

## Research Article

**Cite this article:** Song Z, Yao H, Tian D, Zhan G and Gu Y (2025). Segmentation method of U-net sheet metal engineering drawing based on CBAM attention mechanism. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **39**, e14, 1–14  
<https://doi.org/10.1017/S0890060424000301>

Received: 21 January 2024

Revised: 23 July 2024

Accepted: 12 December 2024

### Keywords:

intelligent manufacturing; welding engineering; convolution block attention module; U-net; double-pool convolution

### Corresponding author:

Yajing Gu;

Email: [guyj90@zju.edu.cn](mailto:guyj90@zju.edu.cn)

# Segmentation method of U-net sheet metal engineering drawing based on CBAM attention mechanism

ZhiWei Song<sup>1</sup> , Hui Yao<sup>2</sup> , Dan Tian<sup>2</sup> , Gaohui Zhan<sup>2</sup>  and Yajing Gu<sup>1</sup>

<sup>1</sup>Ocean College, Zhejiang University, Zhoushan, Zhejiang, 316021, China and <sup>2</sup>School of Mechatronic Engineering, Xi'an Technological University, 710032, China

## Abstract

In this paper, an improved U-net welding engineering drawing segmentation model is proposed for the automatic segmentation and extraction of sheet metal engineering drawings in the process of mechanical manufacturing, to improve the cutting efficiency of sheet metal parts. To construct a high-precision segmentation model for sheet metal engineering drawings, this paper proposes a U-net jump structure with an attention mechanism based on the Convolutional Attention Module (CBAM) attention mechanism. At the same time, this paper also designs an encoder jump structure with vertical double pooling convolution, which fuses the features after maximum pooling+convolution of the high-dimensional encoder with the features after average pooling+convolution of the low-dimensional encoder. The method in this paper not only improves the global semantic feature extraction ability of the model but also reduces the dimensionality difference between the low-dimensional encoder and the high-dimensional decoder. Using Vgg16 as the backbone network, experiments verify that the IoU, mAP, and Accu indices of this paper's method in the welding engineering drawing dataset segmentation task are 84.72%, 86.84%, and 99.42%, respectively, which are 22.10, 19.09 and 0.05 percentage points higher compared to the traditional U-net model, and it has a relatively excellent value in engineering applications.

## Introduction

Heavy industry equipment generally adopts customized manufacturing. In such projects, large sheet metal parts need to be cut out according to the content of customized engineering drawings and finally manufactured by welding and stamping operations. Efficiency and precision are critical to the manufacture of customized heavy industrial equipment. The specific process of the traditional way to obtain sheet metal parts in the manufacturing process of heavy industry equipment is manual recognition of engineering drawings (Ablameyko and Uchida, 2007; De et al., 2016) manual redrawing of specific graphics (Tovey, 1989; Madsen and Madsen, 2016)-sheet metal cutting and stamping based on CAD/CAM integrated system (Pan and Rao, 2009; Lu et al., 2021; Favi et al., 2022). The process of obtaining specific sheet metal parts in traditional ways is cumbersome and inefficient.

In recent years, with the development of artificial intelligence, the industrial field is also more inclined to use deep learning methods to solve engineering problems. Some studies have shown that this method has an excellent performance in solving engineering problems. For example, Lau et al. (2020) utilized deep learning image segmentation to replace the traditional manual road crack detection and achieved extremely high detection accuracy in road defect detection using the ResNet-34 pre-trained model. Tabernik et al. (2020) and Hou et al. (2017) utilized deep learning image segmentation techniques to achieve the automatic detection of defects on metal surfaces and the automatic detection of defects in the welding process, respectively. Zhang et al. (2024) improved the ability of computers to recognize the contents of 2D engineering drawings using deep learning data enhancement techniques. Li et al. (2023) used a self-learning semi-supervised deep learning approach to achieve high-precision semantic segmentation of remote sensing images. Lu et al. (2023) achieved high-precision automated screening of hybridoma cells using the U-net segmentation model with a residual network attention mechanism. The commonly used deep learning method for plate part recognition is the graphical neural topology method, which is mainly applied to 3D part recognition and classification and is not yet able to satisfy the segmentation and extraction of the content contour of 2D engineering drawings (Ma and Yang, 2024).

As we all know, many excellent segmentation networks already exist in deep learning image segmentation. The fully convolutional segmentation model (FCNs) of deep learning was first proposed by Long et al. (2015). FCNs abandon the traditional fully connected layer, and the overall network structure uses a fully convolutional method to achieve end-to-end pixel-level

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

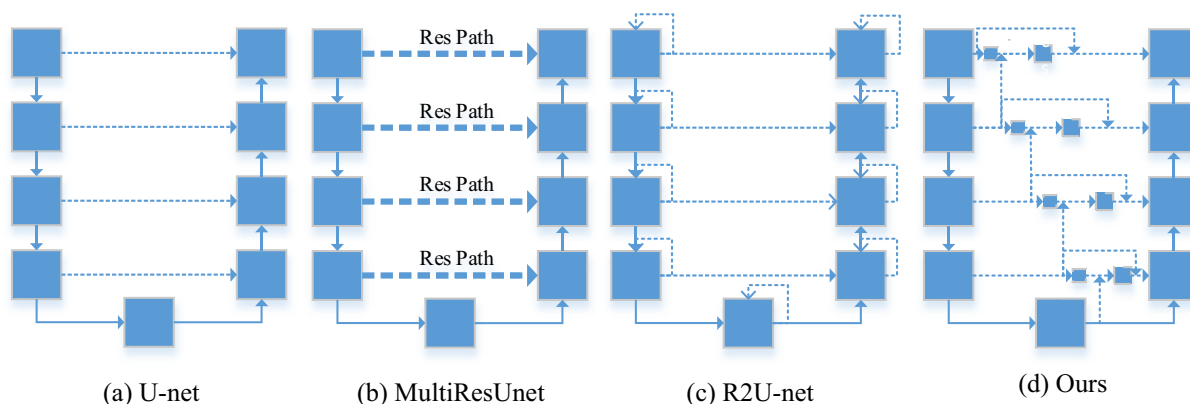
dense prediction of image features, which is suitable for more complex global semantic feature segmentation tasks. The fully convolutional segmentation model uses transposed convolution operations to obtain semantically segmented images of the original size through upsampling. The input of FCNs is an RGB image of any size and the output is the same size as the input. At the same time, it proposes a classic skip connection for fusing features from deep layers (including classification information) and intermediate layers (including location information) to improve feature accuracy output. U-net (Ronneberger and Fischer, 2015) can be considered a variation of FCNs, which still uses the encoder-decoder and skip structure. Compared with the former, the unique skip connection architecture of the U-net enables the decoder to obtain more spatial information lost during the pooling operation and restore a complete spatial resolution. The semantic difference between the encoder and decoder is reduced to achieve better segmentation performance. U-net mainly has two cores: (1) dimensional difference problem between the low-dimensional encoder and high-dimensional decoder in the process of semantic information fusion. How to effectively reduce the dimension difference in the image fusion process of the encoder and decoder? (2) The problem of image spatial position information, how can the encoder and decoder realize the learning of image spatial position information? Researchers have introduced many methods to solve the above problems to reduce the incompatible feature differences between these two groups.

Deeplab-v1, Deeplab-v2, Deeplab-v3, and PSPNet (Chen et al., 2014, Chen et al., 2017a, 2017b; Zhao et al., 2017) use null convolution and pooling with different step sizes to improve the resolution and accuracy of image segmentation, respectively, with more computational parameters in their processes. SegNet (Badrinarayanan et al., 2017) removes the fully connected layer and proposes the up-sampling model structure of the conv layer instead of the Deconv layer, which improves the resolution and greatly reduces its model's operation parameters. HRNet (Sun et al., 2019) adopts the design of a multi-resolution parallel tributary architecture, which better realizes the fusion of different depth semantic features. The segmentation networks focus on acquiring global image semantic features and spatial location information. Therefore, the model mainly integrates the global upper and lower semantic features in image segmentation and learns spatial location feature information. Currently, many improved models have emerged based on U-net, such as R2UNet, R2U++Net, CAggNet, MultiResUNet, NonlocalUNets, and UCTransNet (Alom et al., 2018;

Ibtehaz and Rahman, 2020; Wang et al., 2020, 2022; Cao and Lin, 2021; Mubashar et al., 2022). Such networks are improvements to the U-net hopping structure, to allow better integration of global contextual feature information in the encoder and decoder and thus improve segmentation performance. The unique hopping structure of the U-net model can better realize the fusion of global semantic information, which is characterized by high accuracy, low complexity, and high resolution in image segmentation tasks. As mentioned in UNet++ (Zhou et al., 2018), the dimensionality of the encoder semantic features at the front end of the jump structure is lower than that of the decoder semantic features at the back end, and thus, the large difference in feature dimensions in the jump structure makes the segmentation performance not good enough.

