

# Visual Relationship Detection with Visual-Linguistic Knowledge from Multimodal Representations

Meng-Jiun Chiou, Roger Zimmermann, *Member, IEEE*, Jiashi Feng *Member, IEEE*,

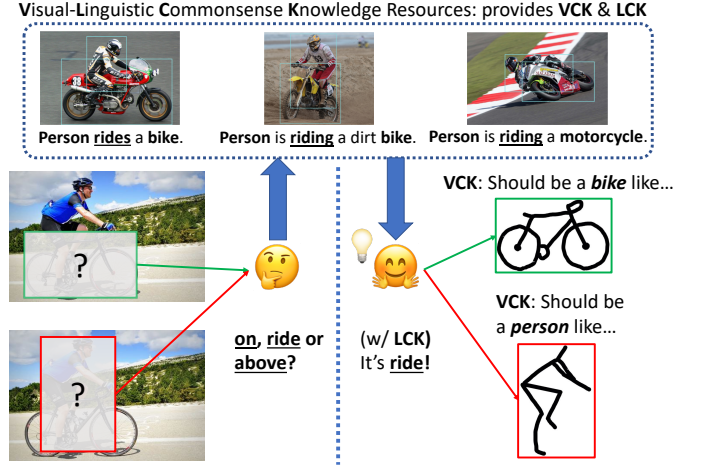
**Abstract**—Visual relationship detection aims to reason over relationships among salient objects in images, which has drawn increasing attention over the past few years. Inspired by human reasoning mechanism, it is believed that external visual commonsense knowledge is beneficial for reasoning visual relationships of objects in images, which is however rarely considered in existing methods. In this paper, we propose a novel approach named *Relational Visual-Linguistic Bidirectional Encoder Representations from Transformers* (RVL-BERT), which performs relational reasoning with both visual and language commonsense knowledge learned via self-supervised pre-training with multimodal representations. RVL-BERT also uses an effective spatial module and a novel mask attention module to explicitly capture spatial information among the objects. Moreover, our model decouples object detection from visual relationship recognition by taking in object names directly, enabling it to be used on top of any object detection system. We show through quantitative and qualitative experiments that, with the transferred knowledge and novel modules, RVL-BERT achieves competitive results on two challenging visual relationship detection datasets. The source code will be publicly available soon .

**Index Terms**—Visual Relationship Detection, Scene Graph Generation, Commonsense Knowledge, Multimodal Pre-training

## I. INTRODUCTION

**V**ISUAL relationship detection (VRD) aims to detect objects and classify triplets of subject-predicate-object in a query image. It is a very crucial task for enabling an intelligent system to understand the content of images, and has received much attention over the past few years [1]–[18]. Based on VRD, Xu et al. [19] proposed *scene graph generation* (SGG) [11], [20]–[31], which is essentially an interchangeable term of VRD and targets at extracting a comprehensive and symbolic graph representation for all visual relationships in an image, with vertices and edges denoting instances and relationships respectively. We use the term VRD throughout this paper for consistency. VRD is beneficial to various downstream tasks including but not limited to image captioning [32]–[34], visual question answering [35], [36], image synthesis [37], image retrieval [38], [39], etc.

To enhance the performance of VRD systems, some recent works incorporate the external *linguistic* commonsense knowledge from structured knowledge bases [15], raw language corpora [8], etc., as priors, which has taken inspiration from human reasoning mechanism. For instance, for a relationship



**Fig. 1:** Illustration of human reasoning over visual relationships with external visual and linguistic knowledge. With commonsense knowledge, a human is able to “guess” the visually blurred regions and prefer *ride* rather than *one* or *above*. **VCK:** Visual Commonsense Knowledge. **LCK:** Linguistic Commonsense Knowledge.

triplet case *person-ride-bike* as shown in Figure 1, with linguistic commonsense, the predicate *ride* is more accurate for describing the relationship of *person* and *bike* compared with other relational descriptions like *on* or *above*, which are rather abstract. In addition, we argue that the external *visual* commonsense knowledge is also beneficial to lifting detection performance of the VRD models, which is however rarely considered previously. Take the same *person-ride-bike* in Figure 1 as an example. If the pixels inside the bounding box of *person* are masked (zeroed) out, humans can still predict them as a *person* since we have seen many examples and have plenty of visual commonsense regarding such cases. This reasoning process would be helpful for VRD systems since it incorporates relationships of the basic visual elements; however, most previous approaches learn visual knowledge only from target datasets and neglect external visual commonsense knowledge in abundant unlabeled data. Inspired by the recent successful visual-linguistic pre-training methods (BERT-like models) [40], [41], we propose to exploit both *linguistic* and *visual* commonsense knowledge from Conceptual Captions [42] — a large-scale dataset containing 3.3M images with coarsely-annotated descriptions (alt-text) that were crawled from the web, to achieve boosted VRD performance. We first pre-train our backbone model (multimodal

Meng-Jiun Chiou and Roger Zimmermann are with the Department of Computer Science, National University of Singapore (email: {mengjiun,rogerz}@comp.nus.edu.sg). Jiashi Feng is with the Department of Electrical and Computer Engineering, National University of Singapore (email: elefjia@nus.edu.sg).

BERT) on Conceptual Captions with different pretext tasks to learn the visual and linguistic commonsense knowledge. Specifically, our model mines visual prior information via learning to predict labels for an image’s subregions that are randomly masked out. The model also considers linguistic commonsense knowledge through learning to predict randomly masked out words of sentences in image captions. The pre-trained weights are then used to initialize the backbone model and trained together with other additional modules (detailed at below) on visual relationship datasets.

Besides visual and linguistic knowledge, spatial features are also important cues for reasoning over object relationships in images. For instance, for A-on-B, the bounding box (or it’s center point) of A is often above that of B. However, such spatial information is not explicitly considered in BERT-like visual-linguistic models [40], [41], [43]. We thus design two additional modules to help our model better utilize such information: a mask attention module and a spatial Module. The former predicts soft attention maps of target objects, which are then used to enhance visual features by focusing on target regions while suppressing unrelated areas; the latter augments the final features with bounding boxes coordinates to explicitly take spatial information into account.

Moreover, our model is fairly flexible and can be placed on top of any object detection system. Previous VRD approaches are divided into two-stage and one-stage ones. The two-stage approaches perform object detection first, and feed detection results to a dedicated relationship classifier. For instance, Lu et al. [2] used pre-trained object detectors to generate object bounding boxes followed by performing predicate classification. On the other hand, the one-stage methods often achieve good performance by combining object detection and relationship classification. For example, Hung et al. [14] embedded entities and relationships in low-dimensional vector spaces and incorporated contextual information of the bounding boxes for simultaneous object detection and relationship classification. One-stage approaches usually achieve boosted performance by training with additional object classification loss, while they also suffer low flexibility in application as they require to re-train the whole model when migrating to different object detectors. In this work, We adopt two-stage design so that it can be flexibly cascaded with different state-of-the-art object detectors.

