Adaptively Clustering-Driven Learning for Visual Relationship Detection

An-An Liu, Yanhui Wang, Ning Xu, Weizhi Nie, Jie Nie, and Yongdong Zhang

***Abstract*—Visual relationship detection aims to describe the interactions between pairs of objects, such as *person-ride-bike* and *bike-next to-car* triplets. In reality, it is often the case that there exist some groups of strongly correlated relationships, while others are weakly related. Intuitively, the common relationships can be roughly categorized into several types such as geometric (e.g., next to), action (e.g., ride), and so on. However, previous studies ignore the relatedness discovery among multiple relationships, which only lie on a unified space to leverage visual features or statistical dependencies into categories. To tackle this problem, we propose an adaptively clustering-driven network for visual relationship detection, which can implicitly divide the unified relationship space into several subspaces with specific characteristics. Particularly, we propose two novel modules to discover the common distribution space and latent relationship association, respectively, which map pairs of object features into translation subspaces to induce the discriminative relationship clustering. Then, a fused inference is designed to integrate the group-induced representations with the language prior to facilitate the predicate inference. Especially, we design the Frobenius-norm regularization to boost the clustering. To the best of our knowledge, the proposed method is the first supervised framework to realize *subject-predicate-object* relationship-aware clustering for visual relationship detection. Extensive experiments show that the proposed method can achieve competing performances against the state-of-the-art methods on the Visual Genome dataset. Additional ablation studies further validate its effectiveness.**

***Index Terms*—Adaptively clustering-driven learning, translation embedding unit, visual relationship detection.**

I. Introduction

THE task of visual relationship detection aims to detect and localize pairs of interactive objects in an image and infer the predicate or interaction between them [1]. As shown in Fig. 1, visual relationship can not only capture the spatial and semantic information of objects *man* and *laptop*, but also infer the pair-wise relationship *looking at*. Due to the structured description and the rich semantic space, visual relationship detection can contribute to diverse high-level vision tasks, such as complex-query image retrieval [2], [3], image captioning [4], visual reasoning [5], [6], image generation [7], and VQA [8], [9].

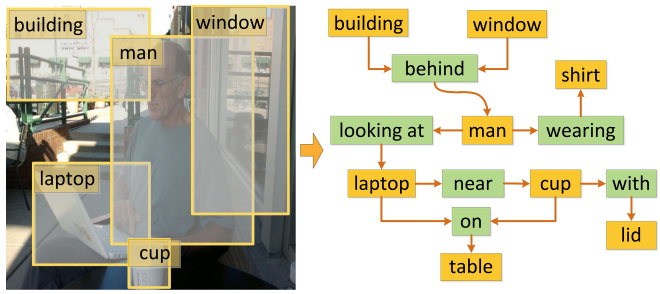

Fig. 1. Visual relationship detection aims to detect and localize objects (yellow nodes) and pair-wise relationships (green nodes) in the image.

Benefiting from the rapid development of deep learning, promising results of visual relationship detection have been achieved. For examples, Zhang *et al.* [10] proposed to formulate the predicate problem as a translation vector between object and subject, avoiding learning the diverse appearances of relationships. Li *et al.* [11] designed the visual phrase guided by the message passing structure to establish the connection among relationship components. Yang *et al.* [12] proposed an attentional graph convolutional network to propagate contextual information for relationship modeling.

A. Motivation and Overview

Although superior performances have been achieved on this task, the existing methods still have two key dilemmas.

- *Ignoring the group information among relationships.* In reality, it is often the case that there exist some groups of strongly correlated relationships, while others are weakly related. For examples, *riding/eating/carrying* can be considered as a highly correlated group for *human-object* interaction, while *on/near/under* can be considered as another group for *spatial orientation* modeling. Hence, the clustering-driven learning can strengthen the relatedness discovery and facilitate the predicate inference. However, previous studies cannot explicitly capture high-level

Manuscript received December 16, 2019; revised July 8, 2020; accepted November 25, 2020. Date of publication December 10, 2020; date of current version December 9, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1406602, in part by the National Natural Science Foundation of China under Grants 61772359, 61525206, 62002257, and 61702471, and in part by the Grant of Tianjin New Generation Artificial Intelligence Major Program (19ZXZNGX00110, 18ZXZNGX00150). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianguo Zhang. *(Corresponding authors: Ning Xu; Jie Nie.)*

An-An Liu, Yanhui Wang, Ning Xu, and Weizhi Nie are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300 072, China (e-mail: anan0422@gmail.com; wangyanhui@tju.edu.cn; ningxu@tju.edu.cn; weizhinie@tju.edu.cn).

Jie Nie is with the College of Information Science and Engineering, Ocean University of China, Qingdao 266 100, China (e-mail: niejie@ouc.edu.cn).

Yongdong Zhang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: zhyd73@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3043084.

Digital Object Identifier 10.1109/TMM.2020.3043084

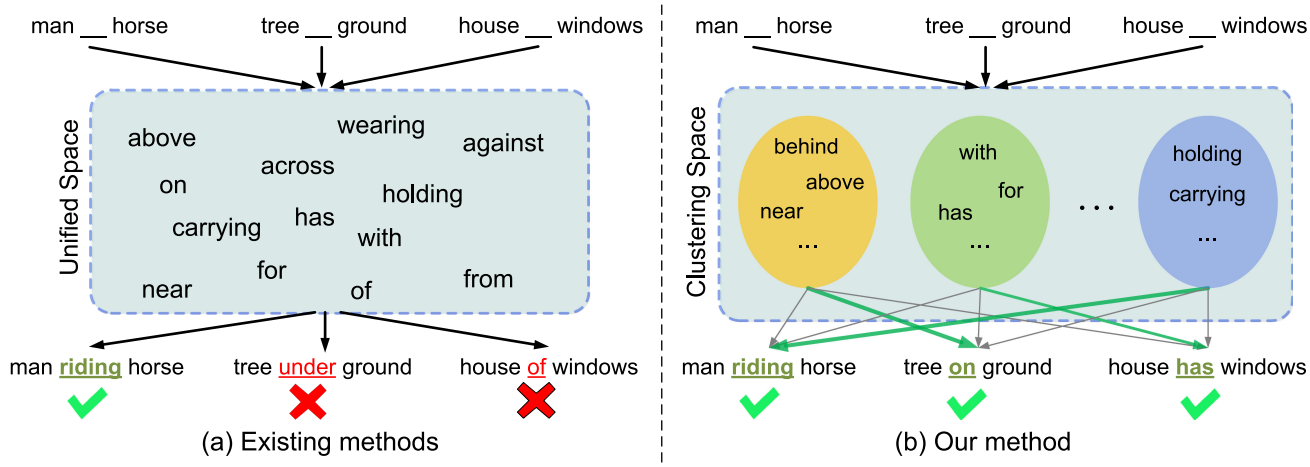

Fig. 2. Comparison between the existing methods and our method. Different from previous methods, which lie on a unified relationship space to directly infer predicates, we implicitly divide the relationship space into several subspaces by adaptively clustering. Given a pair of objects, each subspace will make a decision to infer predicate while the category with the highest score (green line) is selected as the final result.