On the other hand, Fei Wang (Wang et al., 2017) proposed a residual attention network using an encoder-decoder approach. Based on this, Sanghyun Woo (2018) proposed the CBAM (Convolutional Block Attention Module) module to realize the entire convolution channel semantic and spatial information calculation. UCTransNet is a recently proposed attention module inspired by the Self Attention Mechanism and Multi-Head Attention mechanisms in Transformer (Vaswani et al., 2017), and its purpose is to enable the encoder-decoder to obtain more global information fusion. In this paper, an improved model based on U-net is proposed for the segmentation of specific units in sheet metal engineering drawings for heavy equipment manufacturing. The three models (a) (b) (c) in Figure 1 are all examples of model structures that improve the segmentation accuracy of different tasks by improving the U-net jump structure. (d) Figure 1 is a sheet metal engineering drawing outline extraction method based on the CBAM attention mechanism proposed in this paper, which is used for sheet metal segmentation in industrial equipment manufacturing.

By using traditional U-net to conduct feasibility experiments on the sheet metal welding pattern segmentation task, and using the attention mechanism to further study and analyze the structure and principle of U-net (Wang et al., 2017; Woo et al., 2018; Mohan and Bhattacharya, 2022). This paper proposes a dual-pooling convolutional fusion attention mechanism model based on CBAM, which considers the information fusion of channel and spatial dimensions as well as the dimensional difference between encoder and decoder features. Besides, this article proposes global information linkage between U-net encoders, which includes feature cluster integration between vertical encoders and vertical and horizontal double-pooling convolutions. The output



**Figure 1.** Comparison of our method (a) Skip connection structure of the original U-net model; (b) U-net model with residual network used as the skip connection; (c) U-net model with recurrent structures added to both the encoder and decoder; (d) with skip structure schemes of other models. Dashed lines denote skip connections.

features are fused with the original features through the attention module, and the whole jumps to the high-dimensional feature layer of the decoder to achieve secondary fusion. The improved U-net model can not only better extract global semantic features but also reduce the semantic differences in the process of encoder-decoder feature fusion. Compared with traditional models, this research method can better realize the global feature information fusion of encoder and decoder features, effectively reduce the dimensional difference between the low-dimensional encoder and high-dimensional decoder in the feature fusion process, and can better realize the segmentation of specific units in sheet metal engineering drawings. Through experimental verification, the method in this paper has better performance in the sheet metal graphics segmentation task. The main contributions of this study are as follows:

- Propose the automatic cutting technology of sheet metal parts based on U-net for heavy industry equipment manufacturing.
- The segmentation method combining CBAM and U-net is proposed to be suitable for (non-human-assisted) high-precision segmentation of sheet metal graphics.
- A double pooling + convolutional skip structure is proposed to reduce the dimensional difference between encoding and decoding features.
- A skip structure of vertical coding features is proposed to improve the fusion of global semantic information.
- The improved model has been verified for its high-performance segmentation capability of sheet metal graphics.

## Related Work

### CBAM-U-net model

The Convolutional Block Attention Module (CBAM) (Woo et al., 2018) can improve the ability of the convolutional network to express the image of the feature layer, in the process of extracting image features, pay more attention to the feature factors that have a greater impact on the target, and suppress the expression of non-important features that have no obvious impact. Input the original sheet metal engineering drawing  $F$ , which is transformed into a featured image  $X$  through pooling + two convolution operations. The shallow feature map  $X \in \mathbb{R}^{C \times H \times W}$  is input to the CBAM module, which infers the channel attention map  $T_c(X)$  and the

planar spatial attention map  $T_s(X')$ , as shown in Figure 2. The overall attention process of the module is roughly as follows:

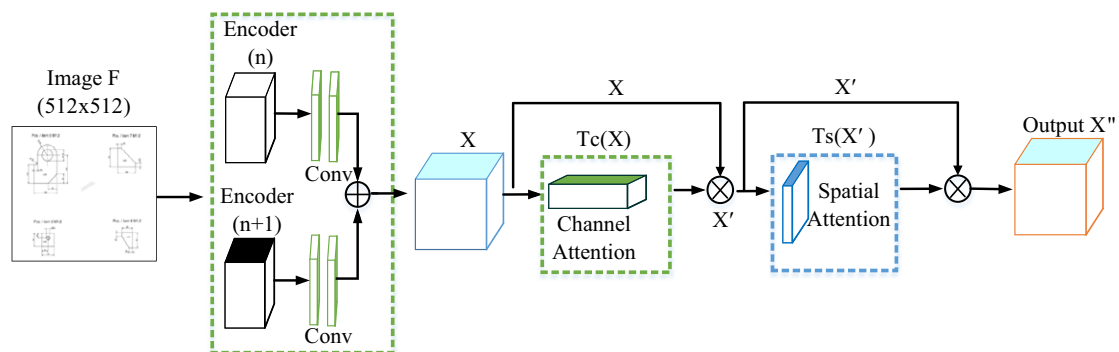
$$X' = T_c(X) \otimes X \quad (1.1)$$

$$X'' = T_s(X') \otimes X' \quad (1.2)$$

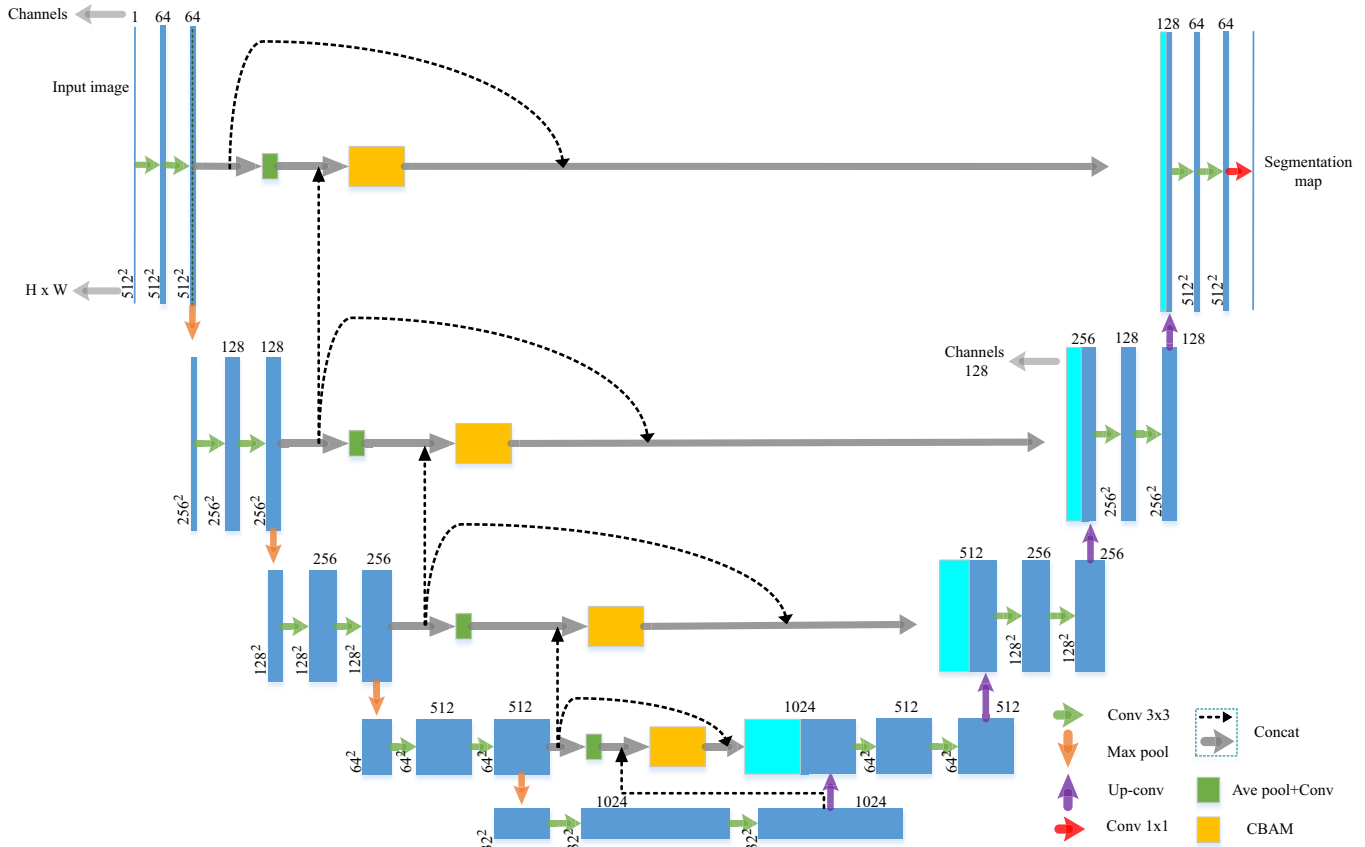
where  $\otimes$  means element-wise multiplication. In the operation process, the CBAM module can continuously obtain the 1D channel attention feature map  $T_c(X) \in \mathbb{R}^{C \times 1 \times 1}$  and the 2D spatial attention feature map  $T_s(X') \in \mathbb{R}^{1 \times H \times W}$  according to the input feature map  $X \in \mathbb{R}^{C \times H \times W}$ . In image feature extraction, the CBAM module assigns corresponding attention weights to the image features that have a greater influence on the target task (the attention weights in the spatial direction are propagated in the channel direction, and vice versa). The feature map  $X''$  marked with channel-spatial attention weights calculated and output by the CBAM module is finer than the image features output by traditional U-net. At the same time,  $X''$  is upsampled by 2x2 and fused with the low-dimensional feature cluster of the horizontal encoder, and the final feature  $F'$  is output. Image features  $F'$  are combined with decoder upsampled graph feature clusters to reduce global contextual semantic differences. The overall improved U-net network structure is shown in Figure 3.

