Visual Relationship Detection: A Survey

Jun Cheng, *Member, IEEE*, Lei Wang, *Member, IEEE*, Jiaji Wu, *Member, IEEE*, Xiping Hu, *Member, IEEE*, Gwanggil Jeon, *Member, IEEE*, Dacheng Tao, *Fellow, IEEE*, and Mengchu Zhou, *Fellow, IEEE*

***Abstract*—Visual relationship detection (VRD) is one newly developed computer vision task, aiming to recognize relations or interactions between objects in an image. It is a further learning task after object recognition, and is important for fully understanding images even the visual world. It has numerous applications, such as image retrieval, machine vision in robotics, visual question answer (VQA), and visual reasoning. However, this problem is difficult since relationships are not definite, and the number of possible relations is much larger than objects. So the complete annotation for visual relationships is much more difficult, making this task hard to learn. Many approaches have been proposed to tackle this problem especially with the development of deep neural networks in recent years. In this survey, we first introduce the background of visual relations. Then, we present categorization and frameworks of deep learning models for visual relationship detection. The high-level applications, benchmark datasets, as well as empirical analysis are also introduced for comprehensive understanding of this task.**

***Index Terms*—Deep learning, detection, neural networks, visual relation.**

I. Introduction

HUMAN learn to understand the environment through multiple ways, including auditory, visual, tactile, and olfactory sensations. Among these sensations, vision provides most of the needed information. Visual cognition techniques based on computer vision have obtained considerable progress in recent years especially with the advancement of deep learning.

Visual cognition techniques include image classification [1]–[5], object detection [6]–[10], semantic segmentation [11]–[13], instance segmentation [14], image caption [15], [16], visual question answer (VQA) [17], visual relationship detection (VRD) [18], etc. VRD, the main concern of this article, is to predict the relations or interactions between paired objects in one image.

Visual cognition techniques enable computers to understand the environment for target tasks. Visual relations, such as ⟨wheel on motorcycle⟩, ⟨person ride bike⟩, ⟨person hold camera⟩, and ⟨laptop on table⟩ in Fig. 1, reveal in-depth semantic meaning of images, and provide a further step to the goal of holistic image understanding, since images are more than the sum of their parts. The detection of visual relations is one kind of mid-level vision task that can obtain information from low-level vision task (object detection and recognition), and be helpful for high-level interpretations (e.g., image retrieval [18], VQA [19], [20], image caption [21], visual reasoning [22], etc.).

VRD is one challenge problem due to the following reasons [23]. First, objects in images may not be completely localized. Second, relations between given paired of objects may not be completely annotated, since every object can be related with each other even in a rare or infrequent type, as well as related with lots of attribute types. Therefore, it is impossible to exhaustively search all possible relationships. Thirdly, relations can be defined in different ways, and their appearance changes drastically. As a result, the distribution of relations is much more long-tailed than objects. Relation is also represented as "*predicate*" [18], [24], so for N objects and K predicates, the complexity of learning relations is $O(N^2K)$, but it is difficult to obtain sufficient training examples for all possible relationships. As a result, only a handful of relationships has been detected in the early works [25]–[27].

Deep learning has achieved great advance for many tasks [28] due to its good feature representation

Manuscript received January 9, 2020; revised August 30, 2020 and July 20, 2021; accepted January 7, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20487 and Grant 61775175; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515012637; in part by the CAS Key Technology Talent Program, Colleges and Universities Key Laboratory of Intelligent Integrated Automation under Grant 201502; in part by the CAS Key Laboratory of Human–Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences under Grant 2014DP173025; in part by the Guangdong-Hong Kong-Macao Joint Laboratory of Human–Machine Intelligence-Synergy Systems under Grant 2019B121205007; in part by the Shenzhen Engineering Laboratory for 3D Content Generating Technologies under Grant [2017] 476; and in part by the Shenzhen Technology Project under Grant JCYJ20170413152535587 and Grant JCYJ20180507182610734. This article was recommended by Associate Editor P. Shi. *(Corresponding author: Lei Wang.)*

Jun Cheng and Lei Wang are with the Shenzhen Institute of Advanced Technology, CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, and Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: jun.cheng@siat.ac.cn; lei.wang1@siat.ac.cn).

Jiaji Wu is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: wujj@mail.xidian.edu.cn).

Xiping Hu is with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 510275, China (e-mail: huxiping@mail.sysu.edu.cn).

Gwanggil Jeon is with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: gjeon@inu.ac.kr).

Dacheng Tao is with the JD Explore Academy, JD.com, Beijing 100176, China (e-mail: dacheng.tao@sydney.edu.au).

Mengchu Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: mengchu.zhou@njit.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3142013.

Digital Object Identifier 10.1109/TCYB.2022.3142013

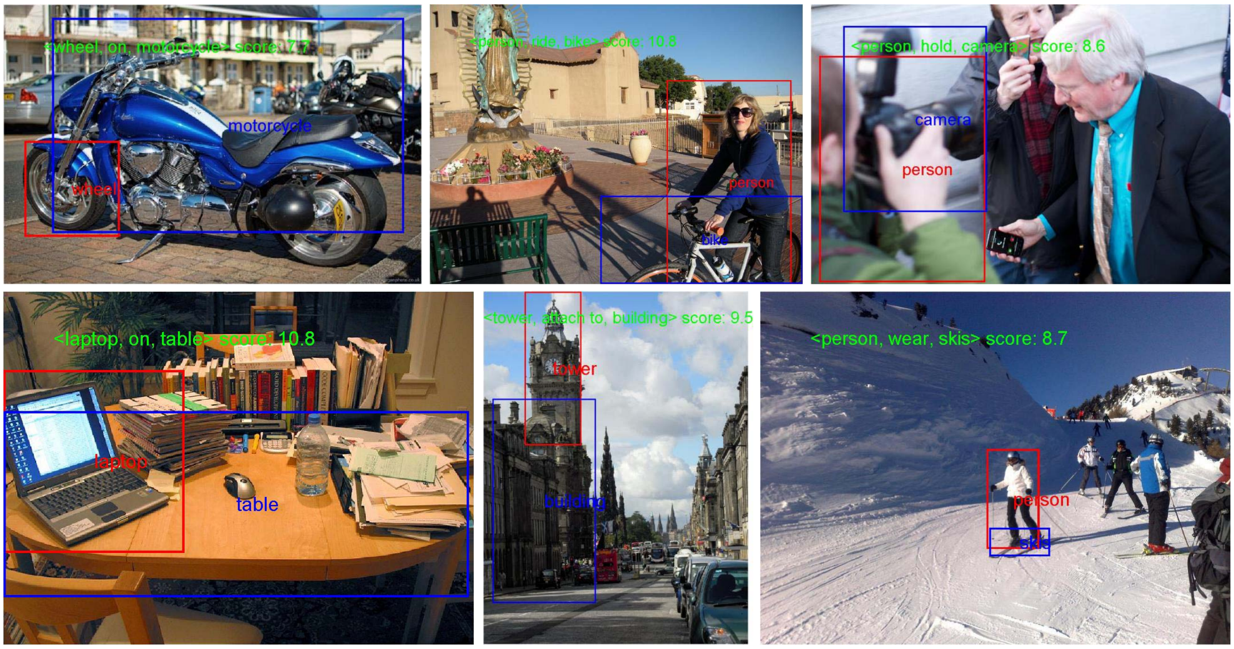

<wheel, on, motorcycle> score: 7.7
motorcycle
wheel
<person, ride, bike> score: 10.8
person
bike
<person, hold, camera> score: 8.6
camera
person
<laptop, on, table> score: 10.8
laptop
table
<tower, attach to, building> score: 9.5
tower
building
<person, wear, skis> score: 8.7
person
skis

Fig. 1. Visual relationships in images.

ability [29], [30], and it performs well in visual-related applications, such as recognition [31]–[33]; medical image segmentation [34]; pose estimation [35], [36]; image captioning [37]; video captioning [38]; image style transfer [39]; and cross-modal retrieval [40]. Various methods based on deep learning have also been proposed for visual relation detection with different models, corresponding target functions, and algorithms. Besides the object detection, relationship detection is investigated through learning modules, such as the conditional random field (CRF) network [41], relation proposal network [42], visual translation embedding (TransE) [24], and knowledge-based feature refinement [43].

