Hindawi
Scientific Programming
Volume 2022, Article ID 5985392, 9 pages
https://doi.org/10.1155/2022/5985392

Research Article

A Symmetric Fusion Learning Model for Detecting Visual Relations and Scene Parsing

Xuan Liu,[1,2] **Xiaochuan Jing**,[2] **Zhong Zheng**,[2] **Wanru Du**,[1,2] **Xingxing Ding**,[1] **and Quan Zhu**[1]

[1]*China Aerospace Academy of Systems Science and Engineering, Beijing, China*
[2]*Aerospace Hongka Intelligent Technology (Beijing) CO LTD, Beijing, China*

Correspondence should be addressed to Xuan Liu; xuan0414@sina.com

Received 31 March 2022; Accepted 25 May 2022; Published 27 June 2022

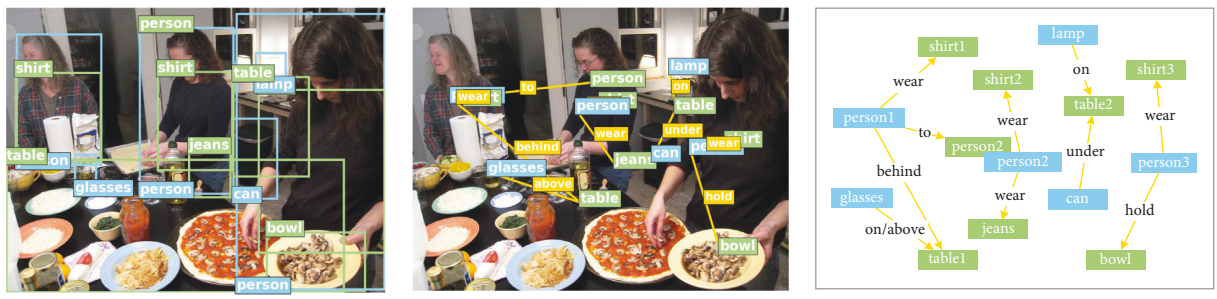
Academic Editor: Liang Zhao

Visual relationship detection (VRD) aims to locate objects and recognize their pairwise relationships for parsing scene graphs. To enable a higher understanding of the visual scene, we propose a symmetric fusion learning model for visual relationship detection and scene graph parsing. We integrate objects and relationship features at visual and semantic levels for better relations feature mapping. First, we apply a feature fusion for the construction of the visual module and introduce a semantic representation learning module combined with large-scale external knowledge. We minimize the loss by matching the visual and semantic embeddings using our designed symmetric learning module. The symmetric learning module based on reverse cross-entropy can boost cross-entropy symmetrically and perform reverse supervision for inaccurate annotations. Our model is compared with other state-of-the-art methods in two public data sets. Experiments show that our proposed model achieves encouraging performance in various metrics for the two data sets investigated. The further detailed analysis demonstrates that the proposed method performs better by partially alleviating the impact of inaccurate annotations.

1. Introduction

The rapid development of the computer vision community pushes forward object detection and semantic segmentation over a short time. These advancements are driven by the deep neural network baselines, such as R-CNN and fully convolutional network (FCN) frameworks for object detection and semantic segmentation. Advanced deep convolutional neural networks (CNNs) have achieved optimal performance in the fields of visual tasks such as image classification [1], object detection [2], and visual relationship detection [3]. Nevertheless, these CNNs need to be trained in a fully supervised learning manner, requiring manually annotated data sets, such as ImageNet [4], MS-Coco [5], and Pascal VOC [6]. Most existing VRD models detect semantic relationships in the VisualGenome (VG) [7] and Visual Relationship Detection (VRD) data sets [3]. However, collecting and labeling a multimodal data set is costly and easy to make errors in actual engineering. Inaccurate and insufficient labels are common noise in manual annotations. Even high-quality data sets likely contain incorrect labels. Therefore, training accurate neural networks in the presence of manual annotations have become a task with crucial practical significance in deep learning.

Image understanding research has gradually developed from low-level feature extraction to high-level semantic learning. The next step is to start inferences on the semantic relationship between multiple objects, which could help many multimodal tasks such as visual question answering [8], image captioning [9], visual commonsense reasoning [10], human-centered activity recognition [11], and intention recognition [12]. Johnson et al. [13] proposed the scene graphs, which give a platform to infer the visual scene. Given an image, the scene graph generation (SGG) task essentially parses the fully connected graphs, and it considers pairwise interactions of nodes (objects) as edges. These interactions can be spatial, comparative, or action-based and are expressed as the subject-relationship-object (SRO) triplets such as < person-ride-horse >



(a) (b) (c)

Figure 1: An example of predicting the semantic relationship between multiple objects from an image and parsing the scene graph: (a) an input image with bounding boxes, (b) the interactions between objects in an image, and (c) an example of parsing a scene graph.