### U-net attention module

The residual attention network (Wang et al., 2017) adopts pre-activated residual units ResNeXt (Xie et al., 2017) and Inception (Szegedy et al., 2017) as a two-branch parallel network structure stacked with residual attention modules. Zhou et al. (2016) and Hu et al. (2018) used average pooling to aggregate and count spatial information, respectively. The convolutional attention module uses inter-channel feature relationships to compress the spatial dimension of the input feature map to compute channel attention. Moreover, it is verified that the average pooling and max pooling simultaneously can improve the feature network's representation ability. What is mentioned here is that the traditional CBAM directly performs the maximum pooling and average pooling operations on the image input. CBAM uses average pooling and maximum pooling to fuse the spatial information of semantic features to generate two different global semantic space information expression features  $F_{ave}^c$  and  $F_{max}^c$ . In improving the



**Figure 2.** CBAM architecture. This module comprises a channel module and a spatial attention module consecutively. The encoder feeds the double-pooled and convolutional features into this module, and the CBAM generates global features with channel and spatial location information.  $(n)$  and  $(n + 1)$  respectively represent the encoders of different vertical layers of U-net.



**Figure 3.** The improved model overall architecture is proposed in this paper (input raw image pixel 512x512). Green squares represent average pooling and two convolution operations ((all abbreviated as 'Ave' in the ablation experiments for ease of writing)), Orange squares represent the CBAM attention mechanism module, and indigo squares represent feature cluster integration. Different colored arrows indicate different operations.

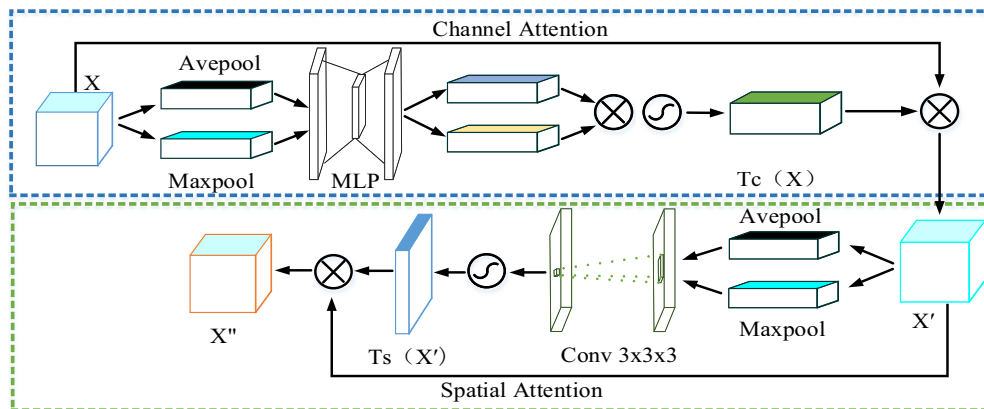
U-net model structure, we made a small change to this. Instead of directly inputting the sheet metal image  $F$  to the CBAM module, we input the features extracted and fused by the vertical encoder. Feed the fused features  $X$  into a shared multi-layer perceptron (MLP). Therefore, at this time, CBAM uses average pooling and maximum pooling operations to generate two different global semantic space information expression features:  $X_{ave}^c \in \mathbb{R}^{C \times 1 \times I}$  and  $X_{max}^c \in \mathbb{R}^{C \times 1 \times I}$ . At the same time, the shared multi-layer perceptron performs integration + ReLU operation on the input feature clusters to generate channel attention feature maps  $Tc(X) \in \mathbb{R}^{C \times 1 \times I}$ .  $Tc(X)$  contains the attention feature relations between the various channel axes of the input feature  $X$ . Then,  $Tc(X)$  is fused with the input feature  $X$  (Note:  $X$  is the image feature after the original image  $F$  has been pooled and convolved by the upper and lower encoders) to generate a feature map  $X'$ . According to the semantic difference between the Upper-encoder and Lower-encoder features, the convolution and linear rectification unit (ReLU) are used to continue to calculate the spatial information relationship of the feature  $X'$ , and generate the attention space feature map  $Ts(X')$ . The channel attention feature  $X'$  is merged with the spatial position attention feature  $Ts(X')$  to generate a globally informative feature  $X''$  with channel-spatial dual attention. The detailed calculation process of the attention module is as follows.

$$\begin{aligned} Tc(X) &= S(M(\text{Maxpool}(X)) + M(\text{Avepool}(X))) \\ &= S(W_2(W_1(X_{ave}^c)) + W_2(W_1(X_{max}^c))) \end{aligned} \quad (1.3)$$

where  $Tc(X)$  is a 1-dimensional channel attention image feature,  $S$  is a sigmoid activation function, and  $M$  represents a multi-layer perceptron (MLP) shared layer.  $\text{Maxpool}(X)$  and  $\text{Avepool}(X)$  are the secondary pooling operations of horizontal low-dimensional encoder features and vertical high-dimensional encoder features, respectively ( $X$  is the result of vertical maximum pooling and horizontal average pooling).  $W_1$  and  $W_2$  represent the shared weights of the input multi-layer perceptron (MLP). The output of the ReLU activation layer is  $W_1$ . The extraction process of spatial feature information in the CBAM structure is similar to that of channel feature extraction. The difference is that the sheet metal graphic feature  $F$  is input into the CBAM module as  $X$  after the convolution operation. The CBAM module still uses maximum pooling and minimum pooling to fuse image feature semantic information to generate two 2D images:  $X_{ave}^s \in \mathbb{R}^{I \times H \times W}$  and  $X_{max}^s \in \mathbb{R}^{I \times H \times W}$ . And perform a convolution operation on it to generate an image  $Ts(X')$  containing spatial feature information. (It is worth noting that we use three 3x3 convolutions instead of 7x7 convolutions in CBAM to reduce calculation parameters). The image spatial feature information is calculated as follows:

$$\begin{aligned} Ts(X') &= S(f_{3 \times 3 \times 3}(Tc([Avepool(X); \text{Maxpool}(X)]))) \\ &= S(f_{3 \times 3 \times 3}([X_{ave}^s; X_{max}^s])) \end{aligned} \quad (1.4)$$

$Ts(X')$  is a 2-dimensional spatial position information feature, and  $f_{3 \times 3 \times 3}$  represents three 3x3 convolution operations. Among



**Figure 4.** Principle of attention mechanism.  $X$  is used as input to the multi-layer perceptron (MLP), and the feature  $X'$  with channel attention information is generated through feature cluster multiplication and Softmax operation.  $X'$  output features  $X''$  with channel spatial information through a similar operation of the spatial attention module.

them, the spatial attention feature layer performs average pooling and maximum pooling operations on the channel attention feature layer  $T_c(X)$  during the calculation process, instead of inputting the welding graph  $F$ . The detailed structure of the attention module is shown in Figure 4. Experiments were conducted only on sheet metal engineering drawings of heavy equipment provided by MCC (a partner company). Vgg16 and ResNet50 are used as the backbone networks to validate the modeling performance of this paper's method, respectively. From the computational analysis results in Table 1, it can be concluded that the method proposed in this paper, in terms of average accuracy, is significantly better than both Vgg16 and ResNet50 as the baseline. In addition to that, the Vgg16 model using this paper's method is better than ResNet50 in terms of global segmentation effect.

This study improves the U-net segmentation network, the main purpose of which is to use artificial intelligence to apply it to the manufacturing process of heavy industry equipment welding engineering, so that it can automatically cut the entire sheet metal, thereby liberating labor and improving enterprise efficiency. Encapsulate the improved model after training in an integrated cutting and processing center with visual functions to realize the automatic cutting of sheet metal parts in the process of intelligent manufacturing. The specific process is shown in Figure 5. At

present, this research has been initially applied to the cutting processing center of China Metallurgical Equipment Corporation to realize a simple sheet metal cutting test.

