Exploring the Semantics for Visual Relationship Detection

Wentong Liao[1*] Cuiling Lan[2†] Wenjun Zeng[2] Michael Ying Yang[3] Bodo Rosenhahn[1]

[1]Institute für Informationsverarbeitung, Leibniz Universität Hannover, Germany [2]Microsoft Research Asia
[3]Scene Understanding Group, University of Twente, Netherlands

{liao, rosenhahn}@tnt.uni-hanover.de {culan, wezeng}@microsoft.com michael.yang@utwente.nl

Abstract

*Scene graph construction / visual relationship detection from an image aims to give a precise structural description of the objects (nodes) and their relationships (edges). The mutual promotion of object detection and relationship detection is important for enhancing their individual performance. In this work, we propose a new framework, called semantics guided graph relation neural network (SGRN), for effective visual relationship detection. First, to boost the object detection accuracy, we introduce a source-target class cognoscitive transformation that transforms the features of the co-occurent objects to the target object domain to refine the visual features. Similarly, source-target cognoscitive transformations are used to refine features of objects from features of relations, and vice versa. Second, to boost the relation detection accuracy, besides the visual features of the paired objects, we embed the class probability of the object and subject separately to provide high level semantic information. In addition, to reduce the search space of relationships, we design a semantics-aware relationship filter to exclude those object pairs that have no relation. We evaluate our approach on the Visual Genome dataset and it achieves the state-of-the-art performance for visual relationship detection. Additionally, Our approach also significantly improves the object detection performance (*i.e. *4.2% in mAP accuracy).*

1. Introduction

In recent years we have witnessed the important breakthroughs in object-centric visual scene understanding, such as object detection [26, 25] and semantic segmentation [19, 36]. The further analysis of the relationship between objects is central to a rich understanding of our visual world. The task of visual relationship detection aims to infer the relations of pairwise objects within an image as illustrated in Fig. 1. A relationship is defined as a triplet ⟨*subject-predicate-object*⟩ [20]. All the relationships in an image can be described by a scene graph as a collection of objects (nodes), and their relationships (edges).

Figure 1: Illustration of visual relation detection and its scene graph representation. Among the detected objects (left), relations of pairwise objects are detected (middle). The visual relation can also be represented by a scene graph (right) based on the detected objects (nodes) and their relationships (edges).

For visual relationship detection, one naive way is to train a generic detector for handling each triplet combination [27, 2], namely *visual phrase*. Then, $\mathcal{O}(N^2K)$ categories from ⟨subject-predicate-object⟩ combinations need to be considered for N object classes and K predicates. However, each class needs to be fed with sufficient training data, which is hard to collect in reality because of the limitation of human efforts, and scarcity of many relationships (*i.e.* long-tail problem [20, 29]). To address this problem, most works train detectors for the object and predicate, respectively. The detection results are aggregated to have triplet combinations [20, 31, 1, 16]. In fact, objects and relations are not semantically independent. For example, a person and a horse are more likely to be in relation "ride" than "eat". In order to improve the prediction accuracy, many

*This work is done when Wentong Liao is an intern at MSRA
†Corresponding author

recent works [29, 12, 13, 32] consider this dependence and jointly infer the objects and their relationships.

Two important problems arise. **i)** One is how object detection and relation detection can mutually enhance each other. Physically, objects and relations are semantically dependent. To jointly infer the object and predicate categories, recent works [12, 13, 30] distill the visual representation of objects and phrases, by passing messages to each other to capture their context dependencies. In those models, they learn weights to determine the amount of information collected from other objects or phrases [13, 30]. Features of different types of objects are different. To avoid the misleading caused by merging features of other types of objects, those models tend to give large weights to the objects of the same type and exclude the contribution from objects of other types. They are not optimal without making full use of objects of other types in the context for feature enhancement. **ii)** The other is how to effectively tackle the superabundant $\langle$subject, object$\rangle$ combinations from the object proposals for further recognizing the relation. Among the $N(N-1)$ combined object pairs from N object proposals, the feasible combinations having meaningful relations are rare/sparse. Such superabundant phrases will make the training process ineffective/intractable, and deteriorate the final prediction performance. Most of the previous works use the strategies of randomly sampling the object pairs [27], or design simple criteria to filter some object pairs [12, 1], or use fewer object proposals [14, 29]. They are far from satisfactory.

In this work, we propose a semantics guided graph relation neural network (*SGRN*) to extensively exploit the dependence of objects and relations for effective visual relationship detection. In our model, semantic embedding are introduced to complement the visual appearance information to (1) enable the source-target class aware visual feature refinement by transforming the features of the source domain to the target domain; (2) help filter the redundant object pairs; (3) enable effective predicate prediction through awareness of "subject" and "object" respectively.

Our **contributions** are five-fold.

- We propose a new semantics guided graph relation neural network (SGRN). The important semantic information is extensively exploited for enhancing both object detection and relation detection.

- We propose a source-target-aware transformation that transforms the features of the co-occurent objects to the target object domain to refine the visual features. Similarly, source-target cognoscitive transformations are used to refine features of objects from features of relations, and vice versa. Both the performance of object detection and relationship recognition are significantly improved.

- We propose a semantics guided relation proposal subnetwork (SRePN) to remove the object pairs with low semantic relatedness to reduce the redundancies and result in a sparsely connected graph, which facilitates the training and relationship detection.

- For relation detection, besides the visual features, we introduce the object and subject class embedding to explicitly provide semantic dependence between object categories and predicate types.

- We extensively explore the effectiveness of embedding semantics in the joint inference models for object detection and relationship recognition. The ablation studies of semantic embedding is a reference for future studies in this field.

2. Related Work

As an important topic of scene understanding, visual relationship detection has attracted increasing attention in recent decades. In recent years, deep learning technologies facilitate more accurate detection of objects as well as visual relationships [20, 1, 15].

Context for Visual Reasoning. There is rich context for entities in the real-world visual scene. It is a piece of important information to improve the visual prediction performance[10, 21, 6, 18]. There is also rich semantic inter-dependence between object categories and predicate types, which is an important cue for accurate visual relationship detection. The joint inference models for scene graph generation are designed to learn the semantic dependence and exploit the context between object proposals and phrases [29, 12, 14, 1, 13, 30]. Dai *et al*. [1] propose a CRF-like model [11] to exploit the statistical dependencies between objects and their relationships and refine the predicted labels. Li *et al*. [14] learn the context from region captions for scene graph generation. Xu *et al*. [29] use two sub-graphs to process objects and relationships respectively, and messages are propagated between these sub-graphs to collect the context. The graph-based message passing models are used in [13, 30] to learn the semantic dependence between objects and relationships and refine their representation in order to improve the object and predicate detection accuracy. Attention mechanism is used in their models to control the amount of information collected from other entities. Zellers *et al*. [32] utilize a set of LSTM models [5] to explore the motifs (*i.e*. regularly occurring graph structures). Interestingly, they built a strong baseline which directly predicts relationships using frequency priors from the dataset.