We integrate all above designs into a novel VRD model, named *Relational Visual-Linguistic Bidirectional Encoder Representations from Transformers* (RVL-BERT). RVL-BERT makes use of the pre-trained visual-linguistic representations as the source of visual and language knowledge to facilitate the learning and reasoning process on the downstream VRD task. It also incorporates a novel mask attention module to actively focus on the object locations in the input images and a spatial module to capture spatial relationships more accurately. Moreover, RVL-BERT is flexible in that it can be placed on top of any object detection model. We show through extensive experiments that the commonsense knowledge and the additional spatial and mask attention module effectively improve the model performance, and our RVL-BERT achieves competitive results on two VRD datasets.

## II. RELATED WORK

### A. Visual Relationship Detection

Visual relationship detection (VRD) is a task reasoning over the relationships between salient objects in the images. Recently, *linguistic knowledge* has been incorporated as guidance signals for the VRD systems. For instance, [2] proposed to detect objects and predicates individually with language priors and fuse them into a higher-level representation for classification. [5] exploited statistical dependency between object categories and predicates to infer their subtle relationships. Going one step further, [15] proposed a dedicated module utilizing bi-directional Gated Recurrent Unit to encode *external* language knowledge and a Dynamic Memory Network [44] to pick out the most relevant facts. However, none of these works consider external *visual* commonsense knowledge, which is also beneficial to relationship recognition. By contrast, we propose to exploit the abundant visual commonsense knowledge from multimodal Transformers [45] learned in pre-training tasks to facilitate the relationship detection in addition to the linguistic prior.

Recent one-stage methods achieve good performance by combining object detection and relationship classification. For example, [19] captured contextualized information between object proposals and relationships with graph neural networks, followed by classifying objects and relationships. [14] embedded entities and relationships in low-dimensional vector spaces and incorporated contextual information of the bounding boxes for simultaneous classification. However, these approaches suffer low flexibility in application as they require to re-train the whole model when migrating to state-of-the-art object detectors. In this work, based on BERT models [46] we design a VRD model that is flexible by taking in objects directly.

### B. Representation Pre-training

In the past few years, self-supervised learning which utilizes its own unlabeled data for supervision has been widely applied in representation pre-training. BERT, ELMo [47] and GPT-2 [48] are representative language models that perform self-supervised pre-training on various pretext tasks with either Transformer blocks or bidirectional LSTM. More recently, increasing attention has been drawn to multimodal (especially visual and linguistic) pre-training. Based on BERT, Visual-Linguistic BERT (VL-BERT) [43] pre-trains a single stream of cross-modality transformer layers from not only image captioning datasets but also language corpora. It is trained on BooksCorpus [49] and English Wikipedia in addition to Conceptual Captions [42]. We refer interested readers to [43] for more details of VL-BERT.

In this work, we utilize both visual and linguistic commonsense knowledge learned in the pretext tasks. While VL-BERT can be applied to training VRD without much modification, we show experimentally that their model does not perform well due to lack of attention to spatial features. By contrast, we propose to enable knowledge transfer for boosting detection accuracy and use two novel modules to explicitly exploit spatial features.

### III. METHODOLOGY

#### A. Revisiting BERT and VL-BERT

Let a sequence of  $N$  embeddings  $x = \{x_1, x_2, \dots, x_N\}$  be the features of input sentence words, which are the summation of *token*, *segment* and *position embedding* as defined in BERT [46]. The BERT model takes in  $x$  and utilizes a sequence of  $n$  multi-layer bidirectional Transformers [45] to learn contextual relations between words. Let the input feature at layer  $l$  denoted as  $x^l = \{x_1^l, x_2^l, \dots, x_N^l\}$ . The feature of  $x$  at layer  $(l+1)$ , denoted as  $x^{l+1}$ , is computed through a Transformer layer which consists of two sub-layers: 1) a multi-head self-attention layer plus a residual connection

$$\tilde{h}_i^{l+1} = \sum_{m=1}^M W_m^{l+1} \left\{ \sum_{j=1}^N A_{i,j}^m \cdot V_m^{l+1} x_j^l \right\}, \quad (1)$$

$$h_i^{l+1} = \text{LayerNorm}(x_i^l + \tilde{h}_i^{l+1}), \quad (2)$$

where  $A_{i,j}^m \propto (Q_m^{l+1} x_i^l)^T (K_m^{l+1} x_j^l)$  represents a normalized dot product attention mechanism between the  $i$ -th and the  $j$ -th feature at the  $m$ -th head, and 2) a position-wise fully connected network plus a residual connection

$$\tilde{x}_i^{l+1} = W_2^{l+1} \cdot \text{GELU}((W_1^{l+1} h_i^{l+1}) + b_1^{l+1}) + b_2^{l+1}, \quad (3)$$

$$x_i^{l+1} = \text{LayerNorm}(h_i^{l+1} + \tilde{x}_i^{l+1}), \quad (4)$$

where GELU is an activation function named Gaussian Error Linear Unit [50]. Note that **Q** (Query), **K** (Key), **V** (Value) are learnable embeddings for the attention mechanism, and **W** and **b** are learnable weights and biases respectively.

Based on BERT, VL-BERT [43] adds  $O$  more multi-layer Transformers to take in additional  $k$  visual features. The input embedding becomes  $x = \{x_1, \dots, x_N, x_{N+1}, \dots, x_{N+O}\}$ , which is computed by the summation of not only the token, segment and position embeddings but also an additional *visual feature embedding* which is generated from the bounding box of each corresponding word. The model is then pre-trained on two types of pretext tasks to learn the visual-linguistic knowledge: 1) *masked language modeling with visual clues* that predicts a randomly masked word in a sentence with image features, and 2) *masked RoI classification with linguistic clues* that predicts the category of a randomly masked region of interest (RoI) with linguistic information.

#### B. Overview of Proposed Model

Figure 2 shows the overall architecture of our proposed RVL-BERT. For the backbone BERT model, we adopt a 12-layer Transformer and initialize it with the pre-trained weights of VL-BERT for visual and linguistic commonsense knowledge. While based on VL-BERT, our model differs in several important aspects: 1) RVL-BERT explicitly arranges query object pairs in sequences of *subject-predicate-object* (instead of sentences in the original design) and receives an extra answer segment for relationship prediction. 2) Our model is equipped with a novel mask attention module that learns attention-guided visual feature embeddings for the model to attend to target object-related area. 3) A simple yet

effective spatial module is added to capture spatial representation of subjects and objects, which are of importance in spatial relationship detection.

