electronics

MDPI

Article

Few-Shot Object Detection Method Based on Knowledge Reasoning

Jianwei Wang [1,2] and Deyun Chen [1,*]

[1] College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China; 1910400003@stu.hrbust.edu.cn
[2] College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050, China
* Correspondence: chendeyun@hrbust.edu.cn; Tel.: +86-13936132299

Abstract: Human beings have the ability to quickly recognize novel concepts with the help of scene semantics. This kind of ability is meaningful and full of challenge for the field of machine learning. At present, object recognition methods based on deep learning have achieved excellent results with the use of large-scale labeled data. However, the data scarcity of novel objects significantly affects the performance of these recognition methods. In this work, we investigated utilizing knowledge reasoning with visual information in the training of a novel object detector. We trained a detector to project the image representations of objects into an embedding space. Knowledge subgraphs were extracted to describe the semantic relation of the specified visual scenes. The spatial relationship, function relationship, and the attribute description were defined to realize the reasoning of novel classes. The designed few-shot detector, named KR-FSD, is robust and stable to the variation of shots of novel objects, and it also has advantages when detecting objects in a complex environment due to the flexible extensibility of KGs. Experiments on VOC and COCO datasets showed that the performance of the detector was increased significantly when the novel class was strongly associated with some of the base classes, due to the better knowledge propagation between the novel class and the related groups of classes.

Keywords: convolutional neural network (CNN); graph neural network (GNN); few-shot object detection; knowledge reasoning; knowledge graphs (KGs)

Citation: Wang, J.; Chen, D. Few-Shot Object Detection Method Based on Knowledge Reasoning. *Electronics* **2022**, *11*, 1327. https://doi.org/10.3390/electronics11091327

Academic Editor: John Ball

Received: 23 January 2022
Accepted: 10 April 2022
Published: 22 April 2022

1. Introduction

The application of artificial intelligence technology in specific industrial contexts has become more and more common [1]. Object detection is the basis of many computer vision tasks, such as instance segmentation, image captioning, object tracking and so on [2]. Object detection aims to find all interested objects in the image by determining their categories and locating their positions. Due to the different appearance and posture of various objects, coupled with the interference of imaging light, occlusion and other factors, object detection has always been a challenging problem. Driven by big data, the deep learning model can be effectively trained with the help of abundant annotation data. In recent years, the performance of object detection algorithm based on deep learning methods has improved consistently. However, object detection methods based on deep learning show obvious shortcomings in open and complex scenes, partly because of a lack of labeled data. Insufficiently labeled data will lead to overfitting of the trained model. Although simple data enhancement and regularization techniques can alleviate this problem, it has not been completely solved.

Since the data in the real world has the characteristics of long tail distribution, object detection in the open world is an urgent and difficult problem. The performance is often degraded by the scarcity of new data. Human beings can make good use of experience knowledge to learn to solve new problems with the help of only a few examples. Few-shot learning aims to learn like human beings by making use of prior knowledge and only a small number of samples of new problems. In recent years, many researches [3–9] have

been provided on few-shot object detection. Abundant labeled objects are taken as base classes while unlabeled objects are treated as new classes. With the aid of abundant data in base classes, novel objects can be detected by the trained few-shot detector with limited data in novel classes.

This project proposes to establish the relationship between the base classes and the novel classes with the help of external semantic knowledge [10]. The KGs containing text and attribute knowledge takes objects as nodes and the relationship between objects as edges in order to form a graph. The related concepts in the graph structure have statistical strength and can be extended to novel classes. In this work, we study how to use this semantic relationship to integrate knowledge reasoning and visual information into a unified framework to achieve optimal compatibility. Specifically, we represent the concept of each class through semantic embedding, use graph neural network (GNN) as a semantic visual mapping network for knowledge reasoning, and combine image features and semantic features through classifier weight fusion to obtain better detection results.

The remainder of this paper is organized as follows. Section 2 reviews related works and explains our contributions. Section 3 introduces the proposed method. Section 4 shows the experimental results. Section 5 shows the experimental analysis. Section 6 summarizes our conclusions and future works.

2. Related Work and Contribution

2.1. Few-Shot Learning

The concept of few-shot learning first emerged from computer vision field [11]. It has attracted extensive attention in recent years. There are many algorithm models with excellent performance in image recognition tasks [12,13], such as the famous prototypical network [14] and the matching network [15]. The method based on meta learning not only trains the model on the target task, but also learns meta knowledge from many different tasks. Meta knowledge is used to adjust the model so that the model can converge quickly when facing a new task.

