Configurable Graph Reasoning for Visual Relationship Detection

Yi Zhu, Xiwen Liang, Bingqian Lin, Qixiang Ye, *Senior Member, IEEE*, Jianbin Jiao, *Member, IEEE*, Liang Lin, *Senior Member, IEEE*, and Xiaodan Liang, *Member, IEEE*

***Abstract*— Visual commonsense knowledge has received growing attention in the reasoning of long-tailed visual relationships biased in terms of object and relation labels. Most current methods typically collect and utilize external knowledge for visual relationships by following the fixed reasoning path of {subject, object → predicate} to facilitate the recognition of infrequent relationships. However, the knowledge incorporation for such fixed multidependent path suffers from the data set biased and exponentially grown combinations of object and relation labels and ignores the semantic gap between commonsense knowledge and real scenes. To alleviate this, we propose configurable graph reasoning (CGR) to decompose the reasoning path of visual relationships and the incorporation of external knowledge, achieving configurable knowledge selection and personalized graph reasoning for each relation type in each image. Given a commonsense knowledge graph, CGR learns to match and retrieve knowledge for different subpaths and selectively compose the knowledge routed path. CGR adaptively configures the reasoning path based on the knowledge graph, bridges the semantic gap between the commonsense knowledge, and the real-world scenes and achieves better knowledge generalization. Extensive experiments show that CGR consistently outperforms previous state-of-the-art methods on several popular benchmarks and works well with different knowledge graphs. Detailed analyses demonstrated that CGR learned explainable and compelling configurations of reasoning paths.**

***Index Terms*— Graph learning, scene graph generation, visual reasoning, visual relationship detection (VRD).**

I. Introduction

VISUAL relationship detection is required to predict high-level semantic relationships for pairs of detected objects. These relationships could help construct a structured abstraction of the visual scenes, benefiting down-stream tasks, such as visual question answering (VQA) [1], image captioning [2], and image retrieval [3]. In general, visual relationships in real-world scenes are long-tailed distributions with very few samples existing for rare classes. Some works follow the fixed reasoning path of {subject, object → predicate}, abridged as {s, o → p} and take full use of the path's knowledge connections to facilitate recognition of infrequent relationships with limited training data. Their knowledge graphs are summarized from data set-specific statistics to provide initial relationship proposals [4], [5] or fixed bias for the relation classifier [6], [7]. Lack of selecting connections from knowledge graphs often makes these works fail to leverage the commonsense knowledge, which collects more comprehensive relationships but contains more noise. Other works also exploit to distil [8] or retrieve knowledge [9] from knowledge graphs which containing many redundant connections. In addition, the limitation remains obvious due to the fixed multidependent reasoning path, where knowledge retrieval suffers from exponentially grown combinations of labels. The retrieved knowledge connections are not general enough to provide highly predictable evidence for the reasoning of visual relationships. Their performance relies on the quality and capacity of the knowledge graphs.

To explore more common and stable reasoning paths for knowledge incorporation, we propose the configurable graph reasoning (CGR) that employs a set of subpaths, i.e., {s → p}, {o → p}, {s → o}, and {o → s}, to retrieve and match knowledge connections, respectively, and then learns to dynamically compose the knowledge-guided paths for the reasoning of each visual relationship. Specifically, a directed relation graph is built by connecting entities (object and predicate labels) based upon the semantic correlations among them. Then, following multiple subpaths, features of object or relation are first projected to match the entities of the knowledge graph, after which these features are updated in a global perspective by the knowledge connections among the corresponding entities before they are projected back. Through incorporating knowledge connections into the features of different subpaths, the graph reasoning module can generate rich and compact representations for each visual element with the guidance of commonsense knowledge. In that way, the reasoning module can help the recognition of visual relationships, especially the infrequent or unseen ones. CGR's configuration module, acting as a switch learns to make discrete decisions, such as whether or not the representations evolved by graph reasoning ought to be used.

The reasoning and the configuration module work together to acquire knowledge for different dynamic composable subpaths and to recognize relationships in an adaptive manner.

Manuscript received March 29, 2020; revised July 18, 2020 and September 22, 2020; accepted September 24, 2020. Date of publication October 29, 2020; date of current version January 5, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100300, in part by the National Natural Science Foundation of China under Grant 61836012, Grant U19A2073, and Grant 61976233; in part by the Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) under Grant 2019B1515120039; in part by the Nature Science Foundation of Shenzhen under Grant 2019191361; and in part by the Zhijiang Laboratory's Open Fund under Grant 2020AA3AB14. *(Corresponding author: Jianbin Jiao.)*

Yi Zhu, Qixiang Ye, and Jianbin Jiao are with the University of Chinese Academy of Sciences (UCAS), Beijing, China (e-mail: zhuyi215@mails.ucas.ac.cn; jiaojb@ucas.ac.cn; qxye@ucas.ac.cn).

Xiwen Liang, Bingqian Lin, Liang Lin, and Xiaodan Liang are with Sun Yat-sen University, Guangzhou, China (e-mail: liangcici5@gmail.com; bingqianlin@126.com; linliang@ieee.org; xdliang328@gmail.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3027575

2162-237X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

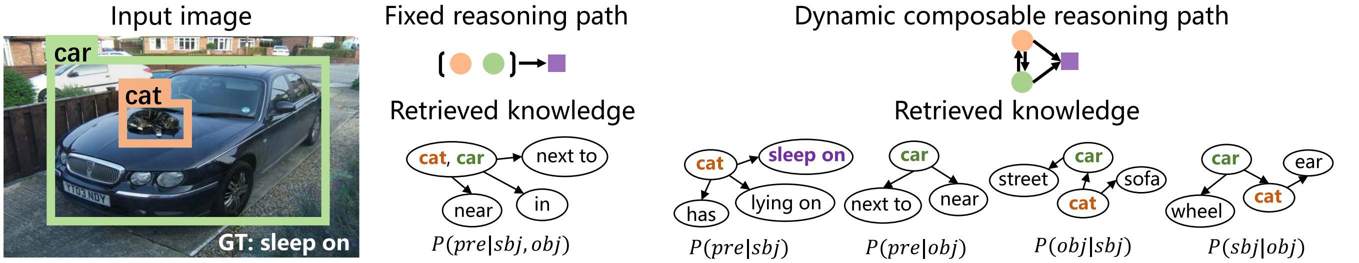


Fig. 1. Comparison of the reasoning over a fixed path and dynamic path. Previous methods with fixed reasoning path failed to retrieve essential knowledge connections for the uncommon relationship {cat, sleep on, car}, while CGR with dynamic composable reasoning paths elaborately hits the knowledge about {cat → sleep on} to facilitate VRD. Best viewed in color.

The method is compatible with different knowledge graphs as long as the graphs contain basic common facts about correlations among entities to facilitate the reasoning over the subpaths. Fig. 1 shows that the knowledge retrieved for the fixed reasoning paths are limited to depicting the prior connections between {cat, car}, and "sleep on" which is uncommon in the current knowledge base. In contrast, CGR selects the knowledge on $P(pre|sbj = \text{“}cat\text{”})$ to provide common and strong evidence predicting the relationships along the dynamic reasoning paths of {s → p}.

Three advantages of CGR stand out. 1) Instead of considering a fixed multidependent reasoning path CGR performs reasoning over multiple subpaths where the dependencies are common and stable. 2) CGR dynamically composes the knowledge incorporation for the reasoning of visual relationships and achieves better generalization ability to the long-tailed relationships. 3) Our method is compatible with knowledge graphs from any data collections as long as they contain basic commonsense about the subpaths.

Extensive experiments on visual relationship detection (VRD) and visual genome (VG) data set show that CGR significantly outperformed the previous state-of-the-arts on several benchmarks of VRD. Evaluation of infrequent relationships and unseen relationships validates that our method is able to extract supplementary information from external knowledge graphs to improve the relationship prediction. Knowledge generalization experiments verified that CGR can not only select knowledge from universal relationships but also is resistant to noise and redundant connections.

The main contributions of this article are as follows.

1) The CGR that learns to tackle the challenging long-tail problem of VRD by dynamically composing the reasoning paths incorporated with visual commonsense knowledge.
2) The exploration to extract commonsense knowledge of universal visual relationships from different data collections and different domains.
3) A demonstration of the fact that the proposed method works well for the infrequent relationships and the unseen ones and achieves significant improvement over previous state-of-the-art.

