samples. We believe that this could be achieved by **(1)** preserving the connectivity of the original graph and **(2)** extensively pruning irrelevant annotations.

3.1. Connectivity

This work uses the following notation: a graph $G = (V, E)$ represents all relations in a given image with a set of edges E and a set of vertices V. It is important to notice here that G is not necessarily fully connected, as some vertices or edges could have been removed from the original annotations. We denote the average graph size on a set of n graphs as: $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} |E(G_i)|$.

The average graph size in the original annotations of Visual Genome is high, with $\bar{s} = 19.02$. However, when pruning the dataset to keep only the top-k object and predicate classes, a large number of annotations were removed leading to $\bar{s} = 6.98$ in the VG150 split [37]. Figure 2 shows the number of relations with respect to the number of images in VG150, where we can see that the distribution of graph size is long-tailed with more than 28% of samples that only contain one relation. The average vertex degree $d(v)$ is also low, with an average of 2.02 against 2.34 for VG80K. These figures can easily be explained by the applied pruning strategy which selected object and predicate classes based only on their overall frequency over the dataset. We believe that this negatively affects the performance of SGG approaches, especially methods that explicitly model the context of every relation using, for instance, Iterative Message Passing [37] or bipartite matching [19]. This is solved in this work by selecting annotated samples based on their inter-connectivity rather than overall frequency.

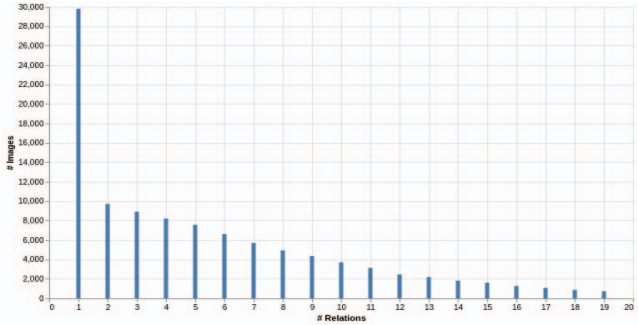


Figure 2. Graph size $\bar{s}$ per image in VG150.

3.2. Irrelevant Relationships

Besides a connectivity issue, the annotations of Visual Genome are also biased with the over-representation of certain triplets. In fact, we observed that some invariant relations (such as $< man, has, head >$) are over-represented in the dataset, creating a bias for models that will always select those relations with over-confidence compared to others. Because the current evaluation metrics Recall@k and meanRecall@k [32] are ranking metrics, this leads to poor performance of the task. Previous work [45] enumerated 3 different relation categories in Visual Genome, as follows: geometric, possessive, and semantic (i.e. actions or activities). The geometric category represents spatial relations such as $< cup, on, table >$; possessive relations are composed of an entity and an artifact such as $< car, has, wheel >$; finally, semantic relations represent activities or actions such as $< man, riding, bike >$. In this context, we categorised the top 50 triplets to see if we can experience a certain pattern in over-represented relations. We followed the same classes as in [45] except that we made a distinction between invariant possessive relation (i.e. *part-whole*) and possessive attributes such as clothing (denoted as the *possessive* category here). Figure 3 shows the number of relations with respect to the number of images in VG150, where we can see that part-whole relations were prevalent with 55.1% of the total number of occurrences for the top 25 more represented relations. As explained in [22, 50, 35], these relations may be biasing the learning process because they are true in the general sense and do not depend on visual features of the scene. We call this *invariant relationship bias*. In order to verify this assumption, we conducted an experiment on predictions obtained by the Motifs-TDE

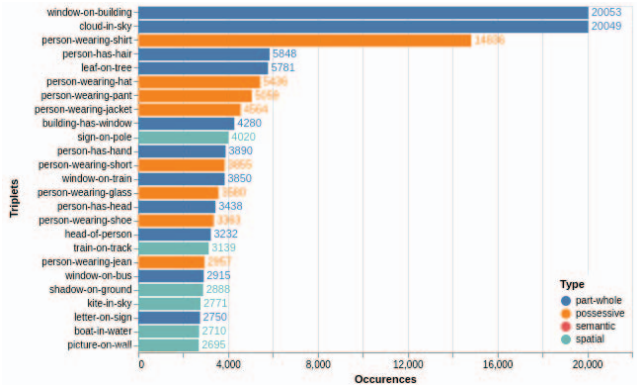


Figure 3. Distribution of the top 25 relations in VG150.

