Pattern Recognition 132 (2022) 108848

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning multimodal relationship interaction for visual relationship detection

Zhixuan Liu, Wei-Shi Zheng*

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Article history:
Received 2 June 2021
Revised 22 March 2022
Accepted 12 June 2022
Available online 14 June 2022

Keywords:
Visual relationship detection
Scene graph generation
Relationship context
Multimodal relationship interaction

ABSTRACT

Visual relationship detection aims to recognize visual relationships in scenes as triplets ⟨*subject-predicate-object*⟩. Previous works have shown remarkable progress by introducing multimodal features, external linguistics, scene context, etc. Due to the loss of informative multimodal hyper-relations (*i.e.* relations of relationships), the meaningful contexts of relationships are not fully captured yet, which limits the reasoning ability. In this work, we propose a Multimodal Similarity Guided Relationship Interaction Network (MSGRIN) to explicitly model the relations of relationships in graph neural network paradigm. In a visual scene, the MSGRIN takes the visual relationships as nodes to construct an adaptive graph and enhances deep message passing by introducing Entity Appearance Reconstruction, Entity Relevance Filtering and Multimodal Similarity Attention. We have conducted extensive experiments on two datasets: Visual Relationship Detection (VRD) and Visual Genome (VG). The evaluation results demonstrate that the proposed MSGRIN has empirically performed more effectively overall.

1. Introduction

Images we see from most scenarios consist of multiple objects which are interacting with each other. These interactions facilitate our understanding of images. Visual relationship detection is a task that aims to detect relationships among entities (subjects and objects). In the task, a visual relationship is represented as a ⟨*subject-predicate-object*⟩ formed triplet. The predicate could be "*left to*", "*taller than*" or "*hit*", which conveys significant information that contributes to scene understanding. The visual relationships can construct the relational reasoning scene graph for the subsequent visual question answering [1] and image caption [2] tasks. Also, the visual relationships summarize the content of a complex image and thus can be an effective tool to measure semantic similarity between images for image-to-image retrieval [3].

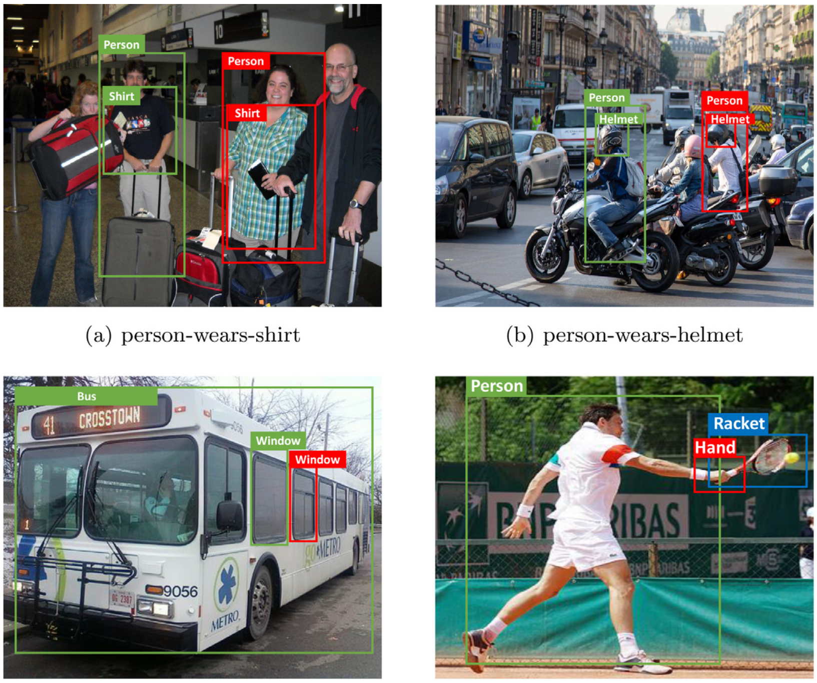
One of the biggest challenges in visual relationship detection is that relationships in a real-world scenario are very complicated. The detection task has to recognize vaguely defined predicates. The variance is unnegligible of intra-class relationships like "*person-ride-horse*" and "*person-ride-bike*", and might lead to misclassification of predicates, whereas the inter-class variance can be small and "*person-ride-bike*" might appear to be the same relationship as "*person-in front of-bike*". Another challenge is the lack of adequate relationship annotations for training. An image can contain thousands of relationships, labeling all of them is high cost and hard to implement. For example, the relationship "*person-wears-jacket*" about the man on the right in Fig. 1(a) is not annotated in the Visual Relationship Detection [4] dataset.

Rare methods in the existing literature attach importance to the relationship-wise interactions which imply lots of high-level semantics. In most cases, visual relationships coexist with each other in the scene and the coexistence is not trivial. For example, *shirt* and *pants* (clothing), or *eye* and *nose* (body parts) usually appear together as the *object* in relationships "*person-has-object*". These universal coexistences imply the related patterns of relationships and describing such informative associations has the potential to explore the scene context further which helps to detect confusing relationship instances. Also, learning by analogy is an important ability of human beings. Taking the scene in Fig. 1(b) as an example, when taught about the relationship "*person-wears-helmet*" in green, we can recognize the relationship in red as "*person-wears-helmet*" as well. What's more, human can also infer "*person-holding-racket*" from "*hand-holding-racket*" and "*person-has-hand*" in the scene Fig. 1(d). We believe that such reasoning ability is closely related to learning interactions among relationships and achieving this will help to tackle the challenges in visual relationship detection.

To further expose the underlying relationship interaction pattern, we investigate the coexistence of relationships in Visual Genome [5] dataset (VG). For any *predicate1* and *predicate2* in VG,

* Corresponding author.
E-mail addresses: liuzhx8@mail2.sysu.edu.cn (Z. Liu), wszheng@ieee.org (W.-S. Zheng).

https://doi.org/10.1016/j.patcog.2022.108848
0031-3203/© 2022 Published by Elsevier Ltd.



(a) person-wears-shirt

(b) person-wears-helmet

(c) window-on-bus

(d) holding-racket

Fig. 1. In most real-world scenarios, relationships coexist with each other. Humans often learn to acquire knowledge through association (*e.g.* analogy).

we calculate the ratio of the co-occurrence frequency of *predicate1* and *predicate2* to the occurrence frequency of *predicate1* and denote it as Coexistence Rate. For example, (7th row, 5th column) in Fig. 2 represents the ratio of images containing both "*behind*" and "*near*" to images containing "*behind*". As illustrated in Fig. 2, besides the trivial coexistence caused by the long tail distribution of relationships (*e.g.* "*on*" appears in most scenes), there exists repetition of the same predicate as well as meaningful interactions of different predicates. For example, "*holding*" and "*wearing*" indicate person status, while "*in front of*" and "*near*" indicate entity position. This finding motivates us to more deeply extract and utilize the potential relationship interactions.

In this work, we explore a novel model to solve the problem by incorporating multimodal cues and capturing the interactions among relationships in graph neural network (GNN) [6] paradigm dubbed Multimodal Similarity Guided Relationship Interaction Network (MSGRIN). In the MSGRIN, the candidate relationships represented by subject-object pairs act as nodes and the interactions among them act as edges; the relationship detection task is equivalent to node classification in the graph. Note that in the rest of the paper, node and edge will follow the above definition. The proposed MSGRIN mainly contains three components: entity-wise relevance module, multimodal affinity generation module and multimodal feature augmentation module. To be more specific, the entity relevance module calculates the probability that two entities (subject and object) are related and filters out the irrelevant noisy nodes. The multimodal affinity generation module firstly calculates the latent similarities among nodes (*i.e.* visual relationships) from the perspectives of three modalities (*i.e.* appearance (A), spatial (S) and linguistic (L) cues) and then aggregates the multimodal similarities to a relational affinity matrix in attention mechanism [7]. Based on the multimodal affinity matrix, node features are enhanced by informative relationship-wise context. The main difference compared to previous work is that we develop an explicit multimodal model to simulate human reasoning ability (*e.g.* analogy) for relationship-wise interaction capturing, which provides a learnable structural prior to make further use of the information contained in relationship coexistence for better relationship detection.