II. Related Works

A. Visual Relationship Detection

As one of the most challenging problems in computer vision, VRD has been extensively investigated [3], [4], [6], [10]–[17]. Xu *et al.* [10] proposed to first build fully connected graphs to connect the visual objects (nodes) with their relationships (edges) and then perform iterative message passing on the graphs to infer the object and relationship labels. To reduce the computational cost of propagating on densely connected graphs, some works attempt to prune the connections between objects in advance, by computing relation proposals [17], dividing the graphs into subgraphs for hierarchically inference [16], or assuming prior structures, such as Tree LSTMs [18]. They are limited in reasoning on the complex dependencies across visual elements based on the fixed prior structures. CMAT [19] formulates the scene graph generation problem as a sequential decision-making process of choosing nodes and edges on an initial graph. A graph-level metric is proposed to assign larger loss for the hub object nodes which contribute more to the global semantic of the scene graph. In contrast, CGR learns to dynamically choose prior connections from commonsense knowledge graphs.

Other works [4], [6], [7], [14] are built following the reasoning path {subject, object → predicate} based on the strong regularities that the label of relationships are highly predictable when the labels of subjects and objects are identified. Deep structure learning [20] is proposed to predict visual relationships considering independent substructures, but fail to select valid paths and construct dynamically composable reasoning paths over the commonsense knowledge. The statistical dependencies in data set about object pairs and their relationships can be utilized to train a deep relational network [4], or to act as an additive bias for the classifier module [6], [14], but such frequency statistics are usually dominated by the most common relationships and often fail to predict the infrequent ones. Besides the frequency statistics on the distribution of relationships between pairs of object labels, more correlation types (e.g., attribute and co-occurrence) between visual elements are explored as correlations are embedded as knowledge graphs to guide the training of graph models of VRD [5]. Yu *et al.* [8] collected external linguistic knowledge from Wikipedia and distilled this knowledge into a deep model. KB-GAN [9] retrieved commonsense facts $\langle s, p, o\rangle$ from ConceptNet using the predicted label s with the triplets were embedded into knowledge units and fused via attention mechanism.

Our approach is different from these works in the following three aspects. 1) Reasoning path. In CGR, a set of dynamically composable subpaths is used for the reasoning of visual relationships, rather than considering [5], [6], [9] only single fixed

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

paths. 2) Knowledge matching. Our reasoning module matches knowledge for each subpath by projecting the object features to the entities of knowledge graphs, rather than as in [9] where the predicted labels are not fault-tolerant and incorrect labels can cause wrong matching results. 3) Knowledge selection. Our method incorporates the knowledge into the subpaths in a first-enhance-then-select manner. In contrast to [9], which uses knowledge in a first-select-then-enhance manner and different from [8], which distilled the knowledge into the deep model without explicitly ensure valid enhancement.

B. Graph Learning

Some researchers effort to model domain knowledge graph for excavating correlations among labels or objects in images, which has been proved effective in many tasks [3], [21]. Graph structures are specialized in modeling a set of objects (nodes) and their relationships [22]. In graph neural networks (GNNs), information of one node can be propagated to other nodes via message passing along edges of the graph. Most of the variants of GNNs [23]–[26] aim to design proper propagation schemes to model various relationships in graph data. Since attention mechanism has been successfully applied in many sequence-based tasks, graph attention networks (GATs) [27] are proposed to incorporate the attention mechanism into the propagation step, which computes the hidden states of each node by attending over its neighbors. To improve the long-term propagation of information across the graph structure, gate mechanism, such as GRU [28] or LSTM [29], are further used in the propagation step. In this work, we propose a configurable graph learning to dynamically and seamlessly integrate knowledge reasoning to compose adaptive reasoning paths and bridge cross-domain semantic gap between knowledge and scene graph of each image.

C. Visual Reasoning

Our method is also related to the visual reasoning methods that usually exploit to introduce different forms of knowledge between objects or scenes, e.g., the shared attributes [30], [31] and the relationships between objects. Early research had explored the similarity of the attributes via linguistic embeddings [32], [33] or has used the knowledge of object relations in a postprocessing step [34], [35]. More recent works propose to reasoning over graph structures [21], [36]–[39] by encoding nodes and edges of the graph by attending to their connected neighbors. These works typically apply a fixed graph, which limits the generalization ability to particular scenarios. The method in [40] uses linguistic knowledge about object attributes as external supervisions to exploit two kinds of knowledge forms. In the works mentioned earlier, each prediction is made with previously defined reasoning patterns. In contrast, our CGR could incorporate various possible knowledge graphs and choose different kinds of reasoning patterns for each relationship in a generalizable and dynamically switchable way.

III. Method

The proposed CGR can be divided into the following steps: 1) generate feature representation for object proposals and relation proposals; 2) incorporate visual commonsense knowledge into the reasoning along different subpaths; 3) select subpaths with valid knowledge enhancement for the reasoning of each visual relationship instances; and 4) predicting the labels of objects and predicates based on the knowledge enhanced features.

A. Feature Representation

We first generate feature representation for the elements of visual relationships, acting as node features for the graph reasoning over knowledge graphs.

1) Object Proposal Feature: Given an image I, we use the faster region based convolutional neural network (R-CNN) [41] to extract a set of object proposals $B = \{b_i\}_{i=1}^{\hat{N}}$ where $b_i = [x_i, y_i, w_i, h_i]$ is the coordinate of the box and the corresponding region feature vector is v_i.

2) Relation Proposal Feature: For any two different objects $[b_i, b_j]$, there are $\hat{N}(\hat{N}-1)$ possible directed relations between them. We denote N as the maximum number of object pairs to be selected in each image to keep computation efficiency. In our experiment, we set $N = 512$. We build a separate network branch to extract region features v_{ij}^p for the closed boxes of the subject-object pairs $[b_i, b_j]$ with the same structure as the region of interest (ROI)-align layers in the detector network, similar to [7]. This branch aims to generate visual features for predicates to focus on the interactive areas of subjects and objects.

3) Node Features for Subpaths: For the reasoning of visual relationship $\{s, p, o\}$ of object pair (b_i, b_j), CGR first decouples the reasoning path of visual relationships into a set of subpaths, including two types of paths, they are paths with homogeneous nodes, e.g., $\{s \rightarrow o\}$ and $\{o \rightarrow s\}$), and paths with heterogeneous nodes, e.g., $\{s \rightarrow p\}$ and $\{o \rightarrow p\}$). The node features for s and o in path $\{s \rightarrow o\}$ and $\{o \rightarrow s\}$ are v_i^s and v_j^o, which are region features of box b_i and b_j, while for s and o in $\{s \rightarrow p\}$ and $\{o \rightarrow p\}$, the node features are $v_{ij}^{\mathrm{sp}} = \mathrm{concat}(v_i^s, v_{ij}^p)$ and $v_{ij}^{\mathrm{op}} = \mathrm{concat}(v_i^o, v_{ij}^p)$ which implies the context information about the predicate.

B. Graph Reasoning With Commonsense Knowledge

The CGR is proposed to introduce high-level semantic and linguistic information for each subreasoning path. Information propagation between visual elements is performed based on the semantic connections from the commonsense knowledge graph, resulting in rich and compact representations for relationship recognition. Let $X = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^{D^d}$ denotes the feature of the start nodes from the subpath of the same type in each image. The graph reasoning module that updates the input features for each subpath can be denoted as $Y = g(X)$. In this section, we describe a general version for a subpath for mathematical simplicity. To perform the VRD, the graph reasoning module $g(\cdot)$ is instantiated with different semantic correlations of the knowledge for the reasoning over different subpaths in Section III-D.

1) Knowledge Embedding: The commonsense knowledge graph is used to show distinct correlations between entities in general. The graph can be formulated as $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k)$, where

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

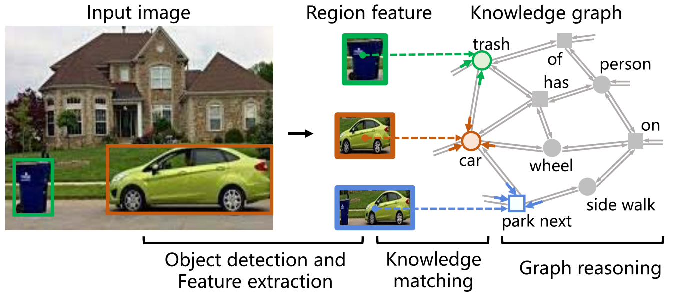


Fig. 2. CGR matches each visual element in the input image to the entities of the knowledge graph, building connections between local regions to knowledge with a global perspective. The graph reasoning is performed over the knowledge graph w.r.t different sub-paths and then the evolved features will be projected back to enhance local representations.

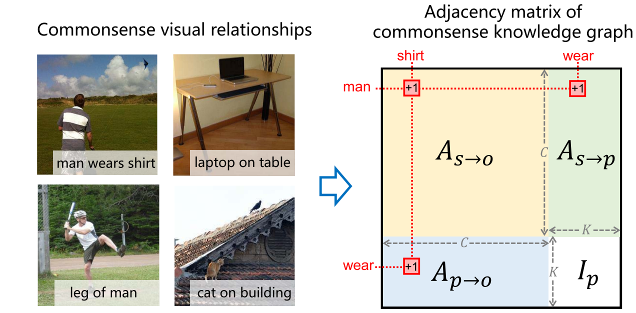


Fig. 3. Commonsense visual relationships are collected from image data sets that have relationship annotations. The commonsense knowledge graph is generated to provide prior knowledge about the semantic correlations among the object and predicate nodes. The connections between each two of the elements in visual relationships are counted in the adjacency matrix.