In our work, we propose SGRN to extensively exploit the semantic dependence among objects (nodes) and relationships (edges). Specifically, our model passes messages

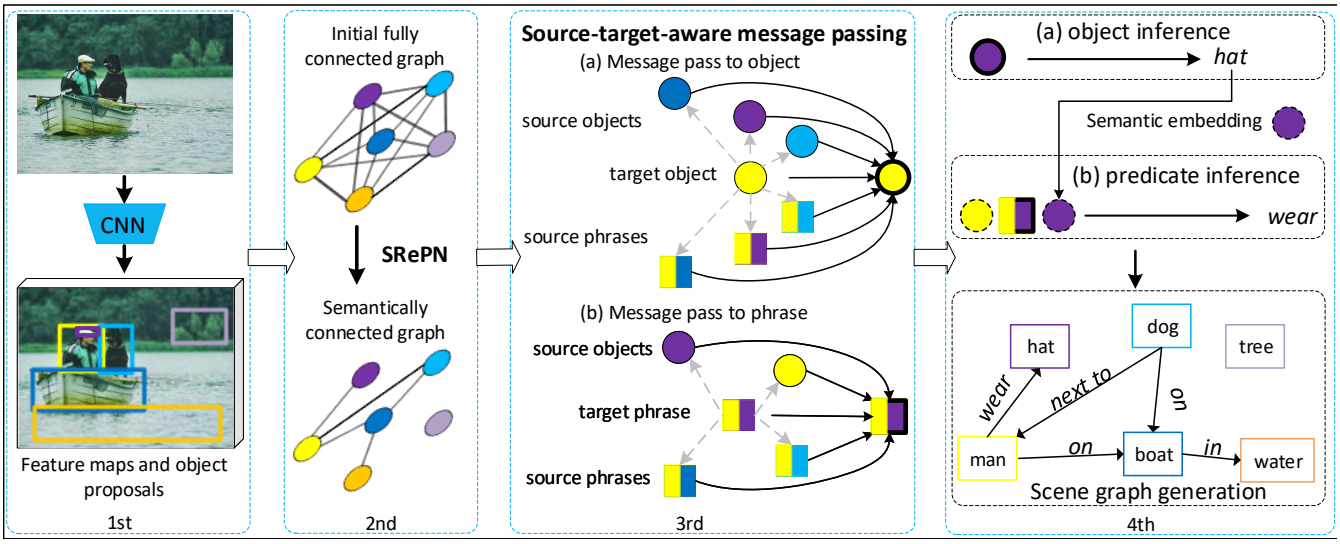

Figure 2: Overview of our SGRNN. (1) Object detection. We use Faster R-CNN for object detection. (2) Relation filtering. A fully-connected graph is formed by associating any two objects as a possible relationship. We remove the connection between a pair of objects that are semantically weakly dependent through our SRePN. (3) Source-target-aware refinement. To refine features, source-target-aware message passing is performed by exploiting contextual information from the objects and relationships that the target entity (object or relationship) is semantically correlated with for feature refinement. (4) Object and relation prediction. Objects are predicted from the refined object features. The predicates are inferred based on the the refined relationship features, and the separate embedding of object and subject probabilities. Note that solid circles and squares refer to object and relationship/phrase features respectively. The colors indicate different entities. Dashed arrows indicate that the target object or phrase is used together with the source object/relation for the source-target-aware transformation to the target object or phrase domain.

in a carefully designed *source-target-aware* manner. We are the first to include both the source and target features and semantics as the input of transformation function to enable effective message passing.

Semantic Embedding for Visual Tasks. Embedding technologies are popular in NLP communities and have been resorted to for visual tasks in recent decades, such as image caption [8, 7] and image retrieval [4]. Motivated by the success of embedding technologies, some works attempt to learn embedding for visual relationship detection. Semantic word embedding is used to explore the language priors between objects in order to improve the relation prediction accuracy [20, 16, 31, 37]. Zero-short visual relationship detection can also benefit from the language priors [20, 35]. Zhang *et al*. [33] learn relation translation vectors from visual triples by embedding object and subject respectively. To deal with the appearance variation of visual relations, some works learn the visual phrase embedding [35, 24]. All these works either directly use language priors in semantic word embedding or learn visual embedding.

In our work, we leverage the embedding of object class information for effective relation proposal, source-target-aware passage passing, and predicate prediction to explicitly use the available predicted class information.

Object Pairs Proposal. To handle the intractable number of possible pairwise object combinations, Dai *et al*. [1] use a simple filter to remove many of the unnecessary object pairs. Li *et al*. [13] cluster the phrase regions into some important ones and pass messages between them. The most related work to ours is [30] which also proposes a relation proposal network to estimate the relatedness of each object pair based on the predicted class probabilities but without semantic embedding. Different from their work, our SRePN uses semantic embedding to choose the most semantically inter-dependent object pairs.

3. Semantics Guided Graph Relation Neural Network

An overview of the proposed framework *Semantics Guided Graph Relation Neural Network* (SGRN) is shown in Fig. 2. Given an image, our model generates a scene graph by jointly reasoning the object categories and their relationships. More precisely, (1) objects are detected via Faster R-CNN [26]. (2) A fully connected graph is built with all the detected objects as nodes, and the possible relationships of each pair as edges. Then, our SRePN estimates the semantic dependence of each object pair and retains the

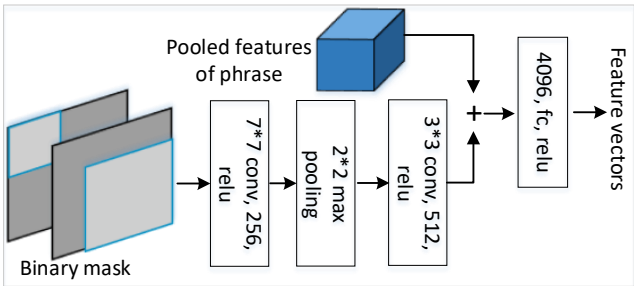

Figure 3: Illustration of the process of extracting the representation of relationship from its visual appearance features and the spatial information.

ones being semantically highly dependent, *i.e*. those that are likely to have meaningful relations. (3) Source-target-aware message passing is applied to exploit contextual information from the objects and relationships that the target entity (object or relationship) is semantically correlated with. (4) The classifiers learn to recognize object categories using the refined features. The predicates are inferred based on the refined relationship features, and the separate embedding of object and subject probabilities. In the following, we will discuss each component of our framework in details.