In this article, we make a survey about visual relationship detection, in the purpose of making an analysis on the topic to provide a comprehensive summary for researchers who are interested in this area. We will introduce the background of visual relationship detection, including its brief history, comparison with visual phrase, and scene graph. Also, we will investigate the approaches proposed to tackle this challenge, especially recent deep neural network (DNN)-based methods. For relationship detection, different modules have been integrated into the DNN framework, such as prior and posterior statistics, knowledge regularization, TransE, attention models, and targeted costs and losses. We will introduce these categories and their frameworks. Specific datasets are necessary for this task to train the modules, so we also introduce them. A brief summary of visual relationship has been depicted in Fig. 2.

The remainder of this article is organized as follows. In Section II, we present the background of visual relationship detection. Section III will introduce categorization and frameworks for visual relationship detection. High-level applications will be briefly introduced in Section IV. Datasets and empirical analysis will be described in Sections V and VI, respectively. Finally, we provide summary and discussions in Section VII.

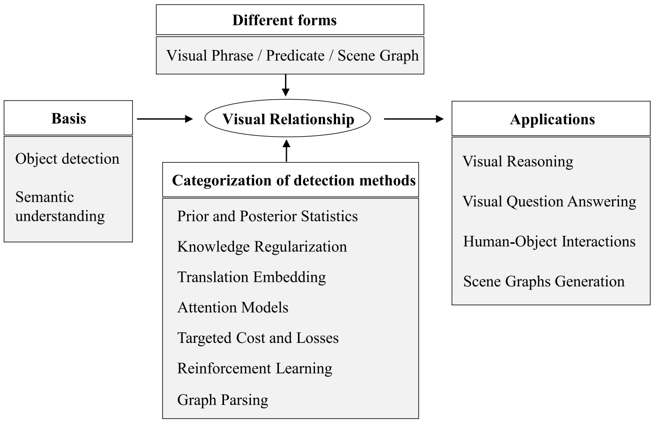


Fig. 2. Brief summary of visual relationship, including its basis, different forms, categorization of its detection methods, and applications.

II. Background

In this section, we outline the background of visual relation detection, including its brief history, handful of visual relationships, and comparison with visual phase and scene graph.

A. Brief History of Visual Relationship Detection

In the early works, visual relations have been learned to improve the object detection performance [25]–[27]. Although these works are not designated for VRD and only a small number of relations have been studied, they have introduced common relations between object pairs, such as *locations* and *comparison of size*, which are still learned by current methods. So we will introduce handful of visual relations in the next section, including *spatial object–object interaction*, *preposition and comparative adjective relations*, as well as human-object interactions (HOIs). To detect these relations, machine learning algorithms have been used, such as max-margin learning [25], Bayesian network [26], and structured learning [27].

With the development of deep learning, a large number of methods based on DNNs are proposed for visual relationship detection. The first was presented by Lu *et al.* [18], which used RCNN to detect objects and predicates with language priors to leverage predicted relationship's likelihood. Since then, there have been increasing improvements on DNNs-based visual relationship detection. A variety of modules has been proposed, including *prior and posterior statistics* [41]; *knowledge regularization* [44]; *TransE* [24]; *attention models* [45], [46]; *targeted cost and losses* [23], [47]; and reinforcement learning (RL)-*based framework* [48]. Details of the categorization of these modules are introduced in Section III.

B. Handful of Visual Relationships

The spatial object–object interaction is one kind of intuitive relation, which has been discussed in [25]. In their work, a single model predicts a structured labeling of the entire image, which incorporates interactions, including within-class (*textures of objects, NMS,* and *expected counts*) and between-class (*spatial cueing, mutual exclusion,* and *co-occurrence*).

The second kind of common interaction includes prepositions and comparative adjectives relation between objects [26]. In their training set, each image is annotated with nouns and relationships between subsets of pairs of these nouns. Relationship types are represented by a vocabulary of prepositions and comparative adjectives, according to their locations and appearances. A vocabulary of 173 nouns and 19 relationships has been defined, including *locations, size, color,* and *appearance* (such as *above, larger, greener,* and *more textured*).

HOI is another kind of visual relation. In the work of [27], six activity classes have been explored, including: 1) ⟨cricket − defensive shot⟩; 2) ⟨cricket − bowling⟩; 3) ⟨croquet − shot⟩; 4) ⟨tennis−forehand⟩; 5) ⟨tennis − serve⟩; and 6) ⟨volleyball − smash⟩. They proposed a random field model to encode the mutual context of objects and human poses in HOI activities. Then, the task is casted as a structure learning problem.

The above-mentioned three examples can be viewed as handful of visual relationships.

C. Visual Relationship Versus Visual Phrase

Visual phrase can be viewed as an early version of the visual relationship. According to [49] and [50], visual phrase corresponds to chunks bigger than objects and smaller than scenes, such as ⟨person riding horse⟩ and ⟨person sitting on sofa⟩. Phrase recognition was first proposed in [49] for the detection of complex visual composites and reasoning about relations between component objects. Farhadi and Sadeghi [50] used the deformable part models to detect objects and visual phrases. The visual phrase detector performs well for multiclass objects detection and relation reasoning in their dataset. When visual relationship is viewed as a phrase, its detection can be formulated as a three interconnected recognition problem [51]. To detect ⟨subject, predicate, object⟩ simultaneously, a visual phrase guided convolutional neural network (ViP-CNN) has been proposed [51], in which a phrase-guided message passing structure (PMPS) is designed to explore the connections of relationship components.

Visual phrase is still included in relationship detection experiments, which will be introduced in the Section VI.

D. Visual Relationship Versus Scene Graph

According to the definition in [52], a scene graph is a data structure that describes the contents of a scene. Relationships between objects, together with objects and their attributes, are viewed as parts of a scene graph [48], [53]–[55]. Scene graph aims at formalizing a representation for images with a structured formalization similar to that of knowledge bases (KBs). In the scene graph, nodes stand for objects and they are connected via edges, which stand for pairwise relationships [56]. Details of scene graph will be introduced in Section IV.

III. Categorization and Frameworks

To learn rich variety of relationships, many methods have been proposed especially in recent years with the development of deep learning, and more kinds of relations have been exploited. The first stage of VRD is always object detection, and the convolutional neural network (CNN)-based detectors are usually used to explore object pairs, followed by relationship learning and prediction. To explore the specific characteristics of relationship, different modules have been designed, such as prior and posterior statistics, knowledge regularization, TransE, attention models, and targeted costs and losses. Object detection always takes advantage of the state-of-the-art methods, so we will focus on the relationship detection modules. A brief overview of typical VRD methods has been given in Table I.

A. Prior and Posterior Statistics

First, we introduce some methods based on prior and posterior statistics. Since the latent semantic space of possible relationships is very huge, their prior or posterior statistical distribution will be beneficial for detection from a few examples. Relationships are always related to one another, such as ⟨person − ride − horse⟩ and ⟨person − ride − elephant⟩, which are semantically similar. So language priors have been used to project relationships into an embedding space where similar relationships are optimized to be close together [18].

The relationship projection function in [18] is defined as

$$f\big(\mathcal{R}_{\langle i,k,j\rangle}, \mathbf{W}\big) = \mathbf{w}_k^T\big[\text{word2vec}(t_i), \text{word2vec}\big(t_j\big)\big] + b_k \quad (1)$$

where word2vec() represents the converting function, and t_j is the word of the jth object category. $\mathbf{W} = \{\{w_1, b_1\}, \ldots, \{w_k, b_k\}\}$ is a set of predicates. Ranking loss

TABLE I
OVERVIEW OF THE TYPICAL VRD METHODS

Category	Method	Features	Pros	Cons
Statistics	VR-LP [18]	Language priors embedding	The first for large-scale relationships	Only prior is highlighted
	VRL [48]	Semantic graph for correlation	First proposed language priors	Higher computational cost
	DR-Net [41]	Based on posterior statistics	Formulate the posterior into relation	Ignores co-occurrence of predicates
Knowledge Regularization	DLK [44]	Teacher-student Distillation	Employ a teacher-student scheme	Enlarge parameter space
	SK [57]	Semantic knowledge	Semantic knowledge distillation	Higher computational cost
	KB-GAN [43]	Knowledge-base refinement	Generator and discriminator for KB	More training time
Translation Embedding	VTransE [24]	Visual translation embedding	Concise and end-to-end framework	Insufficient for context learning
Attention Models	AP+C+CAT [45]	Context-aware interaction	Context-aware attention	Inefficient for semantic
	Referring Relation [46]	Symmetric stacked attention shift	Unambiguously identify entities	Focus on entities localization
	LGRAN [58]	Language-guided graph attention	Node and edge attention mechanism	Less experiments on VRD
	HGAT [59]	Hierarchical graph attention	Attend object and triplet dependencies	Using multiple modules
Targeted Cost and Losses	VRD-DSR [23]	Deep structured ranking	A new ranking objective function	Ignores triplet's global context
	Image-Language-Cues [47]	Comprehensive language-image cues	Appearance/size/position/adjective/spatial	Needs multiple networks
	Spatial-Distribution [60]	Spatial distribution	Spatial with visual and concept	Inefficient on relationship
	RelDN [61]	Graphical contrast loss formulation	Class-agnostic/entity/predicate loss	Long training time
Reinforcement Learning	VRL [48]	Using Deep Q-Network	Reinforcement learning	Needs variation-structured traversal
Graph Parsing	MSDN [64]	Multi-level Scene Description	Jointly modeling three tasks	Insufficient for global context
	MotifNet [65]	Stacked Motif Network	BiLSTM for global context	Learning bias of the dataset
	RelDN [61]	Margin-based triplet loss	Disambiguate proximal relationship	Less robustness for other tasks