## Experiments

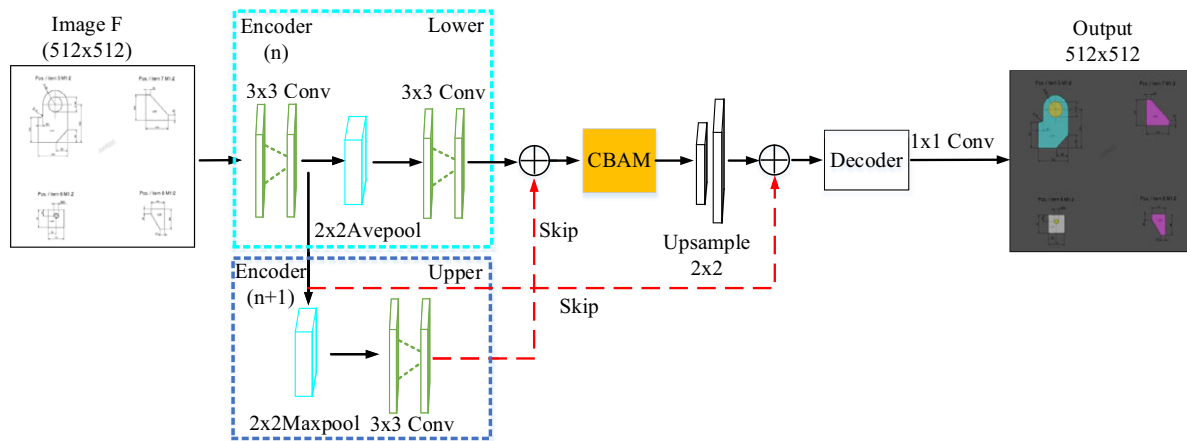
### Dataset

The training data set is a non-public engineering atlas of complex welding structures used in the manufacture of heavy industry equipment provided by the cooperative heavy-pressure riveting and welding company, as well as some public welding engineering atlases in the United States and Japan. Its quantity is shown in Table 2. It is well known that the size of the dataset directly affects the training results, and the network may overfit when there are few training samples. To avoid the problem of biased training results due to the small number of data sets, data set enhancement processing is performed on the provided data sets. First of all, this study uses manual annotation to select 600 welding engineering graphics provided by China Metallurgical Group, the United States, and Japan from the welding equipment engineering atlas collection. The dataset was augmented by cropping, mirroring, deflecting, adding noise, etc., resulting in a dataset of 4094 annotated engineering drawings.

**Table 1.** Training cost analysis and model mean accuracy comparison between Vgg16 and ResNet50 with different model structures

Model	Loss	mAP@0.5	mAP@0.70	Time(s)
Vgg16 (Simonyan and Zisserman, 2014)	0.0183	0.7652	0.6200	43.00 ± 0.761
Vgg16 + Ave-v1	0.0601	0.8113	0.7156	45.00 ± 0.238
Vgg16 + CBAM-v2	0.0211	0.9623	0.8130	146.00 ± 0.099
Vgg16 + Ave + CBAM-v3 (Ours)	0.0730	0.9992	<b>0.9004</b>	150.00 ± 0.021
ResNet50 (He et al., 2016)	0.0189	0.9591	0.4090	<b>17.00 ± 0.102</b>
ResNet50 + Ave-v1	0.1084	0.9604	0.5875	17.00 ± 0.368
ResNet50 + CBAM-v2	0.1248	0.9667	0.6433	40.00 ± 0.994
ResNet50 + Ave + CBAM-v3 (Ours)	0.0929	<b>0.9999</b>	0.6981	41.00 ± 0.006

Note: Same benchmark, bold font means excellent.



**Figure 5.** Improve the U-net model by cutting sheet metal specific contour mechanism. Segmentation extracts the specific unit of welding engineering graphics, and the cutting device automatically cuts the corresponding parts on the whole sheet metal relying on vision.

**Table 2.** Sources of welding engineering datasets and the number of datasets after data enhancement processing

Dataset	MCC	America	Japan	Total
Training set	1637	844	794	3275
Validation set	205	106	99	410
Test set	205	105	99	409
Total number	2047	1055	992	4094

### Evaluation metrics

$$IoU = \frac{\text{Area}(Pp \cap Pgt)}{\text{Area}(Pp \cup Pgt)} \quad (1.5)$$

$Pp$  is the prediction frame,  $Pgt$  is the ground truth frame, and  $IoU$  is the intersection area of the  $Pp$   $Pgt$  regions divided by the union area.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1.6)$$

TP This means that the sample is positive and the predicted value is positive.

FP This means that the sample is negative and the predicted value is positive.

FN This means that the sample is positive and the predicted value is negative.

TN This means that the sample is negative and the predicted value is negative.

Accuracy is used to judge the accuracy of the prediction results, the total number of correct predictions the total number of samples. There are two cases when the prediction is accurate: the sample is positive, the prediction is positive, and the sample is negative.

$$AP = \frac{\sum \text{Precision}}{N} \quad (1.7)$$

Precision is the sample precision, and  $N$  is the total number of samples.

$$mAP = \frac{\sum AP}{N_{\text{class}}} \quad (1.8)$$

AP is the average precision and  $N_{\text{class}}$  the number of classes. AP The average precision rate is the sum of the precision rates for each sample (of a particular category) divided by the total number of samples.  $mAP$  is the mean, and the average precision is the sum of the AP values of all categories divided by the number of categories (note:  $mAP$  in the table is the  $mAP$  of the validation set).

### Implementation details

This experiment is carried out on the environment framework environment of Anaconda3, using the GPU accelerated training method of CUDA parallel computing architecture. Vgg16 (Simonyan and Zisserman, 2014) serves as the backbone network for the whole structure, the input graph size is 512 x 512, and the optimizer is trained using the Adam optimizer with internal momentum = 0.9. Considering memory issues, a pre-trained model with a total number of Epochs of 100 is used during the experiments and frozen experiments are performed. The initial learning rate is set to 0.0001 and the minimum learning rate is 0.000001 in cos descent. Since this segmentation experiment has 9 classes, Dice loss is not used during the experiment. Model hyperparameters used for all experiments are shown in Table 3.

In addition, by combining the actual requirements and comparing the use of the Focal loss with better results, the main purpose is to reduce the weight of easily distinguishable samples and focus on samples that are difficult to distinguish. The experimental comparison results are shown in Table 4. When the network is trained to the 50th epoch, the network starts to load and evaluate the validation set. The K-fold cross-validation method ( $K = 5$ ) was used to validate the model as shown in Figure 6.

Sheet metal cutting techniques in the manufacture of heavy equipment must comply with international standards. The method in this paper is used to segment and extract specific unit contours from sheet metal engineering drawings, and the segmentation accuracy of the graphic boundaries meets the weld width and fineness required by MCC-Shaanxi Pressure Company. The technical requirements of the sheet metal welding drawing are shown in Table 5.

The performance of the improved welded graph U-net segmentation model was evaluated by using the Intersection Over Union (IoU), Accuracy (Accu), and Mean Prediction Precision (mAP) metrics. Its training loss curve is shown in Figure 7.

**Table 3.** Hyperparameter values are used for all training

Hyperparameter	Value
activation function	Relu
Learning_rate	0.0001
num_classes	9.0
backbone	Vgg16
input_shape	512
Freeze_Epoch	100
Freeze_batch_size	1.0
UnFreeze_Epoch	50.0
Unfreeze_batch_size	1.0
optimizer_type	adam
momentum	0.9
decay_type	cos
num_workers	1.0
freeze_layers	17.0
nbs	16.0

**Table 4.** The performance comparison results of various loss functions used by CBAM-U-net to deal with imbalanced datasets

Network	Loss function	Epoch	IoU	Accu
CBAM-U-net	CE Loss	100	0.8252	0.9810
	BCE Loss	100	0.8312	0.9793
	Poly Loss	100	0.8400	0.9902
	Focal Loss	100	<b>0.8472</b>	<b>0.9942</b>

Note: Same benchmark, bold font means excellent.

**Figure 6.** The welding engineering atlas adopts a K-fold cross-training verification process, the training set and verification sets are 4:1, and the stratification factor is K = 5.

During the experimental training process, the Loss plot verifies that the model structure of this paper is stable and there is no model collapse. To directly demonstrate the segmentation effect of our proposed U-net improved model on welding engineering drawings, the results are compared with the traditional U-net segmentation effect through ablation experiments. Its visualization effect is shown in Figure 8. By comparing the traditional U-net segmentation performance with the horizontal encoder average pooling + double convolution vertical encoder jump

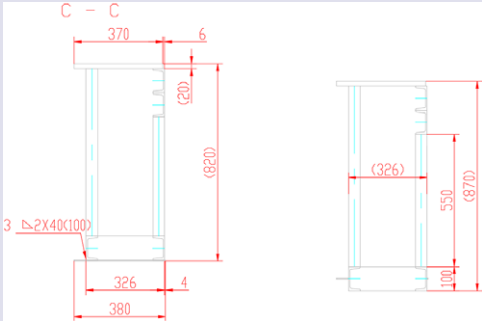
















structure fusion method proposed in this paper, a comparative experiment is carried out. After that, we continue to compare the traditional U-net segmentation performance with the crossbar encoder jump structure of the attention mechanism CBAM proposed in this paper. The results of the comparative experiments are shown in Table 6.