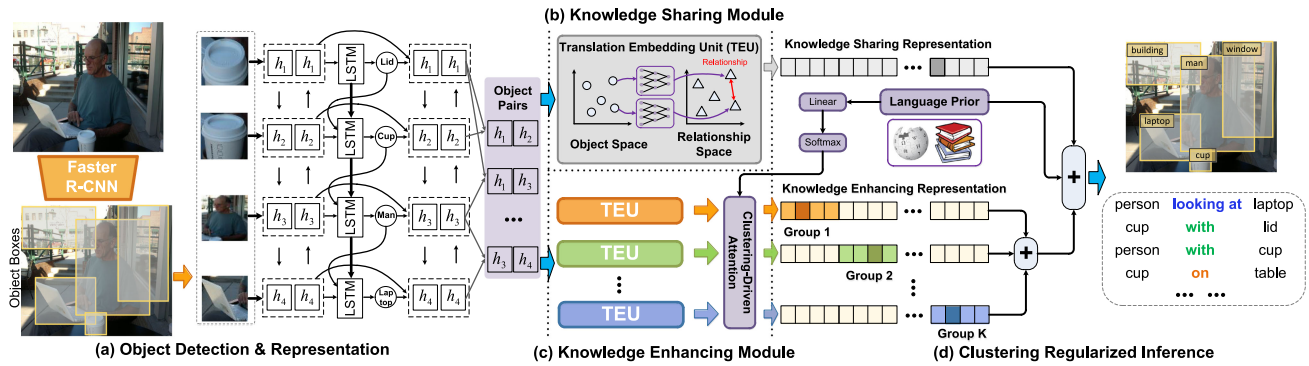

Fig. 3. Overview of the proposed model. (a) Object Detection & Representation aims to recognize objects and provide the object representations based on contextual information for the next step. (b) Knowledge Sharing Module can model the common distribution space of multiple relationships, where the Translation Embedding Unit (TEU) learns a shared low-dimensional space to bridge objects and predicates. (c) Knowledge Enhancing Module can discover the relationship association and induce the discriminative clustering via the multiple attention-guided translation subspaces. (d) Clustering Regularized Inference integrates the group-induced representations with the language prior to facilitate the predicate inference. Grey refers to the group-sharing learning process. Orange, Green, and Blue refer to different group-enhancing learning process. "+" denotes the element-wise addition.

clustering information, but rather lie on a unified space to leverage visual features or statistical dependencies into categories (Fig. 2(a)), which might have negative impact on the performances.

- *Difficulty in group information discovery.* How to design an adaptively clustering method is a fundamental challenge for relationship modeling. Different from object or action detection [13], which can leverage some universal clustering methods such as K-Means [14], GMMs [15] and Mean-Shift [16], visual relationship detection involves a pair of object detection connected via a predicate inference, which goes beyond individual instance modeling. Besides, the huge object & predicate semantic space augments the difficulty in relationship clustering.

To tackle these problems, we propose an adaptively clustering-driven network for visual relationship detection. As shown in Fig. 2(b), the proposed method aims to implicitly divide the relationship space into several subspaces with specific characteristics. Given a pair of objects, each subspace can make a decision to infer predicate while the category with the highest score can be selected as the final decision.

The pipeline is shown in Fig. 3. The proposed method contains four key modules:

1) *Object Detection & Representation:* Given an image, it aims to recognize objects and provide object representations based on contextual information. Specifically, with object proposals from Faster-RCNN [17], we incorporate global context through two layers of bidirectional LSTMs, while another LSTM is used to predict object labels.
2) *Knowledge Sharing Module:* It aims to extract the common distribution space from multiple relationships. Particularly, we utilize the Translation Embedding Unit (TEU) [10] to obtain relationship representations by mapping pairs of object features into a shared translation space.

3) *Knowledge Enhancing Module:* It aims to discover the relationship association and strengthen the discriminative clustering among highly-related ones. Particularly, we construct several separated translation subspaces to learn the relationship groups. Meanwhile, the clustering-driven attention equipped with language knowledge is proposed to regularize the clustering procedure.
4) *Clustering Regularized Inference:* This module can integrate the relationship representations from different translation spaces with language prior to facilitate the predicate inference. Especially, we design the Frobenius-norm regularization to boost the relationship clustering.

The proposed method is evaluated on the popular Visual Genome [18] dataset. Experiment results show that the proposed method can achieve competing performances against the state-of-the-art methods on three standard tasks: Predicate Classification, Phrase Classification, and Relationship Detection. Additional ablation studies further validate its effectiveness.

B. Contributions

The key contributions are listed as follows:

- We propose an adaptively clustering-driven network for visual relationship detection. To the best of our knowledge, the proposed method is the first supervised framework which explores *subject-predicate-object* relationship-aware clustering for this task.
- We propose two novel modules to discover the common distribution space and latent relationship association for discriminative clustering. Then, a fused inference is designed to integrate the group-induced representations with the language prior to facilitate the predicate inference.
- The proposed method is evaluated on the Visual Genome dataset. Experimental results show competing performances against the state-of-the-art methods. Additional ablation studies showcase the efficacy of the proposed modules.

II. Related Work

A. Visual Relationship Detection

Earlier works on visual relationship detection considered the relationship as the visual phrase [19], which combine objects and predicates as a distinct class for prediction. However, this method scale poorly since it is significantly dependent on sufficient training data. To alleviate it, researchers proposed to separate object and predicate detection into individual branches [1], [20]–[25]. For examples, Lu *et al.* [1] first detected subjects and objects, and then classified their predicates individually. Yu *et al.* [26] leveraged language knowledge from the training annotations and the public text corpus for predicate inference. But these methods ignore the visual context information, which can capture the fruitful semantic cues to make better predictions. Recently, more studies [23], [27]–[34] utilized visual context to provide a powerful inductive bias for the object and predicate detection. For examples, Dai *et al.* [35] leveraged the appearance, spatial configurations, as well as the statistical relations between objects and predicates for context modeling. Xu *et al.* [27] proposed the RNN-based message passing mechanism to incorporate context cues for scene graph generation. Li *et al.* [29] aligned object, phrase, and caption regions with a dynamic graph to pass messages in different semantic levels. Yang *et al.* [12] adopted the graph convolution network to iteratively refine features. Zellers *et al.* [23] employed LSTMs to encode global context for predicate inference.