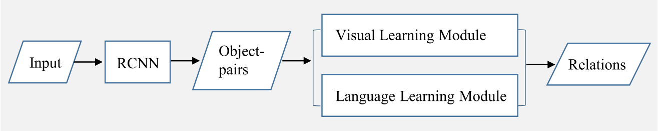


Fig. 3. VRD pipeline of [18].

functions have been defined for training to choose correct relationships. The learned model is used to predict relationships as

$$\mathcal{R}^* = \arg\max_{\mathcal{R}} V(\mathcal{R}, \Theta|\langle O_1, O_2\rangle) f(\mathcal{R}, \mathbf{W}). \tag{2}$$

The VRD pipeline has been shown in Fig. 3. The relationship embedding space learned from language is used to fine-tune the likelihood of a predicted relationship. Their model can scale to predict thousands of types of relationships from a few examples. In the work of [48], language priors have been used to build a semantic graph, in which the nodes are nouns, attributes, and predicates, connected by directed edges that represent semantic correlations.

Based on posterior statistics, one integrated network—deep relational network (DR-Net), has been proposed to learn the relations between *object categories* and *relationship predicates* [41]. Statistical dependency between the relationship predicate and the object categories will help to restrict the relationship to be more reasonable. Statistical relations have been exploited based on DR-Net in [41] for VRD to resolve the ambiguities caused by visual or spatial cues. The posterior probability for relation r is formulated as

$$\mathbf{q}_r = \sigma\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right) \tag{3}$$

where σ represents the activation function, and $\mathbf{W}_{rs} = \varphi_{rs}(r, s)$ and $\mathbf{W}_{ro} = \varphi_{ro}(r, o)$ represent potentials that capture the statistical relations among the relationship predicate r, the subject category s, and the object category o. $\mathbf{x}_r$ represents the compressed pair feature that combines both the appearance of the enclosing box and the spatial configurations. The generation of s and o adopts a similar method, and the updated probability vectors can be obtained as

$$\mathbf{q}_s' = \sigma\left(\mathbf{W}_a\mathbf{x}_s + \mathbf{W}_{sr}\mathbf{q}_r + \mathbf{W}_{so}\mathbf{q}_o\right) \tag{4}$$

$$\mathbf{q}_r' = \sigma\left(\mathbf{W}_r\mathbf{x}_r + \mathbf{W}_{rs}\mathbf{q}_s + \mathbf{W}_{ro}\mathbf{q}_o\right) \tag{5}$$

$$\mathbf{q}_o' = \sigma\left(\mathbf{W}_a\mathbf{x}_o + \mathbf{W}_{os}\mathbf{q}_s + \mathbf{W}_{or}\mathbf{q}_r\right). \tag{6}$$

DR-Net is realized by unrolling this iterative updating procedure into a network with a sequence of computing layers (i.e., these updating formulas).

The prior and posterior statistics will leverage word embeddings in large-scale relationship detection, and language prior has been used as a basic component in many methods. The structured correlations can be further exploited between objects and relationships.

B. Knowledge Regularization

Commonsense knowledge helps human being to reason about visual relations, so it can be used to refine the features of objects and relations.

The knowledge of linguistic statics has been used to regularize visual model learning, which is obtained by mining from training annotations (internal) and publicly available text, for example, Wikipedia (external), followed by computing the conditional probability distribution of a predicate given a ⟨subject, object⟩ pair [44]. The knowledge is distilled into a deep learning model and trained in a teacher–student knowledge distillation framework. We use T-Net and S-Net to represent the teacher and student network, respectively. After constructing a T-Net, its optimization function is defined as the KL-divergence of T-Net and S-Net prediction distributions as follows:

$$\min_{t\in T} \mathrm{KL}(t(Y)||s_\phi(Y|X)) - \lambda\mathbb{E}_t[L(X, Y)] \tag{7}$$

where $t(\cdot)$ and $s_\phi(\cdot)$ represent predictions of the T-Net and S-Net; ϕ is the S-Net's parameter set; and $L(\cdot)$ is a constraint function. λ is a balancing term. The knowledge distillation framework has been shown in Fig. 4.

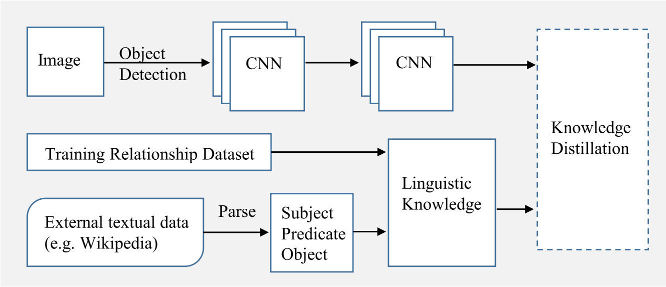


Fig. 4. Linguistic knowledge distillation framework.

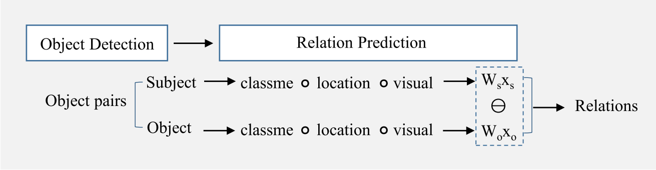


Fig. 5. VRD Framework based on VTransE. ○ and ⊖ denote vector concatenation and elementwise subtraction, respectively. W_s and W_o are two projection matrices learned by VTransE from the feature space to the relation space. x_s and x_o are features of subject and object, respectively.

Nevertheless, important knowledge cannot be distilled for every predicate by the method in [44] when taking into account a context of thousands of classes. So another semantic knowledge distillation scheme has been proposed in [57] to treat extensive classes and strengthen scalability by restricting the burden of directly using large external corpora. The network can make better use of the statistical and semantic dependencies between object and predicate classes, which is extracted from precomputed models and training annotations.

Knowledge-based feature refinement is present to improve the feature representation by taking advantage of the commonsense relationships [43]. External KB is used, which includes semantic entities. By retrieving from KB using the object label, the most common relationships can be identified.

The advantage of this kind of method is its ability to address the long-tail problem in the relationship detection, since the external knowledge consists of statistics about the words used to represent the relationship between subject and object pairs.

C. Translation Embedding

TransE can represent $\langle \text{subject} - \text{predicate} - \text{object} \rangle$ by low-dimensional vectors, $\boldsymbol{s}$, $\boldsymbol{p}$, and $\boldsymbol{o}$ respectively. $\boldsymbol{s} + \boldsymbol{p} \approx \boldsymbol{o}$ represents a translation of the relation when it holds, and $\boldsymbol{s} + \boldsymbol{p} \neq \boldsymbol{o}$ otherwise.

A visual TransE network (VTransE) is proposed for visual relation detection in [24], the framework of which is shown in Fig. 5. The visual relations are modeled by mapping the features of objects and predicates in a low-dimensional space. To project feature space to the relation space, matrices $\mathbf{W}_s$ and $\mathbf{W}_o$ are learned by VTransE. When $\mathbf{x}_s, \mathbf{x}_o \in \mathbb{R}^M$ represent M-dimensional features of *subject* and *object*, $\boldsymbol{s}$ and $\boldsymbol{o}$ can be rewritten as $\boldsymbol{s} = \mathbf{W}_s \mathbf{x}_s$, and $\boldsymbol{o} = \mathbf{W}_o \mathbf{x}_o$, respectively. So the visual translation can be formulated as

$$\mathbf{W}_s \mathbf{x}_s + \mathbf{t}_p \approx \mathbf{W}_o \mathbf{x}_o \tag{8}$$

where $\mathbf{t}_p \in \mathbb{R}^r (r \ll M)$ represents the relation translation vector to be learned. A loss function is defined as follows:

$$\mathcal{L}_{\text{rel}} = \sum_{(s,p,o)\in\mathcal{R}} - \log \text{softmax}\Big(\mathbf{t}_p^T (\mathbf{W}_o \mathbf{x}_o - \mathbf{W}_s \mathbf{x}_s)\Big) \tag{9}$$

where the *softmax* is computed over p. One feature extraction layer is designed after Faster-RCNN to incorporate knowledge transfer between objects and relations, including class probabilities, locations (i.e., bounding boxes' coordinates and scales), and ROI visual features.

