Hindawi
Advances in Multimedia
Volume 2023, Article ID 9763562, 1 page
https://doi.org/10.1155/2023/9763562

*Retraction*

# Retracted: Target Recognition Technology of Multimedia Platform Based on a Convolutional Neural Network

**Advances in Multimedia**

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023



This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] J. Liu and J. Zhang, "Target Recognition Technology of Multimedia Platform Based on a Convolutional Neural Network," *Advances in Multimedia*, vol. 2022, Article ID 8188936, 10 pages, 2022.

Hindawi
Advances in Multimedia
Volume 2022, Article ID 8188936, 10 pages
https://doi.org/10.1155/2022/8188936



*Research Article*

# Target Recognition Technology of Multimedia Platform Based on a Convolutional Neural Network

**Jie Liu**[1] **and Jiamin Zhang**[2]

[1]*Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan Shanxi 030008, China*
[2]*University Medical school, Zhangjiakou Hebei 075000, China*

Correspondence should be addressed to Jie Liu; liuj@tit.edu.cn

Received 29 August 2022; Accepted 27 September 2022; Published 18 November 2022

Academic Editor: Tao Zhou



With the rapid development of the Internet, network media, as a new form of information dissemination, has penetrated into people's daily life. In recent years, with the rapid transformation of Chinese social structure and the rise of self-media platforms, various social contradictions have been highlighted in the form of online public opinion. Especially on online multimedia platforms, the spread of online public opinion is more rapid, which can easily lead to social hotspots. In order to effectively supervise the public opinion information on the Internet, it is necessary to identify the target of the information on the multimedia platform and effectively screen the information, so as to control the network public opinion in the development stage. Aiming at the above problems, we propose a multitarget retrieval method based on a convolutional neural network, which uses multitarget detection algorithm to locate multitarget regions and extract regional features and uses cosine distance as a similarity measure for multitarget recognition. In view of the slow feature extraction speed of VGG model, a lightweight mobile network model is proposed to replace the original VGG model on the mobile phone to reduce the retrieval time and realize the recognition of specific targets on the multimedia platform, and it is applied to the verification of image recognition on the multimedia platform. The results show that the algorithm proposed in this paper has great advantages in multitarget recognition tasks.

## 1. Introduction

Since the 18th National Congress of the Communist Party of China, in order to strengthen and improve the propaganda and ideological work, the General Secretary Xi has put forward such important expositions as "taking the work of online public opinion as the top priority of the propaganda and ideological work," "carrying out in-depth online public opinion struggle," "mastering the initiative in this public opinion battlefield as soon as possible," and "creating a clean and honest cyberspace for the vast number of Internet users." He made the important judgment that "the Internet has become the main battlefield of public opinion struggle" from a strategically advantageous position [1]. This requires the government to create a clear Internet environment for hundreds of millions of people with a high sense of political responsibility and mission and build a good Internet Ecology for realizing the Chinese dream of the great rejuvenation of the Chinese nation. Therefore, strengthening network public opinion monitoring work should be an important part of government work [2].

At present, the Internet is changing the public opinion pattern with unprecedented depth and breadth, affecting people's ideas and value judgments. The main body, mode, channel, and influence of public opinion communication have undergone fundamental changes. The open and interactive communication on PC side has continuously increased the impact on traditional media, and three-dimensional communication has rushed in, which has brought severe challenges to propaganda and ideological

work, especially online public opinion work. Online public opinion venues mainly include websites, web pages, and forums that publish various types of information. Remarks reflect the opinions of netizens through browsing, threading, and forwarding. The characteristics and changes it presents are changing. Controlling the position where public opinion occurs is the basis for improving the monitoring mechanism of public opinion outbreak and the stage of control and guidance [3]. Whether online speech becomes online public opinion is largely determined by the sensitivity and activity of the topic. In the process, it is gradually developing towards network public opinion. This stage is the best time to control and guide. Through the target identification of the Internet multimedia platform, it is possible to effectively guide the potential guidance in advance.

At present, the government's system is not perfect. Due to the imperfect early warning mechanism, governments at all levels are mostly passive in responding to online public opinions when major social incidents occur. When government departments are dealing with online public opinion work, they are mostly individual soldiers. According to the division of responsibilities, each department is responsible for monitoring the content within the scope of its own department's responsibilities, this method is extremely inefficient and has poor accuracy, and it is difficult to effectively provide a comprehensive and accurate reference for decision-making [4]. Study the monitoring analysis and governance strategy of Internet public opinion, and formulate a reasonable monitoring mechanism and governance process. On the one hand, we can find Internet public opinion information at the fastest speed and take corresponding countermeasures; on the other hand, it can also take the initiative to guide public opinion and form a good situation of concerted efforts and harmonious development.

Target recognition is a key technology for Internet multimedia platform to supervise internet content such as network public opinion. Through the target recognition of the multimedia platform, strategies such as improving the quality of decision-makers, increasing participation in decision-making, improving the decision-making system, and optimizing the research and judgment mechanism of network public opinion can effectively supervise the network [5].

Therefore, this paper proposes a multitarget retrieval method based on convolutional neural network, which uses multitarget detection algorithm to locate multitarget areas and extract regional features and uses cosine distance as the similarity measure of multitarget recognition to realize the recognition of specific targets on the multimedia platform and applies it to the verification of image recognition on the multimedia platform. The results show that the algorithm proposed in this paper has great advantages in the task of multitarget recognition.