Nevertheless, existing studies only lie on a unified semantic space to leverage visual context or statistical dependencies. Different from the unified space modeling, the propose method can implicitly divide the semantic space into several subspaces, which focuses on relationship relatedness discovery and realize *subject-predicate-object* relationship-aware clustering.

B. Clustering Algorithms

As one of the most widely fundamental tasks of data analysis, the clustering algorithm aims to classify each data point into the specific groups, where similar points are aggregated together while dissimilar ones should belong to different groups. It had been widely used in many computer vision tasks such as image classification [15], action recognition [14], [36] and object detection [37]. For examples, Perronnin *et al.* [15] proposed the improved fisher kernel to boost the clustering procedure for the large-scale image classification. Zhang *et al.* [38] designed a dropout k-means based method to extract the hierarchical spatial feature for the hyperspectral image classification. Simonyan *et al.* [39] formulated the image classification as a multi-layer fisher encoding problem. These approaches leverage diverse clustering strategies to enrich image representations, which can benefit the problem of large-scale image classification. Moreover, in the field of action recognition, Peng *et al.* [40] proposed to preserve mid-level information with the stacked fisher vectors. Murtaza *et al.* [41] extended standard VLAD encoder to efficiently reduce the inter-class similarity. Ubalde *et al.* [42] presented the Citation-kNN-based method to model skeleton sequences. Generally, clustering algorithms are always used to deal with the intra-class diversity and the inter-class similarity of categories for action recognition task.

Different from image or action classifications, the *subject-predicate-object* relationships go beyond individual instance modeling, which augments the clustering difficulty. In this paper, we equip the attention module with language knowledge to drive the complex relationship clustering.

III. Approach

A. Problem Formulation

Given an image $\mathbf{I}$, the goal of visual relationship detection is to detect and localize pairs of objects and classify the predicate in-between [1]. In this paper, we formulate the output of this task as follows:

- A bounding box set $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N\}$, where N is the number of detected objects; $\mathbf{b}_n \in \mathbb{R}^4$ denotes the bounding box of the n-th object.

- An object label set $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_N\}$, where $\mathbf{l}_n \in \mathcal{O}$ denotes the category of the n-th object; $\mathcal{O}$ is the set of object categories.
- A relationship label set $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M\}$, where M is the number of recognized relationships; $\mathbf{p}_m \in \mathcal{P}$ is the category of m-th relationship; $\mathcal{P}$ is the set of predicate categories. We add the extra 'background' category, which denotes no relationship between objects.

B. Object Detection & Representation

To fully explore contextual information, we utilize [23] for object representations (Fig. 3(a)). First, Faster R-CNN [17] is adopted to generate object proposals from the given image $\mathbf{I}$. For each proposal, we can obtain a bounding box $\mathbf{b}_i$, a vector of object label probabilities $\mathbf{s}_i$, and a feature vector $\mathbf{v}_i$. Similar to [23], we sort the object proposals left-to-right by the central x-coordinate. Then, Bidirectional LSTM [43] is used to sequentially encode all object proposals to obtain the object context representations $\mathbf{G}$:

$$\begin{aligned} \mathbf{x}_i^1 &= [\mathbf{v}_i; \mathbf{W}_1 \mathbf{s}_i] \\ \mathbf{g}_i &= \mathrm{biLSTM}_1(\mathbf{x}_i^1) \end{aligned} \tag{1}$$

where $\mathbf{x}_i^1$ is the i-th input vector to the bidirectional LSTM; $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^N$ is the set of hidden states of LSTM and $\mathbf{g}_i$ corresponds to the i-th input object; $\mathbf{W}_1$ is the learned parameter; $[;]$ denotes the concatenation operation. Then, we use another LSTM to predict the object label $\mathbf{l}_i$ (one-hot) depending on $\mathbf{g}_i$ and the previously detected label $\mathbf{l}_{i-1}$:

$$\begin{aligned} \mathbf{x}_i^2 &= [\mathbf{g}_i; \mathbf{l}_{i-1}] \\ \mathbf{h}_i &= \mathrm{LSTM}(\mathbf{x}_i^2) \\ \mathbf{l}_i &= \mathrm{argmax}\left(\mathbf{W}_2 \mathbf{h}_i\right) \in \mathbb{R}^{[\mathcal{O}]} \end{aligned} \tag{2}$$

where $\mathbf{x}_i^2$ is the i-th input vector to LSTM; $\mathbf{h}_i$ is the hidden state of LSTM; $\mathbf{W}_2$ is the learned parameter. Since the label embedding can boost the relationship inference [1], we further combine the detected object label and the context representation by another bidirectional LSTM:

$$\begin{aligned} \mathbf{x}_i^3 &= [\mathbf{g}_i; \mathbf{W}_3 \mathbf{l}_i] \\ \mathbf{o}_i &= \mathrm{biLSTM}_2(\mathbf{x}_i^3) \end{aligned} \tag{3}$$

where $\mathbf{x}_i^3$ is the i-th input vector to the bidirectional LSTM; $\mathbf{O} = \{\mathbf{o}_i\}_{i=1}^N$ is the set of hidden states of LSTM and $\mathbf{o}_i$ corresponds to the i-th input object; $\mathbf{W}_3$ is the learned parameter. Finally, we use the combined features $\mathbf{O}$ as object representations for the next step.

C. Knowledge Sharing Module

The proposed *Knowledge Sharing Module* aims to build the common distribution space of all relationships. The distribution space formulation has been developed for many years in the field of Multi-Task Learning [44] for action or object recognitions [13], [45]. However, different from the above tasks, visual relationship detection aims to jointly realize the object and predicate inferences, which face two key challenges: 1) the relationship inference consists of both objects and predicates, which leads to a long-tail distribution problem. For N objects and M predicates, there are very few, even zero, training examples for the vast majority of *subject-predicate-object* relationships $O(N^2 M)$ [19], [46]; 2) the complicated interactions between pair-wise objects make this task extremely challenging. For examples, the same object pair with a *person* and a *bicycle* might involve different predicates such as the person *riding*, *pushing*, or even *falling off* of the bicycle. Besides, the same predicate *ride* can be determined by different object pairs such as *person-ride-bike* and *person-ride-horse*. Hence, it is significantly more difficult to extract the shared information from relationships than from objects.

