Semantic Segmentation Algorithm of Animal Husbandry Image Based on an Improved U-Net Network

Jia Li, Jinjing Zhang, Fengjiao Jiang*
Department of Intelligent Agricultural Engineering,
Shanghai Vocational College of Agriculture and Forestry, Shanghai 201699, China

***Abstract*—Due to the limitations of unclear edges and fuzzy features in image segmentation tasks, this study proposes an enhanced U-Net semantic segmentation network utilizing the local and global fusion attention module in response to the drawbacks of fuzzy features and unclear edges in image segmentation tasks. Firstly, a feature extraction module combining convolution and Transformer is introduced in the bottleneck layer, so that the network can fully simultaneously capture local and global features, and effectively promote the fusion of local and global features. Secondly, the CBAM attention module is added to the skip connections between the encoder and decoder. Finally, the output feature map is processed using the ASPP module to enhance focus on target features and improve segmentation performance. Experiments conducted on four animal husbandry segmentation datasets show that the LCA_Net model proposed in this study achieves an IoU score of 90.19% and a Dice score of 94.83%, compared with U-Net and other mainstream segmentation networks, it has improved. This study offers effective technical support for advancing aquaculture status monitoring and lays a foundation for further development in this field.**

***Keywords*—Machine vision; semantic segmentation; feature fusion; attention mechanism**

I. Introduction

Image segmentation is the accurate extraction of target regions of interest from an image, which refers to assigning each pixel in the image to a different category in order to achieve semantic understanding and region recognition of the image. Traditional image segmentation techniques [1] have low efficiency and some defects in the segmentation results, which cannot achieve the expected results. The rapid improvement in computer hardware performance encourages the rapid development of deep learning technology. Deep learning-based methods have achieved excellent results in image segmentation. Its excellent feature extraction and expression ability improve segmentation accuracy and speed, which is superior to traditional machine learning and computer vision approaches [2].

A convolutional neural network significantly enhances image segmentation algorithms. Its end-to-end pixel-level image segmentation networks have created applications in the field of semantic segmentation. These networks can classify images at the pixel-level [3]. Later research, such as SegNet [4], improved the accuracy and precision of segmentation by decoding the feature index generated by the pooling layer, and further promoted the development of semantic segmentation technology. In 2015, the U-Net model was proposed [5], and its unique symmetric structure and excellent performance were quickly known. It has strong adaptability to medical image segmentation tasks by retaining information at different levels through skip-connections, and has been widely used in a variety of different segmentation tasks. Subsequently, Zhou et al. [6] optimized the U-Net as a UNet++ version and solved the semantic gap problem caused by direct connections in the original U-Net. Deeplab series [7-9] used hole convolution and pooling modules with varying dilation rates to obtain more contextual information and improved network performance.

In 2018, Zhou et al. [10] used the multi-scale watershed segmentation algorithm to segment the sheep from the images collected in the real and complex breeding environment. In 2020, Zhang [11] proposed a horse body segmentation method based on YOLACT, and segmented the horse body edge contour through edge detection. Qin et al. [12] focused on fish body image segmentation technology, which is a method based on object detection and edge assistance, and the method achieved remarkable results in fish image segmentation. This method not only significantly improves the accuracy of image segmentation, but also provides important support for the development of animal husbandry intelligent technology [13].

Singh et al. [14] proposed an improved DeepLabV3+ CNN model, which demonstrated significant accuracy in cattle body part segmentation. Feng et al. [15] introduced an improved DeepLabV3+ network segmentation model, achieving higher segmentation accuracy for cattle regions. Xie et al. [16] proposed a multi-scale dual-attention U-Net method for detecting sheep hind limb segmentation. These methods perform well in segmentation tasks for specific livestock species. However, all of these approaches focus on segmentation for a single species. In contrast, our proposed network model is capable of adapting to multiple livestock species, offering greater versatility and broader application potential.

Although numerous image segmentation network models exist, [17-21] traditional convolutional networks are limited by their receptive field size when handling image segmentation tasks with intricate details. This limitation reduces their ability to effectively capture global information, making it challenging to achieve optimal results. In addition, it also faces problems such as occlusion caused by complex background, changes in

*Corresponding author.

lighting conditions, and overlap. This study proposes an image semantic segmentation algorithm that improves the U-Net local-global attention network. While individual components like Transformers or attention mechanisms have been explored in isolation, the novelty of our LCA_Net lies in the synergistic integration strategy. We strategically position these modules to address specific shortcomings of the U-Net at different stages of the network, creating a cohesive pipeline that systematically enhances both local-global context modeling and feature refinement. On four different animal datasets, the experimental results show that the network performs more accurately than other popular network models, which is of some reference significance for the research in image segmentation.

II. Network Architecture

Fig. 1 shows an improved network model based on the U-Net proposed in this study, which has a symmetrical encoder-decoder structure.

Three main reconstructions were carried out on the original U-Net structure to form the LCA-Net structure proposed in this study. First, in order to generate feature maps with varying resolutions, the input image's features are progressively downsampled during the encoder stage, reducing spatial resolution. A local global feature fusion module is designed into the bottleneck layer of the encoder and decoder link, which efficiently combines local and global information. In addition, in order to reduce the redundancy of information between skip connections, an attention mechanism module is filled in. In the decoding phase, the high- and low-level features are spliced and upsampled at the same time. Finally, the image pixels are classified through the ASPP module and the Softmax activation function to obtain the segmentation results.

The proposed LCA-Net's design is grounded in three principles: 1) Hierarchical feature interaction: CSWin captures both local texture details and global shape context, overcoming the U-Net's limited receptive field; 2) Attention-guided fusion: CBAM suppresses spatially redundant features before skip-connections, mitigating semantic gaps; 3) Scale-aware decoding: ASPP complements CSWin by explicitly modeling multi-scale object variations. This layered approach fundamentally differs from prior works that apply these modules in isolation.