## 2. Materials and Methods

The frequent use of many platforms, such as Facebook, Twitter, Weibo, and WeChat, has resulted in data sources high output, which makes the effective management and efficient retrieval of image information resources particularly important. How to accurately and efficiently retrieve and return the images and videos required by users from a large-scale multimedia resource library with rich visual and semantic information is the current research with a wide range of applications in areas such as intelligent video surveillance, robot environment perception, and large-scale image retrieval [6]. Currently, the target method based on a convolutional neural network is a main target detection and recognition method, which uses learned recognition. The advantage of convolutional neural network is that it can make use of the advantage of self-learning features to make the feature expression ability and classification ability better than traditional target detection and recognition methods. Therefore, the target detection and recognition method based on convolutional neural network has achieved high accuracy.

*2.1. Target Detection and Recognition Methods.* Target detection and recognition is a very important research direction in the field of artificial intelligence. Its purpose is to use computers to identify the types of targets in an image and give the location of their bounding boxes. The recognition accuracy will be reduced due to the occlusion of the target or the change of the angle of view. At the same time, the optimization of convolutional neural network and regional suggestion network and the computing capacity of the computer will affect the speed of target recognition and detection: to detect which type for object the object is and target recognition is to detect which category of the target to be detected is. Target detection will be like this. In practice, target detection and target recognition algorithms usually have generality and continuity [7].

(1) Such main idea of using template matching for target detection and recognition is to first make a small number of target images as template images and then match the subimages in the detection images with the template images, and then count whose similarity exceeds a certain threshold as the target. (Figure 1)

The working mode of template matching is roughly as follows: by translating the template image block on the matching image, matching is performed between the actual image block and the template image. This seems very simple, but in practice, we must consider how to deal with the spatial coordinate transformation between images. For example, if a rotation transformation occurs, the template image cannot find an image with the same size and shape but different angles on the matching image, which undoubtedly makes the matching problem more complicated and difficult to solve.

For template matching and postprocessing [8]: in the first step, the preprocessing process is to perform image denoising, image enhancement, color space conversion, and other operations on the detected image. The second step is the process. The third step is the process of feature extraction, which uses a specific algorithm to extract features from subimages. In the fourth step, the template matching process

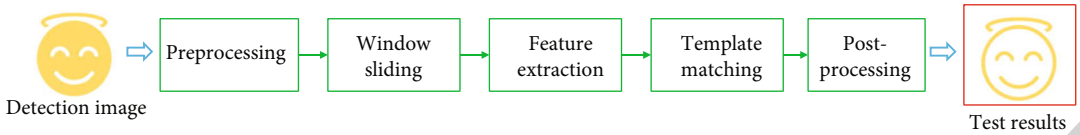


Figure 1: Schematic diagram of the target detection method based on template matching.

uses a specific distance measurement algorithm to calculate the similarity between the subimage features and the template image features and determines the candidate regions; otherwise, the candidate regions are excluded. In the fifth step, the postprocessing process is to merge the candidate regions intersecting with the same type of template and calculate the bounding box of each target.

Although template matching has received more attention in the early days, with the deepening of research, template matching is limited in application scenarios, mainly for the following three reasons [9]. First, because this type of method usually selects features that describe the shape of the target as the template matching features, such as edge features, texture features, and gradient features, the matching effect of targets with complex internal textures is poor. Then, because this type of method of two images is strictly according to the distance metric method, the matching effect of the target with local deformation is poor. Finally, because this type of method needs to collect some representative images for each category and make them as template images, if the detected target categories are added, template images will be added accordingly, slower. This increases the likelihood of false positives. Therefore, target detection is only suitable for simple scenarios [10].

(2) In the object detection method based on image classification

Image classification refers to the task of determining categories for input images from a set of fixed categories. Image classification is an important basic problem in computational vision and is the basis of computer vision tasks such as target detection, saliency segmentation, behavior analysis, and target tracking. Face recognition in the security field, traffic scene recognition in the traffic field, and image recognition in the medical field are all specific applications of image classification.

In complex scenes, researchers have proposed method image classification. Determine the category of each subimage in the detected image, and finally determine the subimage whose output value exceeds a certain threshold as the target [11] (Figure 2).

Recognition based on image classification is mainly divided into six steps: preprocessing, window sliding, feature extraction, feature selection, feature classification, and postprocessing [12]. The process of the first three steps is the same as the template matching-based method. The fourth step, the process of feature selection, is to select representative features from the feature vector, improve the robustness of the feature, and reduce the dimension of the feature. The fifth step, the process of feature classification, is to use a specific classifier to classify the features. If the output value exceeds a certain threshold, the candidate area is the target and the category of the target is determined; otherwise, the candidate area is the background. The sixth step, the postprocessing process, is to merge the candidate regions that are determined to be intersected by the same category; then, calculate it.

Such image classification focuses on how to improve the expressive ability and antideformation ability of features during feature extraction and how to improve the accuracy and calculation speed of classifiers during feature classification. As a result, researchers have proposed a wide variety of features and various forms [13]. First, because the convolutional neural network usually performs convolution and pooling operations on the input image alternately, the features undergo multiple nonlinear transformations and are gradually abstracted and features have a certain degree of separability. Third, because the dataset on which the convolutional neural network is trained contains samples of multiple target classes, the features are not specific to a specific target class. To sum up, since the features extracted by convolutional neural networks overcome the shortcomings of hand-crafted features, target detection and recognition methods are based on convolutional ones [14].

