IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2020, accepted February 23, 2021, date of publication March 8, 2021, date of current version March 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3064180

A Lightweight Multiscale Attention Semantic Segmentation Algorithm for Detecting Laser Welding Defects on Safety Vent of Power Battery

YISHUANG ZHU, RUNZE YANG, YUQING HE, JUNXIAN MA, HAOLIN GUO, YATAO YANG, AND LI ZHANG

College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

Corresponding authors: Yatao Yang (yatao86@szu.edu.cn) and Li Zhang (zhang_li@szu.edu.cn)

ABSTRACT At present, in order to improve the safety performance of power battery, a safety vent is welded on the battery cover to avoid unpredictable explosions. It is vital to detect the laser welding defects on safety vent effectively for product quality. In this paper, a lightweight multiscale attention semantic segmentation algorithm with high accuracy and efficiency was proposed. We built an experimental dataset of safety vent welding defects with a total of 7263 original images, which were collected from a battery manufacturing production line. The main framework of the proposed model consists of four modules: the improved Res2Net serving as the feature extraction sub-module, an attention mechanism, a localization block and a boundary anti-aliasing module. This architecture can segment defects of different sizes and shapes in real-time and get more refined segmentation results simultaneously. To evaluate the method, experiments concerning mean IOU and pixel accuracy were conducted, and an average validation accuracy of 99.4% and the mean IOU of 84.67% were achieved respectively. Furthermore, comparison experiments using some outstanding algorithms on safety vent's welding defects test dataset were performed. It proves that our method achieved the best performance in terms of model size, computational complexity, efficiency and detection accuracy. Specifically, the model size is only 3.8 MB, and the frames per second (FPS) is 132.3. In brief, the proposed model is suitable for laser welding quality detection on safety vent in an industrial environment. Additionally, our study can provide a reference for designing relevant defect detection tasks using semantic segmentation method.

INDEX TERMS Laser welding defects, convolutional neural network (CNN), multiscale attention, semantic segmentation.

I. INTRODUCTION

In recent years, as our country attaches great importance to environmental protection and adopts the corresponding strong national policies, the technology of power battery for new energy vehicles has been developed rapidly. Power battery is one of the most important core components of new energy vehicles, whose quality is directly related to the users' life and security, as well as the service life of the vehicle [1]. Therefore, the safety performance of power battery requires special attention. To avoid the potential explosion hazard of the power battery during use, a safety vent is usually fixed on the battery cover. When the internal pressure of the power battery exceeds the threshold value, the safety vent will burst to release the pressure, thus avoiding an accident [2]. Currently, due to the fast, accurate, delicate weld seam characteristics, laser welding is the mainly adopted technique for safety vent welding. In actual production, laser welding for the safety vent of the power battery and its battery cover is completed in the automatic production line.

However, due to equipment or human reasons, the surface of the safety vent after welding will inevitably present some appearance defects such as cracks, collapse, holes, and incomplete welding during manufacturing process. These defects appear unaesthetic, also may cause potential dangers for battery usage. Consequently, the detection of welding defects on the surface of safety vent is very important. At present, some factories still use manual vision to conduct

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

detection tasks, which consumes a lot of labor but is inefficient and behaves poorly in real time. Simultaneously, the detection accuracy is easily influenced by human experience and subjective judgment, and it is quite prone to lead to miss detection or misjudgment. To overcome the problems of manual inspection, some new automatic detection techniques like pattern recognition and machine vision, are gradually replacing the manual inspection of product's surface defects [3]. Generally, defect detection based on machine vision is to identify whether the image of the product contains defects. In the past few decades, some image processing methods have been widely used in image defect detection, for example, thresholding based [4], segmentation based [5], edge detection [6], Sobel or Canny operator [7], [8] and neural networks [9], support vector machines [10], KNN (K Nearest Neighbor) [11], etc. A similar approach is adopted in [12], [13]. However, there are still some problems to be resolved with these approaches, for instance, being susceptible to light, environment or noise can result in poor anti-interference ability. Additionally, the diversity of the product defects, complex background, and other factors make it difficult to identify the defect target and the recognition rate is low, which requires extensive testing and experienced engineers for further feature selection. Recently, with the great increase in computing power and the rapid development of artificial intelligence, deep learning is making a splash in the field of computer vision [14]. Deep learning can use multi-layer deep neural networks and a large number of data samples to automatically learn implicit relationships in data. It can combine low-level data features to obtain high-level feature representations, thereby improving the accuracy of subsequent recognition and classification. Compared with manual inspection and machine vision detection, deep learning learns characteristics of the defects from a deeper level and can adapt to more complex and changeable production environments, now, it has surpassed human and machine vision methods in some defect detection tasks. As a result, more and more deep learning methods are being applied to product defect detection field and have achieved remarkable success.

Generally, applications of deep learning in surface defect detection fall into three main categories: image classification, object detection, and semantic segmentation. For the image-classification-based method, in [15], T. Wang *et al.* sliced the input image and sent it into the deep learning network for recognition, then obtained high accuracy in the product quality control task. In [16], X. Xu *et al.* improved the Inception v3 model and achieved a top-one accuracy of 99.56% in the roller defect classification task. Additionally, surface defect detection can also be conducted using the object detection method. In [17], W. Choi *et al.* detected five damage types including concrete cracks, steel corrosion with two levels (medium and high), bolt corrosion, and steel delamination through the Faster RCNN network. In [18], J. Chen *et al.* Cascaded two detectors in a rough-to-fine pattern including SSD, YOLO to localize the fasteners' defects. Similarly, in [19], J. Zhang *et al.* used a modified SSD network for automatic detection of paint defects on the vehicle body. Finally, for high-precision defect detection in industrial applications, semantic segmentation-based method has a higher requirement. Compared to the aforementioned two methods, it needs to detect the content of the input image, as well as providing pixel-level defect locations. So, this method is more challenging in that the model needs a trade-off between accuracy and speed. In [20], J. Long *et al.* proposed a fully convolutional networks (FCN), which is considered to be a breakthrough in deep learning for image segmentation [21]. Unlike the FCN network, the U-net network [22] was proposed in 2015 to segment the cell wall, and the segmentation details were more precise. Since the advent of the U-net network, the encoder-decoder architecture is very effective in the field of image segmentation. So, many segmentation networks adopted in surface defect detection use it as the basic network framework. For example, in [23], X. Tao *et al.* designed a new auto-encoder structure to locate the pixel-level defect position and then identified them through a classification network. Also in [24], J. Jiang *et al.* applied the model based on U-net to detect the surface defect on the back glass of smartphones. In [25], S. Mei *et al.* used Gaussian pyramids together with semantic segmentation to reconstruct textile defects, and the inference stage completed the fusion by combining contextual information. A similar approach is adopted in [26]–[28]. These methods realized surface defect detection by designing a complex network structure, thus lead to a significant increase in model size and computational complexity, which is far from lightweight and high efficiency required by practical applications. Moreover, in addition to being related to the network structure, an excellent deep learning defect detection model also needs enough defect samples for training.