In the task of few-shot image recognition, Gregory et al. [16] designed a twin network, with identical structure and shared weights, to extract features from two images respectively, and calculated the similarity of the two images. The relationship network proposed by Flood et al. was transformed from a predefined fixed similarity measurement function to a learnable nonlinear similarity measurement function trained by neural network [17].

2.2. Knowledge Graphs

In daily life, if we know some static attributes of new things in advance including color, texture, shape, etc., as well as relationship attributes, such as the relationship with some easily recognizable objects of base classes, it will become easier to learn said new things. Therefore, when visual information is difficult to obtain, this explicit relational reasoning is more important. This relationship can be constructed through knowledge graphs (KGs). The definition of a knowledge graph is usually based on heuristic methods in common sense knowledge rule database [18,19]. For multi-label recognition, [20] provided an object co-occurrence-based knowledge graph. Ref. [21] provided a reasoning method over knowledge graphs and showed that reasoning over knowledge graphs can obtain conclusions from existing data. An increasing number of KGs have been constructed and published recently, by both academia and industry, such as Google Knowledge Graph, Microsoft Satori, and Facebook Entity Graph [22–24].

2.3. Object Detection

The challenges of object detection include but are not limited to the following aspects: different viewpoints, illumination and intra changes of class, scale changes, object rotation, dense and occluded object detection, small objects, and accurate object positioning etc. [2]. For the detection of scarce objects or the object under given conditions, due to

the lack of labeled data, the conventional detection model usually struggles to achieve the ideal accuracy.

Some works [25–28] have focused on the problem detecting objects in limited data scenarios. LSTD [3] proposed a method of promoting the transfer of knowledge from the source domain to the target domain. RepMet adopted distance measurement learning classifiers in ROI classification header [4]. MSPLD proposed to iterate between model training and high confidence sample selection [5]. Meta R-CNN and FSRW proposed using the attention vector of each class to readjust the feature mapping of the corresponding class [6,7]. MetaDet used meta level knowledge about model parameter generation to deal with category specific components of new classes [8]. In FSOD, the similarity between a small number of support sets and query sets was explored to detect new objects [9].

The performance of a few-shot detector is greatly affected by the scarcity of novel objects. However, the semantic relationship between the novel objects and the base objects is constant [10]. This kind of semantic relationship can be easily extracted from a knowledge graph of the real world. Therefore, we proposed a few-shot object detection method based on knowledge reasoning, shown in Figure 1, to detect and infer novel objects when some basic properties of the novel objects and the relationships with base objects are provided in advance.

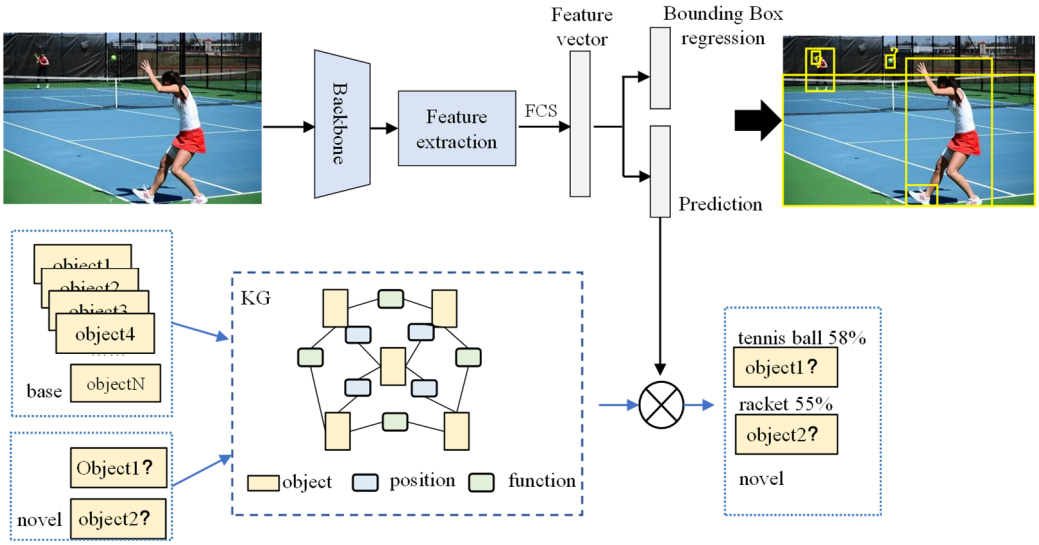

Figure 1. Few-shot object detection framework based on knowledge reasoning. A knowledge subgraph is extracted from the knowledge graph of the real world according to the objects recognized by a CNN recognition model and is applied to infer the unknown objects. ⊗: dot product.

