ABSTRACT
The portrait thangka image is a kind of religious scroll painting that expresses figures’ identity and duties through portraits, sitting platforms, and backlighting. The segmentation of significant semantic objects in the image is one of the essential ways for scholars to study and understand the image’s content. To better understand this content, we elaborately collected a dataset of portrait-like thangkas, which consists of 4086 images covering four object categories. We provide rich annotation for this dataset. In addition, we propose an end-to-end deep learning method that effectively solves the problems of blurred target edges, segmentation errors, and missed segmentation in thangka image segmentation. First, regular convolution and atrous convolution of different sizes are concatenated after the high-level feature output. This method can effectively improve the receptive field of the model while obtaining more image feature information. Then, the attention module is introduced to fully utilize the spatial relationship between the image’s semantic content and enhance the discriminative ability of the feature representation on the thangka image. Finally, cross-layer feature fusion is added to reduce the loss of edge details and improve the accuracy of target edge segmentation. The results show that compared to the base model, the mPA and mIoU indices of the model proposed in this paper reach 90.75% and 85.66%, respectively, which effectively improved the accuracy of the Tangka image segmentation.

INDEX TERMS
Semantic segmentation, DeepLabv3+, attention mechanism, thangka images, deep learning.

I. INTRODUCTION
Thangka is a religious painting unique to Tibet in China, known for its complex and exquisite patterns and rich content. Its content contains all aspects of Tibetan religion, history, politics, culture, and social life. It occupies an important position in Tibetan culture, known as an encyclopedia of Tibetan culture, art, and history, with high research value. It was listed as China’s intangible cultural heritage in 2006. Portrait thangka images are usually presented in a central composition. More of the Buddha statues in the center come from real historical life (such as Lotus Flower, the principal founder of Tibetan Buddhism; Amitabha Buddha, the main deity of the “Western Elysium” and known as the “Three Western Saints” along with Guanyin Bodhisattva and Daishonin Bodhisattva). The forms of these thangka images also vary, with the “Dharma Buddha” sitting on a lotus seat with his mother embracing him. The “Retribution Buddha” holds a bell in his left hand and a pestle in his right. The “Incarnation Buddha” wearing a flower crown on his head, holding a staff in his left hand and a pestle in his right. There are also specific criteria for the content and layout of the thangka. For example, when the main deity is a bodhisattva or dharma, the background is decorated with a pattern consisting of hands in addition to the headlight, backlight, and blooming flowers. In addition, the specific head ornaments and Dharma objects of the Buddha statue can also indicate its identity and merit. These regular features can help us understand the thangka better. With the advancement of information technology, the segmentation of specific objects in thangka images using
image processing techniques can help scholars study the images’ content and thus understand the high-level semantic content [1].

II. RELATED WORK

Semantic segmentation is fundamental to computer vision research tasks such as image classification, image understanding, and scene parsing. It has been widely used in human-computer interaction, geological exploration, medical image analysis, and other fields with important research significance and application value from the 1960s to the present [2]. Traditional image segmentation methods analyze and process images according to their features, such as color and texture information. Otsu [3] proposed a parameter-free and unsupervised automatic threshold selection method for image segmentation to select the best threshold by discriminating the maximum variance between classes; Saint-Mar et al. [4] implemented a method to convert all step edges, roof-like edges, and slope edges after Gaussian smoothing into ideal stepped edges and achieve edge detection by adaptive iterative smoothing. Adams and Bischof proposed to find a seed pixel as a growth point for each region to be segmented, and then merge the pixels in the field around the seed pixel that have the same or similar properties to the seed into the region where the seed pixel is located until no more pixels satisfying the conditions are included [5], [6], traditional image segmentation methods are computationally simple and fast, but have disadvantages such as unsatisfactory segmentation results. With the rapid development of computer hardware, deep learning techniques are widely used in computer vision, which provides strong support for the development of image semantic segmentation techniques. The Fully Convolutional Neural Networks (FCN) proposed by Long et al. [7] uses convolutional layers instead of fully connected layers, which can input arbitrary sizes of images and uses deconvolution to achieve end-to-end training, which is a qualitative leap compared to previous networks. This method breaks the limitations of traditional segmentation methods. It opens the door for the research of semantic segmentation at the pixel level, such as the DeepLab series of models proposed by adding dilated convolution, Atrous Spatial Pyramid Pooling (ASPP), and depth-separable convolution [8], [9], [10], [11], contraction path, and expansion path based U-Net medical image models based on contracting and expanding paths [12], and PSPNet models [13] that incorporate a pyramid pooling module to aggregate contextual information from different regions and thus improve the ability to obtain global information. Most of these models are based on the improvement of FCN. To a certain extent, they solve the problems such as the lack of accuracy and adaptability of traditional semantic segmentation methods. Recently, DAM [14] proposed a dense attention module enabling hierarchical adaptive feature fusion by exploiting inter-channel and intra-channel relationships. The research introduces the adaptive receptive field and channel selection modules that enable the network to tackle variable-sized instances and correlated feature maps [15].

