*mathematics*

MDPI

*Article*

# CMKG: Construction Method of Knowledge Graph for Image Recognition

**Lijun Chen [1,†], Jingcan Li [2,†], Qiuting Cai [2,†], Xiangyu Han [2,†], Yunqian Ma [2,†] and Xia Xie [2,*]**

[1] School of Cyberspace Security, Hainan University, Haikou 570228, China
[2] School of Computer Science and Technology, Hainan University, Haikou 570228, China
* Correspondence: shelicy@hainanu.edu.cn
† These authors contributed equally to this work.

**Abstract:** With the continuous development of artificial intelligence technology and the exponential growth in the number of images, image detection and recognition technology is becoming more widely used. Image knowledge management is extremely urgent. The data source of a knowledge graph is not only the text and structured data but also the visual or auditory data such as images, video, and audio. How to use multimodal information to build an information management platform is a difficult problem. In this paper, a method is proposed to construct the result of image recognition as a knowledge graph. First of all, based on the improvement in the BlendMASK algorithm, the hollow convolution kernel is added. Secondly, the effect of image recognition and the relationships between all kinds of information are analyzed. Finally, the image knowledge graph is constructed by using the relationship between the image entities. The contributions of this paper are as follows. (1) The hollow convolution kernel is added to reduce the loss from extracting feature information from high-level feature images. (2) In this paper, a method is proposed to determine the relationship between entities by dividing the recognition results of entities in an image with a threshold, which makes it possible for the relationships between images to be interconnected. The experimental results show that this method improves the accuracy and F1 value of the image recognition algorithm. At the same time, the method achieves integrity in the construction of a multimodal knowledge graph.

**Keywords:** knowledge graph; image recognition; BlendMask; instance segmentation

**MSC:** 68T07



**Citation:** Chen, L.; Li, J.; Cai, Q.; Han, X.; Ma, Y.; Xie, X. CMKG: Construction Method of Knowledge Graph for Image Recognition. *Mathematics* **2023**, *11*, 4174. https://doi.org/10.3390/math11194174

Academic Editors: Jie Wen, Yongbing Zhang and Lunke Fei

Received: 15 August 2023
Revised: 9 September 2023
Accepted: 12 September 2023
Published: 5 October 2023



## 1. Introduction

With the continuous development of intelligent technology and the exponential growth in the number of images, image detection and recognition technology is becoming more widely used. Computer image recognition is usually divided into several main steps: information acquisition, preprocessing, feature extraction and selection, classifier design, and classification decision. In recent years, the best effect of image recognition has been instance segmentation, which is based on convolution neural network target detection and semantic segmentation. For the initial instance segmentation method in 2014, inspired by the R-CNN [1] two-stage target detection framework, Bharath Hariharan proposed the SDS [2] model, which is also the earliest instance segmentation algorithm, and this laid the foundation for subsequent research.

Ren et al. proposed an instance segmentation model, Mask R-CNN, which has become the basic framework for many instance segmentation tasks and is based on Faster-RCNN [3]; this was added as a branch of semantic segmentation in 2017. In 2020, Chen et al. proposed BlendMask [4]. The method, based on the shortcomings of instance segmentation in one-stage target detection, takes FCOS [5] as the framework and combines the top–down and bottom–up methods [6]. A blender module was designed to fuse high-level features with lower-level features. However, due to the convolution kernel existing in the feature

fusion process, the receptive field in the feature layer of large pixels is too small [7], which will lead to inaccurate feature extraction and fusion. In order to solve the problem of a small receptive field, this paper aims to use hole convolution kernel substitution to further improve the feature extraction and fusion.

As an effective tool for big data mining, knowledge graphs can be applied to all kinds of knowledge topologies under the background of big data [8] because of their excellent ability to organize, manage, and understand data. This paper makes full use of the data association function of knowledge graphs [9] and combines it with the field of image recognition, which can not only analyze, manage, and classify the identified data but also provide a simple and direct infrastructure for the related research of image recognition. Therefore, how does one improve the efficiency of image recognition? At the same time, how does one build a good image knowledge graph?

The contributions of this paper are as follows. (1) In the process of image fusion, the convolution of 3 * 3 is transformed into a hollow convolution kernel to obtain a larger receptive field, which ensures the accuracy of features and reduces the loss. (2) An upsampling algorithm and downsampling algorithm are added to the algorithm to ensure the integrity of the information to the maximum extent. (3) Under the background of the big data network era, this paper considers the application of knowledge graphs to the result management of image recognition. The reasoning ability based on knowledge graphs enhances the ability of the machine to read image information and analyze the results of image recognition. It provides a convenient and concise data knowledge base for subsequent image recognition.

The article is structured as follows. In Section 2, we introduce the related work on the construction of image knowledge graphs. We introduce an improved image recognition algorithm in Section 3. In Section 4 of this paper, we put forward the method of constructing image knowledge graphs. In Section 5, we compare our method with the previous method and prove the effectiveness of the method. Lastly, we draw our conclusions in Section 6.