$\mathcal{V}^k$ and $\mathcal{E}^k$ denote the node set and edge set, respectively. The graph $\mathcal{G}^k$ is a heterogeneous graph in which the node set contains two types of entities, i.e., the object labels (circle nodes in Fig. 2) and predicate labels (square nodes in Fig. 2). The edge between two nodes represents the statistical co-occurrence frequency of the nodes collected from visual relationships in real-world scenes. We denote the adjacent matrix of the knowledge graph $\mathcal{G}^k$ as $A \in \mathbb{R}^{M\times M}$, where $M = C + K$, C, and K are the numbers of the object and predicate classes. $A(i, j)$ indicates the prior frequency of the node j given node i, and thus, we can infer the label of j via the edge $i \to j$. Different parts of the adjacency matrix represents different types of edges, e.g., $A_{s\to o}$ for $s \to o$, $A^T_{s\to o}$ for $o \to s$, $A_{s\to p}$ for $s \to p$, and $A^T_{p\to o}$ for $o \to p$. I_p is the identity matrix. Before performing graph reasoning, the subadjacency matrices for each subreasoning path are normalized by dividing the sum value of the corresponding rows. From the data sets that have rich relationship annotations, we can collect knowledge triplets that contain two objects and the relation between them, e.g., {man, wear, and shirt} in Fig. 3. The co-occurrence of each two element of the triplet is counted once. For each of the visual relationship $\{s, p, o\}$, we increase the value of $A(s, p)$, $A(p, o)$, and $A(s, o)$ by 1. To construct node embeddings, we use the off-the-shelf word vectors [42] as linguistic embedding of each entity in the knowledge graph $\mathcal{G}^k$, denoted as $\mathcal{S} = \{s_n\}$, $s_n \in \mathbb{R}^L$.

2) Knowledge Matching: To incorporate the reasoning of each subpaths with the knowledge graph, CGR projects the representation of the paths into a uniform semantic space composed of the M semantic entities in the knowledge graph. CGR generates the visual representation H^{lt} of the entities in knowledge graph by summarizing the semantic votes from the local visual features X of each image

$$H^{lt} = (\sigma(XW^l))^T XW^a \tag{1}$$

where $\sigma(\cdot)$ is softmax function to normalize the probabilities summarized across different nodes. We denote trainable transformation weights $W^l \in \mathbb{R}^{D^d\times M}$ for projecting input feature to the semantic space with M labels and $W^a \in \mathbb{R}^{D^d\times D^c}$ for converting the input feature to a lower dimension D^c to reduce overfitting, respectively.

3) Graph Reasoning: After building the semantic connection between the knowledge graphs and each visual relationships, the reasoning guided by structured knowledge is employed to leverage semantic constraints from human commonsense to evolve the representations of entity nodes. CGR performs graph propagation over representations H^{lt} of all the entity nodes via the matrix multiplication form, resulting in the evolved features H^{st}

$$H^{st} = \sigma(A[H^{lt}, S]W^s) \tag{2}$$

where $W^s \in \mathbb{R}^{(D^c+L)\times D^c}$ is a trainable weight matrix. Directly use the original A will change the scale of the features. To get rid of this problem, we normalize A as in the graph convolutional networks [38] such that all rows sum to one, i.e., $Q^{-(1/2)}AQ^{-(1/2)}$, where Q is the diagonal node degree matrix of A. This formulation arrives at the new propagation rule

$$H^{st} = \sigma(\hat{Q}^{-\frac{1}{2}}\hat{A}\hat{Q}^{-\frac{1}{2}}[H^{lt}, S]W^s) \tag{3}$$

where $\hat{A} = A + I$ is the adjacency matrix of the graph $\mathcal{G}^k$ with added self-connections for considering its own representation of each node, and I is the identity matrix.

4) Feature Refinement: The evolved representations Hst equipped with relational knowledge, and linguistic information are used to boost the capacity of the feature representation for different subpaths of each visual relationship. As the feature distributions have been changed after matching to the entities of the knowledge graph, we need to find mappings to passing the message about commonsense knowledge from the evolved visual representation of entities H^{st} back to the visual elements X

$$\begin{aligned} H^{rt} &= \sigma([\hat{X}, \hat{H}^{st}]W^q)H^{st}W^p \\ Y &= X + H^{rt} \end{aligned} \tag{4}$$

where a trainable weight matrix $W^q \in \mathbb{R}^{(D^d+D^c)\times 1}$ is used to evaluate the compatibility for H^{st} with the node feature X of the subpaths. To implement this, X is expanded to $\hat{X} \in \mathbb{R}^{N\times M\times D^d}$, and H^{st} is expanded to $\hat{H}^{st} \in \mathbb{R}^{N\times M\times D^c}$. $W^p \in \mathbb{R}^{D^c\times D^d}$ projects the feature to have the same dimension with the input, resulting in H^{rt}. We obtain the refined feature of nodes in subpaths, $Y = \{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^{D^d}$, by summing H^{rt} and a residual connection of X.

As is formulated earlier, CGR performs graph reasoning in three steps: 1) matching the visual elements in each image

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

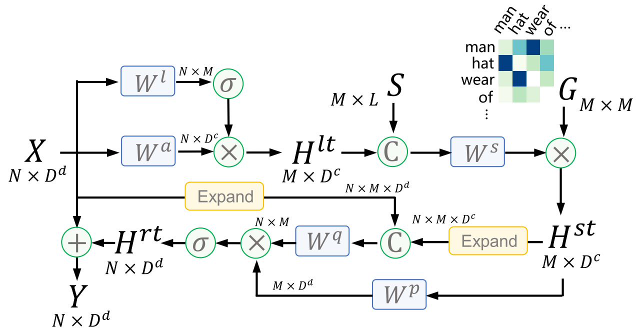
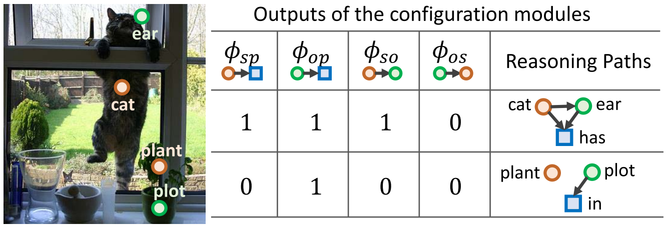


Fig. 4. Illustration of the graph reasoning $g(\cdot)$ with visual commonsense knowledge. The input is first projected to the semantic space of the graph nodes, then enhanced with the linguistic embedding and the knowledge graphs, and finally projected back.



Outputs of the configuration modules

ϕ_{sp}	ϕ_{op}	ϕ_{so}	ϕ_{os}	Reasoning Paths
1	1	1	0	cat ear has
0	1	0	0	plant plot in

Fig. 5. Examples of the outputs of each configuration module. CGR selects knowledge about the most decisive subreasoning paths between visual elements in a dynamic composable manner.

to the entities of the knowledge graph, see equation (1); 2) propagating semantic information among the visual representations of entities and updating the features from a global perspective, see equation (3); 3) adapting the global evolved features back to the local representation of each relationship in each image, see equation (4). Equations (1) and (4) work together to bridge the semantic gap between the knowledge graph and each practical scene graph. The graph reasoning process is illustrated in Fig. 4.

Note that the commonsense knowledge graph $\mathcal{G}^k$ is actually a heterogeneous graph that contains two kinds of nodes (object labels and predicate labels) and different kinds of edges between these nodes (e.g., object → object, predicate → object, and object → predicate). The graph reasoning modules for different subpaths can activate different kinds of connections in $\mathcal{G}^k$. Through the cooperative learning of graph reasoning and configuration, our CGR is able to select semantic connections for benefiting the visual relationship prediction. The reason why we use the comprehensive knowledge graph for the graph reasoning modules is three folds. 1) The commonsense knowledge graph about scene graphs naturally contains the heterogeneous semantic information from objects to objects or to predicates. Different parts of the adjacency matrix A show the semantic correlations for different subpaths, which can be activated independently during the graph reasoning, and the invalid knowledge enhancement would be ignored by the following configuration modules. Therefore, it is not necessary to separate the whole graph into subgraphs with homogeneous connections. 2) During knowledge matching, the input features of the graph reasoning modules are projected to vote to the nodes of the knowledge graph. In the heterogeneous graph, the input features are allowed to vote for both object and predicate nodes. The graph reasoning over subpaths is performed with a global perspective. 3) We had conducted experiments to compare these two settings, the performance of CGR using a heterogeneous graph is comparable to that using a group of homogeneous graphs.