To avoid problems such as the increased variance of the estimated value and mean shift when the network model extracts features. In the U-net improved model, this paper proposes to use horizontal low-dimensional encoder average pooling and vertical high-dimensional encoder maximum pooling to better capture global feature information. After the encoder performs a double pooling operation, its horizontal encoder continues to perform two convolution operations to extract higher-dimensional spatial information of the image. The input image resolution is 512x512, and the encoder uses a repeated convolution operation of 3x3 (same padding), followed by a linear rectification unit (ReLU). The vertical high-dimensional encoder uses max-pooling with stride 2 of size 2x2, and the horizontal encoder uses equal-sized average pooling. Perform two 3x3 (same padding) convolution operations on the average pooled semantic features. The semantic feature clusters of the vertical high-dimensional encoder after max-pooling convolution are fused with the semantic feature clusters of the horizontal low-dimensional encoder after average pooling convolution. The fused feature clusters are input into the jump structure module of the CBAM attention mechanism to realize global context information fusion. The feature information output by the improved model added to the attention mechanism module is fused with the feature information of the low-dimensional encoder (this low-dimensional feature information does not perform any pooling operation). Finally, the dimensionality reduction of the fused semantic features is performed through 1x1 convolution through the jump structure, and the features with a resolution of 512x512 channels and 64 channels after upsampling by the decoder are fused again. Perform 3x3 convolution and 1x1 convolution on the integrated feature map to realize the mapping of each component feature vector class and achieve its segmentation effect.

To further verify the scientificity and effectiveness of the method proposed in this paper, we use CNN to continue to use this method to segment and extract the specific outline of the welding engineering drawing. Abandoning the jumping structure of the traditional U-net, the method proposed in this paper is applied to the traditional convolutional neural network to perform segmentation and extraction experiments on welding graphics. The model uses consecutive 3x3 convolution operations, and the last layer uses the principle of mapping to achieve the segmentation effect. By using double-pooling convolution operations and adding attention mechanism operations in the CNN network model, different CNN models are compared for segmentation performance experiments. Figure 9 shows the Loss diagram of the CNN model with dual pooling convolutional fusion and attention mechanism added simultaneously, from which it can be seen that the model has no collapse phenomenon and the structure is stable for training. The visualization result of the specific contour segmentation of the welding engineering drawing is shown in Figure 10.

Using the traditional multi-layer convolutional neural network as the baseline, the CNN network structure model is constructed using the method in this paper, and different CNN models are experimentally compared on the task of segmenting specific

**Table 5.** Technical requirements detail sheet

Welding parts technical requirements							
	Num	A0	A1	A2	A3	A4	
Size	$L \times B$	841x1189	594x841	420x594	297x420	210x297	
General technical requirements			JB/T5000.3				
Weld quality level not noted.			CS or CK				
Accuracy class			B.F				
Welding seam height			$K = 3 \text{ mm}$				
Annealing stress relief			Not involved				
Surface quality treatment			Sa2 ½				
Channel steel weld contact			$ 0^{+0.1} $				
Material			Q215 or Q235				
							
Line classification	Shape	Width (mm)					
		A0	A1	A2	A3	A4	GB/T 14689;
Thi		0.5	0.5	0.35	0.35	0.35	GB/T 14690;
Dash		0.18	0.18	0.13	0.13	0.13	GB/T 13362.4;
Thin		0.18	0.18	0.13	0.13	0.13	GB/T 13362.5;
Arrow		-	-	-	-	-	GB/T 17450;
Numer	1,2,3,4,5,6...	-	-	-	-	-	GB/T 16675.2;
Welding Annotation	GB/T 324–2008; GB/T4458.4–2003						
V-shaped weld			$C \approx \delta + 3$ ,	Weld seam width requirements for different welding methods. $C$ = width; $\delta$ = thickness(mm); $K$ = height;			
U-shaped weld			$C \approx 0.35\delta + 12.5$				
I-shaped weld			$C = \delta + 2$				
Triangular weld			$K = \delta + 2$				
Spot weld			-				
Groove weld			-				

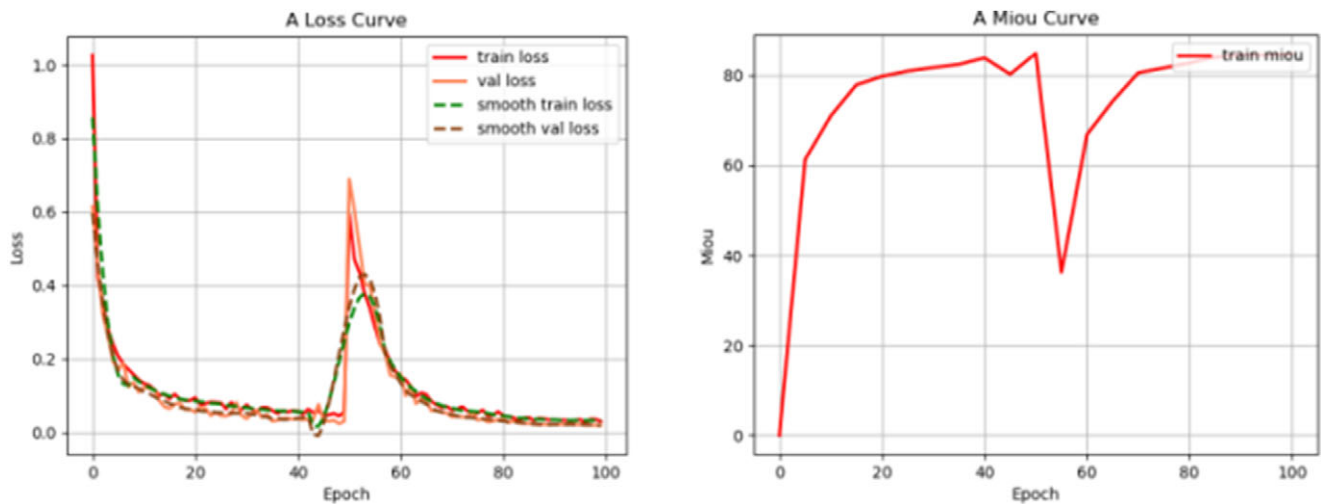
Note: Thick solid lines, thin solid lines, dashed lines, arrow lines, and number lines are denoted by 'Thi', 'Thin', 'Dash', 'Arrow' and 'Numer' are indicated.

contours in heavy equipment welding engineering drawings. The experimental details are shown in Table 7. It can be concluded from Table 7 that the method proposed in this paper has high accuracy in segmenting and extracting specific contours in heavy equipment engineering drawings.

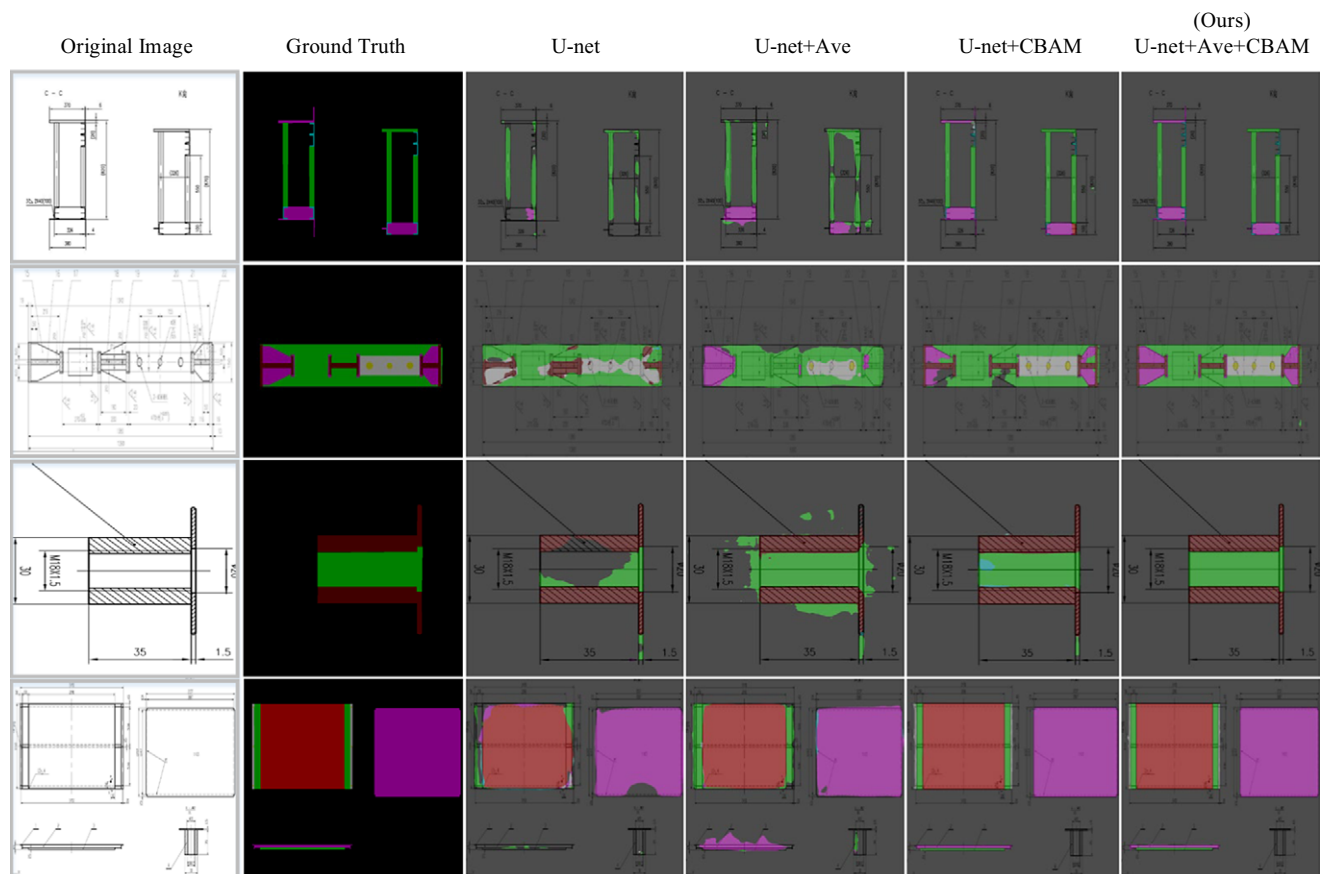
At the same time, ablation comparison experiments are carried out with this paper's method and different excellent segmentation techniques on the task of segmenting sheet metal engineering drawings in intelligent manufacturing. The experiments show that the method of this paper has excellent segmentation accuracy for