Let  $N$ ,  $A$  and  $O$  denote the number of elements for the relationship linguistic segment, the answer segment, and the relationship visual segment, respectively. Our model consists of  $N + A + O$  multi-layer Transformers, which takes in a sequence of linguistic and visual elements, including the output from the mask attention module, and learns the context of each element from all of its surrounding elements. For instance, as shown in Figure 2, to learn the representation of the linguistic element *goose*, the model looks at not only the other linguistic elements (e.g., *to the right of* and *window*) but also all visual elements (e.g., *goose*, *window*). Along with the multi-layer Transformers, the spatial module extracts the location information of subjects and objects using their bounding box coordinates. Finally, the output representation of the element in the answer segment,  $h_{so}$ , is augmented with the output of the spatial module  $C_{so}$ , followed by classification with a 2-layer fully connected network.

The input to the model can be divided into three groups by the type of segment, or four groups by the type of embedding. We explain our model below from the segment-view and the embedding-view, respectively.

1) *Input Segments*: For each input example, RVL-BERT receives a relationship linguistic segment, an answer segment, and a relationship visual segment as input.

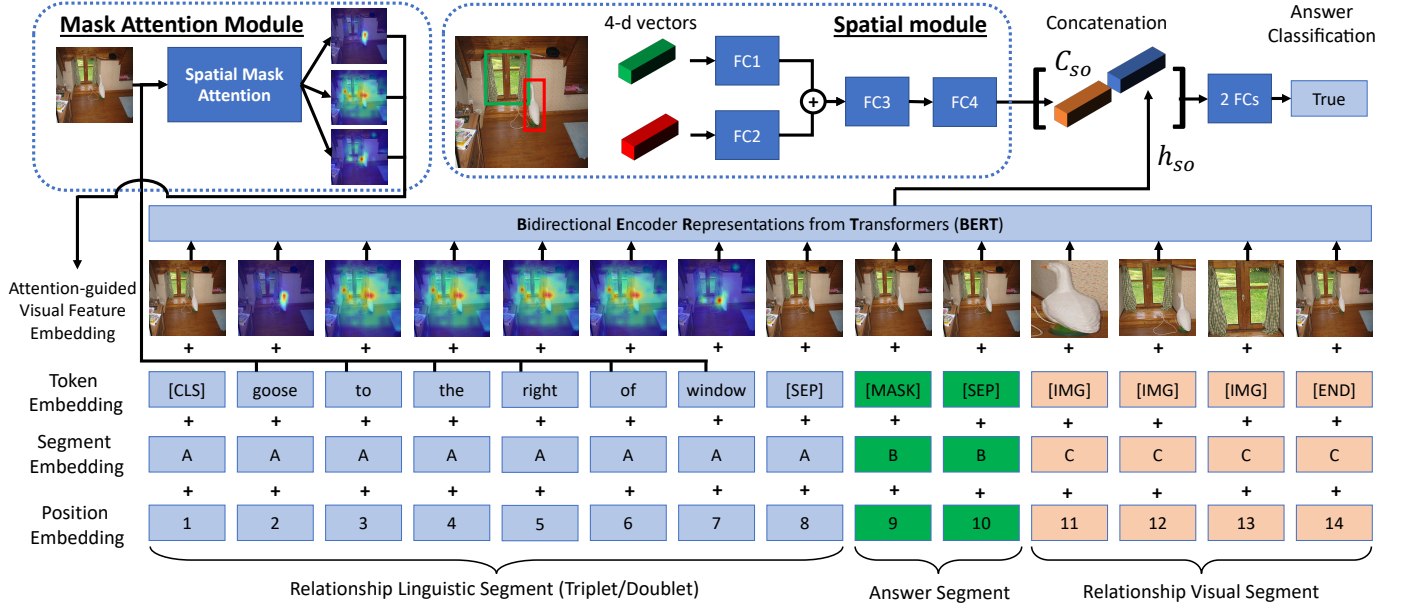
a) *Relationship linguistic segment* (light blue elements in Figure 2) is the linguistic information in a triplet form *subject-predicate-object*, like the input form of SpatialSense dataset [51], or a doublet form *subject-object* like the input in VRD dataset [2]). Note that each term in the triplet or doublet may have more than one element, such as *to the right of*. This segment starts with a special element “[CLS]” that stands for classification<sup>1</sup> and ends with a “[SEP]” that keeps different segments separated.

b) *Answer segment* (green elements in Figure 2) is designed for learning a representation of the whole input and has only special elements like “[MASK]” that is for visual relationship prediction and the same “[SEP]” as in the relationship linguistic segment.

c) *Relationship visual segment* (tangerine color elements in Figure 2) is the visual information of a relationship instance, also taking the form of triplets or doublets but with each component term corresponding to only one element even if its number of words of the corresponding label is greater than one.

2) *Input Embeddings*: There are four types of input embeddings: token embedding  $t$ , segment embedding  $s$ , position embedding  $p$ , and (attention-guided) visual feature

<sup>1</sup>We follow the original VL-BERT to start a sentence with the “[CLS]” token, but we do not use it for classification purposes.



**Fig. 2:** Architecture illustration of proposed RVL-BERT for SpatialSense dataset [51]. It can be easily adapted for VRD dataset [2] by replacing triplets subject-predicate-object with doublets subject-object and performing predicate classification instead of binary classification on the output feature of “[MASK]”.

embedding  $v$ . Among them, the attention-guided visual feature embedding is newly introduced while the others follow the original design of VL-BERT. We denote the input of RVL-BERT as  $x = \{x_1, \dots, x_N, x_{N+1}, \dots, x_{N+A}, x_{N+A+1}, \dots, x_{N+A+O}\}$ ,  $\forall x_i : x_i = t_i + v_i + s_i + p_i$  where  $t_i \in t$ ,  $v_i \in v$ ,  $s_i \in s$ ,  $p_i \in p$ .

a) *Token Embedding*. We transform each of the input words into a  $d$ -dimensional feature vector using WordPiece embeddings [52] comprising 30,000 distinct words. In this sense, our model is flexible since it can take in any object label with any combination of words available in WordPiece. Note that for those object/predicate names with more than one word, the exact same number of embeddings is used. For the  $i$ -th object/predicate name in an input image, we denote the token embedding as  $t = \{t_1, \dots, t_N, t_{N+1}, \dots, t_{N+A}, t_{N+A+1}, \dots, t_{N+A+O}\}$ ,  $t_i \in \mathbb{R}^d$ , where  $d$  is the dimension of the embedding. We utilize WordPiece embeddings for relationship triplets/doublets  $\{t_2, \dots, t_{N-1}\}$ , and use special predefined tokens “[CLS]”, “[SEP]”, “[MASK]” and “[IMG]” for the other elements.