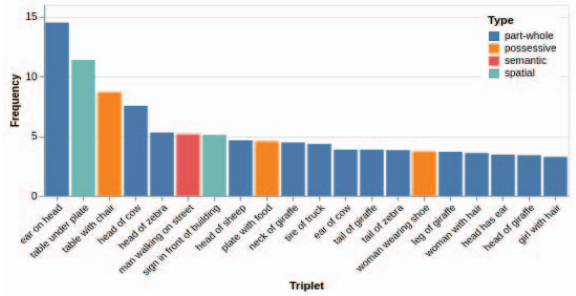


Figure 4. Ratio of predicted triplets over ground truth ones on the test set of VG150. For clarity, we show only the top 20 triplets with more than 20 occurrences. The top-1 triplet is predicted 14,5 more times than the actual ground truth.

model [31] on the test set of the VG150 dataset. Results are shown in Figure 4 where we can see that part-whole relations are overly predicted in comparison to the ground truth annotations. This issue matters in regard to the importance of each relation in the global context of the scene. For instance, it is questionable how possessive relations such as $< man, has, head >$ are actually relevant to describe the scene, especially for usage in downstream tasks such as Visual-Question Answering where the amount of noise in the predicted visual relations is critical. In the next section, we detail our approach to solving these two issues by introducing two novel curation methods.

4. Data Curation

We started with the original version of the Visual Genome dataset that is pre-processed to clean the annotations as described in the sequence. For the object regions, we replicated the approach proposed by [37] to merge bounding boxes with an Intersection over Union (IoU) greater than or equal to 0.9. For the textual annotations, we also followed [37] to remove stop-words and punctuation using the alias dictionaries provided by the authors of the dataset [2]. Finally, we merged synonyms of object classes using WordNet synsets. This process resulted in the **VG80K** version of the dataset that contains 104,832 images annotated and that is similar to the one introduced in [48]. In Section 4.1, we introduce a simple algorithm to improve the number of connected regions. To address the issues of relevance of relations, we focused on categorising and removing irrelevant relations, as detailed in Section 4.2.

4.1. Connectivity

Finding the most connected object ($\hat{\mathbf{o}}$) and predicate ($\hat{\mathbf{p}}$) classes for a set of n graphs can be represented as:

$$\theta(\hat{\mathbf{o}}, \hat{\mathbf{p}}) = \max_{\hat{\mathbf{o}}, \hat{\mathbf{p}}} \sum_{k=1}^{n} |G(u, v, w)|, w \in \hat{\mathbf{p}} \vee [u, v] \in \hat{\mathbf{o}} \quad (1)$$

[2] http://visualgenome.org/api/v0/api_home.html

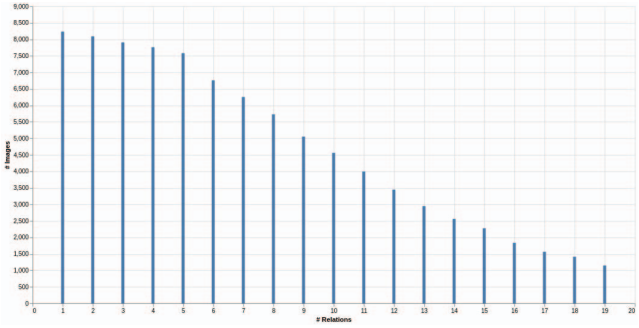


Figure 5. Graph size per number of images in VG150-con.

