Hindawi
Advances in Mathematical Physics
Volume 2021, Article ID 4632070, 12 pages
https://doi.org/10.1155/2021/4632070

Research Article

A Massive Image Recognition Algorithm Based on Attribute Modelling and Knowledge Acquisition

Guohua Li,[1] An Liu,[1] and Huajie Shen[2]

[1]*College of Materials Science and Engineering, Central South University of Forestry & Technology, Changsha, Hunan 410004, China*
[2]*College of Material Science and Engineering, Southwest Forestry University, Kunming, Yunnan 650224, China*

Correspondence should be addressed to Huajie Shen; shenhuajie@swfu.edu.cn

Received 11 October 2021; Revised 3 November 2021; Accepted 5 November 2021; Published 24 November 2021

Academic Editor: Miaochao Chen

In this paper, an in-depth study and analysis of attribute modelling and knowledge acquisition of massive images are conducted using image recognition. For the complexity of association relationships between attributes of incomplete data, a single-output subnetwork modelling method for incomplete data is proposed to build a neural network model with each missing attribute as output alone and other attributes as input in turn, and the network structure can deeply portray the association relationships between each attribute and other attributes. To address the problem of incomplete model inputs due to the presence of missing values, we propose to treat and describe the missing values as system-level variables and realize the alternate update of network parameters and dynamic filling of missing values through iterative learning among subnets. The method can effectively utilize the information of all the present attribute values in incomplete data, and the obtained subnetwork population model is a fit to the attribute association relationships implied by all the present attribute values in incomplete data. The strengths and weaknesses of existing image semantic modelling algorithms are analysed. To reduce the workload of manually labelling data, this paper proposes the use of a streaming learning algorithm to automatically pass image-level semantic labels to pixel regions of an image, where the algorithm does not need to rely on external detectors and a priori knowledge of the dataset. Then, an efficient deep neural network mapping algorithm is designed and implemented for the microprocessing architecture and software programming framework of this edge processor, and a layout scheme is proposed to place the input feature maps outside the kernel DDR and the reordered convolutional kernel matrices inside the kernel storage body and to design corresponding efficient vectorization algorithms for the multidimensional matrix convolution computation, multidimensional pooling computation, local linear normalization, etc., which exist in the deep convolutional neural network model. The efficient vectorized mapping scheme is designed for the multidimensional matrix convolution computation, multidimensional pooling computation, local linear normalization, etc. in the deep convolutional neural network model so that the utilization of MAC components in the core loop can reach 100%.

1. Introduction

With the rapid development of Internet technology, the carrier of information has developed from the traditional textual record to a richer multimedia record. Multimedia carriers such as image, voice, and video contain all kinds of information. Unlike textual records, which contain many abstract concepts, multimedia information content is mostly described as figurative sensory information. How to make artificial intelligence learn to understand multimedia content while correlating abstract textual semantic information with intuitive multimedia content has become a research topic of increasing interest in recent years [1]. In this paper, we focus on learning multimodal correlations between images and text, starting with basic multimodal data association, and automatically constructing large-scale image-text mapping datasets based on the complementarity between images and text to lay the foundation for subsequent research work [2]. First, start with the basic multimodal data association, relying on the complementarity between the image and the text to automatically construct a large-scale image-text mapping dataset, which lays the foundation for subsequent research work. Then, a large-scale weakly supervised data-based feature learning method for the image-text association is introduced to learn both image feature representation and text feature representation in a unified feature space and model the correspondence between them. Finally, two key applications on multimodal association learning are presented: crossmodal image retrieval and multimodal inferential visual quizzing, and a variety of different solutions corresponding to these two applications are provided. With the advancement of deep learning techniques based on images and text in their respective domains, research on the correlation analysis between the two different modalities, images and text, and their corresponding applications have become increasingly important. However, the abstract nature of the text and the figurative representation of images are very different, which makes text-image-based correlation analysis a complex learning task [3].

Imagine that the human brain perceives things by processing visual signals and speech signals simultaneously, combining the two to build cognition. This multimodal interaction is very important in the cognitive and learning process of the human brain. Moreover, this multimodal way of thinking can often directly affect the human brain's reasoning and judgment, for example, vision-based question and answer tasks need to synthesize the meaning of speech signals and visual information to establish a connection and then reason [4]. If we can well solve the multimodal association learning tasks based on images and texts, this will provide two major areas of technology integration and mutual enhancement of images and texts. Study the representation of product design knowledge and modelling of product design tasks. The modelling of design tasks is defined based on knowledge modelling to integrate the design task space and knowledge space effectively. Image semantic modelling technology is a requirement of the times. Facing the massive visual data generated daily on the Internet, efficient data processing and analysis techniques are important research topics that can be widely used in image and video recognition, classification, and retrieval [5]. In image semantic modelling, it is a challenging task to extract discriminative features. Introducing human visual mechanisms into image semantic modelling makes the computer perception of images more closely match human behaviour. The research on this topic is important for image processing and can be applied to many applications [6].

Semantic description of objects in an image is an effective way to address the “semantic gap.” The most important aspect of attribute learning is how to obtain the attribute labels of an image. To obtain objects with semantic information in an image, the traditional way is to manually annotate the image data. However, manual annotation is time-consuming and labour-intensive in the face of large image databases. Therefore, attribute annotation methods based on target detectors or target filters are born. Ideally, the image needs to be scanned using all the target detectors to obtain the responses of different objects so that the semantic attributes of the image can be annotated automatically. However, this process is not achievable in practice; on the one hand, there is not enough research to build sophisticated target detectors for a huge number of generic objects; on the other hand, the semantic hierarchy problem becomes acute as the number of target objects in an image increases, and not all objects in the image contributes to the semantic modelling of the image. With the progress and development of image and text based on deep learning technology in their respective fields, the research on the correlation analysis between the two different modalities of image and text and the corresponding applications has become increasingly important. It is pointed out that it is possible to annotate videos using 3000-4000 objects and achieve satisfactory results. In the context of big data, while paying attention to model accuracy should also pay more attention to the operational efficiency and deployment feasibility of algorithmic models in the big data environment, only algorithms with a better trade-off between algorithm performance and implementation efficiency can meet the practical needs. Big data puts higher demands on the software and hardware environment of computers, and deep learning models mostly require huge computing resources and efficient computing power. The existing software and hardware environment has severely restricted the research and application of deep learning-based image understanding technology, especially for the increasing number of edge devices, so that deep learning algorithms can run on edge devices to make them have intelligence as the current and future development trend in the general environment of IoT. This requires researchers to explore more efficient model training devices and endpoint inference platforms and corresponding software development platforms and efficient algorithm libraries.

