Reasoning-RCNN: Unifying Adaptive Global Reasoning into Large-scale Object Detection

Hang Xu[1*] ChenHan Jiang[2*] Xiaodan Liang[2†] Liang Lin[2] Zhenguo Li[1]
[1]Huawei Noah's Ark Lab [2]Sun Yat-sen University

Abstract

In this paper, we address the large-scale object detection problem with thousands of categories, which poses severe challenges due to long-tail data distributions, heavy occlusions, and class ambiguities. However, the dominant object detection paradigm is limited by treating each object region separately without considering crucial semantic dependencies among objects. In this work, we introduce a novel Reasoning-RCNN to endow any detection networks the capability of adaptive global reasoning over all object regions by exploiting diverse human commonsense knowledge. Instead of only propagating the visual features on the image directly, we evolve the high-level semantic representations of all categories globally to avoid distracted or poor visual features in the image. Specifically, built on feature representations of basic detection network, the proposed network first generates a global semantic pool by collecting the weights of previous classification layer for each category, and then adaptively enhances each object features via attending different semantic contexts in the global semantic pool. Rather than propagating information from all semantic information that may be noisy, our adaptive global reasoning automatically discovers most relative categories for feature evolving. Our Reasoning-RCNN is light-weight and flexible enough to enhance any detection backbone networks, and extensible for integrating any knowledge resources. Solid experiments on object detection benchmarks show the superiority of our Reasoning-RCNN, e.g. achieving around 16% improvement on VisualGenome, 37% on ADE in terms of mAP and 15% improvement on COCO.

1. Introduction

The large-scale detection [18] refers to recognize and localize a large number of categories. The severe imbalance of categories is very common in those tasks (e.g. very few samples for rare classes). Moreover, due to more categories within one image, problems of heavy occlusion, class ambiguities and tiny-size objects become more challenging. Current state-of-the-art object detection methods [3, 8, 43] treat the recognition of each region separately and thus require high-quality feature representations for each region and sufficient labeled data for each category. However, this is not the case for the large-scale detection problem and so such method are inappropriate. Unlike humans who are able to identify objects even in complex situations, current detection systems lack the ability to reason with the help of commonsense knowledge. Therefore, a key issue is how to endow the current detection system with the power of reasoning, in order to mimic human reasoning procedure.

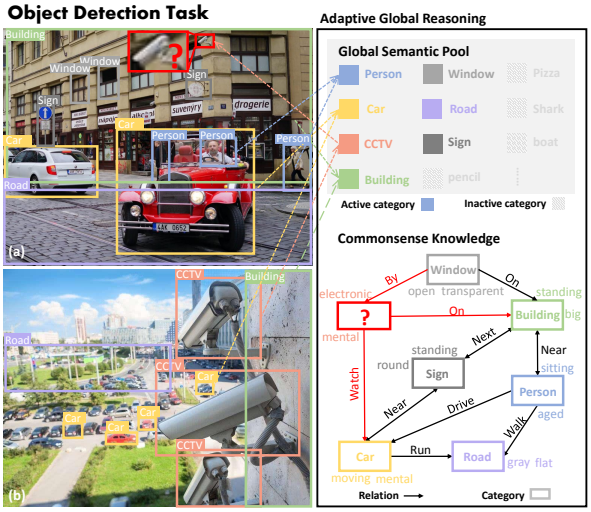

Figure 1. An example of how the proposed adaptive global reasoning can facilitate large-scale object detection, especially for rare and tiny categories. Human can still recognize the tiny object in red frame in (a) as a "CCTV" . This is because: 1) this object looks familiar to the "CCTV" in human memory as we saw before in (b), which inspires our design of Global Semantic Pool; 2) this small electronic mental object is installed on the building and is watching the car running on the road. Thus, it is likely to be a CCTV. Such rich human commonsense can be represented in a knowledge graph and incorporated into our detection pipeline.

*Both authors contributed equally to this work.

†Corresponding Author: xdliang328@gmail.com

Codes and trained model can be found in https://github.com/chanyn/Reasoning-RCNN.

When identifying an object in a scene, reasoning by human common sense can help to make a correct recognition. An example of global knowledge reasoning in Figure 1 would be to identify the tiny-size objects "CCTV" in the red frame of upper image (a). Human will first search his memory for the object in the brain for similar appearance categories (inspiring the design of our Global Semantic Pool), then he will reason by considering global semantic coherency: this small electronic mental object is installed on the building and is watching the car running on the road and thus it is more likely to be a CCTV. Such rich human commonsense can be represented in a knowledge graph and incorporated by visual reasoning in the detection pipeline.

Recent works on visual reasoning can be categorized by different strategies of incorporating knowledge: methods that rely on human prior knowledge and methods that do not. For instance, some works model spatial relationship implicitly from the image itself [5, 19, 47]. These works learn inter-region relationships in an implicit and uncontrollable way thus their boost of performance is limited. Other methods try to incorporate human semantic prior knowledge by defining knowledge graphs in the networks [6, 36]. For example, recently an iterative reasoning approach [6] was proposed to combine both spatial and semantic relationship reasoning. However, they only consider propagating region-wise feature locally in one image by a fixed prior knowledge. In other words, their method will still fail to reason through a bad feature representation when heavy occlusions and class ambiguities exist in the image which is very common in large-scale detection. Furthermore, they used a very complicated structure with three reasoning modules stacked together by GRU. On the contrary, our work aims to develop an in-place and simple global reasoning network which can not only explicitly incorporate multiple kinds of commonsense knowledge but also propagate visual information globally from all the categories to refine both classification and bounding box regression.