C. Graph Reasoning Configuration

Since we cannot ensure that all the graph connections in the commonsense knowledge from open domains could benefit the reasoning of visual relationships, the invalid knowledge should be discarded to avoid introducing noisy distributions into the reasoning. We explore to make decisions about whether to incorporate knowledge graph reasoning into each subreasoning path.

1) Configuration Module: We hope that the gradients can be back propagated through discrete decision making of module configuration for end-to-end training. Therefore, we select the knowledge graph reasoning for each path based on nonlinear transformation $f(\cdot)$ followed by the Gumbel–Max trick with its continuous softmax relaxation [43]–[46]. The samples $\boldsymbol{z}$ can be drawn from a categorical distribution $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_k\}$ as

$$\boldsymbol{z} = \text{one\_hot}(\text{argmax}_{i \in \{1,\ldots,k\}}(\log(\pi_i) + o_i)) \tag{5}$$

where $\boldsymbol{o} = -\log(-\log(\boldsymbol{u}))$ with $\boldsymbol{u} \sim \text{Uniform}(0, 1)$. The categorical variable $\boldsymbol{z}$ here is the one-hot vector of the k dimension. In CGR, the possible decision for using commonsense knowledge is binary; therefore, we set $k = 2$. $\boldsymbol{z}$ is a 2-dim one-hot vector for the discrete decision, where the binary element z_0 is encoded as the Boolean output to determine whether to update the input feature with commonsense knowledge. The softmax relaxation of Gumbel–Max trick is to replace nondifferentiable argmax operation in (5) with the continuous softmax function

$$\hat{z} = \text{softmax}((\log(\pi_k) + \boldsymbol{o})/\tau) \tag{6}$$

where temperature τ of the softmax function is empirically set to 1 in our work. During the training stage, a one-hot vector $\boldsymbol{z}$ was obtained as the discrete sample from, see (5) for forwarding propagation, and compute gradients w.r.t. $\boldsymbol{\pi}$ in, see (6) for backpropagation, whereas at the test stage, we draw the sample with the largest probability without Gumbel samples $\boldsymbol{o}$. The simplest way to configure different subreasoning paths is to directly make decisions about whether to use the output $y_i \in Y$ of each reasoning path, in other words, by estimating a e_i via the Gumbel-Softmax (GS) for each y_i

$$e_i = GS(W^{gp} y_i) \tag{7}$$

where $W^{gp} \in \mathbb{R}^{D^d \times 2}$ denotes the learnable parameters. e_i is a $2-d$ one-hot vector for determining whether to use the knowledge enhanced feature y_i for the final prediction. We denote the configuration module as $\phi(\cdot)$. The input is the output of the graph reasoning modules y_i, and the output is either y_i or x_i

$$\phi(y_i) = e_{i,0} x_i + e_{i,1} y_i. \tag{8}$$

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

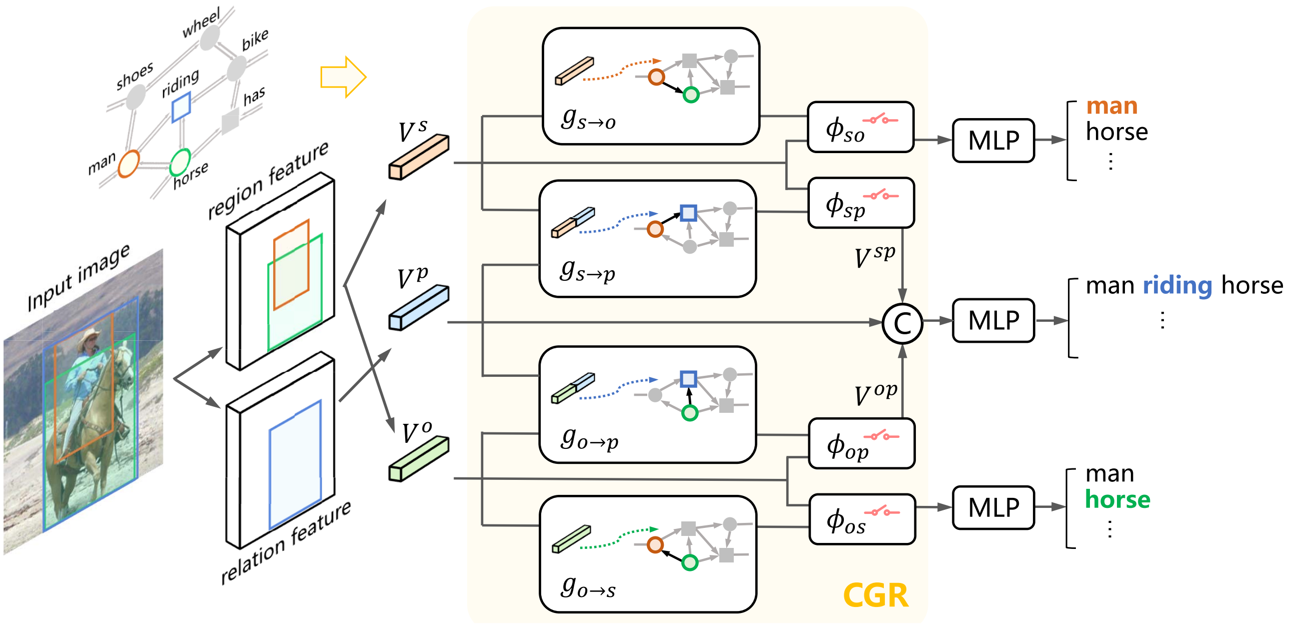


Fig. 6. Overview of the CGR for VRD. Images are fed to a standard object detector to obtain object proposals and their region features. The relation features are generated based on the closed boxes of subject-object pairs. The reasoning modules ($g(\cdot)$) are instantiated with the commonsense knowledge graph about visual relationships, followed by the learnable configuration module ($\phi(\cdot)$) to make decisions about whether to introduce knowledge into the representation of visual elements of each relationship. After dynamically composing the reasoning paths, the labels of subject-object pairs and relationships can be obtained by multilayer perception (MLP) classifiers.

If $e_{i,1} = 1$, we use y_i as input feature for the classifiers, otherwise we use x_i which incorporates commonsense knowledge for benefiting VRD. We implement our configuration module based on the gumbel_softmax function in PyTorch and set hard=True. CGR has four configuration modules for the different subreasoning paths. They are inserted after each of the graph reasoning modules to decide whether to introduce knowledge enhanced features for each subject, object, and relation features in an image. Such a path-level configuration module determines the contribution of the knowledge graph reasoning over different subpaths for relation prediction between visual elements, leading to dynamically switchable incorporation of different knowledge connections.

2) Discussion: Fig. 5 shows the outputs of the configuration modules and the selected knowledge connections for the relationships in an image. It can be seen that CGR selects knowledge about the most decisive dependencies between visual elements in a dynamically composable manner. In the first row, the decisions made by the configuration modules of each subpath are $\phi_{\text{sp}} = 1$, $\phi_{\text{op}} = 1$, $\phi_{\text{so}} = 1$, and $\phi_{\text{os}} = 0$, which indicate that the knowledge enhanced subpaths $\{cat \rightarrow ear\}$, $\{cat \rightarrow has\}$, and $\{ear \rightarrow has\}$ are selected to form the adaptive reasoning path for the recognition of the visual relationship $\{cat, has$, and $ear\}$. In the second row, the path $\{plot \rightarrow in\}$ is selected, while the other paths are discarded.

The configuration module brings advantages for the knowledge-based graph reasoning of CGR in three aspects: 1) the module selected from the representation before and after introducing knowledge, ensuring valid knowledge enhancement and discarding biased and noisy knowledge connections; 2) it explores personalized graph reasoning paths for each relationship, cooperating with highly determinable in an adaptive manner; and 3) since the module is able to filter out keynotes from commonsense knowledge graphs, the CGR generalizes well to knowledge collected from different domains.

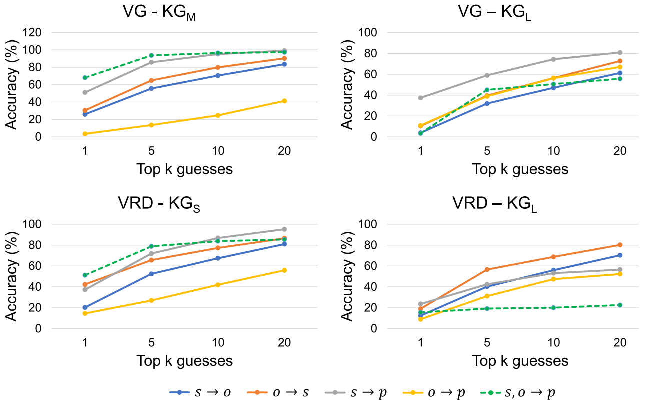


Fig. 7. Accuracy of the top-k guessing results, for the subject, object, or predicate labels in a scene graph, following the different reasoning paths. The generalization ability of the path $s, o \rightarrow p$ is poor when using a cross-domain knowledge graph (KG$_L$), while the subpaths are more robust for different knowledge collections.