In recent years, research on thangka image segmentation has mainly focused on broken region segmentation [16], [17], [18], and only a few scholars have segmented semantic objects in thangka images. In the literature [19], a headdress segmentation algorithm based on circular region localization for portrait-like thangka images is proposed, which uses a circle detection algorithm to locate the approximate location of the headlight based on the characteristics of the circular headlight region and then uses the spatial distribution of headlight color and the exterior contour features of the headdress to achieve the primary dzong headdress segmentation, which cannot be detected for thangka images without a circular headlight, impure headlight color, and headdress with similar headlight color. The literature [20] proposed to use the maximum interclass variance method to obtain the threshold value to segment the image, and then the overall features are inscribed according to the Euler number of the headdress region and the color distribution inside the perspective contour; the literature [21] proposed the Mean-Shift-based thangka segmentation algorithm according to the problems such as the high computational complexity of tangka images and the difficulty of large-scale image segmentation, which combines the quadratic watershed The algorithm combines the quadratic watershed algorithm with the clustering theory of the improved weight matrix algorithm, which effectively reduces the computational complexity of the traditional clustering method; literature [22] proposes a headdress detection algorithm based on saliency mapping of thangka figures, which uses the attention model to calculate the saliency value of each pixel, and then detects the segmented headlight by finding the backlight region of the thangka; The segmentation of thangka images with high complexity using traditional image segmentation methods is not satisfactory; therefore, feature extraction of thangka images using deep learning methods and pixel classification using classifiers can result in highly accurate image segmentation. Literature [23] uses a line drawing enhancement module and a halving region generation network to segment thangka images. The halving region generation network method is designed based on the structural features of the thangka and is adequate but limited.

In this paper, we propose an improved DeepLabv3+ thangka image semantic segmentation model for thangka images with complex structure and content, rich texture, and color mixing, the main contributions of the paper are reflected as follows:

1. Collected and constructed the first pixel-level semantic annotation portrait-like thangkas dataset, consisting of 4086 images covering four object categories: figure, backlight, headlight, and pedestal. All high-definition thangka images are annotated with high quality. These images can help us gain a deeper understanding of the algorithm performance and can be applied to many vision tasks, such as object localization, semantic edge detection, and style migration;

2. The features of RFB (Receptive Field Block) [24] are used to improve the spatial pyramid ASPP (Atrous Spatial Pyramid Pooling) module in the original network to obtain
more image feature information. The use of depth-separable convolution reduces the parameters of the model;

(3) An improved attention mechanism module DCBAM (Dilation Convolutional Block Attention Module) is added before up-sampling of high-level features to improve the accuracy of network segmentation;

(4) Add cross-layer feature fusion to solve problems such as roughness of target edges;

The paper is organized as follows: Section 1 introduces the characteristics of thangka images; The application and development trend of image segmentation techniques are introduced in section 2; Section 3 describes the process of collecting and producing the dataset of this paper; Section 4 describes the structure of the proposed model and the details of the proposed improvements; In Section 5, experimental methods are used to demonstrate the validity of the developed model, and the conclusions are given in Section 6.

III. PROPOSED DATASET
A. IMAGE COLLECTION

In order to construct a high-definition thangka image dataset, we scanned and cropped the high-precision portrait thangka images from the thangka books (in a hard form) in the university library: Gesar Thousand Thangkas (Gesar Thousand Thangkas is a thangka book published by Sichuan Nation Press) and Tibetan Thangkas (Tibetan Thangkas is a hardcover scroll painting published by Heritage Press), and at the same time, used search engines to obtain relevant high-definition thangka images from the Internet. We initially collected a high-resolution collection of 454 portrait thangka images.