## 2. Related Work

This paper proposes the construction of image knowledge graphs in order to have as much and as accurate information as possible to judge the relationship between image entities. We analyze the current research on image recognition and choose a better method to optimize it. After that, the relevant scheme of constructing image knowledge graphs is put forward with the hope of providing better ideas for its development.

### *2.1. Target Detection*

In 2001, Viola et al. proposed the Haar feature extraction method and combined the AdaBoost [10] classification algorithm to achieve face detection. In 2005, Dalal et al. proposed the histogram of oriented gradients algorithm to carry out feature recognition through edge features and realize pedestrian detection by combining it with an SVM classifier. In 2008, Bay et al. proposed the SURF [11] algorithm based on the improvement of the SIFT algorithm, which greatly reduced the running time of the program and increased the robustness of the algorithm. In 2015, Redmon et al. [12] proposed YOLO, which was a new approach to object detection. In this paper, the object detection frame was defined as the boundary box of space separation and the regression problem of correlation probability. From 2016 to now, the YOLO series has changed from YOLOv1 to YOLOv7. In object detection, the anchored regression method and non-anchored regression method are dominant, but they also have certain shortcomings. In 2020, Bin et al. [13] proposed CPM R-CNN, which contains three efficient modules to optimize the anchor-based point-guided method. In recent years, unsupervised pretraining methods have been designed for target detection, but they usually have defects in image classification. Enze et al. [14] proposed a simple and effective self-supervised target detection method, DetCo.

### *2.2. Classical Model*

In 2014, Szegedy et al. designed the GoogLeNet [15] network model and proposed the structure of inception and its branch structure; the model achieved a new low error rate on the image datasets (ImageNet). At the same time, Simonyan et al. proposed the VGG-Net [16] model. In 2015, He et al. proposed the ResNet [17] network model, which achieved an error rate of only 3.6% on ImageNet. Convolutional neural networks (CNNs) have gained remarkable success in many image classification tasks in recent years. Wen-Shuai et al. [18] proposed an automatic CNN architecture design method by using genetic algorithms to effectively address the image classification tasks. In order to suppress the uncertainty of ResNet, in 2021, Kai et al. [19] proposed a simple yet efficient Self-Cure Network (SCN) based on ResNet18, which prevents deep networks from over-fitting uncertain facial images. In 2022, Jing et al. [20] introduced a regulator module as a memory mechanism to extract the complementary features of the middle layer and feed them further to ResNet. Recently, the size of the convolution kernel has also changed. In 2022, Xiaohan Ding et al. [21] proposed RepLKNet, a pure CNN architecture whose kernel size was as large as 31 * 31, in contrast to the commonly used 3 * 3. In 2023, combined with MobileNet and the ResNet-18 model, Lee et al. [22] proposed a block processing strategy, which effectively improved the efficiency of facial expression processing.

### *2.3. Instance Segmentation*

In 2015, Dai et al. proposed an Instance-sensitive Fully Convolutional Network (FCN) [23] in order to make up for the translation invariance defect of the Fully Convolutional Network (FCN) [24], and it completed the task of instance segmentation. Faster R-CNN proposes an RPN network based on the R-CNN series of algorithms to obtain accurate candidate regions. It is an end-to-end detection model for multi-object classification and localization.

In 2019, Bolya et al. proposed the YOLACT model to add the mask branch to the existing one-stage target model based on Mask R-CNN operating the same as Faster-RCNN but without explicit localization steps. To classify based on the instance center points problem and the dense distance regression problem, Xie et al. proposed the PolarMask [25] model. Based on the instance category of the quantized object center position and object size, Wang et al. proposed the SOLO model [26] to identify a single pixel: not a single output category but a category with location information in 2019. In the same year, based on the principle that inspection and segmentation should promote each other, Wang et al. proposed the RDSNet [27] to improve the performance of instance segmentation by making full and reasonable use of the information interaction between the target detection and instance segmentation. In 2022, Lu et al. proposed the Segmenting Objects from Relational Visual Data [28] that promoted the development of image segmentation. In 2023, Lei et al. [29] modeled the image formation as the composition of two overlapping layers and used the double-layer structure to model the occlusion relationship, which naturally decoupled the boundaries between instances and effectively solved the image segmentation problem in the case of occlusion.

### *2.4. Knowledge Graph*

As early as 1960, semantic networks were proposed as a method of knowledge expression, which was mainly used in the field of self-speech language understanding. In 2006, Tim introduced linked data to highlight the essence of the semantic web to establish links between open data. In recent years, the application of knowledge graph technology in various industries has become an important trend [30], such as the Baidu knowledge graph and the Google knowledge graph in the search field. In the medical field, there is a knowledge graph of traditional Chinese medicine [31]; there is JD.com's e-commerce field, and so on. These fully illustrate the universality of the knowledge graph.

The main contents of this paper include the acquisition of datasets, instance segmentation of images using the improved BlendMask algorithm, and the construction

of knowledge graphs. The traditional BlendMask algorithm is mainly studied, and an improved BlendMask algorithm is proposed; knowledge graphs are constructed using the instance segmentation results. The research content involves image segmentation, knowledge graphs, and other related theoretical knowledge.