Turn to the safety vent's welding defect detection, the size and shape of each weld defect vary. Also, the defect features have a certain degree of randomness, which makes the precise segmentation of the safety vent defects a major problem. Aiming at the current research status and the existing problems of surface welding defect detection, our group proposed an optimized VGG (Visual Geometry Group) 16 classification algorithm in [29], which achieved 99.89% accuracy on the task of laser welding defect classification. In [30], we continued to subdivide the defects into seven classifications based on two classifications, and by optimizing the SqueezeNet structure, a top-one accuracy of 95.58% in the seven classifications task for laser welding defects had been achieved. In this paper, using welding defect images data collected from factory as an experimental dataset, we proposed a lightweight multiscale attention semantic segmentation algorithm. As aforementioned, U-net has become a milestone in the application of deep learning for image segmentation based on the encoder-decoder symmetric structure network, our research is also based on U-net. The main contributions of this paper are as follows. First, we constructed a safety vent welding defect dataset with a total of 7263 images, and performed relevant preprocessing and labeling of the collected

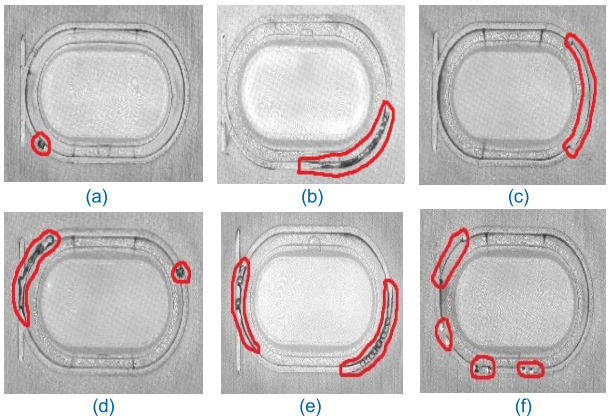


FIGURE 1. Four types of welding defects: (a) P. (b) WC. (c) MW. (d) P and WC. (e) Two sets of WC. (f) P and MW. (d) (e) (f) are mixed defects (MD).

images. A data augmentation strategy is also adopted to avoid over-fitting and enhance the robustness of the model. Then, the weld defects were segmented based on a multi-scale network, and the hole convolution [31] is used in the downsampling part to propose a segmentation model fused with multiple receptive fields. Additionally, the network used the improved Res2Net [32] as the feature extraction sub-module, which greatly reduced the parameter size and calculation complication of the model. Besides, a multiscale attention mechanism network is proposed, which is called MSAN in this paper. It can not only improve the robustness of features to scale changes, but also be capable of suppressing the noise and redundancy in the feature map through the generated mask. Finally, the localization block and boundary anti-aliasing module are proposed to make the model get more refined segmentation results.

II. DATASET ACQUISITION

In the actual safety vent welding process, the qualification rate of the product reaches more than 90%, which makes it difficult to collect samples with different shapes of defect types, and the acquisition cost will increase. To address the problem of data shortage, our group spent two months collecting a total of 7263 original images of safety vent welding defects from a cooperative factory. An AOI system embedded in a laser welding machine is applied in the factory, which consists of a CMOS industrial camera and an LED stable light source for obtaining high-quality images [29]. According to the shape of the defects, this paper classifies the obtained defects into four categories, namely porosity (P), welding collapse (WC), missing weld (MW), and mixed defects (MD), as shown in Figure 1.

Table 1 exhibits the details of the original dataset. We divided the images into three parts, with the training set accounts for 80%, the validation set accounts for 15%, and the test set accounting for 5%.

The original image has a resolution of 1800×1200, which has been normalized for preprocessing. In virtue of the small size of the input image, the subtle defects in the image will be covered, as the example porosity type shown in Figure 1 (a), which makes up only a small proportion of the entire image. Conversely, if the input image is too large, the parameter amount of the model will increase and the training and testing time will prolong correspondingly. Thus, a letterbox transformation is taken to transform the input image to 576×416. After preprocessing, the labeling tool LabelMe is used to provide a pixel-level annotation for each image. In the subsequent training phase, a data augmentation strategy is also applied to improve the detection model's immunity to interference, thereby making the application scenario more practical.

TABLE 1. The detailed information of the safety vent's defects dataset.

Dataset	P	WC	MW	MD	Total
Train	1373	1394	1364	1680	5811
Valid	257	261	256	315	1089
Test	86	87	85	105	363

III. METHODS

A. NETWORK ARCHITECTURE

Attention mechanisms and multiscale features are two important means used to optimize convolutional neural network (CNN) structure and improve the capacity of network feature expression. At present, most networks based on attention mechanisms are single scale ones. For example, the SE (Squeeze-and-Excitation) model [33] only processes the feature channel dimensions, accordingly, the generated mask cannot effectively focus on the multiscale information in the feature. Although ordinary multiscale networks can acquire multiscale features, the redundancy and noise in the features will affect the overall performance of the network. To deal with these problems, we proposed a multiscale attention network defect detection model based on U-net, which we called MSAN_Unet, and the main structure of this model is shown in Figure 2 (a). Figure 2 (b) and Figure 2 (c) depict the localization block (LB) and the boundary anti-aliasing module (BA) in the model respectively.

