Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition

Guojun Yin[1,2], Lu Sheng[2], Bin Liu[1], Nenghai Yu[1], Xiaogang Wang[2], Jing Shao[3(✉)], and Chen Change Loy[4]

[1] Key Laboratory of Electromagnetic Space Information,
University of Science and Technology of China,
The Chinese Academy of Sciences, Hefei, China
gjyin@mail.ustc.edu.cn, {flowice,ynh}@ustc.edu.cn
[2] The Chinese University of Hong Kong, Shatin, China
{lsheng,xgwang}@ee.cuhk.edu.hk
[3] SenseTime Research, Beijing, China
shaojing@sensetime.com
[4] Nanyang Technological University, Singapore, Singapore
ccloy@ieee.org

Abstract. Recognizing visual relationships ⟨subject-predicate-object⟩ among any pair of localized objects is pivotal for image understanding. Previous studies have shown remarkable progress in exploiting linguistic priors or external textual information to improve the performance. In this work, we investigate an orthogonal perspective based on feature interactions. We show that by encouraging deep message propagation and interactions between local object features and global predicate features, one can achieve compelling performance in recognizing complex relationships without using any linguistic priors. To this end, we present two new pooling cells to encourage feature interactions: (i) Contrastive ROI Pooling Cell, which has a unique deROI pooling that inversely pools local object features to the corresponding area of global predicate features. (ii) Pyramid ROI Pooling Cell, which broadcasts global predicate features to reinforce local object features. The two cells constitute a *Spatiality-Context-Appearance Module (SCA-M)*, which can be further stacked consecutively to form our final *Zoom-Net*. We further shed light on how one could resolve ambiguous and noisy object and predicate annotations by Intra-Hierarchical trees (IH-tree). Extensive experiments conducted on Visual Genome dataset demonstrate the effectiveness of our feature-oriented approach compared to state-of-the-art methods (Acc@1 11.42% from 8.16%) that depend on explicit modeling of linguistic interactions. We further show that SCA-M can be incorporated seamlessly into existing approaches to improve the performance by a large margin.

1   Introduction

Visual relationship recognition [22,30,38] aims at interpreting rich interactions between a pair of localized objects, *i.e.*, performing tuple recognition in the form of ⟨*subject-predicate-object*⟩ as shown in Fig. 1(a). The fundamental challenge of this task is to recognize various vaguely defined relationships given diverse spatial layouts of objects and complex inter-object interactions. To complement visual-based recognition, a promising approach is to adopt a linguistic model and learn relationships between object and predicate labels from the language. This strategy has been shown effective by many existing methods [6,26,30,47–49]. These language-based methods either apply statistical inference to the tuple label set, establish a linguistic graph as the prior, or mine linguistic knowledge from external billion-scale textual data (*e.g.*, Wikipedia).
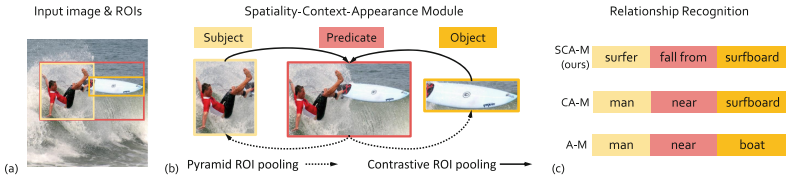

Fig. 1. Given an image '*surfer fall from surfboard*' and its region-of-interests (ROI) in (a), traditional methods without mining contextual interactions between object (subject) and predicate (*e.g.*, Appearance Module (A-M)) or ignoring spatial information (*e.g.*, Context-Appearance Module (CA-M)) may fail in relationship recognition, as shown in the two bottom rows of (c). The proposed Spatiality-Context-Appearance Module (SCA-M) in (b) permits global inter-object interaction and sharing of spatiality-aware contextual information, thus leading to a better recognition performance.

In this paper, we explore a novel perspective beyond the linguistic-based paradigm. In particular, contemporary approaches typically recognize the tuple ⟨*subject-predicate-object*⟩ via separate convolutional neural network (CNN) branches. We believe that by enhancing message sharing and feature interactions among these branches, the participating objects and their visual relationship can be better recognized. To this end, we formulate a new spatiality-aware contextual feature learning model, named as **Zoom-Net**. Differing from previous studies that learn appearance and spatial features separately, Zoom-Net propagates spatiality-aware object features to interact with the predicate features and broadcasts predicate features to reinforce the features of subject and object.

The core of Zoom-Net is a **Spatiality-Context-Appearance Module**, abbreviated as **SCA-M**. It consists of two novel pooling cells that permit deep feature interactions between objects and predicates, as shown in Fig. 1(b). The first cell, *Contrastive ROI Pooling Cell*, facilitates predicate feature learning by

inversely pooling object/subject features to a matching spatial context of predicate features via a unique deROI pooling. This allows all subject and object to fall on the same spatial 'palette' for spatiality-aware feature learning. The second cell is called *Pyramid ROI Pooling Cell*. It helps object/subject feature learning through broadcasting the predicate features to the corresponding object's/subject's spatial area. Zoom-Net stacks multiple SCA-Ms consecutively in an end-to-end network that allows multi-scale bidirectional message passing among subject, predicate and object. As shown in Fig. 1(c), the message sharing and feature interaction not only help recognize individual objects more accurately but also facilitate the learning of inter-object relation.

Another contribution of our work is an effective strategy of mitigating ambiguity and imbalanced data distribution in ⟨*subject-predicate-object*⟩ annotations. Specifically, we conduct our main experiments on the challenging Visual Genome (VG) dataset [22], which consists of over 5,319 object categories, 1,957 predicates, and 421,697 relationship types. The large-scale ambiguous categories and extremely imbalanced data distribution in VG dataset prevent previous methods from predicting reliable relationships despite they succeed in the Visual Relationship Detection (VRD) dataset [30] with only 100 object categories, 70 predicates and 6,672 relationships. To alleviate the ambiguity and imbalanced data distribution in VG, we reformulate the conventional one-hot classification as a n-hot multi-class hierarchical recognition via a novel Intra-Hierarchical trees (IH-trees) for each label set in the tuple ⟨*subject-predicate-object*⟩.

