Grounding Consistency: Distilling Spatial Common Sense for Precise Visual Relationship Detection

Markos Diomataris[1,2,†], Nikolaos Gkanatsios[3,†], Vassilis Pitsikalis[1,†], Petros Maragos[2]
[1]deeplab.ai, [2]National Technical University of Athens, [3]Carnegie Mellon University*
{m.diomataris, vpitsik}@deeplab.ai, ngkanats@andrew.cmu.edu, maragos@cs.ntua.gr

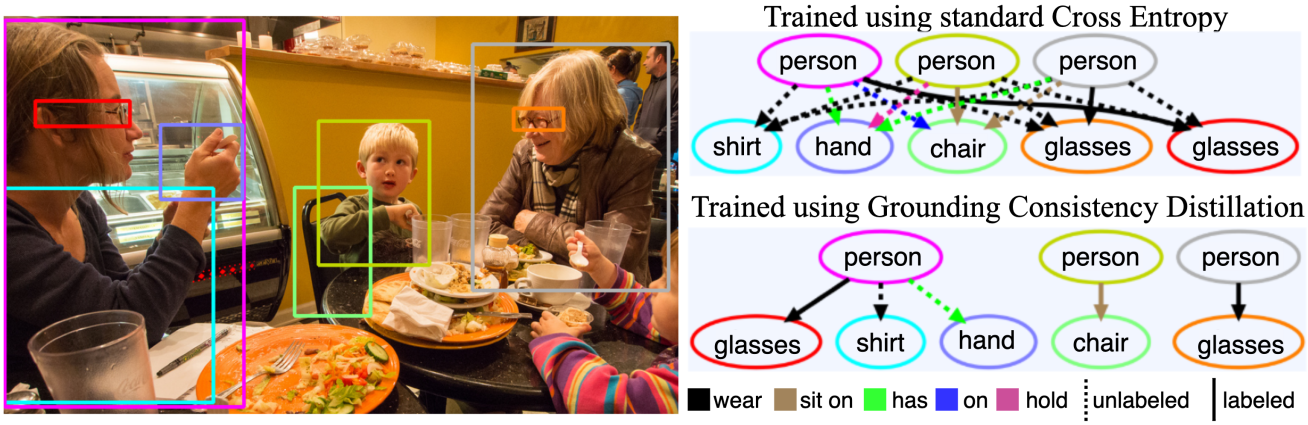
Trained using standard Cross Entropy
person
person
person
shirt
hand
chair
glasses
glasses
Trained using Grounding Consistency Distillation
person
person
person
glasses
shirt
hand
chair
glasses
■wear ■sit on ■has ■on ■hold ⋮unlabeled |labeled

Figure 1. Even a state-of-the-art model [16] overfits the object context and ignores spatial common sense, e.g. it predicts overconfident ($p > 0.7$) `wear` connections between every `person-glasses` pair, simply because `wear` is the predicate with the most samples in this context. When semi-supervised by our Grounding Consistency Distillation (GCD) scheme, the same model is able to overcome such biases, resulting in more precise scene graphs. However, current recall metrics ignore unlabeled pairs and get satisfied with both graphs, failing to capture their obvious differences. To improve visibility we omit predictions of spatial relations. Best viewed in color.

Abstract

Scene Graph Generators (SGGs) are models that, given an image, build a directed graph where each edge represents a predicted `subject predicate object` triplet. Most SGGs silently exploit datasets' bias on relationships' context, i.e. its subject and object, to improve recall and neglect spatial and visual evidence, e.g. having seen a glut of data for `person wearing shirt`, they are overconfident that every `person` is `wearing` every `shirt`. Such imprecise predictions are mainly ascribed to the lack of negative examples for most relationships, which obstructs models from meaningfully learning predicates, even those that have ample positive examples. We first present an in-depth investigation of the context bias issue to showcase that all examined state-of-the-art SGGs share the above vulnerabilities. In response, we propose a semi-supervised scheme that forces predicted triplets to be grounded consistently back to the image, in a closed-loop manner. The developed spatial common sense can be then distilled to a student SGG and substantially enhance its spatial reasoning ability. This Grounding Consistency Distillation (GCD) approach is model-agnostic and benefits from the superfluous unlabeled samples to retain the valuable context information and avert memorization of annotations. Furthermore, we demonstrate that current metrics disregard unlabeled samples, rendering themselves incapable of reflecting context bias, then we mine and incorporate during evaluation hard-negatives to reformulate precision as a reliable metric. Extensive experimental comparisons exhibit large quantitative - up to 70% relative precision boost on VG200 dataset - and qualitative improvements to prove the significance of our GCD method and our metrics towards refocusing graph generation as a core aspect of scene understanding. Code available at `https://github.com/deeplab-ai/grounding-consistent-vrd`.

1. Introduction

"Multiple people `wearing` the same shirt, `sitting` on the same chair and `having` the same hand". Embarrass-

*Work done while N. Gkanatsios was with deeplab.ai.

†This is funded by deeplab.ai

Proceedings of International Conference on Computer Vision (ICCV-2021), Oct. 2021

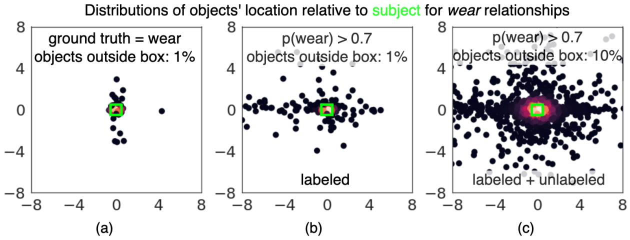

Figure 2. Objects' location distribution relative to their subject (green box) for all subject-wear-object triplets. When using ground truth (a) or a state-of-the-art model's [11] predicted high confidence ($p > 0.7$) triplets on labeled samples (b), only 1% of objects lies outside the subject's box. Incorporating unlabeled samples in evaluation (c) unveils an important misalignment between predictions and ground truth that was previously unobservable. Dimensions are normalized w.r.t. the subject's box.

ingly, as Fig. 1 implies, this is how a current state-of-the-art Scene Graph Generator (SGG) perceives our world. Performing inference on unlabeled object pairs reveals that all architectures, simple or sophisticated, lack a fundamental level of understanding of relationships. Instead, they heavily rely on dataset *context bias*, i.e. the statistical priors between predicates and subject-object categories, to overfit a handful of frequent predicates and minimally improve recall metrics that are unable to capture this fragile behavior.