In this paper, we propose a novel Reasoning-RCNN network to endow any detection networks with the capability of adaptive global reasoning to exploit diverse human commonsense knowledge. Unlike some existing works which only propagate the visual features on the image directly, we globally evolve the high-level semantic representations of all categories to avoid distracted or poor visual features in the image. To achieve this, our method first generates a global semantic pool over all the categories by collecting the weights of previous classification layer. Note that this avoids the computational burden in contrast to traditional methods [27] that take average or clustering all over the data for each category. Then a category-wise knowledge graph is designed to encode certain linguistic knowledge (e.g. attributes, co-occurrence, and relationships). High-level semantic contexts of different categories in the global semantic pool are evolved and propagated according to the connected nodes in the knowledge graph being considered. Rather than propagating information from all semantic information that may be noisy, our adaptive global reasoning further encodes the current image adaptively by an attention mechanism [46] to automatically discover most relevant categories for feature evolving regarding each object. Next, the enhanced categories contexts are mapped back to the regions by a soft-mapping mechanism which enables refinement of inaccurate classification results from previous stage. Finally, each region's new enhanced features are concatenated to the original features to improve the performance of both classification and localization in an end-to-end manner. We experiment with two kinds of knowledge forms in this work: relation knowledge such as co-occurrence and object-verb-subject relationship, and the attribute knowledge (e.g. color, status).

Our Reasoning-RCNN thus enables adaptive global reasoning over categories with certain relations or similar attributes. Recognition of difficult regions with heavy occlusions, class ambiguities and tiny-size problems can thus be remedied by the enhanced features which contains the adaptive context from the global semantic pool. Moreover, the problem of imbalanced categories can then be alleviated by sharing and distilling essential characteristics among frequent/rare categories.

The proposed Reasoning-RCNN outperforms current state-of-the-art detection methods, including Faster R-CNN [43], RetinaNet [31], RelationNet [19], and DetNet [29]. We observe consistent gains on the base detection network Faster R-CNN on object detection benchmarks, i.e., VG (1000/3000 categories), ADE (445 categories), MS-COCO (80 categories), and Pascal VOC (20 categories). In particular, Reasoning-RCNN achieves around 15% of mAP improvement on VG (1000 categories), 16% on VG (3000 categories), 37% on ADE, 15% on MS-COCO, and 2% on Pascal VOC.

2. Related Work

Object Detection. Object detection is a core problem in computer vision. Significant progress has been made in recent years on object detection task using CNN. Modern object detection methods may be categorized in two groups: one-stage detection methods such as SSD [33] and YOLO [41] and two-stage detection methods such as Faster R-CNN [43] and R-FCN [8]. Usually very few categories are considered: 20 for PASCAL VOC [9] and 80 for COCO [32]. These methods are performed separately for each proposal region without any reasoning.

Visual Reasoning. Visual reasoning is intended to combine different information or interactions between objects or scenes. Examples can be found in the task of classification [36], object detection [6] and visual relationship detec-

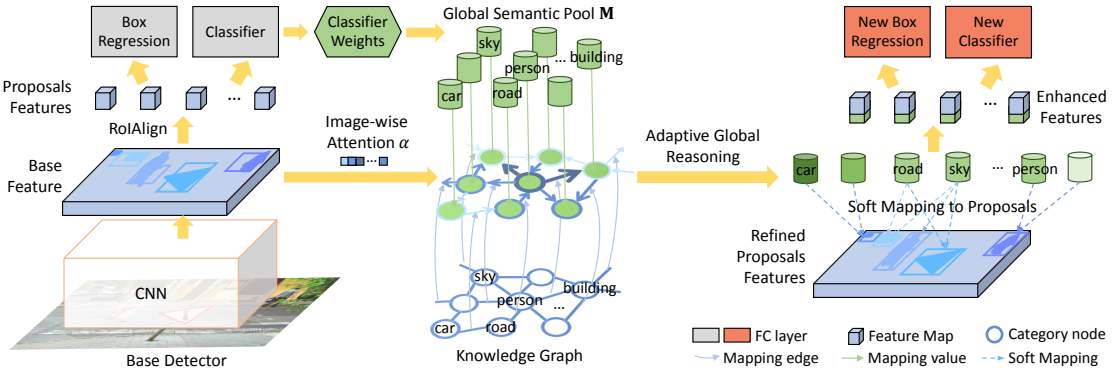

Figure 2. The overview of the proposed Reasoning-RCNN object detection framework. The proposed Reasoning RCNN can be stacked on any existing base detector such as Faster RCNN. The weights of the previous classifier are collected to generate a global semantic pool over all categories, which is fed into our adaptive global reasoning module. The enhanced category contexts (i.e., output of the reasoning module) are mapped back to region proposals by a soft-mapping mechanism. Finally, each region's enhanced features are used to improve the performance of both classification and localization in an end-to-end manner.

tion [7]. Several aspects such as relationships and shared attributes among objects can be considered [1, 2, 26, 38]. Other methods [11, 35, 42] rely on finding similarity such as the attributes in the linguistic space. For incorporating relationship, most early works use object relations as a post-processing step [10, 13, 39]. Recent works consider a graph structure [6, 7, 24, 36] to incorporate knowledge. In these works, they usually consider region-wise reasoning which will fall to reasoning through a bad feature representation when heavy occlusions and class ambiguities exist in the image which is common in large-scale detection. Our method propagates over all categories to avoid this problem and ensure an adaptive global reasoning.

Few-shot Recognition. Few-shot recognition aims to understand a new concept with a few annotated examples, which shares the similar objectives with us. Early works focus on learning attribute embedding to represent categories [22, 26]. Most recent works use knowledge graph such as WordNet [37] to distill information among categories [36, 48, 49]. Gidaris et al. [14] made use of distilling classifier weights to help few-shot task. In contrast, our module benefits from a dynamically updated global semantic pool and explicit prior knowledge.