To be consistent with VG150, we chose $|\hat{\mathbf{o}}| = 150$ and $|\hat{\mathbf{p}}| = 50$. As this is a complex optimisation problem, a satisfying solution can be found by first optimising $\theta(\hat{\mathbf{o}})$ and then $\theta(\hat{\mathbf{o}}, \hat{\mathbf{p}})$ with a fixed set of classes $\hat{\mathbf{o}}$. We applied this method to the original data and obtained a new split that we call VG150-con. This split possesses a significantly higher number of relations (22% more than VG150), with an average graph size $\bar{s}$ of 8.37 versus 6.98 for VG150, see Table 3. Figure 5 displays the top-20 distribution of graph size per image. By comparing this distribution with the original one in Figure 2, we see a clear improvement of the long-tail distribution, proving that our method results in a more connected dataset than the original. More interestingly, we see a net improvement in the average vertex degree (see Table 3), moving up from 2.02 to 2.2. This shows that relations are also more interdependent and thus should benefit the context learning of SGG models. We further analysed the performance of SGG models on this new split of the data, as presented in Section 5.

4.2. Irrelevant Relationships

Building upon our new classification of relevant relations, we employed a new approach to filter the dataset from *part-whole* triplets. To filter categories of relations, previous approaches rely on handcrafted predicate categories [45]. However, this categorisation only takes into account

the semantics of predicates, which suppose that annotations are consistent in the dataset. This assumption is wrong, given the semantic ambiguity introduced in [46]. We give a clear example given the following two triplets from Visual Genome:

$$man \xrightarrow{\text{has}} nose \tag{2}$$

$$man \xrightarrow{\text{has}} surfboard \tag{3}$$

When we look at image samples that contain these relations, we see that Formula 2 refers to a *part-whole* relation that would be described as *nose is part of man* while in Formula 3, the relation is *semantic* and could be described as *man is carrying surfboard*, even though they share the same predicate *has*. On the other hand, it has been noticed that there is a strong correspondence between the knowledge embedded in the Visual Genome annotations and in linguistic commonsense knowledge sources such as ConceptNet [10, 14]. Thus, instead of manually labeling every triplet in VG, we chose to compare triplets annotations with a subset of ConceptNet [29] that contains only part-whole relations. If a relation has a significant similarity with one from ConceptNet, then we can filter all its occurrences from the original data. We used the *part-whole* subset of ConceptNet, following the ontology introduced in [12] with the relations 'PartOf', 'HasA', and 'MadeOf'. Then, we used different approaches to categorize relations between part-whole and non-part-whole from textual annotations only. To evaluate the performance of this filtering, we manually annotated a subset of 1000 random relations from Visual Genome. First, we evaluated the filtering using lexical similarity between $< subject, object >$ pair in Visual Genome and ConceptNet. Second, we compared the average of $< subject, predicate, object >$ Glove embeddings with those from ConceptNet using the cosine similarity. Third, we used different pre-trained Sentence Transformers [27] models to generate sentence similarity embeddings. Finally, we compared those approaches with the predicate-only classification proposed by [45] in which 50 predicate types were classified within semantic, geometric, and possessive classes. Results displayed in Table 2 show that the classification by [45] resulted in the lowest score, this was due to the inconsistency in predicate annotations, as explained before. Sentence-Transformers approaches, as they have been pre-trained on a large corpus of texts, are able to easily generalized and give the best performance. In the choice of embeddings, we prioritized precision over recall as we do not want to discard anything else than *part-whole* relations. The *all-mpnet-base-v2* [3] model has shown the best performance in the task, giving satisfactory trade-off between precision and F1 score. This result is consistent with previous work as this model is ranked 5 in the task of Sentence Similarity [4].

Method	**Recall**	**Precision**	**F1**
Predicate only [45]	0.43	0.62	0.51
Lexical similarity	0.81	0.53	0.64
Glove 6B 300d *(cos=0.7)*	**0.88**	0.5	0.64
RoBERTa-large-v1 *(cos=0.7)*	0.75	0.58	0.66
MiniLM-L6-v2 *(cos=0.7)*	0.74	0.67	0.7
MpNet-base-v2 *(cos=0.75)*	0.64	**0.83**	**0.72**

Table 2. Part-whole relations filtering by comparing with ConceptNet, evaluation on a set of 1000 random samples.

