IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 March 2024, accepted 20 May 2024, date of publication 27 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405957

RESEARCH ARTICLE

Multi-Convolutional Channel Residual Spatial Attention U-Net for Industrial and Medical Image Segmentation

HAOYU CHEN AND KYUNGBAEK KIM, (Member, IEEE)
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Kyungbaek Kim (kyungbaekkim@jnu.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the Innovative Human Resource Development for Local Intellectualization Support Program supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2023-RS-2022-00156287 (50%); and in part by IITP under the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by Korean Government (MSIT) under Grant IITP-2023-RS-2023-00256629 (50%).

ABSTRACT Image segmentation has demonstrated immense potential in computer vision. In particular, the U-Net architecture, built on fully convolutional networks, is highly suitable for image segmentation tasks. Its encoder-decoder structure effectively captures both local and global features. This approach has achieved remarkable outcomes across various sectors, most notably in medical diagnostics and industrial quality control. However, U-Net, by employing skip connections, fuses different low-level and high-level convolutional features between the encoder and decoder, limiting its ability to effectively integrate useful features and harness contextual information. To address these feature disparities between the encoder and decoder, this paper introduces a novel network structure named Multi-Convolutional Channel Residual Spatial Attention U-Net (MCRSAU-Net). Designed for industrial and medical image segmentation, this model is anchored on the U-Net architecture. It replaces the traditional skip connections with channel attention residual paths featuring multiple convolutions, retaining more low-level features. Moreover, spatial attention module is incorporated in the decoding path to ensure the model concentrates on crucial regions of the input space, enhancing its segmentation capability across varied tasks. The proposed method was subjected to 5-fold cross-validation and testing on three public datasets: Mvtec AD, CHASE DB1, and Kvasir SEG. MCRSAU-Net achieved average Dice coefficients of 0.7755, 0.7651, and 0.8958 for defect segmentation of bottles, woods, and tiles, respectively, with average accuracies reaching 0.9751, 0.9815, and 0.9841. For retinal blood vessel and colon polyp segmentation, it exhibited superior performance, achieving average Dice scores of 0.8540 and 0.7053, and average accuracies of 0.9465 and 0.9195, respectively. These results not only underscore MCRSAU-Net's strong performance in image segmentation tasks but also demonstrate its significant potential in addressing specific challenges encountered in industrial and medical image segmentation.

INDEX TERMS Image segmentation, computer vision, deep learning, U-Net, attention.

I. INTRODUCTION

In recent years, with the continuous development of computer vision and deep learning technologies, image segmentation techniques have been widely applied in various fields, especially in the industrial and medical domains, where the use of image segmentation techniques provides novel solutions to some complex segmentation issues.

Image segmentation technology is a significant area of research within computer vision, primarily involving the categorization of pixels using semantic labels, the segmentation of individual objects, or both, to mark various object

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

types at the pixel level [1]. Particularly, the application of convolutional neural networks (CNNs) has shown significant improvements and validated effectiveness in identifying highly complex image patterns [2]. In the industrial sector, the applications of image segmentation technology mainly include product surface defect detection, autonomous driving, image synthesis, among others [3]. Especially in product surface defect detection, image segmentation technology can be employed to detect cracks, wear, and other defects on product surfaces [4], such as in the fields of steel surface defect segmentation [5], wood surface defect segmentation [6], fabric surface defect segmentation [7], and ceramic tile surface defect segmentation [8].

In the medical field, image segmentation technology has found widespread application in medical imaging analysis [9], serving to boost both the efficiency and accuracy of doctors' diagnoses. Fundus fluorescein angiography (FFA), a critical technique for assessing retinal diseases, benefits from training on annotated FFA images using various CNN networks to achieve automated standardized marking of FFA images [10]. Circulating tumor cells (CTCs), which are closely associated with the aggressiveness and metastatic potential of cancer, can be automatically identified using machine learning-based algorithms to minimize human error and enhance accuracy [11]. Endoscopic examinations play a vital role in diagnosing and treating tumor lesions, with the pyramid ORB algorithm being used to stitch endoscopic images together, addressing the issue of endoscopes often failing to provide comprehensive information in a single image [12]. In the domain of medical robotics operating on human organs, soft tissue surface feature tracking methods based on deep matching networks are employed for feature matching in medical images [13]. Mathematical methods of soft tissue modeling can simulate the surface and partial internal features of soft tissues, enabling the model to perform specific deformations for realistic simulation, thus facilitating the application of tactile perception in medical robotics operations on human organs [14]. Medical segmentation of images such as CT and MRI allows for clearer identification and quantification of various lesions [15], for instance, in breast tumor segmentation [16], lung segmentation [17], pancreas segmentation [18], kidney stone segmentation [19], spine segmentation [20], and liver segmentation [21]. Furthermore, image segmentation technology is crucial for surgical planning and intraoperative navigation [22].

Despite the significant achievements of image segmentation in industrial and medical applications, challenges persist, such as dealing with complex backgrounds, improving segmentation accuracy, and enhancing algorithm stability. To address these challenges, we optimized the famous U-Net framework in the field of image segmentation [23]. In this paper, we propose a new CNN, Multi-Convolutional Channel Residual Spatial Attention U-Net (MCRSAU-Net). It replaces U-Net's skip connections with convolution kernels of different kernel sizes and channel attention residual paths to capture features of different scales and automatically pay attention to important channels, which tends to be more effective for complex image segmentation tasks. We used spatial attention mechanisms in the decoding path because spatial information is often critical in image segmentation tasks. With spatial attention mechanisms, the network can automatically focus on important areas in space, allowing the network to pass more important features in the skip connections and reduce the influence of noise features. Additionally, the network loss function employs a binary cross-entropy loss function to handle segmentation tasks in unbalanced scenes, thereby enhancing the network's segmentation capabilities. The main contributions of this paper are as follows:

- Introduce channel attention mechanisms between multiple convolutions of various kernel sizes to adapt to features of different scales.
- Incorporate multi-convolution channel attention residual paths to achieve deeper integration of attention mechanisms.
- Systematically apply spatial attention in the decoding phase to focus on important areas and extract salient features.

The remainder of this paper is arranged as follows: Section II reviews related work. Section III introduces the proposed network. Section IV details the experimental setup. Section V evaluates the experimental results. Section VI discusses the experimental scheme, and Section VII concludes the paper.

II. RELATED WORK

A. U-NET NETWORK

With the continual evolution of CNN, significant improvements have been witnessed over traditional semantic segmentation systems. Currently, semantic segmentation tasks play a crucial role in image processing [24], [25], [26]. In 2015, Long et al. [27] were the first to introduce a Fully Convolutional Network (FCN) into the semantic segmentation network for end-to-end segmentation of natural images, marking a significant shift from traditional machine learning-based methods to deep learning approaches. In 2015, Badrinarayanan et al. [28] proposed a deep learning network called SegNet, which performed efficient feature learning and image reconstruction through an encoder-decoder architecture and max-pooling indices. Subsequently, Ronneberger et al. [23] proposed the U-Net network, which utilized convolutional layers to perform semantic segmentation tasks.