the segmentation of the contents of sheet metal engineering drawings for engineering manufacturing, and also fully verifies the scientificity of the method of this paper. The details of the SOTA experiments are shown in Table 8.

To further verify the scientific validity and rigor of the proposed method in this study, the CBAM-U-net, CBAM-CNN, U-net, and CNN models of Focal loss were used to classify multiple types of lines in welding engineering drawings, respectively. The confusion matrix is used to realize the visualization of multi-type line classification in welding engineering graphics, and the result is shown in



**Figure 7.** The loss curve graph during training and the 50th epoch model reaches a state of convergence. When training to 50 epochs, the network starts to unfreeze the evaluation model. The model will be reloaded from its original form, and fluctuations will have no effect.



**Figure 8.** Visual comparison of segmentation effects between different methods. The original input is a welded structure drawing, and the second column is the ground truth mask. Where 'Ave' is denoted as the average pooling and convolution operations as green squares in Figure 3, CBAM is the attention module, as shown in the orange court in Figure 3.

**Figure 11.** Column elements in the confusion matrix represent true label values for different types of lines, and rows represent true predicted values for different types of lines. Use “Thi,” “Thin,” “Dash,” “Arrow” and “Numer” to represent the thick solid line, thin solid line, dashed line, arrow line, and numbered line in the

content of the welding engineering drawing, respectively. From the classification visualization results, it can be concluded that the improved model CBAM-U-net, CBAM-CNN proposed in this paper improves the prediction performance of different types of lines relative to the traditional U-net, CNN network model, and the

**Table 6.** Comparison of welding engineering map segmentation by different methods, 'Ave' is denoted as average pooling and convolution operations, and CBAM is denoted as attention module

Method	IoU	mAP	Accu
Base (U-net) (Ronneberger and Fischer, 2015)	0.6262	0.6775	0.9937
Base +Ave	0.6324	0.6947	0.9605
Base+CBAM	0.7299	0.7549	0.9745
Base+Ave + CBAM	<b>0.8472</b>	<b>0.8684</b>	<b>0.9942</b>

Note: Same benchmark, bold font means excellent.

results are mainly reflected in the prediction effect of thick solid, thin solid, dashed and arrow lines.

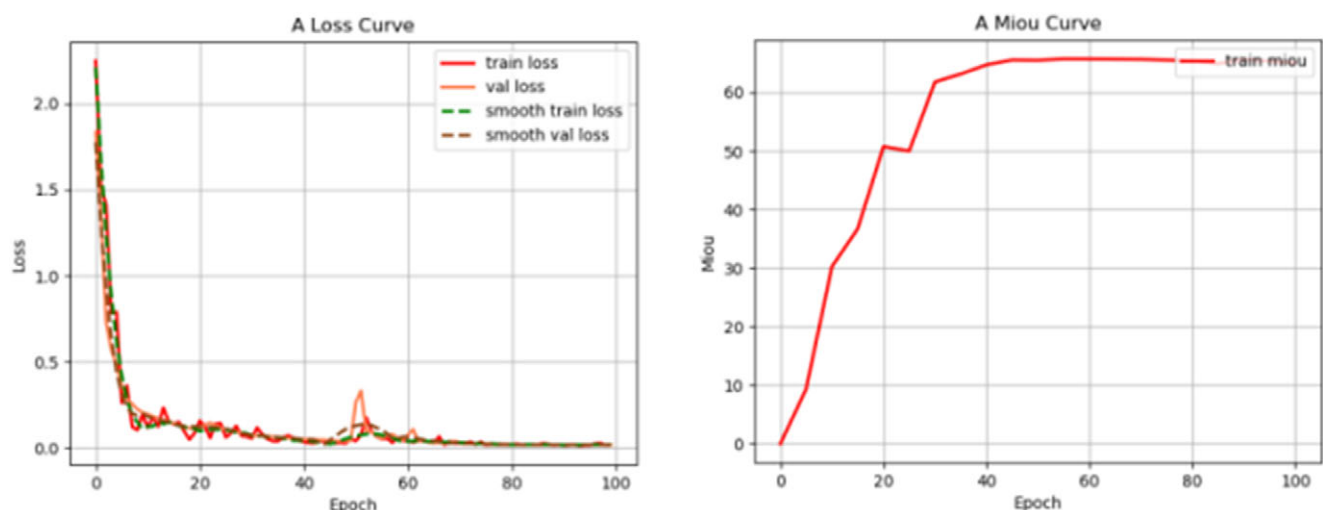
### Results and discussion

The results are shown in Table 6. The original U-net model is applied to the segmentation and extraction of specific units of welding structure engineering drawings, and the average accuracy of the intersection-over-union ratio (IoU) can reach 62.62%, and the category means (mAP) and accuracy (Accu) are 0.6775 and 99.37%, respectively. The third column in Figure 8 is the segmentation visualization result of the traditional U-net. It can be seen from the figure that the segmentation accuracy of the traditional U-net in the welding engineering-specific unit has some shortcomings. To solve this problem, this paper proposes to add average pooling + two convolution operations in the horizontal encoder jump structure, and at the same time proposes a vertical jump fusion of high-dimensional vertical encoder feature clusters and low-dimensional horizontal encoder feature clusters-model-one.

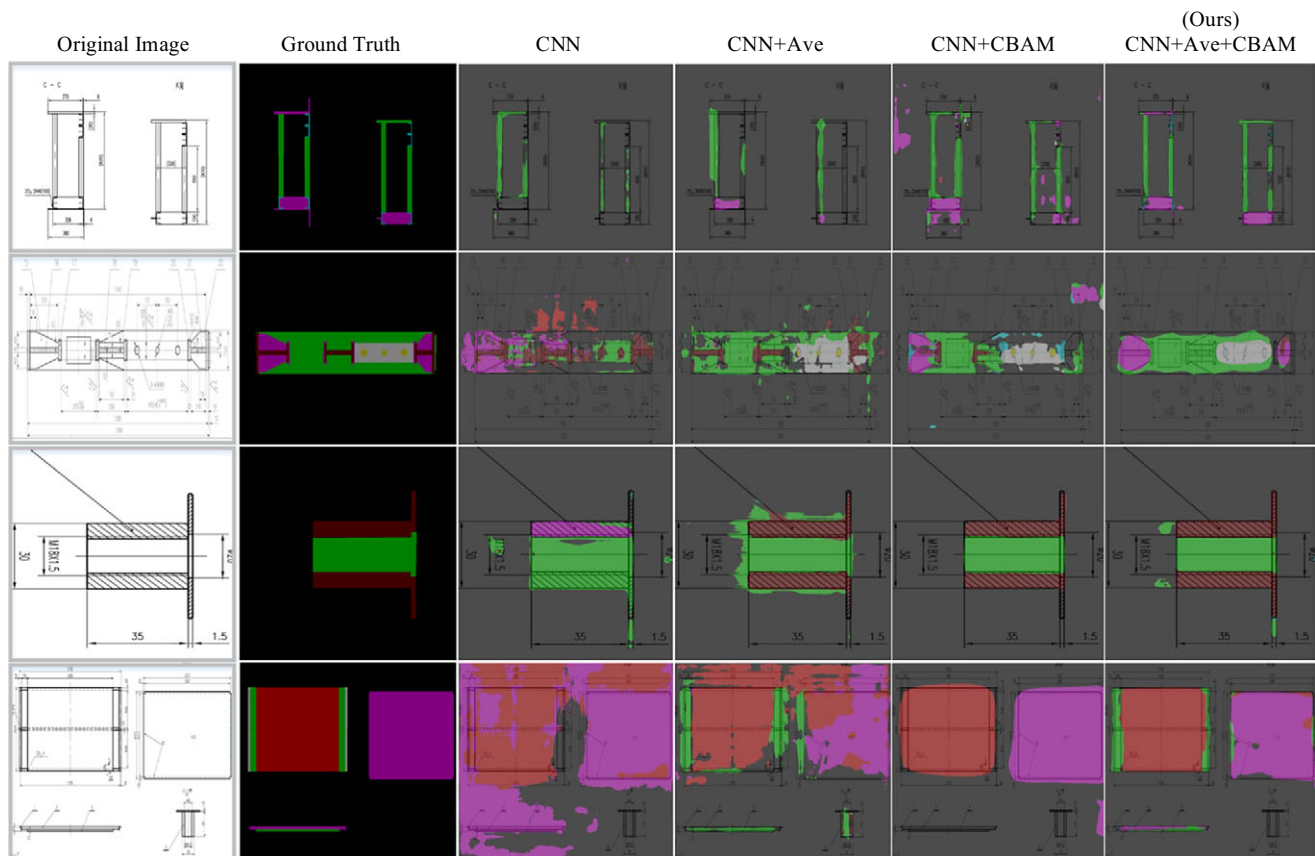
Model one adopts the double-pooling convolution jump structure of the horizontal encoder and the vertical encoder, which reduces the dimensional difference between the encoder and the decoder based on improving the fusion of global semantic features, and realizes the improvement of its segmentation performance. The intersection-over-union ratio (IoU) accuracy of

