Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Research on Image Classification And Semantic Segmentation Model Based on Convolutional Neural Network

Muqing Li[1, a], Ziyi Zhu[2, b], Ruilin Xu[3, c], Yinqiu Feng[4, d], Lingxi Xiao[5, e]

[1] University of California San Diego, USA
[2] New York University, USA
[3] The University of Chicago, USA
[4] Columbia University, USA
[5] Georgia Institute of Technology, USA

[a]MUL003@ucsd.edu; [b]zz1831@nyu.edu; [c]harveytsui915@gmail.com; [d]yf2579@columbia.edu; [e]Lingxi.xiao@gatech.edu

Abstract: This paper investigates convolutional neural network (CNN)-based approaches for image classification and semantic segmentation, with a focus on addressing spatial detail loss and multi-scale feature integration issues prevalent in semantic segmentation. The introduced EDNET model tackles these challenges through the incorporation of spatial information branches and the design of efficient feature fusion mechanisms. It further enhances performance via the use of global pooling and boundary refinement modules. Evaluations on the PASCAL VOC 2012 dataset reveal an 11.67% increase in mean intersection-over-union (IoU) compared to standard fully convolutional networks, demonstrating substantial improvement over comparable techniques. These results confirm the efficacy and practicality of the EDNET framework.

Keywords: Image Classification; Semantic Segmentation; Convolutional neural network.

1. Introduction

In today's highly informationized society, image processing technology, as a link between the physical world and the digital space, has an increasingly significant strategic significance [1-3]. The exponential growth of big data, coupled with substantial advancements in computational capabilities, has fueled a transformative wave in deep learning [4-5], notably through the leveraging of Convolutional Neural Networks (CNNs), which have dramatically reshaped the landscape of computer vision technology [6-9]. Within this domain, the three fundamental pillars comprise image categorization, object recognition, and image partitioning tasks, among which image classification aims to assign category labels to images, object detection requires accurate location of objects, and semantic segmentation pursues accurate classification at the pixel level [10], requiring extraordinary recognition of object boundaries [11].

Despite this, CNN's accomplishments in image classification, exemplified by its groundbreaking performance in the Large-scale Visual Recognition Challenge (ILSVRC), stand as pivotal milestones within the discipline, it is still a major challenge for current research to achieve both high-speed and high-precision semantic segmentation in complex scenes [12]. This requires the model not only to have a strong ability to distinguish categories but also to have a fine spatial analysis power to ensure the accuracy of pixel-level classification. Therefore, the development of an efficient and robust CNN model to realize the efficient integration of image classification and semantic segmentation has become an urgent research focus.

This research is devoted to exploring the image classification and semantic segmentation model based on CNN. By integrating the basic theory of CNN, the latest architectural progress, and its practical application in two types of tasks, we aim to build an innovative model framework that combines the dual advantages of advanced feature extraction and pixel-level accurate positioning. The goal of this research is not only to contribute new theoretical insights and practical methods to image analysis technology, but also to promote these advanced technologies to cross traditional boundaries and achieve wide application in emerging fields such as medical Detection [13-15], geography [16-17], and prediction [18-19], where deep learning is already widely used. This approach aims to drive the continuous expansion and innovation of computer vision technology.

2. Correlational research

Image segmentation, as the core element of image processing, is rapidly becoming a key research focus in this discipline. The technology can be broadly divided into two schools: traditional methods and modern strategies based on deep learning. The traditional approach relies on pixel feature association analysis for region division, which has obvious advantages in simplicity, but in complex scenes, the segmentation accuracy is poor and the application limitations are obvious. Recently, advancements in hardware capabilities, notably accelerated by the advent of Fully Convolutional Neural Networks (FCNs), have propelled semantic segmentation techniques to unprecedented levels. These innovations have surpassed conventional methods in both segmentation precision and versatility, ushering in a new era of development for the sector.

Long et al. pioneered the application of full convolutional networks (FCN) to image semantic segmentation [20], laying the foundation for the application of CNN in this field. Unlike traditional CNN, which uses a fully connected layer to transform feature maps, FCN uses a convolutional layer

instead to directly process input images of any size, resulting in an end-to-end segmentation solution. Badrinarayanan et al. introduced an encoding and decoding architecture through SegNet[21], which cleverly recorded pixel positions in the downsampling process by using pooled index, thus effectively recovering edge information during upsampling and improving segmentation quality. Furthermore, Ronneberger et al. introduced U-Net[22], a symmetric architecture inspired by FCNs, tailored for medical image segmentation tasks. This design aimed to proficiently manage the intricacies of medical image segmentation, thereby reinforcing the inventive capacity of deep learning within the realm of image segmentation.