Previous approaches [36, 32, 41] attribute bias to the long-tail distribution of predicates: the frequent ones overshadow the rare. Thus, they develop techniques that aim to boost recall on the tail-classes. However, Fig. 1 reveals an other implication of bias: models seem to seriously lack *spatial common sense*, even for some of the head classes commonly found in popular datasets such as `wear` and `on`.

Our work explores the *effect* and *origin* of context bias as well as the *most suffering classes*. To mitigate it, we then introduce a *semi-supervised distillation* training scheme called Grounding Consistency Distillation (GCD). In GCD, a teacher SGG network is further constrained to predict relationships that can be grounded back to the image, through a pretrained grounding network. The *spatial common sense* knowledge developed by the teacher is then distilled to a student SGG model. This model-independent scheme forces models to additionally reason for unlabeled samples, supplying out-of-distribution examples that challenge the network's perception of the dominant classes. We further contribute two *negative graph completion rules* used during testing to generate negative labels for unlabeled samples and support metrics that are more reflective of the models' ability to interpret predicates. Lastly, we re-implement and evaluate six state-of-the-art models, that *demonstrate profound gains* when adopting our scheme, even over related alternatives. Our experiments emphasize the importance of precision as a long sidelined aspect of scene graph generators that would encourage their deployment in real-world scene understanding problems.

2. Experimental Evidence on Context Bias

As a springboard for our investigation we examine the level of understanding models have for `wear`, a head class in most popular datasets. Human common sense dictates that for a subject to `wear` an object spatial proximity must apply, i.e. subject's and object's boxes should be intersecting. Most state-of-the-art models achieve close to 100% recall on `wear` but do they possess the aforementioned spatial common sense? Fig. 2 proves that ignoring predictions on unlabeled samples falsely leads us to believe that they do. In fact, even when keeping only high confidence predictions ($p > 0.7$) 10% of them appear to be wrong.

Intrigued to further probe the detectors' incapability of interpreting visual predicates, we contrive a toy *sliding box experiment*: for a given subject-predicate-object triplet, we slide the object's bounding box upon the image to extract a binary map indicating whether the predicate `wear` is predicted in that position. Fig. 3 depicts three alarming facts. First, a state-of-the-art model [54] that aggregates visual, semantic and spatial information, predicts the `person` is `wearing` the `shirt` regardless of the latter's location and appearance. Second, a baseline using only visual features [12] also suffers from these limitations and predicts `wear` when the `shirt`'s box is placed upon any `person`. Third, an even weaker *spatial baseline* [12], aware only of the two bounding boxes, is the most precise and predicts `wear` when the two boxes overlap. These observations indicate that semantic and visual features are both responsible for the memorization of the context bias and the lack of *spatial common sense*, e.g. `wearing` a `shirt` while having zero intersection with its box is not plausible [6].

Nonetheless, context information is itself a measure of a relationship's plausibility. Considering a `person` and a `chair`, humans have an instinctively high prior for `sitting on`, before even viewing the image. Subsequently, human annotators "apply their own subjective judgments on what to ignore and what to mention" [28], causing a reporting bias [38]. Relations that are more useful for scene understanding are far more likely to be annotated, e.g. we rarely expect a `person next to shirt` to be a salient concept, despite being equally observable to a `person wearing shirt`.

Towards validating this, we illustrate predicate distributions for different subject-object labels in Fig. 4a,b for two popular datasets, VRD [25] and VG200 [47]. For the `man-chair` case, most of the annotations concern `sitting on` and its *synonyms* `in`, `on` which share the same meaning in this context [12]. Under these constraints, a frequency baseline achieves a deceptively high recall score [54]. Frustratingly, state-of-the-art models only slightly deviate from - or even build on - this frequency baseline [23] and, as we

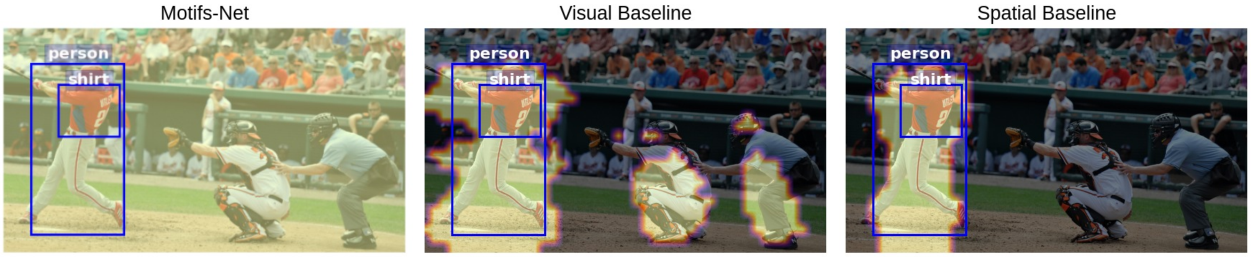

Figure 3. Sliding box experiment for three models: we fix the `person`'s box, slide the `shirt`'s box upon the image and visualize a binary heatmap representing whether `wear` is predicted. Motifs-Net [54] seems to neglect visual and spatial evidence, predicting `wear` almost everywhere. The visual baseline from [12] confuses different instances of `person`. Unaware of the classes of the referred objects, the spatial baseline from [12] employs common sense: the two boxes should intersect to predict `wear`.

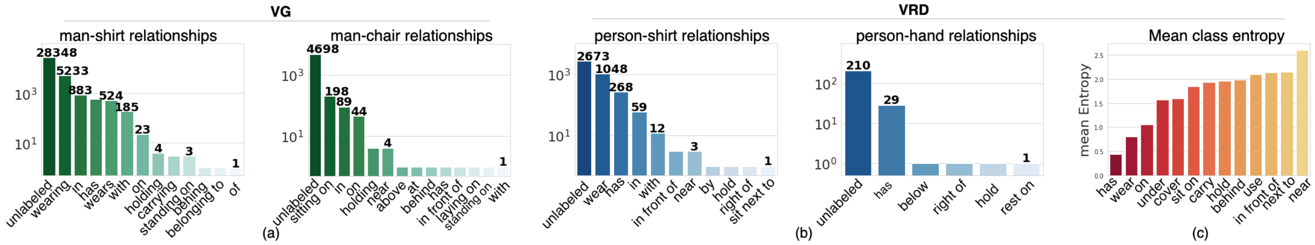

Figure 4. (a) and (b): Context bias is a result of reporting bias. Most subject-object pairs are not annotated with a predicate, only those that the annotators subjectively consider significant in the scene's description. This results in biased conditional predicate distributions where a cluster of synonyms, e.g. `wear`, `has` and `in` in the case of `person-shirt`, dominates other classes. (c) If we measure the mean entropy for the contexts at which a class is prevailing, we observe that predicates which demand a spatial *proximity* between the involved objects, e.g. `has`, have lower entropy values, meaning that they create stronger bias.

will show in section 5, achieve disconcertingly low precision scores. On the other hand, there are many unlabeled samples (97% of VG and 87% of VRD) that could serve as negative examples, yet they remain unexploited.