model one segmentation to extract specific units of welding engineering graphics is 63.24%, and the category means (mAP) and accuracy (Accu) are 0.6947 and 96.05%, respectively. However, as can be seen from Figure 8, model one still performs poorly. Immediately, an improved model of adding an attention mechanism to the traditional U-net jump structure was proposed-model two. Model two pays more attention to the channel and spatial information of global semantic features, and reduces the difference of global feature information to achieve better segmentation. The intersection-over-union ratio (IoU) accuracy of model two segmentation to extract specific units of welding engineering graphics is 72.99%, and the category means (mAP) and accuracy (Accu) are 0.7549 and 97.45%, respectively. To make the model better extract the global feature information, the improved model of average pooling, vertical jumping, and attention module is added to the traditional U-net horizontal jumping structure at the same time – Model three. The improved model three can not only reduce the global information ambiguity but also reduce the information dimension difference between the encoder and the decoder and greatly improve its segmentation performance as a whole. The intersection-over-union ratio (IoU) of the specific unit of welding engineering graphics extracted by model three segmentation is 84.72%, and the category means (mAP) and accuracy (Accu) are 0.8684 and 99.42%, respectively. All in all, the performance of the three improved models proposed in this paper for segmenting specific units of welding engineering graphics has been significantly improved compared with the traditional U-net. Among them, the IoU, mAP, and Accu of model three are improved by 22.10%, 19.09%, and 0.05%, respectively, compared with the traditional U-net in the segmentation task of the specific unit of the welding pattern. To prove the scientificity and rigor of the method proposed in this study, a series of multiple 3x3 convolutional networks was used to continue further verification. From the experimental results in Figure 10 and Table 7, it can be concluded that the method in this paper can effectively improve the specific unit segmentation performance of its model for welded structure engineering graphics.

This paper discusses the application of the U-net-based improved model to the segmentation and extraction task of



**Figure 9.** A graph of the loss curve of a continuous convolutional CNN. When the training has gone through 45 epochs, the model reaches the state of convergence.



**Figure 10.** Visualization of segmentation results for successive convolution operations. 'Ave' is represented as the average pooling and convolution operation of the green square in Figure 3, and CBAM is the attention module, as shown in the orange area in Figure 3.

**Table 7.** In the comparison of different methods for the segmentation results of specific welding engineering units, 'Ave' is expressed as the average pooling and convolution operation, and CBAM is the attention mechanism

Method	IoU	mAP	Accu
CNN(Base)	0.4143	0.5755	0.8867
Base+Ave	0.4633	0.6217	0.9130
Base+CBAM	0.5698	0.7101	0.9377
Base+Ave + CBAM	<b>0.6509</b>	<b>0.7294</b>	<b>0.9518</b>

Note: Same benchmark, bold indicates excellent.

specific units in welding engineering graphics. It mainly realizes the automatic cutting of sheet metal parts through artificial intelligence machine vision and improves the manufacturing efficiency of heavy industry equipment. However, there are still many deficiencies in the positioning of the cutting device, the calculation of the utilization of the sheet metal, and the cutting seam. For example, in the process of cutting sheet metal parts by machine vision, the cutting device cannot calculate the utilization rate of the whole sheet metal, and the positioning redundancy is large, which leads to material waste. In addition, the method in this paper is only applicable to the cutting of sheet metal for heavy equipment, because the welding process of heavy equipment has a large weld seam and low welding accuracy, so the cutting accuracy requirements are not high, and the current cutting

method can fully meet the precision of heavy equipment manufacturing. For high-precision segmentation tasks in the manufacturing industry, the ability of the segmentation model to extract semantic features and fuse semantic features needs to be further increased, while at the same time, the dimensional difference between low-dimensional and high-dimensional features needs to be further eliminated. The attention model with cyclic semantic extraction and fusion will likely be better applied to industrial high-precision segmentation tasks.

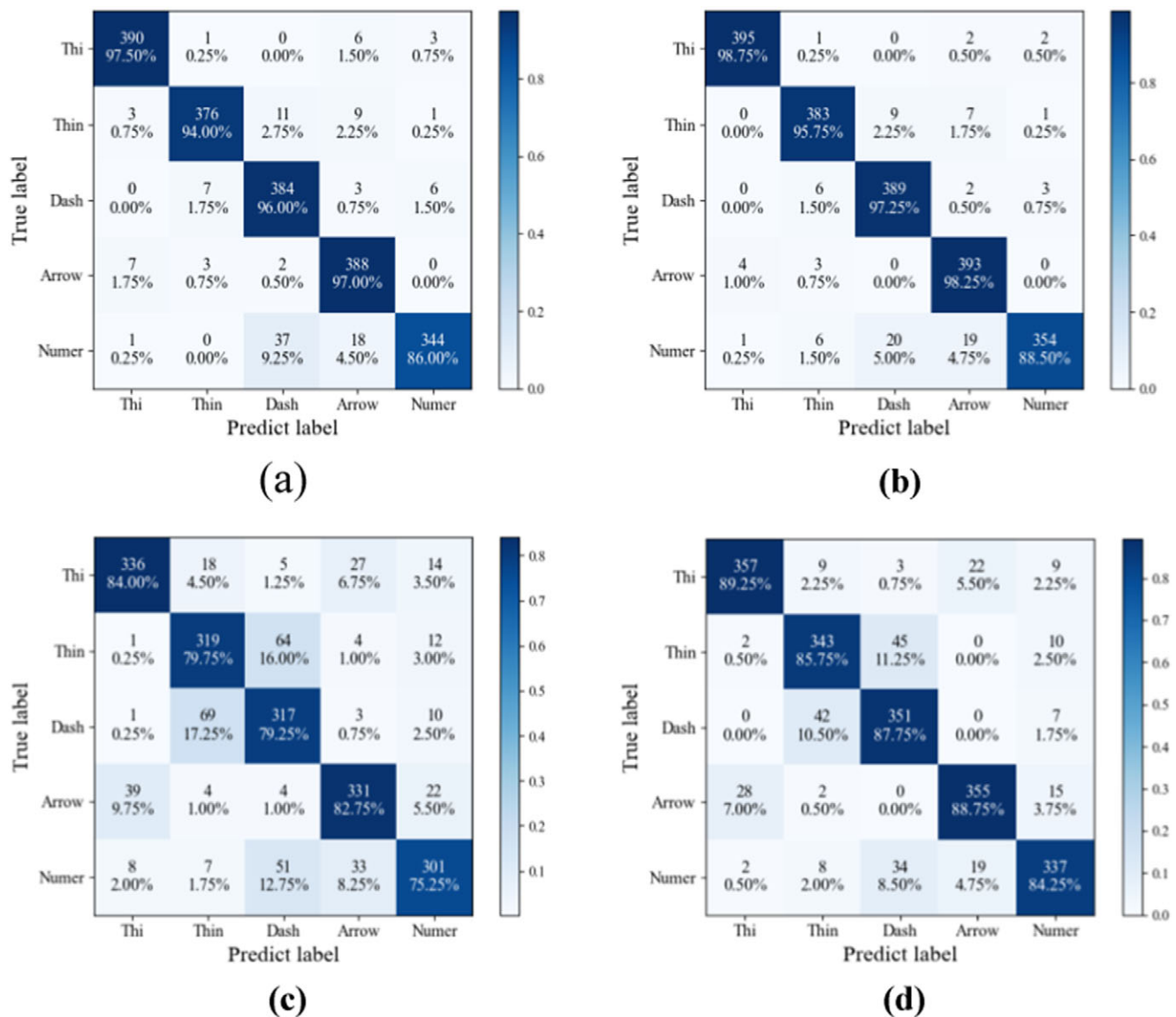
## Conclusion

Deep learning image segmentation technology has achieved excellent performance in many engineering fields. In this study, it is proposed to use the U-net network to realize the segmentation and extraction of specific units of welding structural engineering graphics in heavy industrial equipment manufacturing, so that the cutting device can automatically cut sheet metal parts by machine vision, thereby improving manufacturing efficiency. Based on the research and analysis of the existing U-net improved model, we propose to add the CBAM module and design the double pooling jump model structure of the upper and lower encoders to realize the global semantic fusion of image features. Not only that, this paper performs two convolution operations on the semantic feature clusters after double pooling to reduce the dimensional difference between the encoder and decoder. The proposed method is trained and validated on a dataset of engineering graphics of complex welded structures.

