From Node to Graph: Joint Reasoning on Visual-Semantic Relational Graph for Zero-Shot Detection

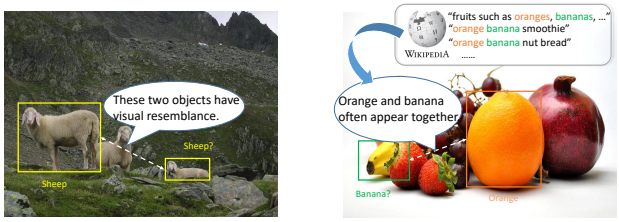
Hui Nie[1,2], Ruiping Wang[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
hui.nie@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Abstract

*Zero-Shot Detection (ZSD), which aims at localizing and recognizing unseen objects in a complicated scene, usually leverages the visual and semantic information of individual objects alone. However, scene understanding of human exceeds recognizing individual objects separately: the contextual information among multiple objects such as visual relational information (e.g. visually similar objects) and semantic relational information (e.g. co-occurrences) is helpful for understanding of visual scene. In this paper, we verify that contextual information plays a more important role in ZSD than in traditional object detection. To make full use of such information, we propose a new end-to-end ZSD method **GR**aph **A**ligning **N**etwork (**GRAN**) based on graph modeling and reasoning which simultaneously considers visual and semantic information of multiple objects instead of individual objects. Specifically, we formulate a Visual Relational Graph (VRG) and a Semantic Relational Graph (SRG), where the nodes are the objects in the image and the semantic representations of classes respectively and the edges are the relevance between nodes in each graph. To characterize mutual effect between two modalities, the two graphs are further merged into a heterogeneous Visual-Semantic Relational Graph (VSRG), where modal translators are designed for the two subgraphs to enable modal information to transform into a common space for communication, and message passing among nodes is enforced to refine their representations. Comprehensive experiments on MSCOCO dataset demonstrate the advantage of our method over state-of-the-arts, and qualitative analysis suggests the validity of using contextual information.*

1. Introduction

Object detection [16, 15, 42, 26, 49, 10, 4, 44, 9, 21, 36, 40, 29, 41, 27, 46] has been greatly developed with the convolutional neural network (CNN) in the past few years. Although these detection methods have achieved good performance, they can only detect objects seen in the training set. Recently, several works [39, 2, 11] propose Zero-Shot Detection (ZSD) task which aims to simultaneously localize and recognize unseen novel objects. Semantic information among categories is exploited to bridge the gap between seen and unseen classes. Nevertheless, the visual information is simply treated by separately matching the semantic information of each category to identify the isolated objects, resulting in no essential difference between ZSD and Zero-Shot Recognition (ZSR) [14, 23]. It should be noted that the target of ZSD is not a simple image containing merely one single object, but a natural scene with multiple related objects. Therefore, context is naturally useful information.



(a) Visual Relational Information (b) Semantic Relational Information

Figure 1. The illustrative diagrams of leveraging two kinds of contextual information: (a) Visual relational information can help detect small objects of the same category; (b) Oranges and bananas often appear together, hence we can use this semantic relational information to detect the occluded banana under the condition that an orange is detected.

The importance of context in ZSD is illustrated in Fig.1. The small object *sheep* and the occluded *banana* will probably fail to be detected in ZSD because of insufficient visual and semantic information of such individual objects. In such case, contextual information will help a lot. For ZSD, we can use contextual information including visual re-

lational information (visually similar objects) and semantic relational information (co-occurrences) to assist detection. For example, in Fig.1 (a) large objects can be used to help detect small visually similar objects $sheep$ as illustrated; in Fig.1 (b) given that the $orange$ is detected well, the semantic relational information of co-occurrences can be utilized to detect the occluded $banana$. These examples suggest that we should not only emphasize the information of the object itself but also exploit the contextual information.

In this work, we verify that contextual information among multi-objects plays a more important role in zero-shot detection than in traditional object detection through experimental study. Graph has shown to be a superior tool to model the visual and semantic relevance in many tasks. Thus, to leverage such contextual information, we propose a novel ZSD approach named **GR**aph **A**ligning **N**etwork (**GRAN**) based on graph modeling and reasoning. Specifically, for graph modeling, we devise a Visual-Semantic Relational Graph (VSRG) to comprehensively use both visual and semantic relational information. We first construct a Visual Relational Graph (VRG) and a Semantic Relational Graph (SRG) where the nodes are the objects in the image and the semantic representations of classes respectively and the edges are the relevance between nodes in each graph. To characterize mutual effect between two modalities, each node in VRG establishes a connection with all nodes in SRG to further formulate a heterogeneous VSRG by simply matching between visual and semantic information. For graph reasoning, modal translators are designed for these two graphs to transform the node states of different modals to a common space for communication. To update the representation of the individual node with information of other nodes, each node first decides which messages to send, then receives visual and semantic messages from other nodes on the VSRG that are highly relevant. Based on this new ZSD framework, zero-shot detection results are not only determined by visual and semantic information of individual objects but also affected by the visually related objects in the image and semantic relevance of prior.

Comprehensive experiments on MSCOCO dataset suggest that the visual and semantic relational information indeed improve the performance of ZSD with more desirable and reasonable outputs.