TransE is efficient in modeling relational data and can be applied in social network analysis and recommender systems. The idea of learning embedding for visual relation is concise, but its performance still needs to be further exploited.

D. Attention Models

The attention mechanism has been widely used in many tasks, and it is also investigated for visual relationship detection. Context-aware attention models have been designed in [45], and they are imposed on the feature maps to select the interaction-specific discriminative feature regions. Based on these models, an interaction recognition framework is proposed for visual relationship detection. The "*predicate*" in those datasets is equivalent to the "*interaction*" in this literature. The context is encoded via *word2vec* into a semantic space, which will benefit zero-shot generalization. Attention and predicate shift modules have been designed in the work of [46], aiming to locate a specific category in one image, and learn to move attention from one entity to another, respectively. An iterative model is introduced to localize two entities (i.e., *subject* and *object*) in the referring relationship. The attention module is formulated as an initial estimate of the *subject* and *object* localizations by approximating the maximizers with the soft attention Att(·) as follows:

$$\widehat{x}^0 = \text{Att}(\mathbf{f}, S) = \text{ReLU}(\mathbf{f} \cdot \text{Emb}(S)) \tag{10}$$

$$\widehat{y}^0 = \text{Att}(\mathbf{f}, O) = \text{ReLU}(\mathbf{f} \cdot \text{Emb}(O)) \tag{11}$$

where Emb(·) embeds the entity into a C-dimensional semantic space, and $\mathbf{f}$ represents feature maps extracted from the image.

A graph-based, language-guided attention mechanism is present in [58] for referring expression comprehension. This mechanism consists of node and edge attention components, designed for object regions and relationships, respectively. These methods always ignore the triplet-level dependencies, so a hierarchical graph attention network is introduced to further capture the dependencies on triplet level [59].

The core of the attention mechanism is to focus on key ares instead of the entire image. The above attention-based methods have explored the soft attention, while some others can be further explored, including local attention, general attention, or multihead attention.

E. Targeted Cost and Losses

Since the predicate samples within a same category are highly diverse, multiple cues have been used to reduce the ambiguity for relationship, with targeted cost and loss functions. A collection of linguistic and visual cues has been used for localization of phrases and detection of relationships [47], and specific cost functions have been designed.

Multicues have been integrated as input to a deep structural ranking (DSR) framework [23], including *visual appearance*, *spatial location*, and *semantic embedding* cues, and a ranking objective function is designed with a structure ranking loss to learn the related predicate for a pair of localized objects, by enforcing the annotated relationships to have higher relevance scores. The loss is defined as

$$\mathcal{L}(x) = \sum_{r\in\mathcal{R}}\sum_{r'\in\mathcal{R}'}\big[\Delta(r,r') + \Phi(x,r') - \Phi(x,r)\big]_+ \qquad (12)$$

where x is the input image, $r = (s,p,o)$ is the relationship instance, $\mathcal{R} = (s,p,o)|(s,o) \in \mathcal{P} \wedge p \notin \mathcal{P}_{s,o}$ represents the visual relationships existing in one image, and $\mathcal{R}' = (s',p',o')|(s',o') \in \mathcal{P} \wedge p' \notin \mathcal{P}_{s',o'}$ is unannotated relationship instances. $[\,\cdot\,]_+ = \max(0,\cdot)$ works to retain positive parts. $\Delta(\cdot,\cdot)$ is a margin function to measure the incompleteness of visual relationship detection, defined as

$$\begin{aligned}\Delta(r,r') &= \Delta\big(s,p,o,s',p',o'\big)\\ &= 1 + P(p|c_s,c_o) - P\big(p'|c_{s'},c_{o'}\big)\end{aligned} \qquad (13)$$

where c_s and c_o represent *subject*'s and *object*'s category. Φ is a function to measure the compatibility between x and r as follows:

$$\Phi(x,r) = \Phi(x,s,p,o) = \mathrm{w}_p^T f(x,s,o) \qquad (14)$$

where w_p represents parameters to be learned for the pth predicate. This framework aims to facilitate the co-occurrence of relationships and mitigate the incomplete annotation problem.

Based on the idea that spatial distribution reflects positional relation (PR) of objects and describes structural information between objects, it is used to facilitate VRD in [60], combined with visual and concept features. The spatial distribution includes *PR*, size relation (SIR), *distance relation (DR)*, and shape relation (SHR). Graphical contrastive losses have been designed to target two types of errors, that is: 1) *entity instance confusion* and 2) *proximal relationship ambiguity* [61]. A predicate class-aware loss has also been designed for *proximal relationship ambiguity* in a similar way.

The design of loss functions is important since it will affect the training procedure of the learning modal. Different cost and loss functions will lead the modal to learn different kinds of visual relations.

F. RL-Based Framework

RL aims to learn an optimal or near-optimal policy to maximize the "reward function" that accumulates from the immediate rewards. Deep RL has been successfully applied in some areas, including games and robotics planning. In the work of [48], a deep variation-structured RL (VRL) framework is proposed aiming at capturing global semantic dependency between relationships and attributes. The key components of VRL include the *directed semantic action graph*, *variation-structured traversal scheme*, *state space*, and *reward functions*. The VRL framework formulates the problem of detecting visual relationships and attributes as a sequential decision-making process. It makes predictions by incorporating global context cues and semantic embedding of previously extracted phrases. The deep Q-network (DQN) [62], [63] is used to estimate network weights θ_a, θ_p, and θ_c of attributes $\mathcal{A}$, predicate categories $\mathcal{P}$, and object categories $\mathcal{C}$, respectively. The reward functions include $\mathcal{R}_a(\mathbf{f},\mathbf{g}_a)$, $\mathcal{R}_p(\mathbf{f},\mathbf{g}_p)$, and $\mathcal{R}_c(\mathbf{f},\mathbf{g}_c)$, which reflect the detection accuracy of taking action ($\mathbf{g}_a$, $\mathbf{g}_p$, $\mathbf{g}_c$) in state $\mathbf{f}$. The greedy strategy is used to select actions $\mathbf{g}_a$, $\mathbf{g}_p$, and $\mathbf{g}_c$, in the variation-structured actions in the training stage, and then the best actions with the highest estimated Q-values are selected in the testing stage to discover objects, relationships, and attributes. The RL-based methods will search for the optimal results in a wide range, but the complexity is high due to its multicomputation of forward search. The core elements (value function, policy, reward, planning, etc.) of RL should be further researched.

G. Graph Parsing

Visual relationship has also been exploited in the scene graph, which outlines the entities, attributes, and relationships as a structured representation.

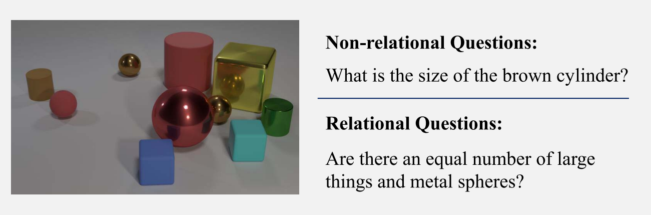
A multilevel scene description network (MSDN) [64] is proposed for graph parsing to jointly model three vision tasks, including: 1) object detection; 2) VRD; and 3) region captioning. The connections between objects, phrases, and region captions are built based on the spatial and semantic relationships. A stacked motif network (MotifNet) is introduced in [65] for graph parsing. Based on the analysis on one dataset, the authors found that half of images have at least one motif involving two relationships. So the MotifNet is designed to encode the global context to inform local predictors, and bidirectional LSTMs have been used. RelDN [61] is proposed to detect relationships with the graphical contrastive loss in graph parsing. Scene graph is a higher level form to represent semantic information including relationships. The graph-related theory and methods will benefit the parsing problem.

IV. High-Level Applications of Visual Relation

In this section, we will introduce some high-level applications of visual relations.