A. Local Global Fusion Module

Accurately distinguishing the target pixel from the background pixel in an image can be difficult, requiring automatic segmentation and feature extraction at both local and global scales to capture remote interactions. Therefore, the work of this study focuses on adding a local-global attention fusion module into the bottleneck layer of the U-Net architecture, and uses convolution and Transformer to transfer different information to the feature map, enabling the interaction and fusion of local and global information. Fig. 2 illustrates the structure of the local-global fusion module.

Downsampling is applied in one of the execution paths to obtain the LGAF input features, and convolution is used to learn the weights of the convolutional kernels. The local feature in the input image x is extracted to obtain a local receptive field, and then the convolution feature layer is normalized (LN) [22]. The stability of the algorithm has been improved, and the scale differences from different routes have been reduced. After the above operations are performed in turn, the X_L representing the aggregated local detail feature is obtained, as shown in Eq. (1):

$$X_L = LayerNorm(Conv(X)) \quad (1)$$

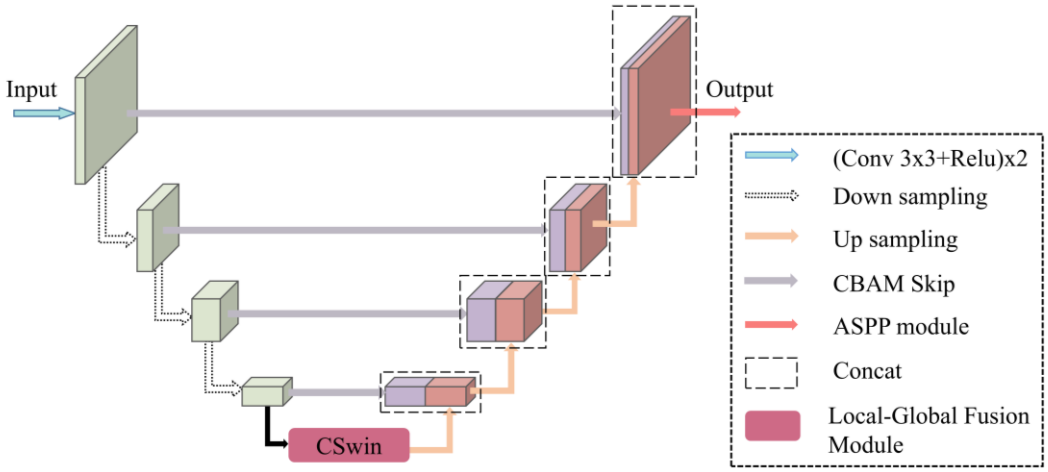


Fig. 1. Architecture of the proposed LCA-Net.

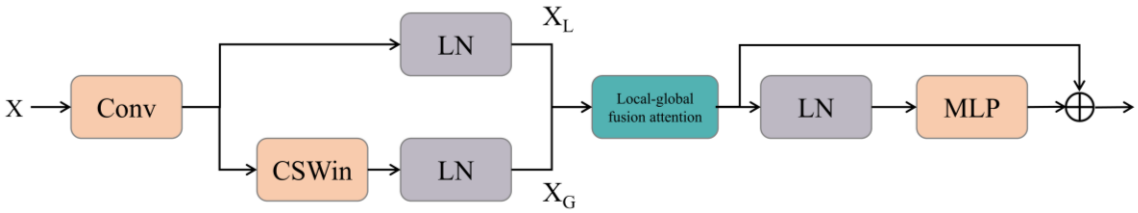


Fig. 2. Local-global attention fusion module.

where, the input feature is denoted by X, the convolution operation by Conv, and the layer normalization by LayerNorm.

In another execution path, self-attention calculation is performed using a convoluted feature map. The Cross-Window converter block (CSWin block) allows for the perception and interaction of global and non-local information [23]. Fig. 3 illustrates the structure of the CSWin block.

After processing previous information, the feature map X_G obtained by this module has its extracted global receptive field, and its specific calculation is shown in Eq. (2):

$$X_G = LayerNorm(CSWin(Conv(X))) \quad (2)$$

The local-global attention module (as shown in Fig. 4) simultaneously accepts and effectively fuses two types of parameter information: one containing local information (X_L) and the other containing global information (X_G).

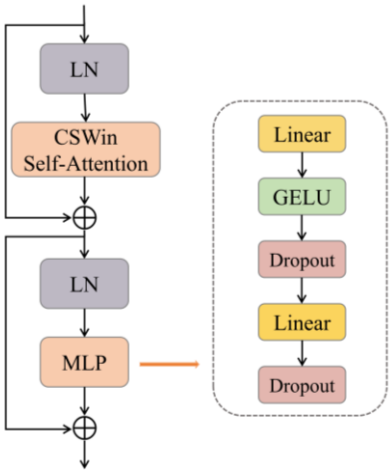


Fig. 3. CSWin module structure.

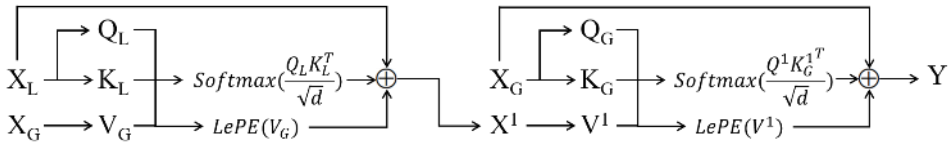


Fig. 4. Local-global attention module.