Contributions. Our contributions are summarized as follows:

(1) *A general feature learning module that permits feature interactions* - We introduce a novel SCA-M to mining intrinsic interactions between low-level spatial information and high-level semantic appearance features simultaneously. By stacking multiple SCA-Ms into a Zoom-Net, we achieve compelling results on VG dataset thanks to the multi-scale bidirectional message passing among subject, predicate and object.
(2) *Multi-class Intra-Hierarchical tree* - To mitigate label ambiguity in large-scale datasets, we reformulate the visual relationship recognition problem to a multi-label recognition problem. The recognizability is enhanced by introducing an Intra-Hierarchical tree (IH-tree) for the object and predicate categories, respectively. We show that IH-tree can benefit other existing methods as well.
(3) *Large-scale relationship recognition* - Extensive experiments demonstrate the respective effectiveness of the proposed SCA-M and IH-tree, as well as their combination on the challenging large-scale VG dataset.

It is noteworthy that the proposed method differs significantly from previous works as Zoom-Net neither models explicit nor implicit label-level interactions between ⟨*subject-predicate-object*⟩. We show that feature-level interactions alone, which is enabled by SCA-M, can achieve state-of-the-art performance. We further demonstrate that previous state-of-the-arts [26] that are based on label-level interaction can benefit from the proposed SCA-M and IH-trees.

2   Related Work

Contextual Learning. Contextual information has been employed in various tasks [1,13,15,25,34,40,42], *e.g.*, object detection, segmentation, and retrieval. For example, the visual features captured from a bank of object detectors are combined with global features in [5,24]. For both detection and segmentation, learning feature representations from a global view rather than the located object itself has been proven effective in [3,23,32]. Contextual feature learning for visual relationship recognition is little explored in previous works.

Class Hierarchy. In previous studies [8,9,11,17,33], class hierarchy that encodes diverse label relations or structures is used to improve performances on classification and retrieval. For instance, Deng *et al.* [11] improve large-scale visual recognition of object categories by forming a semantic hierarchy that consists of many levels of abstraction. While object categories can be clustered easily by their semantic similarity given the clean and explicit labels of objects, building a semantic hierarchy for visual relationship recognition can be more challenging due to noisy and ambiguous labels. Moreover, the semantic similarity between some phrases and prepositions such as *walking on a* versus *walks near the* is not directly measurable. In our paper, we employ the part-of-speech tagger toolkit to extract and normalize the keywords of these labels, *e.g. walk*, *on* and *near*.

Visual Relationship. Recognizing visual relationship [38] has been shown beneficial to various tasks, including action recogntion [7,15], pose estimation [12], recognition and object detection [4,36], and scene graph generation [27,44]. Most recent works [6,18,27–30,35,45,48,50] focus on measuring linguistic relations with textual priors or language models. The linguistic relations have been explored for object recognition [9,31,43], object detection [37], retrieval [39], and caption generation [16,20,21]. Yu *et al.* [46] employ billions of external textual data to distill useful knowledge for triplet ⟨*subject-predicate-object*⟩ learning. These methods do not fully explore the potential of feature learning and feature-level message sharing for the problem of visual relationship recognition. Li *et al.* [26] propose a message passing strategy to encourage feature sharing between features extracted from ⟨*subject-predicate-object*⟩. However, the network does not capture the relative location of different objects thus it cannot capture valid contextual information between subject, predicate and object.

3   Zoom-Net: Mining Deep Feature Interactions

We propose an end-to-end visual relationship recognition model that is capable of mining feature-level interactions. This is beyond just measuring the interactions among the triplet labels with additional linguistic priors, as what previous studies considered.
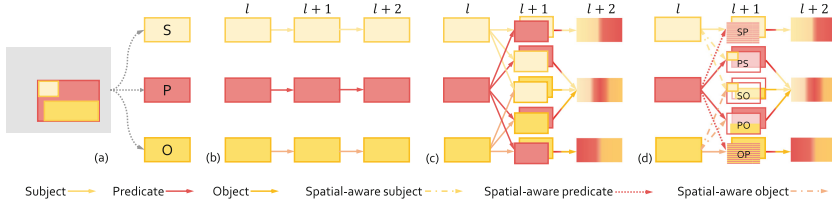
Fig. 2. (a) Given the ROI-pooled features of subject (S), predicate (P) and object (O) from an input image, (b) An Appearance module (A-M) separately processes these features without any message passing, (c) a Context-Appearance module (CA-M) attempts to capture contextual information by directly fusing pairwise features. The proposed SCA-M in (d) integrates the local and global contextual information in a spatiality-aware manner. The SP/PS/SO/PO/OP features are combined by channel-wise concatenation. For instance, SP feature is the result of combining subject and predicate features.

3.1   Appearance, Context and Spatiality

As shown in Fig. 2(a), given the ROI-pooled features of the subject, predicate and object, we consider a question: how to learn good features for both object (subject) and predicate? We investigate three plausible modules as follows.

Appearance Module. This module focuses on the intra-dependencies within each ROI, *i.e.*, the features of the subject, predicate and object branches are learned independently without any message passing. We term this network structure as Appearance Module (A-M), as shown in Fig. 2(a). No contextual and spatial information can be derived from such a module.