A. Visual Reasoning

Visual reasoning is one direct application of the visual relations. This task aims to infer the intrinsic and implicit informations in one image, such as counting and comparing [66], [67]. One example from the compositional language and elementary visual reasoning (CLEVR) dataset [68] has been shown in Fig. 6. The CLEVR diagnostics dataset has been built with the goal of enabling detailed analysis of visual reasoning [68]. This dataset contains $100K$ rendered images and about one million automatically generated questions. There are two types of relationships in the CLEVR dataset: 1) *spatial* and 2) *same-attribute*. *Spatial* relationships include "*left*," "*right*," "*behind*," and "*in front*." The



Non-relational Questions:

What is the size of the brown cylinder?

Relational Questions:

Are there an equal number of large things and metal spheres?

Fig. 6. One example from the CLEVR dataset of relational reasoning.

same-attribute relation is defined for two objects if they have equal attribute values for a specified attribute.

The relation network (RN) has been proposed in [69] as a plug-and-play module for relational reasoning with the philosophy of constraining the functional form of a neural network to capture the core common properties of relational reasoning. One simplest form of RN is defined as

$$\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta\left(o_i, o_j\right)\right) \tag{15}$$

where $f_\phi(\cdot)$ and $g_\theta(\cdot)$ are multilayer perceptrons with learnable parameters ϕ and θ, respectively. O represents the input to the network, and it is a set of objects formulated as $O = \{o_1, o_2, \ldots, o_n\}$, with the ith object $o_i \in \mathbb{R}^n$. A "*relation*" can be output from g_θ to infer how two objects are related.

In [66], inferring and executing programs have been proposed for visual reasoning, both of which are realized by neural networks and trained using backpropagation and reinforce. The inferring program constructs an explicit representation of the reasoning process. The executing program executes the resulting program to produce an answer. A graph neural network architecture is proposed in [70] for visual reasoning that deals with the composite object relationships in the scene as a graph traversal problem.

B. Visual Question Answering

VRD will also benefit visual question answering (VQA) [20], [71]–[74]. VQA is always viewed as a classification problem, formulated as

$$\hat{a} = \arg\max_{a\in\Omega}\ p(a|Q, I; \Theta) \tag{16}$$

where $\hat{a}$ stands for the most possible answer, Q represents a question, I is a related image, Θ denotes one model's parameters, and Ω is the collection of candidate answers. In [75], visual relation facts have been learned for VQA, and a Relation-VQA dataset has been built based on the visual genome (VG) dataset. A relation detector is used to predict visual question-related relation facts. The feature representations v and q of image and question are obtained from convolution and recurrent networks, as $v = \mathrm{Meanpooling}(\mathrm{CNN}(I))$ and $q = \mathrm{GRU}(Q)$, respectively. Then, they are transformed by linear and nonlinear function as $f_v = \tanh(W_v v + b_v)$ and $f_q = \tanh(W_q q + b_q)$, respectively. To fuse the image and question representations, another active function is used as $h = \tanh(W_{vh} f_v + W_{qh} f_q + b_h)$, so that a joint semantic feature embedding can be learned.To predict the classification results of *subject*, *relation*, and *object*, the *softmax* function is used as

$$p_{\text{subject}} = \mathrm{softmax}(W_{hs}h + b_s) \tag{17}$$

$$p_{\text{relation}} = \mathrm{softmax}(W_{hr}h + b_r) \tag{18}$$

$$p_{\text{object}} = \mathrm{softmax}(W_{ho}h + b_o). \tag{19}$$

In [76], a visual knowledge memory network (VKMN) is proposed for VQA, which incorporates structured human knowledge and deep visual features into memory networks. The knowledge consists of triples with the structure of ⟨subject, relation, target⟩. Their visual KB is built up on knowledge entries from one VQA dataset [72] and VG relationship dataset. A multimodal relational network for VQA is proposed in [71], and region relations have been exploited to reason over multiple objects that interact with each other. Spatial and semantic concepts will be learned, such as *on top of, left, right, wear, hold,* etc. Two question-adaptive visual relation (binary and trinary) attention modules have been designed in the multimodal relation attention network (MRA-Net) [20]. To overcome the flaws of popular benchmarks (such as real-world priors and statistical biases within the answer distribution), a new dataset GQA is present recently for visual reasoning and compositional question answering [77]. More than 22 million novel and diverse questions have been generated.

Visual reasoning is one kind of general intelligent behavior, which requires the ability of sophisticated reasoning but not only mapping inputs to outputs with black-box architectures that always exploit biases in the data. The situation is same for VQA, while short-term memory [78]–[81], disentangled representation [29], as well as different patterns of reasoning [82], [83] will help the research in this area.

C. Human-Object Interactions

HOI is one kind of special visual relation, in which predicates always represent actions. Except for the early work of [27], there are some other approaches.

In [84], a human-centric approach is proposed, taking the appearance of a person (i.e., *pose, clothing,* and *action*) as a powerful cue to localize the objects with which the person is interacting. A framework based on Fast-RCNN is introduced to detect interactions in [85]. The framework includes three branches, that is: 1) *object detection*; 2) *human-centric*; and 3) *interaction*. To tackle the challenge of long-tail distribution of HOI categories, a knowledge graph is constructed in [86] based on ground-truth annotations of training dataset and external source. To retrieve verbs that describe the detected ⟨*human*, *object*⟩ pair, a semantic structure-aware embedding space is learned to leverage semantic similarity. This method has been tested on the V-COCO [87] and HICO-DET [88] datasets. Human are always the centric role in images, so HOI is important in many applications, such as human–machine interaction and robotics.

D. Scene Graphs

As introduced in Section II, a scene graph is a data structure that describes the contents of a scene. One simple example of

TABLE II
STATISTICS OF VARIOUS DATASETS

	Visual Phrases	Scene Graph	VRD	Visual Genome
Images	2,769	5,000	5,000	108,077
Objects Categories	8	266	100	33,877
Visual Phrases	17	–	–	–
Relationship Types	13	23,190	6,672	42,374
Relationship Instances	–	109,535	37,993	2,269,617
Attribute Categories	–	145	–	68,111
Predicates per Obj. Category	–	2.3	24.25	–

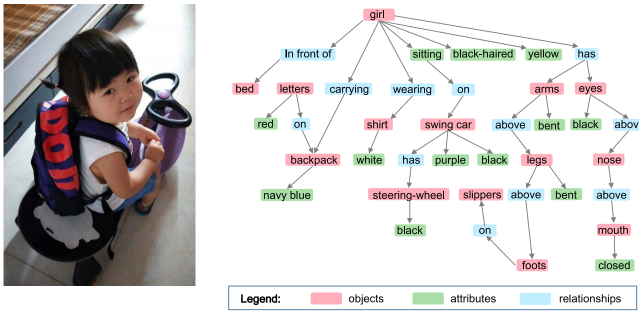


Fig. 7. Simple example of a scene graph.

the scene graph is shown in Fig. 7. The scene graph captures the detailed semantics of visual scenes and describes them without relying on the unstructured text.

A scene graph can be represent as a tuple $G = (O, E)$, where $O = \{o_1, \ldots, o_n\}$ is a set of objects, and $E \subseteq O \times \mathcal{R} \times O$ is a set of edges. Each object has a form of $o_i = (c_i, A_i)$, where $c_i \in \mathcal{C}$ is the object's class and $A_i \subseteq \mathcal{A}$ are attributes of the object. When using a set of bounding boxes B to represent an image, a map $\gamma : O \to B$ represents a grounding of a scene graph. Then, the objective can be formulated as

$$\gamma^* = \arg\max_{\gamma} \prod_{o \in O} P(o|\gamma_o) \prod_{(o,r,o') \in E} P\big(\gamma_o, \gamma_{o'}|o, \gamma, o'\big) \quad (20)$$

where $P(o|\gamma_o)$ is used to measure the consistency between γ_o and o, while $P(\gamma_o, \gamma_{o'}|o, \gamma, o')$ is to model how well the bounding-box pair $(\gamma_o,\ \gamma_{o'})$ represents the tuple (o, γ, o'). Features can be extracted to encode the relative position and scale

$$f(\gamma_o, \gamma_{o'}) = \big((x - x')/w, (y - y')/h, w'/w, h'/h\big) \quad (21)$$

where $\gamma_o = (x, y, w, h)$ and $\gamma_{o'} = (x', y', w', h')$ represent coordinates of the bounding boxes.