3.1. Proposals of Objects and Relationships

In our model, the N objects are detected by Faster R-CNN [26] from an input image. Each detected object o_i is associated with a bounding box $b_i = [x_i, y_i, w_i, h_i]$ that indicates the center coordinate (x_i, y_i) of the box and its width w_i and height h_i, its initially predicted class distribution over all C classes $p_i^o \in \mathbb{R}^{1\times C}$, and its pooled visual feature vectors x_i^o. Then, a scene graph is initially constructed with the n detected objects as nodes and $\mathcal{O}(n^2)$ edges between all object pairs, as shown in Fig. 2 (2nd block).

However, as discussed in previous sections, only a small fraction of object pairs among all the $\mathcal{O}(n^2)$ pairs are likely to have meaningful relationships. A large number of object pairs will make the training and inference intractable, and the redundant relationship proposals will deteriorate the recall performance. To reduce the number of object pairs, we propose a semantics guided relation proposal network (SRePN) which is trained to estimate the semantic dependence of an object pair based on the real-world regularities. In this paper, we exploit the semantic dependence between a pair of objects using word embeddings learned by a Word2vec model [23] trained on the region caption annotations of *visual genome* [9] dataset. The embedding matrix is denoted as W_e, where each row is an embedding vector for an object class. Then, the semantic embedding representation of o_i is obtained as:

$$e_i^o = p_i^o \cdot W_e, \tag{1}$$

which yields a *soft embedding*. Compared with the *hard embedding*, which takes the embedding vector corresponding to the class that has the highest predicted probability, *soft embedding* considers the uncertainty of the object class prediction given by the generic Faster R-CNN. It is capable of alleviating the negative effect introduced by the errors of the object classifier. Then, for an object pair (o_i, o_j), their relationship is represented as the concatenation of the subject embedding, spatial information, and object embedding: $x_{ij} = [e_i^o, \tilde{b}_i, \tilde{b}_j, e_j^o]$. $\tilde{b}_i$ is the bounding box parameters of o_i renormalized with respect to the union box of (o_i, o_j). A multi-layer perceptron (MLP) takes as input x_{ij}, and outputs an estimated semantic dependence score which ranges from 0 to 1 regularized by a sigmoid function. The score indicates how possible it is for the object pair (o_i, o_j) to have a meaningful relationship. Then, the object pairs, whose semantic dependence score falls in the top K scores and is larger than a threshold, are selected. Different from [12, 30], we do not apply non-maximal suppress operation (NMS) on the object pairs to reduce the number of relationship proposals because NMS will decrease the recall rate of the relationship proposals, which matters for the final evaluation metric in recall of relationship detection. Instead, we use K to limit the maximum number of relationship proposals in order to improve the training effectiveness and use a score threshold to reduce the redundancy. The set of selected relationships are denoted as $R = [r_1, \ldots, r_K]$.

Representation of Phrase. The visual appearance information is the most important information for visual tasks, and the spatial configuration of subject and object reflects their relationship directly. Therefore, a relationship should be represented by its visual appearance information and the relative spatial information of the involved pair of objects together. This kind of representation is extracted as shown in Fig. 3. First, the relative spatial information is represented by using a two-layer binary mask which indicates the positions of subject (first layer) and object (second layer) within the union bounding box. On each layer, the pixels that are within the bounding box of the object are assigned 1, otherwise 0. Then, the mask is fed to an MLP to learn the relative spatial features. The spatial features are added to the visual appearance features, and then a fully connected layer is used to fuse them sufficiently to obtain the phrase representation x^r.

3.2. Source-taregt-aware Refinement

We have achieved a graph including detected objects O (nodes) and selected meaningful phrases (edges) R. Each object has visual appearance features x^o and semantic embedding e^o obtained by Eq. (1). Each phrase has the representation x^r. From the graph, we also know that an object can be in multiple relationships with other objects and each relationship involves two objects.

Our message passing principle is based on the assumption that the objects which have relationships are semantically strongly dependent, and the relationships which have overlap object(s) are also semantically related to each other. For instance, as shown in Fig. 2 (2nd block), the yellow node has a relationship with the purple, blue and cyan ones respectively. The relationship between yellow and purple nodes overlaps with the relationships of between yellow and blue nodes, and yellow and cyan nodes. It shares the yellow node with the other two relationships. The phrases contain context information for each entity within it, and objects are the core components of phrases. Our message passing manner is unlike, *e.g.* [13, 30], which transforms the source object feature with ignorance of the target object, *i.e.* takes only the source object as the input of transformation function, we take both the source and target object features as the input of transformation function.

Pass Message to Target Objects. We denote that an object o_i is involved in a set of phrases R_{o_i}, and it has meaningful relationships with the set of objects O_{o_i}. Given a target object o_i, our SGRN learns to collect context from other objects in O_{o_i}, and phrases in R_{o_i} to refine x_i^o and enhances its representation ability. More formally, this message passing is defined as:

$$\hat{x}_i^o = \sigma\Bigg(x_i^o + \frac{1}{2\|O_{o_i}\|}\sum_{O_{o_i}}\big(f^{(o\to o)}([x_i^o, e_i], [x_j^o, e_j]) + \frac{1}{2\|R_{o_i}\|}\sum_{R_{o_i}} f^{(r\to o)}(x_i^o, x_j^r)\big)\Bigg), \quad (2)$$

where $\sigma()$ is an activation function ReLU, "[,]" is the feature concatenation operation, and "$\|.\|$" is the number of the entities of the set. $f^{(o\to o)}$ denotes the message passing function from other objects O_{o_i} to the target object o_i, while $f^{(r\to o)}$ denotes the message passing function from phrases R_{o_i} to o_i. Each $f()$ is an individual MLP. The process of this message passing targeting on an object is illustrated in Fig. 2 (3rd block (a)). The sum of transformed features is averaged over the number of sources to limit the magnitude of the value of the features. Division of two means that objects and relationships contribute equally to the target object. The target's features are added to ensure that the main information of the target is reserved and not overwhelmed by the transformed information from others. $f(a, b)$ can be mathematically understood as a conditional transformation operation: conditioned on target a, extract information from the sources b, *i.e. source-target-aware* message passing. Particularly, for the second term $f^{(o\to o)}$, target and sources are represented by their visual appearance features and semantic embedding features together. Therefore, $f^{(o\to o)}$ fully exploits the semantic dependence between objects using visual information learned by the CNN and the language priors in the semantic embedding. Our source-target-aware message passing is essentially different from the other message passing works, such as [13, 30]. They attempt to learn a shared projection matrix to project the source features to a shared target domain disregard what the target is. In other words, a source will be projected to the same representation even the target is different. For instance, as shown in Fig. 2 (1st block), to refine the features of "boat" and "hat" from the source of "water", "water" provides identical information to both of them. Then, they use attention approaches to compute weights to determine how much information will be collected from other objects to the target.