(action), < plate-on-table > (spatial), and < person1-taller-person2 > (comparative). As Figure 1(b) shows, the interaction between objects in an image could generate a scene graph to explore multiple objects' relationships. It shows that SGG plays a vital role in the high-level understanding of images. Object entities are frequently semantically connected; Chao et al. [14] utilized improving class-representative visual features as the semantic embedding to achieve better object recognition; and several methods have been proposed to learn semantic associations between the visual and semantic modules. Most of them [3, 15–17] mainly followed the pipelined method. The pipelined method detects visual relationships in two separate steps. First, the entities in a figure are detected. Next, the relations between entities are predicted by running classification. However, the accuracy of the classification will be affected by preliminary errors using these pipelines. To address this, we employ a manner that maintains visual similarity to detect the SRO triplets instead of similarity-based relations retrieval. We design a visual and a semantic module that learns the mapping from the visual feature space to the semantic embedding space.

In this work, we propose a symmetric fusion learning model for detecting visual relations. Instead of modeling objects and relations as discrete labels, we can precisely detect visual relationships by matching the visual and semantic embedding space. We utilize fusion learning to design the structure of the visual module, and we introduce a semantic representation learning module combined with large-scale external knowledge. Besides, inaccurate and insufficient labels are common noise in manual annotations. Luo et al. [18] employed an adaptive loss function to mitigate the effects of noises in their video semantic recognition task. Inspired by the symmetric cross-entropy learning loss function [19], we propose a symmetric learning module boosting cross-entropy symmetrically using reverse cross-entropy, to perform reverse supervision for inaccurate annotations and better parsing scene graphs. We demonstrate that our model is highly competitive on the VisualGenome (VG) data set, which contains 108,249 images where each with an average of 35 objects, 26 attributes, and 21 pairwise relationships. Furthermore, we also evaluate our model on the Visual Relationship Detection (VRD) data set, showing that our model can significantly improve visual relationship prediction in scene graphs.

The key contributions are summarized as follows:

(i) We built a symmetric fusion learning model, which can precisely detect visual relations by matching the visual and semantic embedding space

(ii) We propose a symmetric learning module boosting cross-entropy symmetrically to perform reverse supervision for inaccurate annotations and better parsing scene graphs

(iii) Experiments on the two public data sets show that our model achieves encouraging performance and consistent improvements in various metrics obtained by effectively handling the visual relationships detection issue

2. Related Work

2.1. Visual Relationship Detection. Recently, many visual tasks have focused on visual relationship detection for better parsing a scene graph. Early work mostly focused on predicting specific types of predicates, such as predicting the spatial relationship of image objects [20] and detecting the human-interaction relationships [21, 22]. As a mid-level visual task, VRD benefits many high-level visual tasks, such as visual question answering [8], image captioning [9], and visual commonsense reasoning [10].

Early VRD methods used specific phrases to detect the relationship; Lu et al. [3] first employed the "language prior" from semantic word embeddings to predict visual relationships. Zhuang et al. [15] applied feature representations to characterize the interaction pattern based on the context-aware interaction classifier. Like these methods, many other works detected objects and pairwise relationships separately [3, 15–17]. Unlike these approaches, we utilize a fusion learning manner that integrates the subject and object features to design the visual module. We want to learn the mapping from the visual feature space to the semantic embedding space.

Context information learning is another attempt considered by researchers. Yu et al. [23] integrated the prior distribution obtained from external linguistic knowledge into the visual relationship prediction model. Liang et al. [24] proposed a deep neural network model with structural ranking loss to model objects and predicates separately. Subsequently, the feature interactions and message sharing

were discussed by Yin et al. [25]; they formed a spatiality-aware contextual feature learning model Zoom-net to promote feature interactions.

The VRD methods mentioned above focus on detecting predicate relationships. Recently, researchers considered three components of each relationship triplet, detecting object pairs that contain specific predicates. Zhang et al. [26] embedded object pairs and predicate separately to the independent semantic spaces for object and relation. Zhan et al. [27] improved visual relationship detection by utilizing undetermined relationships. Furthermore, Zhan et al. [28] correlated object detection, significance detection, and predicate detection for better visual relationship prediction. Unlike these methods, we employ symmetric learning to adjust the representation of pairwise relationships to maintain stable scene parsing performance.