Chen et al. introduced the DeepLabV1 framework and adopted the expansive convolution technology to broaden the receptive field by filling zeros inside the convolution kernel, which can effectively capture a wider range of context information without increasing parameters or computational burden, significantly improving the efficiency and accuracy of the model [23]. In view of the variable size of objects, some studies try to enhance the model performance by adjusting the input image size and integrating multi-scale feature maps, but this is accompanied by high computational costs [24]. Zhao et al. proposed PSPNet to solve the segmentation problem of small and medium-sized objects in complex scenes. This network integrates global context information by integrating multi-scale feature pyramid, effectively improving the segmentation accuracy of objects of different sizes [25]. On the other hand, Papadomanolaki et al. integrated the unsupervised segmentation algorithm with the full convolutional network, and significantly enhanced the accuracy of pixel-level segmentation by adding segmentation loss function optimization network [26]. The super-resolution semantic segmentation module designed by Pereira et al. successfully realized high-quality segmentation of low-resolution images and proved its efficient segmentation ability in the case of limited resources [27]. These works together demonstrate the key role of optimizing receptive field management, multi-scale feature fusion, and context information utilization through technological innovation in improving semantic segmentation performance.

However, the image segmentation technology based on FCNs still faces several challenges, including insufficient effective fusion between feature levels, insufficient use of spatial and contextual information, and long inference time, which limit the performance of the model in real-time and accuracy. Therefore, exploring model optimization strategies to improve feature alignment, deepen multi-scale feature fusion, enhance context understanding, and speed up the inference process has become a key direction to improve the efficiency of image segmentation.

3. Method

3.1. Convolutional Neural Networks, CNN

Deep learning has seen convolutional neural networks (CNNs) rise to prominence, particularly in computer vision, due to their distinct architectural advantages. However, the utilization of CNNs in the specialized domain of electricity load forecasting is still a relatively untapped area. A standard CNN design encompasses five fundamental segments: an input layer as the point of origin, followed by a sequence of convolutional layers, interspersed with a pooling layer, succeeded by a fully connected layer, and culminating in an output layer for predictions. At the heart of CNNs lies the convolution layer, comprised of three essential facets: convolution kernels, associated layer parameters, and an activation function. Each layer harbors numerous convolutional units, where individual units resemble neurons in a feedforward network, equipped with their own weight coefficients and bias terms. A pivotal benefit of CNNs lies in their ability to automatically extract and learn features during training with a constrained set of weights, a trait that proves highly efficacious for handling high-dimensional datasets.

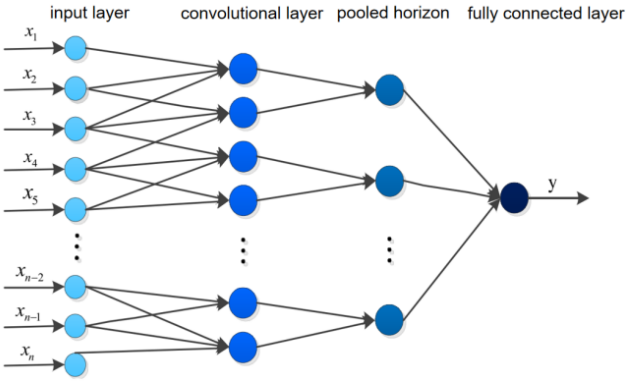

Figure 1. CNN network structure diagram

In the structure shown in Figure 1, the input variables xi are represented as (x1, x2, x3,... , xn) a column vector that integrates multiple information sources such as weather conditions, seasonal changes, date types, and other load data series. The input layer is processed using a linear identity function. In the convolution layer part, there are multiple feature maps which, based on the information received from the input layer, scroll across all the data through each convolution kernel and produce multiple sets of output data. It is worth noting that the weight parameter ω used in the calculation of the same convolution kernel is the same as the threshold b. Then, the multiple sets of data generated by these convolution operations are passed to the subsequent pooling layer by nonlinear transformation. The function of the pooling layer is to aggregate statistical operations on data within a preset range, that is, to replace a series of values in the region with the average or maximum value, so as to reduce the data dimension. Lastly, the data that has undergone dimensionality reduction is concatenated and processed through the fully connected layer to produce the ultimate prediction outcomes.