Intuitively, our method of message passing formally described by Eq. (2) also can be described as to learn the residual representation of x_i^o, which is expected to be the context. Therefore, the representation $\hat{x}_i^o$ contains its visual appearance information and the contextual information around it.

Pass Message to Target Phrases. The representations of phrases are also refined in a similar approach as below:

$$\hat{x}_m^r = \sigma\Bigg(x_m^r + \frac{1}{2\|R_{r_m}\|}\sum_{R_{r_m}} f^{(r\to r)}(x_m^r, x_n^r) + \frac{1}{2\|O_{r_m}\|}\sum_{O_{r_m}} f^{(o\to r)}(x_m^r, x_j^o)\Bigg) \quad (3)$$

where R_{r_m} is the set of phrases, which have shared objects with r_m. O_{r_m} is the object pairs involved in phrase r_m. $f^{(r\to r)}$ and $f^{(o\to r)}$ are the source-target-aware message passing to the target phrase from the source phrases and objects, respectively. They also consist of MLP. The process of this message passing targeting on a phrase is illustrated in Fig. 2 (3rd block (b)).

For a phrase r_m, the most important information is the participating pair of objects, and the other objects affect the recognition of the relationship negligibly. For instance like us human, when we attempt to understand the interaction of two objects before us, we focus on the two objects but not others. Therefore, O_{r_m} only contains the pair of objects. The other phrases that overlap with r_m provide context.

3.3. Relationship Recognition

An object classifier is trained to classify object category using the refined features $\hat{x}^o$, and output the predicted distribution $\hat{p}^o$. The object is assigned with the label with the highest predicted probability. With the predicted objects labels, we need to recognize the predicate types. To improve the accuracy of relationship recognition, we embed the pair objects similar to Eq. (1). Here we study two embedding approaches: i) two embedding matrices are learned for "subject" and "object" respectively, formally defined as Eq. (4); ii) the pair of objects share an embedding matrix,

Method	SGGen			SGCls			PredCls		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
MESSAGE PASSING [29]		3.4	4.2		21.7	24.4		44.8	53.0
Graph R-CNN [30]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
MSDN [14]	-	10.7	14.2	-	24.3	26.5	-	65.2	67.1
ASSOC EMBED [22]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4
MotifNet [32]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
MotifNet†	13.7	17.4	19.7	25.4	31.7	34.8	47.4	59.8	65.2
SGRN (w/o fine-tune, w/o SRePN)	19.5	24.2	29.0	30.2	33.8	35.7	57.3	62.4	65.5
SGRN (w/o SRePN)	22.4	29.3	33.1	34.2	37.3	38.3	**61.8**	**66.5**	**67.7**
SGRN (w/o SRePN) + freq	21.0	28.4	31.1	32.7	35.1	35.8	61.5	66.0	67.6
SGRN	**23.8**	**32.3**	**35.4**	**35.1**	**38.6**	**39.7**	60.3	64.2	66.4

Table 1: Comparison with existing methods on Visual Genome test set [29]. All numbers in %. The results of **ASSOC EMBED** [22], **MP** [29], and **Motifnet** [32] are taken from [32]. They are reimplemented in [32] using the authors' object detection backbone, and output better results than their original reports. **Motifnet**† is our reimplementation of the full code provided by the authors. The results of the last 4 rows (all about SGRN) are achieved by using the same object detection backbone as **Motifnet**†. This basic object detection backbone (without being jointly trained with our SGRN) has the object detection performance of mAP 16.6% (COCO metric).

formally defined as Eq. (5).

$$e^{sub} = \hat{p}^o_{sub} \times W^{sub}_{emb},\ e^{obj} = \hat{p}^o_{obj} \times W^{obj}_{emb}, \quad (4)$$

$$e^{sub} = \hat{p}^o_{sub} \times W_{emb},\ e^{obj} = \hat{p}^o_{obj} \times W_{emb}. \quad (5)$$

Note that the embedding matrix W^{sub}, W^{object}, W_{emb} are randomly initialized, which is different from the embedding matrix W_e in Eq. (1) initialized from a pretrained Word2vec. A predicate classifier is trained to recognize the predicate types by using the predicate features $x^p = [e^{sub}, \hat{x}^r, e^{obj}]$. The relationship is assigned with the predicate label with the highest predicted probability.

4. Experiments

In this section, we will introduce the experimental settings and implementation details of SGRN. Additionally, ablation studies will be conducted to validate the effectiveness of different parts of the framework. Our SGRN will be compared with state-of-the-art methods in terms of accuracy of object detection and relationship recognition.

4.1. Datasets

Our proposed method is evaluated on the Visual Genome dataset [29] that is the standard large-scale datasets used for visual relationship detection and scene graph generation [29, 14, 32]. However, the data preprocessing strategies and data split are inconsistent in different works. In our experiments, we use the most commonly used data split and data preprocessing proposed by [29]. In their data preprocessing, the most-frequent 150 object categories and 50 predicate types are selected. The whole dataset is split into training ($75,651$ images) and test ($32,422$ images) set.

4.2. Implementation Details

Faster R-CNN [26] with a VGG16 [28] backbone using the codebase in [32] is implemented as our underlying detector and visual feature extractor. The input images are scaled and then zero-padded to be 592×592. ROI-pooling [3] is applied to extract features of object and phrase from the shared feature maps output by the backbone network. Then the pooled object features are fed to two FC layers which generate a 512-d feature vector. Pooled phrase features are fused with the spatial configuration to learn a 4096-d relationship feature vector, as shown in Fig. 3. The embedding matrix used in the RePR module is initialized with the 300-d Word2vec provided by [20]. A two-layer MLP is used in the RePR module and outputs a 1-d vector which is then go through a sigmoid function to squash the predicted score in $(0, 1)$. RePR retains at most 256 phrases, and its threshold used to select the relationships in SRePN is empirically set as 0.55. All MLPs in the message passing module consist of two FC layers which generate 512-d feature vectors for objects and 4096-d feature vectors for relationships, respectively. The loss is the sum of the cross entropy for predicates and cross entropy for objects predicted by their own classifiers. SGD ($lr = 5 \times 10^{-3}$) with momentum is applied to optimize the network parameters.