We summarized the contributions as follows:

(1) A few-shot object detection method based on knowledge reasoning was proposed. It applied knowledge graphs together with the visual information to the novel object detection.
(2) We designed a general expression pattern of knowledge graphs, which can be flexibly applied to express the relationship between visible objects, and has good scalability.
(3) By using GNN, a novel object can be recognized by the method of knowledge reasoning. The proposed methodology achieves state-of-the-art performance on object detection.

3. Methodology

3.1. Few-Shot Object Detection

The set of the known object classes is denoted as U, where $U = \{C_1, C_2, \cdots, C_N\}$ and N is the number of the recognized object classes in the image. We assume that there exist unknown object classes in the image, and the set of unknown classes is denoted as V,

where $V = \left\{C_1^?, C_2^?, \cdots, C_j^?, \cdots\right\}$. It is assumed that there are K_1 object instances with their class labels and locations and K_2 unknown object instances with their locations in the input image Im. The i-th object instance is denoted as O_i, where $O_i = [l_i, x_i, y_i, w_i, h_i]$, $l_i \in U$ and x_i, y_i, w_i, h_i denote the bounding box center coordinates, width, and height, respectively. The j-th unknown object instance is denoted as $O_j^?$, where $O_j^? = \left[l_j^?, x_j, y_j, w_j, h_j\right]$, $l_j^? \in V$ and x_j, y_j, w_j, h_j denote the bounding box center coordinates, width, and height respectively.

In a dataset of novel classes, the number of objects for each class is k for k-shot detection task. The few-shot detection model is constructed on the base of a two-stage detection framework. At the second stage, the labels of some uncertain object instances can be inferred by KGs.

3.2. Few-Shot Detector

There are two training phases for a typical few-shot detector, the base training phase on a base dataset and the fine-tuning phase on the union of a base dataset and a novel dataset. Differing from these methods, we designed a CNN model to detect the known object instances with their class labels and locations, and to further infer the class labels of novel objects though knowledge reasoning. Compared with one-stage detectors, two-stage detectors have better open set performance. The framework of our detector is shown in Figure 2, where Faster R-CNN [29] was chosen as the baseline.

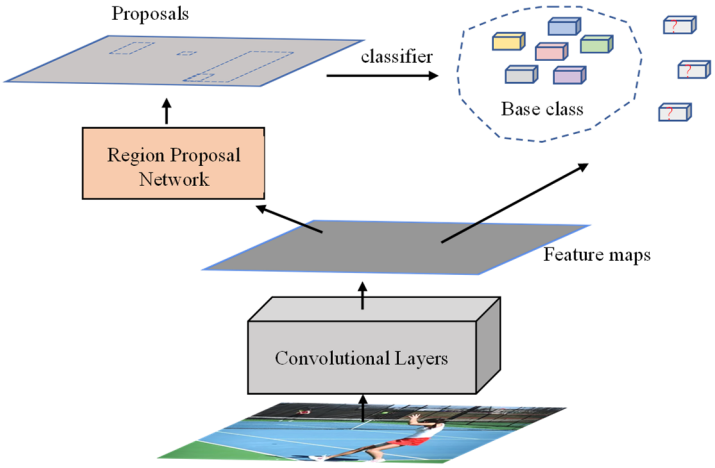

Figure 2. Framework of our detector. Potential regions of objects from the feature maps, which are represented with bounding box coordinates, are proposed by RPN. Objects of base classes are easy to be recognized. It is difficult to recognize objects of novel classes or objects with few features.

The role of Region Proposal Network (RPN) is to search for numerous candidate anchors (a set of candidate bounding boxes on the image), and then to determine whether the corresponding area of an anchor has an object prospect or is a background without object. Feature maps are generated after convolution layers. Each point of the feature maps corresponds to multiple anchors. There are k anchors in total, and each anchor needs to be distinguished between foreground and background. The foreground anchors are obtained by softmax classification; that is, the candidate region box is preliminarily extracted. Each anchor has four position offsets corresponding to [x, y, W, H], which are corrected by using the bounding box regression. In fact, RPN has preliminarily realized the object detection and positioning.

3.3. Knowledge Graph for a Scene

3.3.1. Knowledge Graph

A knowledge graph aims to describe various entities and their relationships in the real world. Entities refer to things that are distinguishable and independent. Semantic

class refers to a collection of entities with certain characteristics. The attribute values are assigned by the entities.

Knowledge graph is defined as G_K.

$$G_K =< E, R, AT \Big| E = \{e_1, e_2, \cdots, e_N\} \wedge R = \{r_1, r_2, \cdots, r_M\} \wedge AT = \left\{AT^1, AT^2, \cdots, AT^N\right\} > \quad (1)$$

where E is the entity set and N is the number of entities in knowledge graph. R is the relationship set and M is the number of relationships in knowledge graph. AT is the attribute set of entities.