To calculate attention scores, three vectors-query, keyword, and value vector-with sequential features are calculated. The CSWin-Transformer is employed to process the input features X_L and X_G, and the process is shown in Eq. (3):

$$X^1 = CSWin\text{-}Transformer(X_L, X_G) + X_L \quad (3)$$

Finally, the CSWin-Transformer operation is reapplied to further enhance the global information. By strengthening the global information fusion of the feature map, the capture efficiency of long-distance dependency is improved, and the overall contour and boundary of the object segmentation are clearly recognized. Where V^1 is given by X^1, and Q and K are transformed from X_G. Output Y is represented by Eq. (4):

$$Y = CSWin\text{-}Attention(X^1, X_G) + X_G \quad (4)$$

B. Convolutional Block Attention Module

Typically, during the feature fusion process of the traditional U-Net model, simple cascade operation is involved, but there is no clear selection and enhancement mechanism in the face of key features. This may lead to the over dependence of irrelevant features on the model, and thus the accuracy of segmentation results is limited. Each encoder layer generates a feature map comprising both valuable information and redundant or irrelevant data. However, simple channel compression techniques may lead to the loss of important channel-wise information within these feature maps. To minimize the influence of the background area and make the model focus more on the target segmentation objects, the CBAM attention mechanism [24] is introduced into the skip connection of the network model designed in this study, which can dynamically adjust the weight and filter the effective feature channel. As a result, the model learns the relevant representations more efficiently. Fig. 5 illustrates the CBAM attention module.

Channel attention (CA) and spatial attention mechanism (SA), the two primary components of the CBAM module, are in charge of capturing dependence of channel and space, respectively. The channel and space feature weights can be dynamically changed by combining the two.

In order to compress the feature information, the channel attention module first applies the max pooling and average pooling operations to each channel for the feature map with the input shape of (B, C, H, W). This approach efficiently enlarges the receptive field of the convolutional network while preserving the number of feature channels, producing an output tensor of shape (B, C, 1, 1). The multi-layer perceptron's input should then be the outcomes of the two pooling operations. The multi-layer perceptron can increase the significance of a particular channel by learning parameters and using the sigmoid nonlinear activation function to assign corresponding feature weights to each channel.

For the spatial attention mechanism, the weight assignment methods of each channel are different. It only weights the local part, and then performs the max pooling, average pooling, and a 7 × 7 convolution operation on the channels within the same spatial area. Finally, the weights for each channel are obtained through the nonlinear activation of Sigmoid.

The following are the steps in the CBAM module's calculation:

The input feature map first generates the characteristic layer F_C with channel weight through the CA module, as shown in Eq. (5):

$$Fc = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (5)$$

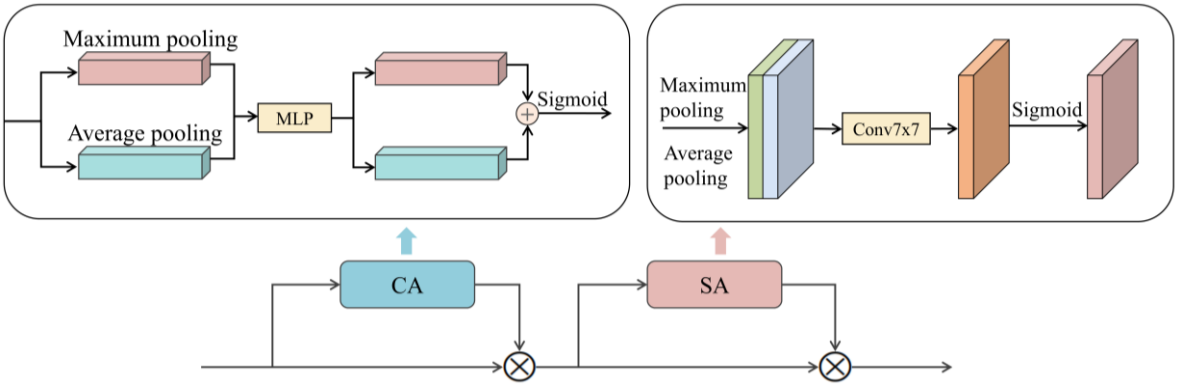


Fig. 5. CBAM attention module.

where, σ denotes the Sigmoid activation function. The number of multi-layer perceptrons is indicated by MLP. Avgpool and Maxpool refer to the global average and max pooling processes, respectively, and F represents the input feature map.

F1 is formed by multiplying the input original feature map F by the channel's weight F_C, which is obtained by F through the CA procedure, as shown in Eq. (6):

$$F_1 = F_C \bullet F \tag{6}$$

Spatial attention is applied to average pooling and max pooling along the channel for the feature map with the input shape of (B, C, H, W). It can compress the channel dimension of the input tensor to 1, then the global channel information integrates into a single channel characteristic graph that has the shape of (B, 1, H, W), and effectively extracts the spatial relationship. In order to allow information interaction between different spatial descriptors, the spatial merged features are connected. The Sigmoid activation function is then employed to acquire the spatial features after the combined features have been transformed into a spatial feature map using the convolution layer Conv, as shown in Eq. (7):

$$F_S = \sigma(Conv([SAavg + SAmax])) \tag{7}$$

where, SAavg and SAmax are the average and maximum pooled spatial feature descriptors. Eq. (8) represents the CBAM module's final output Y.

$$Y = F_1 \bullet F_S \tag{8}$$

The correlation of feature channels is greatly enhanced by this advanced dual attention mechanism, allowing each channel to more effectively capture the most relevant features of its task. Simultaneously, this approach notably improves feature expression, which increases the model's accuracy in understanding and processing the input data. The CBAM module of the skip connection part enhances the feature interaction. The model's accuracy and representational capacity are enhanced by reducing the influence of noise and redundant information through the combination of channel attention and spatial attention.

C. ASPP Module

ASPP was first proposed by Chen et al. [7] in Deeplabv2. Its idea comes from spatial pyramid pooling. Its purpose is to enhance the recognition ability of the network for targets of different scales through pooling operations of different scales. ASPP is a combination of a cavity convolution and a spatial pyramid pooling layer. Specifically, convolutions with different dilation rates are applied within a single branch to extract features. This approach enables the precise capture of multi-scale contextual information through receptive fields of varying sizes, which are subsequently fused to produce the final feature map. It ensures that the receptive fields are increased while maintaining the image resolution, avoiding the problem of a large amount of calculation in traditional convolution.