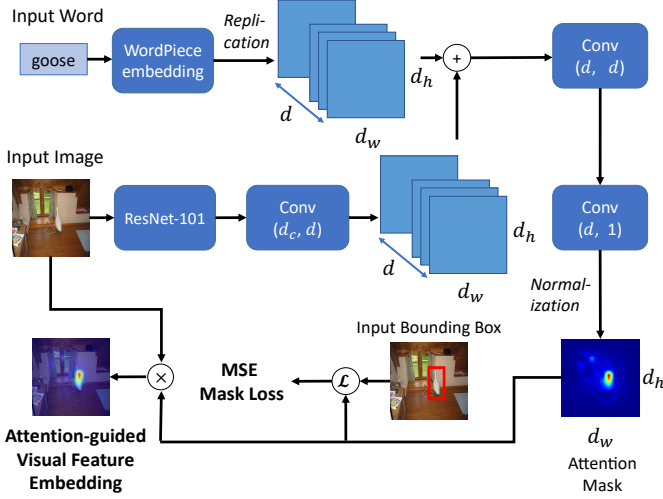
b) *Segment Embedding*. We use three types of learnable segment embeddings  $s = \{s_1, \dots, s_{N+A+O}\}$ ,  $s_i \in \mathbb{R}^d$  to inform the model that there are three different segments: “A” for relationship linguistic segment, “B” for answer segment and “C” for relationship visual segment.

c) *Position Embedding*. Similar to segment embeddings, learnable position embeddings  $p = \{p_1, \dots, p_{N+A+O}\}$ ,  $p_i \in \mathbb{R}^d$  are used to indicate the order of elements in the input sequence. Compared to the original VL-BERT where the position embeddings of the relationship

visual segment are the same for each RoI, we use distinct embeddings as our RoIs are distinct and ordered.

d) *Visual Feature Embedding*. These embeddings are to inform the model of the internal visual knowledge of each input word. Given an input image and a set of RoIs, a ResNet-101 [53] is utilized to extract the feature map, which is prior to the output layer, followed by RoI Align [54] to produce fixed-size feature vectors  $z = \{z_0, z_1, \dots, z_K\}$ ,  $z_i \in \mathbb{R}^d$  for  $K$  RoIs, where  $z_0$  denotes the feature of the whole image. For triplet inputs, we additionally generate  $K(K-1)$  features for all possible union bounding boxes:  $u = \{u_1, \dots, u_{K(K-1)}\}$ ,  $u_i \in \mathbb{R}^d$ . We denote the input visual feature embedding as  $v = \{v_1, \dots, v_N, v_{N+1}, \dots, v_{N+A}, v_{N+A+1}, \dots, v_{N+A+O}\}$ ,  $v_i \in \mathbb{R}^d$ . We let subject and object be  $s$  and  $o$ , with  $s, o \in \{1, \dots, K\}$ ,  $s \neq o$ , and let the union bounding box of  $s$  and  $o$  be  $so \in \{1, \dots, K(K-1)\}$ .

For the relationship visual segment  $\{v_{N+A}, \dots, v_{N+A+O-1}\}$  (excluding the final special element), we use  $z_s$  and  $z_o$  as the features of subject  $s$  and object  $o$  in doublet inputs, and add another  $u_{so}$  in between in case of triplet inputs. For the special elements other than “[IMG]”, we follow VL-BERT to use the full image feature  $z_0$ . However, for the relationship linguistic segment  $\{v_2, \dots, v_{N-1}\}$  (excluding the first and final special elements), it is unreasonable to follow the original design to use the same, whole-image visual feature for all elements, since each object/predicate name in the relationship linguistic segment should correspond to different parts of the image. To better capture distinct visual information for elements in relationship linguistic segment, we propose a *mask attention module* to learn to generate attention-guided visual feature embeddings that attend to important (related) regions, which is detailed at below.



**Fig. 3:** The pipeline of Mask Attention Module. Given an input image and the corresponding word embedding(s), the module generates an attention mask (heatmap) and outputs an attention-guided visual embedding.

### C. Mask Attention Module

An illustration of the mask attention module is shown in Figure 3. Denote the visual feature (the feature map before average pooling) used by the mask attention module as  $v_s \in \mathbb{R}^{d_c \times d_w \times d_h}$ , where  $d_c, d_w, d_h$  stand for the dimension of the channel, width, and height, respectively. To generate the feature for an object  $s$  (e.g., *goose* in Figure 3), the mask attention module takes in and projects the visual feature  $v_s$  and the word embedding<sup>2</sup>  $w_s$  into the same dimension using a standard CNN and a replication process, respectively

$$\tilde{v}_s = \sigma(W_1^T v_s + b_1), \quad (5)$$

$$w_s = \text{Replication}(w_s), \quad (6)$$

where  $\text{Replication}(\cdot)$  replicates the input vector of size  $d$  into the feature map of dimension  $d \times d_w \times d_h$ . The above is followed by element-wise addition to fuse the features, two convolutional layers as well as a re-scaling process to generate the attention mask  $m_s$

$$\tilde{m}_s = \sigma(W_2^T (\tilde{v}_s + w_s) + b_2), \quad (7)$$

$$m_s = \text{Norm}(W_3^T \tilde{m}_s + m_3), \quad (8)$$

where the min-max  $\text{Norm}(\cdot)$  applied to each element is defined by  $\text{Norm}(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)}$ . Note that in the above equations all of the  $W$ 's and  $b$ 's are learnable weights and biases of the convolutional layers, respectively. The attention-guided visual feature  $v_s^{\text{att}}$  is then obtained by performing Hadamard product between the visual feature and the attention mask:  $v_s^{\text{att}} = v_s \circ m_s$ . Finally,  $v_s^{\text{att}}$  is pooled into  $v_s^{\text{att}} \in \mathbb{R}^d$  to be used in  $\{v_2, \dots, v_{N-1}\}$ .

To learn to predict the attention masks, we train the module against the Mean Squared Error (MSE loss) between the mask

$m_s$  and the resized ground truth mask  $b_s$  consisting of all ones inside the bounding box and outside all zeros:

$$\mathcal{L}_{\text{mask}} = \frac{1}{d_w d_h} \sum_{i=1}^{d_w} \sum_{j=1}^{d_h} (m_s^{ij} - b_s^{ij})^2, \quad (9)$$

where  $d_w, d_h$  denote the width and length of the attention mask.

### D. Spatial Module

The spatial module aims to augment the output representation with spatial knowledge by paying attention to bounding box coordinates. See the top part of Figure 2 for its pipeline.