Context-Appearance Module. The Context-Appearance Module (CA-M) [26] directly fuses pairwise features among three branches, in which subject/object features absorb the contextual information from the predicate features, and predicate features also receive messages from both subject/object features, as shown in Fig. 2(b). Nonetheless, these features are concatenated regardless of their relative spatial layout in the original image. The incompatibility of scale and spatiality makes the fused features less optimal in capturing the required spatial and contextual information.

Spatiality-Context-Appearance Module. The spatial configuration, *e.g.*, the relative positions and sizes of subject and object, is not sufficiently represented in CA-M. To address this issue, we propose a Spatiality-Context-Appearance module (**SCA-M**) as shown in Fig. 2(c). It consists of two novel spatiality-aware feature alignment cells (*i.e.*, *Contrast ROI Pooling* and *Pyramid ROI Pooling*) for message passing between different branches. In comparison to CA-M, the proposed SCA-M reformulates the local and global
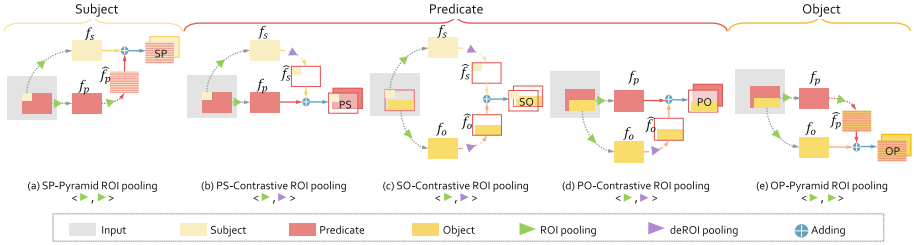
Fig. 3. The Spatiality-Context-Appearance Module (SCA-M) hinges on two components: (i) Contrastive ROI pooling (b–d), denoted as ⟨ROI, deROI⟩, which propagates spatiality-aware features $\hat{f}_s, \hat{f}_o$ from *subject* and *object* into the spatial 'palette' of *predicate* features f_p, and (ii) Pyramid ROI pooling (a,e), ⟨ROI, ROI⟩, which broadcasts the global *predicate* features $\hat{f}_p$ to local features f_s, f_o in *subject* and *object* branches.

information integration in a spatiality-aware manner, leading to superior capability in capturing spatial and contextual relationships between the features of ⟨*subject-predicate-object*⟩.

3.2 Spatiality-Context-Appearance Module (SCA-M)

We denote the respective regions of interest (ROIs) of the subject, predicate and object as $\mathcal{R}_s$, $\mathcal{R}_p$, and $\mathcal{R}_o$, where $\mathcal{R}_p$ is the union bounding box that tightly covers both the subject and object. The ROI-pooled features for these three ROIs are $\mathbf{f}_t, t \in \{s, p, o\}$, respectively. In this section, we present the details of SCA-M. In particular, we discuss how Contrastive ROI Pooling and Pyramid ROI Pooling cells, the two elements in SCA-M, permit deep feature interactions between objects and predicates.

Contrastive ROI Pooling denotes a pair of ⟨ROI, deROI⟩ operations that the object[1] features $\mathbf{f}_o$ are at first ROI pooled for extracting normalized local features, and then these features are deROI pooled back to the spatial palette of the predicate feature $\mathbf{f}_p$, so as to generate a spatiality-aware object feature $\hat{\mathbf{f}}_o$ with the same size as the predicate feature, as shown in Fig. 3(b) marked by the purple triangle. Note that the remaining region outside the relative object ROI in $\hat{\mathbf{f}}_o$ is set to 0. The spatiality-resumed local feature $\hat{\mathbf{f}}_o$ can thus influence the respective regions in the global feature map $\mathbf{f}_p$. In practice, the proposed deROI pooling can be considered as an inverse operation of the traditional ROI pooling (green triangle in Fig. 3), which is analogous to the top-down deconvolution versus the bottom-up convolution.

There are three Contrastive ROI pooling cells presented in the SCA-M module to integrate the feature pairs *subject-predicate*, *subject-object* and *predicate-object*, as shown in Fig. 3(b–d). Followed by several convolutional layers, the

---

[1] Subject and object refer to the same concept, thus we only take object as the example for illustration.

features from subject and object are spatially fused into the predicate feature for enhanced representation capability. The proposed ⟨ROI, deROI⟩ operations differ from conventional feature fusion operations (channel-wise concatenation or summation). The latter would introduce scale incompatibility between local subject/object features and global predicate features, which could hamper feature learning in subsequent convolutional layers.

Pyramid ROI Pooling denotes a pair of ⟨ROI, ROI⟩ operations that broadcasts the global predicate features to local features in the subject and object branches, as shown in Fig. 3(a) and (e). Specifically, with the help of ROI pooling unit, we first ROI-pool the features of predicate from the input region $\tilde{\mathcal{R}}$, which convey global contextual information of the region. Next, we perform a second ROI pooling on predicate features with the subject/object ROIs to further mine the contextual information from the global predicate feature region. The Pyramid ROI pooling thus provides multi-scale contexts to facilitate subject/object feature learning.

3.3   Zoom-Net: Stacked SCA-M

By stacking multiple SCA-Ms, the proposed **Zoom-Net** is capable of capturing multi-scale feature interactions with dynamic contextual and spatial information aggregation. It enables a reliable recognition of the visual relationship triplet ⟨*s-p-o*⟩, where the predicate *p* indicates the relationships (*e.g.*, spatiality, preposition, action and etc.) between a pair of localized subject *s* and object *o*.