The limited cognition of scene graph generators, even for predicates with copious examples, underlines the need to re-evaluate which are the most problematic classes. In response, we measure the entropy of predicate distributions per context and then, for each class, we average the entropy for the contexts at which this class is the most prevalent. A detailed formulation of entropy ranking is presented in the suppl. material. This *entropy ranking analysis* (Fig. 4c) unveils that *proximal predicates*, i.e. predicates that demand a spatial pixel-wise proximity for the subject and the object (e.g. `wear`, `on`, `has`), tend to lead to higher context bias (lower entropy). The three aforementioned classes capture more than 40% of VRD's samples, fact that completely disproves prior belief that only tail-classes suffer from biases [36, 32]. Instead, it is the proximal predicates that display the most severe lack of spatial common sense.

3. Grounding Consistency Distillation

The above analysis highlights three key properties our solution has to incorporate: (1) use unlabeled samples to create a distribution shift against context bias, (2) resolve conflicts between entities of the same category that confuse the network to predict the dominant class, e.g. two `persons` holding the same `umbrella`, (3) be model-agnostic. We address these challenges with a semi-supervised distillation training scheme utilizing three different networks: a Grounder, a teacher SGG and a student SGG. Both the teacher's *spatial common sense* acquisition as well as its infusion to the student are the result of two losses $\mathcal{L}_t$ and $\mathcal{L}_s$ respectively that complement the standard cross-entropy. First, the teacher is trained with $\mathcal{L}_t$ forcing its predictions to be accurately grounded back to the image. Then, during the student's training, $\mathcal{L}_s$ distills [15] this knowledge from the teacher.

Inspired by CycleGANs [60], we call this scheme *Grounding Consistency Distillation* (GCD), since predicted relationships have to be consistent with the grounder's ability to relocate them. GCD is indifferent to the teacher or student model peculiarities and is applicable to unlabeled samples, that serve as out-of-distribution negatives for the per-context dominant classes.

Teacher training As the teacher SGG we employ ATR-Net [11] and assume an existing trained and frozen grounder, i.e. a model that, given a relationship triplet, localizes the bounding boxes of the referring subject-object entities. Training now obtains a closed-loop form, with the teacher predicting a predicate for a pair of entities and the grounder re-estimating their spatial configuration

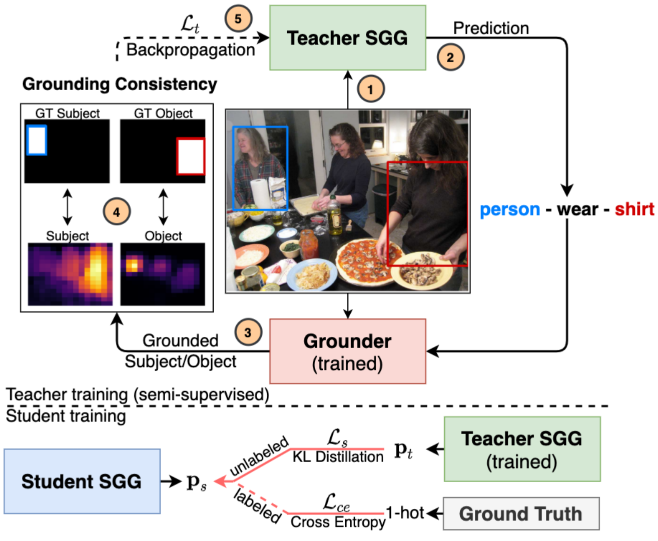

Figure 5. Teacher and Student training pipelines. Teacher: an unlabeled object pair is given as an input (1) to predict a relationship triplet (2), then a Grounder attempts to locate the referring entities, i.e. the subject and the object, back to the image (3). A misguided grounding, e.g. predicting the `person` (blue) is `wearing` the `shirt` (red), leads to a grounding inconsistency (4) that penalizes the relationship classification (5). Student: predictions on unlabeled samples are used to distill knowledge from a trained teacher while standard cross-entropy is applied when labels are available.

(Fig. 5). Based on the grounding quality we penalize or reward a detected relationship, e.g. in Fig. 5, a spatially-inconsistent prediction back-propagates the grounding error to the teacher.

Formally, let $f(S, O) \rightarrow \mathbf{p}_t$ be the teacher relationship detector that maps subject $S = (s_v, s_{sem}, s_{sp})$ and object $O = (o_v, o_{sem}, o_{sp})$ information (visual, semantic, spatial) to a probability distribution on predicates $\mathcal{P}$. If $r = argmax(\mathbf{p}_t^i)$ then, inversely, the grounder is a function g defined as $g(s_{sem}, r, o_{sem}) \rightarrow (\mathbf{h}_s \in \mathbb{R}^{H\times W}, \mathbf{h}_o \in \mathbb{R}^{H\times W})$, that spatially grounds a relationship r to heatmaps measuring the confidence of the localization of each entity upon a $H \times W$ image representation.

We quantify the grounding quality $q \in [0, 1]$ by averaging the maximum confidence value predicted inside the subject's and object's bounding boxes:

$$q = \frac{max(\mathbf{h}_s \odot \mathbf{m}_s) + max(\mathbf{h}_o \odot \mathbf{m}_o)}{2} \tag{1}$$

where $\mathbf{m}_s, \mathbf{m}_o$ are $H \times W$ binary masks that are non-zero inside the ground-truth boxes, $\odot$ is the Hadamard product.