2. Related Work

Contextual information. Naturally, contextual information can help improve object detection, which needs to simultaneously recognize and localize multiple targets. There are many works [1, 8, 13, 19, 34, 47, 48] in the early stage to improve object detection by using context information. Recently, some methods based on CNN utilizing context have been proposed. Several works have tried to use context around an object or scene-level context in detection. ION [3] utilizes recurrent neural networks to exploit information both inside and outside the region of interest. GBD-Net [52] proposes a gated bi-directional CNN to pass messages between features from different support regions. SMN [6] has attempted to use the context of the relationship between objects, and proposes a sequential reasoning model that detects other objects based on existing detected objects. SIN [30] jointly models scene-level and object-object relationships, and proposes a structure inference network to inference object instances in the image via graph. Different from these works, we explore heterogeneous graph cross both visual and semantic modalities to model the relevance of the objects not only in visual image but also in semantic representation.

Graphical Model. Graph is a useful tool to model structure information. In the early years, some works [12, 20, 24, 32, 43, 45, 50] use a graphical model to improve performance. Graphical model and deep neural network have been integrated into a joint framework for group activity recognition in [12]. S-RNN [20] proposes an approach for combining spatio-temporal graphs and recurrent neural networks for diverse spatio-temporal tasks. GSNN [32] uses a knowledge graph to improve performance on image classification. Teney *et al.* [45] propose to build graphs for scene objects and question words to exploit the structure in these representations. Scene graph is utilized to model objects and their relationship in [50]. [51] constructs two graphs from visual and semantic aspects to solve visual relation detection task. Our work shares a similar spirit as [51] in constructing the graphs, however, the two works have essential differences that in [51] the representation and supervision of predicates are indispensable for graph modeling and reasoning respectively in their scene graph generation task, while our work targets at ZSD which does not need to model predicates into graph.

Zero-Shot Detection. Recently, there have been some attempts at ZSD. Rahman *et al.* [39] extend Faster R-CNN [42] with a semantic alignment network and use meta-classes to learn similarity among categories. Demirel *et al.* [11] adopt hybrid region embedding to detect unseen objects. Bansal *et al.* [2] introduce two background-aware approaches to distinguish unseen objects from background. Rahman *et al.* [38] adopt polarity loss maximizing the gap between positive and negative predictions to improve visual-semantic alignment. BLC [54] considers background word embedding is important for differentiating background from foreground objects and proposes a learnable background word embedding. [53, 17] absorb the advantages of the latest Zero-Shot Recognition (ZSR) methods and utilize a generative model to synthesize unseen class features to achieve great improvement. Some works [56, 55] focus on proposing object bounding boxes and use semantic information to improve recall rates for unseen ob-

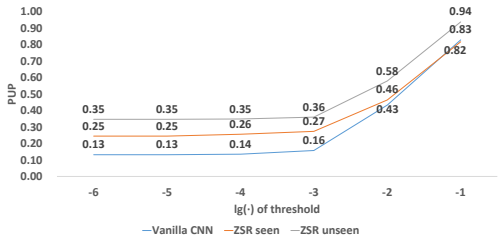


Figure 2. Comparison of *Proportion of Unreasonable Prediction (PUP)* of vanilla CNN with *PUP* of ZSR.

jects. Luo *et al.* [31] also take contextual information into account and propose to infer novel objects surrounded by known objects with inter-objects relation prior, where explicit graph modeling and reasoning with CRF are considered only in visual modal. Different from [31], we consider both visual and semantic modalities and explicitly construct a heterogeneous visual-semantic relational graph to reason unseen objects.

3. Preliminary Studies

While it has been generally recognized that contextual information is important for traditional object detection, we will further show that it is even more important for ZSD. In a real scene containing multiple objects, unreasonable predictions may occur if the information of the object itself is used only, e.g., predicting the mouse which is near the keyboard on the table as a car in an office scene. In order to eliminate unreasonable predictions, contextual information can be leveraged to identify the neighboring objects around an object. The basic idea is that the more unreasonable predictions using only the object's own information will generate, the greater the potential role of context information will be. In other words, the unreasonable predictions will be reduced more by using contextual information.

Since direct analysis between ZSD and traditional object detection is complicated, we relax the constraint to compare Zero-Shot Recognition (ZSR) and traditional image classification. This can be considered as completely accurate localization for ZSD and traditional object detection.

We first construct datasets for the two tasks. In order to reduce the difference from the popular ZSD benchmark, the datasets are constructed from MSCOCO [28] dataset. Specifically, for the ZSR dataset, we divide MSCOCO into 65 seen classes and 15 unseen classes (65/15 split) as proposed by [38], crop the objects within the bounding box area of the image and save them respectively to form the training set *zsr_train* (65 classes) and test set *zsr_test* (15 classes). During the process of cropping, we filter out some objects with small bounding boxes. For the dataset construction of traditional image classification task, the test set *ic_test* is the same as *zsr_test*, while the training set *ic_train* consists of the whole *zsr_train* and the data from the 15 unseen classes, i.e., *ic_train* covers all 80 classes. Each of the 15 categories has 1000 randomly chosen images. In order to more fairly compare the values calculated by co-occurrence times in subsequent stages, the original detection images of these newly added images do not contain any object of seen classes. More details can be seen in supplementary materials.