Ours is a GNN-based approach, since GNN has inherent ability to combinatorial generalization for applying shared computations across the nodes as well as their relations. In this fashion, the MSGRIN can adapt to relational dependency modeling in flexible scenarios. The graph paradigm provides a smoothing way that explicitly encourages alignment between similar labels and nodes with similar properties, which reduces the ambiguities caused by large intra-class variance. In addition, by deeply modeling the multimodal affinity, the MSGRIN explicitly captures the multifaceted interactions of relationships to achieve feature refinement for better relationship discrimination.

Our main contributions are summarized as follows: 1) We explore the interactions among relationships by capturing their multimodal affinities and propose a novel Multimodal Similarity Guided Relationship Interaction Network (MSGRIN), which explicitly models the meaningful context among candidate relationships in the scene. 2) The MSGRIN provides an intuitive and relatively interpretable manner to understand visual relationships in changeable scenes by capturing the multimodal relationship-wise hyper-relations. 3) We have conducted sufficient experiments, and the results show that our method is able to capture the relevancy of relationships in the whole scene and achieves the state-of-the-art performance on the two datasets of Visual Relationship Detection [4] and Visual Genome [5].

2. Related work

2.1. Visual relationship detection

Visual relationship detection involves detecting objects in an image and understanding the relations between them. Early methods like Visual Phases [8] classify relationships based on the tuples. If there are N objects and K predicates, then $O(N^2K)$ unique relationships detectors are supposed to be trained. Thus, these methods only function well for certain use where the categories of ob-

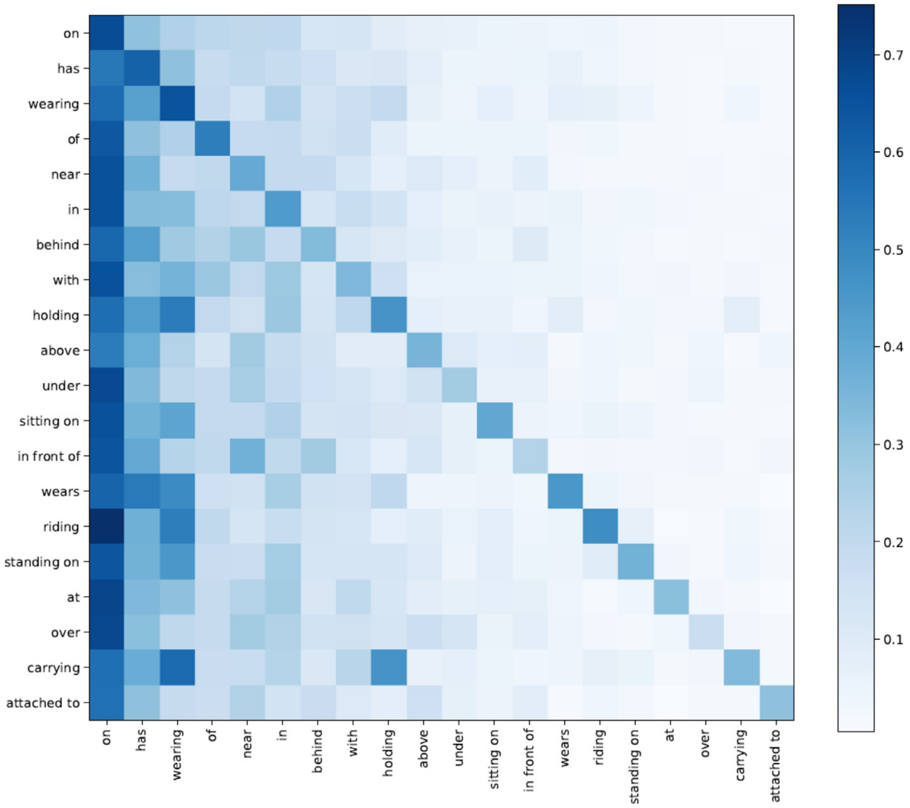


Fig. 2. Coexistence Rate matrix of the top 20 most frequent predicates in VG. Note that the main diagonal is not ones because the same relationship cannot represent the coexistence.

jects and relationships are limited. For instance, the study of Visual Phases [8] only involved a small set of 13 common relationships. Then researchers like Lu et al. introduced more complex datasets and devised an approach to simplify detection models [4]. They have decoupled each tuple into subject-object and predicate, and in this way $O(N+K)$ trained detectors can detect $O(N^2K)$ relationships. Most subsequent studies [9–13] on visual relationship detection as well as our method are based on this approach.

In visual relationship detection, the performance is improved by discovering different modalities. Among them, appearance is the most direct modality used extensively in varies studies, like visual relationship detection [4], semantic segmentation [14], action recognition [15], etc. A typical approach to extract appearance features is to utilize convolutional neural network to encode the full image or part of the image by RoI pooling [16]. Besides appearance, spatial information is considered as a significant modality. Researchers [9–11,17] attempted to learn spatial cues by encoding the absolute and relative location of the objects in an image. Liang et al. explored varies spatial representation approaches [9]. Another modality is the linguistics adopted by [4,9,18,19] to learn relationship from the language, which supplements vision-based approaches. However, these multimodal methods individually detected the relationships and neglected the informative interactions between them. Some recent methods aim to incorporate context information to learn discriminative features for entity and relationship prediction [10,20–23]. Yin et al. introduced contextual information of objects [10]. Motifs [24] investigated the regularly appearing substructures in scene graphs. MR-NET [22] employed the constraint to force the representations of ⟨*sub-obj*⟩ and ⟨*obj-sub*⟩ to be opposite. CISC [23] applied gated recurrent neural networks to memory the relationship context. HCNet [20] aggregated the context from different hierarchies. However, these context-aware methods limited the contextual messages to the entities or the partial relationships, and neglected the fact that the context can be enriched by the complementary multimodal representations. In this work, we focus on multimodal interactions of all candidate relationships in the scene and propose an adaptive and relatively interpretable graph reasoning approach.

2.2. Graph neural network

Graph Neural Network (GNN) is a class of neural network designed to directly operate on the graph structure and it is widely used in natural language processing [25] and computer vision [26,27]. The core component in GNN is the message passing operation. During the message passing phase, hidden states of each node in the graph are updated with the aggregation messages (i.e. interactions) from its neighbors. Justin et al. reformulated existing neural models for graph structured data into a framework called Message Passing Neural Network (MPNN) [28]. Battaglia et al. summarized the GNN methods and pointed out that the good performance is partially due to the combinatorial generalization ability of GNN [6]. Qi et al. proposed a Graph Parsing Neural Network that took humans and objects as nodes and implemented graph inference in Human Object Interaction (HOI) [27]. Jing et al. proposed a Relational Graph Neural Network for situation recognition [29]. When it comes to the scene graph, inter-dependencies of objects and relations are obtained by message passing in the graph. Xu et al. focused on visual context of entities and predicates in the scene and applied message passing with RNNs [30]. Zhou et al. mined the relative location information and constructed a predicate relevance graph based on predefined threshold [31]. Hosenet captured the higher-level behavior patterns of entities based on the connectivity subgraphs [32]. In this paper, we incorporate the complementary information between the three modalities and build dynamic relationship graphs (*i.e.* graphs in which relation-

