IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 March 2024, accepted 15 April 2024, date of publication 19 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391249

RESEARCH ARTICLE

An Attention-Based Improved U-Net Neural Network Model for Semantic Segmentation of Moving Objects

ZHIHAO CUI
College of Physical Education, Pingdingshan University, Pingdingshan, Henan 467000, China
e-mail: 2716@pdsu.edu.cn

ABSTRACT The precise semantic segmentation for moving targets has always been a challenging task in computer vision. The existing methods basically had some limitations in semantic segmentation, such as inability to handle deformation of moving targets, blurred boundaries, and other issues. To address the issue, this paper develops an improved U-Net model based on attention mechanism for this purpose. Firstly, we introduce an attention mechanism to enhance the perceptual ability of the U-Net model. By adding attention modules at different levels between the encoder and decoder, the network can pay more attention to the key features of moving targets at different levels. Then, we add a residual module to improve robustness and complete the capsule network for semantic segmentation of moving targets. By learning the deformation information of moving targets, the network can better adapt to moving targets with different shapes. We have conducted experimental verification on multiple public datasets. The experimental results show that the proposed method has superior performance in semantic segmentation tasks of moving targets. Compared with traditional U-Net-based models, the proposal shows significant improvements in accuracy and robustness.

INDEX TERMS Attention mechanism, semantic segmentation, moving targets, computer vision.

I. INTRODUCTION

The semantic segmentation of moving targets plays a crucial role in the field of computer vision [1]. It is a task aimed at labeling each pixel in an image as a different semantic category, providing key information for applications such as autonomous driving, intelligent monitoring, and augmented reality [2]. However, traditional semantic segmentation methods still face certain challenges due to the changes in scale, shape, and appearance of moving targets [3]. To address this issue, researchers have proposed an improved neural network model - the attention based improved U-Net neural network model [4], [5]. This model combines encoder decoder architecture and attention mechanism, which can overcome some limitations in traditional methods and achieve significant performance improvement in semantic segmentation tasks of moving targets [6], [7].

We have noticed some limitations of traditional U-Net neural network models in handling semantic segmentation of moving targets [8]. The traditional U-Net model uses an encoder to extract image features and a decoder to map these features to the corresponding pixel space [9]. However, traditional U-Net models encounter issues of information loss and blurring when dealing with moving targets, as they ignore the position and contextual information of the target in the image [10]. To overcome these issues, we introduced an attention mechanism to improve the U-Net model. The attention mechanism can automatically learn the important parts of the target in the image, enabling the model to perform semantic segmentation more accurately [35].

We adopted a self-attention mechanism, which calculates the correlation between the target and other pixels in the image, and selects the pixel with the highest correlation as the attention region. The research framework is shown in Figure 1. By introducing a self-attention mechanism, our model can better capture the details of moving targets, thereby improving the accuracy and robustness of

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

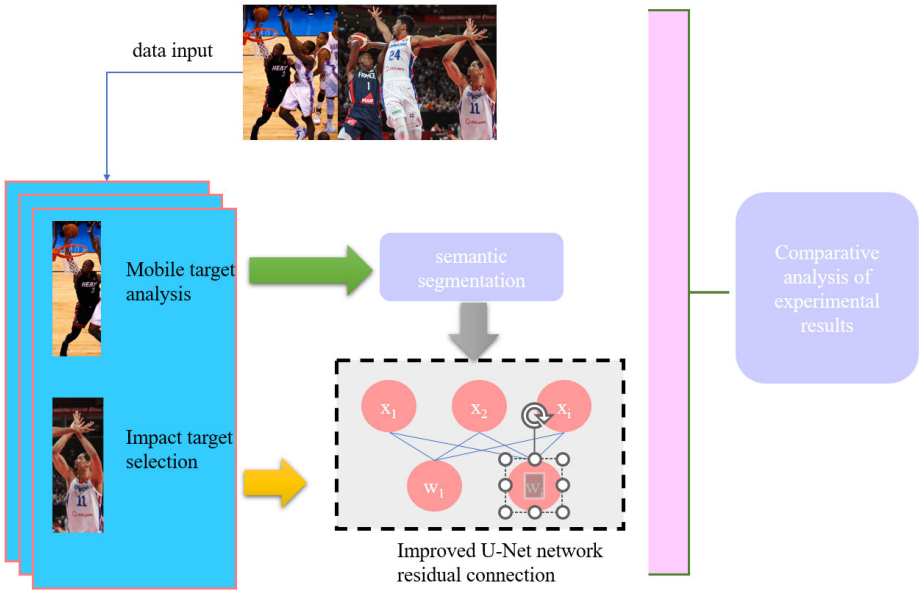

FIGURE 1. Research framework.

semantic segmentation. We will verify the effectiveness and superiority of our model by comparing the performance of the traditional U-Net model and our proposed model on public datasets. The research framework is shown in Figure 1. In addition, we will also explore the impact of attention mechanisms on improving the semantic segmentation performance of moving targets, and discuss possible future improvements and application directions. This article aims to propose an improved U-Net neural network model based on attention to address the limitations of traditional methods in semantic segmentation of moving targets. We believe that by introducing attention mechanisms into the U-Net model, the performance of semantic segmentation tasks for moving targets can be significantly improved, providing more accurate and reliable results for related applications in the field of computer vision.

Main contributions of this paper can be summarized as two points. For one thing, the study introduces attention mechanism in semantic segmentation of moving targets, allowing networks to adaptively learn regions of interest and obtain more contextual information from a global perspective. By adding an attention module in U-Net, our model can better handle the semantic segmentation task of moving targets. For another, we have improved the encoder of U-Net by introducing a residual connection structure, which can effectively reduce the number of parameters and improve computational efficiency, while residual connection can reduce gradient vanishing problems and accelerate the convergence speed of the network.

The structure of the article is organized as follows. Chapter 1 elaborates on the research direction of attention mechanism and U-Net neural network in semantic segmentation of moving targets. Chapter 2 provides an explanation of the research objectives based on previous research conducted by relevant scholars. Chapter 3 presents a multi-dimensional end-to-end attention algorithm with gradient jumps. Chapter 4 elaborates on the complete U-Net network neural model. Chapter 5 explains the data sources and conducts detection experiments. Chapter 6 summarizes the entire text.

II. LITERATURE REVIEW

Scholars in this field, such as Kuan et al. [11] proposed an improved U-Net model based on multi-scale attention mechanism. By introducing a multi-scale attention module in the network, effective fusion and utilization of different scale information of moving targets are achieved, thereby improving the accuracy and robustness of semantic segmentation. The team of researchers Ramadan et al. [12] and others proposed an improved U-Net model that combines spatial attention and channel attention. Through the dual constraints of spatial attention module and channel attention module, it achieves fine segmentation of moving targets, effectively solving the problems of blurring and missed segmentation in traditional U-Net models when segmenting moving targets. Damkliang et al. [13] and other scholars proposed an adaptive spatial attention mechanism to address common challenges such as occlusion and deformation in semantic segmentation of moving targets. During network training, attention weights are dynamically adjusted to improve the accuracy and robustness of the U-Net model for moving target edges and details.