Next, we will define the unreasonable prediction. Let C be the number of classes in the dataset, M be the number of images in the test set and O_i ($i \in [1, M]$) be the i-th image (object). $Index(O_i)$ is the set of indexes of the images (objects) which are from the same detection image with O_i. $Pred(O_i)$ denotes the predicted label of O_i and $GT(O_i)$ denotes the ground truth label. The unreasonable prediction factor of O_i, denoted as a_i, is defined as follows:

$$a_i = \frac{\max_{j \in Index(O_i)} Co(Pred(O_i), GT(O_j))}{\sum_{c=1}^{C} Co(Pred(O_i), c)}, \quad (1)$$

where $Co(\cdot, \cdot)$ computes the co-occurrence times of two classes. Given a threshold τ, if unreasonable prediction factor $a_i < \tau$, then we regard it as an unreasonable prediction under the condition τ. With formulations above, we can define the $Proportion\ of\ Unreasonable\ Prediction\ (PUP)$ under the condition τ as follows:

$$\begin{aligned} PUP_\tau &= \frac{|unreasonable\ prediction|}{|wrong\ prediction|} \\ &= \frac{\Sigma_{i=1}^{M} \mathbb{I}(a_i < \tau)}{N_w}, \end{aligned} \quad (2)$$

where $\mathbb{I}(\cdot)$ is an indicator function, N_w is the number of wrong predictions.

For experiments on traditional image classification, we use ResNet-101 and change the output of the last fully-connected (FC) layer to 80, which is the total number of MSCOCO categories. For experiments on ZSR, we use a recently representative method TCN [22] (with publicly available source codes) that utilizes word embedding of class names as semantic representation to connect the seen and unseen classes.

The greater of *PUP*, the more helpful the contextual information is, which means that the unreasonable predictions of the model can be eliminated with contextual information. As shown in Fig.2, as the threshold decreases, the average *PUP* of ZSR is greater than vanilla CNN. Therefore, these results indicate that contextual information plays a more important role in ZSD than in traditional object detection.

4. Approach

Our goal is to comprehensively exploit visual and semantic relational information to enhance ZSD performance.

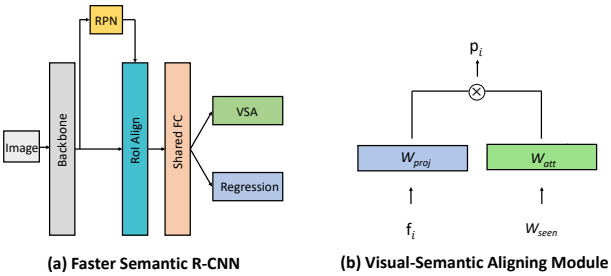


Figure 3. An illustration of our baseline model Faster Semantic R-CNN. (a) is the whole architecture of Faster Semantic R-CNN, VSA represents the Visual-Semantic Aligning Module. (b) is the detailed design of VSA.

To this end, different from existing methods that only exploit visual and semantic information of individual objects, we construct a visual-semantic relational graph to take full advantage of visual relevance in the scene and semantic relevance in semantic representation. Moreover, a **GR**aph **A**ligning **N**etwork (**GRAN**) is designed to pass messages between nodes and update node states in the graph. We will elaborate the whole framework of our method in the following sections.

4.1. Problem Definition

In zero-shot detection, we have N^s seen classes S and N^u unseen classes U, where the seen and unseen classes are disjoint, i.e. $S \cap U = \emptyset$. $\mathcal{X}^s$ is the images of seen classes, $\mathcal{Y}^s$ is the labels of seen classes, $\mathcal{Y}^u$ is the labels of unseen classes. The training data is $\mathcal{D}_{tr} = \{(x_i, y_i, \mathbf{b}_i)|x_i \in \mathcal{X}^s, y_i \in \mathcal{Y}^s, \mathbf{b}_i \in \mathbb{R}^4\}$, where x_i is the image containing multiple objects, y_i is class labels of objects, $\mathbf{b}_i$ represents bounding-box coordinates of objects in the image. No training images are available for unseen classes. Semantic information $\{\mathbf{d}_i\}_{i=1}^{N^s+N^u}$ are available in order to build up the relation between seen and unseen classes. The goal of ZSD is to learn detector of unseen classes $\mathcal{F}_{zsd} : \mathcal{X} \rightarrow \mathcal{Y}^u$ and the goal of Generalized Zero-Shot Detection (GZSD) is to learn more general detector of all classes $\mathcal{F}_{gzsd} : \mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$, where $\mathcal{X}$ denotes test images.

4.2. Baseline Model

In this section, we introduce our baseline model based on Faster R-CNN [42], which is shown to yield better recall in our experiments. It should be noted that our method can be added to any detection model as a plug-and-play module. In our baseline model, partial parameters of regression and classification branches are shared. Unless otherwise specified, we use ResNet-50 as the backbone of the baseline model. In addition, we also use the Feature Pyramid Network [26] to improve the recall for objects. We use RoI Align [18] to get a fixed length feature for each RoI. The process can be written as follows:

$$\mathbf{f}_i = RoIAlign(\mathbf{f}, r_i), \tag{3}$$

where $\mathbf{f}$ denotes the visual feature maps derived from the output of backbone, r_i is the i-th RoI in the image, and $\mathbf{f}_i$ is the feature vector of r_i.

