Bounding-box Channels for Visual Relationship Detection

Sho Inayoshi[1], Keita Otani[1], Antonio Tejero-de-Pablos[1], and Tatsuya Harada[1,2]

[1]The University of Tokyo, [2]RIKEN
{inayoshi, otani, antonio-t, harada}@mi.t.u-tokyo.ac.jp

Abstract. Recognizing the relationship between multiple objects in an image is essential for a deeper understanding of the meaning of the image. However, current visual recognition methods are still far from reaching human-level accuracy. Recent approaches have tackled this task by combining image features with semantic and spatial features, but the way they relate them to each other is weak, mostly because the spatial context in the image feature is lost. In this paper, we propose the bounding-box channels, a novel architecture capable of relating the semantic, spatial, and image features strongly. Our network learns bounding-box channels, which are initialized according to the position and the label of objects, and concatenated to the image features extracted from such objects. Then, they are input together to the relationship estimator. This allows retaining the spatial information in the image features, and strongly associate them with the semantic and spatial features. This way, our method is capable of effectively emphasizing the features in the object area for a better modeling of the relationships within objects. Our evaluation results show the efficacy of our architecture outperforming previous works in visual relationship detection. In addition, we experimentally show that our bounding-box channels have a high generalization ability.

Keywords: Bounding-box Channels, Visual Relationship Detection, Scene Graph Generation

1 Introduction

Although research on image understanding has been actively conducted, its focus has been on single object recognition and object detection. With the success of deep learning, the recognition and detection accuracy for a single object has improved significantly, becoming comparable to human recognition accuracy [17]. Previous works extract features from an input image using Convolutional Neural Networks (CNNs), and then the extracted features are used for recognition. However, single-object recognition and detection tasks for a single object cannot estimate the relationship between objects, which is essential for understanding the scene.

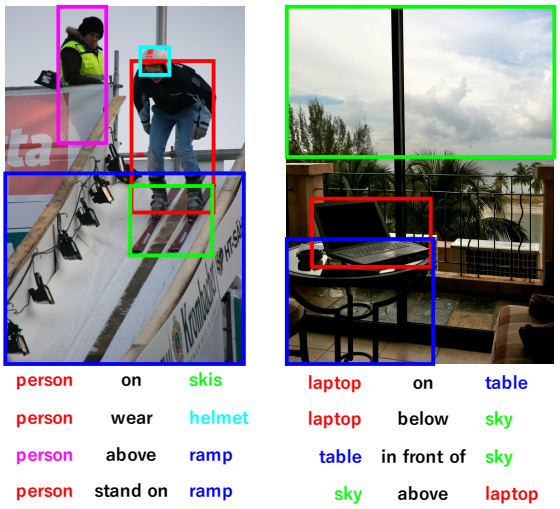

Fig. 1. Visual relationship detection (VRD) is the task of detecting relationships between objects in an image. Visual-relationship instances, or triplets, follow the subject-predicate-object pattern.

Visual relationship detection (VRD) is a very recent task that tackles this problem. The purpose of VRD is to recognize predicates that represent the relationship between two objects (e.g., *person wears helmet*), in addition to recognizing and detecting single objects, as shown in Figure 1. In other words, the purpose of VRD is detecting *subject-predicate-object* triplets in images. When thinking about solving the VRD task, humans leverage the following three features: the image features, which represent the visual attributes of objects, the semantic features, which represent the combination of *subject-object* class labels, and the spatial features, which represent object positions in the image. Therefore, previous research in VRD [1, 6, 9, 15, 24, 26, 27] employs these three types of features. Previous works have attempted to extract these features in a variety of ways. In [6, 9, 13, 15, 25, 26], semantic features are extracted from the label of the detected objects. For spatial features, in [6, 15, 22, 24, 26, 27] the coordinate values of the object candidate area have been used. Alternatively, previous works [1, 9, 23] proposed extracting the spatial features from a binary mask filled with zeros except for the area of the object pairs. In spite to the efforts of the aforementioned approaches, the task of recognizing the relationship between objects it is still far from achieving human-level recognition accuracy. One of the main reasons is that the spatial information contained in the image features has not been successfully leveraged yet. For example, several image recognition approaches flatten image features, but this eliminates the spatial information contained in the image features.

In this paper, in order to improve the accuracy of VRD, we propose a novel feature fusion method, the bounding-box channels, which are capable of modeling

together image features with semantic and spatial features without discarding spatial information (i.e., feature flattening). In our bounding-box channels, spatial information such as the location and overlap of the objects, is represented by adding channels to the image features. This allows for a strong association between the image features of each subject and object and their respective spatial information. Semantic features are also employed in the construction of the bounding-box channels to achieve a strong binding of all three features. Consequently, the relationship estimation network can learn a better model that leads to a better accuracy.