4.3. Metrics

We follow three standard evaluation modes. (1) **Predicate classification** (PredCls): given an image associated with the ground truth bounding boxes and labels of objects, predict predicate labels. (2) **Scene graph classification** (SGCls): given an image associated with the ground truth

Model	MP	ObjE	PredE	Detection	SGGen		SGCls		PredCls	
				mAP	R@50	R@100	R@50	R@100	R@50	R@100
1	-	-	-	16.6	12.7	15.9	26.6	27.4	52.4	54.1
2	✓	-	-	18.5	17.1	19.9	29.7	32.4	58.3	60.4
3	✓	✓	-	20.2	24.7	27.1	33.0	35.1	62.0	64.2
4a	✓	-	✓	18.7	19.2	22.5	30.1	32.8	61.6	64.5
4b	✓	-	✓	18.7	21.5	24.6	31.4	33.8	62.9	65.1
5	✓	✓	✓	**20.8**	**29.3**	**33.1**	**37.3**	**38.3**	**66.5**	**67.7**

Table 2: Ablation studies on SGRN. All numbers in %. **ObjE** denotes the semantic embedding of object during message passing. **PredE** denotes the semantic embedding of subject and object of a relationship. Both of **ObjE** and **PredE** use soft embedding. **MP** denotes whether to pass message between objects and predicates. Model 4a uses Eq. (5) embedding method while model 4b uses Eq. (4) ones. The **Detection** reports the object detection performance (mAP) following COCO metrics [17]. SRePN is not utilized here.

bounding boxes of objects, predict object labels and recognize their relations. (3) **Scene graph detection** (SGDet): given an image, detect objects and recognize their relations. Only when the labels of the subject, predicate, and object of a detected relationship match the ground truth annotation, and the bounding boxes of subject and objects have more than 50% IoU with the ground truth bounding boxes simultaneously, this detection is counted as correct. Following most of the existing works, recall@K is used to evaluate the performance of relationship detection. In addition to most commonly used $R@50$ and $R@100$, we also use the more challenging $R@20$ for a more comprehensive evaluation.

4.4. Results and Comparison

The experimental results are presented in Tab. 1 [1]. Because different works use different data split on visual genome, we compare our proposed SGRN with several recent works, which use the data split [29], including Iterative Message Passing (IMP) [29], Multi-level scene Description Network (MSDN) [14], and the state-of-the-art MotifNet [32] and Graph R-CNN [30]. The results of IMP, MSDN are taken from the paper [32]. The authors reimplemented these methods using their own detector and got better performance. For a fair comparison, we use the same detector provided by [32] in our model. Results of Graph R-CNN are taken from the original paper [30].

Our reimplementation of **MotifNet** results in lower performance than that are reported in the original paper, which means that the object detection backbone we use is at least not better than the one used in their implementation. From Tab. 1 we can see that **SGRN** (w/o fine tune & SRePN) outperforms other methods significantly in most of the evaluation modes, except the originally reported **MotifNet**. Here, the same object detection backbone as **MotifNet**† is implemented but is not jointly trained. The proposed SRePN is not implemented either. To limit the number of object pairs, we naively random sample at most 256 object pairs as the relationship proposals, so as in the experiments of the following two rows. Larger number of proposals makes the training ineffective/intractable because of computation ability. The improvements compared with other methods validates that our proposed source-target-aware message passing is effective in scene graph generation.

In Tab. 1, the last three rows are the results when the whole model is jointly trained till convergence. **SGRN** (w/o SRePN) have higher number than other methods in all evaluation modes consistently. It validates that our method learns powerful representations for object detection and relationship recognition, and what it has learned impacts the backbone network positively with respect to our task. When the proposed SRePN is implemented (the last row), it shows further improvements in **SGCls**, and especially in **SGGen**. The improvements verify the effectiveness of our proposed SRePN, which aims to reduce redundant relationship proposals. For **PredCls**, the model w/o SRePN performs better because SRePN outputs some false negative results. When applying the motif frequency prior [32] on the results predicted by our model, the performance decreases slightly. It indicates that the semantic embedding in our model is able to capture the co-occurent semantic dependence between objects and predicates that is similar to motif frequency prior [32]. However, the frequency prior is obtained by counting using the training dataset, which leads to a strong bias. Even though the learned semantic embedding also has the bias, it is fused into the whole framework as a complement of visual features.

4.5. Ablation Studies

Our proposed SGRNN consists of three important components: ReRN, source-target-aware message passing

[1] More quantitative results and qualitative examples are shown in supplementary material.

(MP), and semantic embedding. The effectiveness of the proposed SRePN has been discussed above. Therefore, we conduct ablation studies to validate how different components affect the final performance w/o SRePN. The results of ablation studies are presented in Tab. 2.

Source-Target Message Passing. Model 1 is the baseline, which directly uses Faster R-CNN as the object detector and to learn shared features. The predicates are predicted based on the concatenated pooled feature vectors of subject, phrase and object. Model 2 adopts the source-target message passing module to refine both of the object and relationship features. Note that the embedding approach is not used in model 2, *i.e.* the embedding items e_i and e_j are removed from Eq. (2). Compared with model 1, we can see that the object detection performance of model 2 increases 1.9% mAP. Model 2 gets significant improvements in the recall accuracy of the three evaluation modes. In particular, there are 5.9% and 6.3% gains on R@50 and R@100 respectively in the PredCls setting, which indicates that our proposed source-target message passing not only learns better representations of objects but also better representations of relationships. The overall improvements in SGGen and SGCls are introduced by better object detection and relationship prediction performance together.

Semantic Embedding of Objects. As discussed in Sec. 3.2, semantic embedding of objects is used as a complement of visual features and guides the message passing between objects. Model 3 adopts semantic embedding of objects on top of model 2, as formally defined in Eq. (2). By comparison, we can see that model 3 gets further improvement on object detection: 1.7% mAP gain. Benefits from better object detection performance, the performance of SGGen and SGCls are improved $3.3\% \sim 7.6\%$. It indicates that semantic embedding of objects can explicitly provide semantic dependence between objects, which associates the visual appearance features to help the model capture the contextual information and pass message between semantically dependent objects. Because of better message passing guided by the introduced semantic information, better representations of relationships are learned, as described in Eq. (3), and results in better PredCls performance.

Semantic Embedding for Representation of Predicates. To achieve better representations of predicates in order to improve the relationship prediction accuracy, we propose to semantically embed the subject and object into embedding vectors, and then concatenate these vectors with the phrase feature vectors as the representation of predicates, as discussed in Sec. 3.3. To validate the effectiveness of semantic embedding of the subject and object for predicate prediction, we adopt the embedding strategy of Eq. (4) on model 4b (*i.e.* two different embedding matrix are learned for subject and object respectively), and adopt the embedding strategy of Eq. (5) on model 4a. Compared with the results of model 2, we can see that both models 4a and 4b achieve improvements in the three evaluation modes. However, model 4b reports better number than model 4a. It indicates that using two individual embedding matrix, which are for subject and object respectively, is better to exploit the semantic dependence between objects and relationships. Note that the object detection performance is improved negligibly, which means the improvements of relationship detection are introduced by the semantic embedding of subject and object for predicates rather than by a better object detector. Especially, this semantic embedding module is a plug-in module and can be adopted by any visual relationship methods because it is based on the object detection results.