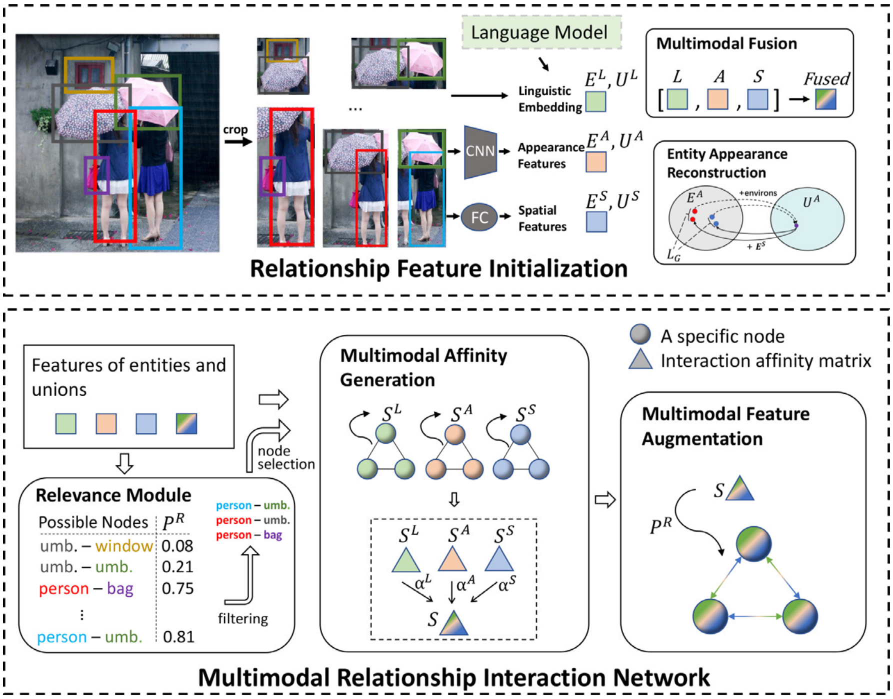


Fig. 3. In Relationship Feature Initialization stage, an object detector (*e.g.* Faster R-CNN) detects entities. Features from three modalities (appearance, spatial and linguistics) are extracted and fused to initialize entities and entity-pairs. In Multimodal Relationship Interaction Network, multimodal features are utilized to generate relevance scores for entity-pairs filtering. Similarities from each modality (S^L, S^A, S^S) are extracted and incorporated to the multimodal affinity (S). The representations of relationships are augmented with their interactions.

ships act as nodes) with adaptive edges to explicitly combine the relationship context for feature refinement.

3. Our approach

We then introduce our proposed Multimodal Similarity Guided Relationship Interaction Network, which contains two phases: Relationship Feature Initialization and Multimodal Relationship Interaction Network. Fig. 3 illustrates the components of the model.

3.1. Multimodal feature extraction

Existing works [9,11] have demonstrated that appearance, spatial and linguistic cues can effectively represent the information of objects and their relationships. The MSGRIN applies multimodal features as well and initially represents the relationships by the fused features. In the following discussion, we denote appearance, spatial and linguistic modality by A, S and L. For instance, spatial features of entities are denoted by E^S. The fused features of entities and pairs are denoted by E and U without superscripts.

– Appearance features. Appearance features are significant to describe visual characteristics. In appearance features extraction, we first detect the bounding boxes of candidate subjects and objects. In our experiment, we utilize Faster R-CNN [16] for this purpose. Then we adopt ResNet-50-FPN [33] as the backbone and extract the RoI Pooling features of entities and subject-object unions from the last convolutional layer. The appearance features of entities and unions are denoted by E^A and U^A, respectively. **– Spatial features.**

Spatial modality information can be used as a supplement to the visual appearance modality which usually represents the relative location between subject and object. Following [11], we employ ($x^s_{\min}$, $y^s_{\min}$, $x^s_{\max}$, $y^s_{\max}$), ($x^o_{\min}$, $y^o_{\min}$, $x^o_{\max}$, $y^o_{\max}$) and ($x^u_{\min}$, $y^u_{\min}$, $x^u_{\max}$, $y^u_{\max}$) to denote the locations of the subject box, the object box and the union box. We denote w^u and h^u as the width and height of the union bounding box where $w^u = x^u_{\max} - x^u_{\min}$ and $h^u = y^u_{\max} - y^u_{\min}$. The spatial feature U^S_{so} is initialized to an 8-dimensional vector as ($\frac{x^s_{\min}-x^u_{\min}}{w^u}$, $\frac{x^s_{\max}-x^u_{\max}}{w^u}$, $\frac{y^s_{\min}-y^u_{\min}}{h^u}$, $\frac{y^s_{\max}-y^u_{\max}}{h^u}$, $\frac{x^o_{\min}-x^u_{\min}}{w^u}$, $\frac{x^o_{\max}-x^u_{\max}}{w^u}$, $\frac{y^o_{\min}-y^u_{\min}}{h^u}$, $\frac{y^o_{\max}-y^u_{\max}}{h^u}$) and transformed to a 128-dimensional feature vector by a two-layer fully connected network. The entity spatial feature E^S_i is initialized in the similar way. The differences are that the aforementioned subject box is replaced by the entity box ($x^i_{\min}$, $y^i_{\min}$, $x^i_{\max}$, $y^i_{\max}$) and the object box is replaced by the entire image (0, 0, W^{img}, H^{img}).

– Linguistic features.

Linguistic representations capture the semantic similarity of different categories of entities with external linguistic knowledge, which provide important supplementary information to the other two modalities. We first predict the categories of entities in the image with Faster R-CNN. The categories are embedded with the language model trained in Wikipedia [34]. Then we input these embeddings to a two-layer feedforward network to obtain the linguistic features. The linguistic features of entities are denoted as E^L and the union linguistic features U^L are obtained by a fully connected layer after concatenation of subjects and objects.

– Feature fusion.

The entity features and the union features are obtained by multimodal feature aggregation. We concatenate the features extracted from three modalities and input them into a two-layer MLP network for multimodal fusion. Both entity features and union features are 256-dimensional after transformation following [9,35]. The multimodal fusion procedures for the entity i (denoted by E_i) and the union of subject s and object o (denoted by U_{so}) are formulated as:

$$E_i = f_E([E^A_i, E^S_i, E^L_i]), \tag{1}$$

$$U_{so} = f_U([U^A_{so}, U^S_{so}, U^L_{so}]), \tag{2}$$

where f_E, f_U are the aggregation functions for entity fusion and union fusion respectively; $[\cdot, \cdot]$ denotes the concatenation operation.

3.2. Entity appearance reconstruction

The union appearance features U^A_{so} incorporate the visual context between the subject and the object. However, directly encoding the union region is susceptible to the noise from irrelevant background. The main reason is that the region sizes of the entities and the background can be huge different. For example, in the union bounding box of the relationship “person-fly-kite”, the “person” box and the “kite” box can be far apart that the background region can be several times greater than the entities. In such cases, the extracted RoI features may overemphasize the background contents and ignore the entity pairs. To mitigate the issue, we prompt the union appearance features to adequately encode the appearance of entity pairs. We apply a constraint in the training phase that the entity appearance features E^A can be reconstructed from the union appearance representations U^A and the entity spatial features E^S. The reconstruction constraint promotes the union features to focus more on the entity pairs and thus can relatively mitigate the irrelevant noise from the background. We find that the reconstruction constraint brings considerable improvement in our method (see ablation results in Section 4.4). The recovered entity appearance G is generated by a fully-connected network f_G:

$$G_s = f_G([U^A_{so}, E^S_s]) \approx E^A_s, \tag{3}$$

$$G_o = f_G([U^A_{so}, E^S_o]) \approx E^A_o. \tag{4}$$

3.3. Multimodal similarity guided relationship interaction network

We develop a graph-based multimodal similarity guided relationship interaction network (MSGRIN) to obtain relationship-wise context aggregation. In the graph, candidate relationships (subject-object pairs) are modeled as nodes. Inspired by the form of dense conditional random field which is widely used in visual recognition [14,36], we take unary and pairwise factors into account and design three modules in the MSGRIN: entity relevance module to calculate the existences of nodes; multimodal affinity generation module to obtain affinity matrices from multimodal cues; and multimodal feature augmentation module to achieve the refinement of node features.