D. Visual Relationship Detection

VRD aims to detect objects and predict relationships for each pair of them. The overall framework of CGR for VRD is shown in Fig. 6. According to Section III-A, the object proposal features $V = \{v_i\}_{i=1}^N, v_i \in \mathbb{R}^{4096}$ are paired to be subject features $V^s = \{v_i^s\}_{i=1}^N$ or object features $V^o = \{v_i^o\}_{i=1}^N$ for visual relationship proposals. And the corre-

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISON OF OUR CGR WITH THE PREVIOUS STATE-OF-THE-ART METHODS ON THE VRD DATA SET. OUR MODEL IS EVALUATED USING THE DETECTOR BACKBONE PRETRAINED ON VRD, IMAGENET AND COCO DATA SET

Pretrain	Method	SGCls			PredCls			SGDet		PhrDet	
		R@20	R@50	R@100	R@20	R@50	R@100	R@50	R@100	R@50	R@100
VRD	IMP [10]	24.0	25.4	25.6	45.2	48.6	49.0	13.4	15.2	27.7	28.2
	MOTIFS [6]	24.2	25.9	26.1	49.5	53.5	54.1	15.4	16.8	26.8	28.1
	CGR (Ours)	**25.8**	**27.6**	**27.8**	**51.3**	**55.3**	**56.0**	**15.8**	**16.9**	**28.2**	**29.3**
ImageNet	RelDN [7]	34.8	34.8	34.8	52.6	52.6	52.6	21.5	26.4	28.2	**35.4**
	CGR (Ours)	**35.8**	**37.3**	**38.1**	**55.5**	**57.0**	**57.7**	**21.7**	**26.5**	**28.7**	**35.4**
COCO	RelDN [7]	34.7	34.7	34.7	52.4	52.4	52.4	28.2	33.2	**34.5**	42.1
	CGR (Ours)	**36.6**	**37.5**	**38.0**	**55.9**	**57.4**	**58.1**	**28.4**	**34.6**	**34.5**	**42.2**

TABLE II

COMPARISON OF MEAN RECALL (mR) PERFORMANCE FOR CGR AND PREVIOUS STATE-OF-THE-ARTS ON VG DATA SET

Method	SGCls		PredCls	
	mR@50	mR@100	mR@50	mR@100
IMP [10]	6.8	7.2	11.9	12.9
MOTIFS [6]	7.7	8.2	14.0	15.3
KERN [5]	9.4	10.0	17.7	19.2
VCTree [18]	10.1	10.8	17.9	19.4
RelDN [7]	11.0	11.0	22.2	22.3
CGR (Ours)	**12.4**	**13.0**	**23.2**	**23.5**
RelDN (X-101-FPN) [7]	11.7	11.7	22.5	22.5
CGR (X-101-FPN)	**13.4**	**13.8**	**24.4**	**24.6**

sponding features for the predicates are $V^p = \{v_i^o\}_{i=1}^N$. In Section III-B, we describe the graph reasoning over the commonsense knowledge graph about visual relationships. The graph reasoning module for a subpath is denoted as $g(\cdot)$ which takes the region features of objects or relations as inputs and updates these features by incorporating semantic correlations extracted from the knowledge graph. As is shown in Fig. 6, we instantiate the graph reasoning module g for different subpaths; they are $g_{s\to o}$, $g_{s\to p}$, $g_{o\to p}$, and $g_{o\to s}$. These modules take different features as input and extract different types of semantic correlations from the commonsense knowledge graph.

1) Object Classification: The features of subjects and objects are evolved by performing graph reasoning $g_{s\to o}$ and $g_{o\to s}$ over paths $s \to o$ and $o \to s$. Following the graph reasoning modules are path configuration modules ϕ_{so} and ϕ_{os}, which make decisions about whether to use the knowledge enhanced feature. MLP layers act as object classifiers to predict labels of subjects and objects $L^s = \{l_i^s\}_{i=1}^N$, $L^o = \{l_i^o\}_{i=1}^N$, and $l_i^s, l_i^o \in \mathbb{R}^C$

$$\begin{aligned} L^s &= \text{MLP}(\phi_{so}(g_{s\to o}(V^s))) \\ L^o &= \text{MLP}(\phi_{os}(g_{o\to s}(V^o))). \end{aligned} \tag{9}$$

Likewise, we instantiate another two graph reasoning modules, denoted as $g_{s\to p}$ and $g_{o\to p}$, to incorporate the paths $\{s \to p\}$ and $\{o \to p\}$ with corresponding visual commonsense knowledge. Two configuration modules $\phi_{\text{sp}}, \phi_{\text{op}}$ are employed to decide whether to use the evolved feature and an MLP classifier is used to predict relationship labels $L^p = \{l_i^p\}_{i=1}^N, l_i^p \in \mathbb{R}^K$ as

$$\begin{aligned} V^{sp} &= \phi_{sp}(g_{s\to p}(\text{concat}(V^s, V^p))) \\ V^{op} &= \phi_{op}(g_{o\to p}(\text{concat}(V^o, V^p))) \\ L^p &= \text{MLP}(\text{concat}(V^{sp}, V^p, V^{op})). \end{aligned} \tag{10}$$

Besides the visual features, we also consider the spatial information of the object pairs and the predicted subject and object labels as in [7], based on which we can predict relation class scores $l_i^{\text{spt}}, l_i^{\text{emb}} \in \mathbb{R}^K$. The final relationship labels are

$$l_i^p = \text{softmax}\left(l_i^p + l_i^{\text{spt}} + l_i^{\text{emb}}\right), \quad i = 1, \ldots, N. \tag{11}$$

A visual relationship prediction contains: 1) subject and object labels, l_i^s and l_i^o. 2) Coordinates of the bounding boxes of subject and objects, b^s, b^o. 3) Predicate labels of the subject–object pair, l_i^p.

IV. EXPERIMENTS

The proposed CGR was evaluated against commonly used benchmarks first by introducing the evaluation details in Section IV-A and then by comparing the proposed CGR with state-of-the-art VRD methods to demonstrate comparative the effectiveness and efficiency of the approach, Section IV-C. In Section IV-E, we instantiated the graph reasoning of CGR with knowledge graphs from different data sets and domains to evaluate CGR's potential for exploring common sense knowledge from universal visual relationships.

A. Experimental Setup

1) Data Sets: To evaluate our approach, two data sets, including VRD [3] and visual genome (VG) [47] are used. VRD contains 5000 images (4000 for training and 1000 for testing) with 100 object categories and 70 predicates, while VG contains 89 189 images (62 723 images for training and 26 446 images for testing) with 150 object classes and 50 predicates.

2) Task: To validate the performance of our model thoroughly, we evaluated it on four tasks: 1) scene graph generation (**SGGen**). Given an image, SGGen is to predict the object bounding boxes, box labels, and edge labels. 2) Visual phrase detection (**PhrDet**). Different from the SGGen that predicts separated bounding boxes for both subject and object, PhrDet detects one union box for <subject, predicate, object> phrases. 3) Scene graph classification (**SGCls**). Given ground

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

TABLE III

COMPARISON OF PERFORMANCE AND TIME CONSUMPTION FOR CGR AND PREVIOUS STATE-OF-THE-ARTS ON VG DATA SET. THE AVERAGE TRAINING TIME (s/Image) FOR VRD IS TESTED ON A TESLA M40 GPU

Method	**SGCls**			**PredCls**			**SGDet**				
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	Mean	Time
IMP [10]	31.7	34.6	35.4	52.7	59.3	61.3	14.6	20.7	24.5	37.2	-
MOTIFS [6]	32.9	35.8	36.5	58.5	65.2	67.1	21.4	27.2	30.3	41.7	0.4
KERN [5]	-	36.7	37.4	-	65.8	67.6	-	27.1	29.8	-	0.9
VCTree [18]	35.2	38.1	38.8	60.1	66.4	68.1	22.0	27.9	31.3	43.1	2.5
RelDN [7]	36.1	36.8	36.8	66.9	68.4	68.4	21.1	28.3	**32.7**	43.9	4.3
CGR (Ours)	**36.2**	**38.9**	**40.8**	**67.8**	**69.3**	**70.3**	**21.3**	**28.4**	**32.7**	**45.1**	1.6
RelDN (X-101-FPN) [7]	38.2	38.9	38.9	67.2	68.7	68.8	22.5	31.0	36.7	45.7	-
CGR (X-101-FPN)	**38.5**	**40.0**	**41.1**	**68.3**	**69.7**	**70.8**	**22.9**	**31.3**	**36.8**	**46.6**	-

TABLE IV

NUMBER OF CORRECTLY DETECTED RELATIONSHIPS UNDER THE ZERO-SHOT SETTING

Dataset	Method	**SGCls**	**PredCls**
VRD	RelDN [7]	35	113
	CGR (KG_L)	**94**	**220**
VG	RelDN [7]	80	471
	CGR (KG_L)	**434**	**1613**

truth boxes, SGCls need to predict box labels and edge labels. 4) Predicate classification (**PredCls**). Given ground truth boxes and box labels, the task of PredCls is to predict edge labels.