Given a set of input signals xi (where i traverses 1 to n), its mathematical expression can be described as follows during the calculation of the convolution layer:

$$x_i^l = f(\sum_{j=i}^{k} x_j^{l-1}\omega_j^l + b_j^l), i = 1,2,\ldots,n-r+1; k = i+r-1 \quad (1)$$

In formula (1), xil represents the output value generated by the i node of layer l; xjl-1 refers to the input signal provided by the JTH node of layer l-1. Here, the stride (r) of the convolutional kernel is introduced, which influences the sampling interval during the convolution operation on input data. The variable j's range from i to k implies a sparse connectivity pattern, meaning that nodes in layer L are connected only to a subset of nodes in the preceding layer. ω jl represents the weight corresponding to xil-1, while bjl signifies the threshold associated with xil-1.

The function f takes the form of the Sigmoid function and can be specifically defined by formula (2).

$$f(I) = \frac{1}{1-e^{-I}} \quad (2)$$

$$I = \sum_{j=i}^{k} x_j^{l-1}\omega_j^l + b_j^l \quad (3)$$

The operation of the pooling layer can be described by the following mathematical expression:

$$x_i^l = h(x_i^{l-1}) + b_i^l \quad (4)$$

In this formula, the function h(~) represents the operation of calculating the average value.

The mathematical expression of the fully connected layer can be expressed as:

$$y = f(I^l), I^l = W^{l-1}x^{l-1} + b^l \quad (5)$$

In this model, Wl-1 represents the weight parameter from the l-1 layer to the L-layer. bl indicates the corresponding threshold. And y represents the final output data.

3.2. The image semantic segmentation model EDNET

The approach presented in this paper is built on a composite network architecture that combines multipath design with basic module stacking, with a core divided into two parts: Encoder and Decoder. As shown in Figure 2, the network layout is clearly divided - the left wire frame is marked as the encoder part responsible for feature extraction, and the right wire frame represents the decoder part that performs feature reconstruction, where the "2×" symbol indicates the double up-sampling operation during the decoding phase, aimed at gradually restoring the original resolution of the input image.

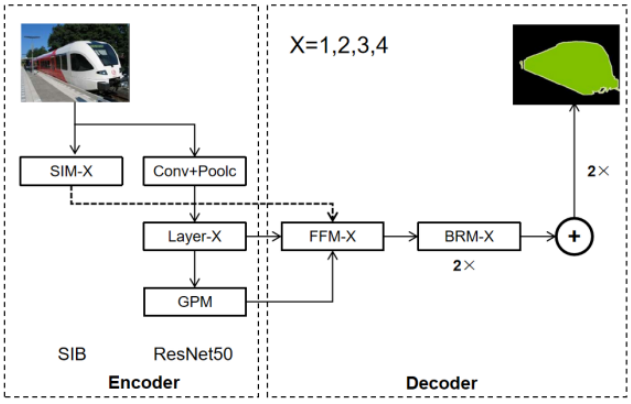

Figure 2. Text framework

In the architecture shown in Figure 2, "SIB" stands for the Spatial Information Branch that processes Spatial Information, and "SIM" is the spatial information Module within the branch that processes spatial features. In addition, "Conv+Pool" and the subsequent "Layer-1 to Layer-4" form the backbone network based on ResNet and are responsible for basic feature extraction. The GPM" Global Pooling Module "is used to capture global context information, and the FFM Feature Fusion Module integrates features from different paths to enhance representation. Finally, the "BRM" Boundary Refinement Module focuses on improving the quality and accuracy of segmentation boundaries.

3.2.1. Encoder network

The encoder network consists of three core components: ResNet50 backbone network [28], spatial information branch and global pooling module. The image input is divided into two tracks: one track captures the spatial fine structure of the image and the target boundary information by means of the spatial information branch, which provides edge enhancement for the decoding stage; The other track uses ResNet50 to deeply mine multi-level features, and then uses global pooling to summarize the macro context information of the entire image.

a. ResNet50 basic network

In this paper, the ResNet50 model is adjusted, and the last pooled layer, fully connected layer and softmax layer are replaced with convolution layer, so as to adapt to the input requirements of different sizes of images. The ResNet50 structure can be divided into five successive stages according to the size and number of channels of the output feature map: initial Conv+Pool, and Layer-1 to Layer-4. The feature map resolution of these stages corresponds to the ratio of 1/4, 1/8, 1/16, 1/32 of the input image resolution in turn. Details about the specific configuration of ResNet50 include input channels (i), output channels (o) and the number of residual units (b) in each stage, as shown in Table 1.