The attribute set of the $i-$th entity is defined as the following:

$$AT^i = \left\{at_1^i, at_2^i, \cdots, at_{nu}^i\right\} \quad (2)$$

where nu is the number of attributes of the i-th entity.

Two basic forms of R in knowledge graph are expressed with Formulas (3) and (4).

$$R_1 = \left\{r_u(e_i, e_j) \middle| e_i \in E \wedge e_j \in E \wedge r_u \in R\right\} \quad (3)$$

$$R_2 = \left\{at_v^k(e_k) \middle| e_k \in E \wedge at_v^k \in AT^k\right\} \quad (4)$$

where $R = R_1 \cup R_2$. There is $r_u(e_i, e_j)$ when e_i and e_j satisfy relation r_u, and there is $at_v^k(e_k)$ when the entity e_k has assigned the value of attribute at_v.

In the knowledge map, relation is a function, which maps K graph nodes (entities, semantic classes, attribute values) to Boolean values.

3.3.2. Scene Graph

In the object detection task, the scene graph describing semantics is defined as following:

$$G_S =< O, C, Edge | O = \{o_1, o_2, \cdots, o_n\} \wedge C = \{c_1, c_2, \cdots, c_m\} \wedge Edge \subseteq O \times R \times O > \quad (5)$$

where O is the set of object instances, C denotes the set of object classes and $Edge$ is a set of edges.

According to the definition of a knowledge graph, the attribute set of object instances is a subset of AT, and the relationship set of object instances is a subset of R. We describe an object instance with a triplet, shown as (6).

$$o_i =< c(o_i), AT(o_i), Loc(o_i) | c(o_i) \in C \wedge AT(o_i) \subseteq AT \wedge Loc(o_i) =< x, y, W, H >> \quad (6)$$

where $c(o_i)$ is the class of the object o_i, $AT(o_i)$ is the attribute set of the object o_i and $Loc(o_i)$ is the position information of the object o_i.

3.3.3. Object Detection

Given the scene graph G_S of the image Im, object instances to be detected are represented by a set of candidate bounding boxes denoted as B. The map is denoted as $\gamma : O \to B$. The initial knowledge graph is built and denoted with an adjacent matrix M_{ob} and a feature matrix M_{AT}, where $M_{ob} \subseteq \{0, 1\}_{n \times n}$, $M_{AT} = \{at_1, \cdots, at_n\}^T \subseteq R_{n \times h}$.

A function $f(\cdot)$ is defined upon graph neural network (GNN) to learn graph and it can be implemented by a method in [30]. For the l-th layer of GNN, the weight matrix is defined as $W(l)$. M_{AT} is used as the initial features of modes. The message passing function is defined as $P(\cdot)$ with the following structure:

$$L(l+1) = P(M_{ob}, L(l), W(l)) \quad (7)$$

where $L(l+1)$ are the nodes embedding after l layers of GNN, and $L(l_1) = M_{AT}$.

3.4. Reasoning Based on Knowledge Graph

A knowledge subgraph is a data structure that describes the semantics of a specific scene. It encodes object instances, attributes of objects, and relationships between objects. The simplest way to extract the knowledge subgraph of a scene from the large knowledge graph describing the objective world is to retrieve the subgraph in the KG according to the recognized object instances and their attributes. The remaining available visual information is used to supplement the knowledge subgraph to enhance the semantic expression of the scene. The knowledge subgraph of Figure 1 is shown in Figure 3.

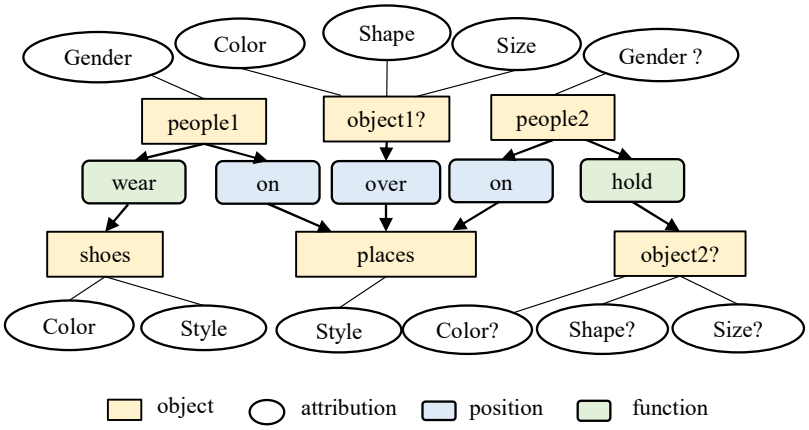

Figure 3. The knowledge subgraph of Figure 1.