3.3.1. Entity relevance module

Visual relationship detection can be decoupled into a two-stage decision problem when subjects and objects are given: firstly to determine whether there exist relationships between them and then to classify what the relationships are. Solving the relevance task is beneficial to the classification task, for its ability to eliminate the entity pairs without any relationship. Unlike most existing approaches that tackle the detection task directly with a final classification loss (*e.g.* multi-label binary cross-entropy loss), our method explicitly integrates the binary relevance classification task.

We obtain relevance embedding R^{emb} between two entities: E_i and E_j through a relevance encoder f^{emb}_R on their concatenated features:

$$R^{emb}_{ij} = f^{emb}_R([E_i, E_j]). \tag{5}$$

The relevance embedding is then fed to a binary classifier to generate predictions of relevance probability P^R. The relevance probability between entity i and entity j is:

$$P^R_{ij} = \sigma(f^{cls}_R(R^{emb}_{ij})), \tag{6}$$

where σ is the sigmoid function.

The entity relevance labels can be generated by the relationship annotations. In the training phase, we directly select the ground-truth pairs for further detection and the relevance module updates its parameters with the entity relevance labels. In the inference phase, based on P^R, the entity pairs with less relevant confidences are filtered out (*i.e.* below to *thres_relevance*), and the top N^R pairs are proposed as the candidate subject-object pairs for further detection.

3.3.2. Multimodal affinity generation

During the propagation in the constructed fully connected graph, node features are updated with their interactions. We model the communication among nodes in a similarity guided [37] mode in which the affinities among nodes are represented by their latent similarities and as the weights to control the interaction message passing. To fully and explicitly incorporate the multimodal information, latent similarities from all the three modalities are extracted and fused. Fig. 4(a) illustrates the process of calculating the affinity matrix.

For each modality $m \in \{A, S, L\}$, we employ monomodal features U^m to generate similarity matrix S^m from the perspective of modality m:

$$S^m_{i'j'} = cos(f^m_S(U^m_{i'}), f^m_S(U^m_{j'})), \tag{7}$$

where i', j' are subscripts for relationships. In Eq. (7), $cos(\cdot,\cdot)$ is the cosine similarity function and f^m_S denotes the transformation to latent similarity measuring space.

Due to the ambiguity of predicate as well as the variance of angle and position, pairs with strong relations do not demonstrate strong associations in all modalities. Thus, max pooling or average pooling may not perform well for similarity fusion. In this work, we devise the Multimodal Similarity Attention based on the importance from different modalities to the multimodal fusion. Inspired by [7,38], the importance score of modality m is computed according to its relevancy to the fused features:

$$e^m = f_\theta(U^m)^T f_\phi(U), \tag{8}$$

where U^m is the aforementioned mono-modal features of modality m and U is the fused features. f_θ and f_ϕ are fully-connected networks for feature space transformation.

Then the importance scores are normalized via a Softmax function to calculate the similarity attention:

$$\alpha^m = \frac{\exp(e^m)}{\exp(e^A) + \exp(e^S) + \exp(e^L)}. \tag{9}$$

The multimodal similarity is finally computed with the attention weights:

$$S = \sum_m \alpha^m \cdot S^m. \tag{10}$$

The attention mechanism obtains dependencies between each modality and the multimodal relationship representations. The modality with higher dependency contributes more to the multimodal similarity.

3.3.3. Multimodal feature augmentation

When we obtain the relevance probabilities and multimodal similarities among relationships, the message to node i' is calculated by encoding and aggregation which is formulated as:

$$M_{i'} = \frac{1}{C_{i'}} P^R_{i'} \sum_{j'} P^R_{j'} S_{i'j'} U_{j'} W^{enc}, \tag{11}$$

where W^{enc} is the learnable message encoding matrix and $C_{i'}$ is the normalization factor. The relevance probability P^R and similarity matrix S jointly guide the message passing.

In our implementation, we use matrix multiplication for computational efficiency. The message matrix of the graph can be formulated in the form of typical graph convolutional network (GCN) [39]:

$$M = D^{-1}\tilde{S}UW^{enc}. \tag{12}$$

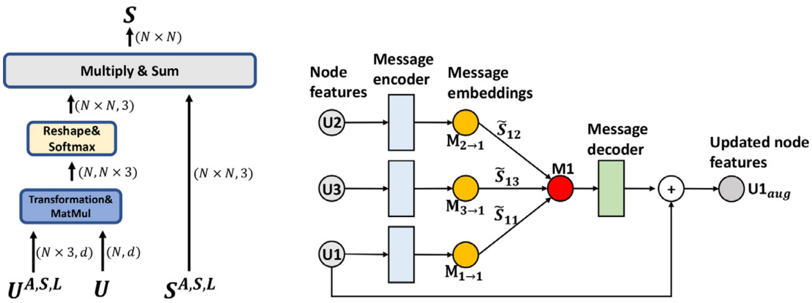


Fig. 4. Illustrations of Multimodal Similarity Attention (left) and Interaction Aggregation (right). In the left subfigure, the shapes are listed to the right of arrows. N and d denote the number and the dimension of nodes. The right subfigure illustrates the augmentation procedure for a node in the three-node graph, as formulated in Eq. (12) and Eq. (13).

In Eq. (12), $\tilde{S} = PP^T \odot S$ is the relevance weighted similarity matrix of the graph, where $\odot$ is the Hadamard product and P^T is the transpose of P. D is the normalization diagonal matrix where $D_{i'i'} = \sum_{j'} \tilde{S}_{i'j'}$ and W^{enc} is the encoding matrix as in Eq. (11). The aggregated messages are decoded via a decoder and node features are updated by the following formula:

$$U_{aug} = U + MW^{dec}. \tag{13}$$

The message aggregation procedure of the MSGRIN is illustrated in Fig. 4(b). The augmented features are finally input to a multi-class classifier to get the predicate scores of subject-object pairs.

3.4. Optimization

In this subsection, we introduce our multi-task setting for training the MSGRIN. Multiple loss functions are applied for tasks of entity category classification, entity-wise relevance classification and relationship detection.

In entity category classification and entity-wise relevance task, we supervise the model with cross-entropy loss:

$$\begin{aligned} \mathcal{L}_E &= CE(P^E, y^E), \\ \mathcal{L}_R &= CE(P^R, y^R), \end{aligned} \tag{14}$$

where P^E and P^R are predictions from object detection and entity-wise relevance module, y^E and y^R are the corresponding labels.

The relationships in visual scenes are too many to be completely labelled and thus there are many truly existing relationships without annotations.The wide-used binary cross-entropy loss might over-suppress the unlabeled relationships. In our method, we apply the structural ranking loss modified from [9], in which $\mathcal{R}$ denotes the set of all annotated relationships in the image. The baseline objective function is formulated as follows:

$$\mathcal{L}_b = \sum_{r \in \mathcal{R}} \sum_{r' \notin \mathcal{R}} [\Phi(r') - \Phi(r) + \Delta(r, r')]_+, \tag{15}$$

where $[\cdot]_+$ represents for $max(0, \cdot)$, $\Phi(\cdot)$ denotes the scores obtained by the MSGRIN and the adaptive margin $\Delta(r, r')$ is computed by the statistical prior probability from the training set as follows:

$$\Delta(r, r') = 1 + Prior(r|s, o) - Prior(r'|s, o). \tag{16}$$

Also, we incentivize the entity appearance generation using the L1 penalty:

$$\mathcal{L}_G = \frac{1}{N} \sum_{s,o} \left\| G_s - E_s^A \right\|_1 + \left\| G_o - E_o^A \right\|_1, \tag{17}$$

where N denotes the number of labeled subject-object pairs in the image.