A scene graph has been used as a query for semantic image retrieval to model multiple modes of interaction between object pairs [52]. A CRF model is designed to model the distribution over all possible groundings of scene graphs. A scene graph dataset is designed, which consists of 5000 samples grounded to images, and it is used to test image retrieval performance. The experiments show that the scene graph method performs better than that use only objects or low-level features for image retrieval.

In computer graphics, graphs have been used to represent structural relationships, which are learned in scenes [53]. This is proposed to solve the problem of data mining in a 3-D scene corpus, since the model density and complexity of the scenes representing virtual environments are rising rapidly. A kernel between these relationship graphs has been defined to compare virtual substructures in two graphs. The framework has been applied to scene modeling problems, such as finding similar scenes, relevance feedback, and context-based model search. In [54], a representation for common sense spatial knowledge is presented, and applied to the task of text-to-3D scene generation. A directed semantic action graph has been built to organize all possible object nouns, attributes, and relationships into a compact and semantically meaningful representation [48]. A 3-D scene graph has been proposed as a construction framework for the environment model, based on the consideration that efficient environment model will take a crucial role in the intelligent agent's autonomy system [89]. Besides similar representations as that of 2-D graphs, the 3-D scene graph extracts physical attributes within environments including 3-D positions. The applicability of the 3-D scene graph has been verified in VQA and task planning in a kitchen simulation environment.

V. DATASETS

To evaluate the performance of visual relation detection methods, several benchmark datasets have been built, and we will introduce them in this section, including visual phrase [49], [50]; scene graph [52]; VRD [18]; and VG [90]. Their statistic information has been summarized in Table II. We have listed some components of these datasets, including the number of *images*, *object categories*, *visual phrases*, *relationship types* and *instances*, *attribute categories*, and *predicates per object category*. GQA [77] is one recent public dataset aiming at visual reasoning and compositional question answering, while the relationship is one important part (51%) in the question semantic types. So we also briefly introduce this dataset. Amazon's Mechanical Turk (AMT) is used for all these datasets except visual phrase.

A. Visual Phrases

The visual phrases dataset contains 17 types of common relationship, and eight object classes [49], [50], which are select from the PASCAL VOC2008 dataset [91]. Objects include *person, dog, horse, car, bike, sofa, chair,* and *bottle*, which are suitable for modeling the interactions between each other. The bounding boxes are manually created with a total number of 5067 (3271 for objects and 1796 for visual phrase).

There are 17 visual phrases, including the interaction between objects or activities of objects. Most visual phrases are person-related activities (such as ⟨person – riding – bicycle⟩), and some are spatial relations (such as ⟨bicycle – next to – car⟩).

B. Scene Graph

The scene graph dataset [52] contains 5000 images (selected from the intersection of the YFCC100M [92] and Microsoft COCO datasets [93]), over 93 000 object instances, and 110 000 instances of attributes. The attribute of objects is one distinctive characteristic of this dataset, which describes *color, shape, appearance*, *material*, *weather*, *human-related*, *state*, *doing-action*, etc. The relation types in scene graph include *activities of the present/past indefinite tense/the present continuous tense*, *spatial*, *comparative*, *prepositions*, etc.

C. VRD

To detect more visual relationships, larger datasets are needed. The visual relationship dataset (VRD) has been present in [18], which contains 5000 images with 100 object categories and 70 predicates, as well as 37 993 relationships with 6672 relation types and 24.25 predicates per object category.

Compared with visual phrase, there are more diverse objects in the VRD dataset, including *person, animals, indoor-objects, building or natural objects, traffic/sports-related, clothes, food, and some others.* The predicates in VRD are similar with the relations in scene graph, such as *activities, spatial, preposition,* and *comparative*. This dataset is always used as a benchmark for evaluation.

D. Visual Genome

To draw the complete information of the visual world, the VG dataset has been built, which contains over 108K images with an average of 35 objects and 26 attributes for each image, and 21 pairwise relationships between objects [90].

Compared with previous datasets (such as Flickr 30K [94], MS-COCO [93], and VRD [18]), much denser and more complete set of descriptions have been provided for different regions and scene in this dataset. Besides the basic type of components (*objects, attributes,* and *relationships*), the mid-level (*region descriptions* and *question answer pairs*) and high-level components (*region graphs* and *scene graphs*) have also been provided in this dataset.

The VG dataset is dense and large. For instance, there are 3.8 million *object instances*, 2.8 million *attributes*, 2.3 million *relationships*, 5.4 million *region descriptions*, 3.7 million *region graphs*, 1.7 million *VQAs*, etc.

E. GQA

GQA is a recently published dataset [77], which is mainly designed for real-world reasoning and compositional question answering. Based on the VG dataset, 22 *million* diverse reasoning questions have been created in GQA by developing a question engine to leverage VG scene graphs (information of *objects*, *attributes*, and *relations*) on 113K images. The questions in GQA measure reasoning ability of *object/attribute recognition*, *transitive relation tracking*, *spatial reasoning*, *logical inference*, and *comparisons*. Although GQA is aiming at visual reasoning and question answering, visual relation is one important semantic type in this dataset.

VI. Empirical Analysis

There are three general tasks for the visual relation detection, including: 1) *predicate detection*; 2) *phrase detection*; and 3) *relationship detection*. The tasks and their corresponding output are listed as follows [18].

1) *Predicate Detection:*
   Output:
   a) A set of possible predicates between pairs of objects.
2) *Phrase Detection:*
   Output:
   a) A label ⟨object1 – predicate – object2⟩.
   b) Localize the entire relationship as one bounding box having at least 0.5 overlap with the ground-truth box.
3) *Relationship Detection:*
   Output:
   a) A set of ⟨object1 – predicate – object2⟩.
   b) Localize both *object1* and *object2* in the image having at least 0.5 overlap with their ground truth boxes simultaneously.

Mean average precision (mAP) is one widely used metric, but it may mistakenly penalize true positives for relationship detection since not all possible relationships can be exhaustively annotated in an image. So we follow most references to use the metric of $Recall@n(R@n)$, which computes the fraction of times the correct relationship is predicted in the top n confident predictions. Taking the VRD dataset (70 predicates and 100 objects) for instance, there will be $100 \times 70 \times 100$ possible relationship predictions, so the $Recall@100$ of random guess will be 0.000143 [18].

The performance of different methods on the VRD and VG datasets has been compared in Tables III and IV, respectively. From Table III, we can see that the *predicate detection* performance has been improved by a large mount from about 47%–86%, while the *relationship detection* performance has a limited improvement from 14% to 20%. Benefitting from the progress of object detection algorithms especially the relationship modules, the performance is improved. The external knowledge regularization and DSR exhibited better efficiency. The results are even lower on the VG dataset with an $R@50$ of 6% for relationship detection. These results illustrate that the relationship detection is a more difficult task than that of predicate, and more difficult in the large dataset due to the long-tail distribution effect.

The initial version of the VG dataset is noisy since it is annotated by crowd workers. Different researchers have processed the dataset by different ways in their experiments. Therefore, we also select methods evaluated on one cleaned version of the VG dataset to make a comparison, in which the relations with less than five samples have been filtered out. After filtering, there are 99 658 images, 200 object categories,

TABLE III
COMPARISON ON THE VRD DATASET (RECALL@N, %)

Methods	Predicate Detection		Phrase Detection		Relationship Detection	
	R@50	R@100	R@50	R@100	R@50	R@100
Visual Phrase [49]	–	–	0.54	0.63	–	–
VR-LP [18]	47.87	47.87	16.17	17.03	13.86	14.70
VTransE [24]	44.76	44.76	19.42	22.42	14.07	15.20
VipCNN [51]	–	–	22.78	27.91	17.32	20.01
VRL [48]	–	–	21.37	22.60	18.19	20.79
PPR-FCN [95]	47.43	47.43	19.62	23.15	14.41	15.72
SD [60]	51.50	51.50	16.94	18.89	14.31	15.77
AP+C+CAT [45]	53.59	53.59	17.60	19.24	15.63	17.39
DLK [44]	54.82	54.82	26.47	29.76	22.68	31.89
MotifNet [65]	65.20	67.10	–	–	–	–
SK [57]	71.02	80.80	–	–	–	–
DR-Net [41]	80.78	81.90	19.93	23.45	17.73	20.88
VRD-DSR [23]	86.01	93.18	–	–	19.03	23.29
FactorizeNet [99]	–	–	26.03	30.77	18.32	21.20
KB-GAN [43]	–	–	27.39	34.38	20.31	25.01

TABLE IV
COMPARISON ON THE VG DATASET (RECALL@N, %)