Qiu et al. [14] proposed an improved method to address the issues of class imbalance and sample sparsity in the U-Net model for semantic segmentation of moving targets. By introducing an attention mechanism, this method can more effectively handle class imbalance situations and improve the accuracy of the model in moving target category recognition and segmentation. De la Sotta et al. [15] attempted to improve the U-Net model by combining reinforcement learning and attention mechanisms to further

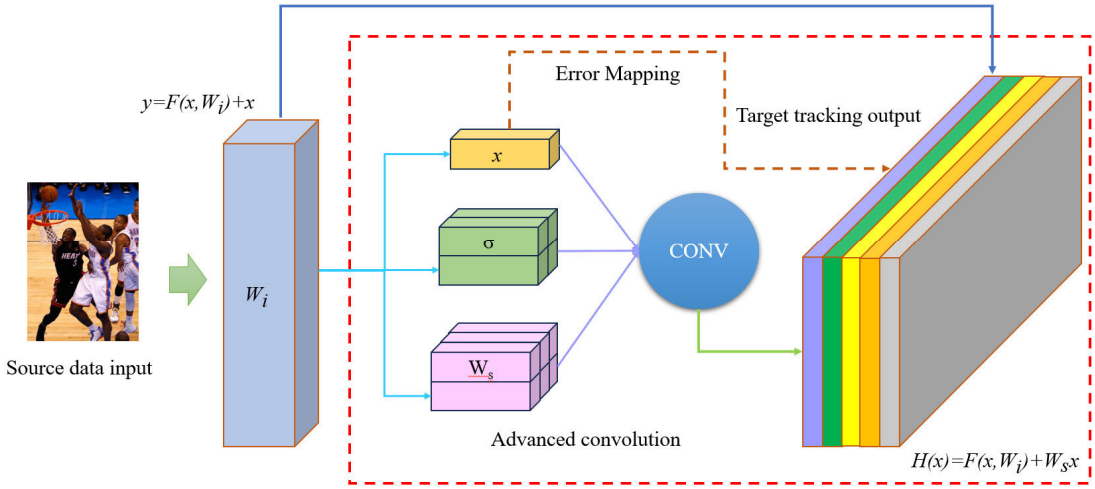

FIGURE 2. Residual module for semantic segmentation.

optimize the semantic segmentation performance of moving targets. They proposed an adaptive attention network based on reinforcement learning, which enables the network to adaptively adjust attention distribution through reinforcement learning algorithms, thereby improving the performance and robustness of semantic segmentation of moving targets.

The recent research results of scholars have demonstrated the potential of attention based improved U-Net neural network models in semantic segmentation tasks of moving targets, providing effective ideas and methods for solving key problems in semantic segmentation of moving targets, which is of great significance for promoting progress in this field.

III. METHODOLOGY

A. RESIDUAL NETWORK WITH GRADIENT JUMPING

We have introduced a multidimensional end-to-end attention mechanism to improve the performance of U-Net [16], [17]. The multi-dimensional end-to-end attention algorithm introduces an attention module to weight and fuse feature maps at different levels, so that the network can automatically focus on important features [18]. This attention mechanism can improve the model's perception of target boundaries and details [19]. In gradient jumping residual networks, attention modules can be inserted between each encoder and decoder to enhance feature transmission and utilization.

Specifically, the attention module can learn and generate a weight vector to perform weighted fusion on feature maps at different levels. This way, the network can better focus on important features, reduce information loss and ambiguity issues. The residual network with gradient jumps also introduces residual connections to more effectively transmit gradient information. Residual connection directly transfers gradient information from the decoder back to the encoder by adding up the features between the encoder and decoder [20]. This can help the network better learn detailed information and target boundaries, and alleviate the problem of gradient vanishing. The residual network of gradient jumps is shown in Figure 2.

The deep residual network framework is not simply fitting multiple layers directly to an ideal latent mapping, but rather using these layers to specifically fit a residual mapping. Assuming x is the input and $H(x)$ is the expected output. Considering traditional network structures, if the accuracy learned has reached saturation, identity mapping learning is required to ensure learning accuracy by making the input x close to the output $H(x)$. The residual network structure uses shortcut links to directly take input x as the output result $H(x) = F(x)+x$. When $F(x) = 0$ and $H(x) = x$, it is an identity mapping. The learning objective of the residual network is $F(x)$: $=$ $H(x)$-x. Therefore, the training objective of the network model is to make the residual approach 0, while the accuracy does not decrease as the number of network layers increases. Utilize residual networks every few layers [21].

$$y = F(x, W_i) + x \tag{1}$$

In the formula, x and y respectively represent the input and output vectors of the network layer, and the function $F(x, W_i)$ is the residual mapping that the model needs to learn. This building block has a total of two layers, where $F = W_2\sigma(W_1x)$, parameters σ This represents the ReLU activation letter.

In the comparison between ordinary networks and residual networks, it is a very important factor to make a fair comparison. We can use a quick connection to make a linear mapping W_s:

$$H(x) = F(x, W_i) + W_sx \tag{2}$$

The speed model encodes the speed sequence previously fed back, and the encoded visual features are used to predict the steering wheel angle. The visual features are recognized together with the feedback speed features, both of which use