Motivated by the advantages of representing large-scale knowledge bases [10], [47], Translation Embedding Unit (TEU) is used to model visual relationships by mapping the object and predicate features in a shared low-dimensional space. TEU offers a linear model to represent the long-tail relationships, where one relationship can be modeled as a vector translation, i.e., subject + predicate $\approx$ object. Moreover, it can avoid difficulty in learning diverse visual patterns of relationships.

Particularly, we denote the features of *subject* and *object* by $\mathbf{o}_i, \mathbf{o}_j \in \mathbb{R}^V$. TEU trains two embedding matrices $\mathbf{W}_{es}, \mathbf{W}_{eo} \in \mathbb{R}^{u \times V}$ from the object space into the predicate space, which can obtain the relationship translation vector $\mathbf{r}_p \in \mathbb{R}^u$:

$$\mathbf{W}_{es}\mathbf{o}_i + \mathbf{r}_p \approx \mathbf{W}_{eo}\mathbf{o}_j \tag{4}$$

In this section, TEU is used to extract the shared information among relationships. To maintain the spatial patterns, we multiply it by the union box feature $\mathbf{v}_{i,j}$ and obtain the relationship-sharing representation $\mathbf{E}_{i,j}^s$.

$$\mathbf{E}_{i,j}^s = \left(\mathbf{W}_{es}\mathbf{o}_i - \mathbf{W}_{eo}\mathbf{o}_j\right) \circ \mathbf{v}_{i,j} \tag{5}$$

Particularly, $\mathbf{v}_{i,j}$ is the RoIAlign feature [48] from the union bounding box of *subject* and *object*.

D. Knowledge Enhancing Module

The proposed *Knowledge Enhancing Module* can strengthen the highly-correlated relationship modeling and induce the discriminative group generation. Till now, how to realize the relationship-enhancing clustering is a fundamental challenge: 1) Few methods have been done to explore the correlation for the *subject-predicate-object* relationships; 2) Language priors capture a wide variety of statistical dependencies between objects and their relationships [1], [23], while there is no method to integrate them to tackle relationship clustering. Hence, in this section, we propose the relationship-enhancing clustering module, to jointly realize the relationship relatedness discovery and the language knowledge transfer. It can be decomposed into four consecutive steps:

- We construct the language prior function $\mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)$ to provide the empirical distribution over relationships between object labels $\mathbf{l}_i$ and $\mathbf{l}_j$ as [23].
- K separated TEUs are jointly trained (as stated in Section III-C), to simultaneously learn individual groups for

relationship clustering. K is the number of groups:

$$\mathbf{e}_{i,j}^{k} = \left(\mathbf{W}_{es}^{k}\mathbf{o}_i - \mathbf{W}_{eo}^{k}\mathbf{o}_j\right) \circ \mathbf{v}_{i,j} (k = 1, 2, \ldots, K) \quad (6)$$

where $\mathbf{e}_{i,j}^{k}$ is the preliminary relationship-enhancing representation in the k-th group.

- To generate the discriminative groups, we leverage the language knowledge to generate the clustering weights to regularize each TEU learning. Especially, by mapping $\mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)$ with learned transformation matrices, we obtain the attentive scores as follows:

$$\boldsymbol{\alpha}_{i,j}^{k} = softmax(\mathbf{W}_{\alpha}^{k}\mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)) \quad (7)$$

where $\mathbf{W}_{\alpha}^{k}$ is the transformation matric for the k-th group; $\boldsymbol{\alpha}_{i,j}^{k}$ denotes the attentive scores of the k-th group between objects $\mathbf{l}_i$ and $\mathbf{l}_j$.

- Then we use the attentive scores as the clustering weight to regularize each TEU learning as follows:

$$\mathbf{E}_{i,j}^{p} = \sum_{k} \boldsymbol{\alpha}_{i,j}^{k} \circ \mathbf{W}_{b}^{k}\mathbf{e}_{i,j}^{k} \quad (8)$$

where $\mathbf{W}_{b}^{k}$ is the learned matric for the k-th group; $\mathbf{E}_{i,j}^{p}$ denotes the regularized relationship-enhancing representation.

E. Clustering Regularized Inference

After aforementioned steps, we get the relationship-sharing and relationship-enhancing representations, $\mathbf{E}_{i,j}^{s}$ and $\mathbf{E}_{i,j}^{p}$. As shown in Fig. 3 (d), the *Clustering Regularized Inference* can leverage both to infer the predicates. Besides, we use the language knowledge to further assist the group modeling. $\mathbf{E}_{i,j}^{s}$ and $\mathbf{E}_{i,j}^{p}$ are fused with the language prior as follows:

$$\begin{aligned}\Pr\left(\mathbf{d}_{i\to j}|\mathbf{B}, \mathbf{L}\right) = \ & \mathrm{softmax}\left(\lambda\mathbf{W}_{r}^{s}\mathbf{E}_{i,j}^{s}\right. \\ & \left. + \beta\mathbf{W}_{r}^{p}\mathbf{E}_{i,j}^{p} + \mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)\right) \end{aligned} \quad (9)$$

where λ and β are the trade-off parameters; $\mathbf{W}_{r}^{s}$ and $\mathbf{W}_{r}^{p}$ are learned parameters. The softmax function normalizes the fused clustering feature to be a distribution over the predicate categories. To enhance the robustness of the model, the set of regularization terms are designed as:

$$\mathrm{Reg}(\cdot) = \sum_{k} \rho_k \|[\mathbf{W}_{es}^{k}; \mathbf{W}_{eo}^{k}]\|_{F}^{2} (k = 1, 2, \ldots, K) \quad (10)$$

where the regularization parameter, ρ_k, controls the importance of the penalty term; $\|\cdot\|_F$ denotes the ℓ_2-norm (Frobenius norm) of the matrix; $[;]$ denotes the concatenation operation.

IV. Experiment

A. Dataset

The proposed method was evaluated on the popular dataset, Visual Genome (VG) [18]. We adopted the training and test splits as [27], with the most frequent 150 object categories and 50 predicate classes. After preprocessing, the number of objects and relationships on each image is 11.5 and 6.2 on average, respectively. We used 70% of images for training (including 5 K images for validation) and 30% of images for testing.

B. Evaluation Settings

Following existing literatures [1], [23], [27], [29], we report the performances of the proposed method on three sub-tasks of visual relationship detection:

- **Predicate Classification** (PredCls): It aims to predict a set of possible predicates between the objects given the ground truth boxes and categories of objects.
- **Phrase Classification** (PhrCls): It aims to predict both the object and pairwise relationship categories, given the ground-truth object bounding boxes.
- **Relationship Detection** (RelDet): It aims to detect objects and recognize their pairwise relationships. The object is correctly detected if it is correctly classified and its overlap with the ground truth box is at least 0.5. A relationship is correctly detected if both the subject and object are correctly detected and the predicate is correctly predicted.