2. Status of Research

Global feature-based modelling algorithms have the advantages of good invariance, computational simplicity, etc., which describe the overall properties of an image, such as colour, texture, and shape features. In general, global features represent the image as a fixed-length feature vector for task learning purposes [7]. Colour- and shape-based feature fusion and Euclidean distance approach is proposed for image retrieval, and the image database used for the experiments contains 150 colour images and 250 grayscale images [8]. The results show that the integrated colour- and shape-based feature representation makes 99% of the images retrieved in the first two positions. Image comparison using colour coherence vectors (CCV) is proposed, which can overcome the drawback of traditional colour histogram-based algorithms that lack spatial information. The algorithm classifies each pixel as coherent or incoherent according to whether each pixel in each colour set belongs to the maximum similar colour region, where the CCV stores the number of coherent and incoherent pixels for each colour. The algorithm can be applied to image retrieval due to its good real-time performance [9]. A holistic representation based on the spatial envelope is proposed to model the image scene, where the spatial envelope is a low-dimensional representation of the image scene. The authors propose five perceptual dimensions, including natural, open, rough, dilated, and solid, which can represent the main spatial structure of the image scene [10].

The spatial envelope model generates a multidimensional space in which scenes with shared members in semantic categories are projected together [11]. A model for generating image descriptions based on a multiple attention mechanism is proposed. Multiple attention modules constructed introducing the focus of human attention on a certain region of an image during image observation into the image description domain [12]. First, an attention module based on image feature encoding is constructed for generating weights for each feature map in the channel direction, explicitly modelling the importance between feature channels; then, a spatial attention module is constructed for focusing on a specific region of the image feature extraction module on the output feature map in the decoding phase; then, a textual attention module is constructed for focusing on the decoding phase to generate utterances correlations exist between them, and the contributions of the three attention modules to the final model are evaluated using ablation experiments; finally, a complete multiple attention model is constructed based on the three attention modules proposed above and is learned using supervised training [13]. The experimental results on several classical datasets show that the proposed model better models the relationships between various objects in images and the correlations between targets and corresponding texts and achieves good experimental results.

However, in the face of more fine-grained visual content understanding tasks, such as the several types of fine-grained visual understanding tasks studied in this paper, there is still much room for improvement in existing deep learning models [14]. First, existing deep models tend to use deeper single models to improve network performance, and these models have the advantage of simple structure and easy end-to-end training. However, single models tend to focus on only a limited number of local features and are unable to understand the dependencies between deeply detailed features, such as the temporal correlation of videos and the spatial correlation of objects [15]. The correlations of these detailed features are crucial for fine visual understanding, so to better solve fine image understanding tasks, the correlations between model-detailed features must be better considered in the models. Second, existing models are often single-stage, where the model reads the input information and outputs the target directly [16]. For fine-grained visual understanding tasks, the output of single-stage models is often inaccurate. For example, when comparing two very similar images, it is a common human practice to compare the most discriminative regions of the two images, and if no conclusion can be reached, then move on to the next detailed region, a process that continues to repeat itself until a conclusion is reached. This process may seem complex, but it encompasses the human idea of parsing fine visual tasks incremental learning. By splitting the single-stage model into multiple incremental stages, it allows the model to better learn detailed information and gradually output results with a higher confidence level. Therefore, the main research idea of this paper is to use the idea of progressive learning to improve the learning problem of relevant features in fine visual understanding tasks.

3. Analysis of Massive Image Recognition Algorithms for Attribute Modeling and Knowledge Acquisition

3.1. Attribute Modelling and Knowledge Acquisition Image Recognition Algorithm Design. Attribute learning can effectively solve the problem of the "semantic gap" generated by underlying visual feature-based algorithms. Attribute learning can describe the semantic information of an image and can be applied to various image processing applications, including image scene classification and image retrieval. Semantic description of objects in an image is an effective way to solve the "semantic gap." The most important aspect of attribute learning is how to obtain the attribute labels of an image. In the second stage, the classifier is used to determine the categories of objects in these boxes. The two-stage fine-grained recognition framework and target detection framework are more complicated than the single-stage framework, but can achieve better performance. To obtain objects with semantic information in images, the traditional way is to manually annotate the image data. However, manual annotation is time-consuming and labour-intensive in the face of huge image databases. Therefore, attribute annotation methods based on target detectors or target filters are born. Ideally, the image needs to be scanned using all the target detectors to obtain the responses of different objects so that the semantic attributes of the image can be annotated automatically. However, this process is not achievable in practice; on the one hand, there is not enough research to build mature target detectors for a huge number of generic objects; on the other hand, the semantic hierarchy problem becomes acute as the number of target objects in an image increases, and not all objects in an image contribute to image semantic modelling. There is a huge amount of image data in the real world, and the existing datasets can only collect and label a very small fraction of it. Thus, for image recognition and classification tasks, the amount of data available for learning is far from sufficient. For the lack of training samples of the target classes in the test set, the algorithm cannot learn effective classification/ recognition features from the available data. Searching the attributes of each object class to train the classifier is a very tedious task and not easy to implement. Therefore, how to efficiently learn high-level attribute features from existing datasets is a popular topic that attracts researchers. Traditional supervised learning-based classifiers can only identify the learned object classes and cannot be used for the classification of other objects. For example, classifiers learned from dog and cat datasets can only be used for dog and cat image classification, but not for horse and cow classification. Since it makes more sense to identify the concept of high-level attributes of images than object categories, we can use existing datasets to learn object attributes across categories.

$$
\begin{aligned}
p(a|x) &= \prod_{m=1}^{N} p\left(a_m \middle| x^2\right), \\
p(a|x) &= \frac{p(y)}{1 + p(a^y)}\left[a^2\right].
\end{aligned}
\tag{1}
$$