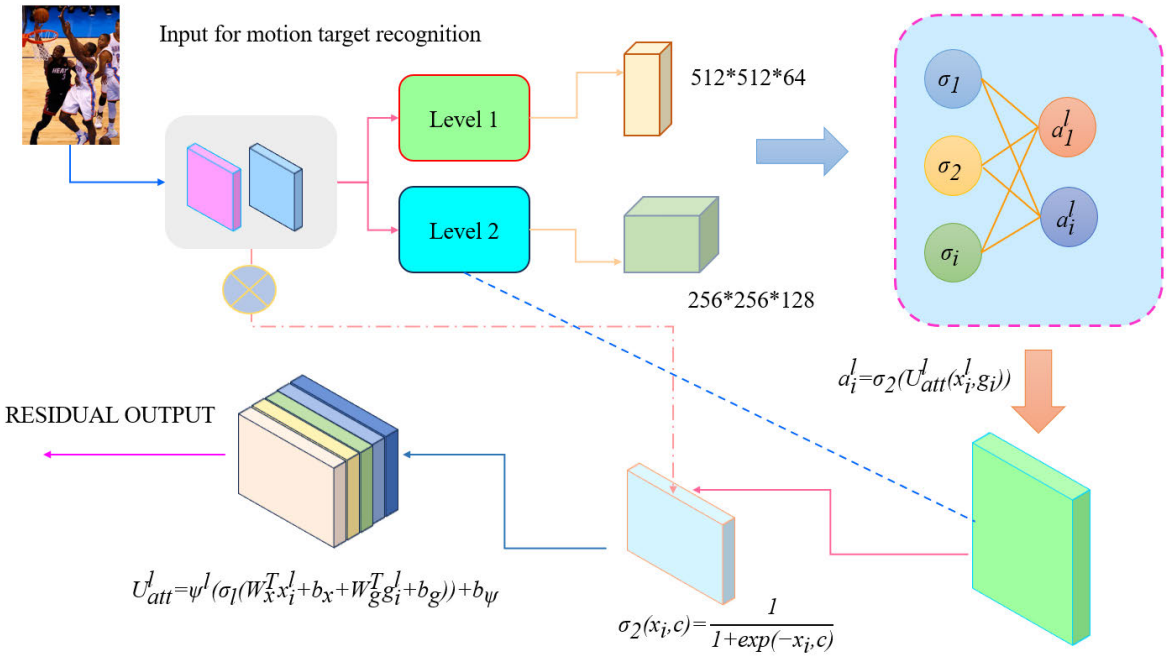

FIGURE 3. Residual module for semantic segmentation.

the average absolute error as the loss function, and a weight parameter is used to adjust the weight between the two losses.

$$a = \frac{s_e - s_s}{c} \tag{3}$$

In the formula, a is the identified acceleration interval s is the initial interval. C is the interval for calculating acceleration.

B. MULTISCALE RESIDUAL MODULE

1) DESIGN OF RESIDUAL MODULE FOR SEMANTIC SEGMENTATION N

We use an attention based improved U-Net neural network model to solve the semantic segmentation problem of moving targets. In the model, we introduce a multidimensional end-to-end attention algorithm and use residual modules to enhance the performance of the network [36]. Specifically, we insert residual modules into the encoder and decoder of U-Net, use residual modules in the encoder to reduce information loss, and use residual modules in the decoder to improve segmentation accuracy [37]. At the same time, we also embed multi-dimensional end-to-end attention modules into the encoder and decoder to improve the expression ability of features [22].

The input is the upper level feature map $x^l \in R^{Fl}$ (where Fl represents the number of corresponding layer feature maps) and the lower level feature map $g_i \in R^{Fg}$. The output is the element level multiplication of the input feature map and attention coefficient $X^l = x^l a_i$, where the feature attention coefficient $a_i \in [0,1]$ is used to strengthen the feature weight of the target area, so that the final output contains activated features related to the target category, as shown in the stained part of the feature map in the figure.

$$U_{att}^l = \psi^l(\sigma_l(W_x^T x_i^l + b_x + W_g^T g_i^l + b_g)) + b_\psi \tag{4}$$

$$a_i^l = \sigma_2(U_{att}^l(x_i^l, g_i)) \tag{5}$$

$$\sigma_2(x_i, c) = \frac{1}{1 + exp(-x_i, c)} \tag{6}$$

In the formula, it represents the use of sigmoid activation function operation, and the feature of Attention Rescue includes a series of parameters: linear operation $W_x \in R^{Fl}$, $W_g \in F^g$, $\psi \in F^{int}$ and the corresponding convolution bias terms b_x, b_g, b_ψ。The module runs as shown in Figure 3.

On this basis, designing residual modules can further improve the performance and stability of the model. Specifically, the residual module adopts skip connections and residual learning methods, which can effectively solve the problems of gradient vanishing and exploding, thereby enabling the model to achieve specific tasks more deeply and accurately [23]. The use of residual modules can better handle specific structures and features in images, thus enabling more accurate semantic segmentation.

2) LOSS FUNCTION

In the multi-dimensional end-to-end attention algorithm for moving targets, the loss function is used to evaluate the degree of difference between the model's predicted results and the actual annotated results. It plays a role in optimizing model parameters during the training process, helping the model perform semantic segmentation tasks more accurately. Usually, multi-dimensional end-to-end attention algorithms for moving targets use the Cross Entropy Loss function as the main loss function [24]. This loss function calculates

the difference between the predicted results and the actual annotated results and minimizes it. In the multi-dimensional end-to-end attention algorithm for moving targets, in order to better handle the semantic segmentation task of moving targets, other loss functions can also be introduced. For example, the authors propose a loss function that combines attention mechanism to capture and emphasize key information of moving targets. The loss function in the multi-dimensional end-to-end attention algorithm for moving targets plays a crucial role in the training process [38].

The study used a binary cross entropy loss function for classification and confidence prediction. After summarizing the methods for handling imbalanced samples, it was decided to use the focal loss function instead of the binary cross entropy loss function in the original classification and confidence loss function [25].

$$P_t = \begin{cases} p & y = 1 \\ 1 - p & other \end{cases} \tag{7}$$

$$H(P_t) = -ln(P_t) \tag{8}$$

where P_t represents the probability that the predicted sample belongs to true, and y represents a label value of $(+1, -1)$. This approach can adjust the weight values of positive and negative samples while reducing the weight of easily classified samples, making the model more focused on difficult to classify samples during training.

$$FL(P_t) = -\alpha(1 - P_t)^r ln(P_t) \tag{9}$$

When a sample is misclassified (i.e. P_t is very small), the modulation factor (1-P_t) approaches 1 and the loss is not affected. When Pt approaches 1 and the factor (1-P_t) approaches 0, the weight values of the divided samples are reduced, making a very small contribution to the overall loss value.

C. SETTING OF SEMANTIC SEGMENTATION INDICATORS FOR MOBILE IMAGES

The setting of semantic segmentation indicators for mobile images is a key factor affecting the accuracy and performance evaluation of neural network models. In image semantic segmentation tasks running on mobile devices, due to limitations in computing power and storage resources, more efficient and reliable indicators are needed to evaluate the performance of the model. The proportion of correctly classified pixels to the total number of pixels. This is one of the most basic indicators, reflecting the evaluation of the overall classification ability of the model. The proportion of correctly classified pixels to the number of pixels predicted by the model to be positive. This indicator measures the accuracy and precision of the model's classification prediction, taking into account the false alarm rate. The proportion of correctly classified pixels to the actual number of pixels corresponding to positive examples. This indicator measures the efficiency and ability of the model to find actual positive examples. Average pixel accuracy refers to the weighted average accuracy of pixels in different categories. This indicator is more suitable for comparing the performance of multi classification problems. The structure of semantic segmentation indicators for mobile images is shown in Table 1.

TABLE 1. Structure of semantic segmentation indicators for mobile images.

Unit	Layer(type)	Output Shape	Layer(type)
Level 1	Conv1	512*512*64	Conv6
	MRConv1	512*512*64	RConv6
	Max_pooling	256*256*64	Sampling
	Conv2	256*256*128	RConv6
Level 2	Conv1	256*256*128	Conv6
	MRConv1	256*256*128	RConv6
	Max_pooling	128*128*128	Sampling
	Conv2	128*128*512	RConv6

The average intersection to union ratio is commonly used in multi classification computer vision research tasks. It calculates the IOU results for multiple classifications and then calculates the average value to mIOU. In the k (1 in $k + 1$ represents the background pixel) classification task, p_{ii} represents the correctly labeled pixel, p_{ij} represents the pixel that should be marked as i and marked as j, and p_{ij} represents the pixel that should be marked as j but is marked as i.