U-Net is an FCN-based network. Its structure is similar to that of FCN and SegNet, utilizing encoders and decoders as well as skip connections. This allows for more precise segmentation on a small number of training images. The U-Net network is characterized by its symmetrical network design, where the left side is an encoder for obtaining contextual information and the right side is a decoder

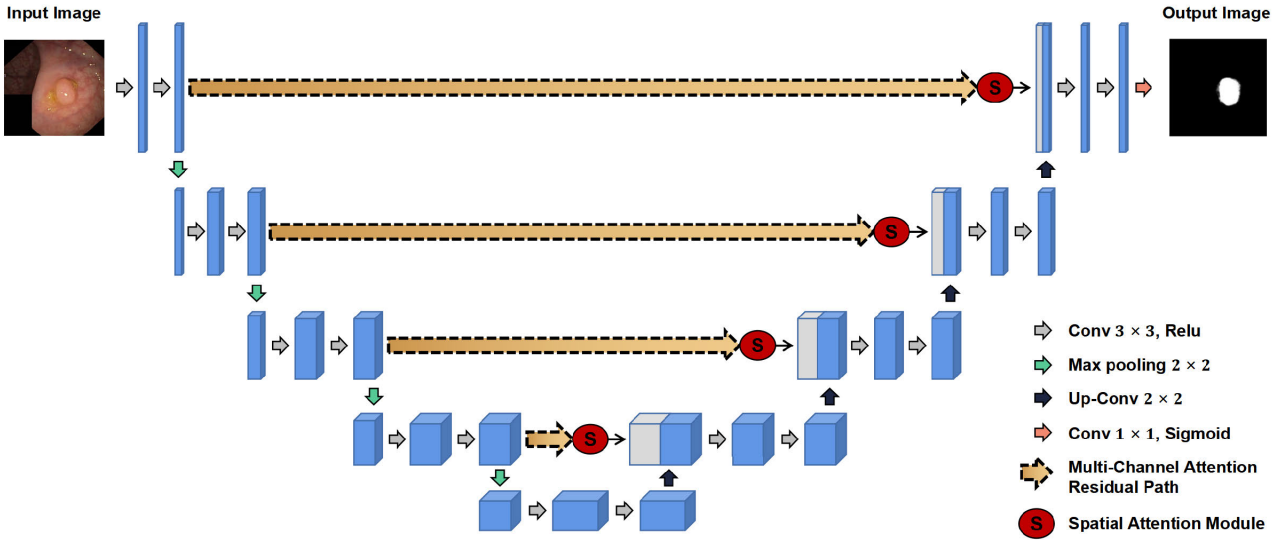

FIGURE 1. MCRSAU-Net architecture.

for precise positioning and restoring the feature map size. By duplicating and cropping the output feature maps of the encoder and fusing them with the deconvolution feature maps of the decoder, they are sent to the next layer for upsampling. During the upsampling phase of the U-Net network, numerous feature channels can pass contextual information to higher resolution levels.

Most notably, in the U-Net network, skip connections link the feature maps of each depth level of the encoder directly to the corresponding depth level of the decoder. The aim of this design is to combine low-resolution, high-level contextual information and high-resolution, low-level detailed information in the decoder section. Therefore, the U-Net network has become the subject of research for many scholars in the field of image segmentation, including Attention U-Net [29], Residual U-Net [30], SA-UNet [31], TransUNet [32] and DUCK-Net [33].

B. ATTENTION MECHANISM

In the field of computer vision, the concept of attention mechanisms has become extremely important. The concept of attention mechanisms is derived from the human visual system, in which humans utilize cognitive information to shift attention to relevant objects in the visual scene while ignoring other information [34]. The primary purpose of introducing attention mechanisms in computer vision is to guide deep learning models to focus more on specific parts of the input data, thereby processing a large amount of visual information more effectively [35], [36]. For image classification tasks, the model may need to focus on a certain object within the image to disregard background noise. In image segmentation tasks, it can help the model to reinforce features related to the target object and suppress information related to the background. This kind of attention mechanism can be divided into spatial attention and channel attention [37]. Spatial attention focuses on the spatial positions of the feature maps, giving more weight to certain parts of the image. On the other hand, channel attention focuses on the channel dimension of the feature maps, enhancing or suppressing certain features by discovering dependencies between different channels. As proposed by Jiang et al. [38], the Multi-scale Attention Convolutional Neural Network introduces Depth CNN-based, Channel, and Spatial Attention Residual Modules (CHARM) into the encoder and decoder of the U-Net network framework to enhance the ability to extract image features. However, it fails to effectively handle the differences between the encoder and decoder caused by traditional skip connections, and attention mechanism processing after multiple convolutions may lead to some important features extracted in the early convolutional layers being ignored before they are fully utilized. Furthermore, the self-attention mechanism in Transformer models has also been widely applied in computer vision, allowing the model to consider other parts of the input sequence when processing each input part, thereby endowing the model with the ability to capture global dependencies [39].

III. PROPOSED METHOD

We propose an image segmentation network, MCRSAU-Net, designed for applications in the industrial and medical domains. An overview of our proposed network is illustrated in Fig. 1. In order to achieve improved image segmentation performance, we have replaced the skip connections in the U-Net architecture with multi-convolutional channel attention residual paths. Additionally, we have incorporated spatial attention modules into various stages of the decoding path to aid the network in focusing on crucial areas of interest. Detailed descriptions of each component of the MCRSAU-Net network are presented in the following subsections.

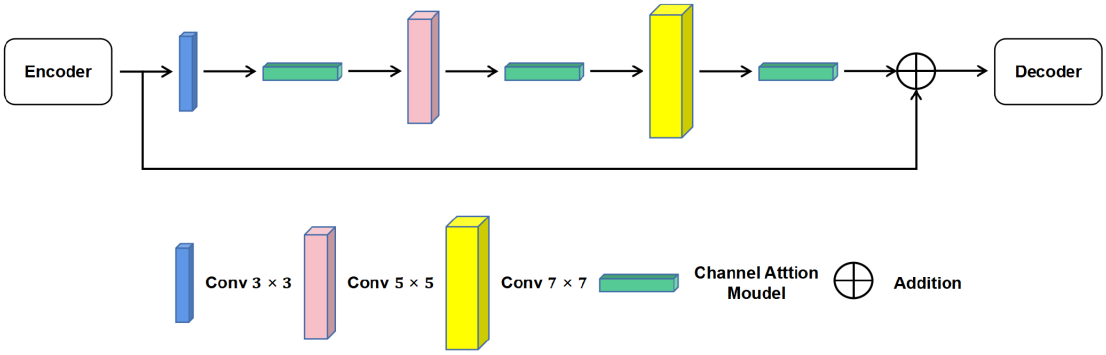

FIGURE 2. Multi-channel attention residual path architecture.

A. ENCODER

In the MCRSAU-Net, the encoder is realized according to the U-Net framework. The encoder section consists of five main stages, each associated with convolutional layers equipped with 32, 64, 128, 256, and 512 filters, respectively. Primarily, two layers of 3×3 convolutions, ReLu activation function, and 2×2 max pooling are adopted. At the onset of each stage, a convolution operation is conducted on the input, primarily aimed at extracting image features, thereby effectively capturing both the contextual information and intricate local details of the image.

B. MULTI-CHANNEL ATTENTION RESIDUAL PATH

The uniqueness of the U-Net architecture lies in the incorporation of skip connections between corresponding layers before max-pooling operations and after deconvolution operations. Before each max-pooling operation, U-Net saves the feature map of the corresponding encoder level and connects it with the feature map of the respective decoder level. The goal of this process is to retain the spatial information lost in the encoder part and pass it on to the decoder part, thereby achieving more precise segmentation results. However, a feature disparity remains between the encoder and decoder, and no existing theory proves they form the best match for feature fusion. To this end, we propose a multi-convolutional channel attention residual path to minimize the information loss between the encoder and decoder.