Finally, a joint loss function is utilized for optimization:

$$\mathcal{L}_{joint} = \mathcal{L}_b + \mathcal{L}_E + \mathcal{L}_R + \alpha \mathcal{L}_G, \tag{18}$$

where α trades off the loss.

4. Experiment

4.1. Experiment setting

– Datasets. We use two public datasets for model validation: the Visual Relationship Detection dataset [4] and the Visual Genome dataset [5].

The Visual Relationship Detection (VRD) dataset consists of 5000 images with 100 object categories and 70 predicate categories. There are 37,993 relationships within 6672 types in VRD in total, with 4000 training images and 1000 test images.

The Visual Genome (VG) dataset is one of the largest relationship detection datasets. The origin VG consists of over 5319 object categories, 1957 predicates and 421,697 relationship types. We evaluate our approach on the widely used split of VG following [24,30,40], which contains 89,169 images with 150 object categories and 50 predicates.

– Evaluation metrics. On VRD dataset, we evaluate our approach in Predicate Detection (Pre.), Phrase Detection (Phr.) and Relationship Detection (Rel.) proposed by Lu et al. [4] in comparison with previous works. Top k predictions for each subject-object pair are taken into consideration. On VG dataset, we evaluate on subtasks of Predicate Classification (PredCls), Scene Graph Classification (SGCls) and Scene Graph Detection (SGDet) with graph constraint following MOTIFS [24]. According to the previous setting, we use Recall as our evaluation metric. The top N recall is denoted as $R@N$. More specifically, $R@N$ is the fraction of ground-truth instances that are correctly recalled in top N predictions from each image and we set N to 50 and 100 in our experiments. Also, by following [41], we respectively compare the results on three high-level types of geometric relationships, possessive relationships and semantic relationships in the PredCls on VG dataset to further evaluate the performance of our method.

4.2. Implementation details

In the experiments, our model uses ImageNet [42] pretrained ResNet-50-FPN as the backbone following [31,43] to extract appearance features. We freeze the weights of all convolutional layers in the pretrained backbone and apply instance normalization to the last convolutional layer. We utilize a pretrained word2vec [34] model to project the entity categories into linguistic embeddings. In our model, we implement f_E, f_U, f_R^{emb}, f_R^{cls}, f_G and f_S functions by two-layer fully-connected networks with leaky rectified linear unit [44]. The affinity matrix S is shrunk by L1 regularization. We use Adam optimizer to train our model with initial learning rate 0.001 and weight decay 0.0005. For relationship detection, we detect object proposals with Faster R-CNN [16] and apply NMS to select at most 50 boxes from the proposals with classification probabilities threshold of 0.3 and IoU threshold of 0.5.

Table 1
Performance (%) comparison on VRD dataset. “-” denotes that the result is unavailable.

	Pre.		Phr.		Rel.	
Model	R@50,	R@50/100,	R@50,	R@50/100,	R@50,	R@50/100,
	k=1	k=70	k=1	k=70	k=1	k=70
VRD-Full[4]	47.9	71.0/84.3	16.2	20.1/24.9	13.9	17.4/21.5
LKD[18]	55.2	85.6/94.7	23.1	26.3/29.4	19.2	22.7/31.9
DSR[9]	-	86.0/93.2	-	- -	-	19.0/23.3
ZoomNet[10]	56.0	89.0/94.6	25.2	29.6/38.4	19.5	22.3/28.5
MF-URLN[11]	58.2	- -	31.5	- -	23.9	- -
RLM[31]	57.1	90.0/96.5	33.2	36.8/46.0	26.5	30.2/37.4
MLA-VRD[35]	-	90.2/95.0	-	23.4/28.1	-	20.5/24.9
HOSE-Net[32]	-	- -	27.0	28.9 36.2	20.5	22.1 27.4
GPS-Net[45]	**58.7**	- -	28.9	- -	21.5	- -
Ours (V)	57.8	90.7/96.6	32.1	36.1/44.6	26.1	29.6/36.6
Ours	57.9	**91.0/96.9**	**33.8**	**37.4/47.3**	**27.2**	**30.8/38.0**

Table 2
Performance (%) comparison on VG dataset. “-” denotes that the result is unavailable.

Model	PredCls		SGCls		SGDet	
	R@50	R@100	R@50	R@100	R@50	R@100
VRD[4]	27.9	35.0	11.8	14.1	0.3	0.5
IMP[30]	44.8	53.0	21.7	24.4	3.4	4.2
GraphRCNN[40]	54.2	59.1	29.6	31.6	11.4	13.7
CISC[23]	53.2	57.9	27.8	29.5	11.4	13.9
MotifNet-LeftRight[24]	65.2	67.1	35.8	36.5	27.2	30.3
MR-NET[22]	65.3	66.5	-	-	12.6	14.3
HCNet[20]	66.4	68.8	36.6	37.3	28.0	31.2
MLA-VRD[35]	67.1	69.2	36.0	37.4	28.1	32.9
GPS-Net[45]	66.9	68.8	39.2	40.1	28.4	31.7
HOSE-Net[32]	66.7	69.2	36.3	37.4	**28.9**	33.3
Ours	**68.5**	**69.4**	**39.4**	**40.2**	28.8	**33.5**

Random translation and scaling within 5% are applied to region proposals for data augmentation. We set *thres_relevance* to 0.25 and N^R to 200 for entity pairs filtering. It takes about 20 epochs to converge on VRD and about 12 epochs on VG. Our model is implemented using PyTorch and trained with a single Nvidia Titan X GPU. Trade-off factor α in Eq. (18) is set to 0.5 in our experiments without further tuning.

4.3. Comparative results

In this subsection, we compare with the state-of-the-art methods to highlight the advantages of the MSGRIN. Firstly, we compare the proposed MSGRIN with nine methods on the VRD dataset. The compared methods include: VRD-Full [4], LKD [18], DSR [9], ZoomNet [10], MF-URLN [11], RLM [31], MLA-VRD [35], HOSE-Net [32] and GPS-Net [45]. In “Ours(V)”, we apply ImageNet pretrained VGG16 as the backbone in Faster R-CNN. The results are summarized in Table 1 and the best results are highlighted in boldface. In comparison to previous state-of-the-art approaches, the MSGRIN improves the recall of predicate detection by 0.8% $R@50, k = 70$ and 0.4% $R@100, k = 70$ compared with the second best. Besides, the MSGRIN achieves the state-of-the-art performance in both phrase detection and relationship detection tasks, even with the vanilla VGG16 backbone. With the VGG16 backbone, our approach still performs well in predicate detection, which indicates the effectiveness of multimodal hyper-relation modeling for feature enhancement. Meanwhile, due to the worse location results from object detection, the performances in phrase detection and relationship detection drop visibly. The results in Table 1 demonstrate the superiority of the MSGRIN compared to previous relationship-independent or monomodal context modeling approaches.

Table 2 reports the comparison of the MSGRIN with existing state-of-the-art methods on the VG dataset. The compared methods are VRD [4], IMP [30], GraphRCNN [40], CISC [23], MOTIFS [24], MR-NET [22], HCNet [20], MLA-VRD [35], GPS-Net [45] and HOSE-Net [32]. The best results are highlighted in bold font. As shown in the results, the MSGRIN achieves the state-of-the-art. From the results, we can see that the improvement in SGDET is not obvious. The main reason is that the larger VG dataset contains more entities per image as well as more unlabeled relationships. As the MSGRIN aggregates relevant relationships, more unlabeled relationships are detected. As illustrated in Fig. 5, the relationships in green are part of the relationship ground-truths in VG, while the red relationships detected by our method (with the top-50 confidence) are not in ground-truths (*i.e.* not counted as correct predictions).