Two types of relationships are defined in knowledge subgraph. R_1 is the set of positional relationships and R_2 is the set of functional relationships, where $R_1 \cup R_2 = R$, $R_1 \cap R_2 = \phi$. Figure 3 shows the knowledge subgraph of Figure 1. In this example, the object instances and their attributions are shown in Table 1. The relationships between instances are shown in Table 2.

Table 1. The object instances and their attributes in Figure 1.

Object Instance	Attributes
people1	Gender—"female"
people2	Gender—"unknown"
places	Style—"tennis court"
shoes	Style—"sports shoes", Color—"white"
object1?	Color—"green", Shape—"round", Size—"small"
object2?	Color—?, Shape—?, Size—?

Table 2. The relationships between instances in Figure 1.

Relationships	Map
r_1—"on"	r_1(people1, places), r_1(people2, places)
r_2—"over"	r_2(object1, places)
r_3—"hold"	r_3(people2, object2)
r_4—"wear"	r_4(people1, shoes)

In Figure 3, object instances (people 1, people 2, places, shoes) are easily detected, and most of the attributes can be recognized (Style of places, Style of shoes, Gender of people 1, the Color, Shape and Size of object 2). Although two object instances (object 1, object 2) cannot be recognized and some attributes (Gender of people2, the Color, Shape and Size of object 2) are unknown, we can clearly see that "people 2 hold object 2", "object 1 is over

tennis court". The set of object instances and the attribute set of object instances are shown as Formula (8) and Formula (9) separately.

$$O = \{obj1, obj2, places, shoes, people1, people2\} \tag{8}$$

$$AT = \{Gender, Color, Shape, Size, Style\} \tag{9}$$

M_{ob} and M_{AT} are defined as Formula (11) and Formula (12) separately.

$$M_{ob} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}_{6\times 6} \tag{10}$$

$$M_{AT}^{T} = \begin{bmatrix} AT^{obj1} & AT^{places1} & AT^{people1} & AT^{shoes} \end{bmatrix} \tag{11}$$

where $AT^{obj1} = \{\text{“}green\text{”}, \text{“}round\text{”}, \text{“}small\text{”}\}$, $AT^{places1} = \{\text{“}tennis \cdot court\text{”}\}$, $AT^{people1} = \{\text{“}girl\text{”}\}$, and $AT^{shoes} = \{\text{“}sports \cdot shoes\text{”}, \text{“}white\text{”}\}$. The attribute values and entities (object instances) can be obtained by visual information and detection algorithms.

3.5. Space Projection

We projected the visual feature into the constructed semantic space to recognize the objects based on both visual information and the semantic relation. In the second-stage of the two-stage object detector, the extracted feature vector region proposals are forwarded to a classification subnet and a regression subnet. In the classification subnet, the feature vector is transformed into a vector denoted as v with d dimensions and forwarded through fully connected layers. Then, v is multiplied by a learnable weight matrix $W \in R_{n\times d}$ to produce a probability distribution, which is shown in (12).

$$Pred = \text{softmax}(W \cdot v + b) \tag{12}$$

where n is the number of classes, b is a learnable bias vector and $b \in R_n$. Cross-entropy loss is used during training.

To reduce the domain gap, semantic embeddings are needed. Learning from the transformer, we implemented a dynamic graph with self-attention architecture [31]. For a new class, it is only necessary to simply insert corresponding embeddings of new classes and fine-tune the detector, because the graph is variable and is constructed according to the word embeddings.

4. Experiments

4.1. Datasets

To evaluate our method, we performed experiments on VOC [32] and COCO dataset [33], which are widely used for pretraining classification models. Before training the few-shot detector, we removed the new classes from the training dataset to initialize the backbone network, and to guarantee that the pretrained model has not seen these novel classes. Corresponding to the novel classes in VOC, the WordNet IDs to be removed are shown in Table 3.

In the COCO dataset, since the classes have the character of long-tail distribution, we selected the data-scarce classes on the distribution tail as the novel classes and used Google Knowledge Graph, Microsoft Satori, and Facebook Entity Graph as the base to extract the knowledge subgraph for a specific scene.

Table 3. The removed new classes.