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{10}$$

In computer vision related tasks, there are generally two types of model evaluation methods. One is the method of calculating intersection to union ratio (IOU) mentioned in the previous section, and the other is based on image pixels. Image segmentation is essentially pixel level prediction. The Mean Intersection Over Union (Mean IOU) refers to the average ratio of the intersection and union between the predicted segmentation and the actual segmentation by the model. This indicator considers the accuracy of classification prediction and the degree of spatial position matching, and is one of the commonly used indicators in mobile image semantic segmentation tasks [26].

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{11}$$

In the formula, p_{ii} represents the correctly labeled pixel, and p_{ij} represents the pixel where the image should have been labeled as i but was incorrectly labeled as j. MPA (mean Pixel Accuracy) is an evaluation metric for multi classification tasks, which calculates the PA (Pixel Accuracy) for each pixel category and then adds it up to take the average value.

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{12}$$

However, in semantic segmentation of moving targets, due to the influence of moving target speed, it is usually necessary

to pay attention to the segmentation results at different time steps. Therefore, when evaluating the segmentation results of moving targets, it is necessary to set specific indicators to consider the segmentation results at different time steps. The Temporal Merged Intersection over Union (tmIoU) can comprehensively reflect whether the algorithm is effective in segmenting moving targets, and has stronger practical value.

D. IMPROVED RESIDUAL STRUCTURE DESIGN FOR U-NET NETWORKS

U-Net is a classic convolutional neural network architecture widely used in semantic segmentation tasks. It has two parts: an encoder and a decoder. The encoder is used to extract features, while the decoder is responsible for restoring image resolution and performing pixel level classification. However, traditional U-Net networks may encounter some problems when dealing with moving targets, such as unclear boundaries and partial occlusion of objects. To address these issues, we can introduce residual structures to improve the U-Net network. The residual structure is achieved by introducing cross layer connections, which can make the network deeper without the problem of vanishing or exploding gradients. Between each layer of the encoder and decoder in U-Net, we can add a residual block to gradually learn richer feature representations from shallow to deep layers.

In the unit mapping, $y = x$ is the observed value, while $H(x)$ is the predicted value, so $F(x)$ corresponds to the residual, which is the residual mapping. Therefore, the general formula for representing the residual block is:

$$y_l = h(x_l) + F(x_l, W_l) \tag{13}$$

Among them, x represents the input vector, y represents the output vector, and function F represents the residual mapping. The input and input dimensions must be consistent, otherwise, linear mapping matching dimensions needs to be performed [27].

$$y_l = W_s h(x_l) + F(x_l, W_l) \tag{14}$$

The meaning of the parameters is consistent with Chapter 1, which is 1 × The convolution operation of 1 is generally only used for dimensionality increase and dimensionality decrease.

We introduce an attention mechanism into the residual structure to enhance attention to moving targets. The attention mechanism can adaptively adjust the weights of feature maps based on the content of the image, improving the representation ability of the target area. The improved U-Net network based on attention has better feature extraction ability and stronger target attention ability, which can achieve better performance in semantic segmentation tasks of moving targets. The design of this network structure can not only improve the accuracy of semantic segmentation, but also maintain clear boundaries and reduce misclassification caused by partial occlusion of the target. The research pseudocode is shown in Algorithm 1.[1]

Algorithm 1 Attention-based Improved U-Net Model

```
Input: The input vector x and output vector y, the final value S_e
of interval velocity, the lower level feature map g_i, the
of true prediction of the sample, the correctly labeled pixel p_ii,
probability P_t the incorrectly labeled pixel p_ij, and i and j are
adaptive variables.
Input x approaches output learning accuracy
y = F(x,W_i)+x
Calculate the acceleration interval for recognition using for-
3: (s_e-s_s)/c
for all i and j = 1 to R do
    Insert the residual module into the encoder and decoder
of U-Net
    Using focal loss function instead of original classification
    Reduce the weight of easily classified samples
    FL(P_t) = -α (1)-P_t)^rln(P_t)
        for p_ii 1: R
            Calculate the average value to mIOU
        if For-10 calculation value does not meet semantic
segmentation requirements
            Calculate the PA for each pixel category
        else
            Evaluate the segmentation results of moving targets
    end for
  end for
```

E. CAPSULE NETWORK FOR SEMANTIC SEGMENTATION OF MOVING OBJECTS

The attention mechanism can make the network pay more attention to the regions of interest in the image during the learning process, thereby improving the accuracy of the model's segmentation of the target. Introducing the concept of capsule networks in semantic segmentation of moving targets can further improve the performance of the model. Capsule network is an alternative neural network architecture that represents the direction and length of feature vectors through capsules. Compared to traditional neural units, capsules can better encode spatial relationships and pose information [28]. By combining capsule networks and U-Net, we can utilize capsule networks to better capture the details of targets and improve the segmentation performance of moving targets.

The input of the capsule network for semantic segmentation of moving targets is an 8 * 8 * 128 feature map extracted from the complex convolutional encoding part of the U-Net network for primary feature extraction. After shape transformation, the 8 * 8 * 1 * 128 feature map is obtained. This output forms the input for the first set of capsules, representing a grid of 8 * 8 capsules, each of which is a 128 dimensional vector. Subsequently, through the subsequent convolutional capsule layer, this process is now extended to any network in the grid, and the resulting network structure is shown in Figure 4.

[1] https://github.com/gzwzzu/EIS-PROJECT

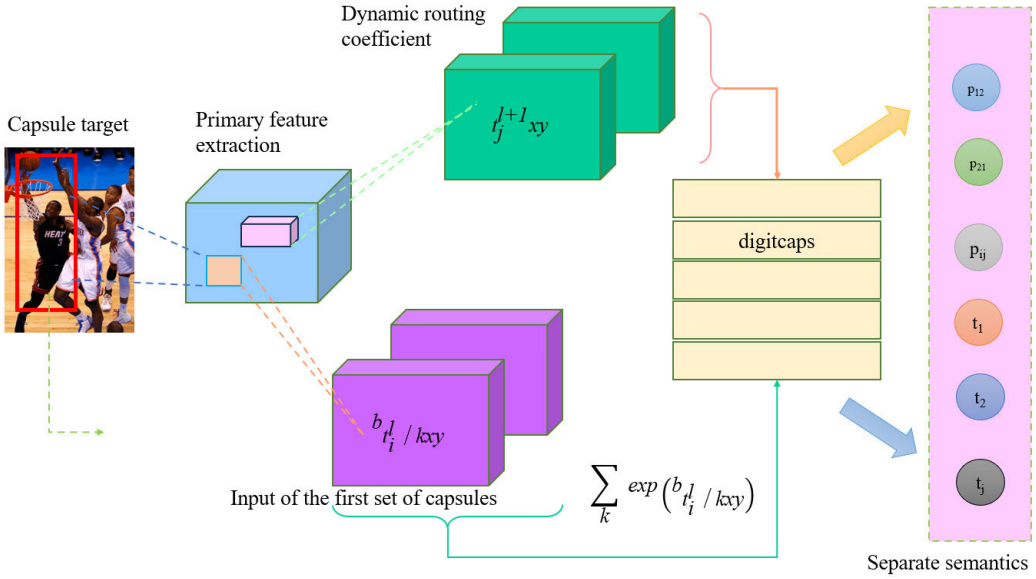

FIGURE 4. Capsule network structure.