Table 1. ResNet50 configuration

Block	i	o	b
Conv+Pool	3	64	--
Layer-1	64	256	3
Layer-2	256	512	4
Layer-3	512	1024	6
Layer-4	1024	2048	3

b. Spatial information branch enhancement

This study incorporates a spatial information branch (SIB) in the encoder specifically designed to capture spatial details, as shown in Figure 3. The branch consists of four continuous spatial information modules (SIM), which are designed to provide rich edge-oriented information for the decoding stage. Each SIM module comprises a 3x3 convolutional layer employing a stride of 2, accompanied by a 3x3 average pooling layer, also utilizing an equivalent stride length. The two work together, although the mechanism of action is different, but can complement each other to extract unique spatial features. As the level of downsampling deepens, the feature map gradually favors high-level semantics at the expense of spatial details, so only four SIMS are deployed in this design to effectively retain and extract the target edge clues. In the specific implementation, SIM first independently applies average pooling and convolution downsampling to the input feature map, and then fuses the features produced by the two through the stitching operation to output the feature map rich in edge information, ensuring the accurate auxiliary positioning of the target contour.

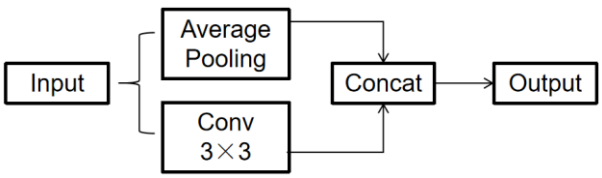

Figure 3. Schematic diagram of SIM

c. Global pooling module

A global pooling module is integrated in the tail of ResNet50 network to capture the macro context features of the whole image and enhance the segmentation performance of the model for large-area targets. As illustrated in Figure 4, the module first performs global averaging pooling to extract the overall information of the image. Subsequently, a 1x1 convolutional kernel is employed to diminish the channel count in the feature map, thereby efficiently condensing information while retaining vital contextual details. Lastly, bilinear interpolation is applied to resample the refined feature map to match the original input image scale (1/32), facilitating effective integration with additional features.

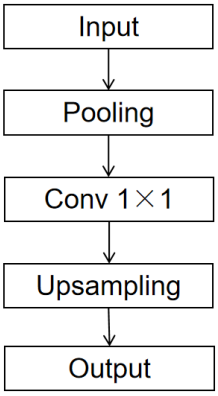

Figure 4. Global pooling module

The basic operation of the global pooling module is global average pooling, which involves summarizing all the position response values of each channel feature map and calculating the average value. The mathematical expression is briefly described as:

$$\mathbf{y}_i = \frac{1}{\mathrm{H}\times\mathrm{W}}\sum_j^{\mathrm{H}}\sum_k^{\mathrm{W}} x_{i,j,k} \quad (6)$$

In Equation (6), yi denotes the output of the (ith) channel following global average pooling; H and W signify the height and width of the input feature map, respectively; while xi,j,k represents the pixel value at channel i, row j, and column k within said input feature map.

3.2.2. Decoder network

The decoder network is responsible for integrating the spatial details, multi-level features and global context information extracted by the encoder, and making comprehensive prediction. This process is realized through bilinear interpolation to improve the resolution of feature map step by step. The core mechanism includes feature fusion module and boundary optimization module.

a. Feature fusion module

The levels of ResNet50 features (layer-1 to Layer-4) carry different levels of information: the shallow Layer is rich in spatial details, while the deep Layer contains deeper semantic content. The branch of spatial information also produces rich spatial features. As a result, integrating ResNet50's multilevel features with the details of the spatial information branches became a challenge. To tackle this issue, we devise a feature fusion module that leverages a channel attention mechanism alongside a residual structure, enabling the seamless integration of the dual information streams with enhanced efficiency.