The contributions of this research are as follows:

- We propose the bounding-box channels, a new feature fusion method for visual relationship detection. It allows strongly combining semantic and spatial features with the image features for a boost in performance, without discarding spatial information.
- Our bounding-box channels follow a clear and straightforward implementation, and can be potentially used as replacements of the semantic and spatial feature extractors of other VRD methods.
- We provide extensive experimentation to show the generalization ability of our bounding-box channels, outperforming previous methods.

2 Related Work

Before the actual visual relationship detection, it is necessary to detect the image areas that contain objects and classify them. This is performed by object detection methods.

2.1 Object Detection

In recent years, the performance of object detectors has improved significantly through the use of deep learning. R-CNN [3] was a pioneer method in using convolutional neural networks (CNNs) for object detection. R-CNN uses a sliding window approach to input image patches to a CNN and performs object classification and regression of the bounding box (rectangular area containing the object). This method allowed for a significant improvement in object detection accuracy, but it has some limitations. First, the computational cost is huge because all windows are processed by the CNN. Second, since the size and position of the sliding window are static, some objects are misdetected. To solve this, Fast R-CNN [2] was proposed. In Fast R-CNN, image segmentation is applied to the whole image, and windows are generated dinamically by selecting regions of interest (RoI). Then, the CNN processes the whole image and classification and regression are performed only in the generated windows. This way, the computation cost is reduced and the accuracy is improved. Later, Faster R-CNN [17] was proposed, further improving computation time and accuracy by using a CNN to generate RoI.

Apart from the aforementioned methods, SSD [12] and YOLO [16], which are faster than the above methods, and FPN [10], which can detect objects of various scales, were also proposed. However, for the sake of comparison with previous VRD methods, directly using the detection results of R-CNN is a common practice [13, 27].

2.2 Visual Relationship Detection

VRD aims to a deeper understanding of the meaning of an image by estimating the relationship between pairs of detected objects. A large number of VRD methods [1, 6, 9, 13, 15, 19, 22–27] have been recently proposed, which share the same three type of features: image, semantic and spatial.

Needless to say, image features play an important role when considering the relationship between objects. For example, in the case of the *subject-object* pair *man-chair*, several possible relationships can be considered (e.g., "*sitting*", "*next to*"), but if an image is given, the possibilities are reduced to one. Previous works in image feature extraction proposed using multiple feature maps of a CNN, such as VGG16 [18]. The image feature map can be also obtained by cropping the feature map of the whole image by the smallest rectangular region containing subject and object using RoI Pooling [2] or RoI Alignment [4]. These feature extraction methods are widely used in various image recognition tasks including object detection, and their effectiveness has been thoroughly validated.

Semantic features are also an important element in VRD. For example, when considering the relationship between a man and a horse, our prior knowledge tells us that the relation "*man rides on a horse*" is more probable than "*man wears a horse*". In order to provide such information, previous works extracted semantic features from the predicted class label of detected objects [9, 13, 15, 22, 25, 27]. Previous works proposed two semantic feature extraction approaches. The first approach uses the word2vec embedding [14]. The class labels of the object pair were embedded in word2vec and processed by a fully-connected layer network, whose output is the semantic features. The second approach used the posterior probability distribution expressed by Eq. 1 [1, 20, 23, 26].

$$P(p|s, o) = N_{spo}/N_{so} \tag{1}$$

In Eq. 1, p, s, and o represents the labels of predicate, subject, and object respectively. N_{spo} is the number of *subject-predicate-object* triplets, and N_{so} is the number of *subject-object* pairs, both are emerged in the training set. This posterior probability distribution was used as the semantic features.

Last but not least, spatial features are crucial to detect spatial relationships like "*on*" and "*under*". For example, when estimating the relationship between a bottle and a table, if the image area of the bottle is located above the area of the table, the relationship is probably "*a bottle on a table*". In order to model these spatial relationships, two spatial feature extraction approaches were mainly proposed in previous work. The first approach used spatial scalar values (e.g., the distance between the centers of the pair of bounding boxes, the relative size

of objects), concatenated them into a vector and passed them through fully-connected layers, to output the spatial features [6, 15, 22, 24, 26, 27]. The second approach created binary images whose pixels are nonzero in the bounding box and zero otherwise, and feed it to a CNN to extract the spatial features [1, 9, 23].

As described above, a lot of approaches for extracting each feature were proposed. However, there was no deep consideration on how to effectively associate them together to learn a better model. Most of previous approaches either modeled each feature separately [13, 15, 25, 26] or simply flattened and concatenated each feature [1, 9, 19, 22, 24]. However, the former approach could not effectively combine the spatial information in the image features and the spatial features, because they process the image features and the spatial features separately. Similarly, the latter approach loses the spatial information contained in the image features due to flattening.

In this work, we propose a novel VRD method capable of effectively modeling image features together with spatial and semantic features. Modeling together instead of separately and concatenating them before flattening the features increases the accuracy and adequacy of the estimated relationship.