There are two main methods for obtaining multiscale features in the convolutional layer. One is to use convolution kernels of different sizes, and the other is to group features along channel dimensions and interactively output features between groups to obtain receptive fields of different scales. Unlike U-net, which uses ordinary convolutional structures to extract features, MSAN_Unet replaces it with the improved Res2Net, named MSAN. Figure 3 (a) presents the original Res2Net structure while Figure 3 (b) exhibits the improved MSAN structure.

MSAN enables the extraction of features at different scales in the downsampling stage, which is responsible for feature multiplexing and also prevents the gradient disappearance phenomenon. To better understand the function of MSAN,

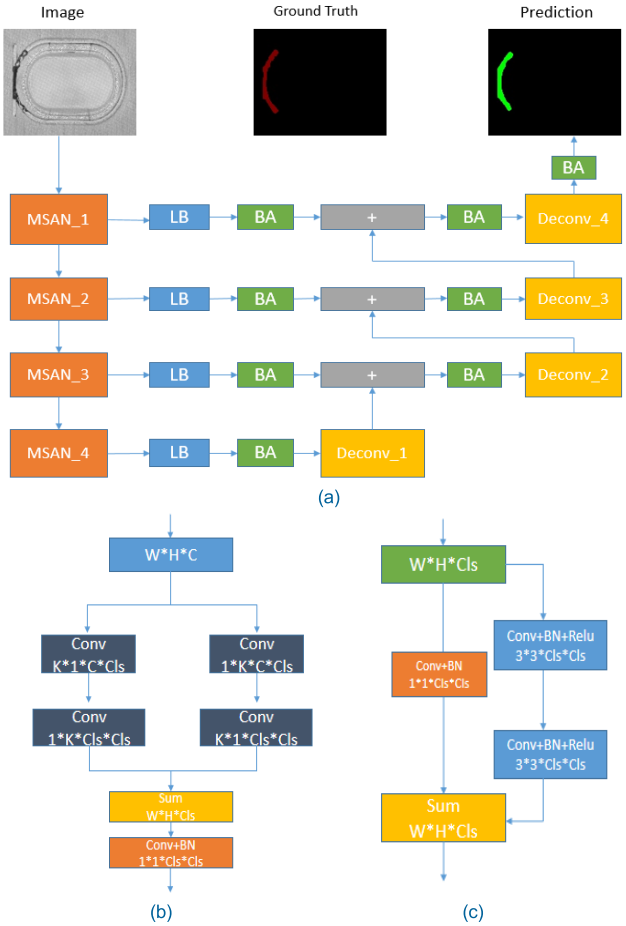


FIGURE 2. (a) MSAN_Unet, MSAN is the multiscale attention network, LB is the Localization block, BA is the boundary anti-aliasing module. (b) LB. (c) BA, Cls is the pixel type, k is the size of the convolution kernel, BN is batch normalization module, Conv is the convolutional layer, ReLU is the ReLU activation function.

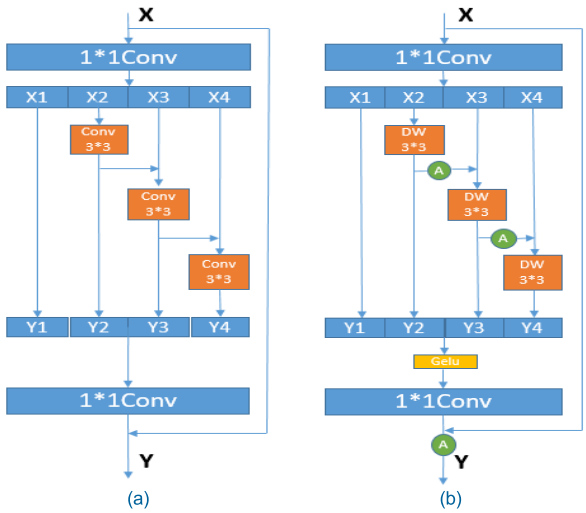


FIGURE 3. (a) Res2Net's network architecture, Conv is the convolutional layer. (b) MSAN, DW is the depthwise separable convolutional layer, A is the attention mechanism, as shown in Figure 5. Gelu is the Gelu activation function.

the architecture details presents here. From Figure 3 (b), after 1*1 convolution, the input feature X will output four sets of feature subsets X_i. Each subset has the same spatial size and

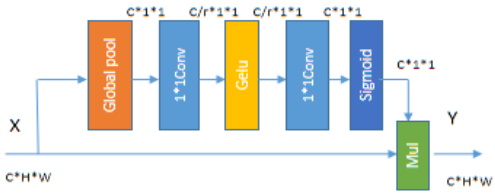


FIGURE 4. Attention mechanism, Global pool is the global pooling layer, Gelu is the Gelu activation function, Conv is the convolutional layer, Sigmoid is the Sigmoid activation function. H and W denote the size of the input features, C is the number of channels, and r is the channel compression ratio.

the number of the channels is a quarter of the original feature except X_1, which does not undergo a convolution operation, thus to maintain the receptive field scale of the input feature, while other groups of features need to undergo corresponding 3*3 convolution. Then, in order to obtain a large receptive field, a 3*3 hole convolution with three different dilation rates are used to convolve the output features of the previous step in parallel, and the dilation rate parameters are one, two, and four respectively, followed by a 1*1 convolution to keep the channels consistent, denoted by f_i. Finally, each group of features is spliced and merged in the channel dimension, and passes Gelu (Gaussian error linear units) [34] function and a 1*1 convolution to obtain model output features with different scale receptive field. If an attention model is introduced on this basis, a mask can be generated based on the features at a different scale, and the network can pay attention to the multiscale information in the features when the mask is fused with the features. However, in the process of generating multiscale features, there is a lot of noise and redundant information in the features of the channels due to different target size and random background interference in the input image, especially in the initial stage of training, when the network is unable to extract characteristics effectively from the input image. Furthermore, by adding the features of adjacent channels, the noise existing in the previous channel will be superimposed on the next channel, which will increase the convergence difficulty of the network and affect the performance improvement of the network. Since the convolution operation only fuses channel and spatial information on the local receptive field during feature extraction, the output features cannot capture contextual information beyond this field. To overcome this shortcoming, MSAN introduces the attention mechanism in the process of inter-group interaction of feature information, with an attention structure as shown in Figure 4.