As shown in Figure 5, the module accepts two feature inputs of the same size (X1 and X2). X1 represents the different scale semantic features extracted by ResNet50, while X2 is derived from the spatial information module or the global pooling module, which is rich in spatial details. First, the number of channels of X1 is reduced by a 1×1 convolution layer to facilitate effective concatenation with X2. Then, the concatenated features are convolved by 1×1 again to generate a preliminary fused feature map. This mapping is then fed into a specialized residual unit, which integrates channel attention mechanisms to dynamically evaluate and highlight important differences between input feature channels, thereby enhancing the expression of key features and achieving optimal information integration.

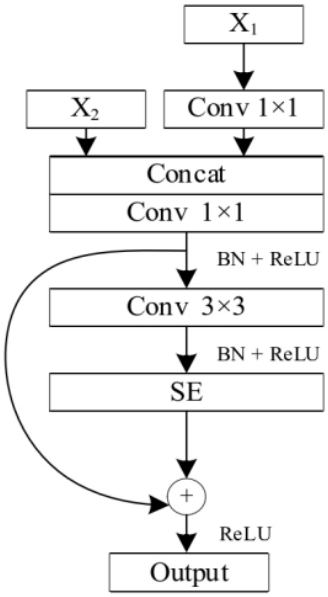

Figure 5. Feature fusion module

b. Boundary refinement module

This paper proposes a boundary refinement module (BRM) with integrated semantic classification guidance, the structure of which is shown in Figure 6. The module usually accepts two inputs: the output of the feature fusion module and the high-level output of the previous stage of BRM (except BRM-4, which uses only the feature fusion output). The two parts of the input are concatenated and combined, and then the feature mapping is carried out by 1×1 convolution, aiming at the preliminary estimation of the boundary. After that, the residual structure is used to further optimize and refine the predicted boundary information, which improves the accuracy of boundary definition.

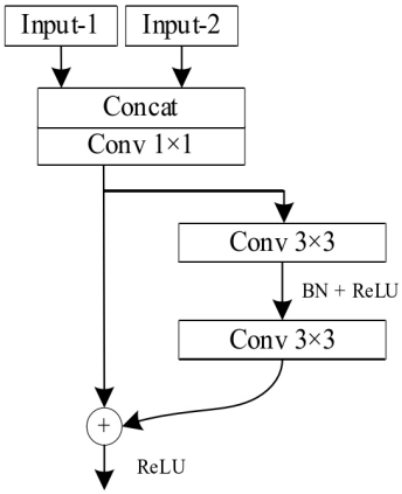

Figure 6. Boundary refinement module

4. Application of model

4.1. Data sets and evaluation indicators

4.1.1. Data set

This research utilizes the PASCAL VOC 2012 dataset, sourced from the 2012 PASCAL Visual Object Classes Challenge, a seminal benchmark in computer vision studies. Renowned for its meticulous annotations and diverse visual content, the dataset is instrumental in evaluating and refining algorithms for tasks including image classification, object recognition, segmentation, and semantic segmentation,

thereby significantly contributing to advancements in the field. The Linked Data methodology is employed to standardize disparate data formats within a dataset, a critical step for scholarly research [29].

The PASCAL VOC 2012 dataset is a diverse resource of approximately 11,500 images taken from various scenes of everyday life, covering 20 categories of target objects (e.g., people, vehicles, animals, etc.) and a background category, which is further subdivided into supercategories[30]. Each image is equipped with detailed annotations, including target boundary boxes (for object detection), pixel-level segmentation masks (for semantic and instance segmentation), and image category labels (for image classification), supporting a variety of visual recognition tasks, making it the preferred platform for comprehensive evaluation of the performance of computer vision algorithms. The data sets were rigorously divided into training, validation and test sets, in particular the test sets were submitted through the official server for standardized scores, ensuring comparability across studies. In addition, the complex background, occlusion, variable pose and scale of the images in the data set are very challenging, requiring the algorithm to have excellent robustness and generalization ability.