Let  $(x_i^0, y_i^0), (x_i^1, y_i^1)$  denote the top-left and bottom-right coordinates of a bounding box of an object  $i$  of an input image, and let  $w, h$  be the width and height of the image. The 4-dimensional normalized coordinate of an object  $i$  is defined by  $C_i = (x_i^0/w, y_i^0/h, x_i^1/w, y_i^1/h)$ . The spatial module takes in coordinate vectors of a subject  $s$  and an object  $o$ , and encodes them using linear layers followed by element-wise addition fusion and a two-layer, fully-connected layer

$$\tilde{C}_{so} = \sigma(W_4 C_s + b_4) + \sigma(W_5 C_o + b_5), \quad (10)$$

$$C_{so} = W_7 \sigma(W_6 \tilde{C}_{so} + b_6) + b_7. \quad (11)$$

The output feature  $C_{so}$  is then concatenated with the multi-modal feature  $h_{so}$  to produce  $f_{so}$  for answer classification:

$$f_{so} = [C_{so}; h_{so}]. \quad (12)$$

## IV. EXPERIMENTS

### A. Datasets

We first ablate our proposed model on VRD dataset [2], which is the most widely used benchmark. For comparison with previous methods, we also evaluate on SpatialSense [51] dataset. Compared with Visual Genome (VG) dataset [55], SpatialSense suffers less from the dataset language bias problem, which is considered a distractor for performance evaluation — in VG, the visual relationship can be “guessed” even without looking at the input image [22], [27].

1) *VRD*: The VRD dataset consists of 5,000 images with 37,993 visual relationships. We follow [2] to divide the dataset into a training set of 4,000 images and a test set of 1,000 images, while only 3,780 and 955 images are annotated with relations, respectively. For all possible pairs of objects in an image, our model predicts by choosing the best-scoring predicate and records the scores, which are then used to rank all predictions in the ascending order. Since the visual relationship annotations in this dataset are far from exhaustive, we cannot use precision or average precision as they will penalize correct detections without corresponding ground truth. Traditionally, Recall@K is adopted to bypass this problem and we follow this practice throughout our experiments. For VRD, the task named *Predicate Detection/Classification* measures the accuracy of predicate prediction given ground truth classes and bounding boxes of subjects and objects independent of the object detection accuracy. Following [2], [6], we use **Recall@K**, or the fraction of ground truth relations that are recalled in the top  $K$  candidates.  $K$  is usually set as 50 or 100 in the literature.

<sup>2</sup>Note that for object labels with more than one word, the embeddings of each word are element-wise summed in advance.



2) *SpatialSense*: SpatialSense is a relatively new visual relationship dataset focusing on especially spatial relations. Different from Visual Genome [55], SpatialSense is dedicated to reducing dataset bias, via a novel data annotation approach called Adversarial Crowdsourcing which prompts annotators to choose relation instances that are hard to guess by only looking at object names and bounding box coordinates. SpatialSense defines nine spatial relationships above, behind, in, in front of, next to, on, to the left of, to the right of, and under, and contains 17,498 visual relationships in 11,569 images. The task on SpatialSense is binary classification on given visual relationship triplets of images, namely judging if a triplet subject-predicate-object holds for the input image. Since in SpatialSense the number of examples of “True” equals that of “False”, the **classification accuracy** can be used as a fair measure. We follow the original split in [51] to divide them into 13,876 and 3,622 relations for training and test purposes, respectively.

### B. Implementation

For the backbone model, we use VL-BERT<sub>BASE</sub> (which is based on BERT<sub>BASE</sub><sup>3</sup>) that is pre-trained on three datasets including Conceptual Captions [42], BooksCorpus [49] and English Wikipedia. For extracting visual embedding features, we use ResNet-101 [53] based Fast R-CNN [56] (detection branch of Faster R-CNN [57]). We randomly initialize the final two fully connected layers and the newly proposed modules (*i.e.*, mask attention module and spatial module). During training, we find our model empirically gives the best performance when freezing the parameters of the backbone model and training on the newly introduced modules. We thus get a lightweight model compared to the original VL-BERT as the number of trainable parameters is reduced by around 96%, *i.e.*, down from 161.5M to 6.9M and from 160.9M to 6.4M when trained on the SpatialSense dataset and the VRD dataset, respectively. ReLU is used as the nonlinear activation function  $\sigma$ . We use  $d = 768$  for all types of input embeddings,  $d_c = 2048$  for the dimension of channel of the input feature map and  $d_w = d_h = 14$  for the attention mask in the mask attention module. The training loss is the sum of the softmax cross-entropy loss for answer classification and the MSE loss for the mask attention module. The experiments were conducted on a single NVIDIA Quadro RTX 6000 GPU in an end-to-end manner using Adam [58] optimizer with initial learning rate  $1 \times 10^{-4}$  after linear warm-up over the first 500 steps, weight decay  $1 \times 10^{-4}$  and exponential decay rate 0.9 and 0.999 for the first- and the second-moment estimates, respectively. We trained our model for 60 and 45 epochs for VRD and SpatialSense dataset, respectively, as there are more images in the training split of SpatialSense. For experiments on the VRD dataset, we followed the training practice in [14] to train with an additional “no relationship” predicate and for each image we sample 32 relationships with the ratio of ground truth relations to negative relations being 1 : 3.

<sup>3</sup>There are two variants of BERT: BERT<sub>BASE</sub> that has 12-layer Transformers and BERT<sub>LARGE</sub> that has 24-layer Transformers.

**TABLE I:** Ablation results for different losses of mask attention module and ways of feature combination. **.3**, **.5** and **.7** denote different  $\alpha$  values in  $f_{so} = \alpha C_{so} + (1 - \alpha)h_{so}$ .

MAM Loss		Feature Combination				Recall@50
BCE	MSE	.3	.5	.7	concat	
✓					✓	53.50
	✓				✓	<b>55.55</b>
	✓	✓				55.46
	✓		✓			54.74
	✓			✓		55.19

**TABLE II:** Ablation results of each module on VRD dataset (Recall@50) and SpatialSense dataset (Overall Acc.) **VL**: Visual-Linguistic Knowledge. **S**: Spatial. **M**: Mask Att.

Model	VL	Spatial	Mask Att.	R@50	Acc.
Basic				40.22	55.4
+VL	✓			45.06	61.8
+VL+S	✓	✓		55.45	71.6
+VL+S+M	✓	✓	✓	<b>55.55</b>	<b>72.3</b>

### C. Ablation Study Results

1) *Training Objective for Mask Attention Module*: We first compare performance difference between training the mask attention module (MAM) against **MSE** loss or binary cross entropy (**BCE**) loss. The first two rows of Table I show that MSE outperforms BCE by relative 3.8% on Recall@50. We also observe that training with BCE is relatively unstable as it is prone to gradient explosion under the same setting.