Datasets	WordNet IDs [10]
aeroplane	n02690373, n02692877, n04552348
bird	n01514668, n01514859, n01518878, n01530575, n01531178, n01532829, n01534433, n01537544, n01558993, n01560419, n01580077, n01582220, n01592084, n01601694, n01608432, n01614925, n01616318, n01622779, n01795545, n01796340, n01797886, n01798484, n01806143, n01806567, n01807496, n01817953, n01818515, n01819313, n01820546, n01824575, n01828970, n01829413, n01833805, n01843065, n01843383, n01847000, n01855032, n01855672, n01860187, n02002556, n02002724, n02006656, n02007558, n02009229, n02009912, n02011460, n02012849, n02013706, n02017213, n02018207, n02018795, n02025239, n02027492, n02028035, n02033041, n02037110, n02051845, n02056570, n02058221
boat	n02687172, n02951358, n03095699, n03344393, n03447447, n03662601, n03673027, n03873416, n03947888, n04147183, n04273569, n04347754, n04606251, n04612504
bottle	n02823428, n03062245, n03937543, n03983396, n04522168, n04557648, n04560804, n04579145, n04591713
bus	n03769881, n04065272, n04146614, n04487081
cat	n02123045, n02123159, n02123394, n02123597, n02124075, n02125311, n02127052
cow	n02403003, n02408429, n02410509
horse	n02389026, n02391049
motorbike	n03785016, n03791053
sheep	n02412080, n02415577, n02417914, n02422106, n02422699, n02423022
sofa	n04344873

4.2. Implementation Details

We trained the KR-FSD on the base of Faster R-CNN with Stochastic Gradient Descent (SGD), and set the batch size to 16. In base training phase, the learning rate was set to 0.02, the momentum was set to 0.9, and the weight decay was set to 0.0001. The learning rate was set to 0.001 in the fine-tuning phase. We sampled the input image randomly from the base set and the novel set with a 50% probability, and then randomly selected an image from the chosen set.

4.3. Results on VOC and COCO Datasets

In Table 4, we show the performance (AP_{50}) of the novel classes on the VOC dataset. We used the same data splits and a fixed list of novel samples provided by [6]. In the VOC dataset, 5 classes were selected as novel classes from 20 object classes. The remaining 15 classes were base classes. Each novel class had only a few annotated object instances, such as 1 annotated object instance, 5 annotated object instances, and 10 annotated object instances. Compared with the state-of-the-art methods (FSRW [6] and Meta R-CNN [7]), our approach can achieve superior performance.

Table 4. Performance (AP50) of the novel classes in the VOC dataset compared with state-of-the-art methods.

Shot	Method	Novel Sets					
		Bird	**Bus**	**Cow**	**Mbike**	**Sofa**	**Mean**
1	FSRW	13.5	10.6	31.5	13.8	4.3	14.8
	Meta R-CNN	6.1	32.8	15	35.4	0.2	19.9
	ours	**35.2**	**49.8**	**56.3**	**61.4**	**22.6**	**45.8**
5	FSRW	31.5	21.1	39.8	40.0	37.0	33.9
	Meta R-CNN	35.8	47.9	54.9	55.8	34.0	45.7
	ours	**42.2**	**55.7**	**60.6**	**62.3**	**41.8**	**52.5**

Table 4. *Cont.*

Shot	Method	Novel Sets					
		Bird	**Bus**	**Cow**	**Mbike**	**Sofa**	**Mean**
10	FSRW	30.0	62.7	43.2	60.6	39.6	47.2
	Meta R-CNN	52.5	55.9	52.7	54.6	41.6	51.5
	ours	**44.5**	**65.8**	**62.7**	**65.8**	**42.5**	**56.3**

Table 5 shows the averaged APs of our method on the COCO dataset. In the COCO dataset, the minival set was used for testing and the rests were used for training. Twenty classes were selected as novel classes from 80 classes. The remaining 60 classes were the base. The novel classes overlapped with the classes in VOC. Each novel class had only a few annotated object instances, such as 10 annotated object instances, 20 annotated object instances, and 30 annotated object instances.

Table 5. The averaged APs of the novel classes on the COCO dataset.

Shot	Method	AP50	AP75	AP
10	FSRW	12.3	4.6	5.6
	Meta R-CNN	19.1	6.6	8.7
	ours	**21.5**	**8.7**	**10.2**
20	FSRW	16.5	6.3	7.8
	Meta R-CNN	22.8	9.1	10.9
	ours	**26.1**	**11.5**	**13.2**
30	FSRW	19.0	7.6	9.1
	Meta R-CNN	25.3	10.8	12.4
	ours	**28.6**	**13.2**	**14.1**

Table 6 shows the ablative performance, where mAP = 50. KR is knowledge reasoning component.

Table 6. Ablative performance (mAP50) on the VOC.

KERRYPNX	KR	1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
Faster R-CN		32.8	44.7	46.1	49.8	55.8
ours	√	45.8	46.2	47.3	52.5	56.3

4.4. Experiments on Relation Reasoning

In order to verify the effectiveness of KG-based reasoning in detection, we selected three groups of data for effectiveness experiments. In the data group, the performance of each class was counted respectively by selecting one class as the novel class and the remaining classes as the base classes.