The results in Table 2 also reveal that the MSGRIN achieves great improvement compared to the recurrent neural network (RNN) based implicit context modeling methods [23,24,30]. The reason may be that our method avoids the context forgetting and the permutation variability of RNNs and provides a suitable structural prior for relationship-wise context modeling. Our approach also stands out from the advanced context modeling methods [20,32], which indicates the strength of incorporating multimodal relationship context in scenes.

Since it’s difficult to build a dataset containing every possible relationship, a visual relationship detection model should be able to perform zero-shot predictions of relationships it has never seen before. In Table 3, we compare our approach with existing methods on VRD zero-shot split [4]. The compared methods are LKD [18], DSR [9] RLM [31] and MLA-VRD [35]. As shown in Table 3, the MSGRIN achieves the best in all subtasks on unseen sets of the VRD dataset, where the relationship triplets are not included in the training set. Our method makes improvement of 1.6% $R@50$ in predicate detection and 1.4% $R@50$ in relationship detection. The results on the zero-shot set suggest that learning the relationship

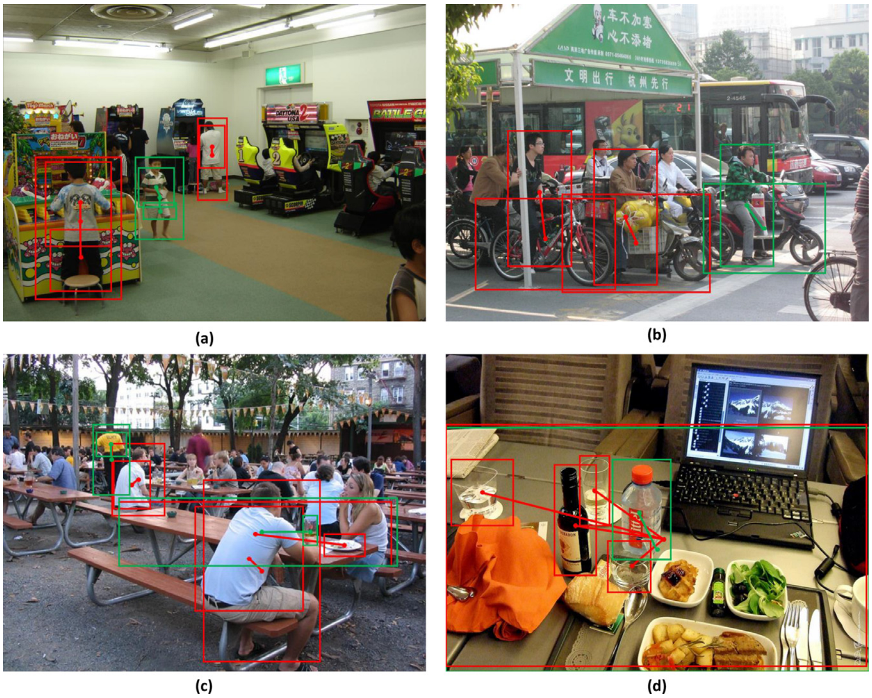


Fig. 5. Illustration examples of incomplete annotations on VG. In the subfigures, green pairs are part of the ground-truths. Red pairs are part of the detected relationships by the MSGRIN, which are counted as incorrect predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Zero-shot performances (%) comparison on VRD dataset. (k=70) .

Model	Pre.		Phr.		Rel.	
	R@50	R@100	R@50	R@100	R@50	R@100
LKD[18]	54.20	74.65	12.96	17.24	12.02	15.89
DSR[9]	60.90	79.81	-	-	5.25	9.20
RLM[31]	72.03	88.11	-	-	-	-
MLA-VRD[35]	73.65	88.96	8.43	13.84	8.08	12.81
Ours	**75.28**	**89.15**	**15.32**	**20.47**	**13.41**	**17.63**

Table 4
Mean recall@50 (in %) of the geometric relationships (*gR*), the possessive relationships (*pR*), the semantic relationships (*sR*) and the average comprehensive recall (*CR*) in the PredCls task of VG.

Model	PredCls			
	gR@50	pR@50	sR@50	CR@50
IMP+[24]	9.05	23.48	3.33	11.95
MOTIFNET[24]	14.69	26.38	6.79	15.95
GFL+Depth [41]	17.01	28.03	12.40	19.15
GFL (fully) [41]	17.37	28.07	**14.55**	20.00
Baseline (Ours)	14.59	26.23	6.12	15.65
MSGRIN (Ours)	**18.23**	**29.42**	13.52	**20.39**

Table 5
Component Analysis (%) on VRD dataset.

	Pre.			Rel.		
	R@50, k=1	R@50, k=70	R@100, k=70	R@50, k=1	R@50, k=70	R@100, k=70
Ablation on model architecture						
Baseline	53.04	86.95	92.68	17.37	21.02	24.79
Baseline+R.	54.16	88.46	94.33	23.44	25.39	27.64
M. w/o. R	56.32	89.21	95.15	21.17	23.48	26.81
MSGRIN	**57.91**	**90.98**	**96.88**	**27.19**	**30.81**	**37.96**
Ablation on the loss function (Eq. (18))						
w/o. $\mathcal{L}_G$	57.46	89.89	96.29	26.01	29.11	35.63
w/o. $\mathcal{L}_R$	56.58	89.74	95.91	22.41	24.53	28.18
Ablation on multi-modality interaction						
w/o. I-A.	56.74	89.82	96.30	26.33	28.66	35.85
w/o. I-S.	56.51	89.45	95.94	25.88	27.56	33.68
w/o. I-L.	56.90	89.96	96.24	26.63	29.14	36.28
Comparison on different affinity modes						
max pooling	56.87	88.75	95.16	24.72	26.75	35.17
fused feat.	57.43	89.84	94.92	25.33	28.34	36.81

interactions in our paradigm provides the combinatorial generalization ability and motivates to reason unseen relationships from existing ones.

To further evaluate the performance, we also report the mean *R*@50 of the geometric relationships, the possessive relationships and the semantic relationships. In our baseline architecture, the multimodal features are directly input to the classifier without interaction refinement. As shown in Table 4, the MSGRIN makes clear improvement to the baseline and outperforms the advanced methods consistently. It is worth noting that our model does not apply focus loss as in [41] for the rare relationship instances. The results reveal that our method can better learn the underlying relationship-wise context from relation co-occurrences and effectively extract high-level associations to enhance different kinds of relationships.

4.4. Component analysis

We conduct a series of experiments in different settings to investigate the importance of various components in the MSGRIN; results are shown in Table 5. In Table 5, we show the ablation results on the model architecture with the joint loss $\mathcal{L}_{joint}$ (Eq. (18)), on the loss function with the full MSGRIN architecture and on the multimodal interactions. We also compare the results of different multimodal affinity generation modes.