3) Metrics: Two metrics are used to evaluate the proposed approach on both VRD and VG data sets. One is the Recall@K (R@K, K=20, 50, 100), which indicates the proportion of the ground truth relationship triplets appearing in the predicted top-K confident triplets in an image. The other is mean Recall@K (mR@K, K=20, 50, 100), which computes the mean value of the R@K scores for each relation category. mR@K can effectively alleviate the impact of the imbalanced distributions of relation categories, which commonly exist in scene graph generation tasks.

B. Analysis of Commonsense Knowledge

1) Statistical Information: In our experiment, we form three commonsense knowledge graphs: KG$_S$, KG$_M$, and KG$_L$ (S, M, and L denoting small, medium, large, respectively), were summarized from popular data sources; they are VRD [3], VG [47], and GQA [48] (a large-scale data set for the VQA task). The knowledge graphs contains 30 355, 439 063, and 3 795 907 relationship instances, 100, 150, and 1073 object categories and 70, 50, and 311 predicate categories.

2) Quality of Knowledge Priors: In Fig. 7, we investigate how much information can be obtained from the commonsense knowledge graphs following different reasoning paths. In particular, we directly use the label statistics as priors to guess the labels of subject (s), object (o), and predicate (p) given the labels of the other elements. Inspired by [6], we evaluate the accuracy of top-k guesses on the test set of VG and VRD. For each of the set, we use knowledge graphs from the same (VRD-KG$_S$ and VG-KG$_M$) and the different domains (VRD-KG$_L$ and VRD-KG$_L$). Higher curves imply that the path is reliable for predicting the labels of elements. It can be seen that the guessing accuracy of $s, o \rightarrow p$ is highly predictable for VG-KG$_M$ and VRD-KG$_S$, where the knowledge graph is collected from the same data sources with the test set. However, when using the cross-domain knowledge graph for the guessing, e.g., VG-KG$_L$ and VRD-KG$_L$, the performance of $s, o \rightarrow p$ drops more significant than the four subpaths, demonstrating the stability and generalization ability of the subreasoning paths.

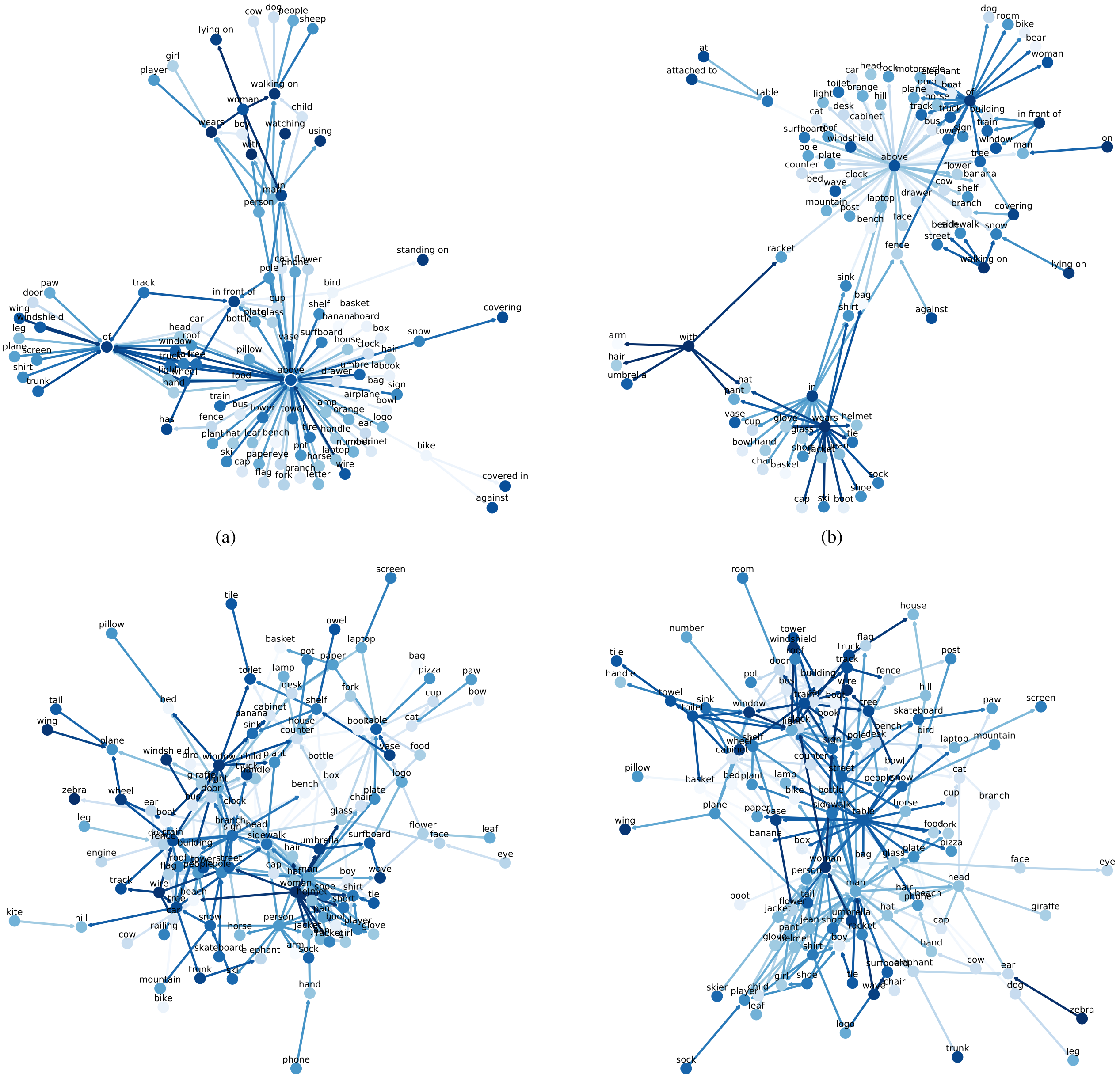
C. Comparison With Previous Methods

1) VRD and VG: The results presented in Tables I and II show that CGR achieves the best visual relationship performance on the VRD and VG data set over all the evaluation tasks in terms of R@20, R@50, and R@100. We also evaluate our models on VRD data set using the faster-RCNN detector [41] with VGG-16 [49] backbone pretrained on ImageNet and COCO, where more training data can be employed. For VG, we use VGG-16 and ResNeXt-101-FPN [50], [51] as our backbone. The improvements on the PredCls and SGCls task are significant, demonstrating the capability of our method to select commonsense knowledge. The improvements of our method on the SGDet and PhrDet tasks are relatively minor since the performance is limited by the quality of the object detector backbones. It is obvious that stronger detectors will achieve significantly better performance on SGDet and PhrDet. In the future works, this issue would be addressed by jointly optimizing the detector backbone and the relationship detection networks.

In Table II, we present the mean Recall@K (mR@K) scores on the VG data set, where the distribution of relationships is extremely unbalanced. It is shown that CGR outperforms the previous methods on both SGCls and PredCls, demonstrating the superiority of our method in predicting infrequent relationships with limited training samples.

2) Time Cost: In Table III, we also compare the time costs of various methods for the future work reference. To get the final detection results, MOTIFS [6] requires two stages of training, i.e., first training the model using the ground truth bounding boxes and then fine tuning the resulting model using object proposals predicted by the detector backbones. In contrast, knowledge-embedded routing network (KERN) [5] requires one more step in the model training procedure. Prior to the training pipeline, KERN trains the models using both ground truth labels and boxes for objects. VCTree [18] first trains the relationship classifiers and then optimize the tree

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

(c)

(d)

Fig. 8. Visualization of the GQA knowledge connections selected by different configuration modules for the VRD on VG data set. Darker edges and nodes indicate that the knowledge about the connected nodes is more frequently selected by the proposed CGR to help guide the reasoning of visual relationships. Best viewed in digital version. (a) Learned knowledge graph w.r.t *subject* → *predicate* (selected by ϕ_{sp}). (b) Learned knowledge graph w.r.t *object* → *predicate* (selected by ϕ_{op}). (c) Learned knowledge graph w.r.t *subject* → *object* (selected by ϕ_{so}). (d) Learned knowledge graph w.r.t *object* → *subject* (selected by ϕ_{os}).

structures of objects and different models are subsequently trained for different evaluation tasks. RelDN and our relation model are directly trained using object proposals detected by the faster-RCNN backbone. As shown in Table III, this model takes less training time than VCTree and RelDN.