As visualized in Fig. 4, we use a shared feature extractor with convolutional layers until `conv3_3` to encode appearance features of different object categories.
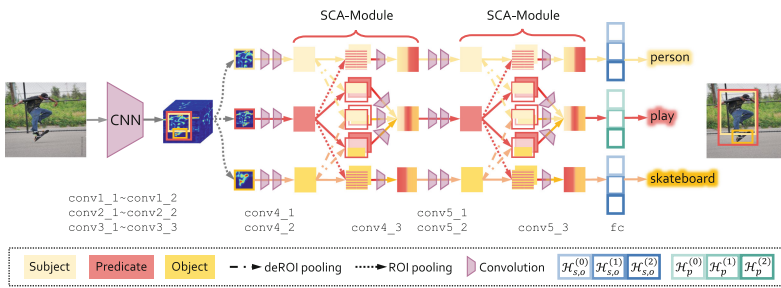

Fig. 4. The architecture of Zoom-Net. The subject (in light yellow), predicate (in red) and object (in dark yellow) share the same feature extraction procedure in the lower layers, and are then ROI-pooled into three branches. Following each branch of pooled feature maps is two convolutional layers to learn appearance features which are then fed into two stacked SCA-Ms to further fuse multi-scale spatiality-aware contextual information across different branches. Three classifiers with intra-hierarchy structures are applied to the features obtained from each branch for visual relationship recognition. (Color figure online)

By indicating the regions of interests (ROIs) for subject, predicate and object, the associated features are ROI-pooled to the same spatial size and respectively fed into three branches. The features in three branches are at first independently fed into two convolutional layers (the `conv4_1` and `conv4_2` layers in VGG-16) for a further abstraction of their *appearance* features. Then these features are put into the first SCA-M to fuse *spatiality*-aware *contextual* information across different branches. After receiving the interaction-augmented subject, predicate and object features from the first SCA-M, $\mathcal{M}_{SCA}^1$, we continue to convolve these features with another two appearance abstraction layers (mimicking the structures of `conv5_1` and `conv5_2` layers in VGG-16) and then forward them to the second SCA-M, $\mathcal{M}_{SCA}^2$. After this module, the multi-scale interaction-augmented features in each branch are fed into three fully connected layers `fc_s`, `fc_p` and `fc_o` to classify subject, predicate and object, respectively.

4   Hierarchical Relational Classification

To thoroughly evaluate the proposed Zoom-Net, we adopt the Visual Genome (VG) dataset[2] [22] for its large scale and diverse relationships. Our goal is to understand the a much broader scope of relationships with a total number of 421,697 relationship types, in comparison to the VRD dataset [30] that focuses on only 6,672 relationships. Recognizing relationships in VG is a non-trivial task due to several reasons:

(1) *Variety* - There are a total of 5,319 object categories and 1,957 predicates, tens times than those available in the VRD dataset.
(2) *Ambiguity* - Some object categories share a similar appearance, and multiple predicates refer to the same relationship.
(3) *Imbalance* - We observe long tail distributions both for objects and predicates.

To circumvent the aforementioned challenges, existing studies typically simplify the problem by manually removing a considerable portion of the data by frequency filtering or cleaning [6,26,46,47]. Nevertheless, infrequent labels like "old man" and "white shirt" contain common attributes like "man" and "shirt" and are unreasonable to be pruned. Moreover, the flat label structure assumed by these methods is limited to describe the label space of the VG dataset with ambiguous and noisy labels.

To overcome the aforementioned issues, we propose a solution by establishing two Intra-Hierarchical trees (IH-tree) for measuring intra-class correlation within object[3] and predicate, respectively. IH-tree builds a hierarchy of concepts that systematically groups rare, noisy and ambiguous labels together with those clearly defined labels. Unlike existing works that regularize relationships across

---

[2] Extremely rare labels (fewer than 10 samples) were pruned for a valid evaluation.

[3] Subject and object refer to the same term in this paper, thus we only take the object as the example for illustration.

the triplet ⟨*s-p-o*⟩ by external linguistic priors, we only consider the intra-class correlation to independently regularize the occurrences of the object and predicate labels. During end-to-end training, the network employs the weighted Intra-Hierarchical losses for visual relationship recognition as $\mathcal{L} = \alpha\mathcal{L}_s + \beta\mathcal{L}_p + \gamma\mathcal{L}_o$, where hyper-parameters α, β, γ balance the losses with respect to subject $\mathcal{L}_s$, predicate $\mathcal{L}_p$ and object $\mathcal{L}_o$. $\alpha = \beta = \gamma = 1$ in our experiments. We introduce IH-tree and the losses next.
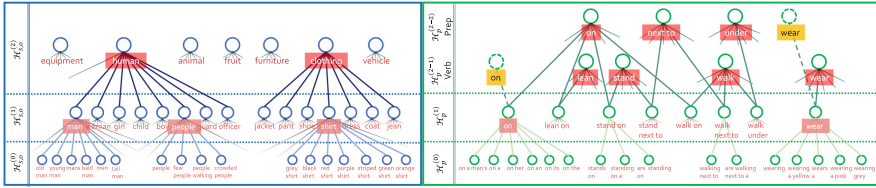

Fig. 5. An illustration of Intra-Hierarchical Tree. Both IH-trees for object (left) and predicate (right) start from the base layer $\mathcal{H}_{s,p,o}^{(0)}$ to a purified layer $\mathcal{H}_{s,p,o}^{(1)}$ but have a different construction in the third layer. The $\mathcal{H}_o^{(2)}$ clusters similar semantic concepts from $\mathcal{H}_o^{(1)}$, while the $\mathcal{H}_p^{(2)}$ separately cluster verb and preposition words from $\mathcal{H}_p^{(1)}$.

4.1 Intra-Hierarchical Tree $\mathcal{H}_o$ for Object

We build an IH-tree, $\mathcal{H}_o$, for object with a depth of three, where the base layer $\mathcal{H}_o^{(0)}$ consists of the raw object categories.