Since the traditional detection model cannot detect unseen classes, we need to modify them to align visual and semantic information for zero-shot detection. We propose a visual-semantic aligning module to replace the classification head of Faster R-CNN to get our baseline model Faster Semantic R-CNN. The details of the module are demonstrated in Fig.3. Motivated by [38, 39], in order to detect unseen objects, we use the semantic representation of the class as supervision to train our model instead of class label. The visual-semantic aligning module contains three main components. After extracting object features, we use a learnable FC layer $W_{proj} \in \mathbb{R}^{v \times d}$ to project feature from visual to semantic space as illustrated in Fig.3, where v and d are the dimensions of visual and semantic space, respectively. The fixed FC layer $W_{seen} \in \mathbb{R}^{d \times (N^s+1)}$ consists of semantic representation (i.e. word embedding) of seen and background classes. For background class, we use the learned word embedding of background as in [54] for its better performance instead of the mean word vector of all classes. $W_{att} \in \mathbb{R}^{(N^s+1)\times(N^s+1)}$ is an adjustable FC layer to perform attention on fixed semantic representation W_{seen}. External vocabulary embeddings, introduced in [38] to enrich the capacity of semantic representation of classes, are not added into the baseline because of no improvement in performance. The visual-semantic aligning module can be formulated as follows:

$$\mathbf{f}_i^{sem} = \mathbf{f}_i W_{proj}, \tag{4}$$

$$\mathbf{f}^{emb} = tanh(W_{seen} W_{att}), \tag{5}$$

$$\mathbf{p}_i = Softmax(\mathbf{f}_i^{sem} \mathbf{f}^{emb}), \tag{6}$$

where $\mathbf{f}_i^{sem}$ denotes projected visual features, $\mathbf{f}^{emb}$ represents the attended word embeddings, and $\mathbf{p}_i \in \mathbb{R}^{N^s+1}$ denotes the prediction score of categories.

4.3. Graph Modeling

In order to simultaneously model the visual and semantic relevance, we construct a visual relational graph (VRG) and semantic relational graph (SRG), respectively.

In the VRG, each node of the graph represents an object in the scene. We link each pair of object nodes with an edge that formulates VRG as a complete graph. The edges represent the visual relevance among nodes in the VRG. To initialize the nodes of the graph, we first generate thousands of dense proposals in the image. We can get some RoIs (Region of Interest) after filtering out duplicate

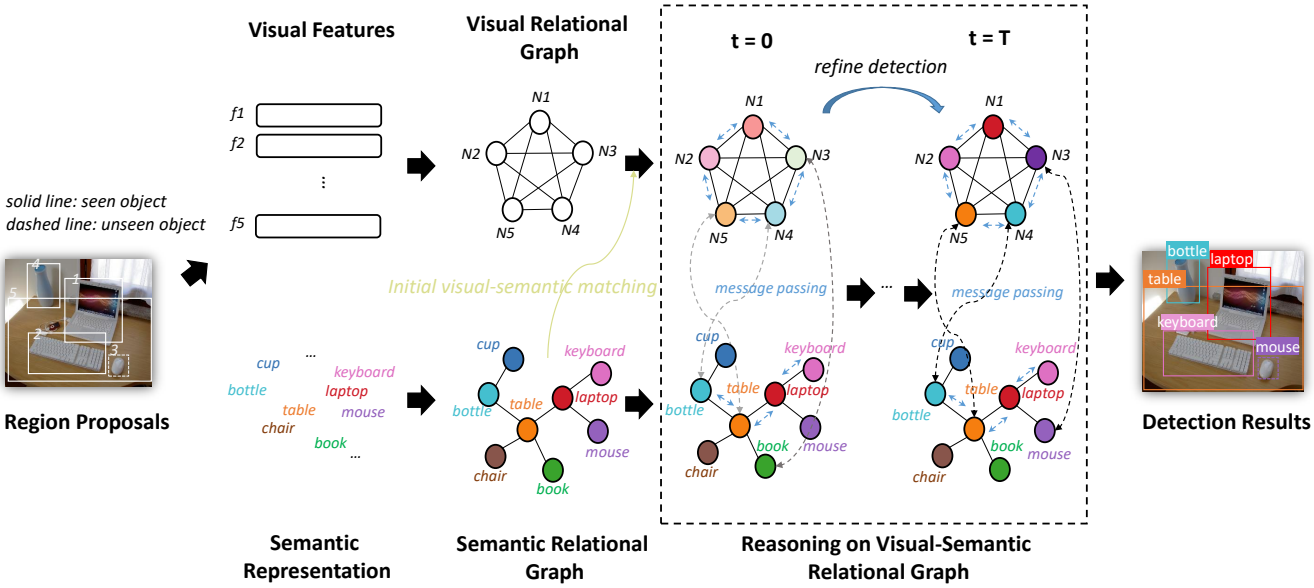


Figure 4. The pipeline of our proposed method. We first get some RoIs from the input image. Each RoI gets fixed-length visual features through RoIAlign and is initialized as node of Visual Relational Graph (VRG). Similarly, word embedding of each category is utilized as semantic representation to initialize node of Semantic Relational Graph (SRG). Next, VRG and SRG are further merged into a Visual-Semantic Relational Graph (VSRG), where edges between two graphs characterize the similarity of nodes through initial visual-semantic matching that is implemented by the visual-semantic aligning module. To reason and update nodes representation on the VSRG, we use GRaph Aligning Network (GRAN) to make nodes interact with each other by passing messages to enrich the representation of nodes for the downstream recognition module.