In the first experiment, we primarily counted the performance of sofa, TV, cat and chair. The object instances (such as TV, chair, and cat) in many scenes are strongly associated with the sofa. We frequently see "TV and sofa together", and "cat sitting on chair" or "cat sitting on sofa". In KGs, the distances between these entities (sofa and TV, cat and sofa, cat and chair) are shorter. Because these classes are closely related, the performance of all classes will be significantly improved when knowledge reasoning is integrated into visual features. The performance is shown in Figure 4.

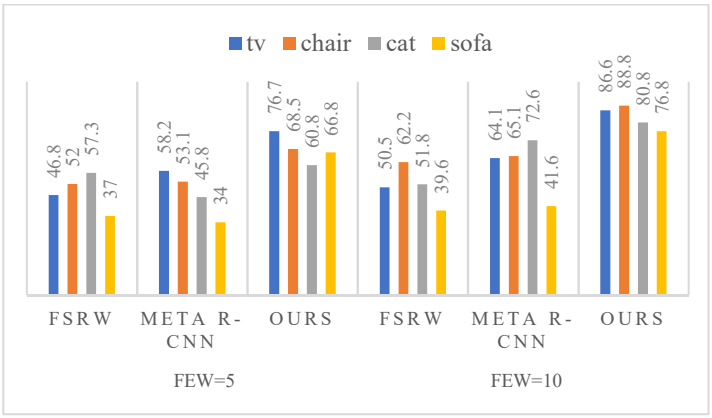

Figure 4. The first experiment on relation reasoning. The classes (sofa and TV, cat and sofa, cat and chair) have strong correlation. The performances of correlated classes are increased.

In other two experiments, sofa almost has no association with other classes (mbike, bus, car, bird, cow, horse). Therefore, the distances between sofa and other entities (mbike, bus, car, bird, cow, horse) are longer. However, we often see bus, car and mbike at the same time, and sometimes see cow and horse together. Figures 5 and 6 show that almost all the performances of correlated classes are increased slightly, due to the better knowledge propagation between the two groups of classes.

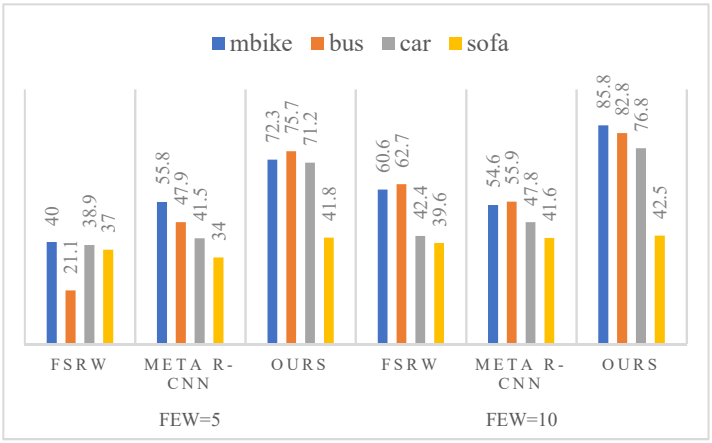

Figure 5. The second experiment on relation reasoning. Sofa has almost no association with other classes (mbike, bus, car). The performances of correlated classes (mbike, bus, car) are increased slightly.

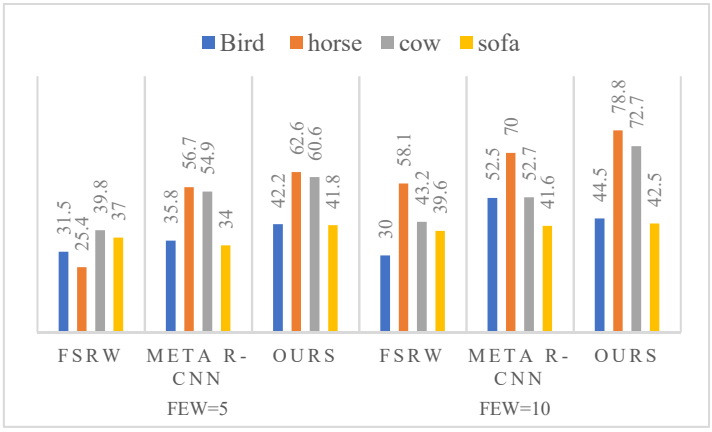

Figure 6. The third experiment on relation reasoning. Sofa almost has no association with other classes (bird, horse, cow). The performances of correlated classes (horse, cow) are increased slightly.