*2.2. Convolutional Neural Network and Its Structure.* When people perceive external things, they usually go from local to global, and there is a similar pattern in the pixel space of an image; that is, the relationship between pixels that are close to each other is closer than that of pixels that are far away. Therefore, the convolutional neural network does not perceive the global pixels but uses the convolution kernel to perceive the image information locally. At the same time, the multilayer convolution is used to summarize the locally learned image. These features improve the performance of the model.

It is downsampling [15]. Such neuron in the network is only connected to some neurons to perceive the local information of the image, and local perception also greatly reduces the weight parameters that need to be trained. Weight sharing means that each group of connections in a convolutional neural network no longer has its own weight but shares a convolution kernel parameter, because a convolution kernel can apply a feature such as an edge after sensing it, where other parts of the image have similar features. Downsampling is because the pooling layer can avoid overfitting.

Typically, there are applications [16] (Figure 3).

A convolutional neural network is different from an ordinary neural network. The unique basic model architecture of

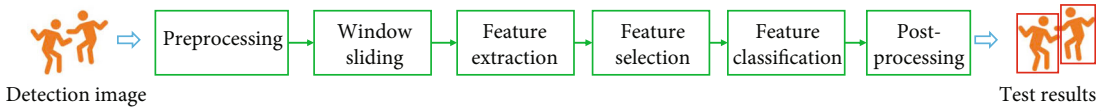


Figure 2: Schematic flow chart of the method of target detection and recognition based on image classification.

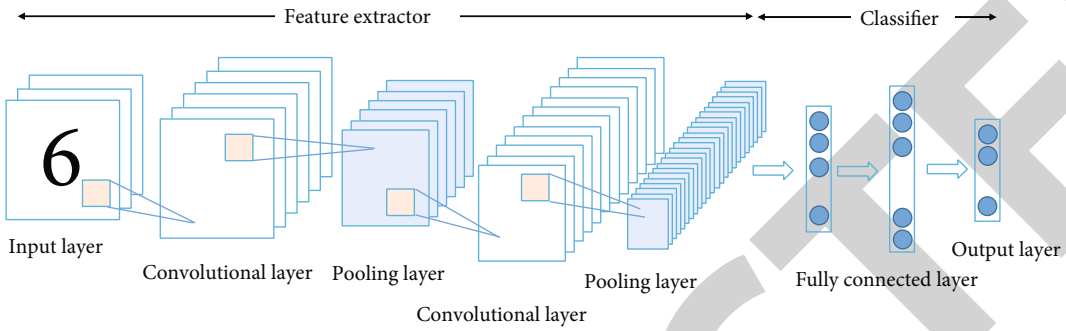


Figure 3: Basic structure of convolutional neural network.

convolutional neural networks includes a convolution layer, pooling layer (also known as lower sampling layer), and fully connected layer (FC). A convolutional neural network includes a feature extractor. The feature extractor is composed of a convolution layer and subsampling layer. A neuron in the convolution layer connects local adjacent neurons. In most cases, a CNN convolution layer contains several feature planes (feature maps). Each feature map is composed of neurons arranged in a rectangle. All neuron weights of a whole feature plane are shared. The shared weights are called convolution cores. Subsampling, also known as pooling, has two forms: max pooling and mean pooling. Subsampling is similar to a special convolution process.

Therefore, according to the function of each layer, the convolutional neural network can be divided into two parts [17] (Figure 4).

Convolution is a special weighted summation method in mathematical analysis, which can also be regarded as a filtering calculation. The two-dimensional convolution process on an image can be understood as multiplying and summing the image using a matrix; such a matrix is often called a window or convolution kernel [18]. Figure 4(a) is a schematic diagram of the process of performing a convolution operation with an input matrix of size $4 \times 4$ and a convolution kernel of matrix size $2 \times 2$ and stride 1. The formula for the convolution operation is

$$h_0 = \frac{h_1 + 2 \times \text{padding\_size\_}h}{\text{stride}}, \tag{1}$$

$$w_0 = \frac{w_1 + 2 \times \text{padding\_size\_}w}{\text{stride}}. \tag{2}$$

The sigmoid function and the tanh the specific forms are as follows:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}, \tag{3}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{4}$$

The image features extracted by convolution operation are linear, but the real samples are often nonlinear. Therefore, a nonlinear function is introduced to solve this problem. Activate the function, so that each pixel can be represented by any value from 0 to 1, simulating more subtle changes. Activation functions are generally nonlinear, continuously differentiable and monotonic. For the activation function, there is also a modified linear unit (ReLU) as the activation function of the convolutional layer. The ReLU function passes positive values directly and sets negative values to zero. Therefore, the ReLU function is a piecewise function, and the expression is as follows:

$$\text{ReLU} = \max(x, 0). \tag{5}$$

Pooling is a downsampling method, and common choices and the result are obtained by sliding a window on the image, but unlike the convolution layer, the parameters in the pooling window matrix are artificially set. Figure 4(b) is a schematic diagram of the max pooling process with a window for a matrix of size $4 \times 4$.

$$L_P(x) = \left( \sum_{x(u,v) \in x} x(u,v)^P \right)^{1/P}. \tag{6}$$

It can be seen from the figure that the pooling operation is simpler and the result is simpler than the result of the convolution output. In the maximum pooling operation, a window is used to slide the image. In each step, the maximum value of all values in the selected area in the window is used as the output of this step, and finally, all the outputs are collected, as the result of max pooling. The average pooling operation uses the average of all values in the window as