2) *Feature Combination*: We also experiment with different ways of feature combination, namely, element-wise addition and concatenation of the features. To perform the experiments, we modify Eqn. 12 as  $f_{so} = \alpha C_{so} + (1 - \alpha)h_{so}$ , and we experiment with different  $\alpha$  values (**.3**, **.5** and **.7**). The last five rows of Table I show that concatenation performs slightly better than addition under all  $\alpha$  values.

The setting in the second row of Table I empirically gives the best performance, and thus we stick to this setting for the following experiments.

3) *Module Effectiveness*: We ablate the training strategy and the modules in our model to study their effectiveness. **VL** indicates that the RVL-BERT utilizes the external multimodal knowledge learned in the pretext tasks via weight initialization. **Spatial (S)** means the spatial module, while **Mask Att. (M)** stands for the mask attention module. Table II shows that each module effectively helps boost the performance. The visual-linguistic commonsense knowledge lifts the **Basic** model by 12% (or absolute 5%) of Recall@50 on VRD dataset, while the spatial module further boosts the model by more than 23% (or absolute 10%). As the effect of the mask attention module is not apparent on the VRD dataset (0.2% improvement), we also experiment on the SpatialSense dataset (Overall Accuracy) and find the mask attention module provide a relative 1% boost of accuracy.

**TABLE III:** Performance comparison with existing models on VRD dataset. Results of previous methods are extracted from [2] and respective papers.

Model	Recall@50	Recall@100
Visual Phrase [1]	0.97	1.91
Joint CNN [2]	1.47	2.03
VTransE [6]	44.76	44.76
PPR-FCN [9]	47.43	47.43
Language Priors [2]	47.87	47.87
Zoom-Net [10]	50.69	50.69
TFR [11]	52.30	52.30
Weakly (+ Language) [7]	52.60	52.60
LK Distillation [8]	55.16	55.16
Jung et al. [12]	55.16	55.16
UVTransE [14]	55.46	55.46
MF-URLN [16]	58.20	58.20
HGAT [18]	<b>59.54</b>	<b>59.54</b>
<b>Ours: RVL-BERT</b>	55.55	55.55

#### D. Quantitative Results on VRD Dataset

We conduct experiments on VRD dataset to compare our method with existing approaches. **Visual Phrase** [1] represents visual relationships as visual phrases and learns appearance vectors for each category for classification. **Joint CNN** [2] classifies the objects and predicates using only visual features from bounding boxes. **VTransE** [6] projects objects and predicates into a low-dimensional space and models visual relationships as a vector translation. **PPR-FCN** [9] uses fully convolutional layers to perform relationship detection. **Language Priors** [2] utilizes individual detectors for objects and predicates and combines the results for classification. **Zoom-Net** [10] introduces new RoI Pooling cells to perform message passing between local objects and global predicate features. **TFR** [11] performs a factorization process on the training data and derives relational priors to be used in VRD. **Weakly** [7] adopts a weakly-supervised clustering model to learn relations from image-level labels. **LK Distillation** [8] introduces external knowledge with a teacher-student knowledge distillation framework. **UVTransE** [14] extends the idea of vector translation in VTransE with the contextual information of the bounding boxes. **MF-URLN** [16] uses external linguistic knowledge and internal statistics to explore undetermined relationships. **HGAT** [18] proposes a TransE-based multi-head attention approach performed on a fully-connected graph.

Table III shows the performance comparison on the VRD dataset.<sup>4</sup> It can be seen that our **RVL-BERT** achieves competitive Recall@50/100 compared to most of the existing methods, while lags behind the latest state-of-the-art such as MF-URLN and HGAT. We note that the use of an additional graph attention network in HGAT and a confidence-weighting module in MF-URLN can be possibly incorporated into our design, while we leave for future work.

<sup>4</sup>Note that for the results other than **Visual Phrases** and **Joint CNN**, Recall@50 is equivalent to Recall@100 (also observed in [2], [8]) because the number of ground truth subject-object pairs is less than 50.

#### E. Quantitative Results on SpatialSense Dataset

We compare our model with various recent methods, including some methods that have been compared in the VRD experiments. Note that **L-baseline**, **S-baseline** and **L+S-baseline** are baselines in [51] taking in simple language and/or spatial features and classifying with fully-connected layers. **ViP-CNN** [3] utilizes a phrase-guided message passing structure to model relationship triplets. **DR-Net** [5] exploits statistical dependency between object classes and predicates. **DSRR** [17] is a concurrent work<sup>5</sup> that exploits depth information for relationship detection with an additional depth estimation model. The **Human Performance** result is extracted from [51] for reference.

Table IV shows that our full model outperforms almost all existing approaches in terms of the overall accuracy and obtains the highest or second-highest accuracy for several relationships. While the concurrent work DSRR achieves a slightly higher overall recall, we expect our model to gain another performance boost with the additional depth information introduced in their work.

#### F. Qualitative Results of Visual-Linguistic Commonsense Knowledge

Figure 4 shows qualitative comparisons between predicting visual relationships with and without the visual-linguistic commonsense knowledge in our model. Especially, the example (a) in the figure shows that, with *linguistic* commonsense knowledge, a person is more likely to wear a shoe, rather than pants to wear a shoe. Same applies to the example (b) (where person-wear-pants is more appropriate than person-in front of-pants) and the example (c) (where tower-has-clock is semantically better than tower-above-clock). On the other hand, as the person in the example (e) is visually occluded, the model without *visual* commonsense knowledge prefers to dog-wear-shoe rather than person-wear-shoe; however, our model with the visual knowledge knows that that part is likely to be a person and is able to make correct predictions. Same applies to the example (d) (where both pillow and sofa are not clear) and (f) (where person is obscure). These examples demonstrate the effectiveness of our training strategy of exploiting rich visual and linguistic commonsense knowledge by pre-training on unlabeled visual-linguistic datasets.