To determine the final input $p_{xy} \in P$ for each microcapsule, calculate the weighted sum of these prediction vectors:

$$\boldsymbol{p}_{t_j^{l+1}xy} = \sum_j r_{t_i^l|t_j^{l+1}xy} \hat{u}_{t_j^{l+1}xy|t_i^l} \tag{15}$$

In the formula, $\sum_j r_{t_i^l|t_j^{l+1}xy}$ is the routing coefficient of the dynamic routing algorithm, which is calculated by the routing Softmas [29].

$$r_{t_i^l|t_j^{l+1}xy} = \frac{\exp\left(b_{t_i^l|t_j^{l+1}xy}\right)}{\sum_k exp\left(b_{t_i^l|Okxy}\right)} \tag{16}$$

Among them, $b_{t_i^l|t_j^{l+1}xy}$ The initial parameter is the logarithmic prior probability, and the prediction vector $\hat{u}_{xy|t_i^l}$ should route it to the sub capsule p_{xy}.

F. MULTITASK SEMANTIC SEGMENTATION FOR MOTION TARGET RECOGNITION

The semantic segmentation of moving targets is an important research direction in the field of computer vision today. When performing semantic segmentation, we usually need to label the target and distinguish different categories. This process can help computers better understand the properties of targets and identify and track them more accurately [30]. However, semantic segmentation of moving targets is not a simple problem [31]. Due to the frequent changes in the appearance and shape of moving targets, such as posture, speed, and lighting, recognition becomes more difficult. In addition, in practical applications, while recognizing a target, it is usually necessary to segment and recognize other related targets, which together constitute a multi task semantic segmentation problem.

The improved U-Net neural network model based on attention is a method for solving multi task semantic segmentation problems. This method applies attention mechanism to U-Net networks to highlight important features and information, thereby improving the processing performance and recognition performance of neural networks. The study adopted a novel attention mechanism to improve the network's ability to detect and segment moving targets. This mechanism can adaptively adjust the weights and parameters of the network, highlight the key features of the target, and thus improve the robustness and accuracy of the entire semantic segmentation system.

In this model, we use two attention modules to improve the processing ability of target dynamic changes and shape deformations. Firstly, a time series attention module is introduced in the encoder to capture the dynamic changes of the target in the time series. At the same time, a shape attention module is introduced in the decoder to capture the shape changes of the target. In addition, we have also improved the feature extractor of the U-Net model to better adapt to the semantic segmentation problem of moving targets. Specifically, we replaced traditional convolutional layers with depthwise separable convolutional layers to improve the efficiency and accuracy of the model.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATA SOURCES

For this study, we used two commonly used datasets, namely the Cityscapes dataset [32], [34] and the PASCAL VOC Challenge dataset [33]. The Cityscapes dataset is a widely used dataset for urban scene segmentation. It contains high-resolution images from Germany and other European cities, totaling approximately 5000 images, each with a resolution of 1024×2048 pixels. These images cover various urban scenes,

TABLE 2. mIoU of different models.

Number of samples collected (k)	DSP	The-Study	FPGA	ViBe	SAR	OpenCV
1	21.26	32.25	15.26	16.26	18.55	21
2	38.25	43.25	16.45	24.26	21.5	36
3	51.25	56.25	25.66	38.56	39.45	48
4	57.36	62.36	48.26	44.65	52.15	50
5	70.23	75.23	58	48.26	62.45	60
6	76.26	67.23	62	64.55	68.26	70
7	78.25	72.15	69	68.46	73.56	72
8	80.26	82.26	72	67.26	74.55	73
9	81.26	83.26	75	72.56	72.15	75
10	83.15	85.66	75	70.45	72.66	75

such as streets, sidewalks, buildings, cars, pedestrians, etc. The Cityscapes dataset provides detailed pixel level labels to identify the semantic category of each pixel in the image, such as roads, buildings, pedestrians, etc. This enables us to accurately learn the semantic information of urban scenes when training our attention based improved U-Net neural network model. We used CDW-2014 in the data cleaning process (http://changedetection.net/). The dataset is used to ensure its scalability.

In addition to the Cityscapes dataset, we also used the PASCAL VOC Challenge dataset, which is a commonly used computer vision dataset primarily used for object recognition, detection, and semantic segmentation tasks. The PASCAL VOC dataset contains approximately 10000 images from 20 different categories, each with varying resolutions. These images cover various different scenes, such as indoor, outdoor, and natural landscapes. The PASCAL VOC dataset also provides pixel level labels to identify the semantic category of each pixel in the image. Using the PASCAL VOC dataset can help us validate the generalization ability of our attention based improved U-Net neural network model in different scenarios.

B. LOSS FUNCTION MOVING TARGET SEMANTIC SEGMENTATION DETECTION

When it comes to the loss function of moving object detection, we usually focus on pixel level annotation in semantic segmentation tasks. The goal of semantic segmentation is to segment an image into different semantic regions and assign

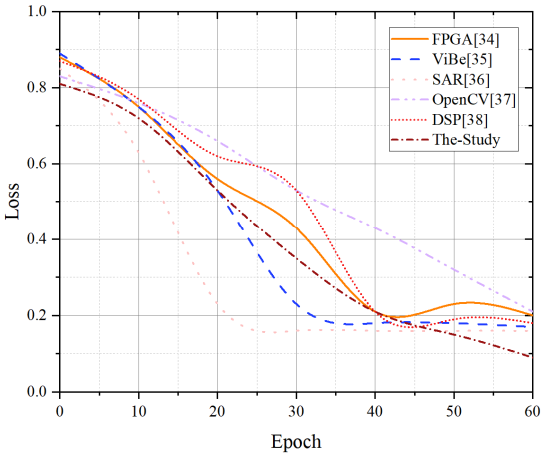

FIGURE 5. Semantic segmentation detection results of moving targets.