Following [1], we used the recall@K (short as R@K) metric as the evaluation metric. R@K computes the fraction of the true relationships that are predicted in the top K confident relationship predictions in an image.

Besides, several works [1], [12], [29] evaluate models with the *constraint* that merely one relationship is predicted for a given object pair. In contrast, other works [23], [49], [50] take the multiple relationships into account, referring to *no constraint*. For the fair comparison, we report the R@K scores with and without constraint on three tasks for evaluation.

C. Implementation Details

Faster R-CNN with a VGG backbone [17] was first pretrained on Visual Genome objects. During the training phase, we integrated the Faster RCNN by freezing the parameters of convolution layers before the ROIAlign, and fine-tuning the fully connected layers for object representation. The other components in our model were randomly initialized. The batch size and the number of groups were set to 6 and 3, respectively. SGD was used to compute the gradient with the initial learning rate 0.001 and the momentum 0.9. The entire model was optimized with the sum of object and relationship XE (Cross-Entropy) losses. Particularly, for RelDet, since the number of all possible relationships were huge, we only considered the relationship between two objects with overlapped bounding boxes as [23]. During the testing phase, we selected the categories with highest scores as the results for objects and predicates. Our work was implemented using PyTorch.

D. Comparison With the State-of-The-Art Methods

We compared the proposed method to two types of representative methods.

1) Separate Inference Model: It first detects subject and object, and then classifies their predicate individually. For examples, **VRD**, **FREQ** and **MOTIFS** used detectors to localize objects and then leveraged language knowledge for predicate classification. **AsscEmbed** extracted feature vectors from the pixel locations and then fed through individual fully connected networks to predict objects and predicates.

TABLE I
PERFORMANCES OF R@K(%) COMPARED WITH THE STATE-OF-THE-ART METHODS WITH AND WITHOUT CONSTRAINT ON VG. THE MEAN IS COMPUTED BY AVERAGING PERFORMANCES ON THE THREE EVALUATION TASKS OVER R@50 AND R@100

	Model	RelDet			PhrCls			PredCls			Mean
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	
Constraint	VRD [1]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0	14.9
	IMP [27]	-	3.4	4.2	-	21.7	24.4	-	44.8	53.0	25.3
	MSDN [29]	-	7.0	9.1	-	27.6	29.9	-	53.2	57.9	30.8
	AsscEmbed [49]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	57.1	55.4	28.3
	Graph-RCNN [12]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1	33.2
	FREQ+ [23]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
	IMP+ [23], [27]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
	MOTIFS [23]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1	43.7
	KERN [50]	-	27.1	29.8	-	36.7	37.4	-	65.8	67.6	44.1
	Ours	**21.7**	**27.5**	**30.7**	**33.7**	**36.7**	**37.6**	**59.4**	**66.2**	**68.3**	**44.5**
No constraint	AsscEmbed [49]	-	9.7	11.3	-	26.5	30.0	-	68.0	75.2	36.8
	IMP+ [23], [27]	-	22.0	27.4	-	43.4	47.2	-	75.2	83.6	49.8
	FREQ+ [23]	-	28.6	34.4	-	39.0	43.4	-	75.7	82.9	50.6
	MOTIFS [23]	22.8	30.5	35.8	37.6	44.5	47.7	66.6	81.1	88.3	54.7
	KERN [50]	-	30.9	35.8	-	45.9	49.0	-	81.9	88.9	55.4
	Ours	**23.1**	**31.0**	**36.4**	**41.6**	**48.9**	**52.4**	**71.1**	**84.4**	**90.9**	**57.3**

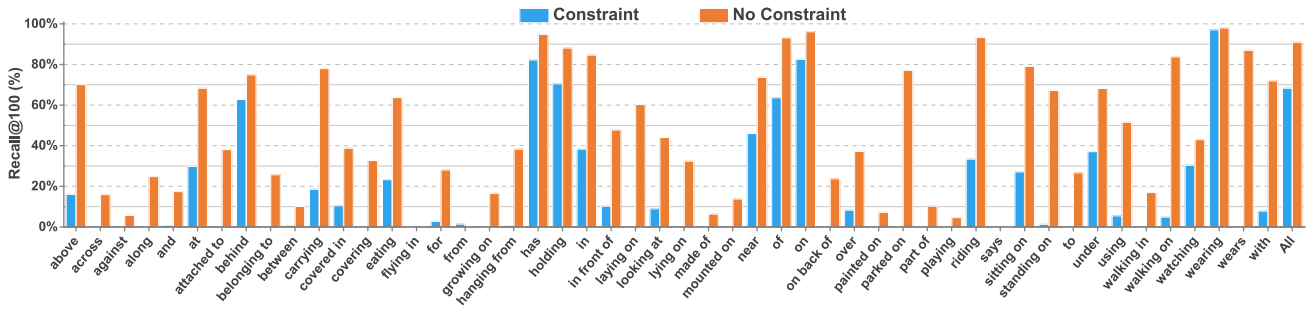

Fig. 4. Category-wise performance comparison w/o constraint. The horizontal axis indicates the relationship indexes and the vertical axis shows the R@100 scores.

2) Joint Learning Model: It can simultaneously detect the objects and their predicates relying on the message passing mechanism. For examples, **MSDN** built a propagating graph with different semantic-level features to benefit the relationship inference. **IMP** leveraged the contextual information to iteratively improve its performances via message passing. **Graph-RCNN** and **KERN** utilized the Graph Convolutional Network to formulate the scene graph.

As shown in Table I, the proposed method can achieve competing performances against the state-of-the-art methods on Visual Genome. We have three key observations:

- The proposed model can achieve competing performances against both separate inference models and joint learning models. Especially, our model can outperform all of the methods on the most challenging RelDet task, which requires the model to jointly detect objects and their relationship with boxes. The improved performances demonstrate that the proposed clustering discovery method can benefit the relationship detection.
- Our model performs better than the strong baseline MOTIFS [23], which use the similar object detection and representation method. It illustrates that the proposed clustering mechanism can provide discriminative cues for relationship modeling.
- Our method uses Translation Embedding Unit (TEU) which only models isolated objects and predicates. Comparatively, KERN employs the advanced Graph Convolutional Network (GCN) to discover the contextual information among them. Nevertheless, our method still outperforms KERN across all metrics, which further shows the advantage of the proposed clustering mechanism.