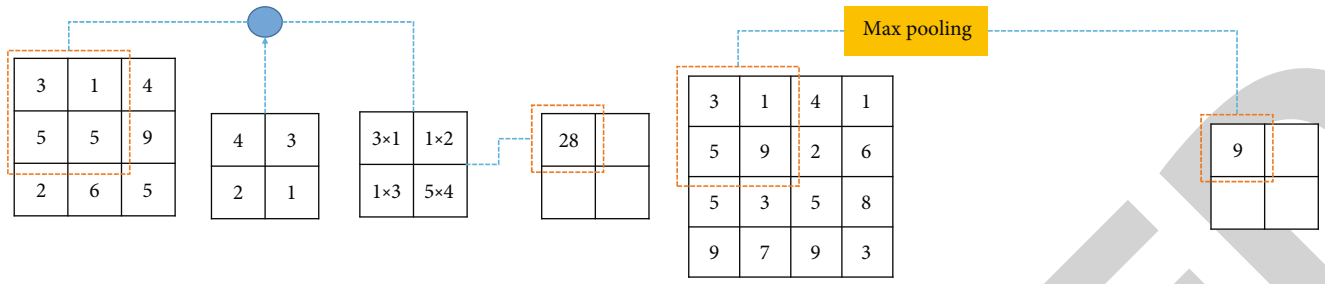


Figure 4: Two operations of convolutional neural network: (a) convolution operation and (b) maximum pooling operation.

the pooling result at each step [19].

$$x_j^l = f\left(w_j^l x^{l-1} + b_j^l\right). \tag{7}$$

Among them, $f(x)$ represents the nonlinear activation function.

The form of the output layer is task-oriented. If the convolutional neural network is used as a classifier, the output layer will generate a prediction vector representing the image category, and the probability of the target being recognized as a certain type of target is obtained by the size of the prediction vector. Assuming that the output layer is set to Softmax regression, then each component $y$ in the prediction vector is calculated as

$$y_j = \frac{e^{-w_l^L x^{L-1}}}{\sum_{i=1}^{M} e^{-w_l^L x^{L-1}}}. \tag{8}$$

*2.3. Target Detection and Recognition Algorithm Based on Convolutional Neural Network.* The most important tasks or recognition algorithms are locating the specific location of the object and identifying the specific category of the object. The target algorithm based on deep learning is designed on the basis of convolutional neural network. This series of algorithms inherits all the advantages of convolutional neural network, such as the effectiveness of convolutional feature extraction and the robustness of convolutional neural network. The commonly used target algorithms will be here, namely, target-based algorithms [20].

From the earliest R-CNN model to Fast-RCNN to the later Faster-RCNN, the R-CNN model has been continuously developed into a series, and these series all belong to the category candidate regions. R-CNN is the first algorithm to apply CNN, and such ability of CNN to extract features is outstanding, which significantly improves the effect of target detection and recognition [21].

(1) On the original image (~2 k)

(2) Stretch of the candidate area into a uniform size, and use CNN to extract convolution features

(3) Input the extracted for target recognition and classification

(4) Use the regressor to predict the bounding box and complete the target positioning

R-CNN has many problems, such as the large number of candidate regions generated on the original image, which increases the amount of calculation and reduces the real-time performance: the stretching of the candidate region leads to the loss of some information, which affects the detection accuracy. Fast-RCNN introduces target region pooling repeated calculation of candidate regions. The classification and localization tasks are simultaneously performed during the training process, which greatly reduces the training and testing time and meets people's requirements for real-time performance.

The YOLO algorithm is a milestone in the history. It integrates the idea of step-by-step detection in R-CNN and uses a convolutional neural one. At the same time, the target classification and regression positioning in R-CNN are also integrated and it absolutely meets people's requirements for real-time performance.

The SSD algorithm has relatively high detection accuracy among all the target detection algorithms mentioned above, but because the SSD algorithm uses low-level features for target detection, many effective small target information is ignored, so there is a very fatal problem [22]. Small targets are less robust. In addition, multitarget feature extraction, and then obtains more feature maps for detection by adding convolutional layers on this basis. The redundant expression of the traditional convolution kernel parameters of the VGG model and the newly added feature extraction layer hinder the improvement of the detection speed to a certain extent [23].

## 3. Results and Discussion

With the introduction of the concept of rapid popularization and convolutional neural networks, it has also proliferated rapidly. Throughout the entire research process, the current research on it. Among them, "two-step" mainly represents the R-CNN series of algorithms. The steps of this series of algorithms are basically the same, generally generating candidate regions to extract convolution features and then using classifiers and regressors for identification and positioning. "One step" mainly represents the YOLO algorithm and the SSD algorithm [24]. This type of algorithm improves the

idea of the R-CNN series of step-by-step detection and uses a CNN for end-to-end target detection, which effectively improves the detection accuracy and speed.

*3.1. Multiobjective Algorithm.* The Mobile Net convolutional neural network replaces the traditional convolution with a depthwise separable convolution and improves the computational speed. It is widely used in mobile terminals.

The multitarget algorithm locates and extracts the regional features and uses the cosine distance as a similarity measure to perform multitarget retrieval. Aiming at the slow feature extraction speed of the VGG model, a lightweight Mobile Net model on mobile phones is proposed to replace the original VGG model to reduce the retrieval time. The multiobjective algorithm is proposed based on the idea of feature fusion, and its network architecture is shown in Figure 5. This framework replaces the basic VGG with a lightweight Mobile Net network, removes the fully connected layer and Softmax layer of Mobile Net, and adds four convolutional layers conv14, conv15, conv16, and conv17.