3) Zero-Shot Learning: A promising model should be capable of predicting unseen relationship, since the training data will not cover all possible relationship types. Lu *et al.* [3] used word embeddings to project similar relationships onto unseen ones, and Liang *et al.* [13] used a large semantic action graph to learn similar relationships on shared nodes. The results reported in Table II demonstrate that the proposed CGR facilitates the prediction of infrequent relationships. Table IV compares CGR with a state-of-the-art method in the zero-shot learning setting to verify the method's ability in detecting unseen relationships (e.g., the test set's subject–predicate–object combinations but not appear in the training set). The VRD data set contains 37 993 relationship instances out of which 1168 relationships occur only in the test set, not in the training set. For the VG data set, we evaluate CGR on 7601 unseen relationships. As shown in Table IV, our method detects more relationships than RelDN on both SGCls and PredCls.

4) Visualization: The configuration module selects knowledge enhanced representation for the subreasoning paths in each scene, meanwhile, the corresponding graph connections in the commonsense knowledge are activated. In Fig. 8, we statistically summarize the knowledge connections

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

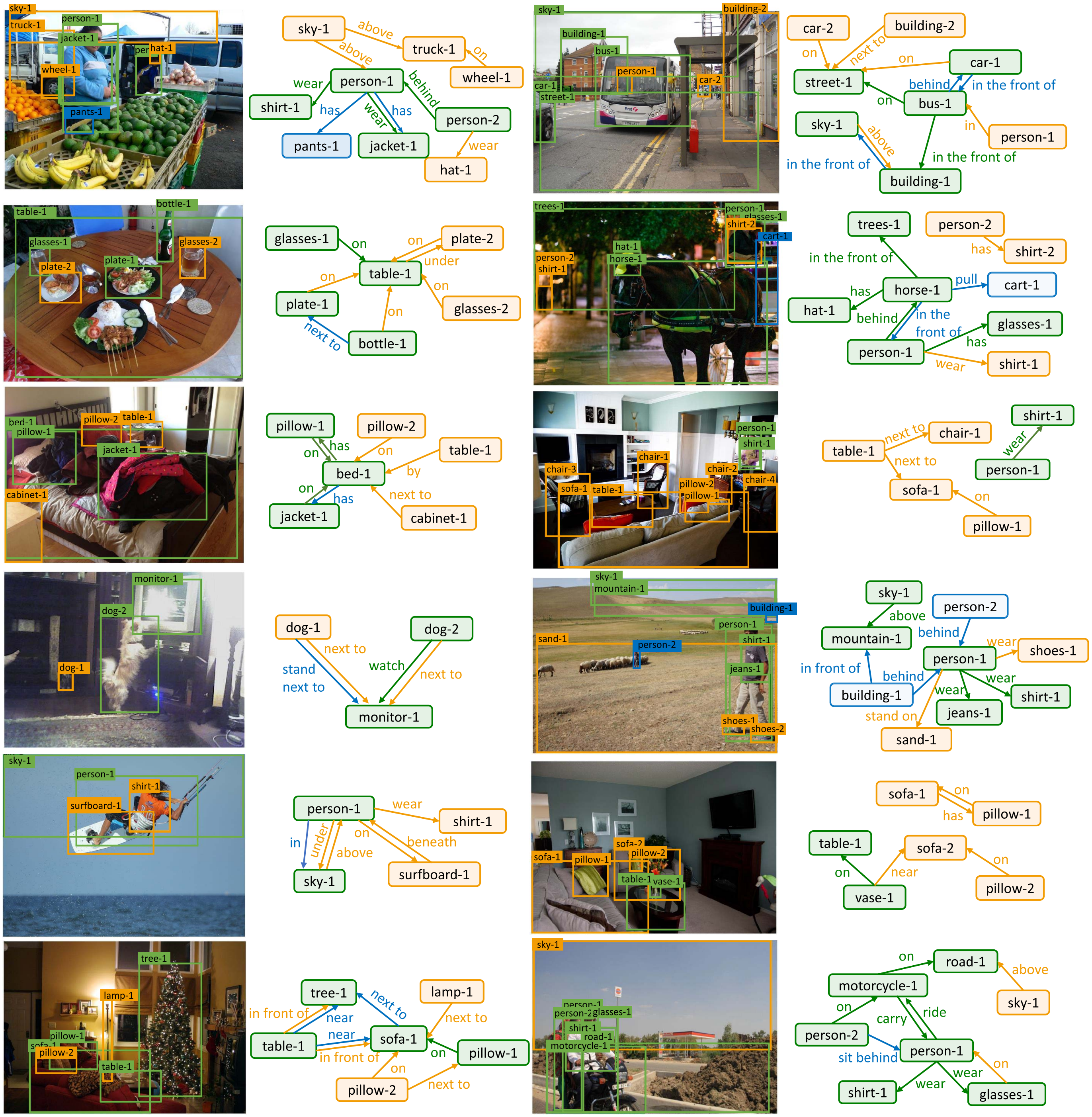


Fig. 9. Qualitative results of our CGR on the VRD data set. Green boxes and edges represent correct predictions (true positives). Blue boxes and edges are missed in the predictions (false negatives). Orange edges indicate reasonable predictions of our model but not annotated as ground truth (false positives).

activated by the configuration modules. Fig. 8(a) presents the learned knowledge graph w.r.t. $\{s \rightarrow p\}$. The knowledge about *woman* and *of* are more frequently selected by ϕ_{sp}, and they are more common and stable than the knowledge connections required by the fixed multidependent reasoning paths.

The knowledge connection $\{woman \rightarrow lying\ on\}$ is frequently selected by ϕ_{sp} to facilitate visual relationship reasoning no matter what the object is. Without considering the diversed object labels (e.g., *sofa* or *bed*), the path $\{woman \rightarrow lying\ on\}$ is more common and stable than the knowledge connections required by the fixed multidependent reasoning paths. In this way, CGR dynamically composes highly predictable reasoning paths for the recognition of each visual relationship. Fig. 9 shows some visualization examples of our method. The visualization results show that our model can recognize various objects and relationships in complex and cluttered scenes. Moreover, from the orange edges in the scene graph examples, we find that our CGR can make reasonable predictions missed in the annotation. For example, in the first picture of Fig. 9,

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

TABLE V

ABLATION STUDY OF THE CONFIGURATION MODULES

Model	SGDet	
	R@50	R@100
CGR (w/o $\phi(\cdot)$)	27.1	33.4
CGR (only ϕ_{so})	28.1	34.2
CGR (only ϕ_{os})	27.9	34.1
CGR (only ϕ_{sp})	28.3	34.3
CGR (only ϕ_{op})	27.6	33.9
CGR (softmax)	28.0	34.2
CGR (ours)	**28.4**	**34.6**

our method correctly predicts "wheel-1 on truck-1" and "sky-1 above truck-1." Besides, our method is able to predict similar relationships, e.g., "sky-1 above building-1" versus the ground truth "sky-1 in the front of building-1." This is because the proposed CGR can dynamically incorporate commonsense knowledge to discover reasonable relationships in the scene.

D. Ablation Study

1) Different Subreasoning Paths: In Table V, we ablate the configuration modules for different subpaths on the VRD data set. The detection backbone models are pretrained on COCO. In the first row, we remove all the configuration modules from the proposed CGR; thus, all the subpaths are incorporated with commonsense knowledge for each image without discarding invalid and noisy knowledge connections. From the second row to the fifth row, CGR employs single configuration modules to dynamically predict whether to use knowledge enhanced features on different subreasoning paths for each sample, achieving consistent improvements on the metric of SGDet. In the last row, CGR has configuration modules for all the subpaths and is able to select the semantic correlations from the knowledge graphs for benefiting the VRD task.

2) Gumbel-Softmax Versus Softmax: Our proposed CGR employs Gumbel-softmax to make discrete decisions about whether to accept the knowledge enhanced features. To investigate the effectiveness of CGR to partially accept the knowledge, we directly replace the Gumbel-softmax with softmax to predict probability weights for combining the features before and after graph reasoning. As is shown in the sixth row of Table V, the performance of CGR (softmax) is slightly lower than that with Gumbel-softmax, which demonstrates the ability of our method to mine knowledge connections that benefit the relationship prediction.

E. Knowledge Generalization

As described in Section III, the graph reasoning is general for commonsense knowledge graphs summarized from universal relationships. The commonsense knowledge, which indicates a semantic correlation between visual elements, is quite general and naturally transferable over different tasks and data sets. We evaluated the proposed CGR with the graph reasoning instantiated by commonsense knowledge graphs summarized from different data sets and different domains.