B. IMAGE PREPROCESSING

As a pre-processing data method, data augmentation plays an essential role in deep learning [25]. In general, effective data augmentation can improve the robustness of the model better and obtain more vital generalization ability. Standard data augmentation methods include rotation, panning, changing contrast, etc. Since annotating each image requires a huge labor cost, data augmentation is used in the experiments to generate additional new images and increase the data volume to solve the problem of the insufficient datasets of portrait-like thangka images. The graph of the effect after data enhancement is shown in Figure 1, where (a) is the original image, (b) is changing the chroma, (c) is changing the contrast, and (d) is obtained by flipping up and down 180°. Finally, it is extended from the initial 454 images to 4086 images. They are used to evaluate the adaptability and generality of the proposed semantic segmentation method in the new dataset and improve the model’s segmentation performance in HD thangka images. Figure 2 shows an example of the thangka dataset.

C. IMAGE DATA ANNOTATION

Using labelme software to label the high-definition portrait-like thangka images following the PASCAL VOC2007 dataset format, different colors are assigned to different categories in the images, such as red for the pixels of the Buddha figure, green for the pixels of the headlight, yellow for the pixels of the pedestal, blue for the pixels of the backlight, and black for the remaining uninteresting pixels. The annotated thangka images were saved in JSON format. Due to the high resolution of the collected thangka images, the annotation time for each image is about 20 minutes. The training set is divided in the ratio of 6:3:1, where the training set is 2451 images, the validation set is 1225 images, and the test set is 410 images.

IV. PROPOSED METHOD
A. BASE MODEL

DeepLabv3+ is a semantic segmentation network developed by Google, absorbing the advantages of the DeepLab family of methods. It uses the residual network as the underlying network. It uses the deep separable convolutional structure for the atrous space pyramid pooling module and the decoding module, which is one of the more popular frameworks in image segmentation.

B. IMPROVED MODEL

The improved DeepLabv3+ implementation in this paper is as follows:

Encoding part: the DeepLabv3+ model uses the lightweight network Xception as the feature extraction network, which uses depth-separable convolution to replace the multi-size convolution of the original Inceptionv3, and then combines it with a residual connection mechanism similar to Resnet to reduce the model parameters while significantly improving the accuracy. The high-level features extracted by Xception are passed through the RFBA module. At the same time, the original atrous convolution is combined with the depth-separable convolution to form three depth-separable
atrous convolutions with atrous rates of 6, 12, and 18, respectively, followed by splicing and fusion of the five multi-scale feature maps generated by the RFBA module in the channel dimension; the 1×1 convolution is implemented in the Before channel compression, the attention mechanism is embedded for the fused features to help the network learn the feature weights of the channel autonomously and make full use of the spatial relationship between the Buddha’s headlight and the backlight. Currently, the feature map resolution size is 1/16 of the original map.

The decoding part uses cross-layer fusion, where the coded output is twice up-sampled by a factor of 2 and then fused separately with lower-layer features, which helps generate detailed information with clear boundaries for high-resolution prediction. The general structure of the model is shown in Figure 3.

C. ENCODER

1) PYRAMID SENSORY FIELD MODULE

Several findings in neuroscience suggest that in the human visual cortex, the size of the groupwise receptive field (pRF) [26] is a function of its retinotopic map expressed in terms of eccentricity, as shown in figure 4(a) pRF size also increases with increasing eccentricity. Also, as shown in figure 4(b), The structure helps to highlight the importance of regions closer to the center, and regions closer to the center have a better role in recognizing objects. Current standard deep learning models usually use regular sampling grids on the feature maps [27], [28], setting the receptive field (RF) to the same size, which may increase the loss of features.

The original aspp used the atrous convolution with different dilation rates in parallel to capture multi-scale information to cope with the diversity of segmentation target scales. Such a design can effectively avoid redundant information and focus directly on the correlation between objects. The rfb module enables the model to simulate better the perception of human vision by connecting the atrous convolution with different dilation rates after the regular convolution at different scales. In this paper, we use the features of RFB to improve aspp and innovatively propose a perceptive field block aspp (the aspp module with the receptive field block, RFBA), as shown in the dashed box in figure 3. Among them, two 3×3 convolutions are used instead of 5×5 convolutions, and 1×7 and 7×1 convolutions are used instead of 7×7 convolutions to reduce the computational effort of the network while improving the ability of the network to represent features nonlinearly. RFB utilizes a multi-branch pool with convolution kernels corresponding to RFS of different sizes, and multiple convolution kernels of different scales are similar to prfs of different sizes. in contrast, the inflated convolution assigns an individual eccentricity, allowing the RF to produce a final spatial array similar to the human eye visual system in figure 5(c).