As shown in Fig. 4, we further provide the performances of each relationship category on the task of PredCls. We observe that several relationships can achieve significantly higher performances without the constraint than with it, such as "above," "at," "carrying," and "lying on". Particularly, for the relationship "above," its performance without constraint (0.70) is 4 times better than with constraint (0.16). Interestingly, we find these relationships have two common characteristics: 1) They always follow the low-frequency distribution in Visual Genome dataset, e.g., "carrying" and "lying on" appear less than 7000 times. Such a small instance space augments the difficulty in relationship

TABLE II
PERFORMANCES OF R@K(%) COMPARISON WITH DIFFERENT VARIANTS FOR THE PROPOSED METHOD

	Model	RelDet			PhrCls			PredCls		
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Constraint	Non-Clustering	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
	Non-Sharing	21.3	27.0	30.0	32.5	35.2	35.9	58.7	65.3	67.1
	Non-Enhancing	21.4	27.2	30.2	32.8	35.6	36.3	58.3	64.9	66.6
	Pre-defined Clustering	21.5	27.2	30.3	32.8	35.6	36.2	58.9	65.5	67.3
	Prior clustering weight	21.4	27.1	30.4	33.5	36.4	37.1	59.1	65.7	67.4
	Visual clustering weight	21.4	27.1	30.3	32.9	35.8	36.5	58.5	65.2	67.1
	Ours	**21.7**	**27.5**	**30.7**	**33.7**	**36.7**	**37.6**	**59.4**	**66.2**	**68.3**
No Constraint	Non-Clustering	-	28.6	34.4	-	39.0	43.4	-	75.7	82.9
	Non-Sharing	22.6	30.4	35.5	37.1	44.0	47.1	66.7	81.3	88.5
	Non-Enhancing	22.7	30.5	35.6	37.8	44.9	48.1	66.2	80.9	88.1
	Pre-defined Clustering	23.0	30.8	35.8	38.3	45.3	48.5	67.0	81.5	88.7
	Prior clustering weight	22.8	30.5	35.8	38.2	45.2	48.4	66.8	81.4	88.7
	Visual clustering weight	22.7	30.5	35.7	37.5	44.6	47.9	66.6	81.1	88.4
	Ours	**23.1**	**31.0**	**36.4**	**41.6**	**48.9**	**52.4**	**71.1**	**84.4**	**90.9**

learning. 2) They are semantically similar with the relationships of high-frequency distribution, such as "on". In detail, "on" and "above" refer to the similar geometric pattern, while the former contains 727 959 instances, nearly 30 times more than the latter. Due to the semantic similarity and imbalanced instances, under the constraint scenario, our model always predicts "on" and ignore "above". However, we can observe that the performances of "above" improve significantly under the no constraint scenario, where the more inferred results are presented. It indicates that our clustering mechanism can make the semantically similar relationships closer to each other, and not affected by the imbalanced distribution.

E. Ablative Studies

In this section, we further validate the efficacy of the proposed method by answering the following questions. **Q1**: Is the idea of clustering-driven learning effective for visual relationship detection? **Q2**: What are the effects of *Knowledge Sharing Module* and *Knowledge Enhancing Module* in the clustering domain? **Q3**: Is it necessary to use the adaptive manner to realize the clustering in the *Knowledge Enhancing Module*? **Q4**: Is there another choice to compute the clustering weights in the *Knowledge Enhancing Module*? **Q5**: What are the effects of the number of groups on the performances?

Effectiveness of Relationship Clustering (Q1). Our model is compared with its variants by removing *Knowledge Sharing Module* and *Knowledge Enhancing Module*. Specifically, we directly remove relationship-sharing and relationship-enhancing representations from 9 as follow.

$$\Pr\left(\mathbf{d}_{i\to j}|\mathbf{B},\mathbf{L}\right)=\mathrm{softmax}\left(\mathbf{w}(\mathbf{l}_i,\mathbf{l}_j)\right) \tag{11}$$

We denote it by Non-Clustering, which falls into FREQ+ [23], the empirical distribution over relationships between objects for inference. As shown in Table II, our model performs better than Non-Clustering, which ignores the relatedness discovery among relationships. It illustrates the proposed clustering method can provide discriminative representation for relationship modeling.

Effectiveness of Relationship-Sharing and Relationship-Enhancing Discovery (Q2). *Knowledge Sharing Module* and *Knowledge Enhancing Module* are two parallel processes. The former can extract the common sharing space while the latter aims to strengthen group-wise subspaces. Both of them can discover different scales of relatedness information. To validate their effectiveness, we ablated the proposed method with two variants:

1) Non-Sharing: We only remove *Knowledge Sharing Module* and 9 is modified by

$$\Pr\left(\mathbf{d}_{i\to j}|\mathbf{B},\mathbf{L}\right)=\mathrm{softmax}\left(\mathbf{W}_r^p\boldsymbol{E}_{i,j}^p+\mathbf{w}(\mathbf{l}_i,\mathbf{l}_j)\right) \tag{12}$$

2) Non-Enhancing: We only remove *Knowledge Enhancing Module* and 9 is modified by

$$\Pr\left(\mathbf{d}_{i\to j}|\mathbf{B},\mathbf{L}\right)=\mathrm{softmax}\left(\mathbf{W}_r^s\boldsymbol{E}_{i,j}^s+\mathbf{w}(\mathbf{l}_i,\mathbf{l}_j)\right) \tag{13}$$

As shown in Table II, we have four observations:

- Both of Non-Sharing and Non-Enhancing performs better than Non-Clustering, which shows the effectiveness of common sharing space modeling and discriminative clustering learning, respectively.
- Our model can consistently outperform these two variants. It indicates that *Knowledge Sharing Module* and *Knowledge Enhancing Module* are complementary to each other in the full model.
- Non-Sharing outperforms Non-Enhancing on PredCls with and without constraints. It confirms that given ground-truth object labels, clustering modeling can strengthen the relatedness discovery and thus provide more discriminative cues for predicate inference.
- Non-Enhancing performs better than Non-Sharing on PhrCls, which shows that learning common distribution space can benefit the joint inference in the *subject-predicate-object* format on PhrCls.