$\mathcal{L}_t$ is the cross-entropy between the grounding quality q and the probability $\mathbf{p}_t^r$ of the predicted predicate r:

$$\mathcal{L}_t = -[q \; log \; \mathbf{p}_t^r + (1-q) \; log \; (1-\mathbf{p}_t^r)] \tag{2}$$

Note, that error backpropagates only through $\mathbf{p}_t^r$. Intuitively, highly probable predicate predictions should ensure high-quality grounding of the referring entities. On the other hand, a spatially implausible predicate causes a mismatch between the estimated heatmaps $\mathbf{h}_s, \mathbf{h}_o$ and the ground-truth boxes that imposes a penalty on the detection. Our total objective is the sum of the standard cross-entropy loss with $\mathcal{L}_t$:

$$\mathcal{L}_t^{total} = \mathcal{L}_{ce} + \alpha(t)\mathcal{L}_t \tag{3}$$

where $\alpha(t)$ is a time-dependent regularizer which increases over time to balance between memorization (recall) $\mathcal{L}_{ce}$ and generalization (precision) $\mathcal{L}_t$.

Student training employs the standard cross-entropy accompanied by $\mathcal{L}_s$: the Kullback-Leibler divergence from $\mathbf{p}_t$ to $\mathbf{p}_s$ (student's output distribution) regularized by a constant hyperparameter λ. The student's total objective is:

$$\mathcal{L}_s^{total} = \mathcal{L}_{ce} + \lambda D_{\text{KL}}(\mathbf{p}_t \parallel \mathbf{p}_s) \tag{4}$$

Since labeled samples already provide training information, both $\mathcal{L}_t$ and $\mathcal{L}_s$ are only applied on unlabeled pairs.

The crux of avoiding directly using the teacher for relationship detection and instead employing distillation is that $\mathcal{L}_t$ is not equally sensitive to all types of misclassification. In fact, since the quality q is not a distribution on $\mathcal{P}$ but rather an isolated plausibility score for r, any class ensuring a high-quality grounding is going to be rewarded. This means that occasionally q may be an overestimate of the prediction probability $\mathbf{p}_t^r$ inducing noise that, as we show in section 5, has a negative effect on models' Recall. The student-teacher scheme attenuates this misbehavior by using the teacher to filter out that noise while doing a better job in distilling its developed *spatial common sense* to the student. That filtering is a result of KL divergence penalizing $\mathbf{p}_s$ proportionally to its deviation from $\mathbf{p}_t$.

Grounding methodology The classic setup for grounding referring relationships [20] matches a subject-predicate-object triplet to the image by detecting both the subject and the object. However, we find that this setup does not handle ambiguous cases where the input triplet can be grounded to more than one pair of entities, e.g. grounding `person wearing hat` on an image showing two `persons` both wearing `hats`. Since we use grounding as a scaffold for learning precise relationships and not to compare to prior literature, we modify the task to resolve such ambiguities by conditioning the object's localization to the subject's ground-truth bounding box and vice versa, thus solving two independent grounding problems.

We break the grounding of each entity into two steps. The first step estimates a plausible box that suits the image scale: "how big should an `elephant` be given that this `person` is `riding` it?". We tackle this question as a regression problem on the box's dimensions. The second step regresses a $H \times W$ heatmap assessing the spatial probability distribution of the position of the estimated box's center.

The GCD formulation is invariant to the exact choice of grounder. Therefore, a more detailed presentation is out of our scope and we refer the reader to our suppl. material.

4. Reorientation of Evaluation with Negatives

As already shown (Fig. 1, Fig. 2, Fig. 4a,b), only when examining unlabeled samples we are able to ascertain the effects of context bias, underscoring the importance of their inclusion in evaluation. However, the most commonly used metric Recall@k (R@k) [25], which measures the portion of true positive relations in the top-k detections, does not penalize mispredictions on the unlabeled pairs. On the other hand, precision metrics that regard unlabeled samples as negatives are pessimistic, since they may penalize correctly predicted relationships that are not annotated [25]. Furthermore, we experimentally prove that measuring precision this way is not insightful. This urges to reexamine how to utilize unlabeled samples in evaluation.

Negative Graph Completion We propose a method to mine and incorporate unlabeled samples into meaningful metrics that reflect context bias and spatial common sense by introducing two *negative graph completion rules* that generate negative labels for proximal predicates. Proximal predicates, which suffer the most from context bias, can be divided in two sets: *possessive* and *belonging*. Possessive predicates denote a possession passing from the subject to object, e.g. having and eating. Belonging predicates have an inverse meaning, with the subject being a part of or living on the object, e.g. of and sitting on. In general, subjects in relations with possessive predicates do not "share" the objects, e.g. in person has hand, the hand belongs only to the referred person. Similarly, subjects connected to objects with belonging predicates are not to be "shared", e.g. a person lying on sofa most probably cannot be simultaneously lying on another sofa. With $\mathcal{R}_p$, $\mathcal{R}_b$ denoting the sets of possessive and belonging predicates respectively, $r(s,o)$ the relationship between the subject s, object o with predicate r, we obtain the following rules:

- Possessive:$\forall r \in \mathcal{R}_p, \forall s, o, s' : r(s,o) \implies \neg r(s',o)$
- Belonging:$\forall r \in \mathcal{R}_b, \forall s, o, o' : r(s,o) \implies \neg r(s,o')$

Examples for these rules are depicted in Fig. 6. Full list of Possessive/Belonging relationships in our supp. material.

The negative labels enrich datasets' test set with targeted and challenging examples that demand models to be more precise, e.g. a model predicting on each time it encounters a jacket and a person would now miss a multitude of negative examples for on. At the same time, precision is not prone to incomplete annotations and can be safely measured on the samples that have a positive or a negative label.

Why not simply rank background edges? Past approaches [54, 56] employ a "background" class and consider the unlabeled samples as negatives for all classes. This approach is inherently flawed, as all the unlabeled samples in fact belong to existing classes. Moreover, the two most prominent pruning strategies manually filter non-intersecting pairs of object boxes as "background" [54] or learn a separate "relatedness" task [11]. Although they partly ameliorate ranking of edges, they conceal the lack of spatial common sense: networks still classify the person (blue box) of Fig. 6 as wearing the shirt (cyan box). Finally, arguing that mispredictions on unlabeled pairs showcase low probability is disproved by Fig. 2c where a large portion of mispredictions is overconfident ($p > 0.7$).

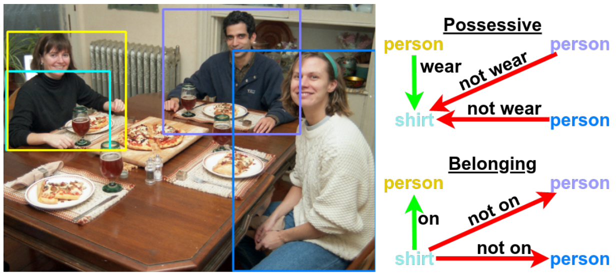

Figure 6. Annotated relationships (green) are used to generate negative edges (red) following specific rules. *Possessive*: since a person (yellow) is wearing the shirt, other persons cannot wear it. *Belonging*: since the shirt is on a person, it cannot be on another person.