## 3. Image Recognition Algorithm

### *3.1. Dense Instance Segmentation*

There are top–down and bottom–up methods for dense instance segmentation. DeepMask was the first model to use top–down instance segmentation. This method predicts a mask representation in each spatial region by a sliding window, but this method has three obvious disadvantages, as follows:

- The association between masks and features (local consistency) is lost, as in DeepMask, where a fully connected network is used to extract masks;
- There are redundant operations in feature extraction. For example, each foreground feature will extract a mask, which makes the model inefficient in DeepMask;
- When the upsampling adopts convolution with a step size greater than 1 (step), the position information of pixels will be lost.

The bottom–up dense instance segmentation method can better preserve low-level features such as details and location information. It first generates embedded features per pixel and then classifies them using post-processing methods. At the same time, it also has the following disadvantages:

- As this method requires a high quality of segmentation, it will lead to suboptimal segmentation results;
- The generalization ability is poor, which may not be applicable in complex scenes with multiple categories;
- It needs post-processing for classification, and the post-processing is cumbersome.

BlendMask uses the high-dimensional instance-level information generated by the top–down method to fuse the per-pixel prediction generated by the bottom–up method. Drawing on the ideas of FCIS and YOLACT, it proposes a Blender module that can blend global information containing instance-level features with low-level features that provide details and location information.

### *3.2. Improvement in the Convolution Kernel in an FPN Network*

In the structure of an FPN [32], the output results will be convoluted (3 * 3) to fuse the upsampling features to eliminate the discontinuity and alias in the upsampling process. Since the convolution kernel of each layer is of fixed size (3 * 3), the loss of image information in the upper convolution layer cannot be solved better.

In order to obtain more pixel features in the upper convolution core, we introduce a (7 * 7) dilated convolution core to strengthen the extraction of a wider range of feature images in the convolution process, reduce the loss of feature information extracted in the high-level feature images, and increase the accuracy of the mask prediction.

The formula for calculating the size of convolutional kernels is:

$$K = k + (k-1)(r-1) \tag{1}$$

where k is the size of the original convolution kernel, r is the parameter voidage (the interval number of kernels), and the standard convolution kernel r = 1, that is, the convolution kernel of 3 * 3.

In this paper, many kinds of convolution kernels of r = 2, r = 3, and r = 4 are compared, and the convolution kernel of 7 * 7 is the best when r = 3 is selected. The convolution kernel diagrams of 3 * 3 and 7 * 7 are shown in Figure 1. The receptive field of the convolution kernel before improvement is smaller and can only obtain pixels in the range of 3 * 3. The convolution kernel of 3 * 3 cannot perceive and solve the problems of discontinuity and

aliasing, while the 7 * 7 convolution kernel can well involve a wide range of image features, and the above problems can also be better solved.

In this paper, based on previous experience, the most suitable dilated convolution kernel size is 7 to 7. The increasing range of the convolution kernel performance of 7 * 7 is larger, the regularity of one convolution layer of 7 * 7 is equivalent to the superposition of three convolution layers of 3 * 3, and the dilated convolution kernel ratio of 7 * 7 is more suitable.

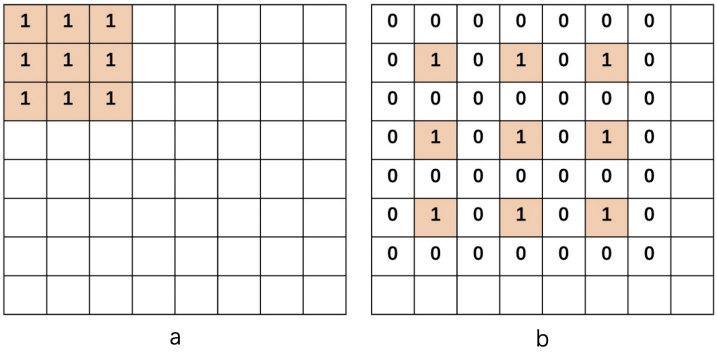



**Figure 1.** The (a) is the original convolution kernel, and (b) is the improved convolution kernel.

## 4. Construction of an Image Knowledge Graph

### *4.1. Introduction to Knowledge Graph*

Now, more and more people are combining a knowledge graph with other areas; for example, Li, in [33], provides the working principle of a Google knowledge graph. In medicine, Maclean, in [34], provided some of the information given by the knowledge graph during COVID-19, and the possibility of applying the knowledge graph in the early stages of drug discovery and reuse is proposed in view of its powerful reasoning and learning ability.

These applications have greatly accelerated the speed at which doctors can make diagnoses and have significantly improved the average level of care. Therefore, based on this idea, we put forward a method of extending the knowledge graph to universal image recognition. In this paper, classical image recognition is combined with knowledge graph construction to analyze and apply the image recognition effect better.

### *4.2. Image Knowledge Graph Construction*

The construction of a traditional knowledge graph is shown in Figure 2. As a graph data structure, the smallest unit of the knowledge graph is two nodes and the relationship between them. In this paper, we combine the image recognition results of the BlendMask algorithm with a knowledge graph. Based on the reasoning ability of the knowledge graph, the machine's abilities to read the image information and analyze the image recognition result are enhanced, which is convenient for the next application.