2.2. Scene Graph Parsing. SGG has attracted extensive attention during the last couple of years due to the significance of parsing scenes in various computer vision tasks. Most context-based modeling methods form the scene graph employing message passing in the local subgraph structure. Subsequently, several scene graph generation methods transfer messages between object pairs and predicates to capture contextual information. Xu et al. [29] proposed a model that passes messages containing contextual information within subgraphs. Li et al. [30] introduced a subgraph-based Factorizable Net that passes the message between object feature vectors and subgraph feature maps. Zellers et al. [31] represented the global context of objects and relationships based on recurrent sequential architecture LSTM. Chen et al. [32] introduced prior knowledge of statistical correlations represented by a knowledge graph to propagate node messages. More recently, Chen et al. [33] employed the generated missing labels to train scene graphs. Yang et al. [34] proposed probabilistic modeling to ease the semantic ambiguity of visual relationship prediction. Saha et al. [35] proposed a context-aware detection method to identify obscured regions of the scene, leading to better visual scene understanding.

Many studies have been proposed to solve various problems existing in the task. We design a semantic module to better infer the semantic relationships between entities; it can project the word vectors of the triplet into an embedding space where the words maintain higher semantic similarity to each other. Furthermore, inaccurate and insufficient labels are common noise in manual annotations. In this work, we propose a symmetric learning module that can alleviate the impact of noisy labeling by reverse supervision. It is straightforward to use and requires minor intervention for training. More importantly, it represents a vital function in tolerating label noise for the manually annotated data sets.

3. Methodology

In this section, we first describe the visual module architecture of our model. Then, based on our visual network structure, we introduce semantic representation learning combined with large-scale external knowledge. Finally, we incorporate symmetric learning against noisy labels into our model for better parsing scene graphs. The brief training process of our model is shown in Figure 2.

3.1. Visual Network Structure. Object Detection: Given an image, we employ faster R-CNN [36] object detector to get better proposals for each image as in previous works. First, we utilize the region proposal network (RPN) to generate a set of object proposals. Each pair of objects is enclosed by a bounding box and obtains the appearance feature. The appearance feature of the bounding box outlines the objects and the surrounding context, which is helpful to predict the relationships. Because the relationship between objects often arises from the visual area where the two objects interact, we extract the features from the union region of object pairs for triplet fusion learning. We utilize a similar process as Zhang et al. [37] to extract the feature for each proposal.

Fusion Learning: For each object region, the feature vector f_s, f_r, and f_o of the subject, relationship, and object, respectively, are extracted by the ROI (region of interest) pooling layer. These features are sent to the fully connected layers to extract and integrate visual information through feature space transformation; then we obtain the implicit semantic embedding h_v^s, h_v^r, and h_v^o through mapping the original features to the hidden node. To jointly identify predicates, the visual feature for the relationship is formed by fusing the hidden features h_v^s, h_v^r, and h_v^o as shown in Figure 3. Later, the fusion learning of the $< s, r, o >$ triplet is carried out through the concatenation of the object feature. Each proposal will be fed into the fusion learning module. Finally, three visual embeddings v^s, v^r, and v^o for a triplet are output by considering independent object features and their fusion embedding.

3.2. Semantic Modeling. On account of semantic correlated relationships to one another, we can infer the correct $\langle s, r, o \rangle$ triplet from similar relationships that occur more frequently. Our approach presents visual relationships by grouping similar language expressions. The semantic module projects the word vectors of the triplet into an embedding space where the words maintain higher semantic similarity to each other. We first introduce the process that maps the word vectors in the embedding space; then we describe the training process that pushes the related relationship closer in the embedding space.

A suitable word vector of objects and relationship labels is essential for fine-tuning. We consider the pretrained word vectors fastText.We obtain semantic knowledge through large-scale public available text data mining. We employ a pretrained word vector fastText trained on Common Crawl [38] to implement our purpose. Unlike the word vector models that ignore the morphological features inside words, the fastText model utilizes a bag of n-grams to obtain the word-internal information. First, we employ pretrained fastText [38] to project the objects and predicates into a word embedding space.