5. Experiments and Results

We evaluate a plethora of state-of-the-art scene graph generators on two datasets aiming to: (1) quantitatively show the context bias effects and validate GCD's efficacy for all tested models, (2) qualitatively explicate GCD's effect towards more precise scene graphs and improved spatial common sense, (3) exhibit the improved ability of our metrics to capture context bias, (4) demonstrate GCD's superiority against other alternatives.

Models, Datasets and Metrics Our model zoo comprises of six re-implemented models, VTransE [57], Motifs-Net [54], RelDN [59], ATR-Net [11], UVTransE [16] and HGAT-Net [27], all employing various feature types and architectures. Implementation details are included in our suppl. material. We train and test all models on VRD [25] and VG200 [47] for predicate detection (PredDet) [25] and predicate classification (PredCls) [47] respectively. In PredCls object categories and boxes are considered known, while in PredDet the additional information of objects interacting is given. We choose those tasks so as to avoid interference with object detection errors. $a(t)$ (eq. 3) is empirically set a unit step function that rises after the first epoch and λ (eq. 4) equal to 80. We report R@50, micro Precision (mP) measured only on labeled samples, mP$^+$ and f-mP$^+$ where $^+$ denotes additional evaluation on our mined negative labels and f- *focusing* measurements only on prox-

Models	VRD (PredDet)					VG200 (PredCls)				
	R@50	mP	mP$^+$	f-mP$^+$	HarMean	R@50	mP	mP$^+$	f-mP$^+$	HarMean
VTransE [57]	53.17	13.11	17.42	26.95	35.77	61.16	2.21	4.57	15.60	24.72
Motis-Net [54]	55.06	13.31	20.67	32.38	40.78	62.54	2.32	4.50	17.98	27.70
RelDN [59]	55.02	13.66	22.94	36.63	43.98	57.83	2.03	4.93	16.82	25.89
ATR-Net [11]	57.69	13.99	23.87	38.78	46.38	63.02	2.25	5.82	20.01	30.30
UVTransE [16]	56.88	13.46	21.63	34.69	43.10	62.69	2.24	4.60	15.57	24.88
HGAT-Net [27]	57.00	13.84	22.46	36.26	44.32	63.30	2.32	5.40	16.82	26.56
VTransE + GCD	54.01	12.92	20.46	36.62	43.65	60.64	2.28	7.18	24.63	34.79
Motifs-Net + GCD	55.12	13.06	25.58	42.43	47.95	63.30	2.27	7.36	25.28	36.08
RelDN + GCD	53.97	12.89	25.22	41.44	46.88	55.49	1.99	7.33	25.32	34.43
ATR-Net + GCD	57.59	13.93	28.98	48.33	52.56	63.35	2.32	7.34	25.17	35.92
UVTransE + GCD	56.72	13.72	28.2	46.77	51.26	62.36	2.28	7.70	26.45	37.04
HGAT-Net + GCD	56.24	13.34	25.8	42.66	48.52	62.83	2.31	7.42	25.50	36.28
Teacher ATR-Net	57.21	13.98	29.43	48.97	52.77	62.78	2.52	7.15	25.55	35.58

Table 1. Results of re-implemented models with and without GCD. We measure Recall@50 (R@50), micro Precision (mP), mP$^+$, f-mP$^+$ and the Harmonic Mean of R@50 and f-mP$^+$. $^+$ indicates additional evaluation on mined negative labels, f- focusing evaluation only on proximal predicates. Teacher included for reference. We conduct experiments for five random initializations. Maximum standard deviation for VRD: R@50 ±0.42, mP ±0.18, mP$^+$ ±0.66, f-mP$^+$ ±1.16. For VG200: R@50 ±0.04, mP ±0.02, mP$^+$ ±0.22, f-mP$^+$ ±0.39.

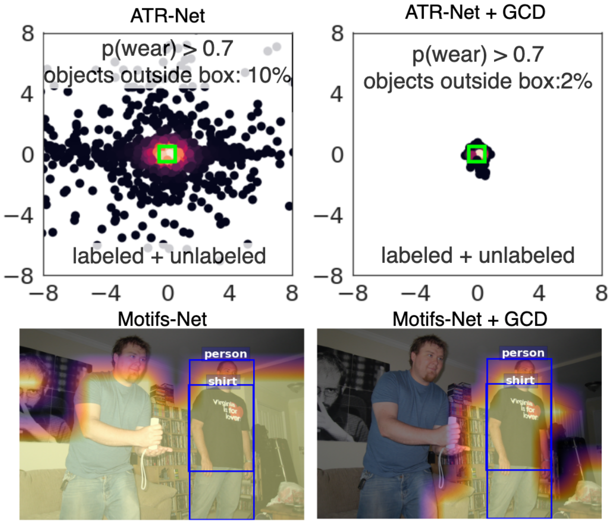

Figure 7. Top: GCD manages to focus the distribution of objects on the subject for ATR-Net's `wear` predictions. Bottom: Sliding box experiment for the phrase `person wearing shirt`. Originally, Motifs-Net predicts `wear` even if the `shirt` is located in background regions. When trained with GCD, it acquires a basic level of spatial common sense and predicts `wear` only upon or very close to the subject.

imal predicates. Lastly, we compute the Harmonic Mean (HarMean) of R@50 and f-mP$^+$ as an overall metric.

Context bias and Grounding Consistency The results for all re-implemented baselines are included in the upper half of Table 1. HarMean changes the ranking between models on both datasets, since models with similar R@50, e.g. UVTransE-Net and HGAT-Net, display significantly different precision gains. The lower half of the Table 1 contains the results when the same models are additionally semi-supervised using the proposed scheme (+GCD). We notice large improvements on mP$^+$ and f-mP$^+$ (up to 35% relative on VRD and 70% on VG200 for UVTransE) with non-substantial R@50 sacrifice. In total, HarMean is increased, up to 22% relative on VRD and 49% on VG200.

Spatial common sense and sparser graphs Models semi-supervised by GCD are able to generate sparser graphs (Fig. 8) and develop a basic level of spatial common sense (Fig. 7). Note, for instance, how ATR-Net (Fig. 8 upper left) is able to perfectly resolve conflicts between all `persons` and `clothes`, indicating an improved understanding of predicates' meanings. For more qualitative results refer to our supp. material.