ObjE	PredE	SGGen		SGCls	
		R@50	R@100	R@50	R@100
h	h	26.1	30.4	33.2	35.6
h	s	26.9	31.7	34.8	36.8
s	h	28.2	32.3	35.2	37.6
s	s	**29.3**	**33.1**	**37.3**	**38.3**

Table 3: Ablation studies on different soft (s) and hard (h) embedding methods. Note that **PredCls** is equivalent to hard embedding, and there is no need to compare here.

Model 5 is the full model of our framework (as illustrated in Fig. 2). It reports the best performance on both object detection and relationship detection. It indicates that each of our proposed modules can be effectively combined together to improve the overall performance of the framework for scene graph generation.

Effectiveness of Different Embedding Strategies. As discussed in Sec. 3.1, there are two embedding strategies: soft embedding and hard embedding. To validate their difference, we experiment on different combination of embedding strategies on model 5 and present the outcomes in Tab. 3. We can see that soft embedding always gets better performance than hard embedding. We analyze the reason as follows. Soft embedding takes the predicted probabilities as the weights to sum the word vectors. It considers the uncertainty of prediction results and tolerates the error prediction to a certain degree. In contrast, hard embedding directly selects the embedding vector corresponding to the highest predicted probability as representation. When the prediction is incorrect, the embedding is also wrong. Therefore, soft embedding provides better performance than hard embedding in our model.

5. Conclusion

In this paper, we introduce a new model for scene graph generation - semantics guided graph relation neural net-

work (SGRN). Our model consists of a semantics guided graph relation neural network that effectively removes the pairs of objects that are semantically not highly dependent, and a source-target-aware message passing module that effectively propagates information across the graph and learns better representations of objects and relationships. We extensively explore the effectiveness of embedding approaches for visual relationship detection and therefrom provide a reference for future study. The experimental results show that our approach significantly outperforms the state-of-the-art methods for scene graph generation.

References

[1] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3076–3086, 2017. 1, 2, 3, 10

[2] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pages 3270–3277, 2014. 1

[3] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 6

[4] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, pages 6589–6598, 2017. 3

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[6] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018. 2

[7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 3

[8] A. Karpathy, A. Joulin, and L. F. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897, 2014. 3

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 4

[10] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, pages 239–253. Springer, 2010. 2

[11] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 2

[12] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, pages 1347–1356, 2017. 2, 4

[13] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *ECCV*, pages 346–363. Springer, 2018. 2, 3, 5, 10

[14] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, pages 1261–1270, 2017. 2, 6, 7

[15] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *ICCV*, pages 848–857, 2017. 2

[16] W. Liao, L. Shuai, B. Rosenhahn, and M. Y. Yang. Natural language guided visual relationship detection. *arXiv preprint arXiv:1711.06032*, 2017. 1, 3

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 7

[18] Y. Liu, R. Wang, S. Shan, and X. Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, pages 6985–6994, 2018. 2

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[20] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016. 1, 2, 3, 6

[21] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, pages 2441–2448, 2014. 2

[22] A. Newell and J. Deng. Pixels to graphs by associative embedding. In *NIPS*, pages 2171–2180, 2017. 6

[23] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 4

[24] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Detecting rare visual relations using analogies. *arXiv preprint arXiv:1812.05736*, 2018. 3

[25] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, pages 6517–6525, 2017. 1

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 3, 4, 6

[27] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752, 2011. 1, 2

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 6

[29] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 1, 2, 6, 7

[30] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Graph r-cnn for scene graph generation. In *ECCV*, pages 690–706, 2018. 2, 3, 4, 5, 6, 7

[31] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, pages 1974–1982, 2017. 1, 3

[32] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 2, 6, 7

[33] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540, 2017. 3

[34] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. M. Elgammal. Relationship proposal networks. In *CVPR*, volume 1, page 2, 2017. 10

[35] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny. Large-scale visual relationship understanding. *arXiv preprint arXiv:1804.10660*, 2018. 3

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1

[37] B. Zhuang, L. Liu, C. Shen, and I. Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, pages 589–598, 2017. 3

Appendix

This appendix to the paper "Exploring the Semantics for Visual Relationship Detection" provides more ablation studies and visualization results.

A. Iterations of Message Passing

For the source-target-aware message passing stage (*i.e.*, the third stage in Fig. 2 of the main manuscript), this message passing procedure can be done once or repeated several times. Tab. 4 presents the model's performance with respect to the number of iteration applied of the message passing. We can see that the message passing of 2 iterations outperforms that of 1 iteration. That enables the use of refined features for the second round message passing. Using larger than 2 (*i.e.*, 3) iterations helps very little and even brings a slight performance decrease on SGGen and SGCls. We suspect this phenomenon is caused by the diffusion of noise of the learned features which is common in message passing models [1, 34, 13]. However, there are some slight gains on PredCls mode in which the performance of object detection is isolated (*i.e.* groudtruth bounding boxes and labels of objects are used). We think that the learning of phrase representations are more difficult than that of individual object representations and more iterations are beneficial for phrase representation refinement. On the other hand, object representations are more susceptible to noise than phrase representations. The weaker object detection decreases the performance of SGGen and SGCls. We use 2 iterations in our final scheme.

B. Effect of Different Information Sources

For the source-target-aware message passing, we study the influences of different information sources to the performance with respect to object detection and relationship recognition respectively (as defined in Eq. (2) and Eq. (3) in the main manuscript).

Tab. 5 presents the experimental results of using different information sources for message passing. Model 0 is a baseline which uses Faster R-CNN as the object detector and uses SRePN to reduce redundant relationship proposals, and uses subject and object embeddings for predicate prediction. But Model 0 does not use any message passing. Note that this baseline Model 0 here is different from the baseline Model 1 in Tab. 2 of our main manuscript. Compared with Model 1 in Tab. 2 of our main manuscript, the gains achieved by Model 0 are attributed to using the SRePN and the embedding method in phrase representation for the predicate prediction. It also indicates the effectiveness of our proposed SRePN and the semantic embedding for phrase representation.