3 Bounding-Box Channels

This section describes our novel proposed bounding-box channels, which effectively combines image, semantic, and spatial features. Due to the structure of CNNs, the image features retain the spatial information to some extent unless they are flattened. For example, the upper right area in an image is mapped to the upper right area in the corresponding image features of the CNN. Contrary to previous methods, we avoid flattening the image features in order to preserve spatial information. In our bounding-box channels, the spatial and semantic information of the objects is modeled in the form of channels that are directly added to the image features. This allows for a strong association between the image features of each subject and object and their respective spatial information, hence the relationship estimation network can learn a better model that leads to a better accuracy.

Image Feature: In our method, the image features $\boldsymbol{F_i} \in \mathbb{R}^{H \times W \times n_i}$ are obtained from the feature map extracted from the backbone CNN such as VGG16 [18] aligned with the smallest area containing the bounding boxes of the subject-object pair by RoIAlign [4]. H and W are the height and width of the image features, and n_i is the number of channels of the image features.

Spatial Feature: As shown in Figure 2, we encode the positions of the objects by leveraging the bounding box of the *subject-object* pair in the image used to extract the image features. We build two tensors $\boldsymbol{C_s}, \boldsymbol{C_o} \in \mathbb{R}^{H \times W \times n_c}$ with the same height H and width W as the image features $\boldsymbol{F_i}$, and a number of channels n_c.

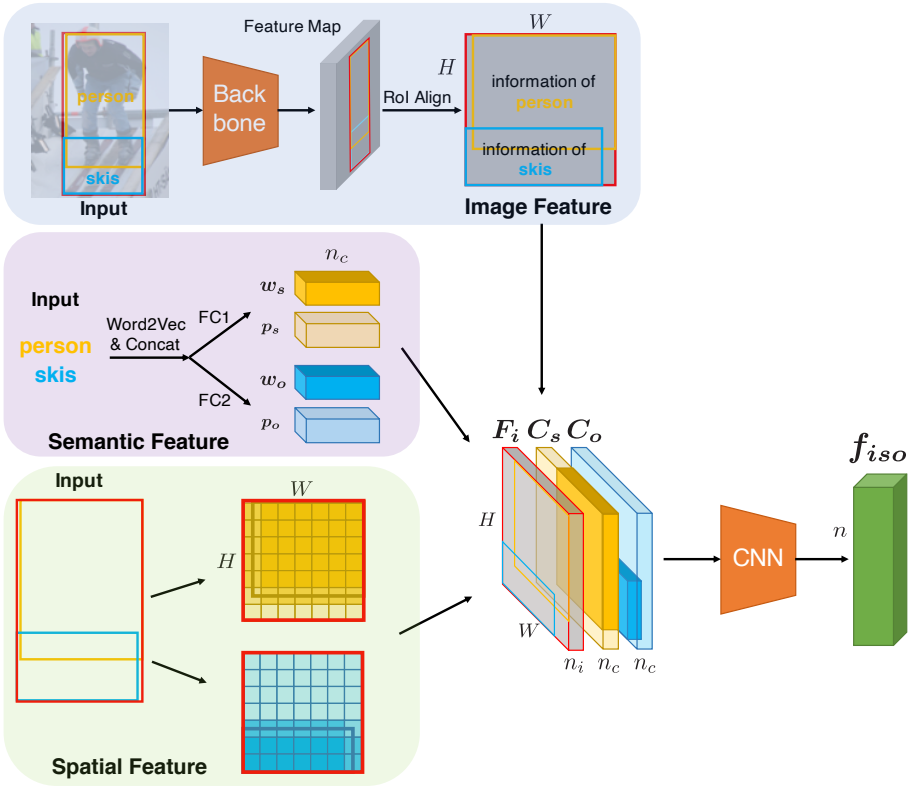

Fig. 2. Overview of how to construct the bounding-box channels. $\boldsymbol{F_i}$ is the aligned image feature. The words are transformed into the semantic features $\boldsymbol{w_s}$ and $\boldsymbol{w_o}$ via word2vec and fully-connected layers. In $\boldsymbol{C_s}$, the inner region of the bounding box is filled with $\boldsymbol{w_s}$, and the outer region is filled with a learnable parameter $\boldsymbol{p_s}$. $\boldsymbol{C_o}$ is filled with $\boldsymbol{w_o}$ and $\boldsymbol{p_o}$ in the same way. Finally, we concatenate $\boldsymbol{F_i}$, $\boldsymbol{C_s}$ and $\boldsymbol{C_o}$ in the channel direction, and fed into CNN to create our bounding-box channels $\boldsymbol{f_{iso}}$.

Semantic Feature: As our semantic features, the words of the detected *subject-object* pair classes are embedded in word2vec [14]. In our implementation, we use the pretrained word2vec model and fix its weights. The obtained pair of word vectors are concatenated and mapped to n_c dimensions by fully-connected layers; we denote them $\boldsymbol{w_s}$ and $\boldsymbol{w_o}$. Also, learn two n_c dimensional vectors, which are $\boldsymbol{p_s}$ and $\boldsymbol{p_o}$, respectively.