**Table 8.** Experimental comparisons and analyses have been carried out using the method of this paper and the current state-of-the-art segmentation technique (SOTA), and the experimental results have been analyzed for different sets of sheet metal welding project drawings

Method	Param (M)	MCC			America & Japan		
		IoU	mAP	Accu	IoU	mAP	Accu
U-net (Ronneberger and Fischer, 2015)	7.2	0.6210	0.6900	0.9959	0.6289	0.7021	0.9900
UCTransNet (Wang et al., 2022)	33.0	0.7996	0.8011	<b>0.9991</b>	0.8101	0.7967	0.9990
U-net++ (Zhou et al., 2018)	37.4	0.6573	0.7103	0.9910	0.6577	0.7100	0.9737
TransUNet (Chen et al., 2021)	52.0	0.7756	0.7840	0.9896	0.6570	0.7840	0.9894
Swin-UNet (Cao et al., 2022)	41.5	0.7814	0.7911	0.9900	0.7814	0.8010	0.9917
DCSAU-Net (Xu et al., 2023)	2.1	0.7207	0.7500	0.9601	0.5122	0.7778	0.9763
ICUnet++ (Li et al., 2023)	42.9	0.6855	0.7332	0.9713	0.4131	0.7479	0.9901
U-net + Ave-v1	10.4	0.6616	0.6881	0.9663	0.5912	0.6994	0.9609
U-net + CBAM-v2	36.1	0.7408	0.7600	0.9906	0.7111	0.7900	0.9855
U-net + Ave + CBAM-v3( <b>Ours</b> )	37.0	<b>0.8519</b>	<b>0.8733</b>	0.9953	<b>0.8224</b>	<b>0.8605</b>	<b>0.9940</b>

Note: Same benchmark, bold indicates excellent.



**Figure 11.** Comparison of confusion matrix results between U-net and CBAM-U-net models ((a) U-net, (b) CBAM-U-net(Ours), (c) CNN, (d) CBAM-CNN).

Experimental results show that our proposed improved model outperforms currently existing state-of-the-art segmentation techniques(SOTA) in segmenting specific cells of welded structural engineering drawings.

**Data availability statement.** All data that support the findings of this study are included within the article (and any supplementary files).

**Author contribution.** All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by [Zhiwei Song], [Hui Yao], [Tian Dan], and [Gaohui Zhan]. The first draft of the manuscript was written by [Zhiwei Song] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding statement.** This paper is supported by the Shaanxi Provincial Innovation Capacity Support Plan: Shaanxi Provincial Bearing Digital Design and Monitoring Technology Innovation Service Platform Project (2022PT-02). Computing resources are provided by the Institute of Advanced Manufacturing, Xi'an Technological University.

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

## References

- Ablameyko SV and Uchida S (2007) Recognition of engineering drawing entities: review of approaches. *International Journal of Image and Graphics* 7(4), 709–733.
- Alom M Z, Hasan M, Yakopcic C, et al. (2018) Recurrent, residual convolutional neural network based on u-net (r2u-net) for medical image segmentation[J]. arXiv preprint [arXiv:1802.06955](https://arxiv.org/abs/1802.06955).
- Badrinarayanan V, Kendall A and Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(12), 2481–2495.
- Cao H, Wang Y, Chen J, et al. (2022) Swin-Unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, pp. 205–218.
- Cao X and Lin Y (2021) Cagnet: Crossing aggregation network for medical image segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 1744–1750.
- Chen J, Lu Y, Yu Q, et al. (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint [arxiv:2102.04306](https://arxiv.org/abs/2102.04306).
- Chen LC, Papandreou G, Kokkinos I, et al. (2017a) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4), 834–848.
- Chen LC, Papandreou G, Kokkinos I, et al. (2014) Semantic image segmentation with deep convolutional nets and fully connected CRFs. *Computer Science* 4, 357–361.
- Chen LC, Papandreou G, Schroff F, et al. (2017b) Rethinking atrous convolution for semantic image segmentation. arXiv preprint [arxiv:1706.05587](https://arxiv.org/abs/1706.05587).
- De P, Mandal S, Bhowmick P, et al. (2016) ASKME: Adaptive sampling with knowledge-driven vectorization of mechanical engineering drawings. *International Journal on Document Analysis and Recognition (IJ DAR)* 19, 11–29.
- Favi C, Campi F, Germani M, et al. (2022) Engineering knowledge formalization and proposition for informatics development towards a CAD-integrated DfX system for product design. *Advanced Engineering Informatics* 51, 101537.
- He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Hou W, Wei Y, Guo J, et al. (2017) Automatic detection of welding defects using deep neural network. *Journal of Physics: Conference Series* 933(1), 012006.
- Hu J, Shen L and Sun G (2018) Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7132–7141.
- Ibtehaz N and Rahman MS (2020) MultiResUNet: Rethinking the U-net architecture for multimodal biomedical image segmentation. *Neural Networks* 121, 74–87.
- Lau SLH, Chong EKP, Yang X, et al. (2020) Automated pavement crack segmentation using U-net-based convolutional neural network. *IEEE Access* 8, 114892–114899.
- Li L, Zhang W, Zhang X, Emam M, et al. (2023) Semi-supervised remote sensing image semantic segmentation method based on deep learning. *Electronics* 12(2), 348.
- Li L, Qin J, Lv L, Cheng M, et al. (2023) ICUnet++: An inception-CBAM network based on Unet++ for MR spine image segmentation. *International Journal of Machine Learning and Cybernetics* 14(10), 3671–3683.
- Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3431–3440.
- Lu J, Ou C, Liao C, et al. (2021) Formal modeling of a sheet metal smart manufacturing system by using petri nets and first-order predicate logic. *Journal of Intelligent Manufacturing* 32, 1043–1063.
- Lu J, Ren H, Shi M, et al. (2023) A novel hybridoma cell segmentation method based on multi-scale feature fusion and dual attention network. *Electronics* 12(4), 979.
- Ma L and Yang J (2024) Adaptive recognition of machining features in sheet metal parts based on a graph class-incremental learning strategy. *Scientific Reports* 14(1), 10656.
- Madsen DA and Madsen DP (2016) *Engineering Drawing and Design*. Cengage Learning.
- Mohan S and Bhattacharya S (2022) Attention W-net: Improved skip connections for better representations. In *26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 217–222.
- Mubashar M, Ali H, Grönlund C, et al. (2022) R2U++: A multi-scale recurrent residual U-net with dense skip connections for medical image segmentation [J]. *Neural Computing and Applications*, 1–17.
- Pan M and Rao Y (2009) An integrated knowledge-based system for sheet metal cutting-punching combination processing. *Knowledge-Based Systems* 22(5), 368–375.
- Ronneberger O and Fischer P (2015) U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, pp. 234–241.
- Simonyan K and Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint [arxiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Sun K, Xiao B, et al. (2019) Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5693–5703.
- Szegedy C, Ioffe S, Vanhoucke V, et al. (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1)*.
- Tabernik D, Šela S, Skvarč J, et al. (2020) Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing* 31(3), 759–776.
- Tovey M (1989) Drawing and CAD in industrial design. *Design Studies* 10(1), 24–39.
- Vaswani A, Shazier N, Parmar N, et al. (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30.
- Wang F, Jiang M, Qian C, et al. (2017) Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3156–3164.
- Wang H, Cao P, Wang J, et al. (2022) Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(3), 2441–2449.
- Wang Z, Zou N, Shen D, et al. (2020) Non-local u-nets for biomedical image segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(04), 6315–6322.
- Woo S, Park J, Lee JY, et al. (2018) Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 3–19.

- Xie S, Girshick R, Dollár P**, et al. (2017) Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1492–1500.
- Xu Q, Ma Z, Duan W**, et al. (2023) DCSAU-net: A deeper and more compact split-attention U-net for medical image segmentation. *Computers in Biology and Medicine* **154**, 106626.
- Zhang W, Joseph J, Chen Q**, et al. (2024) A data augmentation method for data-driven component segmentation of engineering drawings. *Journal of Computing and Information Science in Engineering* **24**(1), 011001.
- Zhao H, Shi J, Qi X**, et al. (2017) Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2881–2890.
- Zhou B, Khosla A, Lapedriza A**, et al. (2016) Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2921–2929.
- Zhou Z, Rahman Siddiquee MM, Tajbakhsh N**, et al. (2018) Unet++: A nested U-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer, pp. 3–11.