2) DILATION CONVOLUTIONAL BLOCK ATTENTION MODULE

Attention refers to the spatial and channel information in the feature channels. In general, we assume that the features obtained from a convolutional network have the same importance. Still, the features are not equally crucial across channels, so using an attention mechanism can help the network learn the feature weights autonomously. The channel attention in the attention mechanism CBAM [29] allows the network to focus on “what” the image is, and the spatial attention complements the channel attention by allowing the network to focus on “where” the object is in the image. In this paper, the 7×7 large convolution kernel in spatial attention is improved by using a 3×3 small convolution with
an atrous rate of 3 instead of the original large convolution kernel to increase the perceptual field and reduce the number of module parameters. The specific process is shown in Figure 6. The fused convolution is mapped by the channel attention to generating the feature map $F$. $F$ is subjected to the parallel operations of maximum pooling and average pooling. The pooling results are combined and input to the multilayer perceptron (MLP) operation and summed to generate the one-dimensional channel attention $M_c$, and the channel attention $M_c$ is multiplied by the original feature map $F$ to obtain the feature map $F$ with channel weights $\delta$, the channel attention is shown in Equation (1). Then the spatial attention mapping performs a series operation of maximum pooling and average pooling on the channel-weighted feature map $F'$, after which it is downscaled into a single-channel feature map using a $3 \times 3$, atrous convolution with an atrous rate of 3. After the activation function two-dimensional spatial attention Ms, the spatial attention is multiplied by $F'$ to obtain the desired spatial feature graph $F''$, and the spatial attention is shown in Equation (2). Finally, the input features are added with the feature map $F''$ after the improved CBAM mechanism to obtain the next layer of convolutional input.

$$M_c(F) = \delta (MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

$$M_s(F) = \delta \left\{3 \times 3_{\text{dilat}} \left[ \text{AvgPool}(F); \text{MaxPool}(F) \right] \right\}$$  \hspace{1cm} (2)

MLP denotes multilayer perceptron, $\delta$ activation function, $3 \times 3_{\text{dilat}}$ denotes a convolution with a convolution kernel size of 3 and a dilation rate of 3.

The improved attention mechanism can enhance the importance of the information carried in the features to the target prediction, attach weight coefficients to each feature channel, make full use of the spatial relationship between headlight, backlight, and sitting platform, highlight useful features, suppress redundant channel information, enhance the learning ability and generalization ability of the model.

![FIGURE 6. DCBAM general structure.](image)

**D. DECODER**

Feature fusion is shown in Figure 7. The coded output features are upsampled by 2-fold bilinear interpolation and fused with the feature map in the backbone network Xception with a resolution of $1/8$ of the original map. The fused features are adjusted to the number of channels and then stitched and fused with the feature map in the backbone network Xception with a size of $1/4$ of the original map by 2-fold up sampling again. The shallow features of the image have more edge details, and the high-level features have rich semantic information. The two reinforce each other. The problem of information loss is reduced by adding cross-layer feature fusion.

![FIGURE 7. Feature fusion.](image)

**V. PERFORMANCE ANALYSIS AND DISCUSSION**

**Experimental Parameters:** This experiment is based on an Intel i7-10700 processor, Nvidia GeForce GTX 1080T 11G graphics card, deep learning framework keras2.0, and programming language python3.7. Pre-training initialization of network parameters is performed by loading Xception on the ImageNet dataset, initializing by migration learning weights are initialized by migration learning, and the model is fine-tuned on the PASCAL VOC2007 and self-built portrait-like tangka datasets, respectively, to speed up the convergence of the model. The model uses an Adam optimizer for updating the network parameters during training, setting the initial learning rate as $1 \times 10^{-4}$ and momentum as 0.9; influenced by the GPU memory capacity, the training samples are input into the network in batches, setting the number of training images per batch value as 4 and Epoch value as 100; this paper uses the cross-entropy loss function. This function evaluates the class prediction of each velocity-limited vector individually and then averages the pixels. In case of sample imbalance, the weights of the small target samples are reset until a better segmentation effect is achieved.