4.1.2. Evaluation indicators

Presently, progress in computer vision has converged upon a standard measure for assessing the efficacy of image semantic segmentation: the Mean Intersection over Union (mIoU). This metric enables an impartial comparison of different models' efficacy. In this research, mIoU is employed as the primary gauge to assess the model's overall performance on a given dataset, computed by averaging category-wise intersection ratios (refer to Equation (7)). Its score spans from 0 to 1, where a higher score signifies greater segmentation precision and superior detail differentiation by the model. Unlike pixel accuracy, mIoU offers a more holistic insight into the model's capability to discern fine dataset features.

$$\mathrm{mIoU} = \frac{1}{\mathrm{k}+1}\sum_{i=0}^{\mathrm{k}} \frac{p_{ii}}{\sum_{j=0}^{\mathrm{k}} p_{ij} + \sum_{j=0}^{\mathrm{k}} p_{ji} - p_{ii}} \tag{7}$$

4.2. Experimental environment and training strategy

4.2.1. Experimental environment

The performance of image semantic segmentation algorithm is not only affected by the ability of model to learn image features, but also closely related to the software and hardware environment used in training. The hardware environment in which the algorithms were run in this study is equipped with high-performance components, including an Intel® Core™ i9-9900K CPU and an NVIDIA GeForce RTX 3080 graphics card with 10GB of video memory to ensure computing power. In terms of software configuration, the compatible 64-bit Windows 10 operating system was selected, CUDA 11.0 was used to achieve GPU acceleration, and Pytorch deep learning framework was adopted to support the efficient execution of the algorithm.

4.2.2. Training strategies

This algorithm utilizes a pre-trained ResNet50 on ImageNet as the encoding foundation, implementing a stratified learning rate approach across network components: the core ResNet50 structure learns at a rate of 0.001, whereas the spatial information module, global pooling module, and decoder network are assigned a rate of 0.01. Training optimization is achieved through stochastic gradient descent (SGD), with batches of 8 samples and a weight decay of 0.0001. Model convergence is reached after 60 epochs and circa 80,000 iterations. Cross-entropy loss guides the learning process, and the "poly" policy, a dynamic learning rate adjustment methodology adjusting rates in accordance with iteration counts, is employed. The mathematical formulation for this adaptive rate reduction strategy is as follows.

$$LR = lr \times \left(1 - \frac{n}{N}\right)^{p} \tag{8}$$

Formula (8) describes the process of dynamic adjustment of learning rate LR with training iteration, where lr represents the initially set learning rate, n refers to the current iteration step, N is the predetermined maximum number of iterations, and p is a attenuation factor, which is set at 0.9 in this experiment. During the training period, the input images were randomly scaled and then randomly cropped with 380×380 pixels as data preprocessing methods to enhance the generalization ability of the model.

4.3. Experimental results and analysis

In the experimental comparison, this study selected the classical full convolutional network (FCN) as the baseline model, and made corresponding adjustments to it: ResNet50 was used as its encoder, bilinear interpolation was used for upsampling, and the learning rate was set at 0.01 to build the control group. As shown in Table 2, the model performance improved significantly with the introduction of different modules one by one. With the successive integration of SIM, GPM, FFM and BRM, the performance of FCN is gradually improved, which directly verifies the utility of these four modules. In particular, the addition of GPM led to a 4.94 percentage point surge in mIoU to 75.06%, highlighting the critical role of global context information in improving model segmentation performance.

Table 2. Influence of different modules on segmentation results

SIM	GPM	FFM	BRM	mIoU(%)
				67.31
√				70.12
√	√			75.06
√	√	√		77.23
√	√	√	√	78.98

The mIoU performance comparison between EDNet and FCN clearly reflects the significant advantages of EDNet. Compared with FCN, the EDNet proposed in this paper greatly enhances the utilization efficiency of multi-scale feature information of the encoder network by integrating spatial information module, feature fusion module, global pooling module and boundary thinning module, thus surpassing the segmentation accuracy and demonstrating the superiority of this method in the field of image semantic segmentation.

To rigorously validate our proposed method, we established a consistent experimental framework and controlled for similar parameter counts, examining the impact of diverse feature fusion tactics on performance enhancement. Table 3 summarizes a comparative analysis incorporating three literature-cited methodologies: Semantic Embedding Branch (SEB) from [31], Residual Convolution Unit (RCU) from [32], and Enhanced Feature Fusion Decoder (EFFD) from [33]. The findings indicate that EDNet outperforms the attention-based schemes of SEB, RCU, and EFFD in boosting model efficacy, furnishing compelling evidence of its supremacy.