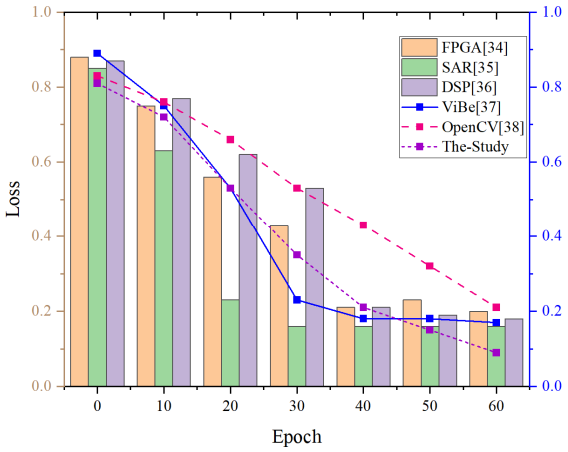

FIGURE 6. Biaxial comparison of detection results.

corresponding labels to each pixel. Moving object detection is the process of detecting and tracking moving targets in sequence videos. The loss function plays a crucial role in mobile object detection tasks, as it is used to measure the difference between the predicted results of the model and the true labels. In semantic segmentation tasks, the design of the loss function needs to take into account the position information of the target area and pixel level category information. For the semantic segmentation task of moving object detection, the design of the loss function needs to consider the position information of the target area and the pixel level category information. The loss function can be improved by introducing attention mechanisms and multitasking learning to enhance the accuracy and robustness of the model, as shown in Figures 5 and 6.

As shown in Figures 5 and 6, the number of iterations has an impact on the loss function loss values of different algorithms. As the number of iterations increases, the loss values of each algorithm's loss function generally show a downward trend. In FPGA and ViBe algorithms, the loss function has relatively small changes in loss values. This indicates that these two algorithms have relatively stable loss functions

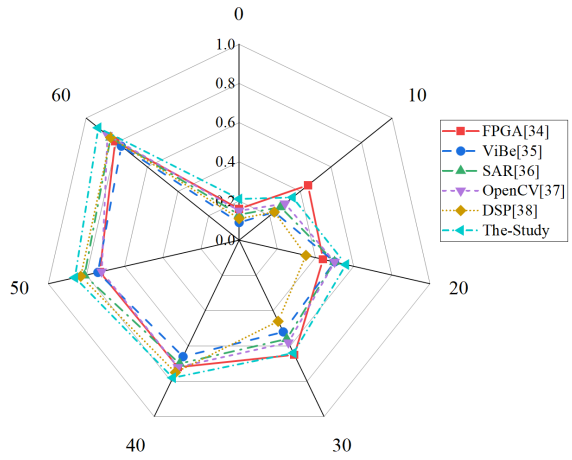

FIGURE 7. Model accuracy detection results.

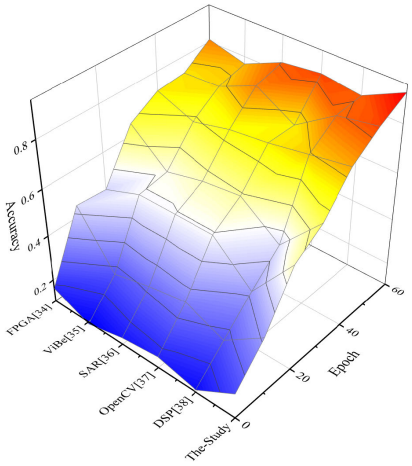

FIGURE 8. Analysis of conditional mapping detection results.

during the iteration process. In SAR and OpenCV algorithms, the loss value of the loss function rapidly decreases after the first few iterations and then tends to stabilize. It may indicate that these two algorithms have achieved better results after iterative optimization in the early stage. The loss function loss values of DSP algorithm and The Study algorithm vary greatly, especially DSP algorithm, whose loss values have a significant decrease in the early stage and then slowly stabilize. This may be related to the design and iterative optimization strategy of the algorithm itself. In this set of data, the loss function loss value of The Study algorithm is generally low, indicating its relatively good performance on this dataset. But we need to note that this is only the result obtained for this specific dataset, and careful consideration is needed for other datasets and practical application scenarios. We can also use cross validation or other evaluation methods to comprehensively evaluate the performance of algorithms.

C. ACCURACY DETECTION

In our study, we evaluated our model through cross validation, which was mainly divided into training and testing sets. We usually randomly divide the dataset into multiple subsets, with one subset used as the test set and the remaining subsets used as the training set for model training. During the training process, we use the data from the training set to train the model, and use the data from the test set to verify the accuracy of the model. We also used some classic evaluation metrics to evaluate the accuracy of the model, which refers to the proportion of correctly classified samples to the total number of samples. In our research, we also employed some data augmentation and noise processing techniques to improve the accuracy of the model. We evaluate the accuracy of the model through cross validation and use classic evaluation metrics to statistically evaluate the accuracy of the model's prediction results, as shown in Figure 7 and Figure 8.

As shown in the results of Figures 7 and 8, the accuracy of the FPGA algorithm is 0.16 when the number of iterations is 0. As the number of iterations increases, the accuracy gradually improves, and finally reaches 0.81 when the number of iterations is 60. The accuracy of the ViBe algorithm is 0.09 when the number of iterations is 0. As the number of iterations increases, the accuracy improves, but the improvement is relatively small. Finally, it is 0.77 when the number of iterations is 60. The accuracy of the SAR algorithm is 0.13 when the number of iterations is 0. As the number of iterations increases, the accuracy gradually improves, and finally reaches 0.84 when the number of iterations is 60.

The accuracy of the OpenCV algorithm is 0.15 at 0 iterations, reaching the highest point of 0.81 at 50 iterations, but decreasing to 0.72 at 60 iterations. The accuracy of the DSP algorithm is 0.11 when the number of iterations is 0. As the number of iterations increases, the accuracy gradually improves, and finally reaches 0.84 when the number of iterations is 60. The accuracy of The Study algorithm is 0.21 when the number of iterations is 0. The accuracy steadily improves with the increase of iterations, and finally reaches 0.92 when the number of iterations is 60. Overall, as the number of iterations increases, the accuracy of most algorithms improves. Among them, The Study algorithm performs best at 60 iterations, achieving an accuracy of 0.92. The accuracy of other algorithms is between 0.77 and 0.86 when the number of iterations is 60. It should be noted that the OpenCV algorithm experienced a decrease in accuracy at 60 iterations, so The Study algorithm is the most stable.