Message Passing to Target Phrase. We analyze the influences of using source objects and source phrases on the message passing to the target phrase for the target phrase feature refinement, respectively (as defined in Eq. (3) in the main manuscript). As shown in Tab. 5, for Model 0 to Model 3, the message passing to the target objects (Eq. (2) in the main manuscript) is not applied for clear comparisons. We have the following observations. By comparing Model 1-3 with Model 0, we can see that refinement of phrase features is not helpful for object detection. The improvements in the three evaluation modes of these three models are attributed to the better predicate representations that are refined by our message passing (to the target phrase). Compared with Model 2, which only uses source phrases (the third term in Eq. (3)) to refine the target phrase, Model 1 which only uses source objects (the second term in Eq. (3)) to refine the target phrase achieves better performance. It indicates that for predicate prediction, information of objects participating in the relationship is more important than the information from external relationships. The gain in Model 2 compared with Model 0 indicates that information of the semantically related phrases is also useful for predicate prediction. Model 3 uses information of both source objects and source phrases to refine the predicate features and achieves better performance.

C. Qualitative Results

Message Passing to Target Object. We study the influence of using source objects and source phrases on the message passing to the target object for the target object feature refinement, respectively (as defined in Eq. (2) in the main manuscript). As shown in Tab. 5, for Model 4 to Model 6,

IteNr.	Object Detection	SGGen			SGCls			PredCls		
	mAP	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
1	19.1	19.3	26.5	30.1	29.8	32.3	34.7	57.6	59.8	62.1
2	**20.8**	**23.8**	**32.3**	**35.4**	**35.1**	**38.6**	39.7	60.3	64.2	66.4
3	20.6	23.1	32.2	35.4	34.4	37.1	**40.1**	**60.5**	**65.7**	**67.8**

Table 4: Ablation study of the effect of the number of iterations of message passing. The full Model is implemented and the whole Model is jointly trained.

Model	T.Obj		T.Phr		Detection	SGGen		SGCls		PredCls	
	S.Obj	S.Phr	S.Obj	S.Phr	mAP	R@50	R@100	R@50	R@100	R@50	R@100
0	-	-	-	-	16.6	14.1	18.5	27.7	30.5	54.2	58.4
1	-	-	✓	-	16.7	14.8	19.7	28.7	32.0	59.7	62.8
2	-	-	-	✓	16.6	14.7	19.6	27.9	31.4	58.2	61.3
3	-	-	✓	✓	16.7	15.1	20.2	29.0	32.3	60.9	63.4
4	✓	-	-	-	19.8	25.5	28.1	32.5	36.8	55.8	59.7
5	-	✓	-	-	16.8	14.2	18.5	28.0	31.1	55.5	59.5
6	✓	✓	-	-	20.2	26.4	28.6	33.8	37.7	56.1	59.8
7	✓	✓	✓	✓	**20.8**	**32.3**	**35.4**	**38.6**	**39.7**	**64.2**	**66.4**

Table 5: Ablation studies on using different sources for message passing. We validate the influence of source objects and source phrase on refining features of the target object and target phrase respectively. **T.Obj** and **T.Phr** denote target object and target phrase respectively, while **S.Obj** and **S.Phr** denote source object and source phrase respectively. Both of Object embedding, and subject and object embeddings are applied (soft embedding). Messages are passed 2 iterations here. Model 0 is the baseline which uses SRePN, subject and object embeddings for predicate prediction but does not use message passing.

Model	ObjE		PredE		Detection	SGGen		SGCls		PredCls	
	init.	train	init.	train	mAP	R@50	R@100	R@50	R@100	R@50	R@100
8	w2v	x	w2v	x	20.7	29.2	32.8	36.2	37.4	62.2	63.9
9a	w2v	x	w2v	✓	20.7	29.3	33.1	36.4	37.3	62.7	64.4
9b	w2v	x	w2v	✓	20.7	31.8	34.9	**38.7**	39.7	63.7	66.0
10	rand	✓	rand	✓	18.3	22.4	25.4	31.9	34.8	62.7	64.2
11	w2v	✓	rand	✓	20.7	32.1	35.4	38.6	**39.8**	**64.3**	66.4
12	w2v	x	rand	✓	**20.8**	**32.3**	**35.4**	**38.6**	39.7	64.2	**66.4**

Table 6: Ablation studies on the initialization and training of the embedding matrices for object embedding (**ObjE**) (for message passing purpose), and subject/object embedding (**PredE**) (for predicate prediction purpose), respectively. **init.** denotes the initialization methods of embedding matrices: using pre-trained word2vector (**w2v**) or randomly initialized matrix (**rand**). **train** denotes whether to train the embedding matrices during training. If **PredE** is initialized by using pre-trained word2vector but is fixed during training, it is equivalent to embedding the subject and object using the same embedding matrix, *i.e*. Eq. (5). Random initialization of any embedding matrix without later training makes no sense in this task and then is not shown. Soft embedding is used for embedding and messages are passed in two iterations here. SRePN is utilized.

and Model 0, the message passing to the target phrases (Eq. (3) in the main manuscript) is not applied for clear comparisons. We have the following observations. Compared with Model 0, Model 4 and Model 6 achieve significant improvements on object detection accuracy (*i.e*. 3.2% and 3.6% in mAP), both of which use source objects (the second term in Eq. (2) in the main manuscript) to refine the features of the target objects. Compared with Model 0, Model 5 brings small improvements on object detection accuracy. It indicates that, the objects which are highly semantically depen-