Aggregation: For $\boldsymbol{C_s}$, we fill the inner and outer regions representing the bounding box of the subject with $\boldsymbol{w_s}$ and $\boldsymbol{p_s}$ respectively. Similarly, for $\boldsymbol{C_o}$, we fill the inner and outer regions representing the bounding box of the object with $\boldsymbol{w_o}$ and $\boldsymbol{p_o}$ respectively. Finally, we concatenate $\boldsymbol{F_i}$, $\boldsymbol{C_s}$, and $\boldsymbol{C_o}$ in the channel direction, and fed into CNN to create our bounding-box channels $\boldsymbol{f_{iso}} \in \mathbb{R}^n$.

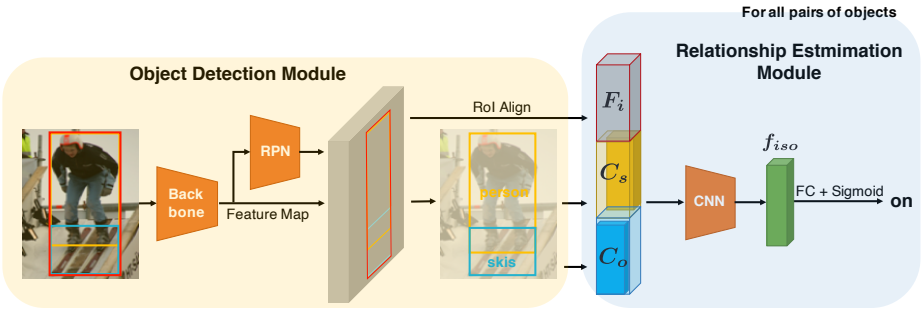

Fig. 3. Overview of the proposed BBCNet for visual relationship detection. First, a candidate set of *subject-object* pairs in the image is output by the object detection module. In our case, we use Faster R-CNN and its region proposal network (RPN). Second, for each *subject-object* pair, we extract the image features $\boldsymbol{F_i}$ from these candidate regions by RoI Align [4] and make $\boldsymbol{C_s}$ and $\boldsymbol{C_o}$ from the results of object detection as explained in Section 3. Finally, relationship estimation is conducted for each set of *subject-object* pair. The image, semantic, and spatial features are modeled together, which allows for a more accurate relationship estimation.

4 Bounding-Box Channels Network

We demonstrate the efficacy of our proposed bounding-box channels with our Bounding-Box Channels Network (BBCNet). Figure 3 shows the pipeline of our proposed BBCNet. The BBCNet consists of an object detection module and a relationship estimation module. First, an object detection module outputs the candidate set of *subject-object* pairs in the image. Second, we extract the image features $\boldsymbol{F_i}$ from the smallest area containing the bounding boxes of these candidate regions by RoI Align [4] and make $\boldsymbol{C_s}$ and $\boldsymbol{C_o}$ from the results of object detection as explained in Section 3. Finally, relationship estimation is conducted for each set of *subject-object* pair.

In previous VRD works, three types of features are leveraged for relationship estimation: image, semantic and spatial features. Our bounding-box channel module computes these three types of features for each candidate set of *subject-object* pair. The bounding-box channel module concatenates the image features extracted from the smallest rectangular region containing subject and object, the semantic features, and the spatial features (see Section 3). This way, the bounding-box channels are built. Our bounding-box channels are fed to a single layer CNN and two fully-connected layers to attain logit scores for each predicate class. As illustrated in Figure 3, we obtain the probability distribution for each predicate class via sigmoid normalization. In multi-class classification tasks, classification is generally performed using softmax normalization. However, in VRD task, multiple predicate may be correct for one *subject-object* pair. For example, in Figure 3, not only *person on bike* but also *person rides bike* can be correct. For such problem settings, not softmax normalization but sigmoid normalization is appropriate.

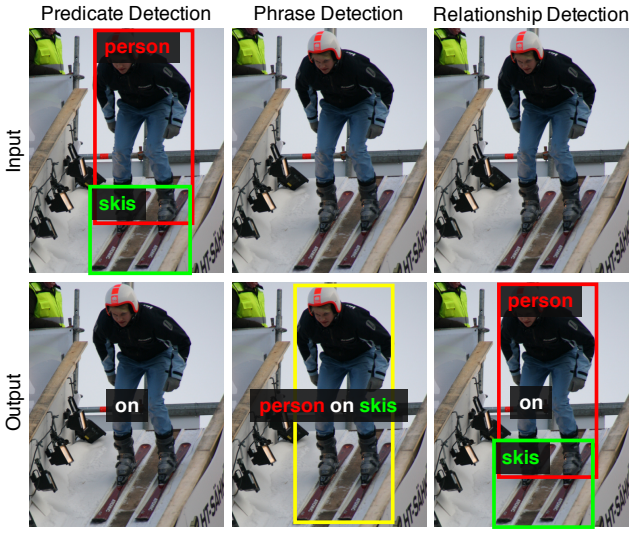

Fig. 4. We evaluate the performance of our method for visual relationship detection in the three tasks proposed in [13]: predicate detection, phrase detection and relationship detection. For predicate detection, classes and bounding boxes of objects are given in addition to an image, and the output is the predicate. Phrase detection and relationship detection take a single image, and output a set containing a pair of related objects or the individual related objects, respectively. In predicate detection, the given pair of objects is always related.