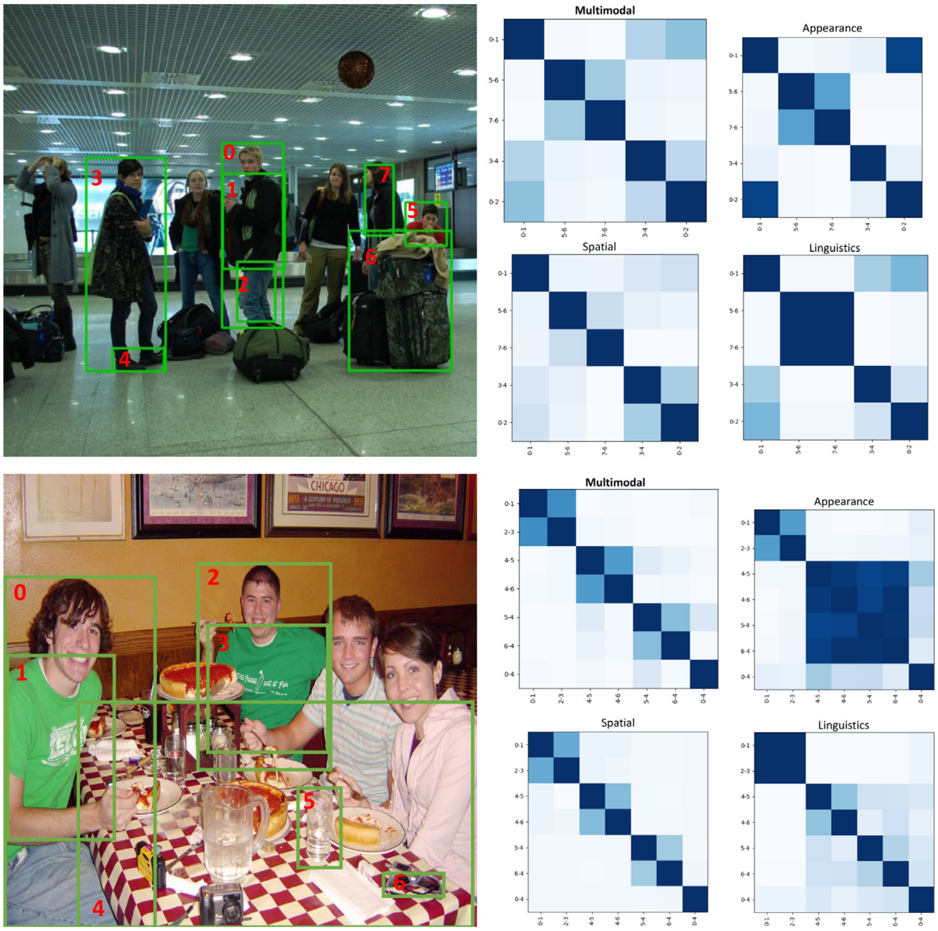


Fig. 6. Visualization examples of multimodal interactions (*i.e.* S, S^A, S^S and S^Lin Eq. (10)) in predicate detection on VRD. In the affinity matrices on the right, coordinates refer to subject-object pairs in the image. Please zoom in for better comparison.

In the baseline architecture, we directly pass the multimodal features to the classifier without feature refinement. The relevance module is applied to the baseline in "Baseline+R.", and "M. w/o R." stands for the MSGRIN without entity relevance module. The results suggest that capturing the interactions among relationships ("MSGRIN") brings at least 2% improvement in predicate detection and 4% in relationship detection compared to the baseline ("Baseline+R."). In addition, the removal of entity relevance module does harm the performance of MSGRIN especially in relationship detection, which shows that filtering out a large number of irrelevant entity pairs by the entity relevance module plays an important role in our approach. Also, it indicates that the entity relevance acts as an effective supplement to the multimodal interaction aggregation (Eq. (11)).

As shown in the ablation results of loss function, applying the appearance reconstruction loss $\mathcal{L}_G$ and the entity-wise relevance loss $\mathcal{L}_R$ respectively gives more that 1.1% and 4.7% improvement in relationship detection. The results reveal that our multi-task setting provides effective supervision on the MSGRIN.

We also perform the ablation experiments on the three modalities in multimodal relationship interactions. The $R@50, k = 70$ results (%) in VRD Relationship Detection decrease from 30.81 to 28.66, 27.56 and 29.14 when we respectively eliminate the appearance, spatial and linguistic term in Eq. (9) and Eq. (10) (*e.g.* "w/o. I-A" means S^A is omitted). The results reveal the importance to incorporate multimodal affinities for relationship interaction modeling.

Moreover, we compare the other two ways for multimodal affinity generation. The "max pooling" takes the maximum similarity among three modalities and the "fused feat." directly applies the fused multimodal features to obtain the relationship-wise affinity by a two-layer multilayer perceptron. The results show that using either the maximum strategy or the implicit neural network for multimodal relationship interaction fusion decreases the performance in all tasks compared to the multimodal attention manner.

4.5. Qualitative results

Fig. 6 shows some visualization examples of interactions captured by the MSGRIN on the VRD dataset. In the affinity matrices on the right, coordinates refer to subject-object pairs in the image, *e.g. 0–1* in the top row represents the *person-jacket* in the middle of the image. Grids in darker color correspond to stronger interactions (S, S^m in Eq. (10)). As shown in the top matrix, in addition to the pair itself, the *person-jacket* (*0–1*) is most related to the *person-pants* (*0–2*) and the *person-luggage* (*7-6*) is most related to the *person-luggage* (*5–6*). Meanwhile, pairs like *0–1* and *0–4* in the bottom row share small linking weights, which limits the message passing and motivate them to remain independent to some extent. From the results, we can see that our model can effectively obtain the interactions among entity pairs.

We also present some qualitative examples of predicate detection on VRD in Fig. 7. Compared to the baseline, the MSGRIN can benefit from the context among relationships and aggregate the relationships with similar semantic or spatial patterns.

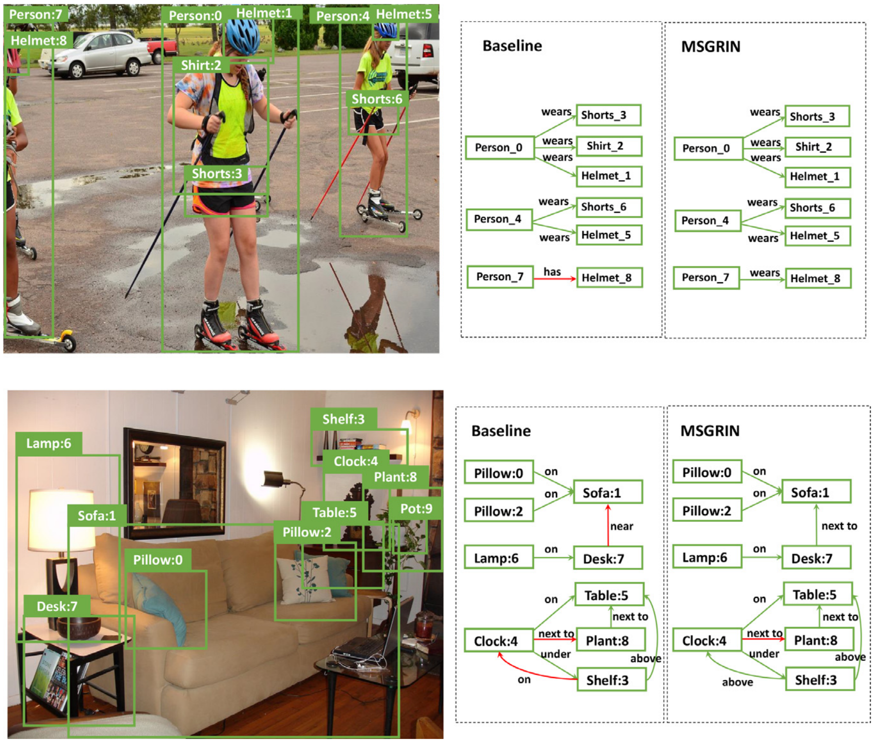


Fig. 7. Examples of top-5 predicate detection results on the VRD dataset. The correct predictions are shown in green. The relationship instance *clock-next to-plant* in the bottom is not in ground truths due to the lack of annotation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

We presented a GNN based multimodal similarity guided relationship interaction network (MSGRIN) which takes subject-object pairs as nodes to explore the role of relationship interaction for visual relationship detection. Compared to existing context modeling methods in visual relationship detection, the MSGRIN directly constructs interactions between arbitrary relationships from the perspectives of three modalities, and thus better benefits from the scene context. We designed relevance module, entity appearance reconstruction and multimodal affinity to mitigate the adverse effects from noisy nodes. Experimental results on two public datasets VRD and VG verify that the MSGRIN is able to dig out the context implied by the relationship co-occurrence and achieve better performance. The MSGRIN presents a relatively interpretable manner by analogy reasoning with GNN. Subsequent works could also benefit from the relationship interaction with the proposed plug-and-play network.