(1) $\mathcal{H}_o^{(0)} \rightarrow \mathcal{H}_o^{(1)}$: $\mathcal{H}_o^{(1)}$ is extracted from $\mathcal{H}_o^{(0)}$ by pruning noisy labels with the same concept but different descriptive attributes or in different singular and plural forms. We employ the part-of-speech tagger toolkit from NLTK [2] and NLTK Lemmatizer to filter and normalize the noun keyword, *e.g.*, "man" from "old man", "bald man" and "men".

(2) $\mathcal{H}_o^{(1)} \rightarrow \mathcal{H}_o^{(2)}$: We observe that some labels have a close semantic correlation. As shown in the left panel of Fig. 5, labels with similar semantic concepts such as "shirt" and "jacket" are hyponyms of "clothing" and need to be distinguished from other semantic concepts like "animal" and "vehicle". Therefore, we cluster labels in $\mathcal{H}_o^{(1)}$ to the third level $\mathcal{H}_o^{(2)}$ by semantical similarities computed by Leacock-Chodorow distance [40] from NLTK. We find that a threshold of 0.65 is well-suited for splitting semantic concepts.

The output of the subject/object branch is a concatenation of three independent `softmax` activated vectors corresponded to three hierarchical levels in the IH-tree. The loss $\mathcal{L}_s$ ($\mathcal{L}_o$) is thus a summation of three independent `softmax` losses with respect to these levels, encouraging the intra-level mutual label exclusion and inter-level label dependency.

4.2   Intra-Hierarchical Tree $\mathcal{H}_p$ for Predicate

The predicate IH-tree also has three hierarchy levels. Different from the object IR-tree that only handles nouns, the predicate categories include various part-of-speech types, *e.g.*, verb (action) and preposition (spatial position). Even a single predicate label may contain multiple types, *e.g.*, "are standing on" and "walking next to a".

(1) $\mathcal{H}_p^{(0)} \to \mathcal{H}_p^{(1)}$: Similar to $\mathcal{H}_o^{(1)}$, $\mathcal{H}_p^{(1)}$ is constructed aiming at extracting and normalizing keywords from predicates. We retain the keywords and normalize tenses with respective to three main part-of-speech types, *i.e.*, verb, preposition and adjective, and abandon other pointless and ambiguous words. As shown in the right panel of Fig. 5, "wears a", "wearing a yellow" and "wearing a pink" are mapped to the same keyword "wear".

(2) $\mathcal{H}_p^{(1)} \to \mathcal{H}_p^{(2)}$: Different part-of-speech types own particular characteristics with various context representations, and hence a separate hierarchical structure for the verb (action) and preposition (spatial) is indispensable for better depiction. To this end, we construct $\mathcal{H}_p^{(2)}$ for verb and preposition label independently, *i.e.*, $\mathcal{H}_p^{(2-1)}$ for action information and $\mathcal{H}_p^{(2-2)}$ for spatial configuration. There are two cases in $\mathcal{H}_p^{(1)}$: (a) the label is in the form of phrase that consists of both verb and preposition (*e.g.* "stand on" and "walk next to") and (b) the label is a single word (*e.g.*, "on" and "wear"). For the first case, $\mathcal{H}_p^{(2-1)}$ extracts the verb words from the two phrases while $\mathcal{H}_p^{(2-2)}$ extracts the preposition words. It thus causes that a label might be simultaneously clustered into different partitions of $\mathcal{H}_p^{(2)}$. If the label is a single word, it would be normally clustered into the corresponding part-of-speech but remained the same in the opposite part-of-speech, as shown with the dotted line in the right panel of Fig. 5. The loss $\mathcal{L}_p$ is constructed similarly to that for the object.

5   Experiments on Visual Genome (VG) Dataset

Dataset. We evaluate our method on the Visual Genome (VG) dataset (version 1.2). Each image is annotated with a triplet $\langle subject\text{-}predicate\text{-}object \rangle$, where the subjects and objects are annotated with labels and bounding boxes while the predicates only have labels. We randomly split the VG dataset into training and testing set with a ratio of $8:2$. Note that both sets are guaranteed to have positive and negative samples from each object or predicate category. The details of data preprocessing and the source code will be released.

Evaluation Metrics. (1) *Acc@N*. We adopt the *Accuracy* score as the major evaluation metric in our experiments. The metric is commonly used in traditional classification tasks. Specifically, we report the values of both Acc@1 and Acc@5 for *subject, predicate, object* and *relationship*, where the accuracy of *relationship* is calculated as the averaged accuracies of *subject, predicate* and *object*.

(2) *Rec@N*. Following [30], we use *Recall* as another metric so as to handle incomplete annotations. Rec@N computes the ratio of the correct relationship instance that is covered in the top N predictions per image. We report Rec@50 and Rec@100 in our experiments. For a fair comparison, we follow [30] to evaluate Rec@N on three tasks, *i.e.*, *predicate recognition* where both the labels and bounding boxes of the subject and object are given; *phrase recognition* that takes a triplet as a union bounding box and predicts the triple labels; *relationship recognition*, which also outputs triple labels but evaluates separate bounding boxes of subject and object. The recall performance is relative to the number of *predicate* per *subject-object* pair to be evaluated, *i.e.*, top k predictions. In the experiments on VG dataset, we adopt top $k = 100$ for evaluation.

Training Details. We use VGG16 [41] pre-trained on ImageNet [10] as the network backbone. The newly introduced layers are randomly initialized. We set the base learning rate as 0.001 and fix the parameters from `conv1_1` to `conv3_3`. The implementations are based on Caffe [19], and the networks are optimized via SGD. The conventional feature fusion operations are implemented by channel-wise concatenation in SCA-M cells here.

5.1   Ablation Study

SCA-Module. The advantage of Zoom-Net lies in its unique capability of learning spatiality-aware contextual information through the SCA-M. To demonstrate the benefits of learning visual features with spatial-oriented and context-aided cues, we compare the recognition performance of Zoom-Net with a set of variants achieved by removing each individual cue step by step, *i.e.*, the SCA-M without stacked structure, the CA-M that disregard the spatial layouts, and the vanilla A-M that does not perform message passing (see Sect. 3.1). Their accuracy and recall scores are reported in Table 1.