**Experiment Database:** We are using the self-built tangka dataset and the open-source PASCAL VOC2007 dataset to compare the proposed method with other segmentation methods.

**Performance Measures:** To quantitatively evaluate the performance of the tangka image segmentation algorithm and verify its effectiveness, mIoU (mean intersection over union), mPA (Mean Pixel Accuracy), and Precision are used as the evaluation criteria for the proposed algorithm.

**A. PERFORMANCE ANALYSIS ON SELF-BUILT PORTRAIT-LIKE TANGKA DATABASE**

1) **PERFORMANCE ANALYSIS OF DIFFERENT MODEL**

The validity of the proposed method is verified by the comparison with the standard semantic segmentation network. The comparison results are shown in Table 1. It can be seen that: the segmentation accuracy and the average cross-merge ratio of the proposed method in this paper are ahead of the comparison networks.
Comparison with other models: In this paper, we conducted training and testing on the self-built portrait-like thangka dataset and compared several more commonly used algorithms in the semantic segmentation field. The visualization results of each model are shown in Figure 8. Compared with the method in this paper, the Segnet model uses inverse pooling to upsample the feature map in the decoder and learns through the position index of the sampled values recorded during downsampling, and the feature map is sparse, and Segnet does not take into account the multi-scale processing of the image, which makes the segmentation results have the problem that acceptable targets cannot be identified, and there is more missed segmentation. PSPnet also has more missed segmentations, and this network introduces the pyramid pooling module on the basis of FCN, which can perform multi-scale pooling operations on the input image, and this has the advantage of using the global spatial context information. However, as shown in Table 2, this network is prone to the problem of insufficient information in a single class of images, while the model directly treats the character outline more smoothly with a more serious loss of details; the DeepLabv3+ network uses the pyramid module to obtain a larger perceptual field, thus incorporating more global features and possessing more accurate results compared to other models.

It can be seen from the figure that the algorithm before the improvement had the problems of incomplete segmentation and blurred segmentation edges for the hands, ears, weapon, and hair ornaments of the Buddha statue, and the improved algorithm can segment the tiny targets existing in the thangka more clearly. The phenomenon of missed segmentation and mis-segmentation has been improved significantly. The reason may be that this paper adds regular convolution with different atrous rates in the RFBA module, which improves the model’s ability to extract features; the attention mechanism with atrous convolution is tandem with RFBA, and the attention mechanism can calculate a weight for each channel, which can let the important channel information be promoted and the minor channel information be suppressed, and strengthen the correlation between channels. At the same time, the spatial position relationship between the headlight, backlight, and pedestal of the thangka image is fully utilized to help the network better extract image features; as the number of network layers increases, the shallow feature information extracted by the model is seriously lost, and by adding a cross-layer feature fusion branch, the network can get a more accurate edge segmentation target. As shown in Figure 9, the loss curves of the DeepLabv3+ network before and after improvement on the training and test sets are shown.

2) PERFORMANCE ANALYSIS OF DIFFERENT IMPROVEMENT SCHEMES

As shown in Table 3, to verify the effectiveness of different improvement schemes (ASPP of fused sensory field blocks, attention mechanism, and multi-scale feature fusion) on the self-built tangka dataset, this paper uses mIoU, mPA, and accuracy as evaluation metrics to compare and analyze the impact of each improvement point on the model.

The DeepLabv3+ model using the migration learning method is set as the base network, and the comparison between the first and second rows of Table 3 shows that the network structure with the addition of the RFBA module improves the mIoU by 0.4%, and the mPA and accuracy

![Figure 8](image-url)

**FIGURE 8.** Comparing the visualization results of the model on the thangka dataset.

**TABLE 1.** Performance comparison of different models on portrait-like thangka dataset.

| Models       | backbone | mPA   | mIoU  | Accuracy | mRecall |
|--------------|----------|-------|-------|----------|---------|
| SegNet       | Xception | 76.31 | 67.79 | 84.56    | 76.31   |
| PSPNet       | Xception | 81.59 | 69.50 | 85.08    | 81.59   |
| DeepLabv3+  | Xception | 90.34 | 94.87 | 93.52    | 90.49   |
| Proposed     | Xception | 90.75 | 85.66 | 93.84    | 90.75   |