For attention mechanisms, first, the output X of the convolution layer is used as the input to the model, where $X \in R^{C*H*W}, X = [X_1, X_2, \ldots, X_C], X_C = [X_C^1, X_C^2, \ldots, X_C^N], \mathrm{N} = H*W$. Since the convolution operation can only process local information of the features, it means that the image is only a collection of a series of local descriptors, lacking global information. So, directly processing the convolution output features cannot effectively model the interrelationship between channels. Usually, this problem is solved by converting the feature space information

into channel descriptors using the global average pooling operation. The i-th channel of Z can be expressed as formula (1) [33]:

$$Z_i = \frac{1}{N}\sum_{j=0}^{N} X_i^j \tag{1}$$

In order to utilize the information collected from feature Z, the subsequent operations not only need to be able to obtain the nonlinear relationship between the feature channels, but also need to be able to learn a non-mutually exclusive relationship to ensure that the output mask allows each channel to get attention. We designed a network based on the bottleneck layer to extract the mask, and mask A(x) can be calculated as formula (2) [33]:

$$A(\mathrm{x}) = \varphi(W_2\phi(W_1 Z)) \tag{2}$$

where φ means the Sigmoid activation function, ϕ represents the Gelu activation function, $W_1 \in R^{\frac{C}{\mathrm{r}}*C}$ and $W_2 \in R^{C*\frac{C}{\mathrm{r}}}$ denote the weights of the 1*1 convolution layer, r denotes the channel compression ratio.

Then, a high-performance neural network activation function Gelu is adopted, and it is expressed as formula (3) [34]:

$$G\mathrm{elu}(x) = xP(X \leq x) = x\varphi(x) \tag{3}$$

where x is the input, and X obeys the standard normal distribution, while $\varphi(x)$ is the probability function of the normal distribution. $P(X <= x)$ determines how much information is retained in x. The feature map is transformed by sigmoid into a mask with a value domain of (0, 1), which is the attention coefficient we need. Finally, after upsampling the mask of size $C*1*1$ to $C*H*W$ equivalently, the output of the attention model can be obtained by fusing the mask with the feature map X through the matrix dot product operation, which can be expressed as follows [33]:

$$Y = A(\mathrm{x}) \otimes X \tag{4}$$

Consequently, the previous set of features (Such as X_2) is converted into a mask by the attention module to suppress the noise and redundancy of the next subset of features (X_3) after the convolution operation, so, the processed features (X_3) have a stronger feature expression ability and can alleviate the pressure of feature learning in subsequent 3*3 convolutional layers. We visualized the area of interest of the input image on the MSAN network, and the result is shown in Figure 5. The highlighted portion of the image represents the network's focus on the image. It can be seen that MSAN can effectively focus on the important areas in the picture compared to not using the attention mechanism. Specifically, the area of attention can evenly cover the whole defect location, indicating that MSAN has a stronger multi-scale capability. From Figure 5 (b) and (c), the feature before mask processing includes other positions besides the defect positions, which means that there is a lot of noise in the feature. Comparatively, the feature after mask processing is highlighted mainly in the defect positions, which means the noise existing in the feature

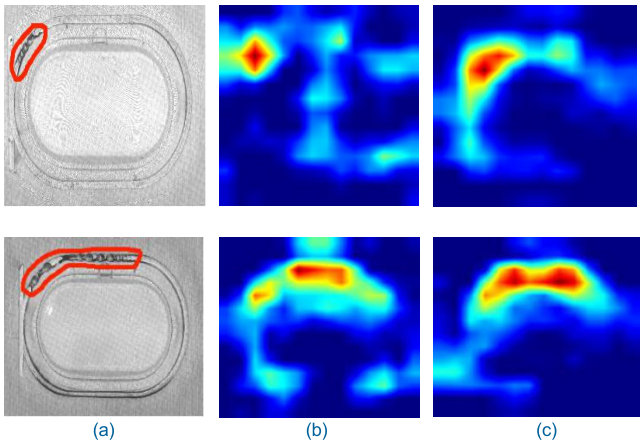


FIGURE 5. Heat map visualization of interest areas of the input image. (a) Original image. (b) Without attention mechanism. (c) With attention mechanism.

is weakened. It verified that the MSAN network can reduce noise and redundancy present in the next grouping features by using the generated masks.

Briefly, MSAN is more capable of handling multiscale features and can improve the expression ability of the model's output characteristics. The output feature Y_i of MSAN can be calculated as formula (5).

$$Y_i = \begin{cases} X_i, & i = 1 \\ CB(f_i(X_i)), & i = 2 \\ CB(f_i(X_i \otimes A(Y_i - 1))), & 2 < i \leq n. \end{cases} \tag{5}$$

Here, CB represents a 1*1 convolution and batch normalization layer. Since the weld defects images of safety vent are edge-rich ones, and the sizes, shapes of the defects vary, also, the location of the defects is not fixed, the defect characteristics have a certain degree of randomness. Concerning defect location, operations like full connection layers or pooling structures can lose location information, in order to retain more spatial position information, full convolution layer is used in our model. In terms of defect classification, we need a larger convolution kernel to make the link of each point on the feature map denser. Then, the features obtained in the downsampling stage pass through the localization block, as shown in Figure 2 (b). Due to the small receptive field of the shallow network, large convolution kernel is composed of symmetrical independent convolution kernel. The convolutions of $1 \times k + k \times 1$ and $k \times 1 + 1 \times k$ are used to replace the original $k \times k$ large kernel convolution according to the matrix decomposition. Because large convolution kernels allow a larger perceptual area, it is more conducive to global feature extraction. Compared with the original large convolution kernel, our design can significantly reduce the parameters and computation. Simultaneously, due to the usage of large convolution kernels in the localization block, the pixel misclassification at the defect boundary is increased, leading to jaggedness phenomenon at the defect edges. Therefore, another module using a small convolution kernel is introduced to do the balance, which is the edge