5. Experimental Analysis

This work combines visual information and knowledge reasoning method in order to recognize novel classes. Experiments on VOC datasets showed that the performance was increased slightly at lower shot levels, such as 1-shot, and the performance was competitive compared with previous state-of-the-art methods at 5-shot and 10-shot, which is shown in Table 4. The average APs of the novel classes on the COCO dataset were increased slightly at 10-shot, 20-shot and 30-shot, which are shown in Table 5. Knowledge reasoning is proved to be meaningful for recognition tasks, as shown in Figures 4–6. When the novel classes were strongly associated with the base classes, the performance was noticeably increased, because there was better knowledge propagation between the novel classes and the related groups of classes.

6. Conclusions

In this paper, we proposed a few-shot object detection method based on knowledge reasoning. Since the semantic relation between the base classes and the novel classes in some scenes can be inferred by KGs, it is helpful to learn the novel concepts by applying knowledge reasoning with the available visual information. We built a few-shot detection model on the base of Faster R-CNN, and applied reasoning to some uncertain object instances at the second stage. To demonstrate the performance, we carried out experiments on VOC datasets and COCO datasets. Compared with state-of-the-art methods, our approach achieved better results at several few-shot detection settings. In future work, we will further carry on research on few-shot recognition and detection driven by knowledge and data.

Author Contributions: J.W. and D.C. designed the study; JW analyzed and interpreted the data; J.W. conducted the experiments, J.W. and D.C. provided the technical and material support. All authors contributed to the writing of the manuscript and final approval. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data is available from authors upon the reasonable request to the corresponding authors.

Acknowledgments: The authors thank the anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influenced the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
KR-FSD	Knowledge Reason Few-Shot Detection
KG	Knowledge Graphs
SGD	Stochastic Gradient Descent
CNN	Convolutional Neural Network
GNN	Graph Neural Network
RPN	Region Proposal Network
ROI	Region of Interest
VOC	Visual Object Classes
COCO	Common Objects in Context
AP	Average Precision

References

1. Xu, X.f.; Hao, J.; Zheng, Y. Multi-objective Artificial Bee Colony Algorithm for Multi-stage Resource Leveling Problem in Sharing Logistics Network. *Comput. Ind. Eng.* **2020**, *142*, 106338. [CrossRef]
2. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
3. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
4. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5197–5206.
5. Dong, X.; Zheng, L.; Fan Ma, Y.Y.; Meng, D. Few-example object detection with model communication. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1641–1654. [CrossRef] [PubMed]
6. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8420–8429.
7. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9577–9586.
8. Wang, Y.-X.; Ramanan, D.; Hebert, M. Meta-learning to detect rare objects. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9925–9934.
9. Fan, Q.; Zhuo, W.; Tang, C.-K.; Tai, Y.-W. Few-shot object detection with attention-rpn and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4013–4022.
10. Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; Savvides, M. Semantic relation reasoning for shot-stable few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8782–8791.
11. Yang, J.; Liu, Y.L. The latest advances in face recognition with single training sample. *J. Xihua Univ. (Nat. Sci. Ed.)* **2014**, *33*, 1–5.
12. Zhang, C.; Cai, Y.; Lin, G.; Shen, C. DeepEMD: Differentiable Earth Mover's Distance for Few-Shot Learning. *arXiv* **2020**, arXiv:2003.06777.
13. Simon, C.; Koniusz, P.; Nock, R.; Harandi, M. Adaptive subspaces for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4136–4145.
14. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *arXiv* **2017**, arXiv:1703.05175.
15. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3630–3638.
16. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015.
17. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.
18. Song, J.; Shen, C.; Yang, Y.; Liu, Y.; Song, M. Transductive unbiased embedding for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1024–1033.
19. Vyas, M.R.; Venkateswara, H.; Panchanathan, S. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In Proceedings of the European Conference on Computer Vision, virtual, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 70–86.
20. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [CrossRef]
21. Chen, X.; Jia, S.; Xiang, Y. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.* **2020**, *141*, 112948. [CrossRef]
22. Google Inside Search. Available online: https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html (accessed on 23 April 2017).
23. Wang, H.; Zhao, M.; Xie, X.; Li, W.; Guo, M. Knowledge graph convolutional networks for recommender systems. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 3307–3313.
24. Ehrlinger, L.; Wöß, W. Towards a Definition of Knowledge Graphs. *SEMANTiCS* **2016**, *48*, 2.
25. Yang, Z.; Wang, Y.; Chen, X.; Liu, J.; Qiao, Y. Context-transformer: Tackling object confusion for few-shot detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12653–12660.
26. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
27. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the European Conference on Computer Vision, virtual, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 456–472.
28. Xiao, Y.; Marlet, R. Few-shot object detection and viewpoint estimation for objects in the wild. In Proceedings of the European Conference on Computer Vision, virtual, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 192–210.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.

30. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
32. Everingham, M.; van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
33. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014. [CrossRef]