**TABLE 2.** mIoU comparison of segmentation results for different classes.

| Models/Class | Backlight | pedestal | Head Light | figure | background |
|--------------|-----------|----------|------------|--------|------------|
| SegNet       | 0.65      | 0.53     | 0.70       | 0.71   | 0.80       |
| PSPNet       | 0.68      | 0.58     | 0.70       | 0.71   | 0.81       |
| DeepLabv3+   | 0.85      | 0.75     | 0.86       | 0.87   | 0.91       |
| Proposed     | 0.86      | 0.75     | 0.88       | 0.88   | 0.91       |
by 0.1% and 0.2%, respectively. The method can reasonably simulate the perception of human eye vision and effectively improve the accuracy of segmentation. As seen from the sixth column of Figure 10a and b, compared with the original model, the problem of wrong segmentation and missed segmentation has been significantly improved by adding the RFBA module.

By comparing the first and third rows of table 3, as can be seen, that embedding the attention mechanism in the model improves mPA and accuracy by 0.3% and 0.2%, respectively, indicating that embedding the attention mechanism in the model can enhance the learning ability of the network and improve the segmentation accuracy. Comparing the first and fourth rows in table 3 shows that multi-scale feature fusion has a positive effect on the segmentation accuracy of the model, which confirms that fusing low-level feature information with high-level feature information can capture richer information. The comparison between the first and fifth rows shows that the model proposed in this paper improves mIoU by 0.79%, mPA by 0.41%, and accuracy by 0.32% compared with the base model, which effectively improves the accuracy of tangka image segmentation and alleviates the problems of wrong segmentation, missed segmentation, and rough segmentation edges of the original network. The visualization results of each ablation experiment are shown in Figure 10 from left to right, the original image, the actual value image, the DeepLabv3+ primary network, the addition of the feature fusion module, the addition of the improved attention mechanism module, the addition of the RFBA module, the network proposed in this paper, and the last column is the result of cropping the mask segmented by the network proposed in this paper in the original image, where the first three rows are the visualization result of the portrait class thangka images. The last two rows are the visualization result of other types of images in the dataset.

We use the confusion matrix in Figure 11 to analyze the network response to the individual classes. Just as its name implies, the confusion matrix measures the accuracy of a classifier’s classification. The diagonal entries correspond to the correctly classified values for each class. We can see that 94% of the Buddha figure pixels in the second row of Figure 11 are truly predicted to be of the Buddha class, with a 4% probability of the Buddha figure being misclassified as background and a 1% probability of being misclassified as pedestal and backlight. From the figure, we can see that the classifier has a better classification effect for each class.

B. PERFORMANCE ANALYSIS OF PASCAL VOC2007 DATASET

Experiments are conducted on the PASCAL VOC2007 dataset to verify the generalization ability of our algorithm before and after the improvement, consisting of 20 categories, with a total of 10,828 images in the training set and 1,205 images in the validation set. The experiments in this chapter are completed based on DeepLabv3+, using Xception as the feature extractor. The experimental parameters are consistent with the previous experiments, and mIoU and mPA are used as evaluation indicators. The experimental results are shown in Table 4. It can be seen that the mIoU of the algorithm proposed in this paper is increased by 2.04% compared with the original network, and the mPA is increased by 2.22%. It can be seen from Table 5 that the algorithm proposed in this paper has improved the detection effect of each category in the data set, which confirms the improvement of this paper. The method can indeed improve the accuracy of the segmentation results.

The prediction results obtained by the model before and after the improvement in the validation set are shown in Figure 12. The first row shows the original image, the second row shows the annotation of the image, and the third row shows the segmentation results of the DeepLabv3+ model.
TABLE 4. Comparison of segmentation accuracy of PASCAL VOC2007 dataset.

| Method   | backbone | params   | mPA   | mIoU   |
|----------|----------|----------|-------|--------|
| DeepLabv3+ | Xception | 41258213 | 87.03 | 72.23  |
| Proposed  | Xception | 41517335 | 89.25 | 74.27  |

TABLE 5. mIoU comparison of segmentation results for different classes.