3. The Proposed Approach

3.1. Overview

In this paper, we introduce Reasoning-RCNN to develop a general model with adaptive global reasoning by incorporating distinct external knowledge to facilitate large-scale object detection. An overview of our Reasoning-RCNN can be found in Figure 2. The proposed Reasoning-RCNN can be stacked on any one/two-stage modern detection framework. More specifically, we first create a global semantic pool to integrate high-level semantic representation for each category by collecting the weights of original classification layer. Then a category-to-category undirected graph $\boldsymbol{G}$: $\boldsymbol{G} =< \mathcal{N}, \mathcal{E} >$ is defined and shared during training and testing, where $\mathcal{N}$ are category nodes and each edge $e_{i,j} \in \mathcal{E}$ encodes a kind of knowledge between two nodes. The region features can be enhanced by propagating semantic contexts over global semantic pool with a particular knowledge graph $\boldsymbol{G}$. Finally, the enhanced features concatenated with the original features are fed into the bounding box regression layer and classification layer to obtain better detection results.

3.2. Adaptive Global Graph Reasoning Module

Our Reasoning-RCNN can be added to any modern dominant detection system for endowing its ability in global reasoning. An overview of our adaptive global reasoning module can be found in Figure 3. Let $\mathbf{f} = \{f_i\}_{i=1}^{N_r}, f_i \in \mathbb{R}^D$ be the visual features of D dimension extracted from the backbone network for all $N_r = |\mathcal{N}|$ region proposals. Our method aims to enhance the original region features $\mathbf{f}$ by exploiting certain commonsense knowledge forms such as pairwise relationship knowledge (e.g. "man rides bicycles") or some kinds of attribute knowledge (e.g. "apple is red."). Specifically, our global reasoning stage evolves the visual object reference in the global semantic pool according to the category-to-category knowledge graph $\boldsymbol{G}$. Attention mechanism is also implemented in order to automatically emphasis more informative and relative categories on each image to enable adaptive global reasoning. $\mathbf{f}$ is then enhanced by the evolved features to improve the performance of both classification and localization.

3.2.1 Global Semantic Pool M

Most existing works [15, 6, 23] usually propagate visual features locally among regions. However, this diagram could lead to failure of graph reasoning because of bad

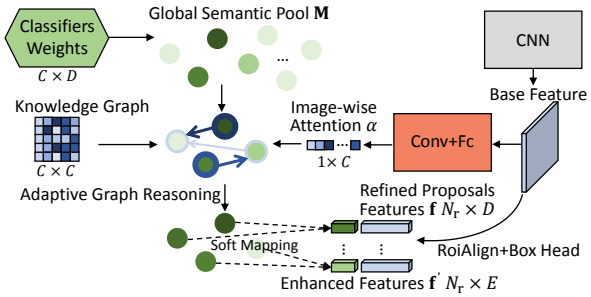

Figure 3. An overview of our adaptive global reasoning module. Global semantic pool $\mathbf{M}$ integrates all high-level semantic representation by the weights of previous detection classifiers for each category. Global reasoning is then performed by propagating all semantic representation in $\mathbf{M}$ according to certain type of knowledge graph. Adaptive Attention is calculated using the image features to automatically discovers most relative categories for adaptive reasoning. The enhanced features are soft-mapped from categories to the proposals to get the region-wise enhanced features $\mathbf{f}'$. Finally, $\mathbf{f}'$ concatenated with region proposal features $\mathbf{f}$ are fed into new bbox regression layer and classification layer to obtain better detection results.

or distracted feature representations when heavy occlusions and class ambiguities exist in the image which is common in large-scale detection. Instead, our method try to propagate information globally over all the categories (not only those categories appearing in the image). To achieve this, we need to create a global semantic pool to store high-level semantic representations for all categories. This is analogous to the memory in the human's brain when human recalls the appearance of one particular category.

To generate this kind of global semantic pool, existing works usually take the average of the features or using the clustering method to find the center as the reference features [27] for each category. However, those approaches record and collect all the information across the whole data which is a huge computational burden. Moreover, those models cannot be trained in an end-to-end style. Inspired by some works in zero/few-shot problem in which they try to train a model to fit the weights of the classifier of an unseen/unfamiliar category [45, 48, 15, 14], we introduce a new way to generate the global semantic pool. The weights of the classifier for each categories actually contains high-level semantic information since they record the feature activation trained from all the images. Formally, let $\mathbf{M} \in \mathbb{R}^{C \times D}$ denote the weights of the previous classifiers (parameters) for all the C categories. The global semantic pool of our model can be obtained by copying the parameters $\mathbf{M}$ from the previous classification layer in the bbox head of the detection networks. Note that the classifiers are updated in each iteration during training so that the global semantic pool $\mathbf{M}$ becomes more accurate from time to time. Furthermore, our model can be trained in an end-to-end style.

3.2.2 Feature Enhanced via Graph Reasoning.

After creating a global semantic pool $\mathbf{M} \in \mathbb{R}^{C \times D}$ for all the C categories, it is natural to propagate the connected categories of $\mathbf{M}$ by the edges $\mathcal{E} \in \mathbb{R}^{\mathcal{C} \times \mathcal{C}}$ in the prior knowledge graph $\boldsymbol{G}$. Thus, the information is shared and propagate globally across all the C categories according to the chosen knowledge denoted as $\mathcal{E}\mathbf{M}$. To enhance features of the regions, we still need to find the mapping between the N_r region proposals and C categories. Intuitively this mapping can be easily obtained from the classification results in the previous stage of the detection networks. Instead of a *hard-mapping* directly from region proposal to categories, we propose a method of *soft-mapping* which is the classification probability distribution $\mathbf{P} \in \mathbb{R}^{N_r \times C}$ over all the C categories. $\mathbf{P} \in \mathbb{R}^{N_r \times C}$ can be calculated by soft-max function over the scores of C categories from previous classifiers. Then this process can be solved by matrix multiplication: $\mathbf{P}\mathcal{E}\mathbf{M}\boldsymbol{W}_G$, where $\boldsymbol{W}_G \in \mathbb{R}^{D \times E}$ is a transformation weight matrix shared for all graphs and E is the output dimension of the reasoning module. Note that the global graph reasoning is based on all categories, which may be noisy. An adaptive reasoning mechanism is needed to incorporate visual patterns of each particular image. That's why we further introduce attentional adaptive reasoning.