In the Inception architecture [40], multiple differently sized convolution operations are used in parallel to better capture the details and contextual information within an image, thereby boosting network performance. As such, we introduce 3×3, 5×5, and 7×7 convolutional kernels between the encoder and decoder to better learn features at different scales. After each convolutional kernel, we add a channel attention module to increase the weight of important channels and reduce the weight of unimportant ones, thereby facilitating automatic feature selection and enhancing feature interaction. This setup enables the extraction of spatial features at different scales and decreases the feature disparity between the encoder and decoder. Following the residual learning network proposed in [41], we utilize residual connections to alleviate the difficulty of network training and preserve more learned features. Therefore, we introduce residual connections in the path, apply Batch Normalization (BN) layers [42], followed by a non-linear Rectified Linear Unit (ReLU) activation function. The structure is illustrated in Fig. 2. Finally, according to [43], it is speculated that as skip connections continue, the difference value between features decreases. In the bottom-up skip connections, we perform 1, 2, 3, and 4 consecutive multi-convolutional channel attention connections, respectively, and use residual connections to enable the network to gain more spatial information. SE-Net [44] is an effective neural network architecture which introduces a squeeze-and-excitation module capable of adaptively calibrating the importance of each channel, thereby optimizing the inter-channel relationship. Therefore, introducing the channel attention module in the model can enhance network performance and generalization ability. The channel attention module in the multi-convolution channel attention residual path is depicted in Fig. 3. Let's assume that the input feature is x, with the shape of $h \times w \times c$, where h is the height, w is the width, and c is the number of channels. The first step is to execute a squeeze operation, performing global average pooling on the input feature to reduce the feature dimension. The execution process of global average pooling is as follows:

$$\text{avg}(c) = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} x[i, j, c]}{h \times w}. \tag{1}$$

where i, j traverse each spatial position.

Next, the features after average pooling are reshaped to fit the input dimensions for the subsequent layer, resulting in a tensor of shape $1 \times 1 \times c$. The first FC layer performs feature extraction and reduces dimensions, employing a ReLU activation function. The second FC layer restores the features to the original feature dimensions. The output dimensions of the two FC layers are c/r and c respectively, where r is the reduction ratio. Then, a sigmoid function is used to compress the feature values between 0 and 1, which can be understood as scoring the importance of each channel.

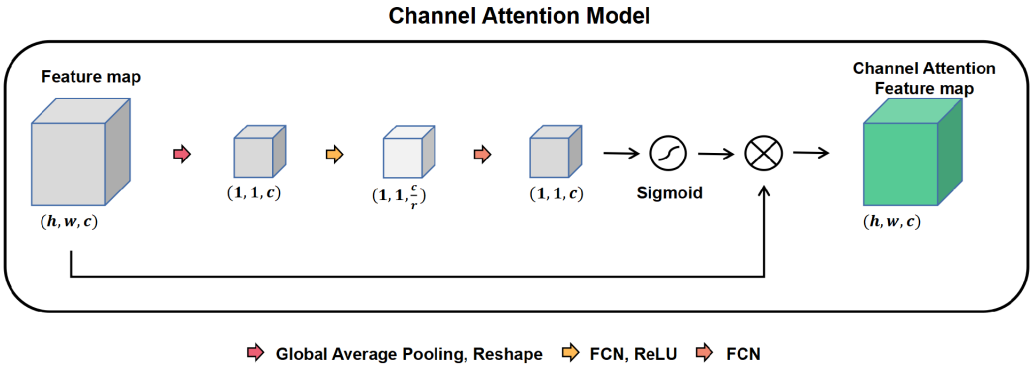

FIGURE 3. Channel attention model architecture.

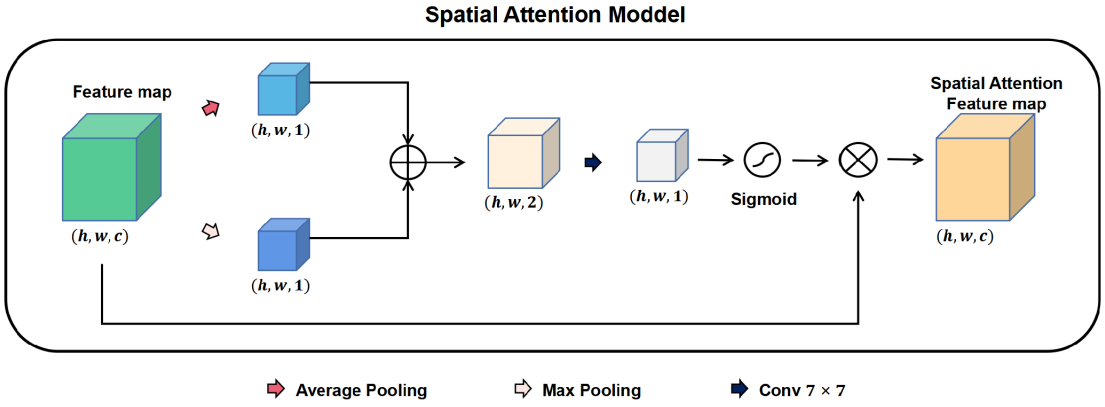

FIGURE 4. Spatial attention model architecture.

A multiplication operation follows, where the result after sigmoid activation is multiplied with the original input feature for corresponding elements. This scales each channel of the original feature with its respective weight, thus completing the channel attention operation. The final output y of this module is as follows:

$$w = \sigma\left(w_2 \cdot \mathrm{ReLU}(w_1 \cdot \mathrm{avg(c)} + b_1) + b_2\right). \tag{2}$$

$$y = \mathrm{x} \odot w. \tag{3}$$

where each channel's weight is w, σ represents the sigmoid function, and the weights of the first and second FC layers are w_1 and w_2, respectively, while b_1 and b_2 denote the bias terms.

Next, let's assume that the input to the multi-convolution channel attention residual path is I, and the 3×3, 5×5, and 7×7 convolutions are denoted by k_3, k_5, and k_7 respectively. The final output O of the multi-convolution channel attention residual path is as follows:

$$O_1 = CA(\mathrm{ReLU}(k_3(I))). \tag{4}$$

$$O_2 = CA(\mathrm{ReLU}(k_5(O_1))). \tag{5}$$

$$O_3 = CA(\mathrm{ReLU}(k_7(O_2))). \tag{6}$$

$$O = BN(ReLU(I + O_3)). \tag{7}$$

where CA represents channel attention, $\odot$ denotes element-wise multiplication, and O_1, O_2, and O_3 are the outputs of the 3×3, 5×5, and 7×7 convolutions with channel attention, respectively.

C. DECODER

In the MCRSAU-Net, the decoder is augmented with a spatial attention module on the basis of the U-Net framework, as depicted in Fig. 4. The decoder part also consists of five stages, each of which corresponds to convolutional layers with 32, 64, 128, 256, and 512 filters, respectively. Predominantly, it employs two layers of 3×3 convolution, ReLu activation functions, and 2×2 up-convolution, with the final step involving a 1×1 convolution and a sigmoid activation function for pixel-level classification. The spatial attention module is appended after the output from each multi-convolution channel attention residual path, and before the convolutional operations at each stage of the encoder. This design allows the model to consider all features globally and to focus on those important features beneficial for the current task, thereby enhancing the model's performance in tasks such as image segmentation.