Using the embeddings produced by *all-mpnet-base-v2*, we were able to extract 36,777 part-whole relations for a total of 416,318 occurrences in VG80K (18% of the annotations). Before removing those annotations from the original samples, we ensured that no other types of relationships were dependent on them. This step is important because by removing some part-whole relations we could lose some of the semantics of the scene. For instance, the sub-graph:

$$person \xrightarrow{\text{has}} hand \xrightarrow{\text{holding}} cup \tag{4}$$

describes a semantic relation between the entity `person` and `cup`, even if the relation $< person, has, hand >$ is classified as a part-whole relation by our method. In this case, the method proposed in this work can be applied as follows: we added a set of weights $w : E \rightarrow \mathbb{R}$ to the original graph $G = (V, E)$ such as $w = 1$ if the edge is a part-whole relation and 0 otherwise. Given this graph, we performed a pruning strategy that iterates through all edges and removed those that were only dependent on other part-whole relations. This removed from the graph relations deemed as irrelevant to the context of the scene.

Finally, we leverage the strategy employed in Section 4.1 to select the most connected object and predicate classes from this new filtered data. This process resulted in a data split with 636,175 filtered relationships, that we call VG150-curated (or VG150-cur). In Table 3 we see that this split possesses fewer samples than VG150-connected. This is the case because we noticed that in the original dataset, part-whole relations are highly connected with each other, i.e. a relation like $< person, has, head >$ will often be associated with $< head, has, hair >$. While other types of relations are more context-dependent, it is harder to find a set of 150 object and 50 relation classes that are highly connected. However, from Table 3 we see that VG150-curated still possesses a higher average vertex degree than VG150, proving that our method is efficient to select inter-dependent relations. VG150-curated also possesses a significantly higher number of triplets, showing that the re-

[3] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[4] `https://huggingface.co/spaces/mteb/leaderboard` *accessed on the 21/11/2022.*

Datasets	Statistics			
	$\bar{d}(v)$	$\bar{s}(G)$	#Rels	#Triplets
VG80K	2.34	19.02	2,316,063	514,526
VG150	2.02	6.98	622,705	35,412
VG150-con	2.20	8.38	799,412	44,851
VG150-cur	2.12	7.14	636,175	41,164

Table 3. Graph's connectivity and size of the different splits; where $\bar{d}(v)$ represents the average vertex degree; $\bar{s}(G)$ the average graph size; #Rels is the total number of relations samples, and #Triplets is the number of different triplets.

lationships represented are more diverse. Without invariant relations, this new split represents a more informative and natural description of scenes. The performance of SGG models with this new split is analysed in the next section.

5. Experimental Setup and Results

To demonstrate that the VG150 benchmark is biased and does not correctly evaluate the performance of approaches in the task, we conducted experiments with baseline SGG models. We follow previous work in the area [37, 45, 32, 31] by evaluating our approach on three distinct (but related) tasks, namely Predicate Classification *PredCls*, Scene Graph Classification *SGCls*, and Scene Graph Generation *SGGen*. *PredCls* concentrates on predicting a relation, given the bounding boxes and $< subject, object >$ pairs. *SGCls* is analogous to *PredCls*, except that $< subject, object >$ pairs are not known *a priori* and they need to be inferred by the model. Finally, *SGGen* assumes no prior knowledge; thus, the task included the prediction of object regions, pairs, and relations. To be consistent with other related work, a selection of the most used baseline models were trained: IMP [37], Motifs [45], and VCTree [32]. For Motifs and VCTree, we trained the TDE version introduced in [31]. As other metrics have proven to be ineffective to measure the performance for both the head and tail classes [32], we used the meanRecall@K metric introduced in [32]. We trained the models on two distinct datasets: **(1)** a highly connected version of Visual Genome, VG150-con, where the goal was to evaluate the impact on the performance of the models with this highly-connected dataset; **(2)** the curated version of VG150 (proposed in this paper) where we removed all part-whole relations, as described in Section 4.2, we call it VG150-cur. This last version represents a highly-connected data split with visually relevant annotations.