Effectiveness of the Adaptive Manner (Q3). *Knowledge Enhancing Module* employs the adaptive manner to cluster diverse relationship categories. Intuitively, is the adaptive manner

TABLE III
RELATIONSHIPS OF THE PRE-DEFINED GROUPS

Group	Relationship			
Action	*carrying holding riding walking in flying in*	*covered in laying on says walking on playing*	*covering looking at sitting on watching using*	*eating lying on standing on wears*
Geometric	*above at hanging from on to growing on*	*across attached to in on back of under*	*against behind in front of over and parked on*	*along between mounted on painted on near*
Possessive	*belonging to of made of*	*for part of*	*from wearing*	*has with*

the best choice for relationship clustering? In this section, we try the fixed manner based on the pre-defined groups. Motivated by [23], where relationships are manually divided into the high-level groups, we pre-defined all relationships into three groups, i.e., *action*, *geometric*, and *possessive*, which is detailed in Table III. Accordingly, we modify the output size of each TEU in *Knowledge Enhancing Module* to match the category number of each group. The outputs of TEUs are denoted by $\boldsymbol{e}^1_{i,j}$, $\boldsymbol{e}^2_{i,j}$, and $\boldsymbol{e}^3_{i,j}$. 8 is replaced by:

$$\boldsymbol{E}^p_{i,j} = [\mathbf{W}^1_b \boldsymbol{e}^1_{i,j}; \mathbf{W}^2_b \boldsymbol{e}^2_{i,j}; \mathbf{W}^3_b \boldsymbol{e}^3_{i,j}] \tag{14}$$

where $\mathbf{W}^1_b$, $\mathbf{W}^2_b$ and $\mathbf{W}^3_b$ are learned transformation matrices of three modified TEUs; $[;]$ denotes the concatenation operation. This variant is named as Pre-defined Clustering. As shown in Tab. II, our method can consistently outperform Pre-defined Clustering under all metrics, which indicates that using adaptive manner can benefit discovering the latent relationship association.

Investigation on Clustering Weight Generation (Q4). The proposed method adopts the language prior to induce the clustering weight generation in the *Knowledge Enhancing Module*. In this section, we designed and evaluated other strategies to compute the clustering weights as follows:

- **Prior clustering weight.** To confirm the effect of proposed clustering-driven attention mechanism, we remove the attentive pipeline and directly use language prior to regularize the relationship clustering, which is denoted by **Prior clustering weight**. Particularly, $\boldsymbol{\alpha}^k_{i,j}$ in 8 is replaced by language prior $\mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)$. As shown in Table II, our method performs better than this variant, which validates that clustering-driven attention mechanism can improve the relatedness discovery in language prior.
- **Visual clustering weight.** To investigate the effect of visual information on relationship clustering, we replace language prior $\mathbf{w}(\mathbf{l}_i, \mathbf{l}_j)$ in 7 by the union box feature $\boldsymbol{v}_{i,j}$ to induce the clustering weight generation. As shown in Table II, the better performances achieved by our model demonstrate the language prior can provide the more reliable guidance than the visual feature on the relationship clustering discovery.

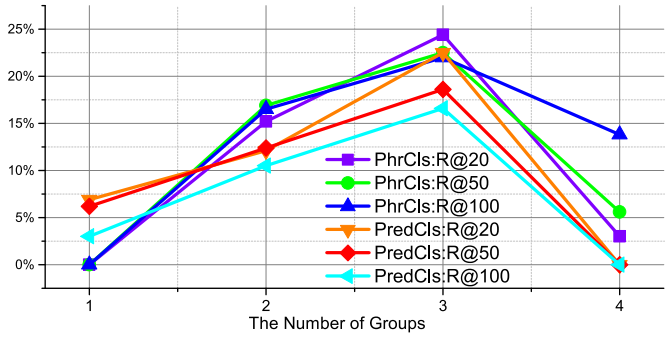

Fig. 5. The effect of the number of groups on the proposed method.

Investigation on the Number of Groups (Q5). In order to show the performances with respect to the number of groups, we tune the number of groups from 1 to 4 for validation. Fig. 5 reports the performances with respect to the different group number K. For the fair comparison, all of evaluation scores are normalized as follows:

$$\tilde{R}_k = \frac{R_k - \min\limits_k\{R_k\}}{\min\limits_k\{R_k\}} \tag{15}$$

where R_k and $\tilde{R}_k$ denote original and normalized performance values. From Fig. 5, we can see that increasing the number of group can generally lead to performance improvement, when K ranges from 1 to 3. It indicates that more distinguished clustering can improve the model capacity. However, when K ranges from 3 to 4, the performances will drop since the relationships are over-divided, which can induce false group information.

In our experiments, the number of group K is empirically set to 3.

F. Qualitative Results

Fig. 6 shows six examples with the corresponding relationships generated by our model. We can observe that the generated relationships are highly correlated with the visual content of the image. These results can qualitatively show that the proposed method can effectively infer objects and predicates by relationship clustering. We can obtain three key points: 1) Our model can predict many seemingly correct edges which do not exist in the ground truth. For examples, our model can make correct directional predictions, i.e., *cat-1-has-head-1* and *head-1-of-cat-1* in Fig. 6(b). But they are treated as false positives, which indicates that the incomplete annotations of Visual Genome may disturb the performances of relationship detection. 2) The failure case occurs when the underlying detector fails, resulting in cascading failure to predict any edges to that object (the blue edges connecting orange boxes). For examples, the failure to predict *lamp*, *desk*, *chair*, and *window* in Fig. 6(f) resulted in three false negative relationships in total. It shows that training more robust object detectors can benefit this task. 3) It is difficult to distinguish between the semantically similar relationships, such as *of* and *on*. Particularly, *of* can be used to indicate the possessive pattern, e.g., *nose-of-man*, while *on*

Fig. 6. Qualitative results from our model on the RelDet task. Detected boxes overlapped with ground truth are shown in green boxes. Orange boxes denote the ground truth with no match. Edges are predicted by our model at R@20 setting. Green edges are true positives, false negatives and false positives are marked by blue and red edges.

always refers to the geometric pattern, e.g., *cup-on-table*. However, it seems correct to replace *of* by *on* at times, since both of *nose-of-man* and *nose-on-man* are acceptable for human. In Fig. 6(g), the ground truth is *nose-1-on-man-1* but our model predicts *nose-1-of-man-1* that is judged to be a false positive. Hence, these similar relationships may disturb the process of adaptive clustering, which brings more challenges for relationship modeling.