3.2.3 Adaptive Attention

Given the evolved global features $\mathcal{E}\mathbf{M}$, we need to emphasize informative and relative categories and suppress less useful ones, thus enables adaptive reasoning for each image. It can be noticed that not all the information of classes is useful for recognizing items in one particular image. Humans only consider several potential categories when identifying items in one scene. In this paper, we make use of the idea of *Squeeze-and-Excitation* [20] to further re-scale the categories being considered. Specifically, in the *squeeze* step, we take the whole image feature $\mathbf{F} \in \mathbb{R}^{W \times H \times D}$ as input and squeeze it into half size by a CNN (with 3×3 kernel, output channel=$D/64$) and a global pooling operation. The *excitation* stage is a fully-connected layer with input $\boldsymbol{z}_s \in \mathbb{R}^{D/64}$. Then a soft-max function is applied to obtain the attention of categories: $\alpha = softmax(\boldsymbol{z}_s \boldsymbol{W}_s \boldsymbol{M}^T)$, where $\boldsymbol{W}_s \in \mathbb{R}^{D/64 \times D}$ is the weights of the fully-connected layer and $\alpha \in \mathbb{R}^C$. Then the enhanced features $\mathbf{f}'$ with adaptive reasoning can be solved by:

$$\mathbf{f}' = \mathbf{P}\left(\alpha \otimes \mathcal{E}\mathbf{M}\right)\boldsymbol{W}_G, \tag{1}$$

where $\otimes$ is the channel-wise product and the rest is the matrix multiplications. $\mathbf{f}' \in \mathbb{R}^{N_r \times E}$ are the enhanced features with E dimension via adaptive global graph reasoning. Flowchart of the adaptive global reasoning can be found in

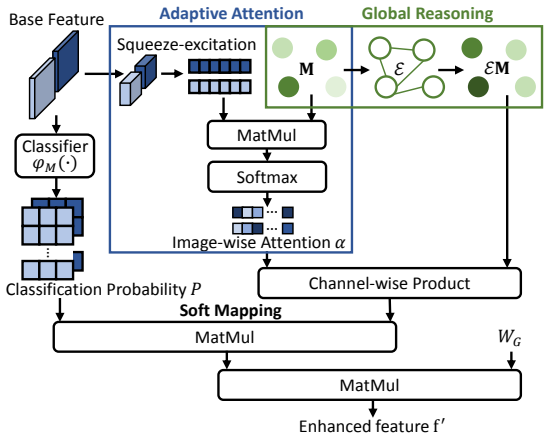

Figure 4. Detailed flowchart of the adaptive global reasoning. The global reasoning is performed on global semantic pool $\mathbf{M}$ according to the prior knowledge graph edge $\mathcal{E}$. An image-wise adaptive attention α is calculated from a Squeeze-and-Excitation of the image base features to emphasize relative categories. Then the adaptive global reasoning with α is obtained by the channel-wise product. Then a soft-mapping from categories to regions is performed according to P. Final enhanced feature $\mathbf{f}'$ is obtained by a matrix multiplication with a fully-connected weights W_G.

Figure 4. Finally, the enhanced features $\mathbf{f}'$ will be concatenated to the original region features $\mathbf{f}$. $[\mathbf{f}; \mathbf{f}']$ will be fed into the bounding box regression layer and classification layer to obtain final detection results. Note that the $\mathbf{f}'$ is a distilled information across the categories with connected edges such as similar attributes or relations. Thus, the problem of few training samples is alleviated by sharing the common features between similar categories. Those proposal regions with heavy occlusions, class ambiguities and the tiny-size problem can also be remedied by adding and discovering adaptive contexts from the global semantic pool guided by external knowledge.

3.3. Model Specification with Relation Knowledge.

The Reasoning-RCNN is flexible enough to incorporate arbitrary types of knowledge. Here, we take relationship knowledge as an example to illustrate how different commonsense knowledge $\boldsymbol{G}$ is specified to obtain distinct graph reasoning behaviors. We also explore another type of knowledge, i.e. the attribute knowledge in the Experiments.

Relationship knowledge $\boldsymbol{G}^{\boldsymbol{R}}$ as one kind of $\boldsymbol{G}$ denotes the pairwise relationship between categories, such as the "subject-verb-object" relationship (e.g. *drive*, *run*), spatial relationship (e.g. *on*, *near*). The global semantic pool will be enhanced with high-level semantic correlations between categories. First, we calculate a $C \times C$ frequent statistics matrix $\boldsymbol{R}^c$ either from the semantic information or simply from the occurrence among all categories pairs. Then, we add the transpose $(\boldsymbol{R}^c)^T$ back to $\boldsymbol{R}^c$. Finally, a column-row normalization is performed to get $\boldsymbol{G}^{\boldsymbol{R}}$: $e^R_{ij} = \frac{R^c_{ij}}{\sqrt{D_{ii}D_{jj}}}$, where $D_{ii} = \sum_{j=1}^{C} R^c_{ij}$. Note that we already include a lot of spatial relationships such as "along", "on", and "near" and that's why we don't consider spatial relationship separately in this paper.