Table 3. Comparison of mIoU based on ResNet50

Model	mIoU(%)
ResNet50-SEB	71.02
ResNet50-RCU	73.95
ResNet50-EFFD	74.46
EDNET	78.98

In the in-depth comparative analysis, we conduct a comprehensive evaluation with the classical segmentation networks in the field from the two dimensions of model parameter scale and semantic segmentation performance (measured by mIoU). Through a series of well-designed experiments, we obtained the detailed data shown in Table 4, where the "Million" index represents the number of trainable parameters of each model. All comparison models are run under uniform experimental conditions, and the learning rate, encoder configuration and other aspects are carefully tuned to achieve the best performance. It is worth noting that the method, FCN and GCN are all built on ResNet50, while DeepLabV3+ uses the more parametric intensive ResNet101 as its core.

Although EDNet has about 5 million more parameters than GCN and FCN, its segmentation performance has been significantly exceeded. From the empirical results presented in Table 4, although DeepLabV3+ 's mIoU performance is slightly better, it is at the cost of nearly 30 million additional parameters, and the mIoU improvement is only a tiny 0.1%. This finding strongly demonstrates that our approach can achieve an efficient use of computing resources while maintaining high accuracy, cleverly balancing precision and efficiency. In contrast, the method proposed in this paper can find a more balanced fulcrum between the number of parameters and the segmentation accuracy, and thus achieve a better segmentation effect.

Table 4. Comparison of segmentation methods

Model	Million	mIoU(%)
GCN	25.40	65.24
FCN	24.23	67.31
DeepLabV3+	59.34	73.57
EDNET	30.11	78.98

5. Conclusions

Addressing spatial detail loss and the challenge of integrating multi-scale features in image semantic segmentation due to subsampling, this paper introduces EDNET, an innovative segmentation strategy focused on multi-scale feature consolidation. By incorporating spatial information branches in up-sampling stages to replenish edge details and employing a custom feature fusion module to effectively merge representations across various scales, it bolsters segmentation precision. Additionally, the integration of global pooling and boundary sharpening modules augments the model's performance. Assessments on the PASCAL VOC 2012 dataset revealed a notable 11.67% increase in mean IoU compared to conventional full convolutional networks. Further, side-by-side evaluations with other sophisticated methodologies under identical conditions confirmed substantial performance advancements, attesting to the potency and real-world applicability of the proposed EDNET framework.

References

[1] Zhu, A., Li, J., & Lu, C. (2021). Pseudo view representation learning for monocular RGB-D human pose and shape estimation. IEEE Signal Processing Letters, 29, 712-716.

[2] Lan, G., Liu, X. Y., Zhang, Y., & Wang, X. (2023). Communication-efficient federated learning for resource-constrained edge devices. IEEE Transactions on Machine Learning in Communications and Networking.

[3] Zi, Y., Wang, Q., Gao, Z., Cheng, X., & Mei, T. (2024). Research on the Application of Deep Learning in Medical Image Segmentation and 3D Reconstruction. Academic Journal of Science and Technology, 10(2), 8-12.

[4] Li, K., Zhu, A., Zhou, W., Zhao, P., Song, J., & Liu, J. (2024). Utilizing Deep Learning to Optimize Software Development Processes. arXiv preprint arXiv:2404.13630."

[5] Lan, G., Wang, H., Anderson, J., Brinton, C., & Aggarwal, V. (2024). Improved Communication Efficiency in Federated Natural Policy Gradient via ADMM-based Gradient Updates. Advances in Neural Information Processing Systems, 36.

[6] Wang, X. S., Turner, J. D., & Mann, B. P. (2021). Constrained attractor selection using deep reinforcement learning. Journal of Vibration and Control, 27(5-6), 502-514.

[7] Li, K., Zhu, A., Zhou, W., Zhao, P., Song, J., & Liu, J. (2024). Utilizing Deep Learning to Optimize Software Development Processes. arXiv preprint arXiv:2404.13630.

[8] Xiao, M., Li, Y., Yan, X., Gao, M., & Wang, W. (2024). Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. doi:10.48550/ARXIV.2404.08279

[9] Zhu, A., Li, K., Wu, T., Zhao, P., Zhou, W., & Hong, B. (2024). Cross-Task Multi-Branch Vision Transformer for Facial Expression and Mask Wearing Classification. arXiv preprint arXiv:2404.14606.

[10] Ning, Q., Zheng, W., Xu, H., Zhu, A., Li, T., Cheng, Y., ... & Wang, K. (2022). Rapid segmentation and sensitive analysis of CRP with paper-based microfluidic device using machine learning. Analytical and Bioanalytical Chemistry, 414(13), 3959-3970.