To more effectively capture and preserve edge detail features and enhance the model's segmentation capability across various target scales, this study uses 8, 12, and 16 as the hole convolution of expansion coefficient, and removes the features obtained by pooling branches. Fig. 6 displays the ASPP structure, which has been optimized.

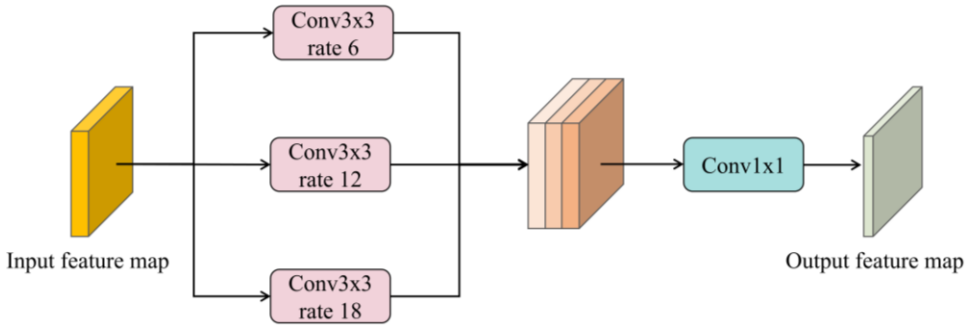


Fig. 6. ASPP module.

III. Experimental Configuration

A. Dataset

All data sets were collected in real animal husbandry environments, including horses and pigeons in outdoor pastures during the day, pigs and cattle in indoor pens. The original image is collected at a resolution of 1920 × 1080 pixels, and then adjusted to 256 × 256 for model input. The label uses LabelMe software to label the target contour to ensure pixel-level accuracy.

The experimental dataset used in this experiment is divided into four parts, which are horse images, pigeon images, cattle images, and pig images. Each image of the horse dataset has been professionally labeled, including the precise contour information of the horse, covering 327 high-resolution images. There are three sections to the dataset: 209 training sets, 53 verification sets, and 65 test sets.

The public pigeon dataset contains 122 high-resolution images and corresponding tags, including 78 training sets, 20 validation sets and 24 test sets. The cattle breeding dataset contains 200 images and corresponding labels, including 128 training sets, 32 validation sets and 40 test sets. The pig farm breeding dataset contains 501 images and corresponding images, including 320 for training, 81 for verification and 100 for test.

B. Experimental Environment and Parameter Setting

The NVIDIA GeForce RTX 3080 Ti GPU, Windows 11 operating system, 16GB of RAM, and Pytorch2.0.1 framework with Python 3.8 programming language constituted the deep learning environment. To ensure the determinism and reproducibility of our experiments, a fixed random seed was used throughout the study for parameter initialization and data shuffling. We employed the Dice loss function as our segmentation loss during the experiment, with a batch size of 8. For data augmentation, we applied a pipeline consisting of random horizontal flip, random vertical flip, random rotation within ±15°, and random brightness/contrast adjustment during the training phase. Additionally, Batch Normalization (BN) was utilized after each convolutional layer and before the ReLU activation function. Dropout was not employed in our architecture.

In this study, all experiments under the same dataset use the same loss function and parameter settings for training, in which the initial learning rate for horse segmentation images is set to 0.0001, with 100 training epochs. For pigeon segmentation images, the initial learning rate is 0.001, and the training consists of 200 epochs. Pig and cattle breeding images have an initial learning rate of 0.0001 and undergo 50 training epochs. Other experimental parameters are consistent with those mentioned above.

C. Evaluation Index

The evaluation indexes chosen for this study are Dice Similarity Coefficient, Precision, Recall, and Intersection over Union (IoU).

The precision indicates the percentage of accurate prediction pixel values in the total pixel values as well as the proportion of correct prediction results in the total predicted value. The accuracy of correctly predicting pixel samples is assessed using this standard, and the Eq. (9) is as follows:

$$Pre = \frac{TP}{TP+FP} \quad (9)$$

The percentage of the actual number of correct total pixel samples that the model correctly predicted is known as Recall. It mainly focuses on the proportion that the target pixel feature is not correctly classified as positive, as shown in Eq. (10):

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

IoU represents the union of the segmentation prediction result and the intersection ratio of the real segmentation label, as shown in Eq. (11):

$$IoU = \frac{TP}{FP+TP+FN} \quad (11)$$

Two samples can be compared for similarity using a function called the Dice similarity coefficient. The similarity between the label image and the predicted image can be computed using it. The value is between 0 and 1, as shown in Eq. (12):

$$Dice = \frac{2TP}{FP+2TP+FN} \quad (12)$$

The positive sample is the target feature, and the negative sample is the background. TP stands for true positive and correctly predicts the target pixel. FP stands for false positive and incorrectly predicts the background pixel as the target pixel. TN stands for true negative and correctly predicts the background pixels. FN stands for false negative, and the target pixel is incorrectly predicted as the background pixel.

IV. Experimental Results and Analysis

A. Comparison of Segmentation Performance under Different Networks

The method suggested in this study and the traditional semantic segmentation algorithm are chosen for comparison experiments in order to confirm the efficacy of this approach. Using the same experimental configuration and the same parameter settings, the comparative experiments with U-Net, ResUnet [25], UNET++ DeepLabv3+, and DCSAU_Net [26] segmentation models are realized, respectively. The model's segmentation performance is assessed using the four datasets listed in this study. The experiment uses four evaluation indexes: Precision, Recall, IoU, and Dice similarity coefficient.