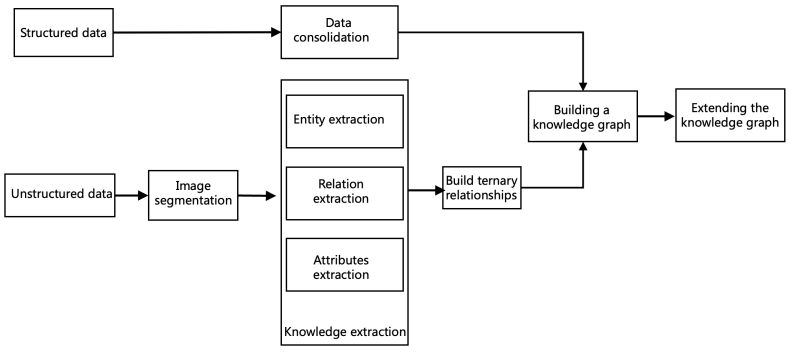



**Figure 2.** Flow chart of image knowledge graph construction.

First of all, the image recognition algorithm can extract the entities in the image. Then, based on the accuracy of the image recognition of objects, we can explore the similarity between instances of the same species to establish the triples of similarity between instances. Finally, we construct an image knowledge graph through these triples; at the same time, we set a certain threshold to be used to determine the same kinds of species similarity: the similarity to the threshold of the entity to judge for the same species. According to the judgment result, we extend the knowledge graph in direct relationship to the entity, and we lay a foundation for the application of the image recognition result later.

Image data acquisition and knowledge fusion: Building a knowledge graph can be segmented into three categories according to the degree of data: the structured data, the semi-structured data, and the unstructured data. For the structured data, there are relatively mature analysis technologies that can transform them. For RDF (Resource Description Framework) data, for example, there is D2R technology. However, the analysis technology using unstructured data is still very complex. For the content of unstructured data, we need to extract information. The information to be extracted includes the structured information such as entities, relationships, and attributes. Related technologies include entity extraction, relationship extraction, and attribute extraction.

As an instance segmentation algorithm, the BlendMask algorithm not only has the characteristics of semantic segmentation accurate to the pixel level but also has the characteristics of target detection and locating different instances of the same types of objects. Since the recognition result is uniform and contains a lot of information image data, this makes the data acquisition process in the knowledge graph construction based on the BlendMask image segmentation algorithm relatively simple and feasible. The image data acquisition process based on the image recognition algorithm is mainly divided into three steps as follows:

Step 1. Information extraction of a single image. We utilize the BlendMask algorithm to obtain the feature matrix of different instances in an image and extract the label and accuracy of the entity in the image recognition result. As shown in Figure 3, each different label represents a category, and each entity has a unique label and accuracy, completing the information extraction of a single image based on the above information;

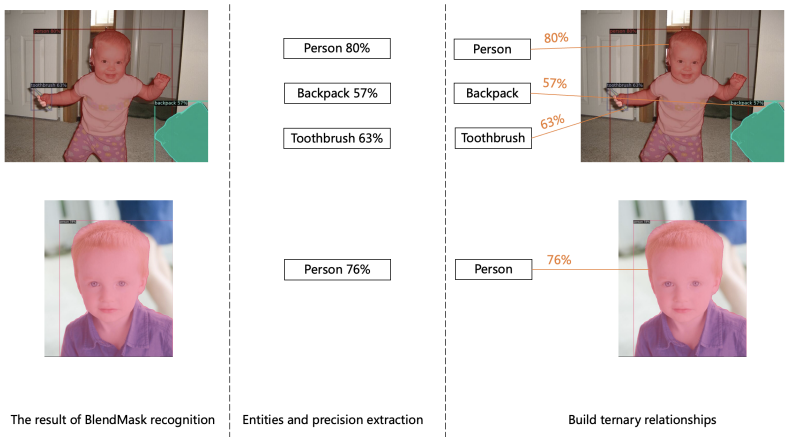



**Figure 3.** Label and accuracy of entities in the recognition result.

Step 2. Information extraction of all images. We repeat step 1 to obtain the entity features of all the images in the image recognition result and extract the label information and accurate information needed to construct the knowledge graph.

Step 3. Here, we classify the entities according to the label information, divide the accuracy of the same class, and finally construct the relationship between the entities according to the accuracy information, thereby constructing the entire image knowledge graph.

After completing the above steps, we have established multiple "category–accuracy–instance" ternary relationships, and we carry out knowledge fusion in the subsequent

stage, which is shown in Figure 4. Firstly, according to the label classification, the entities of the same class are formed into a simple ternary relational network. Secondly, under the same category, we compare the accuracy of each entity. When the accuracy difference between the two entities is less than 0.01, we consider the two instances to be the same species. For example, we add the relationship of the “same species” to them. When the accuracy difference between two instances is 0.01–0.05, we consider the two instances to be highly similar and add the relationship of "similar species" to them. When the accuracy between instances is more than 0.05, we infer that there is no further difference between instances. We obtain all entities and relationships through the above steps and finally build a knowledge graph, which is shown in Figure 5.