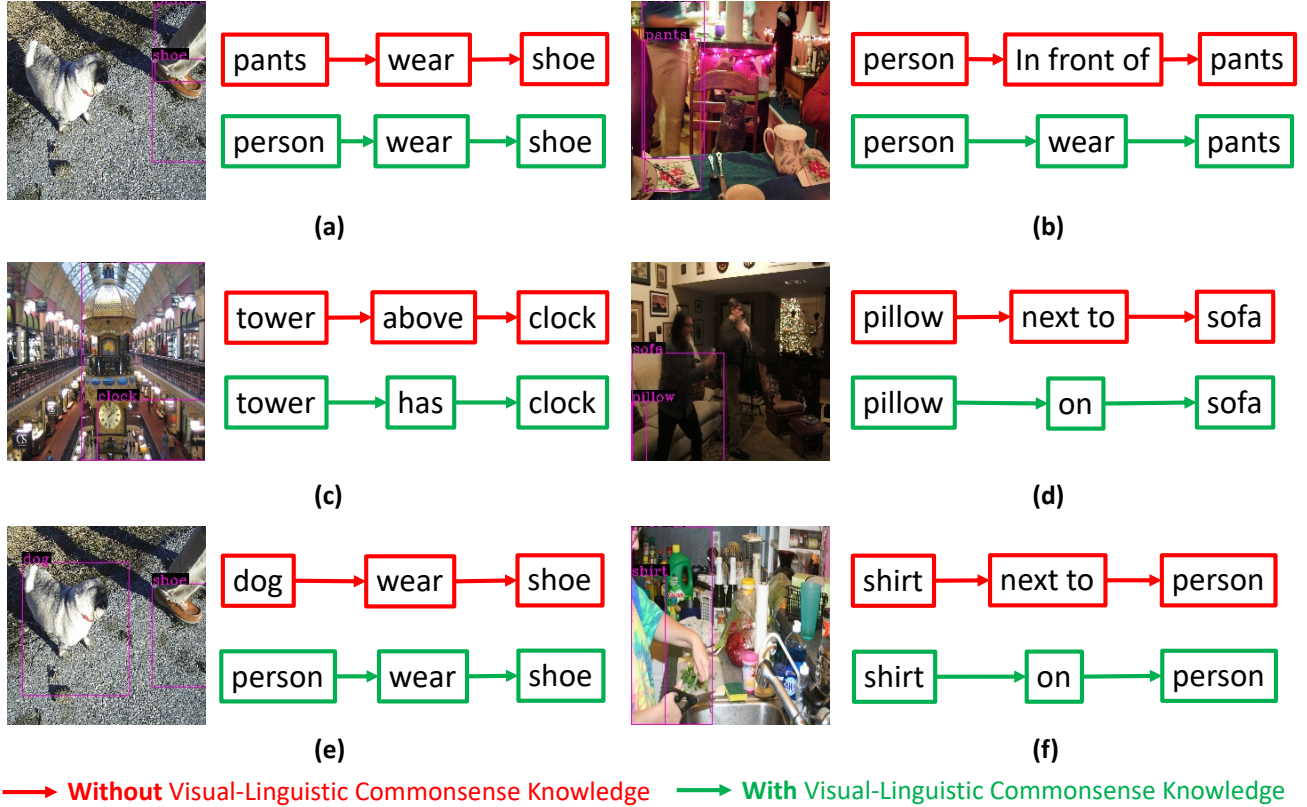
#### G. Qualitative Results of Mask Attention Module

The mask attention module aims to teach the model to learn and predict the attention maps emphasizing the locations of the given object labels. To study its effectiveness, we visualize the attention maps that are generated by the mask attention module during testing on the both datasets in Figure 5. The first two rows show two examples from SpatialSense, while the last two rows show three examples from the VRD dataset. Since the input embeddings of the model for the SpatialSense dataset in the form of triplets subject-predicate-object, and the VRD dataset in the form of doublets subject-object

<sup>5</sup>Originally published in August 2020.

**TABLE IV:** Classification accuracy comparison on the test split of SpatialSense dataset. Bold font represents the highest accuracy; underline means the second highest. Results of existing methods are extracted from [51]. †Note that DSRR [17] is a concurrent work that was published in August 2020.

Model	Overall	above	behind	in	in front of	next to	on	to the left of	to the right of	under
L-baseline [51]	60.1	60.4	62.0	54.4	55.1	56.8	63.2	51.7	54.1	70.3
PPR-FCN [9]	66.3	61.5	65.2	70.4	64.2	53.4	72.0	<u>69.1</u>	71.9	59.3
ViP-CNN [3]	67.2	55.6	68.1	66.0	62.7	62.3	72.5	<b>69.7</b>	73.3	66.6
Weakly [7]	67.5	59.0	67.1	69.8	57.8	<b>65.7</b>	75.6	56.7	69.2	66.2
S-baseline [51]	68.8	58.0	66.9	<u>70.7</u>	63.1	62.0	76.0	66.3	<u>74.7</u>	67.9
VTransE [6]	69.4	61.5	69.7	67.8	64.9	57.7	76.2	64.6	68.5	76.9
L+S-baseline [51]	71.1	61.1	67.5	69.2	66.2	64.8	77.9	<b>69.7</b>	<u>74.7</u>	<u>77.2</u>
DR-Net [5]	71.3	<b>62.8</b>	<b>72.2</b>	69.8	66.9	59.9	<u>79.4</u>	63.5	66.4	75.9
DSRR† [17]	<b>72.7</b>	61.5	<u>71.3</u>	71.3	67.8	<u>65.1</u>	<b>79.8</b>	67.4	<b>75.3</b>	<b>78.6</b>
<b>Ours: RVL-BERT</b>	<u>72.3</u>	<u>62.5</u>	70.3	<b>71.9</b>	<b>70.2</b>	<u>65.1</u>	78.5	68.0	74.0	75.5
Human Perf. [51]	94.6	90.0	96.3	95.0	95.8	94.5	95.7	88.8	93.2	94.1



**Fig. 4:** Qualitative comparisons between predicting visual relationship with or without visual-linguistic commonsense knowledge. Red boxes and arrows denotes predicting with the model without the knowledge, while the green boxes and arrows means predicting with the knowledge. This visualizations are performed during testing on the VRD dataset [2].

are different, three and two attention maps are generated for each example, respectively.