*3.2. Target Recognition Algorithm Based on Multitarget Algorithm.* Such algorithm is used to locate the candidate region and feature extraction of the multitarget, introduces the multitarget region, and designs the similarity measurement method to obtain the target recognition result more reasonably.

In this paper, the extraction of multitarget regions is here. The sizes of the six feature maps are {19, 10, 5, 3, 2, l}, respectively, and the default number of boxes is {3, 6, 6, 6, 6, 6).

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), \quad k \in [1, m]. \tag{9}$$

Such box is as follows:

$$\text{Width} = s_k\sqrt{a_r}, \tag{10}$$

$$\text{Height} = \frac{s_k}{\sqrt{a_r}}. \tag{11}$$

After calculation, the 6 convolution feature maps after the second convolution can finally get 1917 default boxes, but usually, one target will be framed by multiple default boxes; there is a lot of overlap between the default boxes, and there is redundancy in multitarget positioning. Therefore, a nonmaximum suppression method (NMS) is introduced to filter a large number of overlapping default boxes. NMS calculates the confidence values of multiple default boxes for the same target, then arranges the confidence values, and then removes the default boxes with low confidence values and large overlapping areas. The measure of the overlapping area is the intersection-over-union ratio (IoU), and the specific calculation formula is as follows:

$$\text{IoU} = \frac{\text{region}(r_1 \cap r_2)}{\text{region}(r_1 \cup r_2)}. \tag{12}$$

The strategy of multitarget region extraction in this paper is as follows: first, sort the confidence values of all default boxes in descending order, and select the high-scoring default boxes with confidence values higher than 0.6. Then, select the highest score box from high to low in turn, traverse other boxes, and remove the default box with IOU ≥ 0.7 of the current high score box. Keep repeating this operation, and get all the default boxes of the multitarget area after the recursion is completed.

The loss function of the multiobjective algorithm is the weighted sum of the classification loss and the localization loss, namely,

$$L(x, c, l, g) = \frac{1}{N}(L_{\text{conf}}(x, c) + \alpha L_{\text{loss}}(x, l, g)). \tag{13}$$

The classification loss uses the Softmax function, the localization loss uses the smooth L1 function, and the calculation formulas are as follows:

$$L_{\text{conf}}(x, c) = -\sum_{i\in\text{pos}}^{N} x_{ij}^{p} \log\left(c_i^p\right) \sum_{i\in\text{neg}} \log\left(c_i^0\right), \tag{14}$$

$$L_{\text{loss}}(x, l, g) = \sum_{i\in\text{pos}}^{N} \sum_{i\in(cx,cy,w,h)} x_{ij}^{k} \text{smooth}_{L1}\left(l_i^m - g_j^m\right). \tag{15}$$

In the conv_11 feature map, the feature maps corresponding to the $N$ target default regions are found according to the positioning information, and they are taken out. First, all feature maps are subjected to dimensionality reduction processing, and then fully connected into a 1024-dimensional feature vector. Assuming that the $Q$ table represents the query image and $I$ represents the image in the dataset, the cosine distance between $Q$ and $I$ is used as the similarity measure. The larger the cosine distance, the higher the similarity between $Q$ and $'$, and the lower the similarity. In order to facilitate the cosine distance calculation later, $L_2$ normalization is performed on each vector in the multitarget feature vector set feabox. The normalization formula is as follows:

$$\text{Feabox}_i = \frac{\text{feabox}_i}{\|\text{feabox}_i\|_2}. \tag{16}$$

Assuming that the feature set of the target area of $Q$ is $\text{feabox}_0$ and the feature set of the target area of $I$ is $\text{feabox}_1$, the cosine distance calculation formula is as follows:

$$\text{dist} = \cos\left(\theta_{\text{feabox}}\right) = \frac{\text{feabox}_Q \times \text{feabox}_I}{\|\text{feabox}_Q\| \times \|\text{feabox}_I\|}. \tag{17}$$

## 4. Result Analysis and Discussion

*4.1. Algorithm Performance Evaluation Criteria.* The performance evaluation indicators of image retrieval or Mean

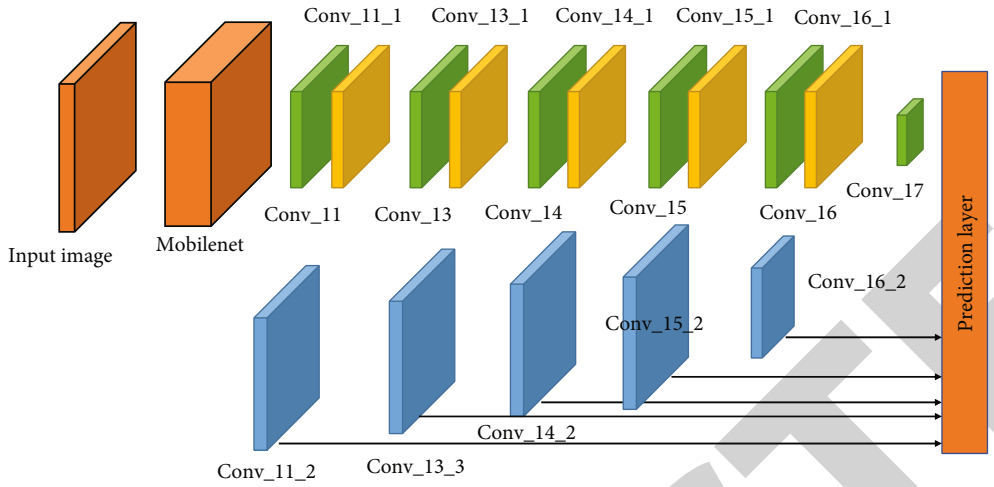


Figure 5: Multitarget detection algorithm network architecture diagram.