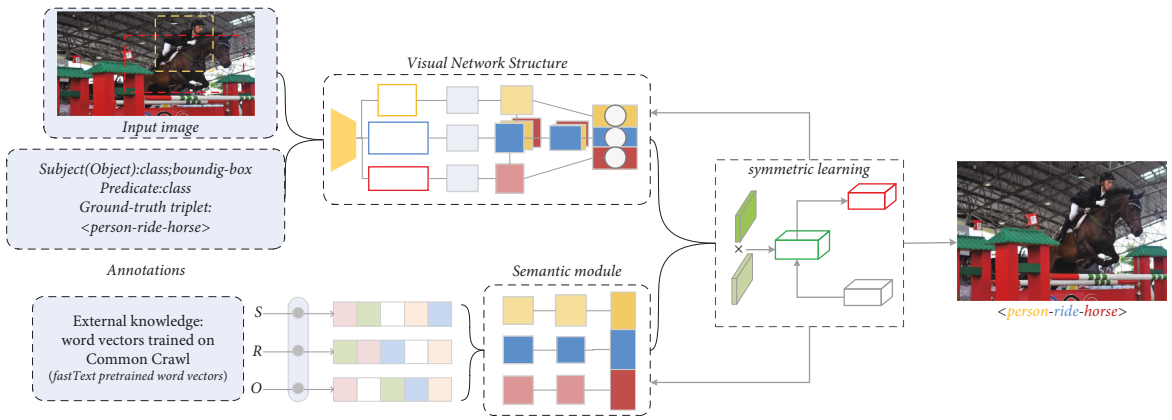


Figure 2: A brief illustration of our proposed approach. We utilize the image annotation from the visual relationship data set and pretrained word vectors from large-scale Common Crawl for training. We predict the visual relationship triplets by matching the embeddings of the visual and semantic modules. Then we apply the symmetric learning module to perform reverse supervision and correction while alleviating noisy labels.

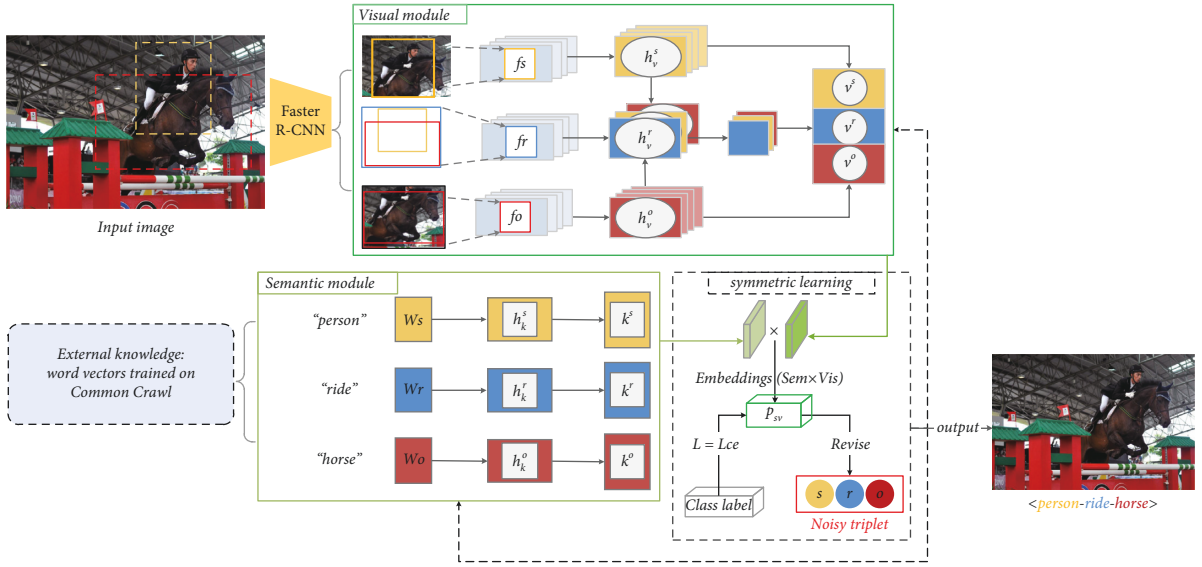


Figure 3: The framework of our proposed approach. We employ faster R-CNN to extract the global feature of an input image via the pretrained VGG-16 backbone and adopt fastText to initialize the word vector to obtain semantic embeddings. Our model contains three modules: (1) a visual module containing SRO branches is used to extract the visual features of object proposals, (2) a semantic module that introduces external information utilizes pretrained fastText word vectors, and (3) a symmetric learning module for alleviating noise via reverse supervision.

$$W_r = \frac{\left(\text{len}\left(r_p\right)+1\right)\cdot F\left(\text{or}_i\right)}{\sqrt{\sum_{i=1}^{d}\left(F\left(\text{ori}\right)\right)^2}},$$
$$W_{s(o)} = \frac{\left(\text{len}\left(r_o\right)+1\right)\cdot F\left(\text{or}_i\right)}{\sqrt{\sum_{i=1}^{d}\left(F\left(\text{ori}\right)\right)^2}}, \quad (1)$$

where W_r and $W_{s(o)}$ are the initial standardized fastText word vector of relationship and objects, respectively; r_o and r_p are the number of the classes of objects and predicates, respectively; and $F(\text{ori})$ represents the raw word vector of pretrained fastText.

Next, the word vectors W_s, W_r, and W_o of the SRO labels are given into an FC layer as shown in Figure 3, which outputs the three intermediate hidden embeddings h_k^s, h_k^r, and h_k^o. The approach aims to generate a word embedding that projects similar relationships closer to one another than the initial fastText word vector space. Finally, we get the three word embeddings k^s, k^r, and k^o of SRO through the FC layer once again.