In comparison to the vanilla A-M, both the CA-M and SCA-M obtain a significant improvement suggesting the importance of contextual information to individual subject, predicate, and object classification and their relationship recognition. Note that contemporary CNNs have already shown a remarkable performance on subject and object classification, *i.e.*, it is not hard to recognize object via individual appearance information, and thus the gap (4.96%) of subject is smaller than that of predicate (12.25%) between A-M and SCA-M on Top-1 accuracy. Not surprisingly, since the key inherent problem of relationship recognition is to learning the interactions between subject and object, the proposed SCA-M module exhibit a strong performance, thanks to its capability in capturing correlation between spatiality and semantic appearance cues among different object. Its effectiveness can also be observed from qualitative comparisons in Fig. 6(a).

Intra-Hierarchical Tree. We use the two auxiliary levels of hierarchical labels $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ to facilitate the prediction of the raw ground truth labels $\mathcal{H}^{(0)}$ for

Table 1. Recognition performances (Acc@N and Rec@N) of Zoom-Net on VG dataset compared with (i) three variants of SCA module, and (ii) Zoom-Net discarding IH-trees. The best results are indicated in bold.

Metrics		@N	Zoom-Net	SCA-M	CA-M	A-M	Zoom-Net w/o $\mathcal{H}^{(1,2)}$	Zoom-Net w/o $\mathcal{H}^{(2)}$
Acc.	Subject	1	**38.94**	37.48	34.84	32.52	36.52	37.88
		5	**65.70**	64.09	61.59	58.28	62.63	63.97
	Predicate	1	**48.73**	48.14	46.81	35.89	47.18	48.26
		5	**77.64**	76.97	75.55	67.05	76.43	77.18
	Object	1	**45.09**	44.13	42.66	41.39	42.52	43.67
		5	**71.69**	70.64	69.55	67.99	69.33	70.35
	Relationship	1	**11.42**	10.51	9.46	6.39	9.92	10.76
		5	**22.80**	21.31	19.70	14.06	20.44	22.08
Rec.	Predicate	50	**67.25**	66.54	65.07	53.94	65.84	66.73
		100	**77.51**	76.92	75.45	66.53	76.30	77.16
	Relationship	50	**19.97**	18.60	17.14	12.23	17.78	18.92
		100	**25.07**	23.51	21.63	15.86	22.53	23.88
	Phrase	50	**20.84**	19.55	18.12	13.05	18.65	19.78
		100	**26.16**	24.70	22.85	16.92	23.62	24.96

the subject, predicate and object, respectively. Here we show that by involving hierarchical structures to semantically cluster ambiguous and noisy labels, the recognition performance *w.r.t.* the raw labels of the subject, predicate, object as well as their relationships are all boosted, as shown in Table 1. Discarding one of two levels in IH-tree clearly hamper the performance, *i.e.*, Zoom-Net without IH-tree experiences a drop of around 1%–4% on different metrics. It reveals that intra-hierarchy structures do provide beneficial information to improve the recognition robustness. Besides, Fig. 6(b) shows the Top-5 triple relationship prediction results of Zoom-Net with and without IH-trees. The novel design of the hierarchical label structure help resolves data ambiguity for both on object and predicate. For example, thanks to the hierarchy level $\mathcal{H}^{(1)}$ introduced in Sect. 4, the predicates related to "wear" (*e.g.*, "wearing" and "wears") can be ranked in top predictions. Another example shows the contribution of $\mathcal{H}^{(2)}$ designed for semantic label clustering, *e.g.* "sitting in", which is grouped in the same cluster of the ground truth "in", also appears in top ranking results.

5.2 Comparison with State-of-the-Art Methods

We summarize the comparative results on VG in Table 2 with two recent state of the arts [6,26]. For a fair comparison, we implement both methods with the VGG-16 as the network backbone. The proposed Zoom-Net significantly outperforms these methods, quantitatively and qualitatively. Qualitative results are

Table 2. Recognition performances (Acc@N and Rec@N) of Zoom-Net on VG dataset compared with the state-of-the-art methods. Results in bold font are the best by a single model, while the underlined results indicate the best performance of a combined model that incorporates the proposed modules into other state-of-the-art architectures.

Metrics		@N	Zoom-Net	DR-Net [6]	ViP [26]	ViP+SCA -M	ViP+IH -tree	ViP+SCA-M +IH-tree
Acc.	Subject	1	**38.94**	30.10	31.10	*37.13*	*34.36*	*38.78*
		5	**65.70**	55.46	57.33	*63.61*	*61.03*	*65.69*
	Predicate	1	**48.73**	44.14	45.17	*48.40*	*46.54*	*49.07*
		5	**77.64**	71.67	74.26	*77.28*	*75.30*	*78.07*
	Object	1	**45.09**	37.91	39.18	*43.09*	*43.18*	*44.96*
		5	**71.69**	64.30	65.68	*69.93*	*69.48*	*71.58*
	Relationship	1	**11.42**	6.69	8.16	*10.65*	*9.97*	*11.79*
		5	**22.80**	13.11	17.01	*21.63*	*20.40*	*23.28*
Rec.	Predicate	50	**67.25**	62.05	63.44	*66.87*	*64.80*	*67.63*
		100	**77.51**	71.96	74.15	*77.22*	*75.29*	*77.89*
	Relationship	50	**19.97**	12.56	14.78	*18.73*	*17.76*	*20.41*
		100	**25.07**	16.06	18.85	*23.67*	*22.35*	*25.55*
	Phrase	50	**20.84**	13.51	15.70	*19.61*	*18.72*	*21.31*
		100	**26.16**	17.23	19.96	*24.70*	*23.50*	*26.66*

shown in the first row of Fig. 6(c). DR-Net [6] exploits binary dual masks as the spatial configuration in feature learning and therefore loses the critical interaction between visual context and spatial information. ViP [26] focuses on learning label interaction by proposing a phrase-guided message passing structure. Additionally, the method tries to capture contextual information by passing messages across triple branches before ROI pooling and thus fail to explore in-depth spatiality-aware feature representations.