| Class/Module | DeepLabv3+ | Proposed |
|--------------|-------------|----------|
| Tvmonitor    | 0.67        | 0.69     |
| Train        | 0.89        | 0.89     |
| Sofa         | 0.56        | 0.56     |
| Sheep        | 0.72        | 0.76     |
| Pottedplant  | 0.39        | 0.42     |
| Person       | 0.80        | 0.80     |
| Cow          | 0.67        | 0.80     |
| Horse        | 0.79        | 0.85     |
| Dog          | 0.84        | 0.87     |
| Motobike     | 0.76        | 0.75     |
| Chair        | 0.48        | 0.56     |
| Cat          | 0.84        | 0.85     |
| Car          | 0.86        | 0.86     |
| Bus          | 0.91        | 0.92     |
| Bottle       | 0.64        | 0.70     |
| Boat         | 0.55        | 0.60     |
| Bird         | 0.83        | 0.85     |
| Becycle     | 0.56        | 0.69     |
| Aeroplane    | 0.79        | 0.80     |
| Diningtable  | 0.57        | 0.61     |

FIGURE 12. Visualization results of the model on the PASCAL VOC dataset.

before the improvement, the fourth row shows the segmentation results of the algorithm proposed in this paper, and the fifth row shows the cropping of the segmentation mask proposed in this paper on the original image. From the first two columns, we can see that the original model has the problem of missed segmentation. For example, the hand of the character in the third row of the first column and the neck of the bird in the third row of the second column, the network proposed in this paper can effectively solve the problem of missed segmentation in the original model; from the third and fourth columns, as can be seen, DeepLabv3+ model has more mis-segmentation, and in the third row of the third column, because the color of the ground is similar with that of the horse. Hence, the original network identifies the ground as the horse as well. In the third row of the fourth column, the original model incorrectly identifies the computer display, while the model proposed in this paper can well solve the mis-segmentation problem in the original model; from the fifth and sixth columns, the segmentation effect of the original model on the feet of birds, and such tiny objects as distant airplanes is not satisfactory. In contrast, the model proposed in this paper, through multi-scale feature fusion and improvement of the perceptual model field, retains more edge details. The proposed model retains more edge details and improves the segmentation ability of the model for small objects.

VI. CONCLUSION

The thangka images contain rich contents, and it not only has great significance to Chinese civilization but also have a profound influence on the history of world civilization. Due to factors such as preservation environment and fallen painting materials, the surviving thangka images have various problems such as cracks and defects, and how to repair the defects in the images is one of the difficulties faced by current researchers.

This paper provides pixel-level semantic annotation for the portrait-like thangka dataset, which can provide experimental data for scholars to understand the content of thangka images. Using the powerful learning ability of deep learning to segment the thangka images is vital for studying the rich connotation of thangka images and the subsequent restoration of thangka images. Aiming at the problems of wrong segmentation, missed segmentation, and rough edges of segmentation targets in Tangka image segmentation, an improved DeepLabv3+ Tangka image segmentation model is proposed, which improves the original ASPP module, adds the DCBAM attention module, and multi-scale feature fusion by connecting the conventional convolution of different scales followed by the atrous convolution of different atrous rates, which effectively improves the algorithm’s image segmentation. Its specially designed search and identification modules can automatically learn rich features, which are crucial for overcoming challenging ambiguities in object boundaries. The experimental data show that the improved model has 85.66% mIoU, 90.75% mPA, and 93.84% accuracy in the self-built thangka image dataset, which has significantly improved the segmentation accuracy compared with the base DeepLabv3+ and improved the final prediction map. The three improvements made in this paper can be used as a general design idea in other networks and have solid practical significance for image segmentation with complex content and color blending.

In future work, we plan to extend our dataset to provide input of various forms, for example, object detection. And we will solve the problems of long network training time and lots of parameters by adjusting the network parameters.

CONFLICTS OF INTEREST

We declare that we have no conflicts of interest.
REFERENCES

[1] J. Qian and W. Wang, “Main feature extraction and expression for religious portrait Thangka image,” in Proc. 5th Int. Conf. Young Comput. Scientists, Hunan, China, Nov. 2008, pp. 18–21.

[2] K. K. Rahimi and S. S. Sudha, “Review of image segmentation techniques: A survey,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 4, pp. 842–845, Jul. 2014.

[3] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst., Man, Cybern., vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[4] P. Saint-Marc, J.-S. Chen, and G. Medioni, “Adaptive smoothing: A general tool for early vision,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 6, pp. 514–529, Jun. 1991.

[5] R. Adams and L. Bischof, “Seeded region growing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 6, pp. 641–647, Jun. 1994.