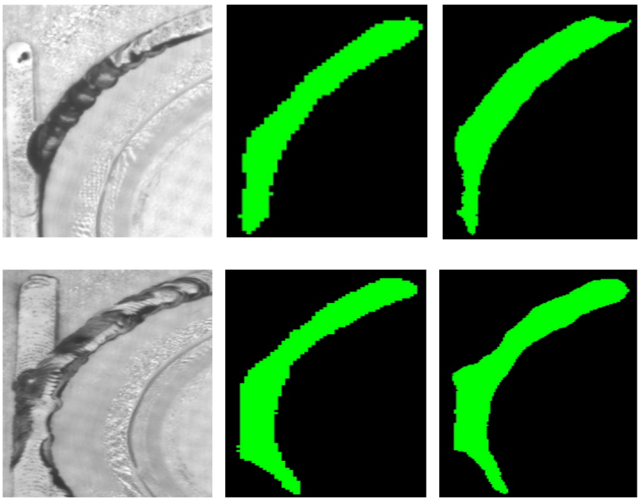


FIGURE 6. The ablation experiments comparison between Using LB and BA modules versus not using them. (a) Original image. (b) Without LB and BA modules. (c) With LB and BA modules.

anti-aliasing module designed based on the residual structure to make the edges of the defect smoother, as shown in Figure 2 (c). The edge anti-aliasing module allows the model to get finer segmentation results. We also conducted ablation experiments on this module, and the results are shown in Figure 6. Obviously, using LB and BA modules can better locate the defects, which allow the segmented image edges smoother and closer to the original image defect positions.

During the upsampling process, transposed convolution [35] is used to replace the deconvolutional layer, and skip connections are also utilized in the symmetrical levels. In this way, the multiple upsampling method allows the final output feature map to fuse more low-level features, as well as features at different scales, thus making the recovered edges finer, consequently, it is suitable for multiscale prediction and depth supervision.

B. EXPERIMENTAL CONFIGURATION AND THE LOSS FUNCTION

All the experiments present in this paper were carried out under Linux Ubuntu 16.04 LTS with CUDA 9.0 library, an Intel Xeon E5-2683 v4 @ 2.10 GHz CPU with 256G of RAM, and the GPU is NVIDIA GTX 2080Ti. Keras with TensorFlow backend is used as the deep learning framework. When designing and training a model, loss function and optimization algorithm are two important factors to be considered. The smaller the loss function is, the better the ability to guide model learning will be. We first consider the cross-entropy loss function, expressed as formula (6) [36]:

$$CE = -\frac{1}{N}\sum_{i-1}^{N}(y_i \log p_i + (1 - y_i)\log(1 - p_i)) \tag{6}$$

where y_i is the true category of the input instance data, and p_i is the probability that the predicted input instance belongs to the correct category. The effect can be improved when the data is relatively balanced. But the training set segmentation is not very good for porosity defects (As seen in Figure 1 (a)), which only occupies a small area of the image. In order to solve the category imbalance problem, we combined the dice loss function with equation (7) [37]:

$$Dice = \sum_{i=1}^{k}\left(\frac{2\sum_{i=1}^{N} p_i t_i + \delta}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i + \delta}\right) \tag{7}$$

where p_i represents the predicted value, while t_i represents the true label value, δ is the unit constant, so the final loss function is:

$$Combine\_Loss = \partial^* CE - (1 - \partial)Dice \tag{8}$$

It is defined as the weighted sum of dice loss and cross-entropy. Among them, ∂ is used to control the effect of dice loss on Combine_Loss function, and the value of ∂ determines the degree of the effect on Combine_Loss. In our experiments, plus-one smoothing is used to prevent zero problems for divide by adding the unit constant δ to both the numerator and the denominator of the Dice loss term [36]. The experiment uses the SGD optimizer, which sets the momentum and weight decay coefficients to 0.9 and 0.0001 respectively. The Keras' built-in LearningRateScheduler function is used for selecting the learning rate, which can dynamically adjust the learning rate at the beginning or the end of each epoch. The best set we found for model convergence are as follows: the initial learning rate is set to 0.0001, the epoch is set to 50, and the batch size is set to 4.

C. DATA AUGMENTATION

Usually, the defect detection model for industrial applications should be adapted to various scenarios, such as different brightness, diverse targets, etc. Increasing the number and diversity of the training samples can improve the robustness of the model and reduce the dependence of the model on certain characteristics. When conduct data augmentation, we have to ensure that the main features of the augmented sample remain consistent with the original sample. In our case, image-space geometric transformations like image flipping and random angle rotation are used. Besides, zero-mean Gaussian noise, Gamma transformation, and contrast changes are also added to enhance the generalization of the model.

D. PERFORMANCE METRICS

The performance of our model architecture is evaluated in terms of mean IOU and pixel accuracy, which are the two most commonly used indicators to evaluate the performance of semantic segmentation [38]. Mean IOU outputs the class prediction accuracy of each pixel, while pixel accuracy measures the overlap rate between the two targets by calculating the ratio of the intersection and union with the ground truth masks. The calculation formulas for the two indicators are as

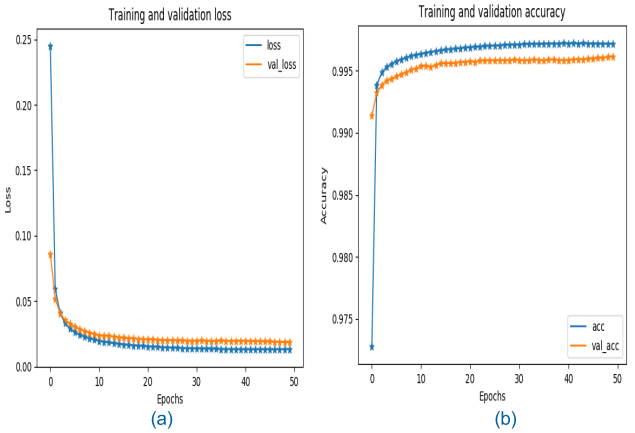


FIGURE 7. (a) Loss versus epochs. (b) Accuracy versus epochs.