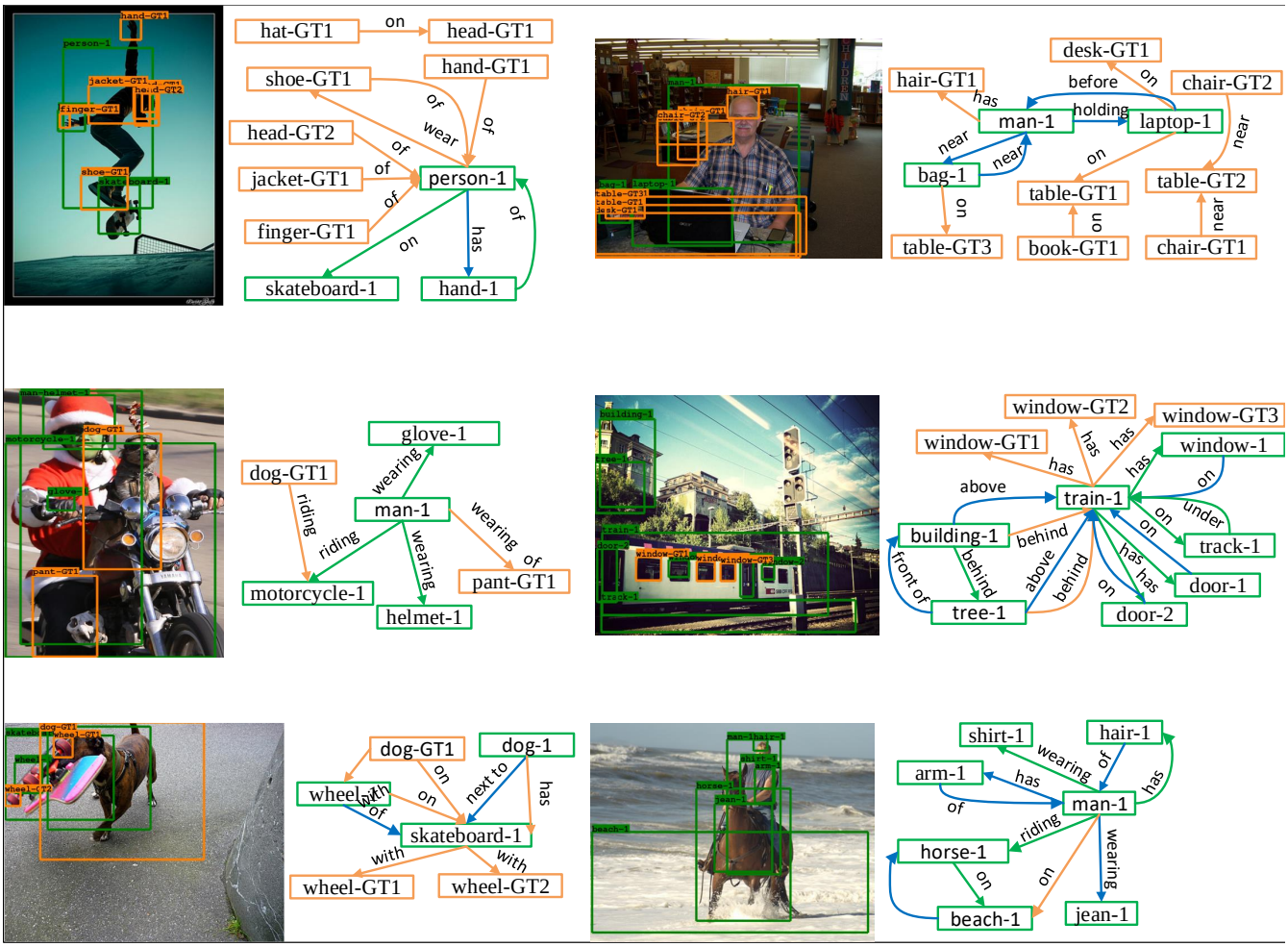

Figure 4: Visualization of relationship detection results from our SGRN. Green boxes denote the correctly detected objects while orange boxes denote the ground truth objects that are not detected. Green edges correspond to the correctly recognized relationships by our model at the R@20 setting, orange edges denote the ground truth relationships that are not recognized, and blue edges denote the recognized relationships that however do not exist in the ground truth annotations. Only predicted boxes that overlap with the ground truth are shown.

dent on the target object, provide more useful information than the semantically related phrases (third term in Eq. (2) in the main manuscript) to refine the features of the target object. Model 4 and 6 achieve significant improvements on SGGen and SGCls but small gain on PredCls. It indicates that the improvements from Models 4 and 6 are mainly brought by the improvements of object detection rather than by obtaining better predicate representations. Model 7 is our full Model which uses both of source objects and source phrase to refine the features of the target objects as well as the target phrases simultaneously and achieve the best performances.

Summary. These ablation study results validate that our proposed message passing method is capable of efficiently propagating information between objects and phrase and exploiting the contextual information among them effectively. The performance of object detection and relationships recognition mutually benefit from each other.

D. Influence of Different Embedding Initializations

For the proposed semantic embedding, we study the influences of different initialization methods for the embedding matrices to the performance with respect to object detection and relationship recognition respectively. Tab. 6 presents the experimental results of using different initialization methods for the embedding matrices of object embedding (ObjE) for message passing purpose, and the subject and object embedding (PredE) for predicate prediction

purpose.

Subject and Object Embedding (PredE) for Predicate Prediction Purpose. It is worth noting that Model 9a and Model 9b are the same (uses pretrianed word2vector (w2v) to initialize both ObjE and PredE) but they result in different performances. Model 9a has similar performance as Model 8, which uses one shared embedding matrix for subject and object of a relationship (*i.e*. Eq. (5)), while Model 9b performs comparably as Model 12, which uses two separate embedding matrices for the subject and object respectively. In fact, we often archive different experimental results when initializing PredE using w2v: system is unstable. Model 9a is the worst we have archived under this setting while Model 9b is the best. We analyze the results as follow. When PredE is initialized using w2v, the two sets of parameters (for subject and object respectively) are similar to each other. During training, these two sets of embedding parameters are updated in a similar tempo and gradient direction. Then, we often get two similar embedding matrices. It is equivalent to that we use a shared embedding matrix for both subject and object. That is why Model 9a has a similar performance as Model 8.

It achieves better performance when using random initialization (*e.g*. comparing Model 8 or Model 9a with Model 12). This is because the semantic embedding in PredE is to embed the subject and object to their corresponding semantic "role" in their relationship rather than use the language prior in w2v.

Object Embedding (ObjE) for Message Passing Purpose. By comparing Model 10 and Model 11, we can see that the randomly initialized ObjE has much worse performance than that initialized using w2v. It is because the pre-trained word2vector contains semantic interdependence between object categories which could guide the message passing for the objects, as discussed in the main manuscript. The randomly initialized embedding matrix does not have any prior information and may be difficult to train.

Comparing Model 11 with 12, we can see that fixing the object embedding matrix has similar or slightly better performance. Fixing the object embedding matrix will preserve the language prior in the w2v which is important to guide the message passing to refine object features effectively.

In summary, using w2v for object embedding and fixing it during training is sufficient to guide the message passing for objects, and randomly initializing PredE could make the training stable and get better performances.

Examples of generated scene graphs from our approach are shown in Fig. 4. For each image, green box denotes the object is correctly detected while orange box denotes the missed detection. Green edges correspond to the correctly recognized relationships by our model, orange edges denote the ground truth relationships that are not recognized, and blue edges denote the recognized relationships that however do not exist in the ground truth annotations. With our proposed SRePN, semantic embedding and source-target-aware message passing, our model is able to generate scene graphs with high recall. However, for this task, we think there is still large space for further enhancing of the performance. First, we observe that the object detection performance is still not perfect even though our message passing solution can enhance it. This could limit the quality of the generated scene graphs. As the examples shown in the first row of Fig. 4, many objects are not detected and then only a partial of the scene graph is correctly built. Second, the quality of annotations would bring difficulty to the model training. For example, some objects are repeatedly annotated and some annotate classes are ambiguous (*e.g*. "wear" and "wearing", "with" and "has", "person" and "man").