3.3. Symmetric Learning Module. We can get the output embedding $v \in \mathscr{V} = \{v^s, v^r, v^o\}$ and $k \in \mathscr{K} = \{k^s, k^r, k^o\}$ from the above two modules in the training process. Here, we minimize the loss by matching the visual and semantic embeddings using our designed symmetric learning module; P_{sv} is the output from matching the visual and semantic embedding as shown in Figure 4. We employ the cross-

FIGURE 4: Qualitative examples of detecting visual relationships by the baseline method and proposed model: (a) a test image from the VRD data set, (b) the extracting visual relationship triplets by the baseline method, and (c) the extracting visual relationship triplets by the baseline model with the *fl* and *sl* module. The entity pairs have the same color as the corresponding bounding boxes. ~~abc~~ indicates the detected error visual relationships. < subject-predicate-object > * represents the detected proper visual relationships different from ground-truth triplets.

entropy (CE) for matching the embedding of triplets, while cross-entropy is the most generally utilized for training deep neural networks. Given an M-class visual relationship data set, $\mathscr{D} = \{(\mathbf{t}, a)^{(i)}\}_{i=1}^{n}$, where $\mathbf{t}$ represents a triplet sample in the multidimensional input space and $a \in \mathscr{A} = \{1, \ldots, M\}$ is ground-truth from the manual annotations. The probability for each triplet $\mathbf{t}$ learning from the ground-truth annotation m is $(p_1(m|\mathbf{t}) = e^{z_m}/\sum_{j=1}^{M} e^{z_j}$; z_j represents the logits. While the $p_2(m| \mid \mathbf{t})$ denotes the ground-truth distribution for the data sets, the CE for the triplet $\mathbf{t}$ is

$$L_c = -\sum_{m=1}^{M} p_2(m|\mathbf{t}) \log p_1(m|\mathbf{t}). \tag{2}$$

For the two distributions in this study, $p_1(m|\mathbf{t})$ is the distribution learned from the training data, and $p_2(m|\mathbf{t})$ denote the ground-truth distribution for the data sets. Kullback–Leibler divergence (denoted as D_{KL}) can be used to calculate the difference between these two distributions:

$$D_{KL}(p_2 p_1) = -S(p_2) + H(p_2, p_1), \tag{3}$$

where $S(p_2)$ is the entropy of the ground-truth distribution for the data sets and $H(p_2, p_1)$ is the cross-entropy (L_c) of p_2 and p_1. In order to make our training model closer to the real distribution, we minimize the D_{KL}.

Various works have confirmed the weakness of the cross-entropy used for deep neural network learning [19]. When there are noisy labels in the data set, it may cause inadequate extraction and ambiguity. A single $p_2(m|\mathbf{t})$ cannot accurately represent the true distribution; instead, the predicted $p_1(m|\mathbf{t})$ can denote the true distribution partly. Consequently, apart from the consideration of $p_2(m|\mathbf{t})$ as the ground-truth, we need to combine the reverse $D_{KL}(p_1 p_2)$ to help the model fit better. Here, we consider the relative entropy of the reverse fitting to obtain the logically symmetric D_{KL} and extend it to the reverse cross-entropy (L_r):

$$L_r = H(p_1, p_2) = -\sum_{m=1}^{M} p_1(m|\mathbf{t}) \log p_2(m|\mathbf{t}). \tag{4}$$

We introduced the reverse cross-entropy boosting cross-entropy symmetrically into our loss function, thus performing reverse supervision for inaccurate annotations. Formally, the final loss function for the symmetric learning module is

$$L = \lambda L_c + \mu \cdot L_r, \tag{5}$$

where λ and μ are hyperparameters, λ is adopted to alleviate the overfitting issue of standard cross-entropy L_c, and μ mitigates label noise by robust adjustment of L_r.

In addition, λ and μ are defined by fine-tuning different modules. In the visual modeling stage, we only use L_c to extract the region of interest ($\mu = 0$). While in the symmetric learning stage, both L_c and L_r are utilized for matching visual and semantic embedding.

4. Experiments and Results

In this part, the performance of relations detection and the effectiveness of noise mitigation are explored. We will first

Table 1: Comparison with state-of-the-art on the VG data set.

Method	SGGen			SGCls			PredCls		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
IMP [29]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
Frequency [31]	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1
Frequency + overlap [31]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
MotifNet-LeftRight [31]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
Graph R-CNN [39]	—	11.4	13.7	—	29.6	31.6	—	54.2	59.1
VCTREE-SL [40]	21.7	27.7	31.1	35.0	37.9	38.6	59.8	66.2	67.9
RelDN [37]	21.1	28.3	32.7	36.1	36.8	36.8	66.9	68.4	68.4
VCTREE + TranstextNet$_A$ [41]	—	28.1	31.7	—	38.3	39.3	—	66.9	68.7
Ours	21.2	29.2	34.7	38.1	38.8	38.8	67.2	68.8	68.8

introduce experimental settings, including data sets, evaluation metrics, and implementation details. Compared results between our model and baseline methods are presented in Section 4.2.