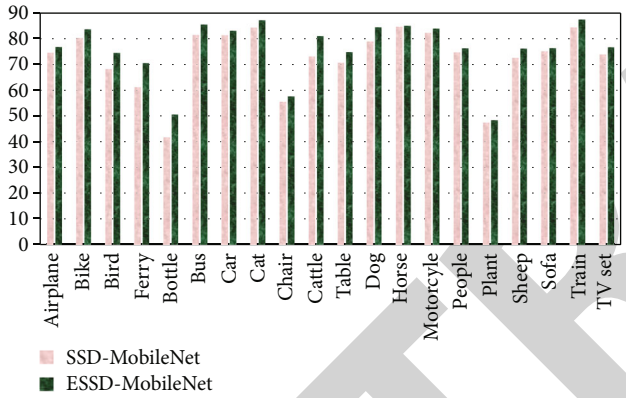


Figure 6: Comparison of AP results for identification of different target categories (changed to grouped histogram).

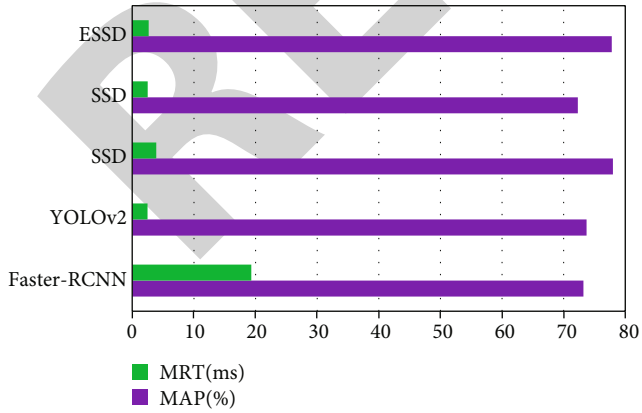


Figure 7: Comparison of image retrieval results of different target recognition algorithms (changed to bar graph).

Retrieval Time (MRT). The specific formulas are as follows:

$$\mathrm{AP} = \frac{\text{Number of target images retrieved}}{\text{The total number of images in the library that contain the target}} \times 100\%, \tag{18}$$

$$\mathrm{MAP} = \frac{\text{Average precision over multiple searches}}{\text{Number of retrievals}} \times 100\%, \tag{19}$$

$$\mathrm{MRT} = \frac{\text{Total retrieval time for the query image}}{\text{Query the number of images}}. \tag{20}$$

At the same time, the performance evaluation of the algorithm also uses the machine learning as the evaluation criteria.

*4.2. Analysis of Experimental Results.* This paper uses the trained SSD-MobileNet model as the pretraining model, and the improved ESSD-MobileNet model is retrained on the training sets of the VOC 2007 and VOC 2012 datasets. The original SSD-MobileNet and the improved ESSD-MobileNet multitarget image retrieval method are experimentally compared and analyzed on the test set of the VOC 2007 dataset. The AP results of 20 target categories are shown in Figure 6.

ESSD-MobileNet method in this paper has increased in 20 target categories with an average increase of 5.5%. Among them, the retrieval accuracy of bird, boat, bottle, and cow small-sized target categories under the improved ESSD-MobileNet model is 74.48%, 70.57%, 50.62%, and 81.30%, respectively, which is a larger increase than that of the traditional method, which is 5.83%, 9.34%, 9%, and 7.84%, respectively. Such feature fusion after convolution of the original feature layer is more effective for small-scale target detection.

In addition, in order to verify the effectiveness of the new ESSD-MobileNet multitarget, under the same experimental conditions, it is compared with the widely used RCNN series

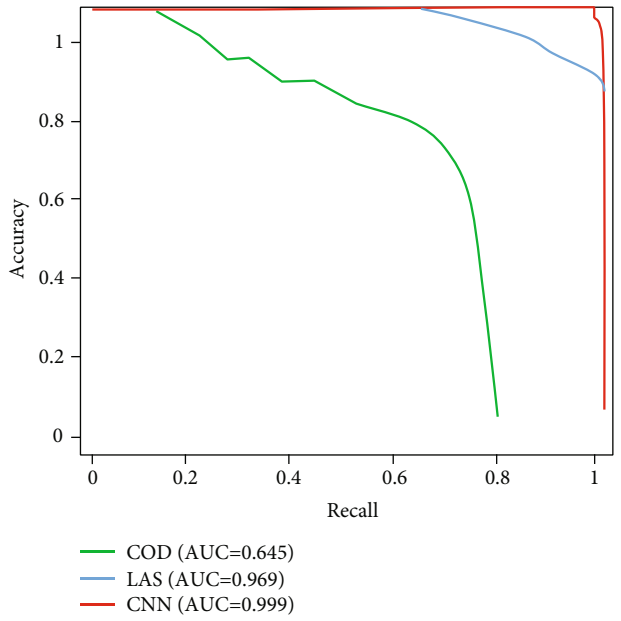


Figure 8: PR curve of the recognition method on the dataset (CNN changed to SSD-MobileNet).