proposals. For each RoI r_i with its bounding box coordinates $\mathbf{b}_i = (x_i, y_i, w_i, h_i)$, we use RoI Align to extract visual feature $\mathbf{f}_i$ from it. A visual-to-semantic layer is deployed to obtain the initial probability prediction $\mathbf{p}_i$. Before the initialization of nodes, we use a visual modal translator layer φ^{vrg}_{proj} to project visual features into a common space for communication with semantic representations. Then the state of each node $\mathbf{n}^{vrg}_i$ in the VRG is initialized with corresponding $\mathbf{f}_i$ as follows:

$$\mathbf{n}^{vrg}_i = \varphi^{vrg}_{proj}(\mathbf{f}_i), \tag{7}$$

where φ^{vrg}_{proj} is a learnable FC layer. For modeling the semantic relevance, we also construct a graph SRG, where each node $\mathbf{n}^{srg}_i$ in the graph represents a category. We also link each pair of category nodes with an edge to formulate SRG as a complete graph, where the edges represent the semantic relevance among nodes in the graph.

Similar to VRG, before initialization of nodes in the SRG, we use a semantic modal translator layer φ^{srg}_{proj} to project word embedding $\mathbf{d}_i$ to the common space as follows:

$$\mathbf{n}^{srg}_i = \varphi^{srg}_{proj}(\mathbf{d}_i), \tag{8}$$

where φ^{srg}_{proj} is also a learnable FC layer. Then we use the projected word embedding of each class to initialize the corresponding node $\mathbf{n}^{srg}_i$ in the SRG.

Up to now, we have two isolated, initialized graphs: visual relational graph and semantic relational graph. In order to make visual and semantic information interact with each other, we need to establish a connection between the two graphs. For each node in the visual relational graph, it has a probability vector $\mathbf{p}_i$ from initial visual-semantic matching through baseline model Faster Semantic R-CNN. We link each node in VRG to all nodes in SRG with its probability vector $\mathbf{p}_i$ as the weights of edges. Same as the nodes in VRG, for each node in SRG, we create a reverse connection to all nodes in VRG. Finally, we form a heterogeneous graph named Visual-Semantic Relational Graph (VSRG) consisting of both visual and semantic relational graphs. The whole pipeline of our proposed model is illustrated in Fig.4.

4.4. Message Passing

After graph modeling, we obtain a heterogeneous graph. To make full use of the information of multiple objects at the same time, we should make nodes interact with each other on the VSRG. Hence we propose a GRaph Aligning Network (GRAN) motivated by GGNN [24] and GB-Net [51], to reason and update node states in the graph. We generate outgoing messages for propagating as in Fig.5. Specifically, each node representation in the heterogeneous graph is fed into a multi-layer perceptron to generate outgo-

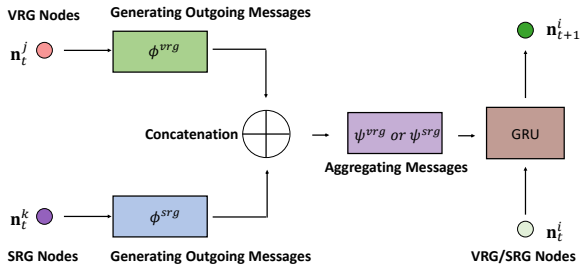


Figure 5. An illustration of message passing. The relational node states in VSRG are selected to send into GRU to update the current nodes.

ing messages as follows:

$$\mathbf{S}_i^{vrg} = \phi_{send}^{vrg}(\mathbf{n}_i^{vrg}), \tag{9}$$

$$\mathbf{S}_i^{srg} = \phi_{send}^{srg}(\mathbf{n}_i^{srg}), \tag{10}$$

where ϕ_{send}^{vrg}, ϕ_{send}^{srg} are learnable sending head which share weights across VRG nodes and SRG nodes, respectively. After generating outgoing messages, we send them through all outgoing edges multiplied by the edge weights. Then for each node $\mathbf{n}_i$, we collect messages from all nodes that have an edge with it, and get $\mathbf{R}_i^{vrg}$, $\mathbf{R}_i^{srg}$ as follows:

$$\mathbf{R}_i^{vrg} = \psi^{vrg}(e_{ij}\mathbf{S}_j^{vrg} + e_{ik}\mathbf{S}_k^{srg}), \tag{11}$$

$$\mathbf{R}_i^{srg} = \psi^{srg}(e_{ij}\mathbf{S}_j^{vrg} + e_{ik}\mathbf{S}_k^{srg}), \tag{12}$$

where e_{ij} represents the edge between node $\mathbf{n}_i$ and node $\mathbf{n}_j$; ψ^{vrg}, ψ^{srg} are learnable weight matrices. When node $\mathbf{n}_i$ and node $\mathbf{n}_j$ are both in VRG or SRG, e_{ij} is set to 1.