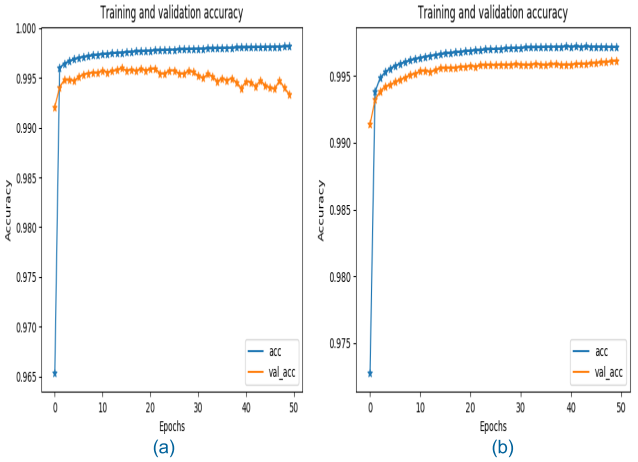


FIGURE 8. Accuracy versus epochs (a) Before data augmentation. (b) After data augmentation.

follows [38]:

$$mean\_IOU = \frac{1}{C}\sum_{i}\frac{p_{ii}}{c_i + \sum_j p_{ji} - p_{ii}} \tag{9}$$

$$pixel\_accuracy = \frac{\sum_i p_{ii}}{\sum_i c_i} \tag{10}$$

where C denotes the number of categories of all the pixels, c_i is the pixels number of category i, and p_{ij} is the pixels number of the category i predicted to be j.

IV. EXPERIMENTAL RESULTS

After trained for 50 epochs, the average pixel accuracy of training is 99.6%, and the validation accuracy can reach 99.4%. The loss and pixel-level accuracy of training and testing versus epoch are plotted as Figure 7. The resulting loss function curve is smooth and close to each other, which indicates that the model is fitted basically and has a fine generalization ability.

We conducted comparative experiments before and after data augmentation, and the result is shown in Figure 8. From Figure 8 (a), when data augmentation is not adopted, after a certain number of iterations, the curve of the verification set gradually deviated from the curve of the training set, means overfitting occurs. It can be seen from Figure 8 (b) that the problem of overfitting is solved after data augmentation.

Table 2 presents the comparisons of the experimental results of the model before and after data augmentation. Obviously, mIOU has been significantly improved from 82.6% to 84.7% after data augmentation, which indicates that data augmentation is an effective method to avoid overfitting.

Table 3 presents the influence of alpha size on experimental results. From Table 3, mIOU reaches the best result when ∂ takes 0.8.

We present the effect of convolution kernel size (K) on localization block (LB) in Table 4. In the process of K raising from seven to nine, the number of model parameters increased but mIOU decreased. Therefore, in the localization block, the convolution kernel size is selected as seven to obtain the optimal result of the model.

TABLE 2. Experimental results comparisons before and after data augmentation.

Data augmentation	mIOU(%)	Pixel acc(%)
Before	82.6	99.2
After	84.7	99.4

TABLE 3. The effect of alpha size on experimental results.

∂	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIOU(%)	79.1	80.3	81.7	82.4	82.9	84.7	83.2

TABLE 4. The effect of K size on LB.

K	3	5	7	9	11	13
mIOU(%)	81.4	82.3	84.7	84.1	82.9	81.7
Parameters(k)	409.4	413.2	417.2	421.1	425.1	428.9

We also did ablation experiments on each module of the MSAN_Unet network, including the improved Res2Net residual module (NewRes2Net), attention mechanism (A), localization block (LB), and boundary anti-aliasing module (BA), and the results are shown in Table 5. When using Res2Net, 77.8% of mIOU was obtained. After replacing Res2Net with NewRes2Net module, mIOU was 82.5%. Additionally, we added the aforementioned modules A, LB and BA successively on the NewRes2Net module based architecture, the final mIOU reached 84.7%. Therefore, each improved module is effective and can increase the accuracy of the network.

TABLE 5. Ablation experiment results for each module of MSAN_Unet.

Method	mIOU(%)
Res2Net	77.8
NewRes2Net	82.5
NewRes2Net+A	83.9
NewRes2Net+A+LB	84.4
NewRes2Net+A+LB+BA	84.7

TABLE 6. Performance comparison of our model with some state-of-the-art models.

Model	FLOPs (M)	Model size(MB)	Parameters (M)	FPS	mIOU (%)
Unet	14.34	57.6	7.17	-	62.86
FCN-8s	7.22	29.1	3.61	-	80.60
BiSeNet[39]	45.64	91.6	22.81	90.6	79.87
Enet [40]	0.76	4.0	0.37	128.5	71.28
ICNet [41]	13.47	54.5	6.74	61.3	61.34
DeepLab-v3+ [42]	82.23	329.1	41.06	-	73.95
PSPNet [43]	8.89	7.8	0.96	-	79.43
SegNet [44]	5.88	23.7	2.94	14.7	83.66
HRNet [45]	19.06	77.2	9.52	66.7	65.43
Ours	**0.83**	**3.8**	**0.42**	**132.3**	**84.67**

Table 6 presents the performance comparisons of our model with some state-of-the-art models on the safety vent welding defect test dataset. The image resolution of 576×416 is utilized. The comparative parameters consist of model size, computational complexity, mIOU, and FPS (Frames Per Second). It demonstrated that our model has great advantages over other models both in speed and accuracy.

We compared the predicted segmentation results with the ground truth images from the safety vent welding defect test dataset, and the results are shown in Figure 9. Column (a) is the original image of the test dataset, column (b) is the predicted image by our model, while column (c) is the ground truth image from the test dataset. Clearly, the predicted segmentation image has smooth edges and is close to the manual labeling result, means that our model can accurately predict welding defects of safety vent of different sizes and shapes.