We retrained every model using the code provided by the authors [31][5], whereby the original parameters were maintained, except for the batch size and learning rate that were fit to our hardware requirements. The Faster-RCNN backbone was trained using the same configuration as [31], and we ensured that the mAp values were similar to that reported in the original paper, in order to guarantee a fair comparison in the SGGen task (respectively 0.24 and 0.27 for VG150-con and VG150-cur, whereas the original Faster-RCNN trained on VG150 has a mAp of 0.28). The training was conducted with a batch size of 32 and a base learning rate of 0.02 on one Nvidia RTX3090 within 20000 iterations (approximately 10 epochs) or 30000 iterations for SGGen. IMP was retrained on the baseline split (VG150) with the above settings, this is why the reported results in Table 4 are slightly different from those reported in the original paper [31]. For comparison, the same training/validation/test split of the original VG150 was used for all datasets.

[5] https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

5.1. Quantitative Results

The results obtained with the experiments conducted in this work are listed in Table 4, where we can see that there was an improvement using VG150-con on the different baseline models (cf. the 6^{th} column of Table 4: *Improv.*). Neural Motifs and VCTree were the two models that benefited the most from the higher connectivity of the dataset for all the tasks. When we compared the statistics on the different data splits in Table 3, we noticed an improvement of 28.3% in the number of the relation samples (train and test splits combined) and 9% in the average vertex degree. This surely benefited the context learning of both Motifs and VCTree. The strong gap in performance between VG150 and VG150-con in PredCls also shows that context learning of baseline models is currently under-exploited. Regarding the performance of the different models trained with VG150-curated, we observed a net improvement in the different tasks in contrast to the results obtained with VG150 and VG150-connected. In particular, there was an improvement of up to 39% for VCTree-TDE model. We believe that the Total Direct Effect (TDE) strategy [31] highly benefited from the removal of irrelevant relationships as these are *invariant* and, thus, biasing the reasoning process employed by the TDE. More concerning, we observe that Motifs-TDE performs best on average on the bias dataset VG150 but worse than VCTree-TDE on our non-bias version VG150-cur. This could impact the current ranking of SGG models on unbiased benchmarks such as VG150-cur. Finally, we also noticed that VG150-cur is very similar in size to VG150 (see Table 3), however, VG150-cur possesses significantly more different triplets, and is thus more challenging to learn from. This result shows that, due to the removal of the invariant relations, the models were not biased into predicting invariant relations with over-confidence, improving by a high margin the meanRecall@K performance.

5.2. Qualitative Results

In Figure 6 we compared predictions from the Motifs-TDE model [31] on the test set of the original VG150 and

Models	Dataset	PredCls mR@20/50/100	SGCls mR@20/50/100	SGGen mR@20/50/100	Improv. (avg.)
IMP [37]	VG150 *	8.8/10.80/11.62	4.63/5.82/6.42	2.76/4.02/5.0	-
	VG150-con	8.8/11.9/13.35	5.63/6.76/7.16	2.59/4.26/5.61	↑ **10.3%**
	VG150-cur	9.61/12.61/13.92	6.99/8.74/9.44	4.09/6.21/7.41	↑ **28%**
Motifs-TDE [31]	VG150	18.5/25.5/29.1	9.8/13.1/14.9	5.8/8.2/9.8	-
	VG150-con	20.38/28.76/34.06	10.3/14.6/17.25	8.15/11.53/13.15	↑ **16.1%**
	VG150-cur	21.38/30.90/36.58	13.75/18.55/21.54	10.49/14.28/17.10	↑ **37%**
VCTree-TDE [31]	VG150	18.4/25.4/28.7	8.9/12.2/14.0	6.9/9.3/11.1	-
	VG150-con	22.5/31.22/37.02	9.38/13.32/15.29	8.56/10.84/13.09	↑ **19.5 %**
	VG150-cur	22.03/32.25/38.24	13.73/18.14/20.70	10.89/14.52/17.09	↑ **39%**

Table 4. Reported performance of baseline models on different datasets, * denotes results reproduced using code by the authors. Improvements are the relative average against the baseline VG150.

our curated version VG150-cur. On all images, we displayed the top 5 predicted relations. On Figures 6(a) to 6(c) we can see that the main element of the image (planes, bears or a bus) are described through their internal components (`wing`, `tail` for the planes; `ear`, `eye` for the bears and `wheel`, `windshield` for the bus) whereas their interconnections with other elements from the image are missing. In the predictions made by training on VG150-curated, see Figures 6(d) to 6(f), we can see that interactions with others elements (`sky`, `wall` and `road`, respectively) are present, giving more information about the scene. This example illustrates the problem with Visual Genome annotations and the bias in the learning process of SGG models: even if all predictions given by models trained on VG150 are correct, they are failing to provide useful information for the downstream tasks. The next section presents experiments on the task of Image Generation to illustrate this point.