4. Experiments

Dataset and Evaluation. Experiments on Reasoning-RCNN have been conducted on large-scale object detection benchmarks with a large number of classes: Visual Genome (VG) [25], ADE [50]. Additionally, we also evaluate on PASCAL VOC 2007 [9] and MSCOCO 2017 [32] to show the performance on common categories (20/80 categories). The task is to localize an object and classify it, which is different from the experiments with given ground truth locations [6]. For Visual Genome, we use the latest release (v1.4), and synsets [44] instead of the raw names of the categories due to inconsistent label annotations, following [21]. We consider two sets of target classes: 1000 most frequent categories and 3000 most frequent categories, resulting in two settings VG_{1000} and VG_{3000}. We split the remaining 92.9K images with objects on these class sets into 87.9K and 5K for training and testing, respectively. In term of ADE dataset, we use 20.1K images for training and 1K images for testing, following [6]. To validate the generalization capability of models, 445 classes that overlap with VG dataset are selected as targets. Since ADE is a segmentation dataset, we convert segmentation masks to bounding boxes for all instances. We also evaluate our Reasoning-RCNN on PASCAL VOC 2007 (20 categories) and MSCOCO 2017 (80 categories) which is prepared following the same protocols in [8]. For PASCAL VOC, training is performed on the union of VOC 2007 trainval and VOC 2012 trainval (10K images) and evaluation is on VOC 2007 test (4.9K images). MSCOCO 2017 contains 118k images for training, 5k for evaluation.

For VG, ADE, COCO evaluation, we adopt the metrics from COCO detection evaluation criteria [32] which is mean Average Precision(mAP) across IoU thresholds from 0.5 to 0.95 with an interval of 0.05 and Average Recall (AR) with different number of given detection per image ($\{1, 10, 100\}$) and different scales (small, medium, big). For PASCAL VOC, we only report mAP scores using IoU thresholds at 0.5 for the purpose of comparison with other existing methods.

Knowledge Graph Construction. We apply general knowledge graphs for experiments on all datasets. A common sense knowledge graphs is generated with the help of the statistics of the annotations in the VG dataset. Specifically, for relationship knowledge graph $\boldsymbol{G}^{\boldsymbol{R}}$, we use top 200 most frequent relationship annotations in VG such as location relationship, subject-verb-object relationship, and

%	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L	#. param
VG_{1000}	Light-head RCNN[28]	6.2	10.9	6.2	2.8	6.5	9.8	14.6	18.0	18.7	7.2	17.1	25.3	74.5M
	Cascade RCNN[3]	6.5	12.1	6.1	2.4	6.9	11.2	15.3	19.4	19.5	6.1	19.2	27.5	91.2M
	Faster-RCNN[43]	6.2	11.3	5.9	1.8	5.9	11.4	14.2	17.8	17.8	4.4	16.1	27.6	54.1M
	Reasoning-RCNN$_R$	$8.1^{+1.9}$	**13.5**$^{+2.2}$	**8.5**$^{+2.6}$	$3.4^{+1.6}$	$8.3^{+2.4}$	**14.0**$^{+2.6}$	**18.6**$^{+4.4}$	**23.4**$^{+5.6}$	**23.5**$^{+5.7}$	$8.8^{+4.4}$	$21.7^{+5.6}$	**32.9**$^{+5.7}$	57.5M
	Faster-RCNN w FPN[30]	7.1	12.7	7.2	3.9	7.6	11.1	14.8	19.7	19.9	10.6	18.8	24.9	61.4M
	Reasoning-RCNN$_R$ w FPN	**8.2**$^{+1.1}$	$13.3^{+0.6}$	**8.5**$^{+1.3}$	**4.4**$^{+0.5}$	**8.9**$^{+1.3}$	$12.9^{+1.8}$	$16.4^{+2.4}$	$22.2^{+2.5}$	$22.5^{+2.4}$	**12.3**$^{+1.7}$	**22.1**$^{+3.3}$	$27.1^{+2.2}$	63.6M
VG_{3000}	Light-head RCNN[28]	3.0	5.1	3.2	1.7	4.0	5.8	7.3	9.0	9.0	4.3	10.3	15.4	78.6M
	Cascade RCNN[3]	3.8	6.5	3.4	1.9	4.8	4.9	7.1	8.5	8.6	4.2	9.9	13.7	97.3M
	Faster-RCNN[43]	3.7	6.4	3.8	1.7	4.6	7.6	8.5	10.5	10.5	4.1	11.6	18.2	58.2M
	Reasoning-RCNN$_R$	**4.5**$^{+0.8}$	**7.3**$^{+0.9}$	**4.7**$^{+0.9}$	$2.2^{+0.5}$	$5.5^{+0.9}$	**9.0**$^{+1.4}$	**10.6**$^{+2.1}$	**12.9**$^{+2.4}$	**12.9**$^{+2.4}$	$5.4^{+1.3}$	**13.8**$^{+2.2}$	**21.9**$^{+3.7}$	65.8M
	Faster-RCNN w FPN[30]	3.7	6.5	3.7	2.1	4.9	6.8	7.6	9.8	9.9	6.8	11.8	14.6	63.4M
	Reasoning-RCNN$_R$ w FPN	$4.3^{+0.6}$	$6.9^{+0.4}$	$4.6^{+0.9}$	**3.2**$^{+1.1}$	**6.0**$^{+1.1}$	$7.9^{+1.1}$	$8.5^{+0.9}$	$11.1^{+1.3}$	$11.2^{+1.3}$	**8.3**$^{+1.5}$	$13.7^{+1.9}$	$16.2^{+1.6}$	68.2M
ADE	Light-head RCNN[28]	7.0	11.7	7.3	2.4	5.1	11.2	9.6	13.3	13.4	4.3	10.4	20.4	72.4M
	Cascade RCNN[3]	9.1	16.8	8.9	3.5	7.1	15.3	12.1	16.4	16.6	6.4	13.8	25.8	89.5M
	Faster-RCNN[43]	8.7	15.5	8.9	3.6	6.9	14.1	11.7	15.9	16.0	6.3	13.9	23.9	52.9M
	Reasoning-RCNN$_R$	$11.5^{+2.8}$	$18.8^{+3.3}$	$11.9^{+3.0}$	$4.6^{+1.0}$	$9.1^{+2.2}$	$18.9^{+4.8}$	$14.8^{+3.1}$	$19.9^{+4.0}$	$19.9^{+3.9}$	$8.2^{+1.9}$	$17.0^{+3.1}$	$30.5^{+6.6}$	55.3M
	Faster-RCNN w FPN[30]	11.3	19.7	11.7	5.9	10.8	17.3	13.4	19.9	20.2	12.9	20.2	28.5	60.8M
	Reasoning-RCNN$_R$ w FPN	**15.5**$^{+4.2}$	**24.6**$^{+4.9}$	**16.3**$^{+4.6}$	**8.8**$^{+2.9}$	**15.5**$^{+4.7}$	**23.5**$^{+6.2}$	**17.5**$^{+4.1}$	**25.9**$^{+6.0}$	**26.6**$^{+6.4}$	**17.2**$^{+4.3}$	**27.6**$^{+7.4}$	**35.4**$^{+6.9}$	62.6M