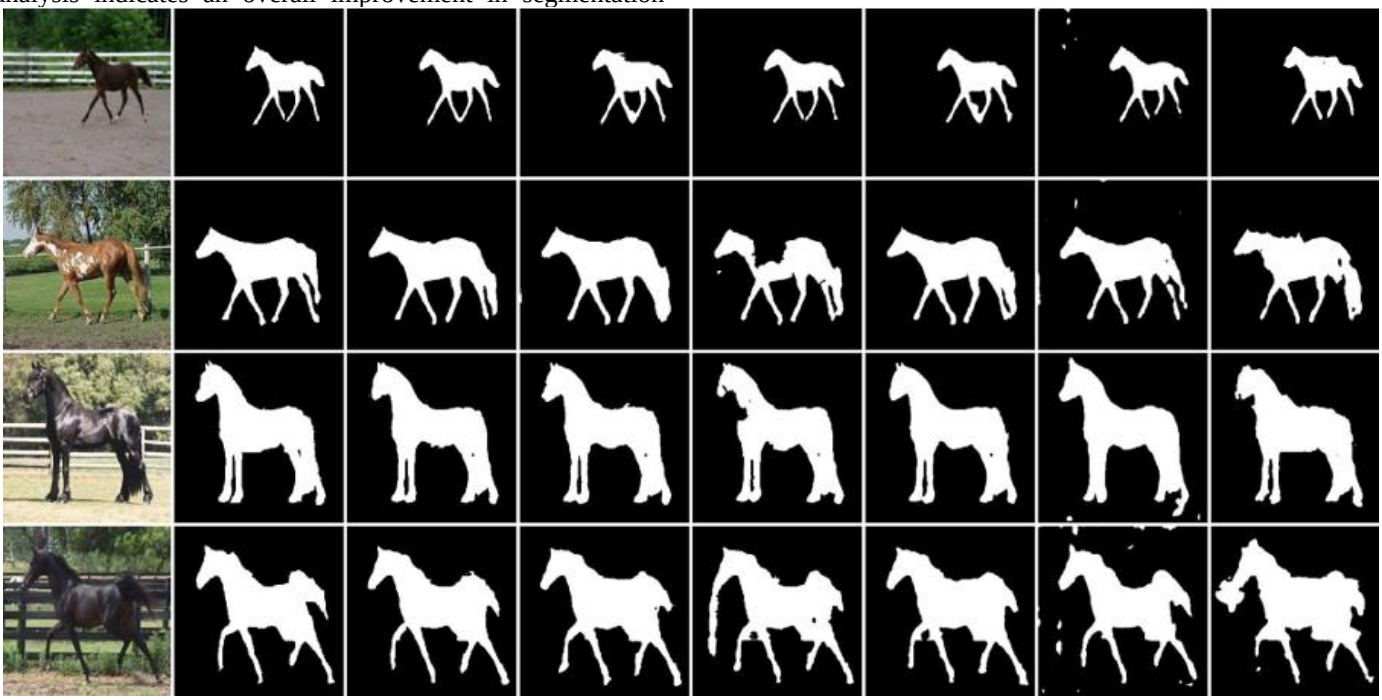
The experimental findings for the dataset of horse image segmentation are displayed in Table I. Bold is the model score that has the highest index. It is evident that the majority of this method's indicators outperform those of other semantic segmentation techniques. The precision, IoU, and Dice similarity coefficients improved by 3.07%, 1.83%, and 1.09%, respectively, in comparison to the traditional U-Net network model.

Fig. 7 visualizes the original image, actual label, and prediction results. The above models' segmentation results are presented in a more intuitive way.

TABLE I. SEGMENTATION RESULTS OF DIFFERENT MODELS IN HORSE SEGMENTATION DATASET

Method	Pre	Rec	IoU	Dice
U-Net	0.9242	**0.9566**	0.8864	0.9385
ResUnet	0.9091	0.9157	0.8375	0.9079
Unet++	0.9421	0.9526	0.8993	0.9458
DeepLabv3+	0.9386	0.9255	0.8734	0.9314
DCSAU_Net	0.9090	0.9300	0.8488	0.9165
Ours	**0.9549**	0.9455	**0.9047**	**0.9494**

In Fig. 7, each row is the segmentation visualization results of different original images and different models. The visual analysis indicates an overall improvement in segmentation accuracy. Comparing the segmentation results reveals that ResUnet has the worst recognition effect on the background and edge, and it can only roughly segment the horse area. For example, in the results of the second and fourth lines, the horse is under-segmented, and the last line is noticeably over-segmented, making it impossible to accurately distinguish the horse's edge from the background. Both U-net and Unet++ achieve satisfactory segmentation of the horse's body as a whole, but lack the accuracy of edge segmentation. The proposed approach performs better overall than other comparison techniques, and the segmentation effect is nearly identical to the label value. Not only the boundary is smooth and accurate, but also there are a few cases of under segmentation and over segmentation.



(a) Image (b) GT (c) Ours (d) U-Net (e) ResUnet (f) Unet++ (g) DeepLabv3+ (h) DCSAU_Net

Fig. 7. Visualization effect of horse image segmentation under different networks.