Suppose the input features to the spatial attention module are denoted as e, having a shape of $h \times w \times c$, where h is

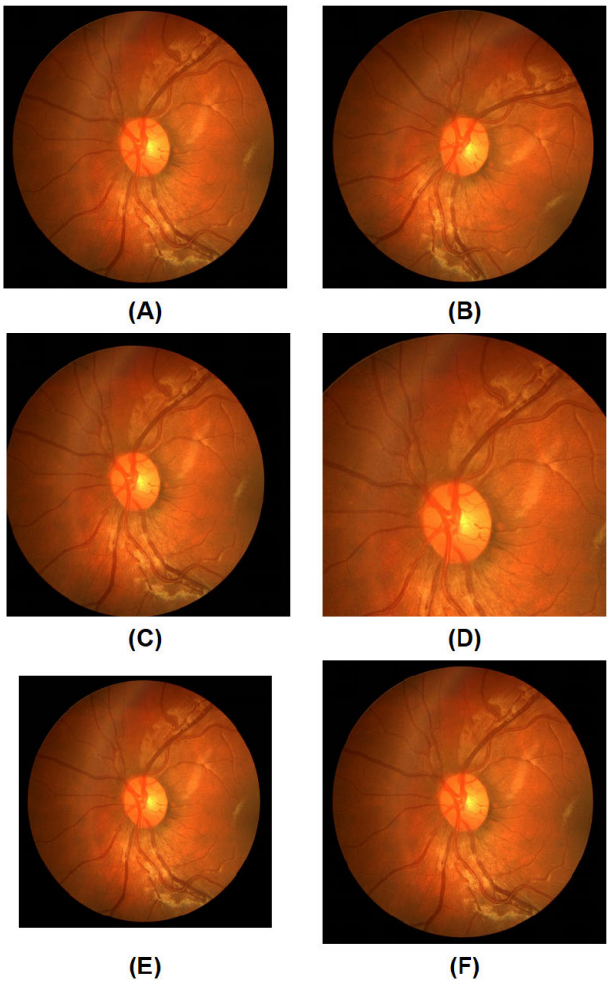

FIGURE 5. Example of image augmentation for the CHASE DB1 dataset. Here, A represents the original image, B is after rotation, C is after translation, D is after cropping, E is after scaling, and F is after flipping.

the height, w is the width, and c is the number of channels. Firstly, average pooling and max pooling are performed on the channel dimension of the input features, which are used to preserve uniform information and capture the most prominent features in the spatial structure, respectively. These operations result in two feature maps of the same size, each with a channel count of 1 The process is outlined as follows:

$$e_{\text{avg}}[i,j] = \frac{1}{c}\sum_{c} X[i,j,c]. \tag{8}$$

$$e_{\max}[i,j] = \max_{c} X[i,j,c]. \tag{9}$$

where i, j traverse each spatial position.

The two pooling results are then concatenated on the channel dimension and serve as the input to the convolutional layer. In this way, the model can consider both the uniform and the salient information of the spatial structure concurrently. Subsequently, a 7×7 convolution is used to obtain the weights for each position. Following this, a sigmoid function generates an attention weight map for each spatial position. This weight map re-weights the input features, emphasizing important spatial regions and suppressing unimportant ones. This weight map is then multiplied element-wise with the original input feature map. The final output H is then expressed as follows:

$$w_s = \sigma(k_7(\text{Concatenate}(e_{\text{avg}}, e_{\max}))). \tag{10}$$

$$H = e \odot w_s. \tag{11}$$

where w_s represents the weight for each spatial position, k_7 denotes a 7×7 convolution, σ stands for the sigmoid function, and $\odot$ signifies element-wise multiplication.

IV. EXPERIMENTAL SETTINGS

In this section, we describe the three different types of datasets used in our experiments, the preprocessing techniques for segmentation tasks, the optimizer and loss function used for training, and the k-fold cross-validation method.

A. DATASETS

In this experiment, we evaluated three publicly available datasets from the industrial and medical fields: MVTec AD [45], CHASE DB1 [46], and Kvasir SEG [47]. Specific descriptions are as follows:

1) MVTEC AD

The MVTec AD (Anomaly Detection) dataset is an industrial image dataset for anomaly detection, released by MVTec Software GmbH in 2019. This dataset contains 5354 images of different types of industrial products (including bottles, cables, capsules, wood, etc.) across 15 categories. Each product type includes a set of normal samples and a set of abnormal samples, with over 70 different types of defects. In our experiments, we used images of abnormal samples from the "bottle," "wood," and "tile" categories for network segmentation training.

The release of the MVTec AD dataset has significantly advanced research in the fields of anomaly detection and image segmentation, becoming one of the most widely used public datasets for industrial anomaly detection. Automated defect recognition in manufacturing and quality supervision is crucial. Traditional manual inspection methods are not only time-consuming but also inefficient, especially when dealing with large-scale product inspections. The MVTec AD dataset brings a diverse collection of defect images, providing valuable resources for developing and testing image segmentation technologies capable of automatically detecting and locating defects, addressing a significant challenge in practical applications. This progress is particularly critical for automated quality control, as the accuracy of image segmentation directly affects the efficiency and reliability of defect detection.

2) CHASE DB1

CHASE DB1 is a lightweight image dataset for retinal vessel segmentation. It consists of 28 color retinal images

from 14 school-aged children, with each image measuring 999 × 960 pixels. Each image has been annotated by two independent human experts.

Accurate analysis of retinal vessels is essential for diagnosing and monitoring various eye diseases such as diabetic retinopathy, glaucoma, and hypertension. The high-definition retinal images contained in the CHASE DB1 dataset lay the foundation for developing algorithms capable of automatically performing such medical analyses, playing a crucial role in improving the accuracy and efficiency of early disease diagnosis. In the field of image segmentation, given the complexity of retinal vessel structures, effective segmentation algorithms need to handle various vessel widths and branching structures while minimizing misidentification of non-vessel structures. The CHASE DB1 dataset serves as a benchmark to assess and compare the performance of different algorithms.

3) KVASIR SEG

The Kvasir SEG dataset primarily consists of 1000 high-quality gastroenterological endoscopic images, with sizes ranging from 720 × 576 to 1920 × 1080 pixels. Each image is accompanied by pixel-level annotations from a certified radiologist.

Endoscopic detection is crucial for diagnosing and evaluating gastrointestinal diseases, and the Kvasir-SEG dataset provides researchers with a rich inventory of high-quality endoscopic images marked by experts, serving as an ideal place for developing and testing innovative image segmentation techniques. The dataset includes images of various pathological states and degrees of lesions, facilitating the development of algorithms capable of dealing with diverse pathological features. Image segmentation models trained using the Kvasir-SEG dataset can assist doctors in more accurately identifying lesion areas, playing a key role in advancing the application of deep learning technologies in the field of medical image analysis.

B. PREPROCESSING AND TRAINING METHODS

We split the original 28 images from the CHASE DB1 dataset into 19 images for training and validation, and 9 images for testing and visual comparison. Given the small number of images for training and validation, we augmented the 19 images used for training and validation to produce 285 enhanced images, including rotations, width and height shifts, cropping, scaling, and horizontal flipping, as shown in Fig. 5. Additionally, we used 900 images from the Kvasir SEG dataset for training and validation, and 100 images for testing and visual comparison. Within the MVTec AD dataset, 56 images from the bottle category were used for training and validation, and 7 images for testing and visual comparison; 54 images from the wood category for training and validation, and 6 images for testing and visual comparison; 75 images from the tile category for training and validation, and 9 images for testing and visual comparison. Finally, when training the network on the MVTec AD, CHASE DB1, and Kvasir SEG datasets, the size of the input images (including the true labels) was adjusted to 224 × 224 pixels and normalized.

As the loss function, we used binary cross-entropy, as its performance in binary semantic segmentation tasks has been proven to be effective, as shown below:

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]. \quad (12)$$

where N represents the number of samples, y_i is the actual label of the i-th sample (0 or 1), and $\hat{y}_i$ is the model's predicted probability of the i-th sample being positive.