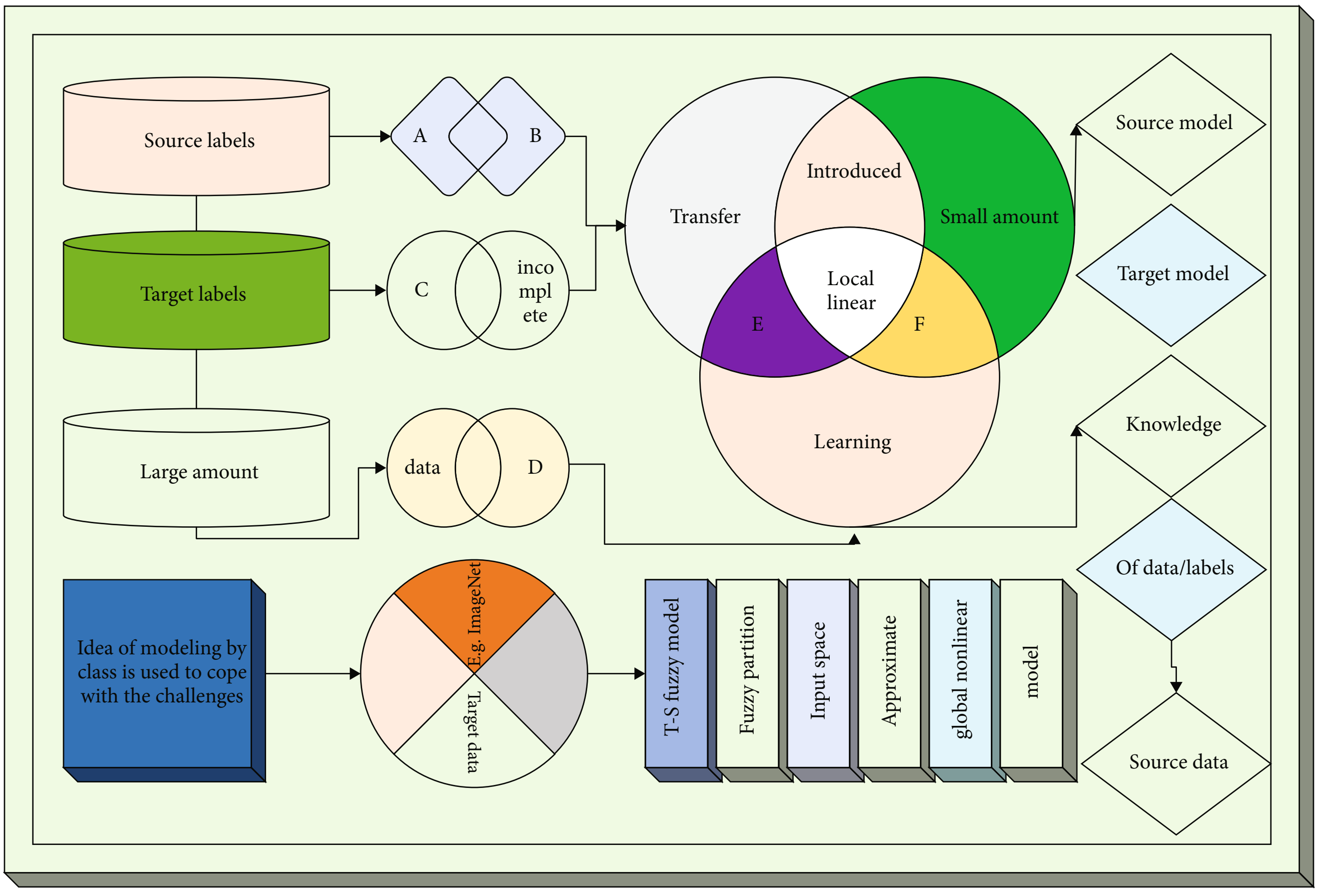

Figure 1: Framework of attribute modelling and knowledge acquisition image recognition algorithm.

To address the variability of attribute relationships between different sample categories, we consider dividing the overall object into several parts, constructing linear models for each part separately, and finally, smoothing these local linear models to obtain the global model [17]. The main idea of the T-S fuzzy model is to fuzzy partition the input space and then approximate the global nonlinear model with the local linear model. Therefore, the T-S model is introduced into incomplete data modelling, and its idea of modelling by class is used to cope with the challenges posed by the variability of attribute relationships between classes and thus improve the fitting accuracy of the regression model. Traditional image recognition dataset construction is generally done in an unsupervised way to obtain data based on direct search in search engines using category names, such as the Tiny Image dataset. However, this method is limited by the performance of the search engine, and the general search engine for image description information is based on the text description around the image in the web page to do document retrieval, and the returned search results are sorted according to user's click behaviour, which leads to the dataset constructed based on this method generally has more noise and more serious bias (as shown in Figure 1).

In this paper, we propose the idea of incremental learning to improve the generic deep learning model, which we call deep incremental learning, and the proposed incremental learning framework differs from the above framework in the following ways. Unlike the generic single-stage deep learning framework, the proposed deep progressive learning framework divides the task into multiple stages, with each stage focusing on information at a certain detail level of the object, and as the stages deepen, the detailed information gradually accumulates, and model's understanding of the object increases, eventually leading to a deep understanding of the object. This multistage design splits the difficulty of the fine visual understanding task and avoids the generic model focusing only on a certain discriminative region, allowing the model to better mine and understand the fine object features. The modelling of design tasks is defined based on knowledge modelling, and the design task space and knowledge space are effectively integrated. Image semantic modelling technology is the requirement of the development of the times. Also, different from the traditional independent multistage framework of computer vision, there is a tight connection between the multiple stages in the deep progressive learning framework. Depending on the task, the relationship between different stages can be flexibly defined. There can be temporal associations, spatial associations, interaction associations, modal associations, etc. between stages, and these associations make the progressive learning framework form an organic whole. In most application scenarios in the subsequent sections, this multistage progressive framework does not affect the end-to-end training of the model.

Thus, the main advantage of the proposed progressive learning framework is that it retains the features of the generic deep learning framework while better mining the fine-grained detailed features of objects, enabling the model to achieve deeper and more fine-grained visual understanding.

$$\begin{aligned} J(W, b, x, y) &= \frac{1}{3}\left\|h_{w,b}(x) + y\right\|^2, \\ \delta_i^{n_i-1} &= \frac{\partial J(W, b, x, y)}{\partial z_i(n-1)}. \end{aligned} \tag{2}$$

Weight sharing means that the weights do not change with position when each convolutional kernel performs sliding window computation in different regions on the picture, i.e., the same convolutional kernel is used to characterize different regions of the picture. The local connectivity and weight-sharing mechanisms can greatly reduce the parameters of the network, allowing for deeper network learning with limited computational resources. Specifically, to obtain an output feature map of a convolution operation, the result of the convolution of the corresponding convolution kernel with each input feature map is first computed; the results are linearly combined and then obtained by an activation function.

$$I_j^i = f\left[\sum_{i=1}^{M^{l-1}} I_i^{l-1} \cdot \left(k_{ij}^l - b_j^i\right)\right]. \tag{3}$$

For image datasets constructed based on deep neural networks, there will be many noisy images that are not in the known category included, but such noisy images can be easily detected by textual information. As the number of target objects in an image increases, the problem of semantic hierarchy becomes acute. Not all objects in the image contribute to image semantic modelling. Similarly, for an image dataset constructed only based on textual information in web pages, many visually irrelevant and semantically ambiguous images can be easily detected by the visual discriminative model obtained from deep neural network learning. Because the confidence level of visually irrelevant noisy images is generally low, the image scenes with textual ambiguities are often very different from the correct category counterparts and can then be detected by a neural network model based on visual information. Considering the complementary nature of text-based and visual information, this paper proposes a new solution for automatic data augmentation: a deep neural network technique based on visual information is organically combined with a text information mining technique based on Internet web pages to automatically construct image datasets.

$$\begin{aligned} \varepsilon_{\mathrm{VTweb}} &= \{<I, c> : f_c(I) \le \alpha\}, \\ T &= \{<T_1, T_2> : f_c(I) \ge \alpha\}, \\ p(y_i = c \,|\, T_i, t_i, d_i) &= \frac{e^{f(yi=c|T_i,t_i,d_i)}}{\sum_{k=1}^{C} e^{f(y=k|T_i,t_i,d_i)}}. \end{aligned} \tag{4}$$

In the research of image analysis and understanding, datasets play an important role; image datasets can be used to test the performance of image feature extraction and detection models, to compare different methods through experiments, and thus, to discover the strengths and weaknesses of different models to help further research improvements, in addition, with the creation of richer, better, and more challenging image databases that continue to drive the computer vision technology development. Image databases are a process from small to large and simple to complex, from the simplest handwritten character font databases to simple image classification datasets to natural image datasets. Their establishment has greatly contributed to the advancement of image understanding techniques in each period of the computer vision development (as shown in Figure 2).