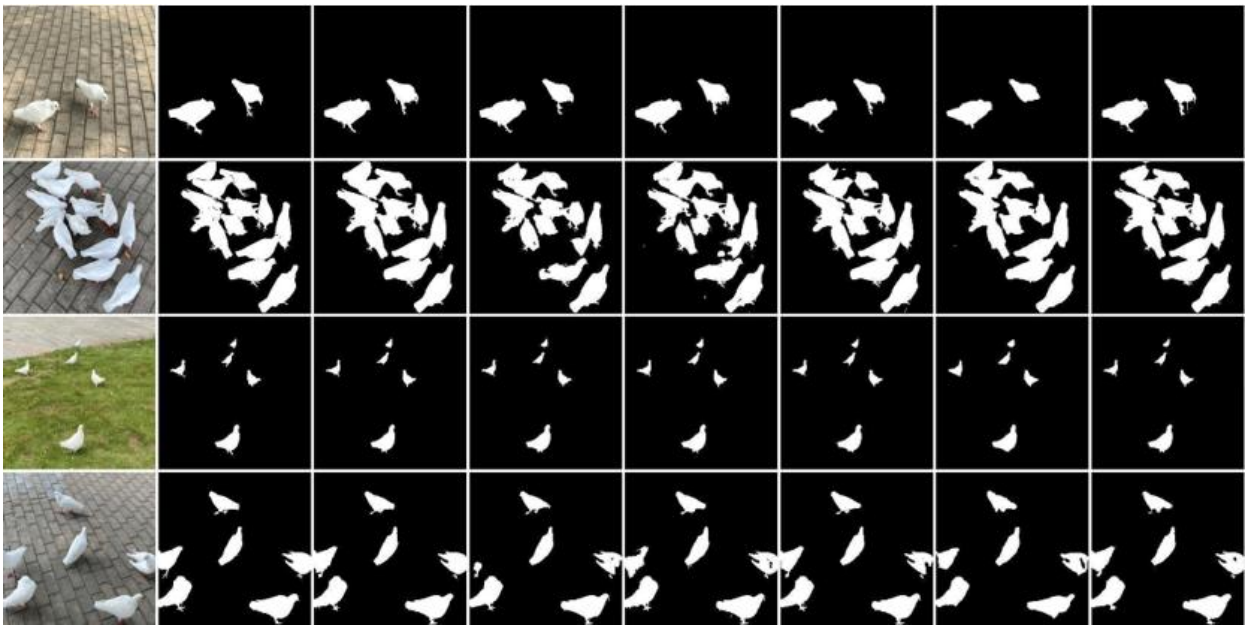
The segmentation performance of the proposed method is more effectively verified through the comparative experiments on the other three datasets. Specifically, the data is showcased in three separate tables, namely Table II, Table III, and Table IV, which are designed to offer a comprehensive view of the various aspects of the experiment.

In the three evaluation indexes of Recall, IoU, and Dice, Table II demonstrates that the LCA_Net network suggested in this study outperforms other networks. The Dice reached 84.62%, which was 0.71%, 1.72%, 0.58% 7.55% and 0.44% higher than that of U-Net, ResUnet, SegNet, Unet++ DeepLabv3+and DCSAU_Net, respectively.

The segmentation results of different models in the pigeon dataset are displayed as illustrated in Fig. 8. The third line shows that when the contrast between the pigeon and the background features is significant, all five models can effectively segment the target features. When the contrast between background features and the overall target is low, other experimental models may experience under segmentation, boundary blurring, and other phenomena. Notably, the model proposed in this study exhibits superior segmentation performance, able to better distinguish subtle pigeon edge features, and the segmentation results are closer to real labels.

TABLE II. SEGMENTATION RESULTS OF DIFFERENT MODELS IN PIGEON SEGMENTATION DATASET

Method	Pre	Rec	IoU	Dice
U-Net	**0.7862**	0.9041	0.7257	0.8391
ResUnet	0.7774	0.9016	0.7139	0.8290
Unet++	0.7774	0.9201	0.7291	0.8404
DeepLabv3+	0.6842	0.8983	0.6369	0.7707
DCSAU_Net	0.7791	0.9194	0.7292	0.8418
Ours	0.7830	**0.9243**	**0.7362**	**0.8462**



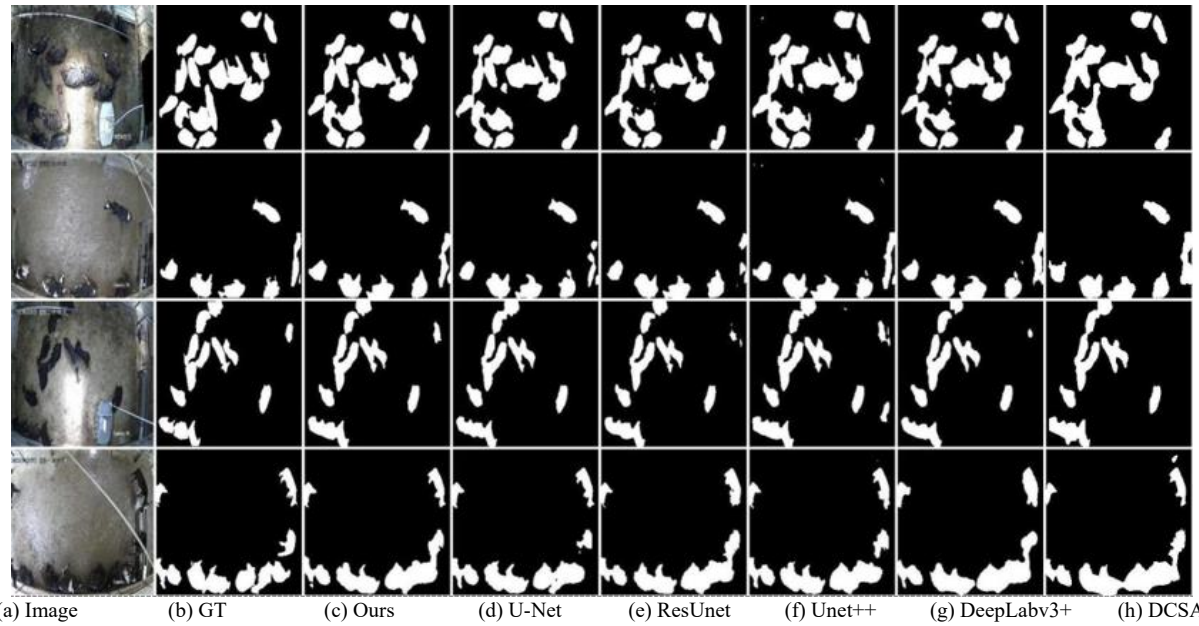
(a) Image (b) GT (c) Ours (d) U-Net (e) ResUnet (f) Unet++ (g) DeepLabv3+ (h) DCSAU_Net

Fig. 8. Visualization effect of pigeon image segmentation under different networks.