The MCRSA-UNet was trained using the Adam optimizer [48] on these three datasets, with an initial learning rate set at 0.001. The training periods for the MVTec AD, CHASE DB1, and Kvasir SEG datasets were 200, 50, and 150, respectively. The batch sizes for the MVTec AD, CHASE DB1, and Kvasir SEG datasets were set to 8. Generally, the network trains faster with smaller batch sizes, as weights are updated after each propagation. All networks were implemented in Keras TensorFlow and trained on 2 NVIDIA A100 SXM4 GPUs equipped with 40GB of memory each.

C. K-FOLD CROSS-VALIDATION METHOD

We employed a 5-fold cross-validation method to evaluate the performance of the networks used in our study. We divided the datasets into five parts, using four parts for training and the remaining one for validation in turn. For the MVTec AD dataset, each time 45 images from the bottle category were used for training, and 11 images for validation; 43 images from the wood category for training, and 11 images for validation; 60 images from the tile category for training, and 15 images for validation. For the CHASE DB1 dataset, each time 228 images were used for training, and 57 images for validation. For the Kvasir SEG dataset, each time 720 images were used for training, and 180 images for validation. This process was repeated five times, with a different part chosen as the validation set each time, to ensure that each data point had an opportunity to serve as validation data. At the end of each fold, the network's performance on the validation set was evaluated and the scores were stored. Finally, the average score across all folds was calculated. By evaluating the model multiple times on different training and validation sets, a more accurate estimate of the model's generalization ability is provided. This helps to reduce bias in the evaluation results due to different ways of data splitting. In situations with limited data, 5-fold cross-validation can make more efficient use of available data, as each portion is used for both training and validation, which is more comprehensive than simple train-test splits. Additionally, we used the dice metric during the validation phase and chose the set of model weights with the highest dice score as the final trained model for testing images.

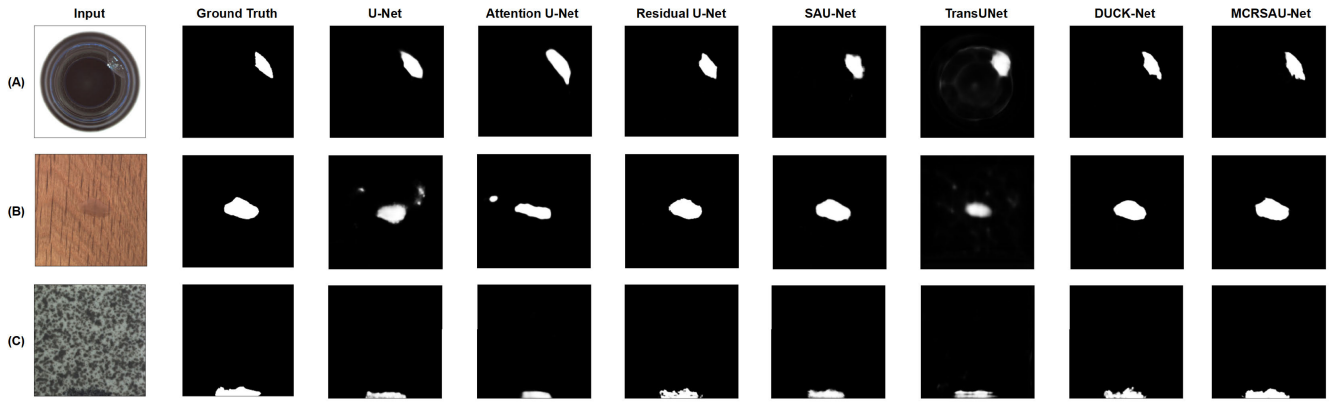

FIGURE 6. The segmentation results of MCRSAU-Net and other networks on selected test images across three categories: bottles, wood, and tiles are presented. Herein, A refers to bottles, B to wood, and C to tiles.

TABLE 1. Comparison of average AC, PR, RE,and F1 between MCRSAU-Net and state-of-the-art methods within three datasets (bottle, wood, and tile) using 5-fold cross-validation. The best performance is highlighted in bold.

Category	Bottle				Wood				Tile			
	AC	PR	RE	F1	AC	PR	RE	F1	AC	PR	RE	F1
U-Net [23]	0.9497	0.8261	0.5248	0.5879	0.9602	0.9593	0.1290	0.1975	0.9712	0.9352	0.7551	0.8229
Attention U-Net [29]	0.9537	0.7706	0.6300	0.6904	0.9612	0.5798	0.3612	0.4063	0.9482	0.8309	0.6708	0.7057
Residual U-Net [30]	0.9738	0.8983	0.7750	0.8295	**0.9821**	**0.9223**	0.7140	**0.8011**	0.9750	0.9023	0.8282	0.8625
SA-UNet [31]	0.9636	0.7541	**0.8452**	0.7936	0.9647	0.6681	0.7099	0.6632	0.9349	0.7331	0.8687	0.7728
TransUNet [32]	0.8582	0.0635	0.0796	0.0664	0.9162	0.5425	0.2345	0.2079	0.9425	0.7747	0.7127	0.7277
DUCK-Net [33]	0.9302	0.7435	0.7459	0.7225	0.9817	0.9162	0.7119	0.7975	0.9767	0.9380	0.8140	0.8627
SU-Net	0.9586	0.6729	0.6445	0.6519	0.9600	0.7602	0.1561	0.2227	0.9690	0.9470	0.7193	0.7994
MCRU-Net	0.9697	0.8472	0.7771	0.8025	0.9806	0.8965	0.7207	0.7898	0.8535	0.7916	0.8854	0.7784
MCRSAU-Net	**0.9751**	**0.8983**	0.7895	**0.8387**	0.9815	0.8984	**0.7281**	0.7972	**0.9841**	**0.9537**	**0.8987**	**0.9247**

V. EVALUATION AND RESULTS

A. EVALUATION METHODS

To evaluate the performance of the network, we utilized the following four assessment metrics, namely Accuracy (AC), F1-Score, Precision (PR), Recall (RE) and Dice Coefficient (DC). Below are the definitions of four terms used to compute these metrics.

- True Positive (TP): The sample labels that are correctly identified by the network as positive.
- True Negative (TN): The sample labels that are correctly identified by the network as negative.
- False Positive (FP): The negative sample labels that are incorrectly identified by the network as positive.
- False Negative (FN): The positive sample labels that are incorrectly identified by the network as negative.

AC is the proportion of samples correctly predicted over the total number of samples.

$$\text{AC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{13}$$

Precision (PR) is used to measure the proportion of samples that are correctly predicted as positive out of all the samples predicted as positive.

$$\text{PR} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{14}$$

Recall (RE) is used to measure the proportion of actual positive samples that are correctly predicted as positive out of all the actual positive samples.

$$\text{RE} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

F1-Score is the harmonic mean of precision and recall. Precision refers to the proportion of actual positive cases among the predicted positive ones, while recall refers to the proportion of positive cases correctly predicted out of all actual positive cases.