Since only the distance between the central superpixel and the adjacent superpixels is calculated during the construction of the micrograph, the size of the micrograph is limited, which leads to the fact that the size and number of superpixels have a great influence on the construction of the micrograph: small superpixels cannot capture larger semantic objects, and large superpixels cannot capture smaller semantic regions [18]. Attribute labelling methods based on target detectors or target filters came into being. Ideally, it is necessary to use all target detectors to scan the image to obtain the response of different objects, so that the semantic attributes of the image can be automatically labelled. The SLIC parameters affect the annotation accuracy; however, since the semantic regions of different images vary in size and number, setting a uniform parameter is not possible. For this reason, we use three SLIC parameters for superpixel segmentation, 100, 150, and 200, i.e., the same image is repeatedly segmented three times with the number of superpixels of 100, 150, and 200 each time.

Then, we set the number of seed points to 10, 15, and 20 for each of these three sizes of images. During further merging of superpixels. In the process of further merging, we discard the micrographs with the number of superpixels less than or equal to 2, because they usually do not contain any semantic regions and do not contribute to the image processing task. We then fuse these micrograph regions, i.e., micrographs acquired for images of either specification are used as candidate regions.

$$A_l = \frac{1}{N}\sum_{i=1}^{N}\frac{m_i}{n_i}. \tag{5}$$

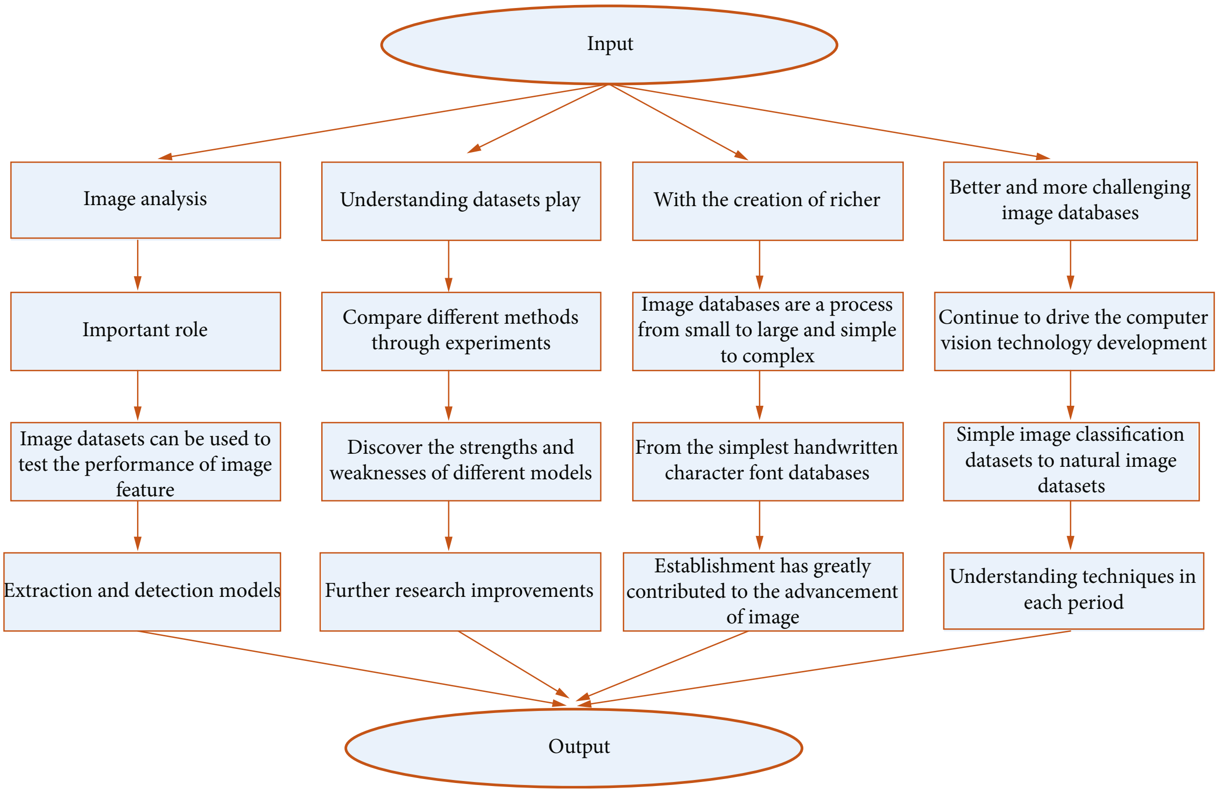
Input
Image analysis
Important role
Image datasets can be used to test the performance of image feature
Extraction and detection models
Understanding datasets play
Compare different methods through experiments
Discover the strengths and weaknesses of different models
Further research improvements
With the creation of richer
Image databases are a process from small to large and simple to complex
From the simplest handwritten character font databases
Establishment has greatly contributed to the advancement of image
Better and more challenging image databases
Continue to drive the computer vision technology development
Simple image classification datasets to natural image datasets
Understanding techniques in each period
Output

Figure 2: Module structure.

The desired default candidate detection frames can be generated on the six output feature maps of the model. First, a minimum square target detection frame and a maximum square target detection frame are generated in the current layers to match the small and large targets at that location. It is stretched or widened to match the long or wide target at that location in the image. The process of generating the default target detection box in the feature map is shown in Figure 2. By traversing each position in the feature map, all the default detection frames at that feature map size are generated, and the above process is repeated for six different sizes of detection maps in the constructed model, and eventually, the model generates all the default detection frames. Since it is traversing all the positions of the feature map, different target detection frames are generated for each position, but not every position in the image has a target, so in the next step, the default generated target detection frames need to be filtered based on the real target position to find the best matching target detection frame.

3.2. Experimental Design of Image Recognition Algorithms. These models tend to achieve relatively good results for general-purpose tasks, but for more fine-grained understanding tasks and higher performance pursuits, single-stage models still have performance bottlenecks. For example, for fine-grained image recognition tasks, currently, popular frameworks generally divided into two phases, with the first phase targeting the localization of foreground objects and filtering out background interference, followed by the second phase classifying the objects. The two-stage target detection model is also a typical multistage framework, where the first stage extracts possible candidate region boxes on the image or feature map, and the second stage then uses a classifier to determine the class of objects within these boxes. Two-stage fine-grained recognition frameworks and target detection frameworks are more complex compared to single-stage frameworks but can achieve better performance [19]. Inspired by existing multistage model frameworks, the proposed deep progressive learning framework in this paper generally contains multiple configurable stages, which can be flexibly configured depending on the task (as shown in Figure 3). For example, for the task of video understanding, the computation of each time node can be divided into one stage; for the task of fine-grained image recognition, we divide the framework into multiple stages according to the understanding of different parts of the object; for the task of interactive action recognition, we divide the model into three stages according to the individual information, the overall information, and the interaction information; and for the task of video generation, we divide the model into two stages: structure generation and pose migration phases. The configurable multistage framework structure splits the overall task difficulty, allowing the individual stages to better focus on different detailed features of the object, thus making it more suitable for fine-grained visual understanding tasks.