TABLE III. SEGMENTATION RESULTS OF DIFFERENT MODELS IN CATTLE FARM SEGMENTATION DATASET

Method	Pre	Rec	IoU	Dice
U-Net	0.9066	0.9199	0.8403	0.9122
ResUnet	**0.9206**	0.9000	0.8345	0.9079
Unet++	0.8826	**0.9428**	0.8372	0.9104
DeepLabv3+	0.8796	0.9212	0.8178	0.8989
DCSAU_Net	0.8764	0.9124	0.8083	0.8926
Ours	0.9102	0.9230	**0.8454**	**0.9151**

The segmentation and visualization outcomes of various models in the cattle farm segmentation dataset are displayed in Table III and Fig. 9, respectively. The experimental indicators are improved compared with other models. It is evident from the visualization results that the model in this study has a better overall segmentation effect than other models, particularly when it comes to the first line of the comparison image. Other models' segmentation performance decreases when the scene's lighting impacts, but the model developed in this study is more in line with the label graph, highlighting the benefits of the model's robustness and anti-interference.



(a) Image (b) GT (c) Ours (d) U-Net (e) ResUnet (f) Unet++ (g) DeepLabv3+ (h) DCSAU_Net

Fig. 9. Visualization effect of cattle image segmentation under different networks.

The segmentation indexes and visualization outcomes of various models in the pig farm segmentation dataset are displayed in Table IV and Fig. 10, respectively. Three performance indices show that the model suggested in this study outperforms other models, and the segmentation effect in night and strong light scenes is obviously superior to other models.

The LCA-Net network model proposed in this study achieved more refined segmentation results through comparative experiments with four datasets and visualization results. These results substantiate the effectiveness of the proposed model.

TABLE IV. SEGMENTATION RESULTS OF DIFFERENT MODELS IN PIG FARM SEGMENTATION DATASET

Method	Pre	Rec	IoU	Dice
U-Net	0.9490	0.9645	0.9184	0.9556
ResUnet	0.9522	0.9644	0.9209	0.9569
Unet++	0.9510	**0.9687**	0.9252	0.9592
DeepLabv3+	0.9399	0.9478	0.8947	0.9435
DCSAU_Net	0.9335	0.9243	0.8719	0.9269
Ours	**0.9578**	0.9635	**0.9263**	**0.9602**

(a) Image (b) GT (c) Ours (d) U-Net (e) ResUnet (f) Unet++ (g) DeepLabv3+ (h) DCSAU_Net

Fig. 10. Visualization effect of pig image segmentation under different networks.

B. Ablation Experiment

Using the same loss function and parameter settings as the comparative experiment for training, the ablation experiment confirms the overall effect of innovative reconstruction on the LCA_Net model. The method is to take U-Net as the baseline and successively accumulate CSwin module, CBAM attention module and ASPP module into it. Finally, ablation experiments are performed on the horse body segmentation dataset using the optimized LCA_Net network. Table V displays the comprehensive study results.

TABLE V. TESTS OF ABLATION FOR DIFFERENT MODULES ON THE HORSE SEGMENTATION DATASET

Method	Pre	Rec	IoU	Dice
U-Net	0.9242	0.9566	0.8864	0.9385
CSwin	0.9378	0.9522	0.8955	0.9435
CBAM	0.9408	0.9571	0.9023	0.9477
ASPP	0.9402	**0.9576**	0.9020	0.9476
CSwin+CBAM+ASPP	0.9549	0.9455	0.9047	0.9494

As shown in Table V, when the CSwin module is introduced alone, Precision and IoU are improved by 1.36% and 0.91% over the baseline, respectively, demonstrating its effectiveness in deep feature extraction. However, Recall slightly decreases. The CBAM module, when introduced alone, performs optimally, with Precision, IoU, and Dice increasing by 1.66%, 1.59%, and 0.92%, respectively, while Recall remains stable. This validates that the attention mechanism helps suppress redundant features and enhance critical areas. The performance of the ASPP module alone is close to that of CBAM, indicating that multiple modules can integrate multi-scale contextual information, but the improvement in Precision is limited. After integrating all three modules, Precision, IoU, and Dice reach their optimal values, improving by 3.07%, 1.83%, and 1.09%, respectively, compared to the baseline. Although Recall slightly decreases, the overall results show that the feature extraction capability of CSwin, along with the refined optimization from CBAM and ASPP, complement each other effectively, notably improving the model's segmentation accuracy and target localization. This suggests that the feature extraction of CSwin, attention enhancement from CBAM, and multi-scale fusion from ASPP

form an effective synergy that drives the systematic optimization of the model's segmentation performance.

Each module's efficacy can be confirmed by the ablation experiment. Every module introduced in this study has the capability to increase the precision and accuracy of segmentation. The CSwin module in particular is essential for enhancing segmentation performance.

V. Discussion

However, there are still some limitations in this study, including the relatively small size of the dataset and the fact that cross-dataset generalization tests have not yet been conducted to fully assess the model's adaptability. Future research directions should prioritize addressing these limitations. Expanding the diversity and volume of training data, implementing rigorous cross-dataset benchmarking, and conducting more sophisticated statistical testing will be essential steps toward strengthening the validity and real-world impact of the LCA-Net framework.

VI. Conclusion

Aiming at the problem of edge blurring in image segmentation, an improved U-Net network combining the CSwin module and channel spatial attention mechanism is proposed. Additionally, the ASPP module is introduced into the decoder output which contributes in improving the quality image segmentation accuracy. Moreover, the LCA-Net proposed by ours is compared with U-Net, ResUnet, Unet++ and Deeplabv3+ networks. On the test sets of four datasets, the LCA-Net's evaluation indexes are essentially superior to those of other networks, achieving better results in segmentation visualization. The LCA_Net network model can accurately segment the target edge, provide a reliable technical means for growth monitoring in the process of animal husbandry, and provide a valuable reference for other similar segmentation tasks.