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{16}$$

DC is computed as twice the size of the intersection of the predicted and ground truth segmentation, divided by the sum of the sizes of the predicted and ground truth segmentation.

$$\text{DC} = \frac{2 \times |\text{PR} \cap \text{GroundTruth}|}{|\text{PR}| + |\text{GroundTruth}|}. \tag{17}$$

B. INDUSTRIAL PART DEFECT SEGMENTATION

In industrial part defect segmentation, we compared MCRSAU-Net against six U-Net and its variants: U-Net [23], Attention U-Net [29], Residual U-Net [30], SAU-Net [31], TransUNet [32], and DUCK-Net [33] using datasets from three categories (bottle, wood, and tile) within the MVTec AD dataset, where the number of filters in the convolutional

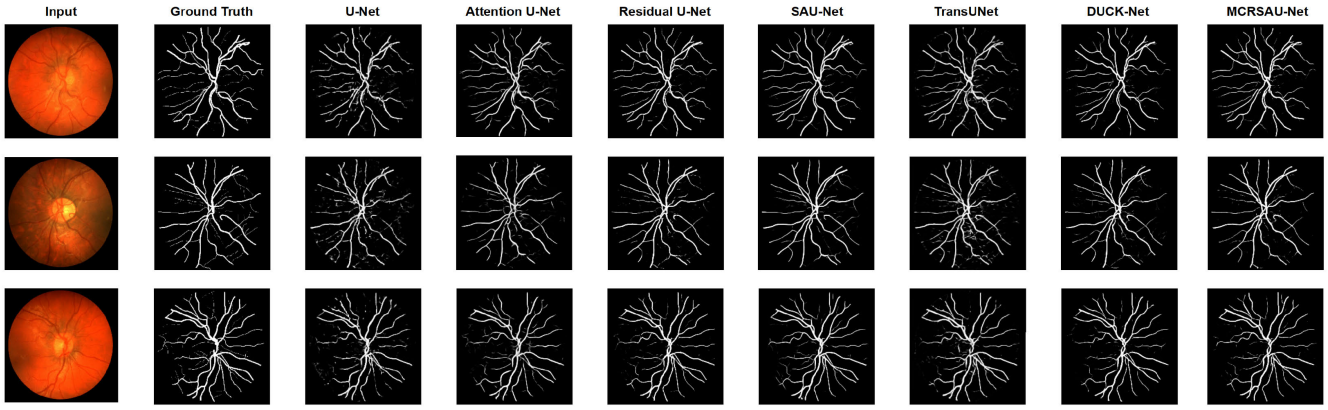

FIGURE 7. The segmentation results of MRSAU-Net and other networks on a subset of test images from the CHASE DB1 dataset.

TABLE 2. Comparison of average DC between MCRSAU-Net and state-of-the-art methods within three datasets (bottle, wood, and tile) using 5-fold cross-validation. The best performance is highlighted in bold.

Method	DC		
	Bottle	Wood	Tile
U-Net [23]	0.5733	0.2124	0.7012
Attention U-Net [29]	0.6305	0.4434	0.5575
Residual U-Net [30]	0.7644	**0.7697**	0.8192
SAU-Net [31]	0.7052	0.6258	0.7136
TransUNet [32]	0.0781	0.1381	0.5744
DUCK-Net [33]	0.6550	0.7304	0.8064
SU-Net	0.5984	0.1868	0.7059
MCRU-Net	0.7200	0.7510	0.7523
MCRSAU-Net	**0.7755**	0.7651	**0.8958**

layers are 32, 64, 128, 256, and 512 respectively. We also included comparisons with U-Net variants featuring a spatial attention with multi-convolutional channel residual path (SU-Net) and a multi-convolutional channel residual path without spatial attention (MCRU-Net). All methods were trained and validated using 5-fold cross-validation and the best network set was used for visual testing on test images, with the average score across all folds calculated to evaluate the networks, ensuring the reliability and generality of the results.

Results, as shown in Table.1 and Table.2, indicate that MCRSAU-Net surpasses its competitors in all aspects except the RE metric in the bottle and tile categories. However, in the wood category, its metrics are slightly below those of the Residual U-Net [30], suggesting a need for deeper network layers or adjusted attention mechanisms to better address the complexity and details of wood textures. In the bottle dataset, the average AC, PR, RE, F1, and DC values reached as high as 0.9751, 0.8983, 0.7895, 0.8387, and 0.7755 respectively. For the tile dataset, they were 0.9841, 0.9537, 0.8987, 0.9247, and 0.8958, and for the wood dataset, they were 0.9815, 0.8984, 0.7281, and 0.7972 respectively.

The visualization of segmentation results in Fig.5 further confirms the exceptional capability of MCRSAU-Net in accurately identifying and segmenting industrial part defects. Especially in bottle defect segmentation tasks, MCRSAU-Net not only accurately marks the defect areas but also demonstrates higher segmentation precision and lower misclassification rates compared to other models. Moreover, in tackling defect segmentation tasks for wood and tile components, MCRSAU-Net's results closely mirror the actual conditions, thereby proving the network's robust capability in segmenting anomalous areas in industrial images.

C. RETINAL VESSEL SEGMENTATION

In the context of pediatric retinal vessel segmentation, we compared the performance of MCRSAU-Net with SU-Net, MCRU-Net, and other U-Net variants using the CHASE DB1 dataset enhanced through data augmentation. All compared methods were trained and validated using 5-fold cross-validation, and the best network set was used for visual testing on test images, with the average score across all folds calculated to evaluate the networks, ensuring the reliability and generality of the results. Results, as shown in Table.3, demonstrate that MCRSAU-Net's performance in pediatric retinal vessel segmentation, in terms of average AC, PR, RE, F1, and DC, were 0.9465, 0.9925, 0.7021, 0.8222, and 0.8540 respectively, slightly below MCRU-Net in RE and F1. MCRSAU-Net significantly outperforms other comparison networks. Results of MCRSAU-Net's pediatric retinal vessel segmentation are displayed in Fig.6. The segmented images reveal our network's precision in segmenting ocular vessels more accurately than other networks. Importantly, compared to other networks, ours performs better in segmenting the ends of retinal vessels, recognizing and segmenting more peripheral branches of retinal vessels, a feat unmatched by other networks. Precise segmentation of retinal vessel terminations holds significant clinical importance in the diagnosis and treatment of eye diseases.

D. COLON POLYP SEGMENTATION

In this study, for the colon polyp segmentation task, we used the widely cited Kvasir-SEG dataset to compare the comprehensive performance of MCRSAU-Net and its variants,

TABLE 3. Comparison of average AC, PR, RE, F1, and DC between MCRSAU-Net and state-of-the-art methods within the CHASE DB1 dataset using 5-fold cross-validation. The best performance is highlighted in bold.

Method	AC	PR	RE	F1	DC
U-Net [23]	0.9420	0.9609	0.6460	0.7725	0.7507
Attention U-Net [29]	0.9401	0.9494	0.6172	0.7479	0.7296
Residual U-Net [30]	0.9444	0.9798	0.6580	0.7872	0.8179
SAU-Net [31]	0.9451	0.9811	0.6977	0.8153	0.8101
TransUNet [32]	0.9323	0.9260	0.4845	0.6343	0.5874
DUCK-Net [33]	0.9457	0.9889	0.6772	0.8039	0.8250
SU-Net	0.9436	0.9721	0.6575	0.7844	0.7850
MCRU-Net	0.9465	0.9924	**0.7024**	**0.8226**	0.8525
MCRSAU-Net	**0.9465**	**0.9925**	0.7021	0.8222	**0.8540**