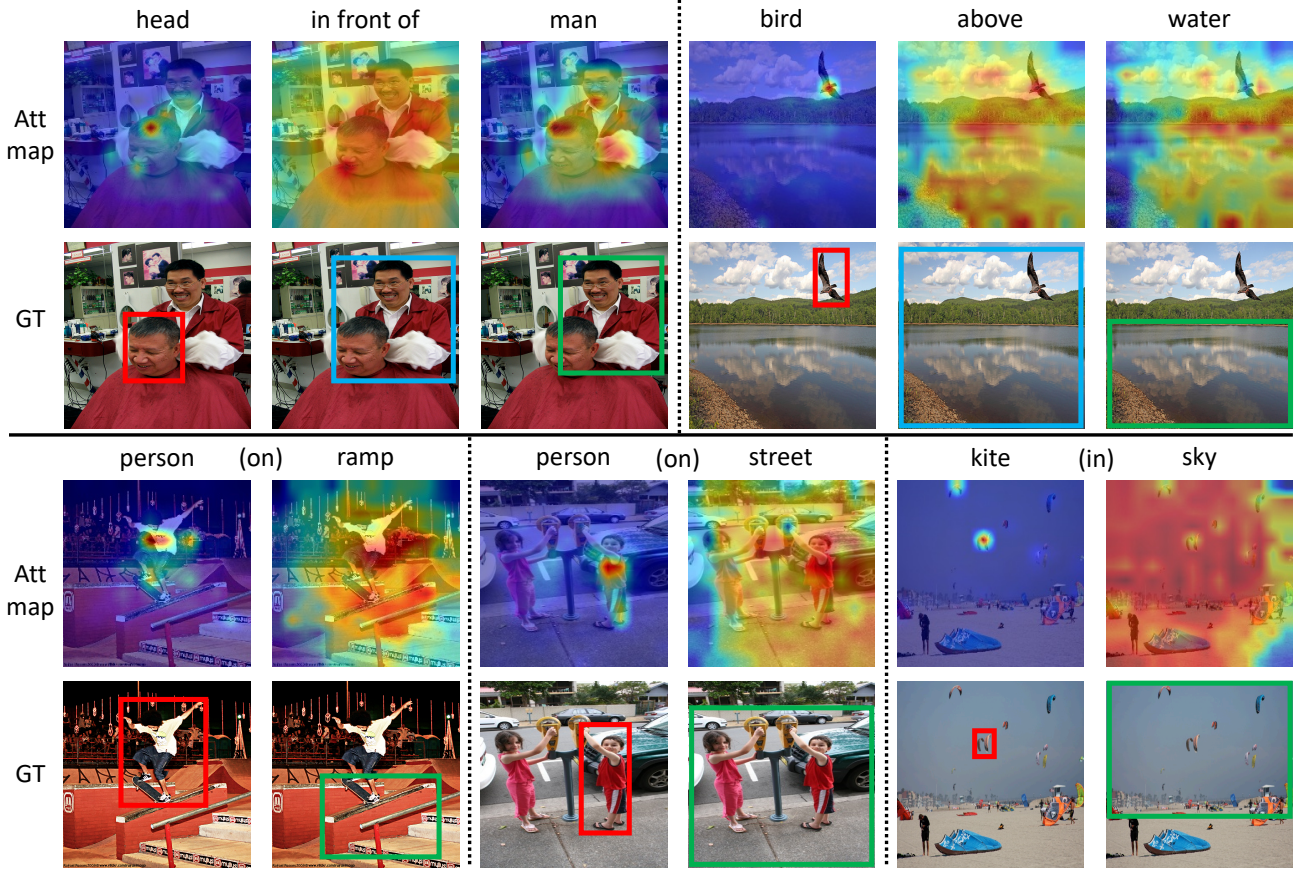
For both datasets, the model is actively attending to the region that contains the target object. Especially for the triplet data from SpatialSense, the model is also looking at the union bounding boxes which include cover both subjects and objects. Overall, we observe that the mask attention module learns better with triplet inputs than doublets inputs and this is assumedly because the additional examples of union boxes

provide more contexts and facilitate the learning process.

## V. CONCLUSION

In this paper, we proposed a novel visual relationship detection system named RVL-BERT, which exploits visual commonsense knowledge in addition to linguistic knowledge learned during self-supervised pre-training. A novel mask attention module is designed to help the model learn to capture the distinct spatial information and a spatial module





**Fig. 5:** Attention map visualization of SpatialSense (the first two rows) and VRD dataset (the last two rows). For each example, the first row shows predicted attention maps while the second shows ground truth bounding boxes.

is utilized to emphasize the bounding box coordinates. Our RVL-BERT is flexible in the sense that it can be solely used for predicate classification or cascaded with any state-of-the-art object detector. We have shown that it achieves competitive results with both quantitative and qualitative experiments on two challenging visual relationship detection datasets.

## REFERENCES

- [1] M. A. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1745–1752.
- [2] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in *European Conference on Computer Vision*. Springer, 2016, pp. 852–869.
- [3] Y. Li, W. Ouyang, X. Wang, and X. Tang, “Vip-cnn: Visual phrase guided convolutional neural network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.
- [4] B. Zhuang, L. Liu, C. Shen, and I. Reid, “Towards context-aware interaction recognition for visual relationship detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 589–598.
- [5] B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3076–3086.
- [6] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relation detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5532–5540.
- [7] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, “Weakly-supervised learning of visual relations,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5179–5188.
- [8] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, “Visual relationship detection with internal and external linguistic knowledge distillation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1974–1982.
- [9] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, “Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4233–4241.
- [10] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy, “Zoom-net: Mining deep feature interactions for visual relationship recognition,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 322–338.
- [11] S. Jae Hwang, S. N. Ravi, Z. Tao, H. J. Kim, M. D. Collins, and V. Singh, “Tensorize, factorize and regularize: Robust visual relationship learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1014–1023.
- [12] J. Jung and J. Park, “Visual relationship detection with language prior and softmax,” *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, pp. 143–148, 2018.
- [13] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, “Large-scale visual relationship understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9185–9194.
- [14] Z.-S. Hung, A. Mallya, and S. Lazebnik, “Union visual translation embedding for visual relationship detection and scene graph generation,” *arXiv preprint arXiv:1905.11624*, 2019.
- [15] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, “Scene graph generation with external knowledge and image reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1969–1978.
- [16] Y. Zhan, J. Yu, T. Yu, and D. Tao, “On exploring undetermined relationships for visual relationship detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5128–5137.

- [17] X. Ding, Y. Li, Y. Pan, D. Zeng, and T. Yao, "Exploring depth information for spatial relation recognition," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020, pp. 279–284.
- [18] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 886–13 895.
- [19] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.
- [20] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *Advances in Neural Information Processing Systems*, 2017, pp. 2171–2180.
- [21] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1261–1270.
- [22] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.
- [23] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 670–685.
- [24] 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 *Proceedings of the European Conference on Computer Vision*, 2018, pp. 335–351.
- [25] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Linknet: Relational embedding for scene graph," in *Advances in Neural Information Processing Systems*, 2018, pp. 560–570.
- [26] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in *Advances in Neural Information Processing Systems*, 2018, pp. 7211–7221.
- [27] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.
- [28] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.
- [29] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive relational networks for mapping images to scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.
- [30] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," *arXiv preprint arXiv:2001.02314*, 2020.
- [31] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.
- [32] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 684–699.
- [33] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 477–485, 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1047320318303535>
- [34] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.
- [35] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1–9.
- [36] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8376–8384.
- [37] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1219–1228.
- [38] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678.
- [39] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the fourth Workshop on Vision and Language*, 2015, pp. 70–80.
- [40] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.
- [41] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 13–23.
- [42] P. Sharma, N. Ding, S. Goodman, and R. Soiccut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.
- [43] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vi-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2020.
- [44] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International Conference on Machine Learning*, 2016, pp. 1378–1387.
- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [47] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.
- [48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [49] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 19–27.
- [50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [51] K. Yang, O. Russakovsky, and J. Deng, "SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, October 2019.
- [52] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.
- [53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [54] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [55] 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," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [56] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.