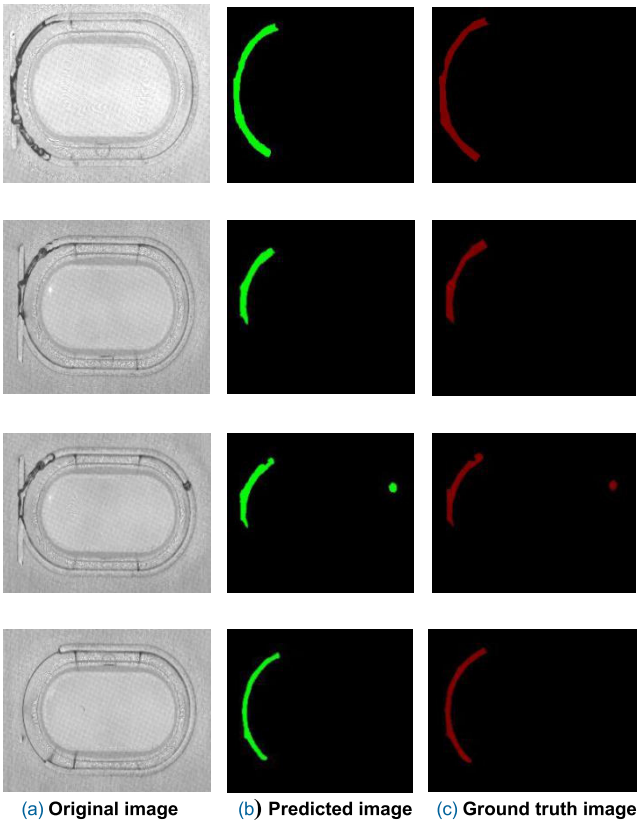


FIGURE 9. The comparison between the predicted segmentation results and the ground truth images.

V. CONCLUSION

This paper proposes a multiscale attention semantic segmentation network for the segmentation of surface welding defects of safety vent in the real industry. First, the network obtains multiscale features by combining feature channel grouping and information interaction between groups, and then gets multiscale attention features by adding attention mechanism in the process of information interaction between groups. Features from the previous set are converted into masks by an attention mechanism to fuse with the next set of features, thereby reducing their presence of noise and redundancy. Second, the network uses the improved Res2Net as the feature extraction sub-module to obtain different receptive fields with different dilation rates of the hole convolution, it improves the perception ability of the network for multiple targets, also greatly reduces the number of model parameters and the computational complexity. At last, this paper uses the localization module and the boundary anti-aliasing module to make the model get more refined segmentation results. Additionally, a data augmentation strategy is adopted, which combines Gaussian noise, Gamma transformations, and spatial geometric transformations to avoid overfitting due to the small dataset. It enhances the robustness and generalization ability of the model simultaneously. The comparison experiments results effectively validate the proposed model, which can segment defects of different shapes and sizes in real-time. In future work, we plan to continue to expand the welding

defect dataset of the safety vent, and further improve the network architecture to achieve higher detection accuracy, faster test speed, and better detection stability.

REFERENCES

[1] Z. Youlong, Y. Wenqiang, R. Kai, and Z. Chengyan, "Research on power battery technology of pure electric vehicle," *Automobile Appl. Technol.*, to be published.

[2] J.-H. Kim, K.-H. Lee, D.-C. Ko, S.-B. Lee, and B.-M. Kim, "Design of integrated safety vent in prismatic lithium-ion battery," *J. Mech. Sci. Technol.*, vol. 31, no. 5, pp. 2505–2511, May 2017.

[3] C. L. S. C. Fonseka and J. A. K. S. Jayasinghe, "Implementation of an automatic optical inspection system for solder quality classification of THT solder joints," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 9, no. 2, pp. 353–366, Feb. 2019.

[4] H.-F. Ng, "Automatic thresholding for defect detection," *Pattern Recognit. Lett.*, vol. 27, no. 14, pp. 1644–1649, Oct. 2006.

[5] H. Oliveira and P. L. Correia, "Automatic road crack segmentation using entropy and image dynamic thresholding," in *Proc. 17th Eur. Signal Process. Conf.*, Aug. 2009, pp. 622–626.

[6] C. Jia, Y. Wang, and J. Xing, "Edge detection of crack defect based on wavelet multi-scale multiplication," *Comput. Eng. Appl.*, vol. 47, no. 15, pp. 219–221, May 2011.

[7] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the Sobel operator," *IEEE J. Solid-State Circuits*, vol. 23, no. 2, pp. 358–367, Apr. 1988.

[8] L. Er-Sen, Z. Shu-Long, Z. Bao-Shan, Z. Yong, X. Chao-Gui, and S. Li-Hua, "An adaptive edge-detection method based on the canny operator," in *Proc. Int. Conf. Environ. Sci. Inf. Appl. Technol.*, Jul. 2009, pp. 465–469.

[9] H. N. Robert, "Theory of the backpropagation neural network," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 1, 1989, pp. 593–605.

[10] C. Saunders, "Support Vector Machine," *Comput. Sci.*, vol. 1, no. 4, pp. 1–28, 2002.

[11] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[12] Z. Liu, B. Gao, and G. Y. Tian, "Natural crack diagnosis system based on novel L-shaped electromagnetic sensing thermography," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9703–9714, Nov. 2020.

[13] B. Gao, P. Lu, W. L. Woo, and G. Y. Tian, "Variational Bayes subgroup adaptive sparse component extraction for diagnostic imaging system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1518–1522.

[14] W. Cao, J. Yuan, Z. He, Z. Zhang, and Z. He, "Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection," *IEEE Access*, vol. 6, pp. 8990–8999, 2018.

[15] T. Wang, Y. Chen, M. Qiao, and H. Snoussi, "A fast and robust convolutional neural network-based defect detection model in product quality control," *Int. J. Adv. Manuf. Technol.*, vol. 94, nos. 9–12, pp. 3465–3471, Feb. 2018.

[16] X. Xu, H. Zheng, Z. Guo, X. Wu, and Z. Zheng, "SDD-CNN: Small data-driven convolution neural networks for subtle roller defect inspection," *Appl. Sci.*, vol. 9, no. 7, p. 1364, Mar. 2019.