Table 1. Main results of test datasets on VG_{1000} , VG_{3000} and ADE. "Reasoning-RCNN$_R$" is our full model empowered with relation knowledge. #. parameters is the number of parameters for the model.

%	Method	backbone	#. param (M)	time (ms)	mAP
PASCAL VOC	SMN[5]	ResNet-101	66.7	-	67.8
	R-FCN[8]	ResNet-101	54.0	111.1	80.5
	DSSD513[12]	ResNet-101	-	156.2	81.5
	Faster-RCNN[43]	ResNet-101	52.0	56.4	80.8
	Reasoning-RCNN$_A$	ResNet-101	53.6	58.8	81.9
	Reasoning-RCNN$_R$	ResNet-101	53.6	58.8	82.5
MS COCO	Relation Network[19]	ResNet-101-FPN	62.8	-	38.8
	RetinaNet[31]	ResNet-101-FPN	56.9	200	39.1
	DetNet[29]	DetNet-59-FPN	-	-	40.2
	Faster-RCNN[43]	ResNet-101	52.2	64.9	34.9
	Reasoning-RCNN$_A$	ResNet-101	53.8	69.1	39.2
	Reasoning-RCNN$_R$	ResNet-101	53.8	69.1	40.5
	Faster-RCNN w FPN[30]	ResNet-101-FPN	60.4	73.0	37.3
	Reasoning-RCNN$_R$ w FPN	ResNet-101-FPN	61.5	75.3	42.9
	Mask-RCNN w FPN[16]	ResNet-101-FPN	63.4	86.6	39.4
	Reasoning-RCNN$_R$ w Mask	ResNet-101-FPN	64.5	89.4	**43.2**

Table 2. Comparison of mean Average Precision (mAP) on PASCAL VOC and MSCOCO. "Reasoning-RCNN$_A$"/"Reasoning-RCNN$_R$" are the Faster-RCNN adding our model with Attribute/Relation knowledge.

count frequent statistics of each pair. We match the categories of ADE, COCO and VOC with VG to obtain their corresponding knowledge graph in order to validate the generalization capability of the common knowledge.

We also consider attribute knowledge in this paper. The attribute knowledge graph $\boldsymbol{G^A}$ is defined as the similarity among categories according to their attributes such as colors, size, materials among object categories. We consider the top 200 most frequent attributes annotations in VG such as color, material,and status of the categories ($C = 3000$), and count their frequent statistics as the class-attribute distribution table. Then the pairwise Jensen–Shannon (JS) divergence between probability distributions P_{c_i} and P_{c_j} of two classes c_i and c_j can be measured as the edge weights of two classes : $e^A_{c_i,c_j} = JS(P_{c_i}||P_{c_j})$.

Implementation Details. We treat the state-of-the-art Faster-RCNN with FPN[4, 30] as our baseline and implement Reasoning-RCNN stacked on it in Pytorch[40]. ResNet-101 [17] pretrained on ImageNet [44] is used as our backbone network. Horizontal image-flipping and multi-scaling augmentations are adopted in training. Following [43], RPN is applied to all the feature maps. The parameters before conv1 is fixed, same with [30]. We sampled a minibatch containing 512 region proposals after NMS, each of which is positive if it has an IoU$>$ 0.7 with the ground-truth regions and it is negative if the IoU$<$ 0.3. After ROI Align, proposals features are avg-pooled and feed to 2 shared FC layers to become the input of the final classifier ($D = 1024$). At the testing time, we keep 2000 region proposals after NMS with IoU threshold$>$ 0.6. Hard example mining is not used in the all experiments. Unless otherwise noted, settings are the same for all experiments.

For our reasoning stage stacked on Faster-RCNN with FPN, we use same operation (average global pooling and shared 2 FC layer) for re-extract region proposal visual feature $\mathbf{f}$. The hyperparameter is $E = 256$ of $\boldsymbol{W}_G$ for any knowledge which is considered sufficient to contain the enhanced feature. We apply synchronized SGD with a weight decay of 0.0001 and momentum of 0.9 to optimize all models. The initial learning rate is 0.02, reduce two