5 Experiments

5.1 Dataset

For our experiments, we used the Visual Relationship Detection (VRD) dataset [13] and the Visual Genome dataset [8], which are widely used to evaluate the performance of VRD methods. The VRD dataset contains 5000 images, with 100 object categories, and 70 predicate (relationship) categories among pairs of objects. Besides the categories, images are annotated with bounding boxes surrounding the objects. In total, VRD dataset contains 37993 subject-predicate-object triplets, of which 6672 are unique. We evaluate our method using the default splits, which contain 4000 training images and 1000 test images. The Visual Genome dataset contains 108073 images, with 150 object categories, and 50 predicate categories among pairs of objects. It is labeled in the same way as the VRD dataset. Our experiments follow the same train/test splits as [19].

5.2 Experimental Settings

VRD: We evaluate the proposed method in three relevant VRD tasks: predicate detection, phrase detection and relationship detection. The outline of each task is shown in Figure 4.

The predicate detection task (left) aims to estimate the relationship between object pairs in an input image, given the object class labels and the surrounding bounding boxes. In other words, this task assumes an object detector with perfect accuracy, and thus, the predicate detection accuracy is not influenced by the object detection accuracy. Next, the phrase detection task (middle) aims to localize set boxes that include a pair of related objects in an input image, and then predict the corresponding predicate for each box. Lastly, the relationship detection task (right) aims to localize individual objects boxes in the input image, and then predict the relevant predicates between each pair of them. In phrase detection and relationship detection, not all object pairs have a relationship and, in contrast to the predicate detection task, the performance of the phrase detection and the relationship detection is largely influenced by the performance of the object detection.

Following the original paper that proposed the VRD dataset [13], we use Recall@50 (R@50) and Recall@100 (R@100) as our evaluation metrics. R@K computes the fraction of true positive relationships over the total relevant relationships among the top K predictions with the highest confidence (probability). Another reason for using recall is that, since the annotations do not contain all possible objects and relationships, the mean average precision (mAP) metric is usually low and not representative of the actual performance, as some predicted relationships, even if correct, they may not be included in the ground truth. Evaluating only the top prediction per object pair may mistakenly penalize correct predictions since annotators have bias over several plausible predicates. So we treat the number of chosen predictions per object pair (k) as a hyper-parameter, and report R@n for different k's for comparison with other methods. Since the number of predicates is 70, $k = 70$ is equivalent to evaluating all predictions w.r.t. the two detected objects.

Visual Genome: We evaluate the proposed method in two relevant VRD tasks: predicate classification (PRDCLS) and scene graph detection (SGDET) [19]. PRDCLS is equivalent to predicate detection, and SGDET is equivalent to relationship detection.

5.3 Implementation Details

In our method, the image, semantic, and spatial features are extracted from an input image (Figure 3).

For object detection, we used the Faster R-CNN [17] structure[1] with ResNet50-FPN backbone [5, 10] pretrained with MSCOCO [11]. First, we input the image into the backbone CNN, and get the feature map of the whole image and the object detection results. Next, as described in Section 3, we create the bounding-box channels. When extracting the semantic features, we leverage the word2vec model [14] pretrained with the Google News corpus and fixed weights. We embed the object class names of subject and object separately using the word2vec

[1] for the sake of comparison, some experiments replace Faster R-CNN with R-CNN [3]

model, which generates a 300-dimensional vector per word, and concatenate them into a 600-dimensional vector. Then, we feed this vector to two separate fully-connected layers, and denominate the output $\boldsymbol{w_s}$ and $\boldsymbol{w_o}$ as our semantic features. In this paper, we set $H = 7, W = 7, n_i = 512, n_c = 256$, and $n = 256$. Our implementation is partially based on the architecture proposed in the work of [27], but our BBCNet results outperform theirs.

We apply binary cross entropy loss to each predicted predicate class. We train our BBCNet using the Adam optimizer [7]. We set the initial learning rate to 0.0002 for backbone, and 0.002 for the rest of the network. We train the proposed model for 10 epochs and divide the learning rate by a factor of 5 after the 6th and 8th epochs. We set the weight decay to 0.0005. During training, from the training set, we sample all the positive triplets and the same number of negative triplets. This is due to the highly imbalance nature of the problem (only a few objects in the images are actually related).