[17] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 9, pp. 731–747, Sep. 2018.

[18] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2018.

[19] J. Zhang, J. Xu, L. Zhu, K. Zhang, T. Liu, D. Wang, and X. Wang, "An improved MobileNet-SSD algorithm for automatic defect detection on vehicle body paint," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 23367–23385, Aug. 2020.

[20] 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.

[21] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Proc. 19th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* (Lecture Notes in Computer Science), vol. 9901. 2016, pp. 478–486.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. 2015, pp. 234–241.

[23] X. Tao, D. Zhang, W. Ma, X. Liu, and D. Xu, "Automatic metallic surface defect detection and recognition with convolutional neural networks," *Appl. Sci.*, vol. 8, no. 9, p. 1575, Sep. 2018.

[24] J. Jiang, P. Cao, Z. Lu, W. Lou, and Y. Yang, "Surface defect detection for mobile phone back glass based on symmetric convolutional neural network deep learning," *Appl. Sci.*, vol. 10, no. 10, p. 3621, May 2020.

[25] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors*, vol. 18, no. 4, p. 1064, Apr. 2018.

[26] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1266–1277, Jun. 2018.

[27] 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.

[28] G. Zhou and H. Sun, "Defect detection method for steel based on semantic segmentation," in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 975–979.

[29] Y. Yang, L. Pan, J. Ma, R. Yang, Y. Zhu, Y. Yang, and L. Zhang, "A high-performance deep learning algorithm for the automated optical inspection of laser welding," *Appl. Sci.*, vol. 10, no. 3, p. 933, Jan. 2020.

[30] Y. Yang, R. Yang, L. Pan, J. Ma, Y. Zhu, T. Diao, and L. Zhang, "A lightweight deep learning algorithm for inspection of laser welding defects on safety vent of power battery," *Comput. Ind.*, vol. 123, Dec. 2020, Art. no. 103306.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[32] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[34] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*. [Online]. Available: https://arxiv.org/abs/1606.08415

[35] S. Falong and Z. Gang, "Fast semantic image segmentation with high order context and guided filtering," May 2016, *arXiv:1605.04068*. [Online]. Available: https://arxiv.org/abs/1605.04068

[36] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, Jan. 2021.

[37] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[39] Y. Changqian, W. Jingbo, P. Chao, G. Changxin, Y. Gang, and S. Nong, "BiseNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. 15th Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science). 2018, pp. 334–349.

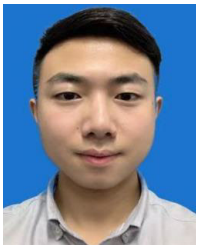
[40] A. Paszke, A. Chaurasia, K. Sangpil, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," Jun. 2016, *arXiv:1606.02147*. [Online]. Available: https://arxiv.org/abs/1606.02147

[41] Z. Hengshuang, Q. Xiaojuan, S. Xiaoyong, S. Jianping, and J. Jiaya, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. 15th Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science). 2018, pp. 418–434.

[42] H. G. Schnack, H. E. H. Pol, W. F. C. Baaré, M. A. Viergever, and R. S. Kahn, "Automatic segmentation of the ventricular system from MR images of the human brain," *NeuroImage*, vol. 14, no. 1, pp. 95–104, 2001.

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[44] 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.
[45] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.



YISHUANG ZHU received the B.S. degree from Huaibei Normal University, China, in 2018. He is currently pursuing the M.S. degree with Shenzhen University. His research interests include deep learning and the optical Internet of Things.

RUNZE YANG is currently pursuing the M.S. degree with Shenzhen University. His research interests include image processing and the optical Internet of Things.

YUQING HE is currently pursuing the M.S. degree with Shenzhen University. Her research interests include data processing and the optical Internet of Things.

JUNXIAN MA is currently the Board Chairman of the Shenzhen Institute of Communications and a Tenured Professor with the College of Electronics and Information Engineering, Shenzhen University. His research interests include micro and nano components of optical fiber communication, and sensor networks.

HAOLIN GUO is currently pursuing the M.S. degree with Shenzhen University. His research interests include deep learning and the optical Internet of Things.

YATAO YANG received the B.Sc. degree in optical instrumentation engineering from Zhejiang University, China, and the Ph.D. degree in fiber optics from Glasgow Caledonian University, U.K. He joined the Institute of Optics and Electronics, Chinese Academy of Sciences, where he was involved in semiconductor equipment and optoelectronic device development. In 1996, he was with the University of Leeds, U.K., as a Research Officer, involved in the areas of optical fiber laser materials. In 1997, he joined Resonance Ltd., Canada, where he developed spectral gas sensors. In 1998, he joined JDSU Corporation, Canada, where he was developing optical fiber devices. In 2000, he joined Chorum Technologies Inc., U.S., where he was developing optical fiber devices. In 2004, he joined JDSU Corporation, U.S., where he was developing optical fiber devices and fiber lasers. In 2009, he joined NeoPhotonics Corporation, as VP of Research and Development, developing optical fiber devices. In 2014, he founded Shenzhen Dade Laser Technology Company Ltd. In 2017, he was appointed as a Distinguished Professor with the College of Electronics and Information Engineering, Shenzhen University, China. In 2018, he became the Head of the Smart IoT Center, Shenzhen University. His research interests include optical networking, optical sensors, optical data transmission and data processing, optical nanomaterials, optical-wireless communications, and lasers.

LI ZHANG received the M.S. degree in communication and information systems from Lanzhou University, China, in 1999, and the Ph.D. degree in optical engineering from the Huazhong University of Science and Technology, China, in 2008. In 2008, she worked with Shenzhen University. From 2016 to 2017, she worked with the Department of Electrical and Computer Engineering, University of California, San Diego, for one year, as a Visiting Scholar. She is currently an Associate Professor with the College of Electronics and Information Engineering, Shenzhen University. Her current research interests include optical communication, surface plasmon polariton (SPP) waveguides and devices, artificial intelligence, and deep learning in the field of optical Internet of Things.