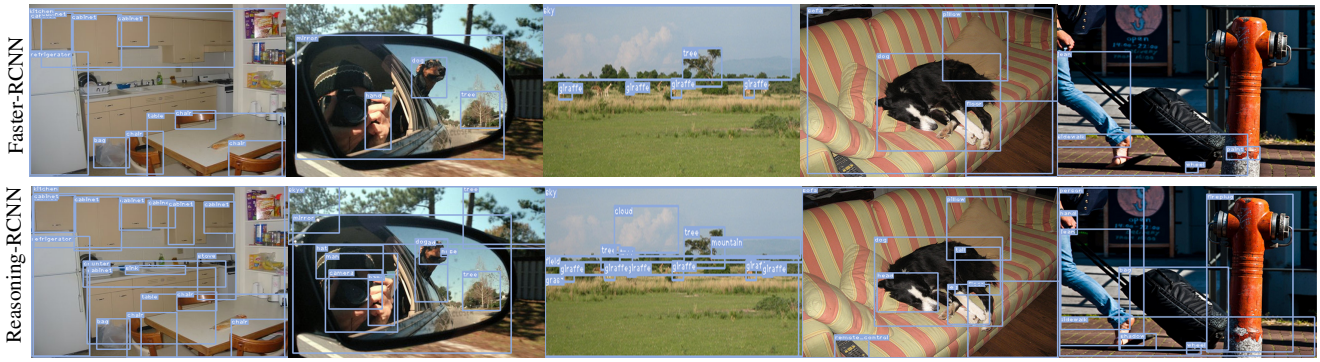

Figure 5. Qualitative result comparison on VG_{1000} between Faster-RCNN and our Reasoning-RCNN. Objects with occlusion, ambiguities and rare category can be detected by our method.

times ($\times 0.01$) after 8 and 11 epochs. We train 32 epochs with mini-batch size of 2 for baseline Faster R-CNN with FPN (Further training after 12 epochs won't increase the performance of baseline.). For Reasoning-RCNN, we use 12 epochs of the baseline as pretrained model and train another 16 epochs with the same settings.

4.1. Comparison with state-of-the-art

We report the result comparisons on VG_{1000} with 1000 categories , VG_{3000} with 3000 categories and ADE dataset in Table 1. We compare with Light-head RCNN[28], Faster-RCNN with FPN [30], three-stage Cascade-RCNN of baseline Faster-RCNN[3] using the public released code. Same settings of hyper-parameters are used for all experiments unless specifically mentioned in our paper. As can be seen, the Reasoning-RCNN with relation knowledge outperforms the baseline Faster-RCNN[43] on all dataset. Our Reasoning-RCNN achieves an overall AP of 8.1% compared to 6.2% by Faster R-CNN on VG_{1000}, 4.5% compared to 3.7% on VG_{3000}, and 11.5% compared to 8.7% on ADE, respectively. Significant performance gap can also be observed for those rare categories with very few samples (See Supplementary materials for details). Moreover, our Reasoning-RCNN achieves significant gains on both classification and localization accuracy than the baseline on all cases (i.e. different scales and overlaps). This verifies the effectiveness of incorporating adaptive global reasoning into local region recognition in large-scale object detection task. Compare to [6], they report a performance gain (AP) of 13% on VG and 20% on ADE compared to the same baseline on the task of only classification based on the ground-truth bounding box. Our method achieves around 30% improvement on VG and 32% on ADE on the harder task of object detection (both localization and classification). Moreover, A negative correlation between the average AP gain of our method with baseline Faster-RCNN and the category frequency is shown in Figure 7.

We also evaluate on PASCAL VOC and MS COCO datasets with only 20/80 categories to compare with the state-of-art methods. The result can be found in Table 2 and the accuracy numbers of the competing methods are directly from the original paper. For PASCAL VOC, we compare with the Spatial Memory Network[5], R-FCN[8] , and DSSD513[12]. As can be seen, our method performs 1.7% better than the baseline Faster-RCNN and all the other competitors. For MS COCO, comparison is made among Faster-RCNN with FPN[30], Relation Network[19], RetinaNet[31] and DetNet[29]. Our method with FPN boosts the mAP from 37.3% to 42.9% and outperform all the other methods. Note that our method can achieve higher performance with auxiliary segmentation task. This demonstrates the Reasoning-RCNN can strongly improve the power of feature representation, due to its ability of global adaptive reasoning. Furthermore, from the comparison of computation cost, the computation overhead is relatively small (less than 2% parameter size and 3 ms) with with input-size of 800×800 pixels on Titan XP for FPN on MS COCO.

Figure 5 shows qualitative results comparison between the baseline model and our Reasoning-RCNN (more examples in Supplementary material.). The results show that the baseline model tends to ignore rare categories and ambiguity objects. For example, it can not detect "stove" in first image and "remote control" in the fourth image. Reasoning-RCNN tend to detect all similar objects, such as "cabinet" in the first image, "giraffe" in the third image. Besides, in the 2nd, 5th image, our Reasoning-RCNN can detect "hat", "man", "hand" and "camera" with obscure and occlusion. More examples and results can be found in supplementary materials.

4.1.1 Generalization capability

From Table 1 and Table 2, the external knowledge graph from VG can actually help to improve the performance of ADE, COCO and PASCAL VOC. Therefore, any datasets with overlap categories can share the existing knowledge graph. Besides, our module can be added to diverse detec-

tion systems easily.

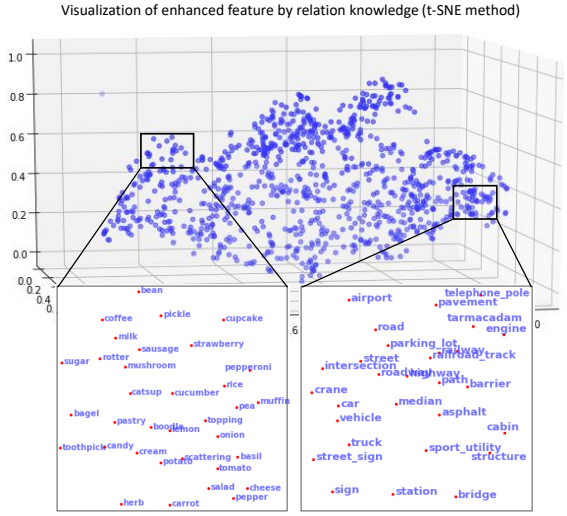

Figure 6. 3-D visualization of $\mathbf{f}'_r$ by t-SNE method [34]. The squared regions are enlarged in bottom panels. The categories with relation are closed to each other. This verifies that our method has global reasoning power over categories according to certain knowledge.