During testing, we rank relationship proposals by multiplying the predicted subject, object, and predicate probabilities as follows:

$$\boldsymbol{p}^{total} = \boldsymbol{p}^{det}(s) \cdot \boldsymbol{p}^{det}(o) \cdot \boldsymbol{p}^{pred}(pred) \tag{2}$$

where $\boldsymbol{p}^{total}$ is the probability of subject-predicate-object triplets, $\boldsymbol{p}^{det}(s), \boldsymbol{p}^{det}(o)$ are the probability of subject and object classes respectively, and $\boldsymbol{p}^{pred}(pred)$ is the probability of the predicted predicate class (i.e., the output of BBCNet). This reranking allows a fairer evaluation of the relationship detector, by giving preference to objects that are more likely to exist in the image.

5.4 Quantitative Evaluation

VRD: As explained in Section 5.2, we compare the proposed method and related works via three evaluation metrics. For phrase detection and relationship detection, we compare our performance with four state-of-the-art methods [13, 15, 25, 27], that use the same object detection proposals reported in [13] using R-CNN [3]. Also, we compare our performance with three state-of-the-art methods [21, 25, 26] using the object detection proposals of a more complex object detector (Faster R-CNN in our case). The phrase detection and the relationship detection performances are reported in Table 1 and Table 2. Also, the predicate detection performance is reported in Table 3. These results show that our BBCNet achieves state-of-the-art performance, outperforming previous works in almost all evaluation metrics on entire VRD dataset.

In particular, Table 2 and the zero-shot part in Table 3 show the results of the performance comparison when using combinations of triplets that exist in the test split but not in the training split (zero-shot data). A high generalization ability for zero-shot data has an important meaning in model evaluation. A poor generalization ability requires including all combinations of subject-predicate-object in the training data, which is unrealistic in terms of computational complexity and dataset creation. Our BBCNet achieves the highest performance in the zero-shot VRD dataset for all evaluation metrics, which shows its high generalization ability.

Table 1. Performance comparison of the phrase detection and relationship detection tasks on the entire VRD dataset. "-" indicates performances that have not been reported in the original paper. The best performances are marked **in bold**. In the upper half, we compare our performance with four state-of-the-art methods that use the same object detection proposals. In the lower half, we compare with three state-of-the-art methods that use more sophisticated detectors. Our method achieves the state-of-the-art performance in almost all evaluation metrics.

	Phrase Detection						Relationship Detection					
	$k=1$		$k=10$		$k=70$		$k=1$		$k=10$		$k=70$	
Recall at	100	50	100	50	100	50	100	50	100	50	100	50
w/ proposals from [13]												
CAI [27]	-	-	-	-	19.24	17.60	-	-	-	-	17.39	15.63
Language cues [15]	-	-	20.70	16.89	-	-	-	-	18.37	15.08	-	-
VRD-Full [13]	17.03	16.17	25.52	20.42	24.90	20.04	14.70	13.86	22.03	17.43	21.51	17.35
LSVR [25]	19.78	18.32	25.92	21.69	25.65	21.39	**17.07**	**16.08**	22.64	19.18	22.35	18.89
Ours	**20.95**	**19.72**	**28.33**	**24.46**	**28.38**	**24.47**	16.63	15.87	**22.79**	**19.90**	**22.86**	**19.91**
w/ better proposals												
LK distillation [22]	24.03	23.14	29.76	26.47	29.43	26.32	21.34	19.17	29.89	22.56	31.89	22.68
LSVR [25]	32.85	28.93	39.66	32.90	39.64	32.90	26.67	23.68	32.63	26.98	32.59	26.98
GCL [26]	36.42	31.34	42.12	34.45	42.12	34.45	28.62	25.29	33.91	28.15	33.91	28.15
Ours	**40.72**	**34.25**	**46.18**	**36.71**	**46.18**	**36.71**	**33.36**	**28.21**	**38.50**	**30.61**	**38.50**	**30.61**

Table 2. Performance comparison of the phrase detection and relationship detection tasks on the zero-shot data (i.e., subject-predicate-object combinations not present in the training split) in the VRD dataset. We compare our performance with four methods which use the object detection results reported in [13] using R-CNN [3]. Our method outperforms the other related works in all evaluation metrics, and demonstrates high generalization ability.

	Phrase Detection						Relationship Detection					
	$k=1$		$k=10$		$k=70$		$k=1$		$k=10$		$k=70$	
Recall at	100	50	100	50	100	50	100	50	100	50	100	50
w/ proposals from [13]												
CAI [27]	-	-	-	-	6.59	5.99	-	-	-	-	5.99	5.47
Language cues [15]	-	-	15.23	10.86	-	-	-	-	13.43	9.67	-	-
VRD-Full [13]	3.75	3.36	12.57	7.56	12.92	7.96	3.52	3.13	11.46	7.01	11.70	7.13
Ours	**8.81**	**8.13**	**16.51**	**12.57**	**16.60**	**12.66**	**6.42**	**5.99**	**13.77**	**10.09**	**13.94**	**10.27**

Visual Genome: As explained in Section 5.2, we compare the proposed method with four state-of-the-art methods [19, 23, 26] via two evaluation metrics. The scene graph detection and the predicate classification performances are reported in Table 5.4, in which graph constraint means that there is only one relationship between each object pair (that is, $k = 1$ in VRD dataset). These results show that our BBCNet achieves the state-of-the-art performance in almost all evaluation metrics for the Visual Genome dataset as well.