D. COMPARISON OF MIOU VALUES FOR SAMPLE COLLECTION

Accurately evaluating the performance of a model on moving targets is crucial in semantic segmentation tasks. One of the commonly used indicators to evaluate the performance of a model is the average intersection to union ratio (mIoU). The mIoU value measures the accuracy of the model by calculating the intersection and union ratio between the predicted segmentation results and the true labels. During the sample collection process, selecting appropriate samples is crucial for obtaining reliable mIoU comparison results. After completing sample collection and training the model,

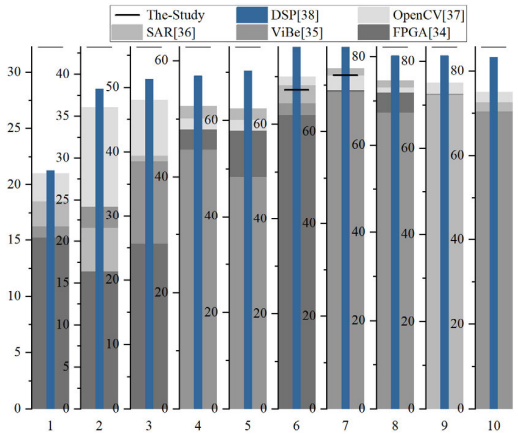

FIGURE 9. Comparison of mIoU values for sample collection.

TABLE 3. F1 values for different models.

Epoch	FPGA	ViBe	SAR	OpenCV	DSP	The-Study
0	0.72	0.79	0.73	0.76	0.65	0.8
10	0.72	0.8	0.71	0.77	0.65	0.8
20	0.71	0.75	0.72	0.74	0.68	0.86
30	0.7	0.78	0.68	0.81	0.69	0.84
40	0.71	0.79	0.71	0.74	0.73	0.79
50	0.69	0.75	0.71	0.76	0.69	0.79
60	0.68	0.76	0.69	0.76	0.69	0.79

we can evaluate the performance of the model by calculating the mIoU (mean intersection to union ratio) value. MIoU is an indicator obtained by adding and averaging the IoU values of all categories, used to evaluate the segmentation accuracy of the model in multi category scenarios. The results are shown in Figure 9.

Figure 9 shows whether the mIoU value of the algorithm shows a stable upward trend with the increase of sample size. As the sample size increases, the mIoU values of most algorithms show an increasing trend, but the speed and magnitude of growth vary. Under the same sample size, it can be observed that some algorithms perform better than others. For example, when the sample size is 10, ViBe algorithm and SAR algorithm have lower mIoU values, while FPGA algorithm and OpenCV algorithm perform relatively well. Some algorithms do not steadily improve as the sample size increases, for example, when the sample size is 6, the mIoU value of the DSP algorithm decreases. There is a certain correlation between the mIoU value of DSP algorithm and The Study algorithm. As the number of samples increases, the mIoU value of DSP algorithm also shows a gradually increasing trend. Specifically, from a sample size of 1 to 10, the mIoU value of the DSP algorithm increases from 21.26 to 83.15. When the sample size of The Study algorithm is from 1 to 4, the mIoU value also shows a gradually increasing trend. However, starting from a sample size of 5, the mIoU value of The Study algorithm fluctuates and does not always increase. Specifically, from a sample size of 1 to 10, the mIoU value of The Study algorithm increased from 32.25 to 85.66. The mIoU value of the ViBe algorithm showed a sharp decrease after the first few sample sizes increased, which may indicate that the ViBe algorithm encountered a performance bottleneck when facing large sample sizes. The mIoU value of the OpenCV algorithm remains relatively stable under changes in sample size, indicating that the performance of the algorithm is not significantly affected by sample size to some extent. Overall, the improved U-Net algorithm proposed in this study has good stability, and the applicability of other algorithms may be smaller than that of the improved U-Net algorithm. Therefore, in practical applications, it is necessary to comprehensively consider the relationship between algorithm performance and data size, and choose the algorithm that is most suitable for specific scenarios.

V. CONCLUSION

This study aims to propose an improved U-Net neural network model based on attention for semantic segmentation of moving targets. By introducing attention mechanism in the U-Net architecture, we can effectively improve the performance of the model in mobile target semantic segmentation tasks. Firstly, starting from the basic U-Net model, we will make improvements to address the potential issues that may arise when dealing with moving targets. The traditional U-Net model has shown good performance in semantic segmentation tasks, but there are certain limitations in processing moving targets. We observe that traditional U-Net models are prone to blurring or information loss in the details of moving targets, and are not sensitive to the correlation between targets. To address these issues, we have introduced an attention mechanism. By adding attention modules in the encoding decoding stage of U-Net, we can guide the model to pay more attention to the important regions of moving targets and better preserve the correlation information between targets. Through this approach, we can effectively improve the performance of the model.

The existing improved U-Net neural network models may require a large amount of computing resources and may encounter difficulties in real-time application on mobile devices. Therefore, further research is needed on how to optimize models and algorithms under limited resources in order to achieve real-time semantic segmentation on mobile devices. Existing models may have limitations when dealing with moving targets of different types, scales, and speeds. Therefore, future research can attempt to improve the model's adaptability to the diversity of moving targets by introducing multimodal information or joint learning. In semantic segmentation of moving targets, multiple sensor information such as cameras, LiDAR, infrared, etc. can be combined to further improve the accuracy and robustness of the model. In the future, attention based improved U-Net models can

be integrated into complete end-to-end automation systems to achieve comprehensive perception and understanding of moving targets.

REFERENCES

[1] C. Li, Y. Tan, W. Chen, X. Luo, Y. He, Y. Gao, and F. Li, "ANU-Net: Attention-based nested U-net to exploit full resolution features for medical image segmentation," *Comput. Graph.*, vol. 90, pp. 11–20, Aug. 2020.

[2] M. Lei, Z. Rao, H. Wang, Y. Chen, L. Zou, and H. Yu, "Maceral groups analysis of coal based on semantic segmentation of photomicrographs via the improved U-Net," *Fuel*, vol. 294, Jun. 2021, Art. no. 120475.

[3] M. R. Ahmed, A. F. Ashrafi, R. U. Ahmed, S. Shatabda, A. K. M. M. Islam, and S. Islam, "DoubleU-NetPlus: A novel attention and context-guided dual U-Net with multi-scale residual feature fusion network for semantic segmentation of medical images," *Neural Comput. Appl.*, vol. 35, no. 19, pp. 14379–14401, Jul. 2023.

[4] Z. Jian, T. Song, Z. Zhang, Z. Ai, H. Zhao, M. Tang, and K. Liu, "An improved nested U-Net network for fluorescence in situ hybridization cell image segmentation," *Sensors*, vol. 24, no. 3, p. 928, Jan. 2024.

[5] H. Yuan, T. Jin, and X. Ye, "Modification and evaluation of attention-based deep neural network for structural crack detection," *Sensors*, vol. 23, no. 14, p. 6295, Jul. 2023.

[6] N. S. Punn and S. Agarwal, "Modality specific U-Net variants for biomedical image segmentation: A survey," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5845–5889, Oct. 2022.

[7] J. Wang, X. Zhang, P. Lv, L. Zhou, and H. Wang, "EAR-U-Net: Efficientnet and attention-based residual U-Net for automatic liver segmentation in CT," 2021, *arXiv:2110.01014*.

[8] M. Yu, Z. Huang, Y. Zhu, P. Zhou, and J. Zhu, "Attention-based residual improved U-Net model for continuous blood pressure monitoring by using photoplethysmography signal," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103581.