The given fine-grained visual understanding task needs to be analysed first, focusing on the problems of the single-stage model in handling such tasks, and the task needs to be split for the existing problems, to reduce the difficulty of

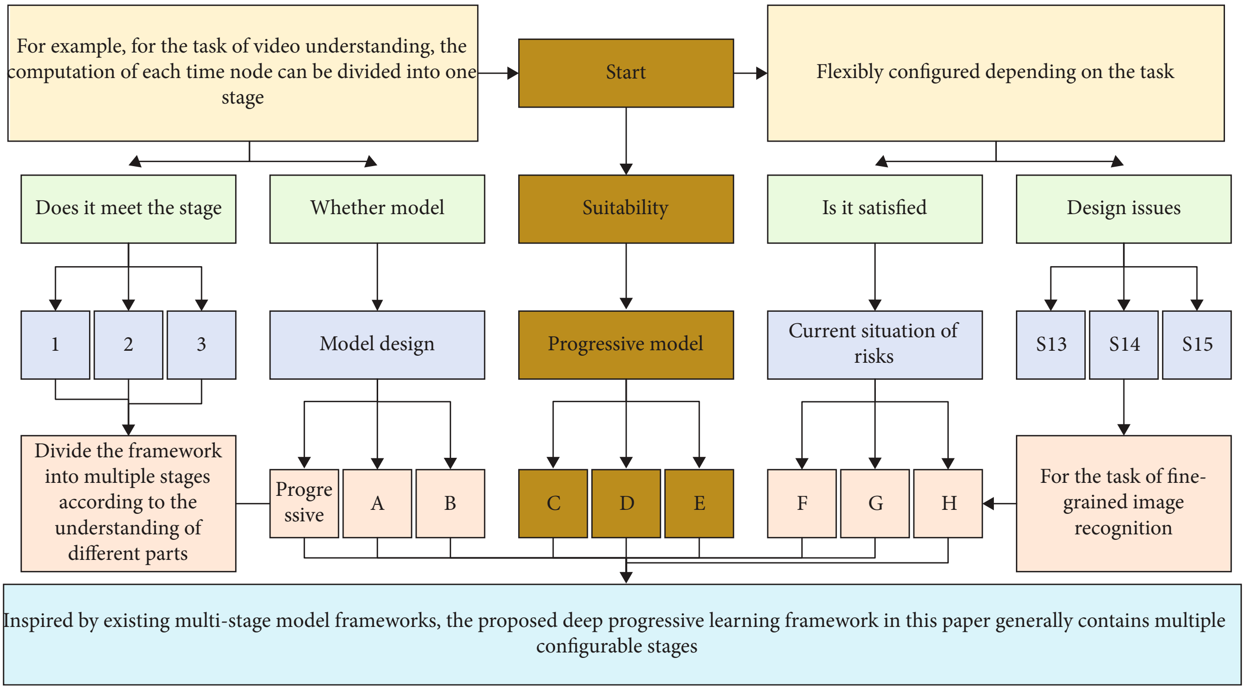
For example, for the task of video understanding, the computation of each time node can be divided into one stage
Start
Flexibly configured depending on the task
Does it meet the stage
Whether model
Suitability
Is it satisfied
Design issues
1
2
3
Model design
Progressive model
Current situation of risks
S13
S14
S15
Divide the framework into multiple stages according to the understanding of different parts
Progre ssive
A
B
C
D
E
F
G
H
For the task of fine-grained image recognition
Inspired by existing multi-stage model frameworks, the proposed deep progressive learning framework in this paper generally contains multiple configurable stages

Figure 3: Progressive learning research methodology.

the task at each stage, and the task needs to satisfy configurability after splitting. The question of how to effectively split fine-grained tasks is an open-ended one, and there is no single definitive answer. In practice, it is necessary to fully think about the requirements of fine-grained tasks, constantly reflect on the shortcomings of existing models, dig deeper into the deep semantic associations that exist in the data, split the task into multiple stages as simply and intuitively as possible, and ensure that there is some semantic association between the stages. Task splitting is the first stage of deep incremental learning, and different splitting methods have a direct impact on subsequent model design, so it is necessary to make bold assumptions, seek proof carefully, and choose the most likely effective solution from the alternatives to try.

$$y_{ij} = w_{i0} - \sum_{l=1, l\neq j}^{s} w_{il} x_{ikl},$$

$$L(x, p, g, c, t) = \frac{1+\alpha}{N} L_{\mathrm{loc}}(x, p, \mathrm{g}) - (\beta - 1) L_{\mathrm{loc}}(x, c, \mathrm{t}). \tag{6}$$

Based on the split task, a suitable progressive model needs to be designed, which not only needs to learn the subtasks at each stage but also needs to satisfy the overall task requirements, i.e., the proposed model needs to satisfy the scalability. Therefore, the model design generally contains two levels, i.e., the subtask level and the overall task level. The model design generally uses generic deep learning models, e.g., the individual subtasks tend to have a single learning goal and are more suitable to be modelled using convolutional neural networks, while integrating subtasks into the overall task is a process of information aggregation and tools such as recurrent neural networks and long and short-term memory networks can be considered. Model design is the core stage of deep progressive learning, and it directly affects the final model performance. When designing the model, it is necessary to consider both the local and overall nature of the task, to reasonably select and designs the model structure for different stages according to the correlation between the subtasks and the overall task, and to actively think about and innovate the traditional model structure, so that the designed model can better meet the task requirements.

$$\sigma = \left(0, 1 + \frac{c_1}{\sum_{\varphi=1}^{k} c_\varphi}, \cdots, 1 + \frac{c_k}{\sum_{\varphi=1}^{k} c_\varphi}\right),$$

$$p_w = p'(Q) \frac{Q}{p(Q)} \prod_{1}^{t-1} \frac{p_d(R_i)}{\sum_d p_d(R_i) \cdot d(R_i)}. \tag{7}$$

In this way, based on the learned category model, we divide each micrograph into known and unknown classes and cluster the unknown classes based on their shape similarity and their position to the surrounding known objects. To model interclass interactions, we use object graph descriptors to encode the layout of the unknown objects. The entire approach does not require knowledge of all object classes in the image but allows for the extraction of useful clues from known objects to better detect new objects.

$$d(g) = [H_0(g), H_1(g), \cdots, H_R(g)]^2. \tag{8}$$

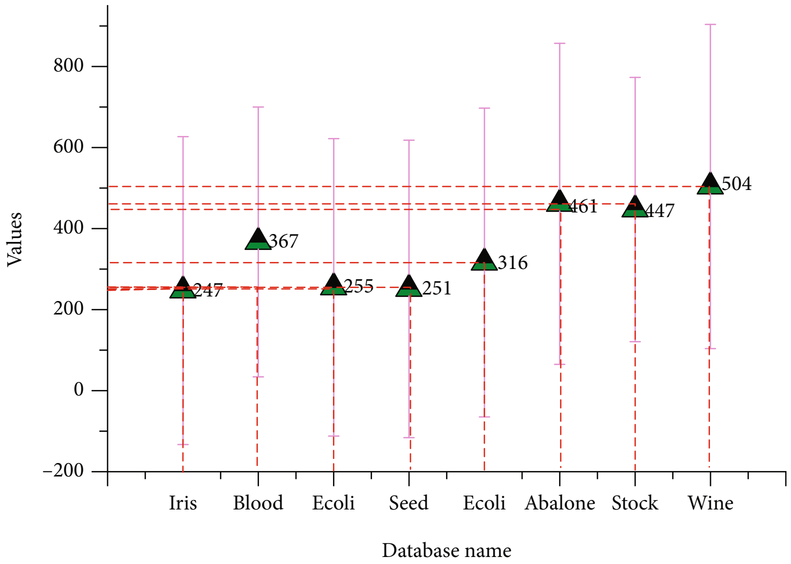

Figure 4: Experimental dataset.