What do models predict in place of the most frequent predicate? A model that penalizes a predicate in favor of a synonym [12], e.g. predicting `person on chair` instead of `sit on` for a sample where `on` is false, is equivalently ignorant of predicate interpretations. Visualizing all edges and predictions upon the graph (top and middle right column of Fig. 8) indicates that models trained with GCD harness implicit spatial features and give reasonable predictions for all samples, e.g. a `person` falsely being `on skis` is now `next to` them (Fig. 8 top right).

Metric comparisons Despite the above qualitative evidence, R@50 or mP do not captivate a quantitative improvement. On the other hand, mP+ clearly quantifies GCD's benefits for all models, due to the employment of targeted negatives in evaluation that penalize relentless biased predictions based on context. We further validate that f-mP$^+$ better captures the models' behavior compared to mP+, without significantly altering the ranking. This can be attributed to the nature of non-proximal predicates: most of

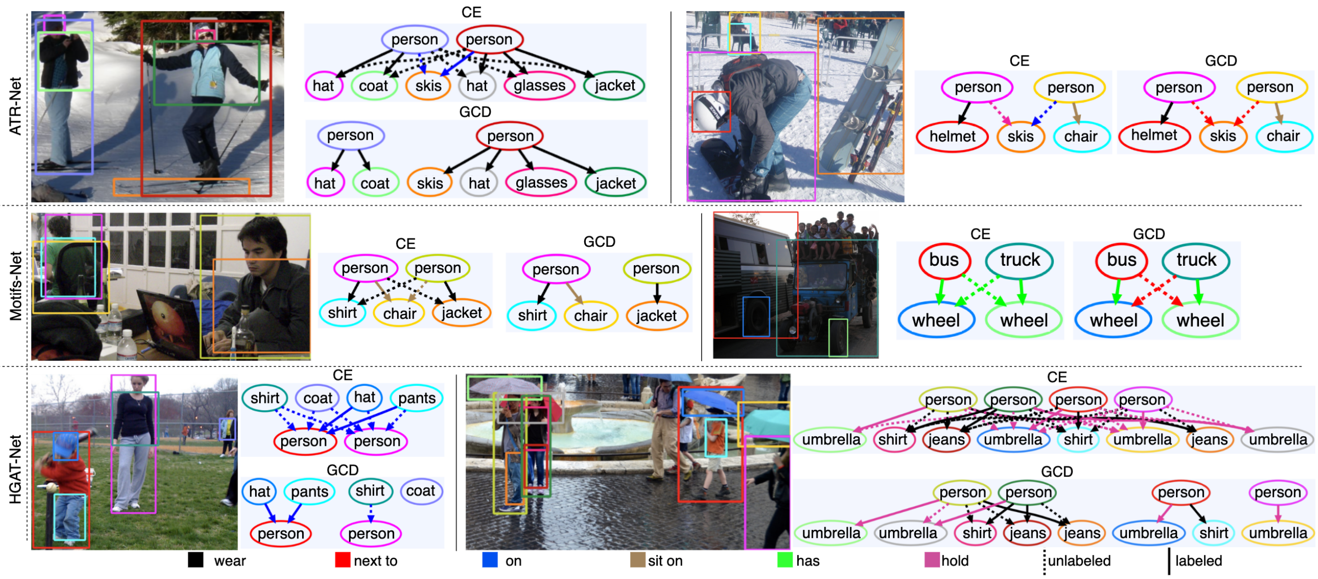

Figure 8. Qualitative comparison of three models' predictions on images with proximal predicates trained with standard cross-entropy (CE) and with our method (GCD). With the exception of top and middle of right column, non-proximal predicates are filtered out for clarity. GCD creates graphs with sparse connected components. Most edges incorrectly classified as proximal predicates are now associated to a reasonable geometric predicate. Best viewed in color.

Method	R@50	mP$^+$	f-mP$^+$	HarMean
GCD-G	**+0.50**	+5.23	+6.24	+4.03
GCD-D	-2.41	+16.36	+20.53	+10.22
GCD	-0.33	**+19.28**	**+26.19**	**+14.61**

Table 2. Ablation on GCD's structure reporting the average relative gains for R@50, mP+, f-mP+ and HarMean across the six baselines of Table 1. Removing grounding (GCD-G) or distillation (GCD-D) will respectively cancel out Precision gains and restrict models from optimally developing spatial common sense.

Method	R@50	mP$^+$	f-mP$^+$	HarMean
Spatial Baseline*	47.08	20.09	32.87	38.71
Oracle Teacher*	56.44	33.13	55.61	56.02
SpatDistill	-0.09	+12.67	+15.10	+8.82
GraphL	**+0.27**	+17.95	+23.25	+13.35
GCD (Ours)	-0.33	**+19.28**	**+26.19**	**+14.66**
oracle with NCE (Ours)	-0.09	+39.03	+45.47	+24.03

Table 3. Average relative performance gains across the six baselines of Table 1. GCD outperforms other approaches in distilling spatial common sense and is comparable to the oracle NCE that uses ruled-based negatives. For models with * the absolute results are reported for reference only.

them are geometric and alternatively used for each other, e.g. `next to`, `near` and `adjacent to`. [12] show that in such cases, models tend to predict the most frequent synonym per context. Resolving that type of bias is a hard problem and outside the scope of this work. Instead, proximal predicates clearly benefit from GCD, as f-mP$^+$ reflects.

Ablation study Combining grounding and knowledge distillation is key to effectively acquiring spatial common sense. To validate this, we perform an ablation study with two structural variations of GCD: removing the Grounder (GCD-G) and applying $\mathcal{L}_t$ directly on the baseline models without an intermediate distillation step (GCD-D).

Table 2 showcases the mean relative performance gain across the six baselines presented in Table 1. GCD-D introduces a high relative Recall drop while having inferior Precision boost in comparison to GCD which manages to both maximize Precision and retain minimal Recall penalty. GCD-D proves that simply using the teacher's soft-labels on unlabeled samples to distill knowledge is not able to improve models' spatial reasoning ability.

Comparison to other approaches While retaining the teacher-student part of GCD we experiment with alternative sources of spatial common sense besides a Grounder. Motivated by our analysis that networks biased to context neglect spatial features, we employ the spatial baseline of Fig. 3. We call this approach SpatDistill. A second approach is to directly use the oracle negatives derived from our rules and apply the Negative Cross-Entropy loss (NCE) of [18] to the teacher (Oracle Teacher). Distillation from the Oracle Teacher serves as an upper bound of GCD since networks do not have to reason, using an imperfect grounder, about whether an example is a negative. Lastly, we compare GCD to the graphical contrastive losses (GraphL) of [59], that learn to rank negative samples based on rules.