[9] Q. Xu, Z. Ma, N. He, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106626.

[10] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[11] T.-W. Kuan, Y. Gu, T. Chen, and Y. Shen, "Attention-based U-Net extensions for complex noises of smart campus road segmentation," in *Proc. 10th Int. Conf. Orange Technol. (ICOT)*, Nov. 2022, pp. 1–4.

[12] R. Ramadan and S. Aly, "CU-Net: A new improved multi-input color U-Net model for skin lesion semantic segmentation," *IEEE Access*, vol. 10, pp. 15539–15564, 2022.

[13] K. Damkliang, P. Thongsuksai, K. Kayasut, T. Wongsirichot, C. Jitsuwan, and T. Boonpipat, "Binary semantic segmentation for detection of prostate adenocarcinoma using an ensemble with attention and residual U-Net architectures," *PeerJ Comput. Sci.*, vol. 9, p. e1767, Dec. 2023.

[14] X. Qiu, "U-Net-ASPP: U-Net based on atrous spatial pyramid pooling model for medical image segmentation in COVID-19," *J. Appl. Sci. Eng.*, vol. 25, no. 6, pp. 1167–1176, 2022.

[15] T. de la Sotta, V. Chang, B. Pizarro, H. Henriquez, N. Alvear, and J. M. Saavedra, "Impact of attention mechanisms for organ segmentation in chest X-ray images over U-Net model," *Multimedia Tools Appl.*, vol. 2023, no. 10, pp. 1–23, Oct. 2023.

[16] R. Arora, B. Raman, K. Nayyar, and R. Awasthi, "Automated skin lesion segmentation using attention-based deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 65, Mar. 2021, Art. no. 102358.

[17] H. Zhang, X. Zhong, G. Li, W. Liu, J. Liu, D. Ji, X. Li, and J. Wu, "BCU-Net: Bridging ConvNeXt and U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 159, Jun. 2023, Art. no. 106960.

[18] R. Xiao and Z. Wan, "GAEI-UNet: Global attention and elastic interaction U-Net for vessel image segmentation," 2023, *arXiv:2308.08345*.

[19] J. Fan, M. Du, L. Liu, G. Li, D. Wang, and S. Liu, "Macerals particle characteristics analysis of tar-rich coal in northern Shaanxi based on image segmentation models via the U-Net variants and image feature extraction," *Fuel*, vol. 341, Jun. 2023, Art. no. 127757.

[20] K. Ding, S. Chen, Y. Wang, Y. Liu, Y. Zeng, and J. Tian, "AAU-Net: Attention-based asymmetric U-Net for subject-sensitive hashing of remote sensing images," *Remote Sens.*, vol. 13, no. 24, p. 5109, Dec. 2021.

[21] K. Sun, Y. Chen, Y. Chao, J. Geng, and Y. Chen, "A retinal vessel segmentation method based improved U-Net model," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104574.

[22] X. Gong, L. Qingge, Q. Liu, and P. Yang, "Improved U-Net-like network for visual saliency detection based on pyramid feature attention," *Wireless Commun. Mobile Comput.*, vol. 2022, Aug. 2022, Art. no. 1108462.

[23] Z. Li, H. Zhang, Z. Li, and Z. Ren, "Residual-attention UNet++: A nested residual-attention U-Net for medical image segmentation," *Appl. Sci.*, vol. 12, no. 14, p. 7149, Jul. 2022.

[24] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A Siamese network based U-Net for change detection in high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2357–2369, 2022.

[25] I. S. Han, "Multimodal brain image analysis and survival prediction using neuromorphic attention-based neural networks," in *Proc. Brainlesion: Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries: 6th Int. Workshop, BrainLes, Held Conjunct (MICCAI)*, Lima, Peru: Springer, 2021, pp. 194–206.

[26] Z. Al-Huda, B. Peng, R. N. A. Algburi, M. A. Al-antari, R. Al-Jarazi, O. Al-maqtari, and D. Zhai, "Asymmetric dual-decoder-U-Net for pavement crack semantic segmentation," *Autom. Construct.*, vol. 156, Dec. 2023, Art. no. 105138.

[27] Y. Gulzar and S. A. Khan, "Skin lesion segmentation based on vision transformers and convolutional neural networks—A comparative study," *Appl. Sci.*, vol. 12, no. 12, p. 5990, Jun. 2022.

[28] N. Subaramani and E. Sasikala, "An attention-based dense network model for cardiac image segmentation using learning approaches," *Soft Comput.*, vol. 28, no. 1, pp. 765–775, Jan. 2024.

[29] H. Al Jowair, M. Alsulaiman, and G. Muhammad, "Multi parallel U-Net encoder network for effective polyp image segmentation," *Image Vis. Comput.*, vol. 137, Sep. 2023, Art. no. 104767.

[30] S. V. Lim, M. A. Zulkifley, A. Saleh, A. H. Saputro, and S. R. Abdani, "Attention-based semantic segmentation networks for forest applications," *Forests*, vol. 14, no. 12, p. 2437, Dec. 2023.

[31] S. Zhao, Y. Wang, and K. Tian, "Using AAEHS-Net as an attention-based auxiliary extraction and hybrid subsampled network for semantic segmentation," *Comput. Intell. Neurosci.*, vol. 2022, Oct. 2022, Art. no. 1536976.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, Sep. 2010.

[34] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in *Proc. CVPR Workshop Future Datasets Vis.*, vol. 2, 2015, p. 1.

[35] L. Zbinden, D. Catucci, Y. Suter, A. Berzigotti, L. Ebner, A. Christe, V. C. Obmann, R. Sznitman, and A. T. Huber, "Convolutional neural network for automated segmentation of the liver and its vessels on non-contrast T1 vibe Dixon acquisitions," *Sci. Rep.*, vol. 12, no. 1, p. 22059, Dec. 2022.

[36] F. Gao, T. Huang, J. Sun, J. Wang, A. Hussain, and E. Yang, "A new algorithm for SAR image target recognition based on an improved deep convolutional neural network," *Cogn. Comput.*, vol. 11, pp. 809–824, Jun. 2019.

[37] S. Kumari, L. Gupta, and P. Gupta, "Automatic license plate recognition using OpenCV and neural network," *Int. J. Comput. Sci. Trends Technol. (IJCST)*, vol. 5, no. 3, pp. 114–118, 2017.

[38] D. Wang, K. Xu, J. Guo, and S. Ghiasi, "DSP-efficient hardware acceleration of convolutional neural network inference on FPGAs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 12, pp. 4867–4880, Dec. 2020.

ZHIHAO CUI received the B.A. and M.S. degrees from Henan Normal University, China, in 2003 and 2010, respectively, and the Ph.D. degree from Wonkwang University, South Korea, in 2020. In 2009, he was a Lecturer with Pingdingshan University, China. His research interests include computing education, deep learning, and computer vision.