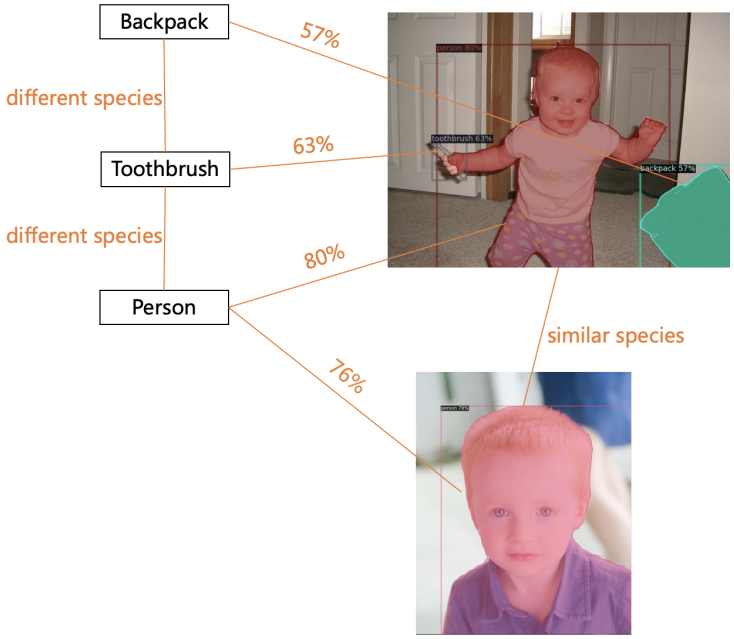



**Figure 4.** Image knowledge fusion.

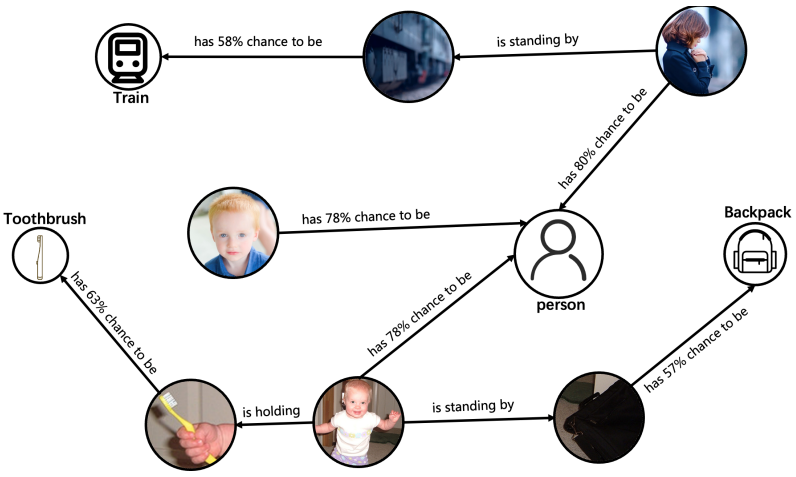



**Figure 5.** Effect graph of the image knowledge graph.

## 5. Experiments

We used the MSCOCO2017 [35] dataset to conduct the experiments. It is a large instance segmentation dataset, providing 80 categories, and the images are mainly taken from complex daily scenes. Our model used train2017 for training with a total of 105,000 images. The ablation experiment used val2017 with a total of 10,000 images.

**Training details:** In our experiment, our backbone network used the pretrained ResNet-50, and the bottom module used DeepLabv3+. Our basic learning rate was 0.01, and the batch size was 32. We conducted training on two GPUs for a total of 90,000 iterations. When iterated to 60,000, the learning rate was reduced to 0.001. In addition, we also adjusted the size of the input image, and we specified that the maximum length of the image was 1333px, and the maximum width was 800 px. All the other hyperparameters were consistent with FCOS.

**Testing details:** In our experimental results, the all-time units were milliseconds, and the experimental results in Tables 1 and 2 were obtained by running each batch of images on a 1080Ti GPU.

### *5.1. Evaluation Indicators*

Accuracy, Recall, and F1 were used for quantitative evaluation. Among them, accuracy calculates the proportion of correct predictions of all predictions, as in formula (1). Precision refers to the proportion of the correct positive predictions to all the positive predictions, as in formula (2). Recall refers to the proportion of correct positive predictions to all the correct predictions, as shown in formula (3):

*TP*: The positive class of the correct object detected;
*TN*: The negative class of the correct object detected;
*FP*: The positive class of detected incorrect objects;
*FN*: The negative class of the detected incorrect objects.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

### *5.2. Experimental Analysis of Mask R-CNN and BlendMask Algorithm*

From Table 1 of the experimental results, it can be seen that ResNet-101 is better than ResNet-50 in both the Mask R-CNN model and BlendMask model. The experimental results of the BlendMask model are generally better than Mask R-CNN. This is because BlendMask uses the high-dimensional information of the instance level generated by the top–down method. In Mask R-CNN, to obtain more accurate mask features, one must increase the resolution of the retooling. This will double the calculation time of the head and the network depth of the head, which is not only slow but also worse.