The resulting mean relative performance gains are provided in Table 3, where GCD has an obvious advantage over SpatDistill and GraphL. Although precise, the spatial baseline is naive (Table 3) and constrains models' ability to learn the good context prior. GraphL deteriorates precision gains as it incorrectly regards all unlabeled pairs as negatives for all classes, yet these do belong to a class; in fact, there are many unlabeled positives even for proximal predicates. Lastly, targeted negatives (NCE) have a great impact on the precision metrics. Note, that, in contrast to GCD, GraphL and NCE depend on rules that, although valid on VRD and VG200, may not generalize for all datasets. See supplementary for the expanded versions of Tables 2, 3.

Limitations of GCD The grounder employed by our pipeline is not perfect: it can be confused by instances that lie too close (Fig. 8 bottom right), while incorrect predicate predictions may result in correct grounding. Nevertheless, our experiments prove that GCD achieves a basic level of spatial common sense and is comparable to the oracle NCE while being semi-supervised.

6. Related work

Visual Relationship Detection and **Scene Graph Generation** (SGG) both refer to detecting objects and classifying the predicate of each pair separately [56, 57, 59, 2, 11, 22, 31, 50] or jointly upon the graph [3, 4, 51, 52, 53, 37, 44, 43, 34]. All these works optimize recall on the datasets' ground-truth and they are cursed to overfit the context bias and sacrifice precision. Closest to ours, [59] also explore qualitative errors attributed to confusing entities and use contrastive losses to improve the average precision on specific classes. However, they treat all unlabeled samples as negatives and provide limited insight on the predictions between "not related" entities. Orthogonal to the above, we thoroughly analyze the context bias effects, mine targeted negatives to enhance the resonance of our metrics and then amplify the precision for all tested models.

Unbiased SGG Concurrent literature imputes the bias to the long-tail distribution of relationships and applies low-shot learning [5, 42, 30] or image manipulation [45, 8, 19] to overcome the lack of samples for tail classes. [12] and [36] expose the effects of mimicking context bias on few-shot generalization. Our analysis on unlabeled samples and the sliding box experiment universalize the inability to properly interpret frequent predicates as well. Thus, we redefine the suffering classes with a context-conditioned entropy rather than a predicate-conditioned frequency ranking.

Grounding refers to the localization of an image region described by a natural language expression [26, 33]. Recent approaches align visual and language scene graphs [48, 24, 40] to parse and disambiguate referring expressions. Closest to ours, [20, 35] explicitly ground referring relationships, i.e. subject - predicate - object triplets, but their task differs in that they aim to detect both objects, while we condition the subject to the object and vice versa. Our grounder also draws inspiration from **spatial common sense** works [49], particularly in breaking the detection in two steps, an image-agnostic inference of objects' layouts given a relationship [6] and then a refinement on the image.

Grounding Consistency is inspired by recent semi-supervised approaches [55, 17]. The original consistency regularization loss [21] minimizes the difference between the predictions $f(x)$ and $f(x')$ for an input image x and its perturbed version x'. Our formulation is reminiscent of the adversarial Cycle Consistency Loss [60], where f and f^{-1} are jointly learned so that $f^{-1}(f(x))$ approximates x. In our case, we approximate f^{-1} with a pretrained grounder. The only scene graph generator that shares a similar consistency logic is that of [13], which auto-encodes images via intermediate scene graph representations, but uses a generative model to reconstruct the image, while we re-ground objects to enhance the detector's spatial awareness.

Limited supervision for SGG has been an answer to the surplus of unlabeled samples due to the sparsity of the annotated scene graphs. A stream of works thus employ weakly-supervised approaches [29, 58, 10, 51, 1] to take advantage of both labeled/unlabeled data. [7, 11, 31, 32, 56, 9, 46] use filters or multi-tasking to rank the labeled samples above the unlabeled ones. Other approaches use semi-supervised learning [5], self-training [2] or distillation [50, 31] to estimate pseudo-labels for unlabeled samples. However, pseudo-labels also suffer from context bias. Diametrically opposite, our semi-supervised approach directly penalizes predictions that are not grounded back to the image.

Scene Graph Completion A few works attempt to populate the graph with predicate edges based on the existing ones [39]. [14] apply rules for transitive and converse relations, while [12] construct synonym classes of relationships. These approaches generate positive examples to assist training, while our rules mine targeted negative examples to enhance the insight of precision metrics.

7. Conclusions

Current state-of-the-art generators are yet far from supporting visual graph reasoning. Instead, they overfit the context bias to satisfy recall metrics of little insight. We design a semi-supervised framework that grounds the predicted relationships back to the image to cultivate a basic level of spatial common sense. We further devise two negative graph completion rules to enhance the test set with meaningful negative examples able to capture context bias and demonstrate significant gains under various setups. However, spatial common sense is a single aspect of interpreting visual predicates. Future detectors should also incorporate concept reasoning as a higher level of knowledge about the physical world, which is not limited to specific types of object interactions. We are confident that our approach motivates a rethinking of the importance of unlabeled data as an inherent element of both scene graph generation and, equally importantly, evaluation.

References

[1] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based Weakly-supervised Learning of Visual Relations with Graph Networks. In *Proc. ECCV*, 2020.

[2] Diqi Chen, Xiaodan Liang, Yizhou Wang, and Wen Gao. Soft Transfer Learning via Gradient Diagnosis for Visual Relationship Detection. In *Proc. WACV*, 2019.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Scene Dynamics: Counterfactual Critic Multi-Agent Training for Scene Graph Generation. In *Proc. ICCV*, 2019.

[4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-Embedded Routing Network for Scene Graph Generation. In *Proc. CVPR*, 2019.

[5] Vincent S. Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Ré, and Li Fei-Fei. Scene Graph Prediction With Limited Labels. In *Proc. ICCV*, 2019.

[6] Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *Proc. AAAI*, 2018.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting Visual Relationships with Deep Relational Networks. In *Proc. CVPR*, 2017.

[8] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D. Hager, Federico Tombari, and C. Rupprecht. Semantic Image Manipulation Using Scene Graphs. In *Proc. CVPR*, 2020.

[9] Mohammed Haroon Dupty, Zhongpei Zhang, and Wee Sun Lee. Visual Relationship Detection with Low Rank Non-Negative Tensor Decomposition. In *AAAI*, 2020.