Transferable SCA-M Module and IH-Tree. We further demonstrate the effectiveness of the proposed SCA-M module in capturing spatiality, context and appearance visual cues, and IH-trees for resolving ambiguous annotations, by plugging them into architectures of existing works. Here, we take the network of ViP [26] as the backbone for its end-to-end training scheme and state-of-the-art results (Table 2). We compare three configurations, *i.e.*, *ViP+SCA-M*, *ViP+IH-tree* and *ViP+SCA-M+IH-tree*. For a fair comparison, the ViP is modified by replacing the targeted components with SCA-M or IH-tree with other components fixed. As shown in Table 2, the performance of ViP is improved by a considerable margin on all evaluation metrics after applying our SCA-M (*i.e. ViP+SCA-M*). The results again suggest the superiority of the proposed spatiality-aware feature representations to that of ViP. Note that the overall performance by adding both stacked SCA module and IH-tree (*i.e.*, *ViP+SCA-M+IH-tree*) surpasses that of ViP itself. The ViP designs a phrase-guided message passing structure to learn textual connections among ⟨*subject-predicate-object*⟩ at label-level. On the contrary, we concentrate
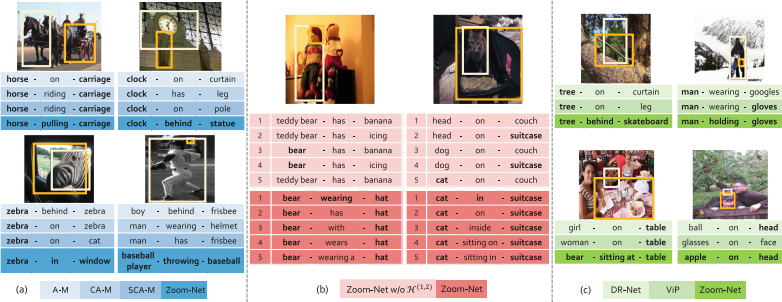
Fig. 6. Qualitative results on VG dataset. (a) Comparison results with the variants of different module configurations. (b) Results by discarding IH-trees. (c) Comparison between Zoom-Net with state-of-the-art methods. (a) and (c) show Top-1 prediction results while (b) provides Top-5 results for each method. The ground truth are in bold.

more on capturing contextual connections among $\langle subject\text{-}predicate\text{-}object\rangle$ at feature-level. Therefore, it's not surprising that a combination of these two aspects can provide a better result.

6   Comparisons on Visual Relationship Dataset (VRD)

Settings. We further quantitatively compare the performance of the proposed method with previous state of the arts on the Visual Relationship Dataset (VRD) [30]. The following comparisons keep the same settings as the prior arts. Since VRD has a clean annotation, we fine-tune the construction of IH-tree by removing the $\mathcal{H}_o^{(1)}$ and $\mathcal{H}_p^{(1)}$, which aim at reducing data ambiguity and noise in VG (details in Sect. 4). For a fair comparison, object proposals are generated by RPN [14] here and we use triplet NMS to remove redundant triplet candidates following the setting in [26] due to its excellent performance.

Evaluation Metrics. We follow [6,46] to report Recall@50 and Recall@100 when $k = 70$. The IoU between the predicted bounding boxes and the ground truth is required above 0.5 here. In addition, some previous works used $k = 1$ for evaluation and thus we report our results with $k = 1$ as well to compare these previous methods under the same conditions.

Results. The results listed in Table 3 show that the proposed Zoom-Net outperforms the state-of-the-art methods by significant gains on almost all the evaluation metrics[4]. In comparison to previous state-of-the-art approaches, Zoom-

---

[4] Note that Yu *et al.* [46] take external Wikipedia data with around 4 billion and 450 million sentences to distill linguistic knowledge for modeling the tuple correlation from label-aspect. It's not surprising to achieve a superior performance. In this experiment, we only compare with the results [46] without knowledge distillation.

Table 3. Comparisons with the referenced methods on VRD dataset. Results in bold indicate the best performance while the underlined results represent the next best. * marks the results of LK without knowledge distillation. ** marks the results of LK with knowledge distillation including large-scale external Wikipedia data.

k	Methods	Predicate		Relationship		Phrase	
		Rec@50	Rec@100	Rec@50	Rec@100	Rec@50	Rec@100
$k = 1$	LP [30]	47.87	47.87	13.86	14.70	16.17	17.03
	VTransE [47]	44.76	44.76	14.07	15.20	19.42	22.42
	VRL [29]	-	-	18.19	20.79	21.37	22.60
	PPRFCN [48]	47.43	47.43	14.41	15.72	19.62	23.15
	SA-Full [35]	50.40	50.40	14.90	16.10	16.70	18.10
	LK [46]*	47.50	47.50	16.57	17.69	19.15	19.98
	LK [46]**	<u>55.16</u>	<u>55.16</u>	<u>19.17</u>	21.34	23.14	24.03
	ViP [26]	-	-	17.32	20.01	22.78	27.91
	CAI [49]	53.59	53.59	15.63	17.39	17.60	19.24
	Zoom-Net	50.69	50.69	18.92	<u>21.41</u>	<u>24.82</u>	<u>28.09</u>
	CAI + SCA-M	**55.98**	**55.98**	***19.54***	***22.39***	***25.21***	***28.89***
$k = 70$	LK [46]*	74.98	86.97	20.12	<u>28.94</u>	22.59	25.54
	LK [46]**	<u>85.64</u>	**94.65**	**22.68**	**31.89**	26.32	29.43
	DR-Net [6]	80.78	81.90	17.73	20.88	19.93	23.45
	Zoom-Net	84.25	90.59	21.37	27.30	<u>29.05</u>	<u>37.34</u>
	CAI + SCA-M	**89.03**	<u>*94.56*</u>	<u>*22.34*</u>	*28.52*	***29.64***	***38.39***