Based on the detected unknown category micrograph, we model it and its surrounding contextual information. To reduce the difficulty of the task at each stage, the configurability must be satisfied after the task is split. How to effectively split fine tasks is an open-ended question, and there is no single definite answer. Specifically, we build a graph representing interactions between objects with nodes that are known objects and edges that connect neighbouring objects. We can then match any two such graphs to determine how well the object-level context agrees for the two candidate regions that may be grouped. Regions with similar contexts will have similar graphs, while regions with dissimilar contexts will produce different graphs. If the classification of image superpixel segmentation and micrographs is precise, then the graphs we construct are idealized in the sense that we can simply count the number and type of known objects and record their relative layout (as shown in Figure 4).

In practice, however, image segmentation and classification algorithms are not perfect, which leads to the fact that we cannot always obtain good classifiers. Although we cannot correct mislabelled known and unknown regions, we can introduce uncertainty into the contextual description of the object, which can make misclassified known regions more robust [20]. Linear regression modelling approaches assume that incomplete data attributes have linear relationships with each other, describing the association between attributes in terms of straight lines, planes, or hyperplanes and solving for the model parameters using least squares. However, the attribute relationships may vary from sample to sample in the actual data, and the overall trend of attribute relationships is nonlinear, then the linear model constructed for the data is bound to have some deviation from the actual regression relationships.

4. Results and Analysis

4.1. Attribute Modelling and Knowledge Acquisition Image Recognition Algorithm Performance. We compare our algorithm with SIFT-Bow, GIST, and SPM algorithms, and deep learning algorithms based on Reset networks after denoising the image semantic labels, we use a stream learning algorithm to automatically pass the image-level labels to the pixel level. After the semantic labelling is completed, the micrographs are proposed and optimized. Since the micrograph is a polygon with irregular edges, we use the outermost box enclosing the box to approximate the representation of the micrograph and scale the micrograph to 112 × 112. Then, we perform deep feature extraction according to the network architecture, and we train a simple support vector machine (SVM) to perform image classification. The experimental results are shown in Figure 5. Images with noisy labels are processed, and the algorithm achieves a classification accuracy of 68.51% for the VOC2012 dataset, which is lower than the classification accuracy of the Reset network and the classification accuracy of our method without noisy labels. Since our method cannot fully complement the missing and eliminate the wrong labels, these noisy labels affect the label delivery and the extraction of the micrographs, which in turn affects the image classification accuracy.

The analysis in Figure 6 shows that those characteristics of the target have a large impact on the detection performance of the model. The area of the anchor frame shows that the model performs significantly better on most targets for large objects than for small ones, and the size of the target is more sensitive to the performance of the model. The microimage is scaled to 112 × 112, and then, deep feature extraction is performed according to the network architecture. We train a simple support vector machine (SVM) to perform image classification. The aspect ratio shows that the model has better detection performance for medium-sized objects, i.e., better detection performance for targets with a

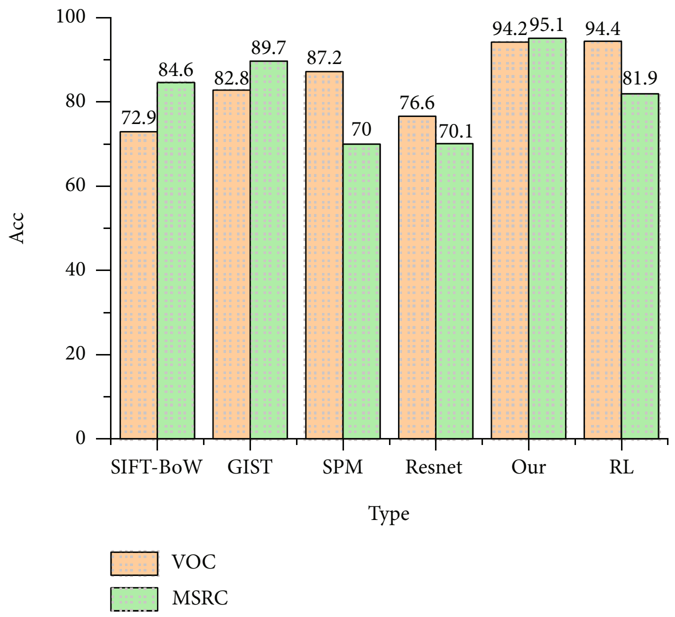

Figure 5: Accuracy of image classification under different methods.

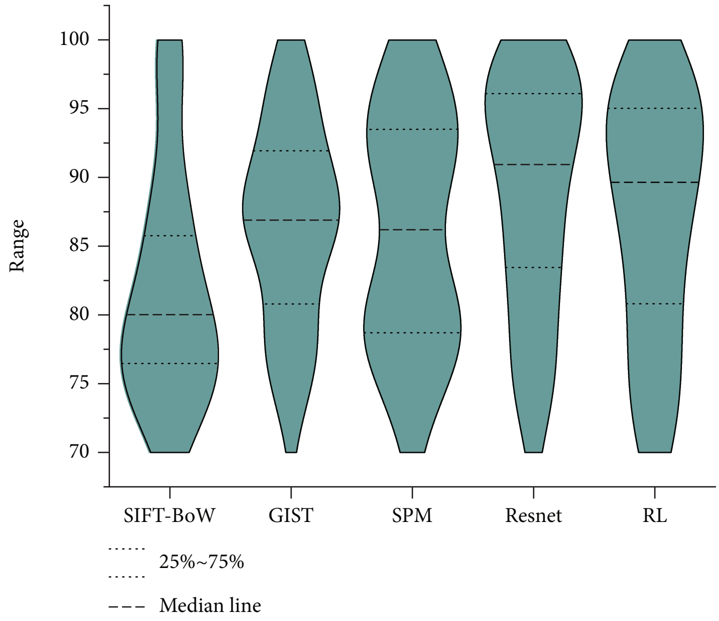

Figure 6: Effect of target characteristics on model detection performance and sensitivity.

bias towards squares, but there are large differences between categories for very high or very wide targets, but overall, the aspect ratio of the target is less sensitive to model's detection performance than the effect produced by the anchor frame area. Although the above analysis does not directly improve the detection performance of the model, it can help us to reasonably evaluate the advantages and disadvantages of several models and give reference to further improve the model performance.