It can also be seen from the experiment that the speed of BlendMask using ResNet-101 was faster than that of Mask R-CNN using ResNet-50. This is because RPN is used in MASK R-CNN, while BlendMask uses focus, which is a first-level detector, saving the calculation of mask features.

**Table 1.** Experimental results of the Mask R-CNN and BlendMask.

| | Parameter | ResNet-50 | ResNet-101 |
|---|---|---|---|
| Mask R-CNN | Accuracy | 0.853 | 0.860 |
| | Precision | 0.902 | 0.906 |
| | Recall | 0.864 | 0.893 |
| | Times (ms) | 97 | 117 |
| BlendMask | Accuracy | 0.912 | 0.930 |
| | Precision | 0.910 | 0.910 |
| | Recall | 0.852 | 0.891 |
| | Times (ms) | 79.1 | 93 |

### *5.3. Experimental Analysis of BlendMask Algorithm before and after Improvement*

In the process of image feature extraction, this paper paid attention to the shortcomings of the fusion in feature extraction, increased the hollow convolution kernel, expanded the receptive field, and improved the efficiency and accuracy of the feature extraction, as shown in Table 2; the image recognition effect is shown in Figure 6.

**Table 2.** Experimental results before and after the improvement of the BlendMask algorithm.

| | Parameter | ResNet-50 | ResNet-101 |
|---|---|---|---|
| BlendMask | Accuracy | 0.912 | 0.930 |
| | Precision | 0.910 | 0.910 |
| | Recall | 0.852 | 0.891 |
| BlendMask_ours | Accuracy | 0.915 | 0.932 |
| | Precision | 0.911 | 0.913 |
| | Recall | 0.860 | 0.895 |

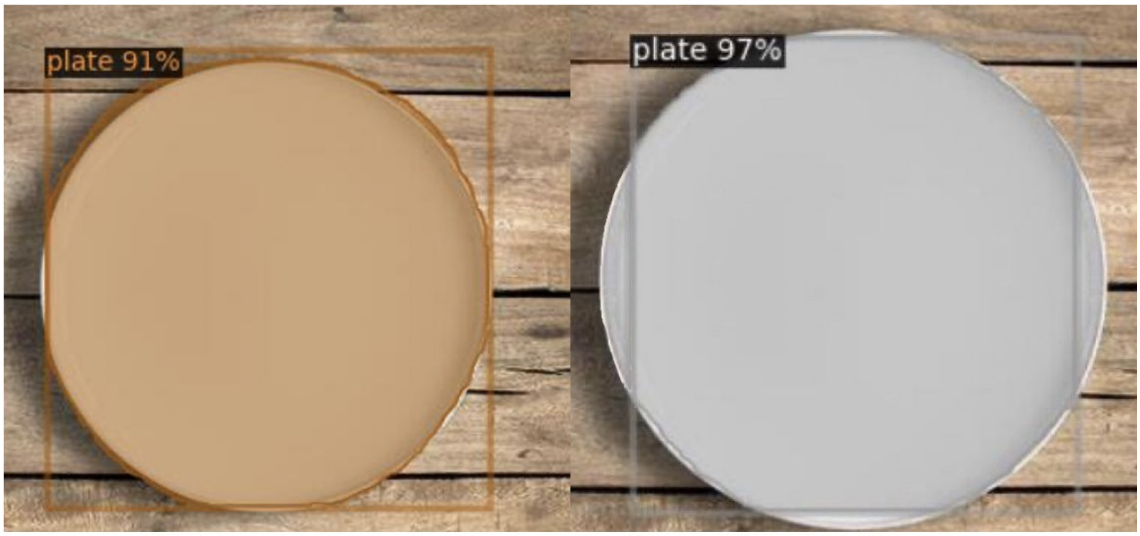
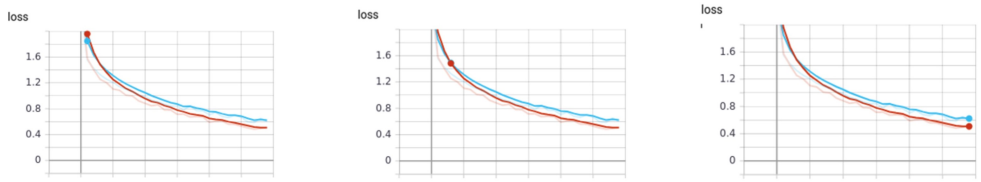



**Figure 6.** On the **left** is the BlendMask, and on the **right** is the improved model.

As shown in Table 1, the improved BlendMASK has improved in accuracy, precision, and recall, because the inflated convolution kernel strengthens the extraction of a larger range of feature images in the convolution process, reduces the loss of feature information extracted from high-level feature images, and improves the accuracy of mask prediction. This can better solve the discontinuity and aliasing effect. Compared with the ResNet-50 network, the ResNet-101 network has higher performance, which is consistent with our previous prediction. In Figure 6, we show their segmentation effect, and we can clearly see that adding the hole convolution BlendMASK can better extract the features of the image.