Net improves the recall of predicate prediction by 3.47% Rec@50 and 3.62% Rec@100 when $k = 70$. Besides, the Rec@50 on relationship and phrase prediction tasks are increased by 1.25% and 6.46%, respectively. Note that the result of *predicate* ($k = 1$) only achieves comparable performance with some prior arts [29, 35, 46, 49] since these methods use the groundtruth of *subject* and *object* and only predict *predicate* while our method predicts *subject, predicate, object* together.

Among all prior arts designed without external data, CAI [49] has achieved the best performance on predicate prediction (53.59% Rec@50) by designing a context-aware interaction recognition framework to encode the labels into semantic space. To demonstrate the effectiveness and robustness of the proposed SCA-M in feature representation, we replace the visual feature representation in CAI [49] with our SCA-M (*i.e. CAI + SCA-M*). The performance improvements are significant as shown in Table 3 due to the better visual feature learned, *e.g.*, predicate Rec@50 is increased by 2.39% compared to [49]. In addition, with neither language priors, linguistic models nor external textual data, the proposed method can still achieve the state-of-the-art performance on most of the evaluation metrics, thanks to its superior feature representations.

7   Conclusion

We have presented an innovative framework Zoom-Net for visual relationship recognition, concentrating on feature learning with a novel Spatiality-Context-Appearance module (SCA-M). The unique design of SCA-M, which contains the proposed Contrastive ROI Pooling and Pyramid ROI Pooling Cells, benefits the learning of spatiality-aware contextual feature representation. We further designed the Intra-Hierarchical tree (IH-tree) to model intra-class correlations for handling ambiguous and noisy labels. Zoom-Net achieves the state-of-the-art performance on both VG and VRD datasets. We demonstrated the superiority and transferability of each component of Zoom-Net. It is interesting to explore the notion of feature interactions in other applications such as image retrieval and image caption generation.

References

1. Alexe, B., Heess, N., Teh, Y.W., Ferrari, V.: Searching for objects driven by context. In: NIPS (2012)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media Inc., Newton (2009)
3. Carreira, J., Li, F., Sminchisescu, C.: Object recognition by sequential figure-ground ranking. IJCV **98**, 243–262 (2012)
4. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: ICCV (2013)
5. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: CVPR (2010)
6. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: CVPR (2017)
7. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS (2011)
8. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR, pp. 785–792. IEEE (2011)
9. Deng, J., et al.: Large-scale object classification using label relation graphs. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 48–64. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_4
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)

11. Deng, J., Krause, J., Berg, A.C., Li, F.F.: Hedging your bets: optimizing accuracy-specificity trade-offs in large scale visual recognition. In: CVPR, pp. 3450–3457. IEEE, June 2012

12. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 158–172. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_12

13. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. IJCV **95**, 1–12 (2011)

14. Girshick, R.: Fast R-CNN. In: ICCV (2015)

15. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R* CNN. In: ICCV (2015)

16. Guadarrama, S., et al.: YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV (2013)

17. Hu, H., Zhou, G.T., Deng, Z., Liao, Z., Mori, G.: Learning structured inference neural networks with label relations. In: CVPR, pp. 2960–2968 (2016)

18. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR (2017)

19. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM (2014)

20. Karpathy, A., Joulin, A., Li, F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: NIPS (2014)

21. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)

22. Krishna, R., et al.: Visual genome: connecting language and vision using crowd-sourced dense image annotations. IJCV **123**, 32–73 (2017)

23. Li, C., Parikh, D., Chen, T.: Extracting adaptive contextual cues from unlabeled regions. In: ICCV (2011)

24. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)

25. Li, Y., Huang, C., Loy, C.C., Tang, X.: Human attribute recognition by deep hierarchical contexts. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 684–700. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_41

26. Li, Y., Ouyang, W., Wang, X., Tang, X.: ViP-CNN: Visual phrase guided convolutional neural network. In: CVPR (2017)

27. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)

28. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent topic-transition GAN for visual paragraph generation. In: ICCV (2017)

29. Liang, X., Lee, L., Xing, E.P.: Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: CVPR (2017)

30. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51

31. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR (2007)

32. Mottaghi, R., et al.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014)

33. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: ICCV, pp. 2768–2775. IEEE (2013)
34. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 241–254. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_18
35. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: ICCV (2017)
36. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV (2007)
37. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: CVPR (2017)
38. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
39. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the Fourth Workshop on Vision and Language (2015)
40. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of the 16th European Conference on Artificial Intelligence (2004)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint (2014)
42. Torralba, A., Murphy, K.P., Freeman, W.T.: Using the forest to see the trees: exploiting context for visual object detection and localization. Commun. ACM **53**, 107–114 (2010)
43. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions. In: BMVC (2009)
44. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
45. Yatskar, M., Zettlemoyer, L., Farhadi, A.: Situation recognition: visual semantic role labeling for image understanding. In: CVPR (2016)
46. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: ICCV (2017)
47. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: CVPR (2017)
48. Zhang, H., Kyaw, Z., Yu, J., Chang, S.F.: PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In: ICCV (2017)
49. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: ICCV (2017)
50. Zhuang, B., Wu, Q., Shen, C., Reid, I., van den Hengel, A.: Care about you: towards large-scale human-centric visual relationship detection. In: arXiv preprint (2017)