5.3. Evaluation

As highlighted in [31], the tasks of Visual Question Answering and Image Captioning rely on particular settings and external datasets with their own acknowledged biases. To remove those biases and outline the full potential of our approach, we chose to evaluate the quality of VG150-curated on the task of Image Generation from Scene Graphs [13]. In contrast to Image Captioning or VQA, Image Generation models can be trained directly from the raw dataset, without inputs of captions or question-answer pairs that could bias the evaluation. Thus, we used a straightforward approach by retraining the popular Image Generation benchmark sg2im[6] [13] with VG150-cur and VG150. We also compared it to the version of Visual Genome used by the original authors that possesses 178 object and 45 predicate classes [13]. We trained the model for 300,000 iterations with a batch size of 64 on one Nvidia RTX3090 GPU with a target image size of 128/128 pixels. To evaluate images generated with the different datasets, we use the

[6] `https://github.com/google/sg2im`

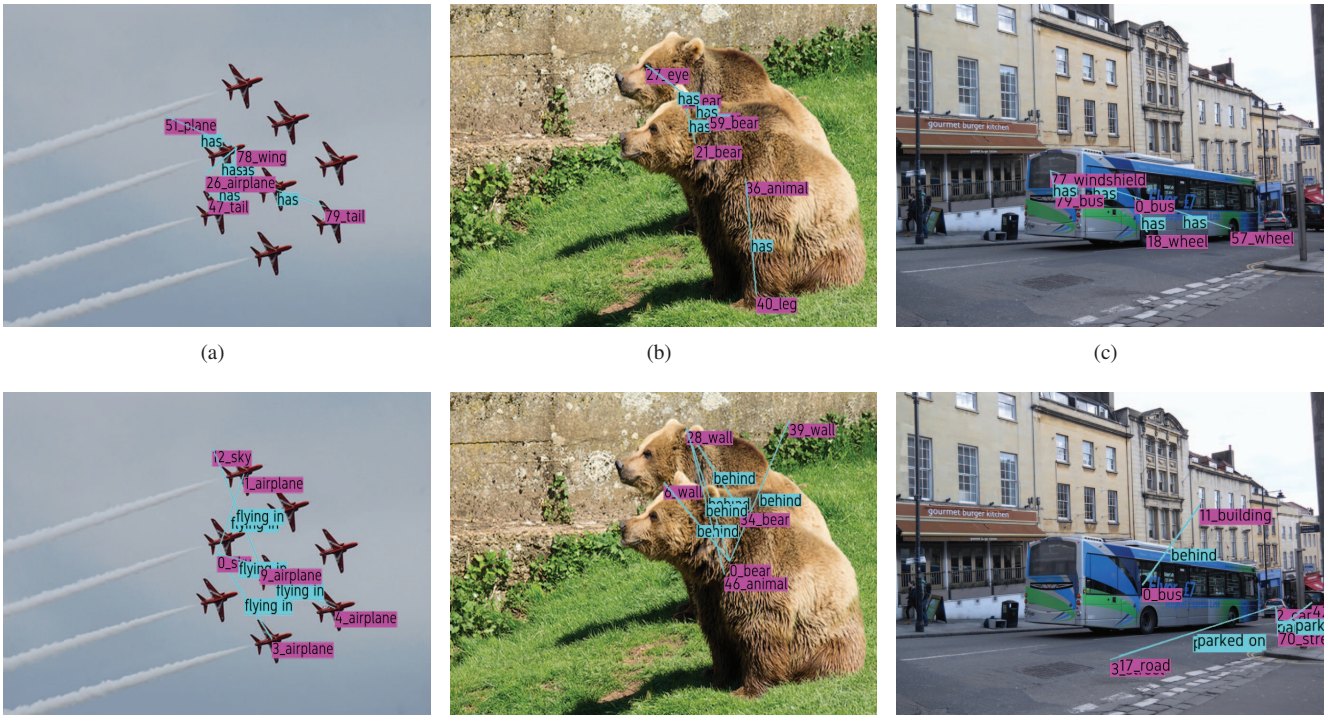
Dataset	FID Score ↓
VG [13]	143.2
VG150	115.2
VG150-curated	96.8