The future work we will focus on the development of a lightweight version of LCA-Net to provide a low-power and high-efficiency solution for the farm mobile monitoring system. At the same time, we will optimize the model architecture for specific scenarios such as livestock individual identification and health monitoring to ensure stable operation under complex conditions.

Acknowledgment

The authors gratefully acknowledge the support from the 2022 Liaoning Provincial Natural Science Foundation Program Key Science and Technology Innovation Base Joint Open Fund(2022-KF-18-04) and Shanghai Municipal Commission of Education (C2024090).

References

[1] Jain S, Sikka G, Dhir R. (2024) A systematic literature review on pancreas segmentation from traditional to non-supervised techniques in abdominal medical images. *Artificial Intelligence Review*, 57(12): 317.

[2] Tian, X. Wang, L. and Ding, Q. (2019). Review of image semantic segmentation based on deep learning. *Ruan Jian Xue Bao/Journal of Software*, 30(2), 440-468. (in Chinese)

[3] Long, J. Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431-3440). Santiago: IEEE.

[4] 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.

[5] Ronneberger, O. Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234-241). Cham: Springer International Publishing.

[6] Zhou, Z. Siddiquee, MMR. Tajbakhsh, N. and Liang, J. (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* (pp. 3-11). Cham: Springer. doi:10.1007/978-3-030-00889-5_1

[7] Chen, L. C. Papandreou, G. and Kokkinos, I. (2017). 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. doi:10.1109/TPAMI.2017.2699184

[8] Chen, L.-C. Zhu, Y. Papandreou, G. Schroff, F. and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801-818). Cham: Springer.

[9] Song, Z., Zou, S., Zhou, W. et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 11, 4294 (2020). https://doi.org/10.1038/s41467-020-18147-8

[10] Yanqing, Z. Heru, X. and Xinhua, J. (2018). Non-contact measurement of sheep body size based on multi-scale Retinex image enhancement. *Journal of China Agricultural University*, 23(9), 156-165. (in Chinese)

[11] Zhang, J-J. (2020). Measurement and design of horse's body size based on deep learning and image processing technology. *Computer Technology and Development*, 30(11), 180-184+189. (in Chinese)

[12] Qin, X. Huang, D. and Song, W. (2023). Fish image segmentation method based on object detection and edge support. *Journal of Agricultural Mechanization*, 54(1), 280-286. (in Chinese)

[13] Yao, C. Ni, F-C. and Li, G-L. (2023). Research progress on the application of image segmentation based on deep learning in livestock and poultry farming. *Journal of Huazhong Agricultural University*, 42(3), 39-46. (in Chinese)

[14] Naseeb, S. Indu, D. and Kuldeep, D. (2024). Development of attention-enabled multi-scale pyramid network-based models for body part segmentation of dairy cows. *Journal of Biosystems Engineering*, 49(2), 186-201.

[15] Tao, F. Yangyang, G. Xiaoping, H. and Yongliang, Q. (2023). Cattle target segmentation method in multi-scenes using improved deepLabV3+ method. *Animals*, 13(15), 2521.

[16] Bin, X. Weipeng, J. Changkai, W. Songtao, H. Fan, Z. Kaidong, L. and Jinlin, L. (2021). Feature detection method for hind leg segmentation of sheep carcass based on multi-scale dual attention U-Net. *Computers and Electronics in Agriculture*, 191, 106482.

[17] Turkmenli, E. Aptoula. and Kayabol, K. (2024). HistSegNet: Histogram layered segmentation network for SAR image-based flood segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21, 1-5. doi:10.1109/LGRS.2024.3450122

[18] Bello, R.-W. Mohamed, A.S.A. and Talib, A.Z. (2021). Contour extraction of individual cattle from an image using enhanced Mask R-CNN instance segmentation method. *IEEE Access*, 9, 56984-57000. doi:10.1109/ACCESS.2021.3072636

[19] Wang, L. Chen, C. Chen, F. Wang, N. Li, C. Zhang, H. Wang, Y. and Yu, B. (2024). UGTransformer: A sheep extraction model from remote sensing images for animal husbandry management. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 5402614. doi:10.1109/TGRS.2024.

[20] Wang, M. Lv, M. Liu, H. and Li, Q. (2023). Mid-infrared sheep segmentation in highland pastures using multi-level region fusion OTSU algorithm. *Agriculture*, 13(7), 1281. doi:10.3390/agriculture13071281

[21] Liu, L. Li, Y. Wu, Y. Ren, L. and Wang, G. (2023). LGI Net: Enhancing local-global information interaction for medical image segmentation.

Computers in Biology and Medicine, 167, 107627. doi:10.1016/j.compbiomed.2023.107627

[22] Roburin, S. de Mont-Marin, Y. Bursuc, A. Marlet, R. Perez, P. and Aubry, M. (2022). Spherical perspective on learning with normalization layers. *Neurocomputing*, 487, 66-74. doi:10.1016/j.neucom.2022.02.051

[23] Dong, X. Bao, J. Chen, D. Zhang, W. Yu, N. Yuan, L. and Chen, D. (2022). CSwin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12124-12134). New Orleans: IEEE.

[24] Woo, S. Park, J. Lee, J. Y. and Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3-19). Munich: Springer.

[25] Zhang, Z., Liu, Q. and Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749-753. doi:10.1109/LGRS.2018.2802944.

[26] Xu, Q. Ma, Z. He, N. and Duan. W. (2023) DCSAU-net: A deeper and more compact split-attention U-Net for medical image segmentation. *Computers in Biology and Medicine*, 154, 106626. doi:10.1016/j.compbiomed.2023.106626