[10] Sarthak Garg, Joel Ruben Antony Moniz, Anshu Aviral, and Priyatham Bollimpalli. Learning to Relate from Captions and Bounding Boxes. In *Proc. ACL*, 2019.

[11] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. Attention-Translation-Relation Network for Scalable Scene Graph Generation. In *Proc. ICCV Workshops*, 2019.

[12] Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. From Saturation to Zero-Shot Visual Relationship Detection Using Local Context. In *Proc. BMVC*, 2020.

[13] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene Graph Generation with External Knowledge and Image Reconstruction. In *Proc. CVPR*, 2019.

[14] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. Learning Canonical Representations for Scene Graph to Image Generation. In *Proc. ECCV*, 2020.

[15] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. *ArXiv*, abs/1503.02531, 2015.

[16] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. *PAMI*, 2020.

[17] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based Semi-supervised Learning for Object Detection. In *Proc. NeurIPS*, 2019.

[18] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. NLNL: Negative Learning for Noisy Labels. In *Proc. ICCV*, 2019.

[19] Matthew Klawonn and Eric Heim. Generating Triples With Adversarial Networks for Scene Graph Construction. In *Proc. AAAI*, 2018.

[20] R. Krishna, Ines Chami, M. Bernstein, and Li Fei-Fei. Referring Relationships. In *Proc. CVPR*, 2018.

[21] S. Laine and Timo Aila. Temporal Ensembling for Semi-Supervised Learning. In *Proc. ICLR*, 2017.

[22] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual Relationship Detection With Deep Structural Ranking. In *Proc. AAAI*, 2018.

[23] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. GPS-Net: Graph Property Sensing Network for Scene Graph Generation. In *Proc. CVPR*, 2020.

[24] Daqing Liu, Hanwang Zhang, Z. Zha, Meng Wang, and Qianru Sun. Joint Visual Grounding with Language Scene Graphs. *ArXiv*, abs/1906.03561, 2020.

[25] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual Relationship Detection with Language Priors. In *Proc. ECCV*, 2016.

[26] Junhua Mao, J. Huang, A. Toshev, Oana-Maria Camburu, A. Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *Proc. CVPR*, 2016.

[27] Li Mi and Zhenzhong Chen. Hierarchical Graph Attention Network for Visual Relationship Detection. In *Proc. CVPR*, 2020.

[28] I. Misra, C. L. Zitnick, Margaret Mitchell, and Ross B. Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proc. CVPR*, 2016.

[29] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-Supervised Learning of Visual Relations. In *Proc. ICCV*, 2017.

[30] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Detecting Unseen Visual Relations Using Analogies. In *Proc. ICCV*, 2019.

[31] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise J. Prêteux. Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. In *Proc. ICME*, 2018.

[32] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise J. Prêteux. Focusing Visual Relation Detection on Relevant Relations with Prior Potentials. In *Proc. WACV*, 2020.

[33] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In *Proc. ICCV*, 2017.

[34] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive Relational Networks for Mapping Images to Scene Graphs. In *Proc. CVPR*, 2019.

[35] Moshiko Raboh, Roei Herzig, Gal Chechik, Jonathan Berant, and Amir Globerson. Differentiable Scene Graphs. In *Proc. WACV*, 2020.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased Scene Graph Generation from Biased Training. In *Proc. CVPR*, 2020.

[37] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Weiwei Liu. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *Proc. CVPR*, 2019.

[38] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. L. Zitnick, and D. Parikh. Learning Common Sense through Visual Abstraction. In *Proc. ICCV*, 2015.

[39] H. Wan, Yonghao Luo, Bo Peng, and W. Zheng. Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding. In *IJCAI*, 2018.

[40] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. In *Proc. WACV*, 2020.

[41] Tzu-Jui Wang, Selen Pehlivan, and J. Laaksonen. Tackling the Unannotated: Scene Graph Generation with Bias-Reduced Models. In *Proc. BMVC*, 2020.

[42] W. Wang, Meng Wang, Sen Wang, Guodong Long, L. Yao, G. Qi, and Y. A. Chen. One-Shot Learning for Long-Tail Visual Relation Detection. In *Proc. AAAI*, 2020.

[43] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. In *Proc. CVPR*, 2019.

[44] Wen-Bin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching Image Gist: Human-Mimetic Hierarchical Scene Graph Generation. In *Proc. ECCV*, 2020.

[45] Xiaogang Wang, Qianru Sun, Tat-Seng Chua, and Marcelo H. Ang. Generating Expensive Relationship Features from Cheap Objects. In *Proc. BMVC*, 2019.

[46] Ruihai Wu, K. Xu, Chenchen Liu, Nan Zhuang, and Y. Mu. Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection. In *Proc. AAAI*, 2020.

[47] Danfei Xu, Yuke Zhu, Christopher Bongsoo Choy, and Li Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *Proc. CVPR*, 2017.

[48] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-Modal Relationship Inference for Grounding Referring Expressions. In *Proc. CVPR*, 2019.

[49] Mark Yatskar, V. Ordonez, and Ali Farhadi. Stating the Obvious: Extracting Visual Common Sense Knowledge. In *Proc. NAACL-HLT*, 2016.

[50] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. In *Proc. ICCV*, 2017.

[51] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging Knowledge Graphs to Generate Scene Graphs. In *Proc. ECCV*, 2020.

[52] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly Supervised Visual Semantic Parsing. In *Proc. CVPR*, 2020.

[53] Alireza Zareian, Haoxuan You, Zhecan Wang, and Shih-Fu Chang. Learning Visual Commonsense for Robust Scene Graph Generation. In *Proc. ECCV*, 2020.

[54] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene Graph Parsing with Global Context. In *Proc. CVPR*, 2018.

[55] Xiaohua Zhai, A. Oliver, A. Kolesnikov, and Lucas Beyer. S4L: Self-Supervised Semi-Supervised Learning. In *Proc. ICCV*, 2019.

[56] Yibing Zhan, Jia Ming Yu, Ting Yu, and Dacheng Tao. On Exploring Undetermined Relationships for Visual Relationship Detection. In *Proc. CVPR*, 2019.

[57] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual Translation Embedding Network for Visual Relation Detection. In *Proc. CVPR*, 2017.

[58] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN. In *Proc. ICCV*, 2017.

[59] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical Contrastive Losses for Scene Graph Generation. In *Proc. CVPR*, 2019.

[60] Jun-Yan Zhu, T. Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proc. ICCV*, 2017.