Section 6 offers a more detailed discussion on the cause of the obtained results.

Table 3. Performance comparison of the predicate detection task on the entire VRD and zero-shot VRD data sets. Our method outperforms the other state-of-the-art related works in all evaluation metrics on both entire set and zero-shot set. This result shows that when the object detection is perfectly conducted, our method is the most accurate for estimating the relationships between *subject-object* pairs.

	Predicate Detection					
	Entire set			Zero-shot		
	$k = 1$	$k = 70$		$k = 1$	$k = 70$	
Recall at	100/50	100	50	100/50	100	50
VRD-Full [13]	47.87	-	-	8.45	-	-
VTransE [24]	44.76	-	-	-	-	-
LK distillation [22]	54.82	90.63	83.97	19.17	76.42	56.81
DSR [9]	-	93.18	86.01	-	79.81	60.90
Zoom-Net [21]	50.69	90.59	84.25	-	-	-
CAI + SCA-M [21]	55.98	94.56	89.03	-	-	-
Ours	**57.87**	**95.98**	**89.43**	**27.54**	**86.06**	**68.78**

Table 4. Comparison with the state-of-the-art methods on Visual Genome dataset. Graph constraint means that only one relationship is considered between each object pair. Our method achieves the state-of-the-art performance in most evaluation metrics.

	Graph Constraint						No Graph Constraint			
	SGDET			PRDCLS			SGDET		PRDCLS	
Recall at	100	50	20	100	50	20	100	50	100	50
w/ better proposals										
Message Passing [19]	4.2	3.4	-	53.0	44.8	-	-	-	-	-
Message Passing+ [23]	24.5	20.7	14.6	61.3	59.3	52.7	27.4	22.0	83.6	75.2
MotifNet-LeftRight [23]	30.3	27.2	**21.4**	67.1	65.2	58.5	35.8	**30.5**	88.3	81.1
RelDN [26]	32.7	28.3	21.1	68.4	68.4	66.9	36.7	30.4	97.8	93.8
Ours	**34.3**	**28.5**	20.4	**69.9**	**69.9**	**68.5**	**37.2**	29.9	**98.2**	**94.7**

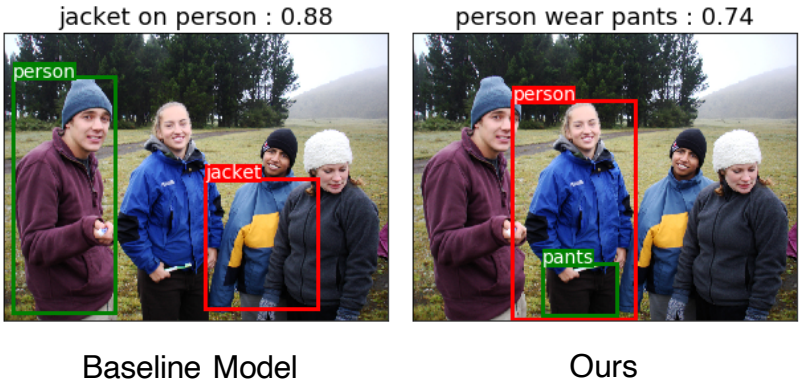

Fig. 5. Qualitative comparison of our method with the baseline (no bounding box channels, as in [27]). The number over the image represents confidence of the output triplet. Thanks to our better modeling of the image-semantic-spatial features, the relationships detected by our method (right column) are more adequate than those of the baseline (left column).