Although our method has made some progress by capturing multimodal relationship-wise interactions, it needs more improvement in detecting rare semantic relationships. The reason may be that the long-tailed annotations bring serious bias to the interaction learning and the sophisticated semantic patterns are still underexplored. In future, we are going to introduce label-distribution-aware designs and fine-grained body part interactions to mitigate the problem. Future work may also focus on a one-stage framework that simultaneously optimizes the object detection network with the hyper-relations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002).

References

[1] Y. Xi, Y. Zhang, S. Ding, S. Wan, Visual question answering model based on visual relationship detection, Signal Process. Image Commun. 80 (2020) 115648.
[2] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 684–699.
[3] S. Yoon, W.Y. Kang, S. Jeon, S. Lee, C. Han, J. Park, E.-S. Kim, Image-to-image retrieval by learning similarity between scene graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 10718–10726.
[4] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 852–869.
[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73. Https://doi.org/10.1007/s11263-016-0981-7.
[6] P.W. Battaglia, J.B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al., Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261 (2018).
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[8] M.A. Sadeghi, A. Farhadi, Recognition using visual phrases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1745–1752.
[9] K. Liang, Y. Guo, H. Chang, X. Chen, Visual relationship detection with deep structural ranking, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7098–7105.
[10] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, C. Change Loy, Zoom-net: Mining deep feature interactions for visual relationship recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2018, pp. 322–338.
[11] Y. Zhan, J. Yu, T. Yu, D. Tao, On exploring undetermined relationships for visual relationship detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 5128–5137.

[12] K. Tang, H. Zhang, B. Wu, W. Luo, W. Liu, Learning to compose dynamic tree structures for visual contexts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 6619–6628.

[13] J. Luo, J. Zhao, B. Wen, Y. Zhang, Explaining the semantics capturing capability of scene graph generation models, Pattern Recognit. 110 (2021) 107427. Https://doi.org/10.1016/j.patcog.2020.107427.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, in: 3rd International Conference on Learning Representations, Conference Track Proceedings, 2015.

[15] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recognit. 47 (10) (2014) 3343–3361. Https://doi.org/10.1016/j.patcog.2014.04.018.

[16] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149. Https://doi.org/10.1109/TPAMI.2016.2577031.

[17] C. Han, F. Shen, L. Liu, Y. Yang, H.T. Shen, Visual spatial attention network for relationship detection, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 510–518.

[18] R. Yu, A. Li, V.I. Morariu, L.S. Davis, Visual relationship detection with internal and external linguistic knowledge distillation, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 1068–1076.

[19] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, Visual translation embedding network for visual relation detection, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 5532–5540.

[20] G. Ren, L. Ren, Y. Liao, S. Liu, B. Li, J. Han, S. Yan, Scene graph generation with hierarchical context, IEEE Trans. Neural Networks Learn. Syst. 32 (2) (2021) 909–915. Https://doi.org/10.1109/TNNLS.2020.2979270.

[21] R. Li, S. Zhang, B. Wan, X. He, Bipartite graph network with adaptive message passing for unbiased scene graph generation, arXiv preprint arXiv:2104.00308 (2021).

[22] Y. Bin, Y. Yang, C. Tao, Z. Huang, J. Li, H.T. Shen, Mr-net: exploiting mutual relation for visual relationship detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8110–8117.

[23] W. Wang, R. Wang, S. Shan, X. Chen, Exploring context and visual pattern of relationship for scene graph generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 8188–8197.

[24] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 5831–5840.

[25] R. Palm, U. Paquet, O. Winther, Recurrent relational networks, in: Advances in Neural Information Processing Systems, 2018, pp. 3368–3378.

[26] G. Monfardini, V. Di Massa, F. Scarselli, M. Gori, Graph neural networks for object localization, in: Proceedings of European Conference on Artificial Intelligence, IOS Press, 2006, pp. 665–669.

[27] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2018, pp. 401–417.

[28] G. Justin, S.S. Schoenholz, P.F. Riley, V. Oriol, G.E. Dahl, Neural message passing for quantum chemistry, in: International Conference on Machine Learning, JMLR. org, 2017, pp. 1263–1272.

[29] Y. Jing, J. Wang, W. Wang, L. Wang, T. Tan, Relational graph neural network for situation recognition, Pattern Recognit. 108 (2020) 107544. Https://doi.org/10.1016/j.patcog.2020.107544.

[30] D. Xu, Y. Zhu, C.B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 5410–5419.

[31] H. Zhou, C. Zhang, C. Hu, Visual relationship detection with relative location mining, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 30–38.

[32] M. Wei, C. Yuan, X. Yue, K. Zhong, Hose-net: Higher order structure embedded network for scene graph generation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1846–1854.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 2117–2125.

[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, Workshop Track Proceedings, 2013.

[35] S. Zheng, S. Chen, Q. Jin, Visual relation detection with multi-level attention, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 121–129.

[36] J. Oh, H.-I. Kim, R.-H. Park, Context-based abnormal object detection using the fully-connected conditional random fields, Pattern Recognit. Lett. 98 (2017) 16–25. Https://doi.org/10.1016/j.patrec.2017.08.003.

[37] Y. Shen, H. Li, Y. Shuai, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in: Proceedings of European Conference on Computer Vision, volume 11219, Springer, 2018, pp. 508–526.

[38] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 7794–7803.

[39] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, Conference Track Proceedings, 2017.

[40] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph R-CNN for scene graph generation, in: Proceedings of the European Conference on Computer Vision, volume 11205, Springer, 2018, pp. 690–706.

[41] J. Jiang, Z. He, S. Zhang, X. Zhao, J. Tan, Learning to transfer focus of graph neural network for scene graph parsing, Pattern Recognit. 112 (2021) 107707. Https://doi.org/10.1016/j.patcog.2020.107707.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[43] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 3716–3725.

[44] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, volume 30, Citeseer, 2013, p. 3.

[45] X. Lin, C. Ding, J. Zeng, D. Tao, Gps-net: Graph property sensing network for scene graph generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 3746–3753.

Zhixuan Liu received the bachelors degree in software engineering from Sun Yat-Sen University in 2018. He is pursuing Ph.D. degree with the School of Computer Science and Engineering in Sun Yat-sen University. His research interests are computer vision and person/object association.

Wei-Shi Zheng is now a Professor with Sun Yat-sen University. He has now published more than 160 papers, including more than 120 publications in main journals (TPAMI, TNN/TNNLS, TIP, TSMC-B, PR) and top conferences (ICCV, CVPR, IJCAI, AAAI). His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. Especially, Dr. Zheng has active research on person re-identification in the last five years. He serves a lot for many journals and conference, and he was announced to perform outstanding review in recent top conferences (ECCV 2016 & CVPR 2017). He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He has ever served as an area chairs/SPC many conferences (such as CVPR, ICCV, BMVC, IJCAI and AAAI). He is an IEEE MSA TC member. He is an associate editor of Pattern Recognition. He is a recipient of Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship of United Kingdom.