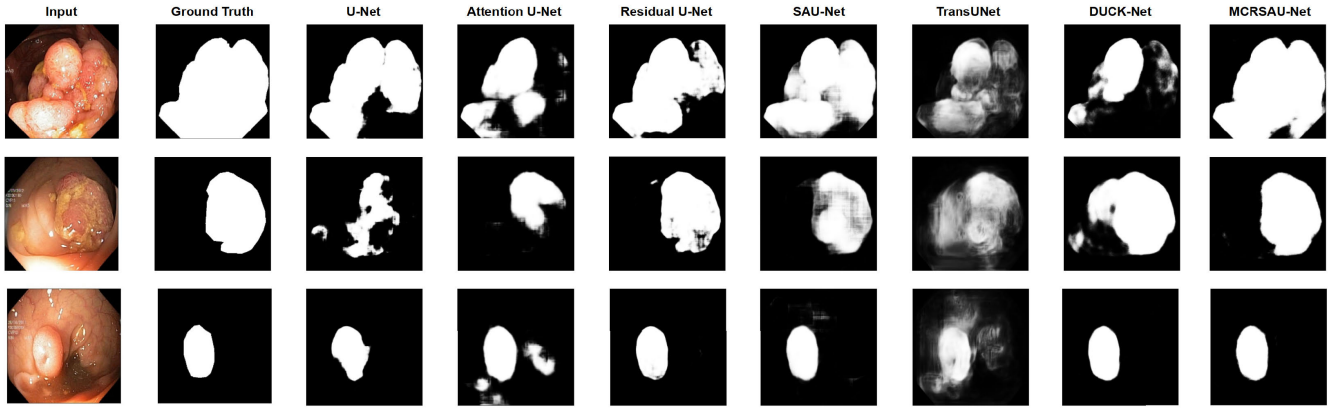

FIGURE 8. The segmentation results of MRSAU-Net and other networks on selected test images from the Kvasir SEG dataset are presented.

TABLE 4. Comparison of average AC, PR, RE, F1 and DC between MCRSAU-Net and state-of-the-art methods within the Kvasir SEG dataset using 5-fold cross-validation. The best performance is highlighted in bold.

Method	AC	PR	RE	F1	DC
U-Net [23]	0.8785	0.6630	0.5282	0.5853	0.5138
Attention U-Net [29]	0.9032	0.7830	0.5673	0.6566	0.6173
Residual U-Net [30]	0.9329	0.8825	0.7010	0.7810	0.7742
SAU-Net [31]	0.9156	0.8594	0.6213	0.7063	0.6864
TransUNet [32]	0.8103	0.4812	0.6600	0.5250	0.4198
DUCK-Net [33]	**0.9406**	0.8853	**0.7565**	**0.8156**	**0.8173**
SU-Net	0.8944	0.7184	0.5918	0.6481	0.6263
MCRU-Net	0.9380	**0.8935**	0.7288	0.8018	0.7958
MCRSAU-Net	0.9195	0.7756	0.7365	0.7549	0.7053

including U-Net, SU-Net, MCRU-Net, and other U-Net variants. To ensure the accuracy and broad applicability of the evaluation, all models were trained and validated through 5-fold cross-validation, ensuring the robustness of the results. Comparative results and visualization of segmentation effects are summarized in Table.4 and Fig.8. MCRSAU-Net's average performance metrics for colon polyp segmentation, including Accuracy (AC), Precision (PR), Recall (RE), F1 score, and Dice Coefficient (DC), reached 0.9195, 0.7756, 0.7365, 0.7549, and 0.7053 respectively. Compared to other competitive networks in the field, although MCRSAU-Net's overall performance was not universally superior, under specific optimal network configurations, its segmentation accuracy for colon polyps was noticeably better than the comparison group. A detailed analysis of the visualization of segmented test images revealed MCRSAU-Net's exceptional proficiency in precisely delineating the contours of individual and multiple polyps. This finding not only reflects our model's advantage in capturing details but also highlights its application potential in complex medical image processing tasks.

E. COMPUTATIONAL COMPLEXITY

In this study, aside from focusing on the model's segmentation performance, inference speed, and model size were also considered crucial metrics for assessing the model's suitability for practical applications (such as industrial inspection and clinical diagnosis). To comprehensively quantify computational complexity, we employed Gigaflops (GFLOPs) and inference speed (i.e., the time required by the model to process a single image) as primary evaluation metrics. Specific computational complexity results are summarized

TABLE 5. Comparison of computational complexity.

Method	U-Net [23]	Attention U-Net [29]	Residual U-Net [30]	SAU-Net [31]	TransUNet [32]	DUCK-Net [33]	MCRSAU-Net
GFLOPs	18.50	19.34	19.59	14.93	55.97	33.13	101.35
Inference speed (ms)	8.39	9.90	8.88	8.27	13.93	15.04	13.64

in Table.5. The design of MCRSAU-Net incorporates multi-convolutional channel attention residual paths, an innovation that significantly enhances segmentation accuracy but also increases the model's computational burden. Specifically, due to the numerous multi-convolutional channel attention residual modules employed, there was a significant increase in the overall computational complexity of the model. In GFLOPs assessment, our network ranked 7th, reflecting a relatively higher computational demand. In terms of inference speed, our model ranked 5th in the comparison, indicating that despite the model's complexity, its processing speed remains within a reasonable range. This balance reflects our consideration of the trade-off between efficiency and accuracy in model design.

F. ABLATION STUDIES

To deeply understand the role of each key component in MCRSAU-Net, we added comparison networks, SU-Net and MCRU-Net, for ablation studies in the MVTec AD, CHASE DB1, and Kvasir SEG datasets. The evaluation and comparison results in Table.1, Table.2, and Table.3 indicate that adding spatial attention modules and multi-convolution channel attention residual paths on top of U-Net can enhance performance, especially with the addition of multi-convolution channel attention residual paths, which significantly improve performance. MCRSAU-Net, which combines spatial attention modules with multi-convolution channel attention residual paths, shows particularly outstanding performance in the MVTec AD and CHASE DB1 datasets. However, the situation is different in the Kvasir SEG dataset. We found that compared to MCRU-Net, which only incorporates multi-convolution channel attention residual paths, the performance of MCRSAU-Net actually decreased. This demonstrates that adding channel attention modules after consecutive convolutions through multi-convolution channel attention residual paths can increase the weight of key channels while reducing the influence of secondary channels, allowing for timely adjustments of the importance of each convolution layer's channel output, focusing the network more on the most important parts and useful feature channels at every stage. Although integrating spatial attention mechanisms throughout the entire decoding phase of the network did not result in as significant performance improvements as expected, it indeed enhanced the model's ability to focus on important features under certain conditions.

VI. DISCUSSION

In this study, we introduce the MCRSAU-Net network, a novel architecture that significantly improves the performance of image segmentation tasks across several applications, including industrial part defects, retinal vessels, and colon polyps. This advancement comes through the incorporation of multi-convolutional channel attention residual paths and spatial attention modules. Despite MCRSAU-Net's exceptional performance, its application in medical contexts, particularly in decision-making diagnoses or treatments, warrants careful consideration. The potential disparities between model-generated and actual medical images underscore the importance of caution to maintain the accuracy and reliability of clinical applications. Therefore, it is imperative to underscore that medical professionals should corroborate medical decisions based on image segmentation results from MCRSAU-Net, ensuring decisions are informed by comprehensive medical knowledge and clinical experience.

While the network demonstrates superior capabilities in capturing intricate details and accurately delineating target edges, it is not without its limitations. The increased computational complexity and dependency on specific datasets highlight potential constraints on the technology and datasets, potentially limiting the model's applicability and evaluation of its generalization capabilities in resource-constrained environments. Future research directions focus on optimizing the network structure to alleviate computational demands, broadening the model's application across a wider array of image segmentation scenarios, and addressing challenges related to performance insufficiency and universality. Furthermore, enhancements to the spatial attention module in MCRSAU-Net aim to fortify its role in improving segmentation accuracy by better focusing on key features, given its variable effectiveness across different datasets.

To conclude, MCRSAU-Net offers significant contributions toward advancing image segmentation technology and expanding its applications. Future endeavors not only pursue technological enhancements and optimizations but also carefully consider the ethical application of these advancements in real-world medical scenarios. Ensuring the responsible deployment of artificial intelligence in healthcare is paramount to safeguarding patient safety and improving the quality of medical services. Additionally, ongoing research delves into refining attention mechanisms and customizing their application to various segmentation tasks, aiming to maximize their efficacy in diverse scenarios and further investigate the broad application prospects of MCRSAU-Net in the medical field.