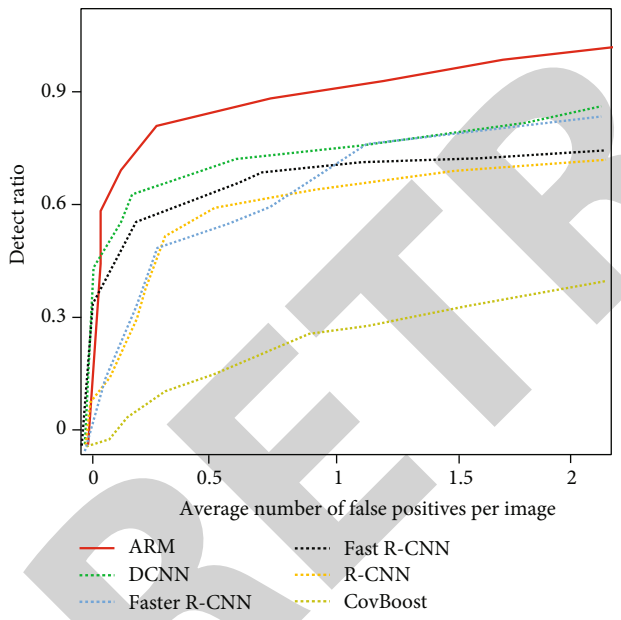


Figure 9: ROC curve of the recognition method on the dataset (ARM changed to SSD-MobileNet).

algorithms and the YOLO algorithm, and the results are shown in Figure 7.

It can be clearly seen from Figure 7 that the time spent by the algorithm in this paper to retrieve an image is slightly higher than that of the SSD-MobileNet and YOLO-Dark Net methods. This is because of the secondary convolution operation and the difference between the original feature map and the secondary convolution feature map. The fusion operation takes time. However, on the MAP index, the algorithm in this paper has achieved a great improvement, with

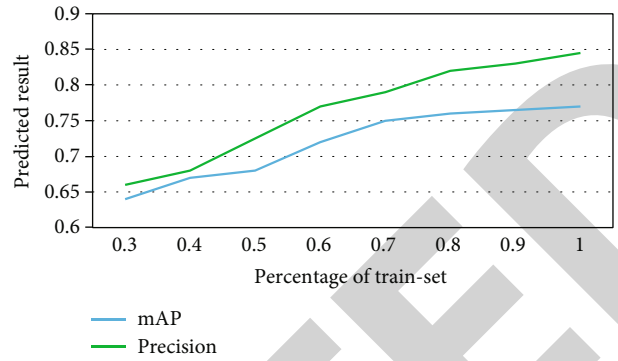


Figure 10: Using the kernel function, the effect of the training set size on the experimental results of target recognition.

an increase of 5.5% and 4.1%, respectively. Compared with the SSD-VGGl6 method, the MAP of the algorithm in this paper is basically the same. On the MRT indicator, the time it takes to retrieve an image is only 2.65 ms, which is reduced by 1.25 ms. This is because the lightweight Mobile Net solves the problem of extracting features from the VGG model slow feature. At the same time, compared with the Faster-RCNN algorithm of the RCNN series, the method in this paper not only guarantees a 4.6% increase in MAP but also reduces MRT by 17.34 ms, which has an obvious performance advantage.

The detection result is judged to be correct if and only if the coincidence rate of the target bounding box and the ground-truth bounding box exceeds 50%. In the experiment, the precision-recall (PR) curves of different target recognition methods were compared, and the area under the PR curve (AUC) was used as the evaluation standard of the recognition model. Figure 8 presents the PR curves of the three algorithms on the dataset. It can be seen from the figure that the recognition method after using the CNN algorithm has a higher PR value and has a greater advantage in the model effect.

The results of verifying the ROC performance of different algorithms on the dataset are shown in Figure 9. Since DCNN is a target detection model based on regression convolutional neural network, by comparing the target recognition effects of SSD-MobileNet and DCNN, it can be verified that Select the effectiveness of available convolution kernels. For a fair comparison, the above methods all use the same model to extract features.

As can be seen from Figure 9, the SSD-MobileNet multi-target recognition algorithm has superior performance and relatively high detection rate.

It can be achieved by controlling the number of training set samples. Specifically, by reducing the input of training set samples, randomly extract them. When using the kernel function (linear kernel), the experiment is carried out in the case, and the change of the specific prediction results is shown in Figure 10.

Such increase of the recognition curve begins to decrease, and when it reaches 80%, the overall trend begins to flatten. The function passes through the data, and its curve still has an upward trend in comparison. This shows

that compared with the data that has not been processed by the kernel function, the prediction result has a higher potential for growth under the large data set after processing with the kernel function.

## 5. Conclusion

With the comprehensive popularity of mobile electronic devices, the output of text, image, video, and audio data is amazing. How to retrieve images that meet the needs of users from the massive image data published and stored on the Internet has become the focus of current scholars. In view of the above problems, this paper proposes a multitarget retrieval method based on convolutional neural network. First of all, this paper uses ESSD Mobile Net multitarget algorithm to locate the multitarget area of the image and extract the features of the multitarget area. The cosine distance is used as the similarity measure to obtain the multitarget image recognition result. Then, the experimental analysis is carried out on the Multitarget Image Database PASCALVOC. The results show that the algorithm proposed in this paper has great advantages in the task of multitarget recognition.

## Data Availability

The figures used to support the findings of this study are included in the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2019L0931) and Taiyuan Institute of Technology Youth Academic Leader Support Program.