Table 5. Results on the Image Generation task using the Fréchet Inception Distance (FID), lower is better.

Fréchet Inception Distance (FID) [11]. In our case, this metric is evaluating the distance between the distribution of the ground truth images from the VG dataset and the one generated using the different graph annotations from VG150 and VG150-cur. We found out that this is the best metric to evaluate the quality of annotations because the more informative elements there are on the input graphs, the closest to the original image the generation should be. Table 5 shows our results obtained by retraining the model on the different datasets. We first observed that VG150 outperformed the Visual Genome split employed by the original authors by a strong margin. This is mainly due to the cleaning process of VG150 which is more elaborate than VG [13] (as described in Section 4). Then, we also noticed a strong improvement by using VG150-curated, this shows the clear benefit of our curation method for downstream tasks. In Figure 7 we display a few generated samples, from where it is worth noting that the image generated using VG150 annotations (7(b)) is far from the target (7(a)), whereas the version generated with the curated dataset proposed in this work (7(c)) has more similar patterns with respect to 7(a).

6. Discussion

The first work that used the Visual Genome dataset proposed a simple curation algorithm that selects relations only based on their overall frequency in the dataset [37]. Since then, this split has been used in the literature with no consideration of the type and interest of the annotation samples it contains. In this work, we showed that this approach was misleading and that Visual Genome contains more biases than the long-tail distribution of predicates [31]. We



(a) (b) (c)

(d) (e) (f)

Figure 6. Top-5 predictions of Motifs-TDE [31] on the test set of the original VG150 dataset (top) and our curated version VG150-cur (bottom). Pink labels represent objects, and blue labels represent predicates. Best viewed in colour.

(a) (b) (c)

Figure 7. Images generated using sg2im [13]. Left: ground-truth image downsample in 128/128, middle: generation using graphs from VG150, right: generation with VG150-curated.

demonstrated that the over-representation of irrelevant relations in the training data leads the baseline models to predict useless relations with high confidence. By removing these relations, by the proposed new curation process, we observed an improvement of up to 39% meanRecall@K on SGG baseline models, even if our new dataset is more diverse than VG150 (see Table 3). This work also showed the limit of the evaluation metric commonly used in the literature: as meanRecall@K is a ranking metric, its performance results can be easily biased by the prediction of invariant relations. These relations are also irrelevant for downstream tasks. The method proposed in this work was evaluated in the task of Image Generation where, with the use of our curated dataset, a better performance was achieved when compared to training the models on VG150. Finally, comparing SGG models on Image Generation has shown to be a more reliable way of comparing results, than the usual Recall or meanRecall metrics that are highly biased by the nature of the ground-truth annotations.

7. Conclusion

In this work, we analysed two biases in the data distribution and annotations of the Visual Genome dataset. We then proposed two novel techniques to alleviate those biases, resulting in two new splits of the data, called VG150-connected and VG150-curated. These splits, in particular VG150-curated, facilitated an improvement in the SGG task using traditional evaluation metrics; thus, providing a fair comparison with respect to the most used dataset in the literature, VG150. We then analysed the obtained results qualitatively and quantitatively and demonstrated the correlation between higher-quality annotations and better representation learning. We hope this last point will help future approaches to obtain higher-quality datasets for the task. Future work will consider leveraging the principles of Efficient SGG for commonsense reasoning in embodied (robotics) agents. This shall include the use of spatial, semantic, and possessive relations for Visual Understanding in Human-Agent Open-Ended Interaction.