With information interaction, we can use visual and semantic relational information to update the states of the nodes and get a more precise representation for the downstream recognition module. We want to update the representations of the nodes with the received information and the previous states of the nodes, which requires modeling in the time series. Recurrent Neural Network (RNN) is a type of neural network used to process sequence data. In our method, we use Gated Recurrent Unit (GRU) [7], a variant of RNN, because of its well-designed memory mechanism and effectiveness. For each node $\mathbf{n}_i$ in VSRG, we have:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{R}_t + \mathbf{U}_z\mathbf{n}_t), \tag{13}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{R}_t + \mathbf{U}_r\mathbf{n}_t), \tag{14}$$

$$\mathbf{h}_t = tanh(\mathbf{W}_h\mathbf{R}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{n}_t)), \tag{15}$$

$$\mathbf{n}_{t+1} = (1 - \mathbf{z}_t) \odot \mathbf{n}_t + \mathbf{z}_t \odot \mathbf{h}_t, \tag{16}$$

where $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_h, \mathbf{U}_z, \mathbf{U}_r, \mathbf{U}_h$ are learnable weight matrices, $\odot$ represents element-wise multiplication, $\mathbf{n}_t$ is the node state in the t-th iteration, $\mathbf{R}_t$ is the aggregated messages in the t-th iteration. We use the updated nodes between VRG and SRG to compute similarities to update the edges between these two graphs. After message passing for T iterations, we obtain final predictions of classification scores $\mathbf{p}_i$ for the node $\mathbf{n}_i^{vrg}$ from the edges between each node $\mathbf{n}_i^{vrg}$ in VRG and all nodes in SRG.

Table 1. Comparison of our method with other methods on evaluation protocol of ZSD on two splits of MSCOCO. Seen/Unseen is the split of the dataset. The recall is evaluated by selecting the top 100 detections over different IoU thresholds from 0.4 to 0.6. For mAP, we use the VOC metric which only requires IoU of 0.5.

Method	Seen/Unseen	Recall@100			mAP
		0.4	0.5	0.6	0.5
SB [2]	48/17	34.46	22.14	11.31	0.32
DSES [2]	48/17	40.23	27.19	13.63	0.54
TD [25]	48/17	45.60	34.30	18.10	-
PL [38]	48/17	-	43.59	-	10.10
BLC [54]	48/17	51.33	48.87	45.03	10.60
Ours	48/17	**58.51**	**55.03**	**50.28**	**11.40**
PL [38]	65/15	-	37.72	-	12.40
BLC [54]	65/15	57.23	54.68	51.22	14.70
Ours	65/15	**65.27**	**62.70**	**58.28**	**14.90**

4.5. Training and Inference

It is worth noting that our proposed GRAN can be easily embedded in various detection models, including one-stage and two-stage models. In this study, we choose the classic two-stage detection framework Faster R-CNN. Our approach only needs one-step end-to-end training.

Training. The loss function L of our model contains two parts: bounding box regression loss L_{reg} and classification loss L_{cls}. The final prediction of the model can be represented as a two-tuple $(\mathbf{p}_i, \mathbf{b}_i)$, where $\mathbf{p}_i$ is the classification score and $\mathbf{b}_i$ is the coordinates of the box. Similarly, the ground truth is $(\mathbf{p}_i^*, \mathbf{b}_i^*)$. The loss function for an image is written as:

$$L = \frac{1}{N_{cls}}\sum_i L_{cls}(\mathbf{p}_i, \mathbf{p}_i^*) + \lambda\frac{1}{N_{reg}}\sum_i L_{reg}(\mathbf{b}_i, \mathbf{b}_i^*), \tag{17}$$

where L_{cls} uses cross entropy loss and L_{reg} uses the smooth L1 loss, N_{cls} is the number of RoIs for classification and N_{reg} is the number of RoIs for regression. λ is set to balance the two losses.

Inference. For seen classes, we can directly obtain classification score same as training. For unseen classes, we follow [39] to use an additional procedure similar to ConSE [35] during inference. The process can be briefly demonstrated as follows:

$$\mathbf{p}_{unseen} = (\mathbf{p}_{seen}\mathbf{d}_{seen}^T)\mathbf{d}_{unseen}, \tag{18}$$

where $\mathbf{p}_{seen}, \mathbf{p}_{unseen}$ represent classification scores of seen

Table 2. Comparison of our method with other methods on evaluation protocol of GZSD on two splits of MSCOCO. Seen/Unseen is the split of the dataset. The IoU threshold is set to 0.5 on evaluation of recall and mAP. HM means the harmonic mean of mAP of seen and unseen.