## References

[1] M. Afif, R. Ayachi, Y. Said, E. Pissaloux, and M. Atri, "An efficient object detection system for indoor assistance navigation using deep learning techniques," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 16601–16618, 2022.

[2] N. S. Akinshin, A. A. Potapov, R. P. Bystrov, O. V. Esikov, and A. I. Chernyshkov, "Building systems for object recognition by multichannel sensing systems based on neural networks and fractal signatures," *Journal of Communications Technology and Electronics*, vol. 65, no. 7, pp. 835–842, 2020.

[3] A. Aslam and E. Curry, "A survey on object detection for the internet of multimedia things (IoMT) using deep learning and event-based middleware: approaches, challenges, and future directions," *Image and Vision Computing*, vol. 106, no. 10, article 104095, 2021.

[4] V. Ayzenberg and S. Lourenco, "Young children outperform feed-forward and recurrent neural networks on challenging object recognition tasks," *Journal of Vision*, vol. 20, no. 11, p. 310, 2020.

[5] C. Federer, H. Xu, A. Fyshe, and J. Zylberberg, "Improved object recognition using neural networks trained to mimic the brain's statistical properties," *Neural Networks*, vol. 131, pp. 103–114, 2020.

[6] N. Dad, N. En-Nahnahi, and S. Ouatik, "Quaternion harmonic moments and extreme learning machine for color object recognition," *Multimedia Tools & Applications*, vol. 78, no. 15, pp. 20935–20959, 2019.

[7] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-based object recognition using multimodal convolutional neural networks: a survey," *IEEE Access*, vol. 7, pp. 43110–43136, 2019.

[8] Z. Gao, D. Y. Wang, Y. B. Xue, G. P. Xu, H. Zhang, and Y. L. Wang, "3D object recognition based on pairwise multi-view convolutional neural networks," *Journal of Visual Communication and Image Representation*, vol. 56, pp. 305–315, 2018.

[9] A. R. Hawas, H. A. El-Khobby, M. Abd-Elnaby, A. El-Samie, and E. Fathi, "Gait identification by convolutional neural networks and optical flow," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 25873–25888, 2019.

[10] N. Kumar and N. Sukavanam, "A cascaded CNN model for multiple human tracking and re-localization in complex video sequences with large displacement," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 6109–6134, 2020.

[11] Q. Lai, S. Khan, Y. Nie, H. Sun, J. Shen, and L. Shao, "Understanding more about human and machine attention in deep neural networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 2086–2099, 2021.

[12] B. Li, Y. Zhang, and F. Sun, "Deep residual neural network based PointNet for 3D object part segmentation," *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11933–11947, 2022.

[13] A. Lukman and C. K. Yang, "An object recognition system based on convolutional neural networks and angular resolutions," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 16059–16085, 2021.

[14] F. Pastor, J. M. Gandarias, A. J. García-Cerezo, and J. M. Gómez-de-Gabriel, "Using 3D convolutional neural networks for tactile object recognition with robotic palpation," *Sensors*, vol. 19, no. 24, p. 5356, 2019.

[15] A. Pertusa, A. J. Gallego, and M. Bernabeu, "MirBot: a collaborative object recognition system for smartphones using convolutional neural networks," *Neurocomputing*, vol. 293, no. 7, pp. 87–99, 2018.

[16] M. Sheeny, A. Wallace, and S. Wang, "300 GHz radar object recognition based on deep neural networks and transfer learning," *IET Radar Sonar ? Navigation*, vol. 14, no. 10, pp. 1483–1493, 2020.

[17] R. D. Singh, A. Mittal, and R. K. Bhatia, "3D convolutional neural network for object recognition: a review," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 15951–15995, 2019.

[18] C. Xu, J. Yang, and J. Gao, "Coupled-learning convolutional neural networks for object recognition," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 573–589, 2019.

[19] R. R. Ziyatdinov and R. A. Biktimirov, "Application of neural networks in object recognition tasks for ADAS systems," *IOP Conference Series: Materials Science and Engineering*, vol. 570, no. 1, p. 012107, 2019.

[20] L. Chen, Z. Wang, L. Zhang, H. Cheng, and D. Qi, "Description technology of fractured-vuggy carbonate reservoir in Halahatang Oil Field, Tarim Basin—take the Ha 7 test area as an example," in *Proceedings of the International Field*

*Exploration and Development Conference 2018*, pp. 1032–1042, Singapore, 2020.

[21] M. Wang, H. Cheng, J. Wei, K. Zhang, D. Cadasse, and Q. Qin, "High-temperature-resistant, clean, and environmental-friendly fracturing fluid system and performance evaluation of tight sandstone," *Journal of Environmental and Public Health*, vol. 2022, Article ID 5833491, 7 pages, 2022.

[22] Y. Guo, Y. Yang, G. Ren, J. Ni, and H. Cheng, "Analysis of pore characteristics of reservoir rock based on CT scanning—taking the Tazhong Block of Tarim Basin as an example," in *Proceedings of the International Field Exploration and Development Conference 2018*, pp. 973–984, Singapore, 2020.

[23] H. Cheng, Y. Dong, C. Lu, Q. Qin, and D. Cadasse, "Intelligent oil production stratified water injection technology," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3954446, 7 pages, 2022.

[24] Y. Bai, K. Yang, T. Mei, W. Y. Ma, and T. Zhao, "Automatic data augmentation from massive web images for deep visual recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 3, pp. 1–20, 2018.