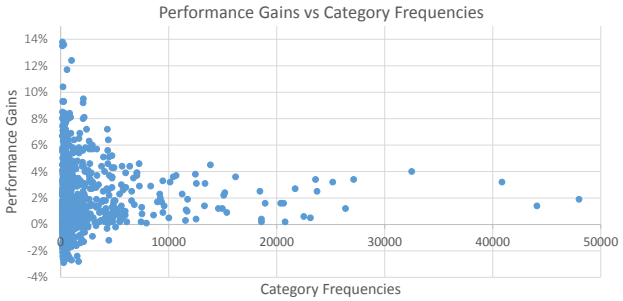

Figure 7. Correlation between performance gains of AP with baseline Faster-RCNN and the category frequencies on VG_{1000}.

4.2. Ablation Studies

To perform a detailed component analysis, we conducted the ablative experiments on MSCOCO.

The effect of commonsense knowledge. Table 3 shows that commonsense knowledge is the most vital component for passing information among categories. If we replace commonsense knowledge with a identity matrix, performance decreases 3.4% mAP on MSCOCO.

The effect of global semantic pool $\mathbf{M}$. Adaptive global reasoning among global semantic pool $\boldsymbol{M}$ can significantly improve the mAP by 2.7. It can be found that the recall for small region boost significantly (about 11% on AR_S compared to pass information among region feature).

The effect of adaptive attention α. We also conduct experiment without adaptive attention, and regard uniform weight of categories ($\frac{1}{C}$) as attention. Note that these two attention mechanisms lead to reason same information per image during testing. The results show the image-wise attention mechanism helps more effective propagation across the graph by discovering more relevant categories, and increase the overall AP by 1.4% to 2%.

%	Method and Modifications	mAP	AP_{50}	AR_S	AR_M
MS-COCO	**Reasoning-RCNN$_R$ with FPN**	**42.9**	**61.8**	**40.2**	**63.9**
	Identity co-occurrence knowledge graph	$39.5^{-3.4}$	$58.3^{-3.5}$	$36.4^{-3.8}$	$60.1^{-3.8}$
	Use region feature instead of global semantic pool	$40.2^{-2.7}$	$59.0^{-2.8}$	$36.2^{-4.0}$	$61.4^{-2.5}$
	No adaptive attention	$40.9^{-2.0}$	$59.5^{-2.3}$	$37.3^{-2.9}$	$61.6^{-2.3}$
	Uniform attention weights	$41.5^{-1.4}$	$60.6^{-1.2}$	$37.5^{-2.7}$	$61.5^{-2.4}$

Table 3. Ablation Studies based on modifications of our final model on MSCOCO. Reasoning-RCNN$_R$ with FPN is our final model. The backbones are ResNet-101 with FPN.

4.3. Analysis of feature interpretability

To better understand the enhanced feature representations that our Reasoning-RCNN actually learned, we record the output $\mathbf{f}'_r$ for our method with relation knowledge and its corresponding real labels from each region of 10000 VG_{1000} images. Then we take average according to the labels and use the t-SNE [34] clustering method to visualize them as shown in Figure 6. From two enlarged regions, we can see that features $\mathbf{f}'_r$ of categories which have the spatial relationship or co-occurrence relationship such as "things on the street" and "foods in the table" are closed to each other. And this speaks well our knowledge reasoning stage successfully incorporates the prior relation knowledge and leads to interpretable feature learning. More gradient visualization results are included in Supplementary materials for better understanding the method.

5. Conclusion

We present a novel adaptive global reasoning network named Reasoning-RCNN. By propagating over a global semantic pool, our Reasoning-RCNN enhances the feature expressions for both classification and localization and coordinate with visual patterns adaptively in each image. We instantiate our method with two kinds of prior knowledge, e.g. relationship and attribute. The solid and consistent detection improvements of the Reasoning-RCNN on all datasets suggest the adaptive global reasoning is required to advance large-scale object detection. For future works, extensions can be made for embedding our reasoning framework into other tasks such as instance-level segmentation.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 3

[2] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 3

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 1, 6, 7

[4] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. mmdetection. https://github.com/open-mmlab/mmdetection, 2018. 6

[5] X. Chen and A. Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, 2017. 2, 6, 7

[6] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, 2018. 2, 3, 5, 7

[7] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 3

[8] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 1, 2, 5, 6, 7

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2, 5

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 3

[11] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 3

[12] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. In *ICCV*, 2017. 6, 7

[13] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 3

[14] S. Gidaris and N. Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018. 3, 4

[15] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu. Frage: Frequency-agnostic word representation. In *NIPS*, 2018. 3, 4

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[18] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, 2014. 1

[19] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *CVPR*, 2018. 2, 6, 7

[20] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[21] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to segment every thing. In *CVPR*, 2018. 5

[22] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014. 3

[23] C. Jiang, H. Xu, X. Liang, and L. Lin. Hybrid knowledge routed modules for large-scale object detection. In *NIPS*, 2018. 3

[24] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2016. 5

[26] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 3

[27] K.-H. Lee, X. He, L. Zhang, and L. Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 2018. 2, 4

[28] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Light-head r-cnn: In defense of two-stage object detector. In *CVPR*, 2017. 6, 7

[29] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Detnet: A backbone network for object detection. In *ECCV*, 2018. 2, 6, 7

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6, 7

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2, 6, 7

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[34] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[35] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015. 3

[36] K. Marino, R. Salakhutdinov, and A. Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, 2017. 2, 3

[37] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3

[38] I. Misra, A. Gupta, and M. Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 3

[39] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 3

[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Workshop*, 2017. 6

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[42] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 3

[43] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 6, 7

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5, 6

[45] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, pages 1481–1488. IEEE, 2011. 4

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[47] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 2

[48] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018. 3, 4

[49] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 3

[50] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 5