At the same time, we also consider the training and verification loss map before and after the improvement of the Blend Mask model, as shown in Figure 7.




step=1 step=3 step=29

**Figure 7.** Loss value of model training times before and after improvement.

In the figure, the red line chart is the loss value change diagram of the dataset of the improved BlendMask, the blue line chart is the loss value change diagram of the dataset of the traditional BlendMask, and Figure 7 left, middle, right show the first step, the third step, and the 29th step of the first epoch, respectively. (The training is 30 epochs at a time, and there are 30 steps for each epoch).

As shown in Figure 7, the loss value after improvement is obviously optimized.

### *5.4. Experiment of Image Knowledge Graph Construction*

We took the image results recognized by the algorithm as the experimental dataset, which contained 5077 images and no more than 10 entities. In addition, we divided the dataset into three parts, in which 70% of the images and their corresponding entities and relationships were used as the training set, 20% of the images were reserved for testing, and 10% of the images were used as the verification set.

Evaluation metrics. In order to verify the effectiveness of our proposed method, we used the traditional SGG index as our evaluation index. We used zR@K and ng-zR@K as our evaluation indicators, and a detailed introduction to them follows:

- ZR@K: Here, zR@K is the abbreviation of zero-shot Recall@K. Recall@K was first put forward by Lu and others, and it is the most accepted evaluation index. In SGG, the entity-to-relationship may be incomplete, and we need to regard the relationship analysis problem as a retrieval problem, so it is very appropriate for us to choose Recall@K as our evaluation index.
- Ng-zR@K: Here, ng-zR@K is the abbreviation of the zero-shot no-graph-constraint Recall@K, and No-graph-constraint Recall@K was first proposed by Newel and others. There are multiple relationships in an entity in SGG, so it is very appropriate for us to choose Recall@K as our evaluation indicator.

Performance evaluations: We chose the NM, TDE, and COACHER-N+P models for experiments; the representation without * uses the original dataset, and the representation with * uses the image recognized by our algorithm. As can be seen from Table 3, our relationship-building method took ZR@K and ng-ZR@K as evaluation indicators, which showed a good improvement in the NM, TDE, and COACHER-N+P baseline models. Our proposed method, based on the accuracy of object image recognition, explores the similarity between instances of the same species to establish a triplet of similarity between instances.

**Table 3.** Performance (%) on the dataset.

| Method | zR@20 | zR@50 | zR@100 | ng-zR@20 | ng-zR@50 | ng-zR@100 |
|---|---|---|---|---|---|---|
| NM | 13.04 ± 0.06 | 19.03 ± 0.22 | 21.98 ± 0.22 | 15.15 ± 0.49 | 28.78 ± 0.57 | 41.54 ± 0.79 |
| NM* | **13.21 ± 0.04** | **19.13 ± 0.18** | 21.84 ± 0.23 | **15.19 ± 0.47** | **28.80 ± 0.55** | 41.49 ± 0.76 |
| TDE | 8.39 ± 0.25 | 14.35 ± 0.27 | 18.04 ± 0.46 | 9.82 ± 0.33 | 19.28 ± 0.56 | 28.99 ± 0.44 |
| TDE* | 8.30 ± 0.25 | **14.40 ± 0.20** | **18.14 ± 0.42** | **9.89± 0.35** | 19.24 ± 0.56 | 28.96 ± 0.42 |
| COACHER-N+P | 13.41 ± 0.28 | 19.33 ± 0.27 | 22.24 ± 0.29 | 15.54 ± 0.27 | 29.30 ± 0.27 | 41.42 ± 0.22 |
| COACHER-N+P* | **13.48 ± 0.30** | **19.34 ± 0.25** | 22.21 ± 0.19 | **15.59 ± 0.27** | 29.30 ± 0.27 | 41.36 ± 0.21 |

## 6. Conclusions

With the continuous development of artificial intelligence technology, the number of images is increasing exponentially, and image detection and recognition technology is more widely used. A knowledge graph is a software tool to transform the vast amount of information on the Internet into more coherent and easier-to-understand knowledge. It also provides great help for human beings to organize, manage, and understand the information on the Internet. In this paper, on the basis of the original hybrid mask algorithm, the idea of a hole convolution kernel is improved, which expands the receiving range and improves the accuracy of mask prediction without changing the convolution result of the convolution kernel.

At the same time, it is managed by using a knowledge graph. This is a major exploration of the future of machine learning, and it is also a vital part of big data management research and development in the era of the Internet of everything. In the following research, the author will explore more feature factors that can be considered in the process of image knowledge graph construction so that the image knowledge graph can be more accurately applied to more image fields.