Methods	Predicate Detection	
	R@50	R@100
Visual Phrase [49]	0.63	0.84
JointCNN [96]	3.06	3.99
VR-LP [18]	53.49	54.05
DR-Net [41]	88.26	91.26
DLK [44]	92.31	95.68

TABLE V
COMPARISON ON THE *Cleaned-VG-v1* DATASET (RECALL@N, %)

Methods		Visual Phrase	VTransE	PPR-FCN	VRD-DSR
Predicate	R@50	–	62.63	64.17	69.06
	R@100	–	62.87	64.86	74.37
Phrase	R@50	3.41	9.46	10.62	–
	R@100	4.27	10.45	11.08	–
Relationship	R@50	–	5.52	6.02	–
	R@100	–	6.04	6.91	–

100 predicates, and 19 237 unique relations. In this article, we name this version of the VG dataset as the *Cleaned-VG-v1*. The dataset is split into 73 801 images for training and 25 857 for testing. The comparison results on the *Cleaned-VG-v1* have been given in Table V, including the methods of visual phrase [49], VTransE [24], PPR-FCN [95], and VRD-DSR [23]. There is another cleaned version of the VG dataset, by using an annotation refinement process to correct, delete, or merge duplicate bounding boxes. We name this version as the *Cleaned-VG-v2*, which contains an average of 25 distinct objects and 22 relationships per image. The most frequent 150 object categories and 50 predicates have been used for evaluation. The results of different methods on this dataset have been listed in Table VI. From Tables V and VI, it can be seen that VRD-DSR [23] and Pixel2Graph [98] perform better for the predicate detection, while KB-GAN [43] is better for relationship detection. External knowledge and image reconstruction loss have been used in KB-GAN [43] to overcome the dataset bias issues.

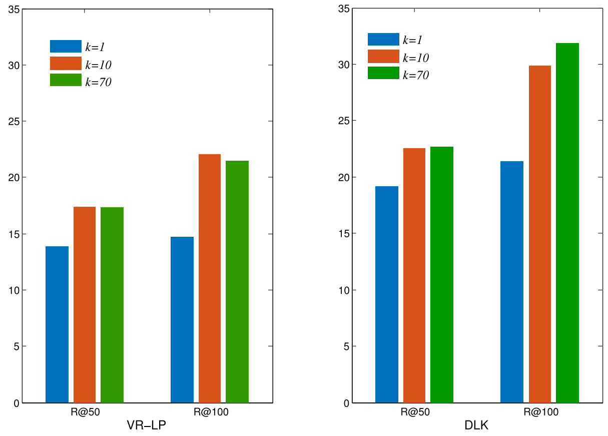


Fig. 8. Relationship detection (%) on the VRD dataset with different k.

A parameter k is introduced in [18] to make the evaluation more effective. First, the confidences of all predicates for a pair of objects are sorted, and then the top k predicates are taken into consideration to calculate the *Recall* without using all predicates in the dataset. Generally, k is set to be 1, 10, and 70 ($k = 70$ considers all predicates, which is equivalent to not filtering predicates). But the parameter has not been reported by most of the aforementioned methods, except for VR-LP [18] and DLK [44]. Their relationship detection results according to different k have been shown in Fig. 8, from which it can be seen that the parameter k affects the final performance significantly.

VII. SUMMARY AND DISCUSSIONS

In this article, we have reviewed about the visual relationship detection, which is a relatively new computer vision task. The purpose of this task is to exploit relations between objects in one image. Visual relationship can be used for high-level tasks, such as scene graph generation, visual reasoning, VQA, and HOI detection, and it will also

TABLE VI
COMPARISON ON THE *Cleaned-VG-v2* DATASET (RECALL@N, %)

Methods	**Predicate Detection**		**Phrase Detection**		**Relationship Detection**	
	R@50	R@100	R@50	R@100	R@50	R@100
IMP [97]	44.80	53.00	21.70	24.40	7.09	9.91
MSDN [64]	63.10	66.40	19.95	24.93	10.72	14.22
Pixel2Graph [98]	68.00	75.20	26.50	30.00	9.70	11.30
MotifNet [65]	41.80	48.80	23.8	27.20	–	–
Graph RCNN [42]	54.20	59.10	29.6	31.60	11.40	13.70
FactorizeNet [99]	–	–	22.84	28.57	13.06	16.47
KB-GAN [43]	–	–	23.51	30.04	13.65	17.57

benefit image retrieval with more details to search target precisely.

From handful of relations, visual phrases, and predicates, to scene graphs, visual relationships have been explored with different modalities in recent years. Many approaches have been proposed for this question, such as deep neural networks with prior and posterior statistics, knowledge distillation, TransE, and attention models.

Different datasets have been built with various kinds of relationships, including visual phrase, scene graph, VRD, and VG. The experimental results of different approaches have been compared on the most popular benchmark VRD and VG dataset. As shown in Section VI, the $R@50$ performance of the predicate detection has exceeded 80%. However, the recall performance on the phrase and relationship detection is still very low, since the dimensionality of the output of ⟨subject, predicate, object⟩ is much higher. Due to the long-tailed bias distribution, the evaluation metric of *Recall@n* may not be very appropriate for phrase and relationship detection, since it tends to be dominated by the performance of that with a large proportion of samples. This problem has also been discussed in [100] and [101], and *mean Recall@n* has been used to evaluate the performance of each relationship more comprehensively. A framework for unbiased scene graph generation has been proposed to address the bias issue [101]. The harmful bias is removed from the good context bias with counter-factual causality. It is pointed out that relationship is a context-based label. For example, whether the relation is ⟨human, on, horse⟩ or ⟨human, riding, horse⟩ depends on "what is on horse?" or "what is he doing?" It will be more important to extract meaningful relations and scene graphs other than to fit the bias distribution of the dataset.

The same relationship may be meaningful or meaningless for different objects, so one direction worthy investigating is the constraint-based value metric design for visual relationship. It will be an interesting approach to efficiently measure the relation's value for one image, although this depends on different targets. From our viewpoint, target-specific relationship detection will be more important, especially in the applications of VQA and visual reasoning. Visual relationships will also benefit the general artificial intelligence, although there will be a long way to go, since the ability of analysis, summary, deduction, and reasoning is also critical for its realization.

REFERENCES

[1] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 1794–1801.

[2] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 143–156.

[3] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1055–1064, Sep. 1999.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[5] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[6] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 936–944.

[11] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1520–1528.

[12] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[15] K. Xu *et al.*, "Show, Attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2015, pp. 2048–2057.

[16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[17] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 21–29.

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

[19] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 913–926, Feb. 2021.

[20] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, "MRA-Net: Improving VQA via multi-modal relation attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 318–329, Jan. 2022, doi: 10.1109/TPAMI.2020.3004830.

[21] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 9959–9968.

[22] 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.*, Salt Lake City, UT, USA, 2018, pp. 7239–7248.

[23] K. Liang, Y. Guo, H. Chang, and X. Chen, "Visual relationship detection with deep structural ranking," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7098–7105.

[24] 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)*, Honolulu, HI, USA, 2017, pp. 3107–3115.

[25] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, 2009, pp. 229–236.

[26] A. Gupta and L. S. Davis, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2008, pp. 16–29.

[27] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 17–24.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[30] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[31] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1–9.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[33] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1791–1802, May 2019.

[34] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Stanford, CA, USA, 2016, pp. 565–571.

[35] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 1653–1660.

[36] Y. Wu, W. Ji, X. Li, G. Wang, J. Yin, and F. Wu, "Context-aware deep spatiotemporal network for hand pose estimation from depth images," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 787–797, Feb. 2020.

[37] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 6837–6844.

[38] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.

[39] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2414–2423.

[40] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1979–1988.

[41] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 3298–3308.

[42] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11205, 2018, pp. 690–706.

[43] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1969–1978.

[44] 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)*, Venice, Italy, 2017, pp. 1068–1076.

[45] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 589–598.

[46] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei, "Referring relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6867–6876.

[47] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1946–1955.

[48] 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)*, Honolulu, HI, USA, 2017, pp. 4408–4417.

[49] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 1745–1752.

[50] A. Farhadi and M. A. Sadeghi, "Phrasal recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2854–2865, Dec. 2013.

[51] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: Visual phrase guided convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 7244–7253.

[52] J. Johnson *et al.*, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3668–3678.

[53] M. Fisher, M. Savva, and P. Hanrahan, "Characterizing structural relationships in scenes using graph kernels," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–12, 2011.

[54] A. X. Chang, M. Savva, and C. D. Manning, "Learning spatial knowledge for text to 3D scene generation," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 2028–2038.