The user click behaviour data generally shows a heavily heavy-tailed distribution. Typically, only a few words appear frequently, while most words appear very infrequently. This also means that the training process frequently passes many similar output signals to the deep convolutional neural network. This results in the response values of the convolutional kernels corresponding to the frequently occurring visual templates being much larger than the response values of the other convolutional kernels, and most of the convolutional kernels in the network will then tend to respond to those visual template inputs that occur more frequently. As a result, there will be many similar convolutional kernels during the training of the neural network at the beginning,

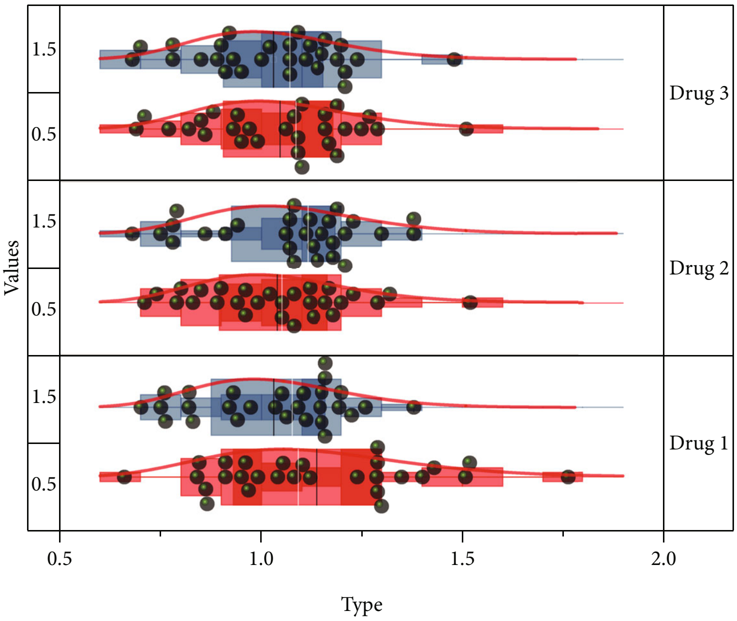

Figure 7: Filling results for single output subnetwork model and linear regression model.

and such many similar convolutional kernels wastes a large amount of parameter space of the neural network, resulting in slow convergence of the neural network.

4.2. Experimental Results. For each missing rate, five incomplete datasets randomly generated for each complete dataset, and the attribute association models were built for the incomplete data based on the subnet iteration method and the linear regression method, respectively, and the missing value filling error MAPE was calculated for the two methods, which measured the accuracy of fitting the attribute relationships of the two models to the incomplete data. Figure 7 shows the filling results of both methods, SONN+IL and LR, for each missing rate of the eight experimental datasets.

The most essential difference between the SONN+IL and LR methods is that the attribute regression model built by the SONN+IL method for incomplete data is nonlinear, while the LR method builds a linear model. Nonlinearity is one of the typical properties of complexity in nature. Compared with linearity, nonlinearity is usually closer to the objective nature itself, so the attribute regression model built based on the subnet iteration method is closer to the real correlation relationship between data attributes, and thus, the accuracy of filling in missing values is higher. In addition, the LR method is based on the least-squares method to solve the parameters of linear equations and find the global minimum by directly deriving the objective function in a noniterative way; in this way, the determination of model parameters is to some extent affected by the quality of prefilling, which is often coarse. Although the SONN+IL method also introduces a prefill link, the missing values participate as system-level variables during neural network training, and their fill values are adjusted in real time according to the model output during iteration, gradually weakening the influence of prefill on model parameter learning.

The mean absolute percentage error MAPE between the missing value filling values and the true values of the three methods is calculated, and this error is used as an evaluation metric to measure the effectiveness of the three methods in modelling incomplete data attributes and missing value filling. Five incomplete datasets were randomly generated for each complete dataset at each missing rate, and the average of these five-filling error MAPEs was taken as the final experimental result. The experimental results of the three methods on the five datasets are shown in Figure 8.

With the rapid development of web technologies and portable mobile devices, a large amount of image data is added to the Internet every day, and how to manage these images quickly has become a pressing problem. Traditional image processing tasks (e.g., image classification) can be performed using unsupervised or supervised algorithms. In general, supervised learning algorithms can achieve better performance than unsupervised learning algorithms; however, supervised learning algorithms require extensive pixel-level annotation of images, which is very impractical in large-scale image applications. This is very unrealistic in large-scale image applications. The image semantic modelling task is dedicated to allowing the machine to "understand" the meaningful objects contained in the image, such as people, animals, and other objects. The task of image semantic modelling is dedicated to making the machine "read" the meaningful objects contained in the image, such as people, animals, and other objects. In the face of massive image data, image semantic modelling can provide solutions for tasks such as image classification, recognition, and retrieval. Image feature extraction is the basis of image modelling. Early image modelling algorithms based on the underlying visual features of images do not reflect well the human visual perception of