4.1. Experimental Settings

4.1.1. Data Sets. We conduct experiments on two public data sets: VisualGenome (VG) [7] and Visual Relationship Detection (VRD) data sets [3].

VG: in our experiments, we use the pruned version of VG [29] that only contains 150 object categories and 50 predicates. We follow the same train/test splits as in Xu et al. [29].

VRD: the VRD data set we used consists of 5,000 images with 100 object categories and 70 predicate categories. We use the same train/test split as in previous work [3].

4.1.2. Evaluation Metrics. *VG*: following Zellers et al. [31], we conduct three metrics to evaluate the performance: scene graph generation (SGGen), scene graph classification (SGCls), and predicate classification (PredCls). SGGen is the mode that needs to predict subject/object boxes and all labels. SGCls predict that all labels are given ground-truth subject and object boxes. PredCls predict predicate labels are given ground-truth subject and object boxes and labels. We use Recall@n (R@20, R@50, and R@100) as the evaluation metrics following previous works. Recall (R@N) is defined as the ratio of the true relationship in the top-N confident relation predictions in an image.

VRD: following Zhang et al. [26], we apply the object detector pretrained on the COCO data set. We follow previous works [23] using Recall@n (R@50, R@100) as the evaluation metrics, which reports R@50 and R@100 for relationship and phrase detection at 1, 10, and 70 predicates per entity pair.

4.1.3. Implementation Details. In our experiments, to ensure compatibility with the structures of previous works, we utilize VGG-16 as the backbone of VG and VRD data sets. For our symmetric learning module, a relative λ is used for achieving better convergence on difficult data sets. The large λ tends to cause overfitting, while the small λ can ease the overfitting of the single CE. Nevertheless, the reverse cross-entropy term is noisy tolerant, but the convergence becomes slow when μ is too large. We use a relatively small λ to avoid overfitting and large μ against noisy labels. The parameters λ and μ are set to 0.1 and 1, respectively.

Our model was optimized using SGD with momentum, and the base learning rate is set to $1e^{-3}$. Moreover, since the labels of subject and object play an important role in predicting visual relationships, we employ the empirical distribution over relations between object pairs to aid in generating scene graphs as in previous work.

4.2. Experimental Results

4.2.1. Compared Results. In this section, we compare our proposed method with the previous state-of-the-art models. We conducted experiments on two data sets (VG and VRD) and compared the performance with previous works. Tables 1 and 2 show the results of different baseline models, together with our framework for two data sets.

VG: We compare our method with eight state-of-the-art (SoTA) methods on the VG data set. The eight methods are IMP [29], frequency [31], frequency + overlap [31], MotifNet-LeftRight [31], graph R-CNN [39], VCTREE-SL [40], RelDN [37], and VCTREE + TranstextNet [41]. Table 1 presents the performances of ours and the other SoTA methods. As shown in Table 1, our method achieves encouraging R@n scores on various metrics in the VG data set. In SGGen, our method performs the best. Compared with the current baseline method VCTREE + TranstextNet$_A$ [41], our method outperforms it by 3% at R@100. In PredCls, our method outperforms the other methods on R@20, R@50, and R@100. In SGCls, our method outperforms the best baseline by 0.5% on R@50 and is only lower than it by 0.5% on R@100. Note that our method has not made outstanding progress on the SGCls and PredCls tasks, the improvement is obvious on the SGGen task compared to the other tasks as we keep on improving relationship prediction capabilities.

VRD: Table 2 presents comparisons on VRD with eight state-of-the-art methods: VRD [3], KL distillation [23], Zoom-net [25], CAI + SCA-M [25], RelDN [37], AVR [42], GPS-Net [43], and SABRA [44]. For a fair comparison of VRD, we adopt the VGG-16 backbone pretrained on ImageNet used for these baselines to train our model. As shown in Table 2, our method consistently achieves superior performance on two metrics. The proposed method

Table 2: Comparison with state-of-the-art on the VRD data set.