TABLE VI

STATISTICAL INFORMATION OF COMMONSENSE KNOWLEDGE GRAPHS

KG	Source	Relationships	#Obj	#Pred
KG_S	VRD [3]	30,355	100	70
KG_M	VG [47]	439,063	150	50
KG_L	GQA [48]	3,795,907	1,073	311

TABLE VII

CROSS-TASK AND CROSS-DATA SET KNOWLEDGE GENERALIZATION ON VRD DATA SET

KG	SGDet		PhrDet	
	R@50	R@100	R@50	R@100
CGR (KG_S)	27.8	33.7	33.7	41.3
CGR (KG_M)	27.6	34.0	33.4	41.6
CGR (KG_L)	**28.4**	**34.6**	**34.5**	**42.2**

The statistical information about the collected knowledge graphs is presented in Table VI. To alleviate the effect of object detection results, we report the R@K scored on the PredCls task, which requires predicting the relation classes between two given boxes with ground truth coordinates and labels. We use L1 distance $d_{l1}(\cdot, \cdot)$ to evaluate the similarity between different knowledge graphs. The calculated knowledge distances are $d_{l1}(\text{KG}_S, \text{KG}_L) = 196.8$, $d_{l1}(\text{KG}_S, \text{KG}_M) = 143.4$. The difference between KG_L and KG_S is larger than the difference between KG_M and KG_S, and we, thus, conduct generalization experiments on VRD using knowledge graphs from the perspective of the cross-data set and cross-task knowledge.

1) Cross-Data Set Knowledge: We first verify the CGR's ability to select useful information from knowledge graphs and to be free from the interference of redundant knowledge. From the first two rows in Table VII, we can see that CGR using KG_M achieves higher performance on SGCls and PredCls than that with KG_S, because KG_M summarized from VG contains more general relationships which could help CGR to explore more complete commonsense to benefit the visual relationship recognition.

2) Cross-Task Knowledge: To further validate whether the method works well with knowledge from other task domains, we trained CGR on VRD data set using KG_L that was derived from the VQA task. From the last row of Table VII, it can be seen that the performance of CGR with KG_L outperformed that of CGR with KG_S and KG_M by significant margins, due to a notable ability in extracting common and stable knowledge connections for the subpaths of the adaptive reasoning of each visual relationship.

V. CONCLUSION

In this work, we present CGR for detecting visual relationships. The CGR employs multiple subpaths to explore adaptive reasoning paths for each visual relationship and selectively incorporates commonsense knowledge for different subpaths of the reasoning about visual relationships. The proposed CGR is compatible with general relationship knowledge summarized from other data sets and domains, without being interfered by enormous irrelevant knowledge.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

Our approach constructs a strong baseline so that a universal knowledge graph about visual relationships can be built using natural language sentences. It also opens a promising direction about reasoning on rich and comprehensive knowledge from real-world scenes.

REFERENCES

[1] D. Hudson and C. Manning, "Learning by abstraction: The neural state machine for visual reasoning," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 5903–5916.

[2] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3107–3115.

[3] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.

[4] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3076–3086.

[5] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6163–6171.

[6] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[7] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11535–11543.

[8] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1974–1982.

[9] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1969–1978.

[10] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5410–5419.

[11] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1261–1270.

[12] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2171–2180.

[13] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 848–857.

[14] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "LinkNet: Relational embedding for scene graph," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 560–570.

[15] G. Yin *et al.*, "Zoom-Net: Mining deep feature interactions for visual relationship recognition," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 322–338.

[16] 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 *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 335–351.

[17] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 670–685.

[18] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6619–6628.

[19] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4613–4623.

[20] Y. Zhu and S. Jiang, "Deep structured learning for visual relationship detection," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7623–7630.

[21] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2014–2023.

[22] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: http://arxiv.org/abs/1812.08434

[23] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.

[24] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," 2018, *arXiv:1805.11724*. [Online]. Available: http://arxiv.org/abs/1805.11724

[25] J. Atwood and D. F. Towsley, "Diffusion-convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1993–2001.

[26] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," 2017, *arXiv:1704.01212*. [Online]. Available: http://arxiv.org/abs/1704.01212

[27] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[28] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 503–510.

[31] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1792–1801.

[32] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[33] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[34] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[35] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[36] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[37] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, "Gated graph sequence neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–19.

[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[39] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7239–7248.

[40] C. Jiang, H. Xu, X. Liang, and L. Lin, "Hybrid knowledge routed modules for large-scale object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1552–1563.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[43] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*, vol. 33. Washington, DC, USA: U.S. Government Printing Office, 1948.

[44] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, *arXiv:1611.01144*. [Online]. Available: http://arxiv.org/abs/1611.01144

[45] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–17.

[46] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen, "Recursive visual attention in visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6679–6688.

[47] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016, *arXiv:1602.07332*. [Online]. Available: http://arxiv.org/abs/1602.07332

[48] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6700–6709.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[50] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[51] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.

Yi Zhu received the B.S. degree in software engineering from Sun Yat-sen University, Guangzhou, China, in 2013. She is currently pursuing the Ph.D. degree in computer science with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China.

Her current research interests include object recognition, scene understanding, weakly supervised learning, and visual reasoning.

Xiwen Liang received the B.S. degree in software engineering from South China Agricultural University, Guangzhou, China, in 2018. She is currently pursuing the master's degree in software engineering with Sun Yat-sen University, Guangzhou, China.

Her research interests include computer vision, pattern recognition, image processing, machine learning, scene graph generation, visual and language navigation, and bad weather restoration.

Bingqian Lin received the B.E. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2016, and the M.E. degree in computer science from Xiamen University, Xiamen, China, in 2019. She is currently pursuing the D.Eng. degree with the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China.

Her research interests include multiview clustering, image processing, and vision-and-language understanding.

Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He has been a Professor with the University of Chinese Academy of Sciences, Beijing, since 2009. He was a Visiting Assistant Professor with the Institute of Advanced Computer Studies, University of Maryland at College Park, College Park, MD, USA, in 2013. He has authored over 50 articles in refereed conferences and journals. He pioneered the kernel support vector machine (SVM)-based pyrolysis output prediction software, which was put into practical application by SINOPEC in 2012. He developed two kinds of piecewise linear SVM methods, which were successfully applied to visual object detection. His current research interests include image processing, visual object detection, and machine learning.

Dr. Ye received the Sony Outstanding Paper Award.

Jianbin Jiao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively.

From 1997 to 2005, he was an Associate Professor with HIT. Since 2006, he has been a Professor with the University of the Chinese Academy of Sciences, Beijing, China. His research interests include image processing and pattern recognition.

Liang Lin (Senior Member, IEEE) served as the Executive Director of the SenseTime Group, Hong Kong, from 2016 to 2018, leading the research and development teams in developing cutting-edge, deliverable solutions in computer vision, data analysis and mining, and intelligent robotic systems. He is currently the Chief Executive Officer (CEO) of Dark Matter Artificial Intelligence (DMAI) Great China, Guangzhou, China, and a Full Professor of computer science with Sun Yat-sen University, Guangzhou. He has authored or coauthored more than 200 articles in leading academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision (IJCV), Computer Vision and Pattern Recognition (CVPR)*, International Conference on Computer Vision (ICCV), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), and The Association for the Advancement of Artificial Intelligence (AAAI).

Dr. Lin is a fellow of IET. He was the recipient of Annual Best Paper Award by *Pattern Recognition* (Elsevier) in 2018, the Diamond Award for Best Paper at the IEEE International Conference on Multimedia and Expo (ICME) in 2017, the ACM Non-Photorealistic Animation and Rendering (NPAR) Best Paper Runners-Up Award in 2010, the Google Faculty Award in 2012, the Award for the Best Student Paper in IEEE ICME in 2014, and the Hong Kong Scholars Award in 2014. He served as the Area/Session Chair for numerous conferences, such as CVPR, ICME, ICCV, and International Conference on Multimedia Retrieval (ICMR). He is an Associate Editor of the IEEE TRANSACTIONS ON HUMAN–MACHINE SYSTEMS and *IET Computer Vision*.

Xiaodan Liang (Member, IEEE) received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2016, under the supervision of Liang Lin.

She was a Post-Doctoral Researcher with the Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA, working with Prof. Eric Xing from 2016 to 2018. She is currently an Associate Professor with Sun Yat-sen University. She has authored or coauthored several cutting-edge projects on human-related analysis, including human parsing, pedestrian detection and instance segmentation, 2-D/3-D human pose estimation, and activity recognition.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on August 04,2022 at 14:03:29 UTC from IEEE Xplore. Restrictions apply.