Method	Seen/Unseen	Seen		Unseen		HM	
		Recall	mAP	Recall	mAP	Recall	mAP
DSES [2]	48/17	15.02	-	15.32	-	15.17	-
PL [38]	48/17	38.24	35.92	26.32	4.12	31.18	7.39
BLC [54]	48/17	57.56	42.10	46.39	4.50	51.37	8.20
Ours	48/17	**66.70**	**43.90**	**54.54**	**4.70**	**60.01**	**8.50**
PL [38]	65/15	36.38	34.07	37.16	12.40	36.76	18.18
BLC [54]	65/15	56.39	36.00	51.65	13.10	53.92	19.20
Ours	65/15	**65.31**	**38.10**	**60.52**	**13.90**	**62.82**	**20.40**

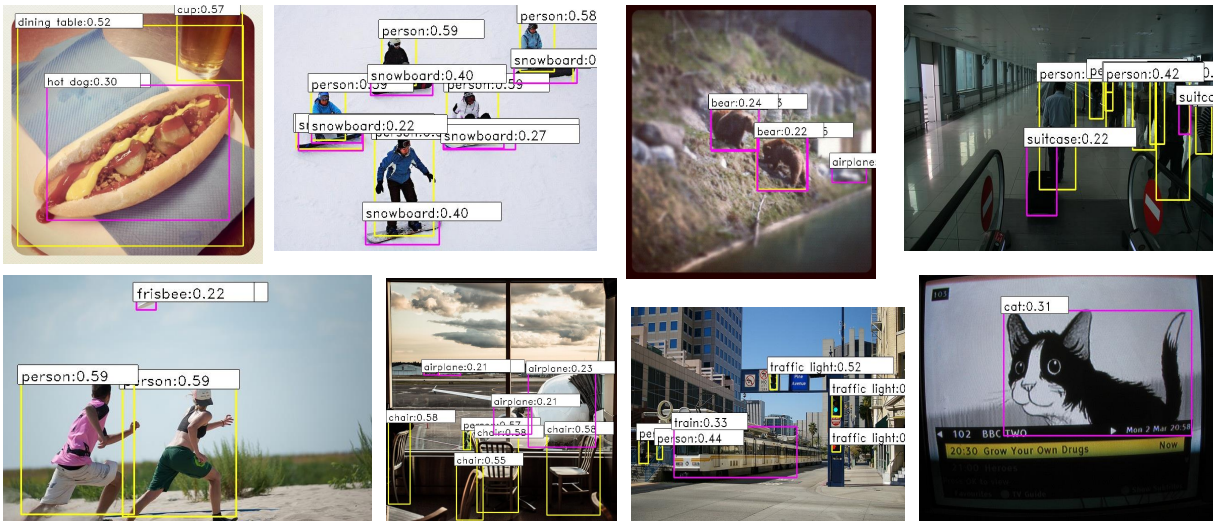


Figure 6. Qualitative examples for testing set detection results of our method. Yellow and pink boxes refer to detections of seen and unseen classes respectively.

and unseen classes respectively, $\mathbf{d}_{seen}, \mathbf{d}_{unseen}$ represent word embeddings of seen and unseen classes respectively.

5. Experiments

5.1. Datasets and Splits

We validate our proposed method on the widely used object detection dataset MSCOCO [28], which includes 82,783 training images and 40,504 validation images of 80 classes. Following the dataset splits of MSCOCO proposed in [2] and [38], we use both two splits of the dataset in experiments: (1) 48 seen classes and 17 unseen classes; (2) 65 seen classes and 15 unseen classes. Note that the seen classes and unseen classes are disjoint.

5.2. Evaluation Protocol

The results of the evaluation protocol on ZSD and GZSD in two splits of MSCOCO are reported. We measure the performance with metrics of recall and mAP. Only the top 100 detections are valid for evaluation.

Table 3. Ablation study of our method in different splits. S/U is seen and unseen split. FSR means Faster Semantic R-CNN which is our baseline model.

S/U	FSR	GRAN	mAP/Recall		
			Seen	Unseen	HM
48/17	✓		44.6/65	4.5/52	8.2/58
	✓	✓	43.9/67	4.7/55	8.5/60
65/15	✓		37.7/62	13.6/58	20.0/60
	✓	✓	38.1/65	13.9/61	20.4/63

5.3. Implementation Details

ResNet-50 is used as the default backbone network with FPN [26]. All experiments are trained with 4 TITAN RTX GPUs (two images for each GPU) for 12 epochs. For optimization, SGD optimizer is applied with momentum is

Table 4. This table shows class-wise recall@100 of unseen classes for two splits of MSCOCO with IoU threshold is 0.5.

48/17 split of MS-COCO

Category	bus	dog	cow	elephant	umbrella	tie	skateboard	cup	knife	cake	couch	keyboard	sink	scissors	airplane	cat	snowboard
Recall	84.1	91.7	82.2	84.6	0.4	0.0	41.1	58.0	43.4	50.5	78.3	34.3	24.3	37.5	60.9	83.4	72.5

65/15 split of MS-COCO

Category	airplane	train	parking meter	cat	bear	suitcase	frisbee	snowboard	fork	sandwich	hot dog	toilet	mouse	toaster	hair drier
Recall	70.4	79.8	21.8	96.9	94.4	62.3	54.9	72.8	48.6	85.6	73.3	69.0	17.5	56.4	4.1

0.9 and weight-decay is 0.0001. The warming-up trick is used to avoid over-fitting in our experiments. The learning rate is set to 0.01 and decreased by 0.1 after 8 and 11 epochs. The L_2 normalized 300 dimensional unsupervised Word2Vec [33], which is trained from large corpora like Wikipedia, is adopted as the semantic representation of MSCOCO classes. Our model is implemented with PyTorch [37] and MMDetection [5] codebase [1].