To give an insight to the discriminative clustering, we further visualized the semantic affinities between the predicate parameter vectors in Fig. 7. The t-SNE method [51] is used to visualize 50 predicate model parameters of MOTIFS (Fig. 7(a)) and the proposed method (Fig. 7(b)). Especially, we color-encode the relationship points corresponding to the learned clustering weights from *Knowledge Specific Module*. As shown in Fig. 7(a), since MOTIFS ignores the relatedness discovery among relationships, it cannot provide the explicit semantic groups, where relationship points almost follow the uniform distribution. In contrast, our model can effectively explore the relatedness information and present an obviously clustering tendency in Fig. 7(b). For examples, most of the human-centered relationships such as *looking at*, *holding*, and *carrying*, are closer for each other (purple points). Meanwhile, several relationships that do not completely correspond to human activities such as *behind*, *under*, and *of*, will keep away from the human-centered ones (green points). In addition, we notice that there exist several points with similar semantics, such as *above* and *on*, are also linked into the same group (yellow points). It illustrates that our

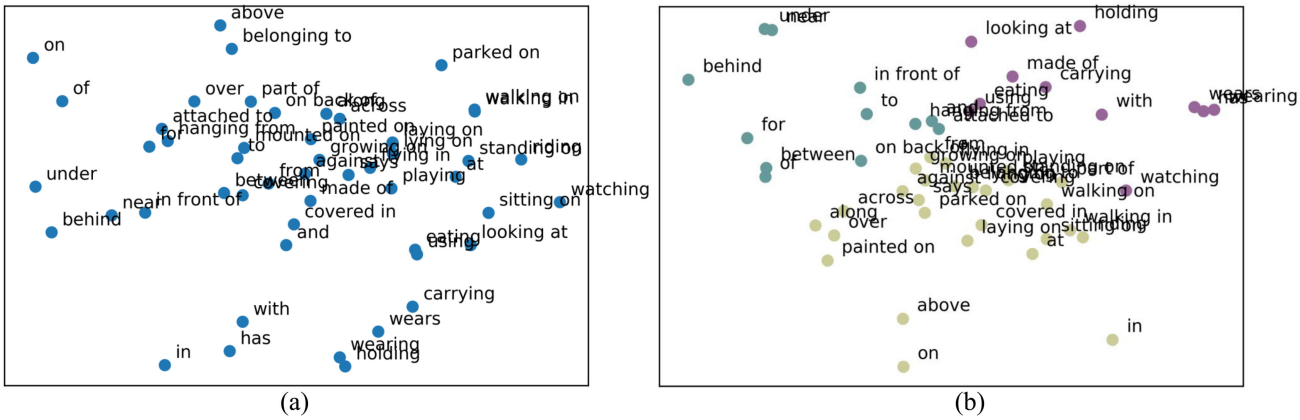

Fig. 7. t-SNE visualizations [51] of 50 predicate model parameters for MOTIFS (a) and our model (b) from VG. Specifically, relationship points of our model are color-encoded by learned clustering weights.

model is more likely to understand the meaning of orientation information.

V. Conclusion

In this paper, we propose an adaptively clustering-driven network for visual relationship detection, which is the first supervised framework to realize *subject-predicate-object* relationship-aware clustering. Particularly, we propose two novel modules to both discover the common distribution space and latent relationship association for discriminative clustering. Then, the clustering regularized inference is designed to integrate group-induced representations with the language prior to facilitate the predicate inference. Especially, we design the Frobenius-norm regularization to boost the clustering. Extensive comparative and ablative experiments on Visual Genome demonstrate the effectiveness of the proposed method.

References

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

[2] J. Johnson *et al.*, "Image retrieval using scene graphs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3668–3678.

[3] C. Liu *et al.*, "Graph structured network for image-text matching," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 918–10 927.

[4] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 711–727.

[5] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8368–8376.

[6] M. Yatskar, L. S. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5534–5542.

[7] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1219–1228.

[8] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8344–8353.

[9] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3233–3241.

[10] H. Zhang, Z. Kyaw, S. Chang, and T. Chua, "Visual translation embedding network for visual relation detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3107–3115.

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

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

[13] A.-A. Liu, Y. Su, W. Nie, and M. S. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[14] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *CVIU*, vol. 150, pp. 109–125, 2016.

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

[16] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: A texture classification example," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 456–463.

[17] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[18] 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.

[19] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1745–1752.

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

[21] L. Zhu *et al.*, "Discrete multimodal hashing with canonical views for robust mobile landmark search," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2066–2079, Sep. 2017.

[22] S. J. Hwang *et al.*, "Tensorize, factorize and regularize: Robust visual relationship learning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1014–1023.

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

[24] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global-local attention for image description," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[25] N. Zhao, H. Zhang, R. Hong, M. Wang, and T. Chua, "Videowhisper: Toward discriminative unsupervised video feature learning with attention-based recurrent neural networks," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2080–2092, Sep. 2017.

[26] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1068–1076.

[27] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3097–3106.
[28] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1271–1278.
[29] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1270–1279.
[30] C. Cui *et al.*, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1209–1220, May. 2019.
[31] Z. Cheng, X. Chang, L. Zhu, R. Catherine Kanjirathinkal, and M. S. Kankanhalli, "MMALFM: Explainable recommendation by leveraging reviews and images," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 16:1–16:28, 2019.
[32] C. Yan *et al.*, "A fast uyghur text detector for complex background images," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3389–3398, Dec. 2018.
[33] S. Li, W. Liu, and H. Ma, "Attentive spatial-temporal summary networks for feature learning in irregular gait recognition," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2361–2375, Sep. 2019.
[34] N. Xu *et al.*, "Scene graph inference via multi-scale context modeling," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.
[35] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3298–3308.
[36] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous domain adaptation through progressive alignment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1381–1391, May. 2019.
[37] A. Holub, M. Welling, and P. Perona, "Combining generative models and fisher kernels for object recognition," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 136–143.
[38] F. Zhang, B. Du, L. Zhang, and L. Zhang, "Hierarchical feature learning with dropout k-means for hyperspectral image classification," *Neurocomputing*, vol. 187, pp. 75–82, 2016.
[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep fisher networks for large-scale image classification," in *Proc. Conf. Neural Inf. Process. Syst.*, 2013, pp. 163–171.
[40] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 581–595.
[41] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "DA-VLAD: Discriminative action vector of locally aggregated descriptors for action recognition," in *Proc. Int. Conf. Image Process.*, 2018, pp. 3993–3997.
[42] S. Ubalde, F. G. Fernández, N. A. Goussies, and M. Mejail, "Skeleton-based action recognition using citation-knn on bags of time-stamped pose descriptors," in *Proc. Int. Conf. Image Process.*, 2016, pp. 3051–3055.
[43] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[44] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 895–903.
[45] Y. Chen, D. Zhao, L. Lv, and Q. Zhang, "Multi-task learning for dangerous object detection in autonomous driving," *Inf. Sci.*, vol. 432, pp. 559–571, 2018.
[46] V. Ramanathan *et al.*, "Learning semantic relationships for better action retrieval in images," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1100–1109.
[47] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. Assoc. Advance. Artif. Intell.*, 2015, pp. 2181–2187.
[48] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[49] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 2171–2180.
[50] T. Chen, W. Yu, and R. Chen, "Knowledge-embedded routing network for scene graph generation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6156–6164.
[51] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.