VII. CONCLUSION

In this paper, we introduce an innovative U-Net network architecture designed to bridge the feature disparity between

the encoder and decoder. This is achieved by integrating residual paths equipped with convolutional kernels of varying sizes and channel attention, enabling the network to retain a richer set of feature information. Furthermore, we incorporate a spatial attention module within the decoder, allowing the network to focus more intently on distinct spatial location features within the input feature map. This not only amplifies the network's semantic comprehension of the image but also bolsters its performance in tasks like image segmentation. Our experiments on three publicly available datasets—MVTec AD, CHASE DB1, and Kvasir SEG—attest to our network's robust feature extraction prowess across an array of intricate images, effectively segmenting images in both industrial and medical contexts. Moving forward, we aspire to refine our network to diminish its parameters, all the while preserving its superior performance. Additionally, we intend to train our network on a broader spectrum of images from varied domains to craft a more universally adept image segmentation network.

ACKNOWLEDGMENT

The authors appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

REFERENCES

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[3] J. Yang, C. Wang, B. Jiang, H. Song, and Q. Meng, "Visual perception enabled industry intelligence: State of the art, challenges and prospects," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2204–2219, Mar. 2021.

[4] Y. Gao, X. Li, X. V. Wang, L. Wang, and L. Gao, "A review on recent advances in vision-based defect recognition towards industrial intelligence," *J. Manuf. Syst.*, vol. 62, pp. 753–766, Jan. 2022.

[5] Z. Huang, J. Wu, and F. Xie, "Automatic surface defect segmentation for hot-rolled steel strip using depth-wise separable U-shape network," *Mater. Lett.*, vol. 301, Oct. 2021, Art. no. 130271.

[6] T. He, Y. Liu, C. Xu, X. Zhou, Z. Hu, and J. Fan, "A fully convolutional neural network for wood defect location and identification," *IEEE Access*, vol. 7, pp. 123453–123462, 2019.

[7] J. Liu, C. Wang, H. Su, B. Du, and D. Tao, "Multistage GAN for fabric defect detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3388–3400, 2020.

[8] G. Dong, S. Sun, N. Wu, X. Chen, P. Huang, and Z. Wang, "A rapid detection method for the surface defects of mosaic ceramic tiles," *Ceram. Int.*, vol. 48, no. 11, pp. 15462–15469, Jun. 2022.

[9] T. Fernando, S. Denman, D. Ahmedt-Aristizabal, S. Sridharan, K. R. Laurens, P. Johnston, and C. Fookes, "Neural memory plasticity for medical anomaly detection," *Neural Netw.*, vol. 127, pp. 67–81, Jul. 2020.

[10] Z. Gao, X. Pan, J. Shao, X. Jiang, Z. Su, K. Jin, and J. Ye, "Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning," *Brit. J. Ophthalmol.*, vol. 107, no. 12, pp. 1852–1858, Dec. 2023.

[11] B. He, Q. Lu, J. Lang, H. Yu, C. Peng, P. Bing, S. Li, Q. Zhou, Y. Liang, and G. Tian, "A new method for CTC images recognition based on machine learning," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 897, Aug. 2020.

[12] Z. Zhang, L. Wang, W. Zheng, L. Yin, R. Hu, and B. Yang, "Endoscope image mosaic based on pyramid ORB," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103261.

[13] S. Lu, S. Liu, P. Hou, B. Yang, M. Liu, L. Yin, and W. Zheng, "Soft tissue feature tracking based on deep matching network," *Comput. Model. Eng. Sci.*, vol. 136, no. 1, pp. 363–379, 2023.

[14] Y. Tang, S. Liu, Y. Deng, Y. Zhang, L. Yin, and W. Zheng, "An improved method for soft tissue modeling," *Biomed. Signal Process. Control*, vol. 65, Mar. 2021, Art. no. 102367.

[15] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, and P. Szczuko, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Inf. Fusion*, vol. 90, pp. 316–352, Feb. 2023.

[16] Y. Yan, Y. Liu, Y. Wu, H. Zhang, Y. Zhang, and L. Meng, "Accurate segmentation of breast tumors using AE U-Net with HDC model in ultrasound images," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103299.

[17] K.-B. Chen, Y. Xuan, A.-J. Lin, and S.-H. Guo, "Lung computed tomography image segmentation based on U-Net network fused with dilated convolution," *Comput. Methods Programs Biomed.*, vol. 207, Aug. 2021, Art. no. 106170.

[18] F. Li, W. Li, Y. Shu, S. Qin, B. Xiao, and Z. Zhan, "Multiscale receptive field based on residual network for pancreas segmentation in CT images," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101828.

[19] A. Nithya, A. Appathurai, N. Venkatadri, D. R. Ramji, and C. Anna Palagan, "Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images," *Measurement*, vol. 149, Jan. 2020, Art. no. 106952.

[20] S. F. Qadri, H. Lin, L. Shen, M. Ahmad, S. Qadri, S. Khan, M. Khan, S. S. Zareen, M. A. Akbar, M. B. Bin Heyat, and S. Qamar, "CT-based automatic spine segmentation using patch-based deep learning," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–14, Mar. 2023.

[21] M. Ahmad, S. F. Qadri, M. U. Ashraf, K. Subhi, S. Khan, S. S. Zareen, and S. Qadri, "Efficient liver segmentation from computed tomography images using deep learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, May 2022.

[22] J. Wallner, M. Schwaiger, K. Hochegger, C. Gsaxner, W. Zemann, and J. Egger, "A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery," *Comput. Methods Programs Biomed.*, vol. 182, Dec. 2019, Art. no. 105102.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, 2015, pp. 234–241.

[24] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[25] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020.

[26] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[29] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[30] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[31] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial attention U-Net for retinal vessel segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1236–1242.

[32] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[33] R.-G. Dumitru, D. Peteleaza, and C. Craciun, "Using DUCK-net for polyp image segmentation," *Sci. Rep.*, vol. 13, no. 1, p. 9803, Jun. 2023.

[34] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.

[35] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[36] M. Guo, T. Xu, J. Liu, Z. Liu, P. Jiang, T. Mu, S. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. media*, vol. 8, no. 3, pp. 331–368, 2022.
[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[38] Y. Jiang, S. Cao, S. Tao, and H. Zhang, "Skin lesion segmentation based on multi-scale attention convolutional neural network," *IEEE Access*, vol. 8, pp. 122811–122825, 2020.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[43] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.
[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[45] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9584–9592.
[46] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.
[47] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *MultiMedia Modeling*. Daejeon, South Korea: Springer, 2020, pp. 451–462.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

HAOYU CHEN received the B.S. degree in mechanical engineering from Chonnam National University, South Korea, in 2022, where he is currently pursuing the M.S. degree with the Department of Artificial Intelligence Convergence. His research interests include image segmentation, image classification, and image recognition.

KYUNGBAEK KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1999, 2001, and 2007, respectively. He is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University. Previously, he was a Postdoctoral Researcher with the Department of Computer Sciences, University of California at Irvine, Irvine, CA, USA. His research interests include intelligent distributed systems, software-defined networks/infrastructure, big data platforms, GRID/cloud systems, social networking systems, AI-applied cyberphysical systems, blockchain, and other issues of distributed systems. He is a member of ACM, IEICE, KIISE, KIPS, KICS, KIISC, and KISM.