Method	Relation detection				Phrase detection			
	$k=1$		$k=70$		$k=1$		$k=70$	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
VRD [3]	17.03	16.17	24.90	20.04	14.70	13.86	21.51	17.35
KL distillation [23]	19.17	21.34	22.68	31.89	23.14	24.03	26.32	29.43
Zoom-net [25]	18.92	21.41	21.37	27.30	24.82	28.09	29.05	37.34
CAI + SCA-M [25]	19.54	22.39	22.34	28.52	25.21	28.89	29.64	38.39
RelDN [37]	18.92	22.96	21.52	26.38	26.37	31.42	28.24	35.44
AVR [42]	22.83	25.41	27.35	32.96	29.33	33.27	34.51	41.36
GPS-Net [43]	21.50	24.30	—	—	28.90	34.00	—	—
SABRA [44]	24.47	29.16	27.27	33.99	30.57	36.80	33.39	41.79
Ours	26.01	29.90	28.63	35.21	32.02	37.31	34.88	43.07

performs the encouraging R@100 ($k=70$) that is 35.21% on relation detection and 43.07% on phrase detection. These improvements again verify the ability of our framework and the necessity of symmetric learning for visual relationship detection. Moreover, these performances verify that our framework can be applied to data sets of different scales, as well as to more complex situations.

4.2.2. Ablation Study. We conduct an ablation study analyzing the contributions of two key components: the structure of fusion learning in the visual module (fl) and the structure of the symmetric learning module (sl). The baseline indicates the prediction model of Figure 3 without using symmetric fusion learning; that is, we take images and word vectors as input to the visual and semantic module without fusing the hidden features, and we minimize the loss by matching visual and semantic embedding only using cross-entropy. We validate the performance of the fl and sl components in the two harder tasks: phrase and relationship detection.

The R@n scores (R@50 and R@100) of phrase detection and relation detection on the VRD data set are chosen as the evaluation metrics. The results of the ablation experiments are summarized in Table 3; we report the phrase and relationship detection performance in R@n scores (R@50 and R@100), where baseline denotes the baseline model and baseline + fl + sl presents our model with all proposed components. From rows in Table 3, we can see that the performance improves consistently when all the components are utilized together.

To further evaluate the ability of our model against label noise, experimental noisy labels are generated by transforming the 20% labels of training samples to one of the other class labels randomly. In Table 4, baseline (20% noisy labels) is the baseline model trained on noisy labels; rows 2 and 4 demonstrated the effectiveness of our model against label noise. We can see that the baseline model with fl and sl can partially alleviate the impact of noisy labels.

4.2.3. Qualitative Results. Figure 4 presents the qualitative analysis of our model on VRD. We implement qualitative statistics and visualizations on the VRD data set to better show the performance improvement of our model.

Table 3: Ablation studies on the key components of our method. We report the phrase and relationship detection performance in R@n scores (R@50 and R@100).

K	Method	Relation detection		Phrase detection	
		R@50	R@100	R@50	R@100
1	Baseline	25.39	29.67	31.50	37.00
	Baseline + fl	25.62	29.92	31.45	37.21
	Baseline + fl + sl	26.01	29.90	32.02	37.31
70	Baseline	27.38	34.33	33.56	41.90
	Baseline + fl	27.38	34.81	33.91	42.46
	Baseline + fl + sl	28.63	35.21	34.88	43.07

Table 4: Ablation studies on the key components of our model. We report the phrase and relationship detection performance in R@n scores (R@50 and R@100).

K	Method	Relation detection		Phrase detection	
		R@50	R@100	R@50	R@100
1	Baseline (20% noisy labels)	18.00	22.44	26.37	31.42
	Baseline + fl + sl (20% noisy labels)	24.09	28.13	29.34	34.51
70	Baseline (20% noisy labels)	21.52	26.38	28.24	35.44
	Baseline + fl + sl (20% noisy labels)	25.06	30.83	30.47	38.07

To verify the effectiveness of the key components of our model, we visualize the extracting results of the two models (baseline and baseline + fl + sl) on test examples in Figure 4. The baseline indicates the prediction model without using symmetric fusion learning. The comparisons with the ground-truth triplets show that our proposed model can properly detect the correct triplets. It proves the effectiveness of the fusion learning structure and the symmetric learning module.

Compared with the baseline, our proposed model utilizing fl and sl component can detect more correct relationships, for example, < tree-to-building >, < grass-beneath-hydrant >, and < train-under-sky >. Our model can correct some mistakes of predicates. For example, the < grass-on-hydrant > is revised to < grass-beneath-hydrant

>, making it more precise to parse the scene graph. It also can be observed that our model processed more complete detection of visual relationships; some triplets that are not in ground-truth are also precisely detected.

5. Conclusion

In this work, we introduce a symmetric fusion learning model for detecting visual relationships and parsing scenes. The visual module is designed by integrating the subject and object representations. We can precisely detect visual relationships by matching the visual and semantic embedding space. Moreover, the model can also alleviate the impact of noise with the symmetric learning module. Comprehensive experimental results on VRD and VG data sets show the effectiveness of our proposal.

Data Availability

All data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

X.L. and X.J. conceptualized the study. X.L. and Z.Z. contributed to data curation. X.L. and X.J. contributed to methodology. X.L., Z.Z., W.D., X.D., and Q.Z. contributed to software. X.L. contributed to formal analysis. X.L., W.D., X.D., and Q.Z. contributed to funding acquisition. X.L. wrote the original draft of the manuscript. X.L. and X.J. reviewed and edited the manuscript.