5.4. Quantitative Results

ZSD Evaluation. We compare GRAN with the state-of-the-art zero-shot detection approaches on both 48/17 [2] and 65/15 [38] splits of MSCOCO in Tab.1. For the 48/17 split, we compare our method with SB [2], DSES [2], TD [25], PL [38], BLC [54]. Our method outperforms all of them in recall@100 and mAP by a significantly large margin. For the split of 65/15, we beats PL [38] and BLC [54], bringing up to 14.7% and 1.4% gains on recall@100 (IoU=0.5) and mAP respectively.

GZSD Evaluation. The GZSD task, which is more difficult and realistic, requires detecting seen and unseen classes at the same time. Same as ZSD evaluation, we compare our method with DSES [2], PL [38], BLC [54] in GZSD setting on two different seen/unseen splits of MSCOCO and report the results in Tab.2. Note that the IoU threshold for recall and mAP is set to 0.5. HM means the harmonic mean of mAP of seen and unseen. It can be observed that our method surpasses other methods on recall and mAP metrics. Class-wise recall@100 of unseen classes can be seen in Tab.4.

Ablation Study. We conduct a controlled study of our proposed method on GZSD evaluation. As shown in Tab.3, the baseline method Faster Semantic R-CNN gives a good foundation and achieves comparable mAP and excellent recall compared with others in Tab.2. Our method is able to consistently improve on both seen and unseen categories.

[1]Source codes are available at http://vipl.ict.ac.cn/resources/codes or https://github.com/witnessai.

5.5. Qualitative results

We show extensive qualitative results in Fig.6. Although these images have various complex natural scenes with multiple objects, our method can simultaneously detect both seen and unseen objects well. These examples suggest effectiveness of utilizing contextual information: the large *airplane* help detect the small *airplane*; the *hot dog* help detect the occluded *cup*. Additionally, our method has good generalization ability of detection even for cartoon images (the bottom right image in Fig.6). More cases can be seen in our supplementary materials.

6. Conclusion

In this work, we find that contextual information is more important in zero-shot detection than in traditional object detection, hence we propose a zero-shot detection method to jointly utilize contextual information containing both visual and semantic relational information. We formulate a visual-semantic relational graph to comprehensively consider these information. A graph aligning network is used to reason and update representation of nodes in the graph with more abundant information for the downstream recognition. Experiments show that visual and semantic relational information are useful for zero-shot detection targeting at natural scenes with multiple objects. Promising results on MSCOCO dataset indicate the potential of the proposed method to be applied in more challenging scenarios with larger detection datasets.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009, and National Key R&D Program of China (2020AAA0105200). Besides, we also thank Chen He and Fengyuan Yang for many helpful discussions.

References

[1] Bogdan Alexe, Nicolas Heess, Yee Teh, and Vittorio Ferrari. Searching for objects driven by context. In *NeurIPS*, 2012.

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018.

[3] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, 2017.

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *EMNLP*, 2014.

[8] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[11] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, 2018.

[12] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, 2016.

[13] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[14] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[15] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[17] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[19] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

[20] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

[21] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018.

[22] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *ICCV*, 2019.

[23] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[24] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *ICLR*, 2016.

[25] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *AAAI*, 2019.

[26] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[30] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, 2018.

[31] Ruotian Luo, Ning Zhang, Bohyung Han, and Linjie Yang. Context-aware zero-shot recognition. In *AAAI*, 2020.

[32] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, 2017.

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[35] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2013.

[36] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *CVPR*, 2019.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[38] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, 2020.

[39] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, 2018.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[41] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *CVPR*, 2017.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[43] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *ICONIP*, 2018.

[44] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection - SNIP. In *CVPR*, 2018.

[45] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, 2017.

[46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.

[47] A Torralba, K Murphy, and WT Freeman. Using the forest to see the trees: Object recognition in context. *Comm. of the ACM*, 2010.

[48] Antonio Torralba, Kevin P Murphy, William T Freeman, and Mark A Rubin. Context-based vision system for place and object recognition. In *ICCV*, 2003.

[49] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-Fast-RCNN: Hard positive generation via adversary for object detection. In *CVPR*, 2017.

[50] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[51] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *ECCV*, 2020.

[52] Xingyu Zeng, Wanli Ouyang, Bin Yang, Junjie Yan, and Xiaogang Wang. Gated bi-directional cnn for object detection. In *ECCV*, 2016.

[53] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. Gtnet: Generative transfer network for zero-shot object detection. In *AAAI*, 2020.

[54] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *ACCV*, 2020.

[55] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero shot detection. *IEEE TCSVT*, 2019.

[56] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, 2020.