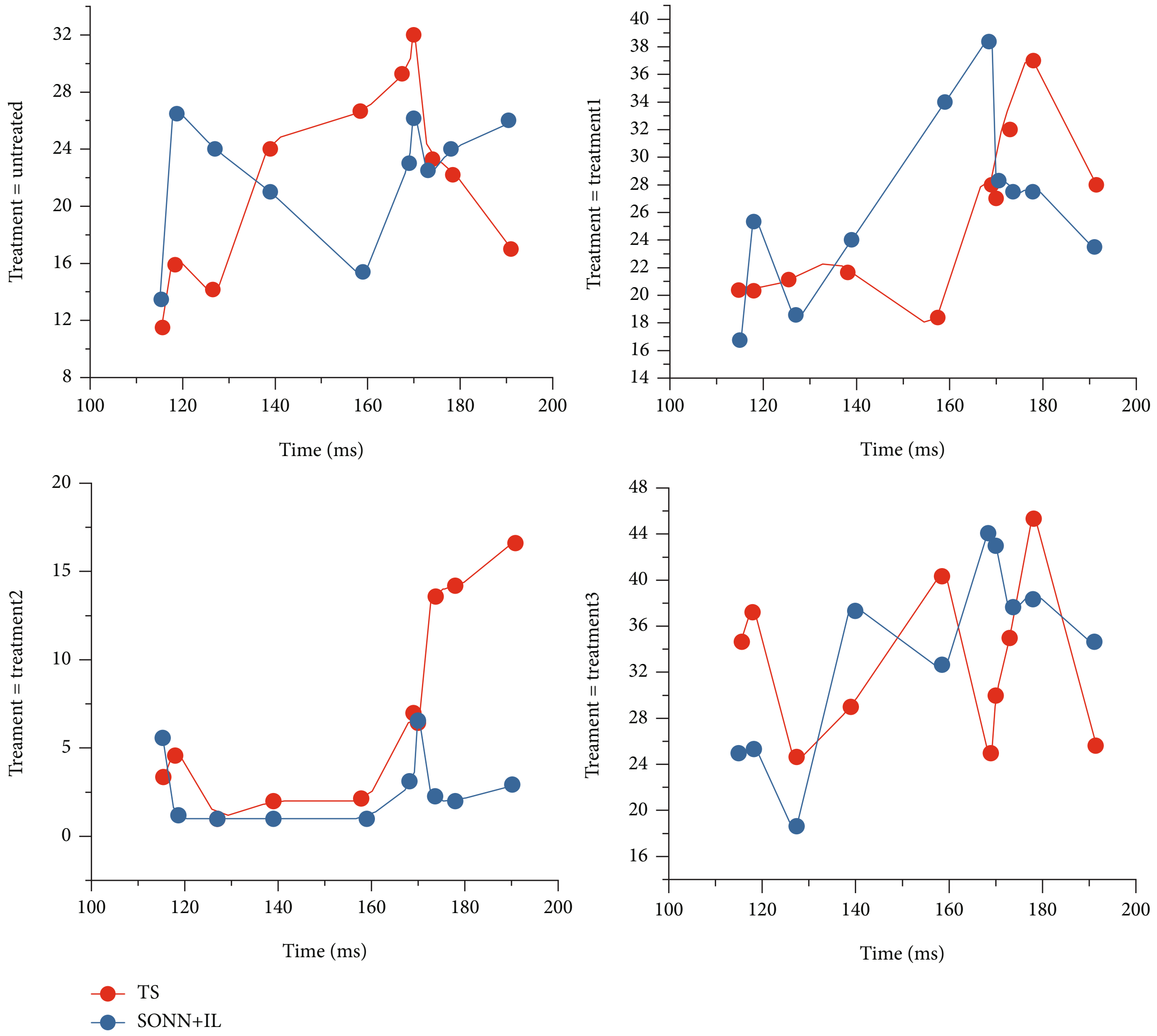

Figure 8: Experimental results.

images. Later, attribute learning-based algorithms came into being. Attributes can be understood as semantic descriptions of image contents, such as object feature descriptions and object name descriptions. However, attribute learning still requires a lot of manual annotation, and many annotation effects still depend on the performance of many external detectors. Also, the selection of appropriate attributes depends on the experience of the engineer.

5. Conclusion

A single-stage multitarget detection and recognition model based on deep learning is proposed. The model uses a commonly used large-scale convolutional neural network with good migration learning capability as the backbone network, and first generates output feature maps of different scales at different stages of the backbone network, and to fuse the detection information on the feature maps of different scales, the output feature maps with strong semantic information at the higher levels are fused with the output feature maps at the bottom levels by transposed convolution to effectively learn the hierarchical structural features of the images. Then, inspired by human visual perceptual fields, a module is constructed to fuse different visual perceptual fields, which fuses output feature maps with different perceptual field information by using convolution kernels and null convolution at different scales and introduces crosslayer connectivity to alleviate the gradient vanishing problem of the model, by introducing two parameters in the category loss function from category imbalance and category probability. By introducing two parameters in the category loss function to weight the model loss function in terms of category imbalance and category probability, respectively, the accuracy of target detection is better improved. Experimental results on several datasets demonstrate the effectiveness of the proposed model, which achieves a better trade-off between detection accuracy and operation speed than classical target detection networks such as two-stage and single-stage. Then, the bilinear difference-based pooling of regions of interest is used to generate fixed-size feature maps for subsequent inputs, and a target detection and classification module, a target instance segmentation module, and a human pose estimation module are constructed; a joint multitask deep learning model is constructed based on the above-proposed modules, and model learning is performed by supervised fine-tuning. Experiments on challenging image datasets and generalized datasets demonstrate that the proposed model can achieve comparable or even better performance on multiple images understanding tasks compared to single-task models.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

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

Acknowledgments

This work was supported by the Yunnan Provincial Science and Technology Department, Yunnan Provincial Innovation and Entrepreneurship Space "Woodcraft innovation space" (no. xctd201801).

References

[1] L. Wang, "Heterogeneous data and big data analytics," *Automatic Control and Information Sciences*, vol. 3, no. 1, pp. 8–15, 2017.

[2] W. Lu, "Improved K-means clustering algorithm for big data mining under Hadoop parallel framework," *Journal of Grid Computing*, vol. 18, no. 2, pp. 239–250, 2020.

[3] L. Zhou, C. Zhang, F. Liu, Z. Qiu, and Y. He, "Application of deep learning in food: a review," *Comprehensive Reviews in Food Science and Food Safety*, vol. 18, no. 6, pp. 1793–1811, 2019.

[4] G. Chen, Q. Weng, G. J. Hay, and Y. He, "Geographic object-based image analysis (GEOBIA): emerging trends and future opportunities," *GIScience & Remote Sensing*, vol. 55, no. 2, pp. 159–182, 2018.

[5] Z. Wang, H. Di, M. A. Shafiq, Y. Alaudah, and G. AlRegib, "Successful leveraging of image processing and machine learning in seismic structural interpretation: a review," *The Leading Edge*, vol. 37, no. 6, pp. 451–461, 2018.

[6] J. R. Saura, "Using data sciences in digital marketing: framework, methods, and performance metrics," *Journal of Innovation & Knowledge*, vol. 6, no. 2, pp. 92–102, 2021.

[7] L. D. Xu and L. Duan, "Big data for cyber physical systems in industry 4.0: a survey," *Enterprise Information Systems*, vol. 13, no. 2, pp. 148–169, 2019.

[8] V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks - a review," *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 4, pp. 415–425, 2019.

[9] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, "Deep learning-based image recognition for autonomous driving," *IATSS Research*, vol. 43, no. 4, pp. 244–252, 2019.

[10] T. R. H. Goodbody, N. C. Coops, and J. C. White, "Digital aerial photogrammetry for updating area-based forest inventories: a review of opportunities, challenges, and future directions," *Current Forestry Reports*, vol. 5, no. 2, pp. 55–75, 2019.

[11] X. Qi, G. Chen, Y. Li, X. Cheng, and C. Li, "Applying neural-network-based machine learning to additive manufacturing: current applications, challenges, and future perspectives," *Engineering*, vol. 5, no. 4, pp. 721–729, 2019.

[12] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019.

[13] Yogesh, A. K. Dubey, R. Ratan, and A. Rocha, "Computer vision based analysis and detection of defects in fruits causes due to nutrients deficiency," *Cluster Computing*, vol. 23, no. 3, pp. 1817–1826, 2020.

[14] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 6–22, 2019.

[15] A. Abid, M. T. Khan, and J. Iqbal, "A review on fault detection and diagnosis techniques: basics and beyond," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3639–3664, 2021.

[16] J. Wen, J. Yang, B. Jiang, H. Song, and H. Wang, "Big data driven marine environment information forecasting: a time series prediction network," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 4–18, 2021.

[17] G. Fernandes, J. J. P. C. Rodrigues, L. F. Carvalho, J. F. al-Muhtadi, and M. L. Proença Jr., "A comprehensive survey on network anomaly detection," *Telecommunication Systems*, vol. 70, no. 3, pp. 447–489, 2019.

[18] S. Wan, Y. Xia, L. Qi, Y. H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1756–1768, 2020.

[19] U. Leicht-Deobald, T. Busch, C. Schank et al., "The challenges of algorithm-based HR decision-making for personal integrity," *Journal of Business Ethics*, vol. 160, no. 2, pp. 377–392, 2019.

[20] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159–177, 2017.