[11] Kim, M., Lee, H., & Cho, S. Attention-Guided Dual-Task Learning for Simultaneous Image Classification and Semantic Segmentation[J]. Computer Vision and Image Understanding, Vol. 214, Article 103041, February 2023.

[12] Misra D, Nalamada T, Arasanipalai A U, et al. Rotate to attend: Convolutional triplet attention module[C]//Proceedings IEEE Winter Conference on Applications of Computer Vision, WACV 2021: 3139-3148.

[13] Dai, W., Tao, J., Yan, X., Feng, Z., & Chen, J. (2023, November). Addressing Unintended Bias in Toxicity Detection: An LSTM and Attention-Based Approach. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 375-379). IEEE.

[14] Xin Chen, Yuxiang Hu, Ting Xu, Haowei Yang, Tong Wu. (2024). Advancements in AI for Oncology: Developing an Enhanced YOLOv5-based Cancer Cell Detection System. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),75-80, doi:10.55524/ijircst.2024.12.2.13.

[15] Yulu Gong, Haoxin Zhang, Ruilin Xu, Zhou Yu, Jingbo Zhang. (2024). Innovative Deep Learning Methods for Precancerous Lesion Detection. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),81-86, doi:10.55524/ijircst.2024.12.2.14.

[16] Yan, C., Qiu, Y., Zhu, Y. (2021). Predict Oil Production with LSTM Neural Network. In: Liu, Q., Liu, X., Li, L., Zhou, H.,

Zhao, HH. (eds) Proceedings of the 9th International Conference on Computer Engineering and Networks. Advances in Intelligent Systems and Computing, vol 1143. Springer, Singapore. https://doi.org/10.1007/978-981-15-3753-0_34.

[17] Wang, X. S., & Mann, B. P. (2020). Attractor Selection in Nonlinear Energy Harvesting Using Deep Reinforcement Learning. arXiv preprint arXiv:2010.01255.

[18] C. Yan, "Predict Lightning Location and Movement with Atmospherical Electrical Field Instrument," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2019, pp. 0535-0537, doi: 10.1109/IEMCON.2019.8936293.

[19] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024). Survival prediction across diverse cancer types using neural networks. doi:10.48550/ARXIV.2404.08713

[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. 2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE Press, 2015: 3431-3440.

[21] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[22] Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation[C]. International Conference on Medical Image Computing and Computer-assisted Intervention, 2015: 234-241.

[23] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. arXiv preprint arXiv:1412.7062, 2014.

[24] Lan, G., Han, D. J., Hashemi, A., Aggarwal, V., & Brinton, C. G. (2024). Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algorithm Design and Convergence Analysis. arXiv preprint arXiv:2404.08003.

[25] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 2881-2890.

[26] Papadomanolaki M, Vakalopoulou M,Karantzalos K.A Novel Object-Based Deep Learning Framework for Semantic Segmentation of Very High-Resolution Remote Sensing Data: Comparison with Convolutional and Fully Convolutional Networks [J]. Remote Sensing,2019,11(6).

[27] Pereira M B,Santos J A D.An End-to-end Framework For Low-Resolution Remote SensingSemantic Segmentation[J].2020,arXiv/abs/2003.07955.

[28] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770- 778.

[29] Li, Y., Yan, X., Xiao, M., Wang, W., & Zhang, F. (2024). Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets. In Proceedings of the 2023 13th International Conference on Communication and Network Security (pp. 77–81). Association for Computing Machinery.

[30] Vicente, S., Carreira, J., Agapito, L., & Batista, J. (2014). Reconstructing pascal voc. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 41-48).

[31] Zhang Z L, Zhang X Y, Peng C, et al. Exfuse: Boosting Semantic Segmentation via Enhanced Feature Integration[C]. Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer-Verlag, 2018: 273-288.

[32] Shah, A., Kadam, E., Shah, H., Shinde, S., & Shingade, S. (2016, September). Deep residual networks with exponential linear unit. In Proceedings of the third international symposium on computer vision and the internet (pp. 59-65).

[33] Ma Z H, Gao H J, Lei T. Algorithm for Semantic Segmentation employing an Augmented Feature Fusion Decoder[J]. Computer Engineering, 2020, 46(5): 254-258.