**Author Contributions:** Conceptualization, L.C. and X.X.; methodology, L.C.; software, J.L.; validation, Q.C., X.H. and Y.M.; data curation, Q.C.; writing—original draft preparation, X.H. and Y.M.; writing—review and editing, L.C.; visualization, J.L.; supervision, L.C. and J.L.; project administration, X.X. 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 used to support the findings of this study can be made available by the corresponding author upon request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

| Abbreviation | Full name |
|---|---|
| CNN | Convolutional Neural Network |
| FCNS | Fully Convolutional One-Stage Object Detection |
| SVM | Support Vector Machines |
| SIFT | Scale Invariant Feature Transform |
| YOLO | You Only Look Once |
| CPM | Calibrating Point-guided Misalignment in Object Detection |
| DetCo | Unsupervised Contrastive Learning for Object Detection |
| SCN | Self-Cure Network |
| FCN | Fully Convolutional Network |
| FCIS | Fully Convolutional Instance-aware Semantic Segmentation |
| YOLACT | You Only Look At CoefficienTs |
| FPN | Feature Pyramid Network |
| RDF | Resource Description Framework |
| D2R | Database to RDF |
| FCOS | Fully Convolutional One-Stage Object Detection |

## References

1. Ren, S.; He, K.; Girshick, R.; Jian, S. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
2. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. *Simultaneous Detection and Segmentation*; Springer: Cham, Switzerland, 2014.
3. Ullah, A.; Xie, H.; Farooq, M.O.; Sun, Z. Pedestrian Detection in Infrared Images Using Fast RCNN. In Proceedings of the 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), Xi'an, China, 7–10 November 2018; pp. 1–6.
4. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Yan, Y. Blendmask: Top-down meets bottom-up for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
5. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT++ Better Real-Time Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1108–1121. [CrossRef] [PubMed]
6. Viola, P.A.; Jones, M.J. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001.
7. Zhang, J.; Ye, G.; Tu, Z.; Qin, Y.; Qin, Q.; Zhang, J.; Liu, J. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* **2022**, *7*, 46–55. [CrossRef]
8. Wu, C.; Shi, S.; Hu, J.; Huang, H. Knowledge-enriched joint-learning model for implicit emotion cause extraction. *CAAI Trans. Intell. Technol.* **2023**, *8*, 118–128. [CrossRef]
9. Zhang, Y.; Zhang, S.; Huang, R. Combining deep learning with knowledge graph for macro process planning. *Comput. Ind.* **2022**, *140*, 103668. [CrossRef]
10. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005.
11. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionm, Boston, MA, USA, 7–12 June 2015.

12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Zhu, B.; Song, Q.; Yang, L.; Wang, Z.; Liu, C.; Hu, M. CPM R-CNN: Calibrating point-guided misalignment in object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3248–3257.
14. Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; Luo, P. Detco: Unsupervised contrastive learning for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 5–9 January 2021; pp. 8392–8401.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
18. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically designing CNN architectures using the genetic algorithm for image classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [CrossRef] [PubMed]
19. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
20. Xu, J.; Pan, Y.; Pan, X.; Hoi, S.; Yi, Z.; Xu, Z. RegNet: Self-regulated network for image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–6. [CrossRef] [PubMed]
21. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
22. Lee, D.H.; Yoo, J.H. CNN Learning Strategy for Recognizing Facial Expressions. *IEEE Access* **2023**, *11*, 70865–70872. [CrossRef]
23. Dai, J.; He, K.; Li, Y.; Ren, S.; Sun, J. Instance-sensitive fully convolutional networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
24. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
25. Zhou, X.; Yang, X.; Ma, J.; Wang, I.K. Energy efficient smart routing based on link correlation mining for wireless edge computing in iot. *IEEE Internet Things J.* **2021**, *9*, 14988–14997. [CrossRef]
26. Zhou, X.; Xu, X.; Liang, W.; Zeng, Z.; Yan, Z. Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart iot. *IEEE Internet Things J.* **2021**, *8*, 12588–12596. [CrossRef]
27. Zhou, X.; Liang, W.; Li, W.; Yan, K.; Wang, I.K. Hierarchical adversarial attacks against graph neural network based iot network intrusion detection system. *IEEE Internet Things J.* **2021**, *9*, 9310–9319. [CrossRef]
28. Lu, X.; Wang, W.; Shen, J.; Crandall, D.J.; Van Gool, L. Segmenting Objects from Relational Visual Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7885–7897. [CrossRef] [PubMed]
29. Ke, L.; Tai, Y.W.; Tang, C.K. Occlusion-Aware Instance Segmentation Via BiLayer Network Architectures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10197–10211. [CrossRef] [PubMed]
30. Wang, E.; Yu, Q.; Chen, Y.; Slamu, W.; Luo, X. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Inf. Fusion* **2022**, *88*, 78–85. [CrossRef]
31. Guo, Q.; Cao, S.; Yi, Z. A medical question answering system using large language models and knowledge graphs. *Int. J. Intell. Syst.* **2022**, *37*, 8548–8564. [CrossRef]
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
33. Nigam, V.V.; Paul, S.; Agrawal, A.P.; Bansal, R. A review paper on the application of knowledge graph on various service providing platforms. In Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India, 29–31 January 2020.
34. Maclean, F. Knowledge graphs and their applications in drug discovery. *Expert Opin. Drug Discov.* **2021**, *16*, 1057–1069. [CrossRef] [PubMed]
35. Zhou, K.; Zhan, Y.; Fu, D. Learning Region-Based Attention Network for Traffic Sign Recognition. *Sensors* **2021**, *21*, 686. [CrossRef] [PubMed]