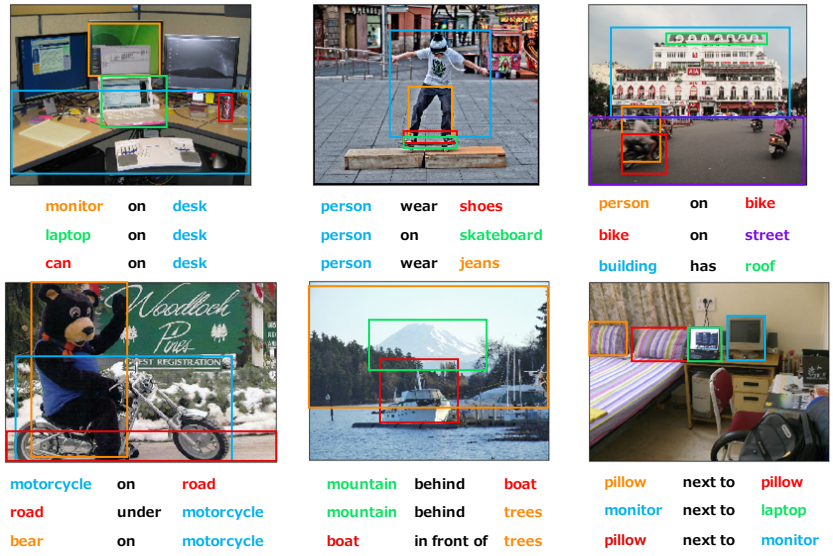

Fig. 6. Qualitative examples of our proposed VRD method. The color of the text corresponds to the color of the bounding box of the object (same color means same object). "*bear on motorcycle*" in the lower-left image is an example of zero-shot data (i.e., *subject-predicate-object* combinations not present in the training split).

5.5 Qualitative Evaluation

In order to understand the contribution of the bounding box channels (BBC), we performed a comparison with the baseline in [27], whose architecture resembles ours, but without the BBC. Figure 5 shows a the VRD results of both our method and the baseline. Whereas the baseline outputs the relationship between a person and different person's belongings, our method outputs the relationship between a person and their own belongings. Similarly, in the other example, the relationship estimated by our method is more adequate than that of the baseline. The reason is that, although the baseline is able to combine the semantic features with the image features, the spatial features do not work for removing inadequate relationships with respect to the objects location. On the other hand, the proposed BBC allows considering the objects position properly when estimating their relationship. Figure 6 shows supplementary results of our method. Our method is able to estimate relationships not present in the training split (zero-shot data), as in the case of "*bear on motorcycle*" in the lower-left image.

6 Discussion

6.1 Quantitative Evaluation

Our method outperforms previous works in the vast majority of the conducted experiments. We can draw some conclusions from these results. First, our bounding-

box channels can model more discriminative features for VRD than previous methods. The reason is that our BBCNet does not lose the spatial information in the image features, and effectively combines image, semantic and spatial features. Second, the word2vec based semantic feature extraction method has high generalization ability, because similar object names are projected on neighbor areas of the feature space. Therefore, the relationships not present in the training split but similar with the relationships present in the training split can be detected. For example, If the *dog under chair* triplet present in the training split, the *cat under chair* triplet is likely to be detected even it is not present in the training split. In contrast, the generalization ability is lower in methods whose semantic feature extraction uses the posterior probability distribution in Eq. 1. This occurs because restricting the semantic features to the posterior probability of the triplets included in the training set, worsens robustness against unseen samples (i.e., zero-shot data).

6.2 Qualitative Evaluation

As explained in Sec. 5.5, our BBCNet without the bounding-box channels (BBC) resembles the architecture of CAI [27]. Thus, these results can also be interpreted as an ablation study that shows the improvement in performance of a previous method by applying our BBC. But our bounding-box channels are potentially applicable not only to the architecture of CAI [27] adopted in this paper but also to other architectures. First, as far as the task of VRD is concerned, the image, semantic, and spatial features can be effectively combined by simply replacing the feature fusion modules with our bounding-box channels. In addition, not limited to VRD, if the task uses an object candidate area and an image, our bounding-box channels can be used to effectively combine both, expecting an improvement in performance.

7 Conclusion

In this paper, we proposed the bounding-box channels, a feature fusion method capable of successfully modeling together spatial and semantic features along with image features without discarding spatial information. Our experiments show that our architecture is beneficial for VRD, and outperforms the previous state-of-the-art works. As our future work, we plan to apply our bounding-box channels to a variety of network architectures, not limited to the VRD task, to further explore the combination of the image, semantic and spatial features.

Acknowledgements. This work was partially supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, and partially supported by JSPS KAKENHI Grant Number JP19H01115. We would like to thank Akihiro Nakamura and Yusuke Mukuta for helpful discussions.

References

1. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
2. Girshick, R.: Fast r-cnn. In: IEEE International Conference on Computer Vision (ICCV) (2015)
3. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2015)
6. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)
8. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (May 2017)
9. Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: Association for the Advancement of Artificial Intelligence (AAAI) (2018)
10. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision (ECCV) (2016)
13. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (ECCV) (2016)
14. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR) (2013)
15. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: IEEE International Conference on Computer Vision (ICCV) (2017)
16. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Neural Information Processing Systems (NIPS) (2015)

18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
19. Xu, D., Zhu, Y., Choy, C., Fei-Fei, L.: Scene graph generation by iterative message passing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: European Conference on Computer Vision (ECCV). pp. 670–685 (2018)
21. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., Change Loy, C.: Zoom-net: Mining deep feature interactions for visual relationship recognition. In: European Conference on Computer Vision (ECCV) (September 2018)
22. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: IEEE International Conference on Computer Vision (ICCV). pp. 1068–1076 (2017)
23. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
24. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
25. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Association for the Advancement of Artificial Intelligence (AAAI) (2019)
26. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
27. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: IEEE International Conference on Computer Vision (ICCV) (2017)