[55] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph property sensing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 3743–3752.

[56] R. Krishna, V. Chen, P. Varma, M. Bernstein, C. Ré, and L. Fei-Fei, "Scene graph prediction with limited labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 2580–2590.

[57] F. Plesse, A. Ginsca, B. Delezoide, and F. Prêteux, "Visual relationship detection based on guided proposals and semantic knowledge distillation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, 2018, pp. 1–6.

[58] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. van den Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1960–1968.

[59] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 13883–13892.

[60] Y. Zhu, S. Jiang, and X. Li, "Visual relationship detection with object spatial distribution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, 2017, pp. 379–384.

[61] 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)*, Long Beach, CA, USA, 2019, pp. 11527–11535.

[62] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," in *Proc. Deep Learn. Neural Inf. Process. Syst. Workshop*, 2013, pp. 1–9.

[63] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[64] 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)*, Venice, Italy, 2017, pp. 1270–1279.

[65] 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. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 5831–5840.

[66] J. Johnson *et al.*, “Inferring and executing programs for visual reasoning,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 3008–3017.

[67] E. Perez, F. Strub, H. Vries, V. Dumoulin, and A. C. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 3942–3951.

[68] J. Johnson, B. Hariharan, L. V. D. Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1988–1997.

[69] A. Santoro *et al.*, “A simple neural network module for relational reasoning,” in *Advances in Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran, 2017, pp. 4967–4976.

[70] M. Haurilet, A. Roitberg, and R. Stiefelhagen, “It’s not about the journey; it’s about the destination: Following soft paths under question-guidance for visual reasoning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1930–1939.

[71] R. Cadene, H. Ben-Younnes, M. Cord, and N. Thome, “MUREL: Multimodal relational reasoning for visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1989–1998.

[72] A. Agrawal *et al.*, “VQA: Visual question answering,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 4–31, May 2017.

[73] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7W: Grounded question answering in images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4995–5004.

[74] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, “Universal weighting metric learning for cross-modal matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13002–13011.

[75] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, “R-VQA: Learning visual relation facts with semantic attention for visual question answering,” in *Proc. Int. Conf. Knowl. Discovery Data Min. (KDD)*, London, U.K., 2018, pp. 1880–1889.

[76] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li, “Learning visual knowledge memory networks for visual question answering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7736–7745.

[77] D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 6700–6709.

[78] A. Graves *et al.*, “Hybrid computing using a neural network with dynamic external memory,” *Nature*, vol. 538, pp. 471–476, Oct. 2016.

[79] A. Joulin and T. Mikolov, “Inferring algorithmic patterns with stack-augmented recurrent nets,” in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 190–198.

[80] J. Weston, S. Chopra, and A. Bordes, “Memory networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[81] C. Xiong, S. Merity, and R. Socher, “Dynamic memory networks for visual and textual question answering,” in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 48. New York, NY, USA, 2016, pp. 2397–2406.

[82] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Learning to compose neural networks for question answering,” in *Proc. Annu. Conf. North Amer. Ch. Assoc. Comput. Linguist. (NAACL)*, 2016, pp. 1545–1554.

[83] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 39–48.

[84] G. Gkioxari, R. Girshick, P. Dollár, and K. He, “Detecting and recognizing human-object interactions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 8359–8367.

[85] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448.

[86] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, “Learning to detect human-object interactions with knowledge,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2019–2028.

[87] S. Gupta and J. Malik, “Visual semantic role labeling,” 2015, *arXiv:1505.04474*.

[88] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, “Learning to detect human-object interactions,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, 2018, pp. 381–389.

[89] U.-H. Kim, J.-M. Park, T.-J. Song, and J.-H. Kim, “3-D scene graph: A sparse and semantic representation of physical environments for intelligent agents,” *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4921–4933, Dec. 2020, doi: 10.1109/TCYB.2019.2931042.

[90] R. Krishna *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[91] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes challenge: A retrospective,” *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[92] B. Thomee *et al.*, “YFCC100M: The new data in multimedia research,” *Commun. ACM*, vol. 59 no. 2, pp. 64–73, Feb. 2016.

[93] T. Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[94] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, 2017.

[95] H. Zhang, Z. Kyaw, J. Yu, and S. Chang, “PPR-FCN: Weakly supervised visual relation detection via parallel pairwise RFCN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4243–4251.

[96] H. Fang *et al.*, “From captions to visual concepts and back,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1473–1482.

[97] 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)*, Honolulu, HI, USA, 2017, pp. 3097–3106.

[98] A. Newell and J. Deng, “Pixels to graphs by associative embedding,” in *Proc. 30th Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 2171–2180.

[99] Y. Li, W. Ouyang, B. Zhou, J. Shi, Z. Cao, and X. Wang, “Factorizable net: An efficient subgraph-based framework for scene graph generation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 346–363.

[100] 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)*, Long Beach, CA, USA, 2019, pp. 6156–6164.

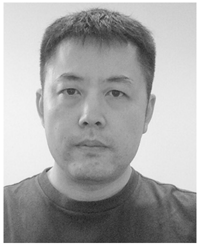
[101] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, “Unbiased scene graph generation from biased training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 3713–3722.

Jun Cheng (Member, IEEE) received the B.E. and M.E. degrees from the University of Science and Technology of China, Hefei, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2006.

He is currently with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, as a Professor, where he is also the Director of the Laboratory for Human Machine Control. His current research interests include computer vision, robotics, machine intelligence, and control.

Lei Wang (Member, IEEE) received the Ph.D. degree in electrical engineering from Xidian University, Xi’an, China, in 2010.

He worked with Huawei Technologies Co., Ltd., Shenzhen, China, from 2011 to 2012, and the University of Jinan, Jinan, China, from 2012 to 2016. From 2014 to 2015, he was with the Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea, as a Postdoctoral Fellow. He is currently an Associate Professor with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He has authored or coauthored around 40 papers in conferences and journals, including IEEE SIGNAL PROCESSING LETTERS, IEEE GEOSCIENCE REMOTE SENSING LETTERS, IEEE International Conference on Computer Vision, and IEEE/CVF Conference on Computer Vision and Pattern Recognition. His research interests include image processing, transforms, machine learning, computer vision, visual semantic understanding, video analysis, 3-D reconstruction, and robotics.



Jiaji Wu (Member, IEEE) received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1996, the M.S. degree from National Time Service Center, Chinese Academy of Sciences, Xi'an, in 2002, and the Ph.D. degree in electrical engineering from Xidian University in 2005.

He is currently a Professor with Xidian University. His current research interests include still image coding, hyperspectral/multispectral image processing, communication, big data, IoT, and high-performance computing.

Xiping Hu (Member, IEEE) received the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada, in 2015.

He is currently a Professor with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China. He has more than 120 papers published and presented in prestigious conferences and journals. He was also the Co-Founder and the CTO of Bravolol Ltd., Hong Kong, a leading language learning mobile application company with over 100 million users, and listed as the top two language education platform globally. His research areas consist of mobile cyberphysical systems, crowdsensing, and affective computing.

Gwanggil Jeon (Member, IEEE) received the B.S., M.S., and Ph.D. (*summa cum laude*) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, South Korea, in 2003, 2005, and 2008, respectively.

From 2008 to 2009, he was with the Department of Electronics and Computer Engineering, Hanyang University from 2009 to 2011, a Postdoctoral Fellow with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, and an Assistant Professor with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, from 2011 to 2012. He is currently a Professor with Incheon National University, Incheon, South Korea, and Xidian University, Xi'an, China. His research interests fall under the umbrella of image processing, particularly image compression, motion estimation, demosaicking, and image enhancement and also computational intelligence, such as fuzzy and rough sets theories.

Dacheng Tao (Fellow, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2002, the M.S. degree from the Chinese University of Hong Kong, Hong Kong, in 2004, and the Ph.D. degree from the University of London, London, U.K., in 2007.

He is currently the Inaugural Director of the JD Explore Academy, Beijing, China, and a Senior Vice President of JD.com. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences.

Dr. Tao received the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a Fellow of the Australian Academy of Science, AAAS, and ACM.

Mengchu Zhou (Fellow, IEEE) received the Ph.D. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He then joined the New Jersey Institute of Technology, Newark, NJ, USA, where he is currently a Distinguished Professor. He has over 900 publications, including 12 books, 600+ journal papers (500+ in IEEE transactions), 29 patents, and 29 book-chapters. His interests are in Petri nets, automation, Internet of Things, and big data.

Dr. Zhou is a Fellow of IFAC, AAAS, CAA, and NAI.