References

[1] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: a comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[2] H. Zhou, Y. Lin, L. Yang, J. Lai, and X. Xie, "Benchmarking Deep Models for Salient Object Detection," 2022, https://arxiv.org/abs/2202.02925.

[3] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proceedings of the European Conference on Computer Vision*, pp. 852–869, Springer, Amsterdam, The Netherlands, October 2016.

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, Miami, FL, USA, June 2009.

[5] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Springer, Zurich, Switzerland, September 2014.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[7] R. Krishna, Y. Zhu, O. Groth et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," 2016, https://arxiv.org/abs/1602.07332.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," 2016, https://arxiv.org/abs/1606.01847.

[9] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11218, pp. 684–699, Munich, Germany, September 2018.

[10] X. Zhu, Z. Li, X. Wang et al., "Multi-Modal Knowledge Graph Construction and Application: A Survey," 2022, https://arxiv.org/abs/2202.05786.

[11] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2019.

[12] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3033–3044, 2019.

[13] J. Johnson, R. Krishna, M. Stark et al., "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668–3678, Boston, MA, USA, June 2015.

[14] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proceedings of the European Conference on Computer Vision*, pp. 52–68, Springer, Amsterdam, The Netherlands, October 2016.

[15] 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*, pp. 589–598, Venice, Italy, October 2017.

[16] 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*, pp. 5532–5540, Honolulu, HI, USA, July 2017.

[17] Y. Zhu and S. Jiang, "Deep structured learning for visual relationship detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, New Orleans, LA, USA, February 2018.

[18] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2017.

[19] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, Seoul, Korea (South), October 2019.

[20] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, AK, USA, June 2008.

[21] B. Yao and L. Fei-Fei, "Grouplet: a structured image representation for recognizing human and object interactions," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9–16, IEEE, Francisco, CA, USA, June 2010.

[22] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 1331–1338, IEEE, Barcelona, Spain, November 2011.

[23] 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*, pp. 1974–1982, Venice, Italy, October 2017.

[24] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 848–857, Honolulu, HI, USA, July 2017.

[25] G. Yin, L. Sheng, B. Liu et al., "Zoom-net: mining deep feature interactions for visual relationship recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 322–338, ECCV), Munich, Germany, July 2018.

[26] 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, pp. 9185–9194, Honolulu, HI, USA, January 2019.

[27] Y. Zhan, J. Yu, T. Yu, and D. Tao, "On exploring undetermined relationships for visual relationship detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5128–5137, Long Beach, CA, USA, June 2019.

[28] Y. Zhan, J. Yu, T. Yu, and D. Tao, "Multi-task compositional network for visual relationship detection," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2146–2165, 2020.

[29] 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*, pp. 5410–5419, Honolulu, HI, USA, July 2017.

[30] 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*, pp. 335–351, (ECCV), Munich, Germany, October 2018.

[31] 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*, pp. 5831–5840, Salt Lake, UT, USA, June 2018.

[32] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6163–6171, Seoul, Korea (South), October 2019.

[33] V. S. Chen, P. Varma, R. Krishna, M. Bernstein, C. Re, and L. Fei-Fei, "Scene graph prediction with limited labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2580–2590, Seoul, Korea (South), October 2019.

[34] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang, "Probabilistic modeling of semantic ambiguity for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12527–12536, Nashville, TN, USA, June 2021.

[35] B. Saha and S. Das, "Catch Me if You Can: A Novel Task for Detection of Covert Geo-Locations (CGL)," 2022, https://arxiv.org/abs/2202.02567.

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[37] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11535–11543, Long Beach, CA, USA, June 2019.

[38] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation*, (LREC, Miyazaki, Japan, May 2018.

[39] 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*, pp. 670–685, (ECCV), Munich, Germany, October 2018.

[40] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6619–6628, Long Beach, CA, USA, June 2019.

[41] G. S. Kenigsfield and R. El-Yaniv, "TranstextNet: transducing text for recognizing unseen visual relationships," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1955–1964, Waikoloa, HI, USA, January 2021.

[42] J. Lv, Q. Xiao, and J. A. V. R. Zhong, "Attention Based Salient Visual Relationship Detection," 2020, https://arxiv.org/abs/2003.07012.

[43] X. Lin, C. Ding, J. Zeng, and D. Tao, "Gps-net: graph property sensing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3746–3753, Seattle, WA, USA, June 2020.

[44] D. Jin, X. Ma, C. Zhang et al., "Towards Overcoming False Positives in Visual Relationship Detection," 2020, https://doi.org/10.48550/arXiv.2012.12510.