[6] J. F. Lu, H. Lin, and Z. Pan, “Adaptive region growing algorithm in medical images segmentation,” J. Comput. Aided Des. Comput. Graph., vol. 17, pp. 2168–2173, Oct. 2005.

[7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2014, arXiv:1412.7062.

[9] L. C. Chen, G. Papandreou, and I. Kokkinos, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Jun. 2017.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 801–818.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI), vol. 9351, Cham, Switzerland: Springer, Nov. 2015, pp. 234–241.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.

[14] T. Ihaya, M. Urmia, A. Khan, and H. Kim, “DAM: Hierarchical adaptive feature selection using convolution encoder decoder network for strawberry segmentation,” Frontiers Plant Sci., vol. 12, Feb. 2021, Art. no. 591333.

[15] T. Ihaya, A. Khan, M. Urmia, Y. Jeong, and H. Kim, “Multi-scale context aggregation for strawberry fruit recognition and disease phenotyping,” IEEE Access, vol. 9, pp. 124491–124504, 2021.

[16] L. Baqrijuan, W. Weilan, H. Wenjin, and L. Wenbin, “Damaged regions segmentation on Thangka image combining color and texture features,” Int. J. Digit. Content Technol. Appli., vol. 10, pp. 131–143, May 2016.

[17] B. Luo, W. Wang, Y. Jia, and W. Gao, “A segmentation method for spotted-partten damaged Thangka image combining grayscale morphology with maximum entropy threshold,” in Proc. 6th Int. Cong. Image Signal Process. (CISP), Hangzhou, China, Dec. 2013, pp. 561–565, doi: 10.1109/CISP.2013.6744060.

[18] H. M. Liu, “Damaged region segmentation of Thangka using combining cloud model and proportion threshold,” J. Fuyang Teachers College, vol. 33, pp. 62–67, Jan. 2016.

[19] X. Fei, C. S. Liu, and F. Miao, “Headdress segmentation algorithm for Thangka portrait image based on circle feature location,” Comput. Telecommun., vol. 9, pp. 27–31, Sep. 2015.

[20] Y. Zhang and W. Wang, “Automatic annotation of portrait-like Thangka images based on headdress features,” J. Northwest Univ. Nationalities, vol. 35, pp. 16–20, Feb. 2014.

[21] X. Yan and W. Wang, “The Thangka image segmentation algorithm based on mean shift,” Autom. Instrum., vol. 1, pp. 130–131, Jan. 2016.

[22] L. Yin and W. Wang, “Headdress detection based on saliency map for Thangka portrait image,” in Proc. MVA, Nara, Japan, Jun. 2011, pp. 13–15.

[23] J. Meng, W. Hu, L. Jia, G. He, and P. Xue, “A semantic segmentation model for headdresses in Thangka image based on line drawing augmentation and spatial prior knowledge,” IEEE Sensors J., vol. 21, no. 22, pp. 25161–25170, Nov. 2021.

[24] S. Liu and D. Huang, “Receptive field block net for accurate and fast object detection,” in Proc. ECCV, Munich, Germany, 2018, pp. 385–400.

[25] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Jan. 2009.

[26] B. A. Wandell and J. Winawer, “Computational neuroimaging and population receptive fields,” Trends Cogn. Sci., vol. 19, no. 6, pp. 349–357, Jun. 2015.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[28] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in Proc. ICCV, Venice, Italy, Oct. 2017, pp. 764–773.

[29] S. Woo, J. Park, Y. J. Lee, and I. S. Kweon, “CBAM: Deformable convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 3–19.

HAO WANG was born in Shaanxi, China. He is currently an Associate Professor with Xizang Minzu University, where he is also the Dean of the School of Information Engineering. His research interests include image processing and image encryption.

JINGYUN HU was born in Anhui, China. She is currently pursuing the degree with the School of Information Engineering, Xizang Minzu University. Her research interests include image segmentation and image processing.

RU XUE was born in Shaanxi, China. She is currently a Professor with Xizang Minzu University. Her research interests include image processing and image encryption.

YUE LIU was born in Chongqing, China. She is currently pursuing the degree with the School of Information Engineering, Xizang Minzu University. Her research interests include image segmentation and image processing.

GUANXIU PAN was born in Shaanxi, China. She is currently a Professor with Xizang Minzu University. Her research interests include image processing and image encryption.