References

[1] Sherif Abdelkarim et al. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021.

[2] David Abou Chacra and John Zelek. The topology and language of relationships in the visual genome dataset. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4859–4867. IEEE, 2022.

[3] Saeid Amiri, Kishan Chandan, and Shiqi Zhang. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters*, 7(2):5560–5567, 2022.

[4] Fernando Amodeo et al. OG-SGG: Ontology-guided scene graph generation—a case study in transfer learning for telepresence robotics. *IEEE Access*, 10:132564–132583, 2022.

[5] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation Transformer for Scene Graph Generation, Aug. 2022. arXiv:2201.11460 [cs] version: 2.

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting Visual Relationships with Deep Relational Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308, Honolulu, HI, July 2017. IEEE.

[7] Vinay Damodaran et al. Understanding the role of scene graphs in visual question answering. *arXiv preprint arXiv:2101.05479*, 2021.

[8] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021.

[9] Xingning Dong et al. Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19405–19414, New Orleans, LA, USA, June 2022. IEEE.

[10] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978, 2019.

[11] Martin Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[12] Filip Ilievski et al. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347, 2021.

[13] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[14] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 466–482. Springer, 2021.

[15] Ranjay Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[16] Soohyeong Lee, Ju-Whan Kim, Youngmin Oh, and Joo Hyuk Jeon. Visual question answering over scene graph. In *2019 First International Conference on Graph Computing (GC)*, pages 45–50. IEEE, 2019.

[17] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18878, 2022.

[18] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19486–19496, 2022.

[19] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming 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*, pages 11109–11119, 2021.

[20] Wei Li et al. PPDL: Predicate probability distribution based loss for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2022.

[21] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. Embodied semantic scene graph generation. In *Conference on Robot Learning*, pages 1585–1594. PMLR, 2022.

[22] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10403–10412, 2019.

[23] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. GPS-Net: Graph Property Sensing Network for Scene Graph Generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3752, Seattle, WA, USA, June 2020. IEEE.

[24] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11556, 2021.

[25] Yichao Lu et al. Context-aware scene graph generation with seq2seq transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15931–15941, 2021.

[26] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Francoise Preteux. Focusing visual relation detection on relevant relations with prior potentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2980–2989, 2020.

[27] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *Advances in neural information processing systems*, 28, 2015.

[29] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[30] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-Based Learning for Scene Graph Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13931–13940, Nashville, TN, USA, June 2021. IEEE.

[31] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased Scene Graph Generation From Biased Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, Seattle, WA, USA, June 2020. IEEE.

[32] 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/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[33] Dalin Wang, Daniel Beck, and Trevor Cohn. On the role of scene graphs in image captioning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 29–34, 2019.

[34] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8180–8189, Long Beach, CA, USA, June 2019. IEEE.

[35] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2020.

[36] Bin Wen, Jie Luo, Xianglong Liu, and Lei Huang. Unbiased Scene Graph Generation via Rich and Fair Semantic Extraction, Feb. 2020. arXiv:2002.00176 [cs].

[37] 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.

[38] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual Communication and Image Representation*, 58:477–485, 2019.

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

[40] Xuewen Yang, Yingru Liu, and Xin Wang. Reformer: The relational transformer for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5398–5406, 2022.

[41] Keren Ye and Adriana Kovashka. Linguistic Structures as Weak Supervision for Visual Scene Graph Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8285–8295, Nashville, TN, USA, June 2021. IEEE.

[42] Kanghoon Yoon, Kibum Kim, Jinyoung Moon, and Chanyoung Park. Unbiased Heterogeneous Scene Graph Generation with Relation-aware Message Passing Neural Network, Dec. 2022. arXiv:2212.00443 [cs] version: 1.

[43] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1274–1280. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021.

[44] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017.

[45] 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.

[46] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 409–424. Springer, 2022.

[47] 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.

[48] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9185–9194, 2019.

[49] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to Generate Scene Graph from Natural Language Supervision. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1803–1814, Montreal, QC, Canada, Oct. 2021. IEEE.

[50] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene Graph Generation: A Comprehensive Survey, June 2022.