Instance Segmentation Model of Wool Cashmere Fiber Based on Mask RCNN

Fen Li[1][0009-0006-2292-5460], Yu wang[2][0009-0004-6180-9222], Xin Shi[1][0009-0000-4747-3727], Kewei Chen[1][0000-0002-3106-4336], Fangyan Dong[1*][0000-0001-5958-5734]

[1]Faculty of Mechanical Engineering & Mechanics,Ningbo University,Ningbo 315211,China
{2211090124,2211090061,chenkewei,dongfangyan} @nbu.edu.cn
*Corresponding author: Fangyan Dong, dongfangyan@nbu.edu.cn
[2]China Academy of Safety Science & Technology ,Beljing,100000,China
48745370@qq.com

Abstract: In order to improve the accuracy of instance segmentation of wool and cashmere fiber, this paper proposed an automatic instance segmentation fiber detection model STC-Mask RCNN based on Mask Region Convolutional Neural Network. Firstly, Swin Transformer is used as the backbone network of the model to improve the feature extraction ability of the model; then, the Convolutional Block Attention Module is introduced into the mask branch to guide the model to focus on the learning channel and spatial dimension; Finally, the ReLu activation function in the detection head and RPN network structure is replaced with the Mish activation function to improve the accuracy of the model. The experimental results show that:The AP value of STC-Mask RCNN is 89.4%, which is 2.1 percentage points higher than that of the baseline model Mask RCNN; Compared with the one-stage SOLOv2 and YOLACT models, the AP value is also improved, which proves that the STC-Mask RCNN model can improve the accuracy of wool and cashmere fiber instance segmentation.

Key words： Wool and cashmere fiber , instance segmentation, Swin Transformer, Convolutional Block Attention Module, Mish activation

1 Introduction

Cashmere fiber is slender and uniform, soft and plush, good surface gloss, known as the "king of fiber" and "soft gold", due to the rare production and high price[1]. Wool and cashmere are very similar in shape characteristics, physical and chemical properties, composition, but the price gap between them is wide. In the interest of driving, many delinquent businesses to cashmere fiber into the value of relatively low wool, shoddily good, adulterated, false standard cashmere and wool content ratio to pursue tremendous profits. Therefore, it is necessary to study the quantitative identification of cashmere and wool in related products.

At present, the quantitative identification methods of cashmere and wool in China mainly include: microscopic projection method, DNA identification method, staining identification method, near infrared spectroscopy, scanning electron microscopy, etc[2-6]. The microscopic projection method is currently the most widely used in third-party institutions, but the accuracy of this method is closely related to the experience of inspectors.

Since the rise of Convolutional Neural Network(CNN)[7] in 2012, researchers have begun to apply CNN to the field of image segmentation. CNN can be trained to automatically extract the feature information in the image, replacing the complicated process of extracting the feature information manually. Girshick et al.[8] proposed RCNN (Regions with CNN), which uses the strategy of RPN to train the model to realize object localization and image segmentation. The Mask RCNN network constructed by He et al.[9] significantly enhances the accuracy of image instance segmentation. With the development of machine learning, more and more researchers have applied deep learning methods to the detection and instance segmentation tasks of wool and cashmere images. Cong Mingfang et al.[10] tried to use Mask RCNN algorithm to realize instance segmentation of wool and cashmere microscopic images. Therefore, this paper improves Mask RCNN to achieve better instance segmentation effect of wool and cashmere.

2 Mask RCNN algorithm

Mask RCNN is a two-stage algorithm capable of performing multiple tasks, including target classification, detection, and instance segmentation. It extends the Faster RCNN[11] framework by incorporating a branch composed of fully convolutional

neural networks to predict target masks. This mask branch is utilized to perform instance segmentation on the regions of interest provided by the RCNN network, enabling simultaneous completion of tasks such as target classification, localization, and segmentation. The network structure of Mask RCNN model is shown in Fig. 1. The structure of the model is simple, and only a small amount of calculation is added to realize the detection of image objects and generate a high-quality mask for each object.

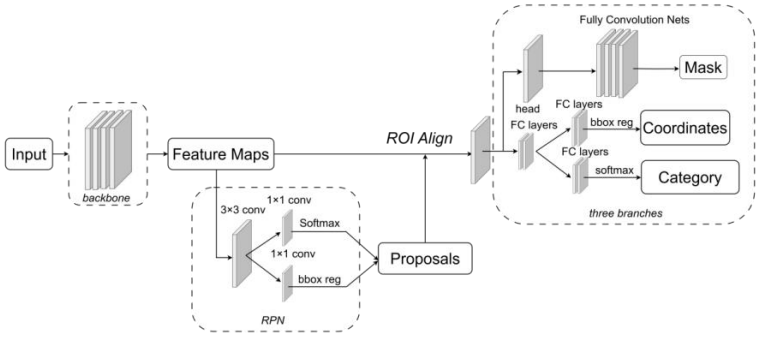


Fig. 1. Mask RCNN network structure

Mask RCNN is a two-stage network structure. In the initial stage: the image is first fed into the backbone network for feature extraction;furthermore, the feature pyramid's hierarchical and lateral structure is leveraged to integrate image features across multiple scales, enabling the incorporation of semantic and positional information at varying scales; then, a predetermined number of anchor boxes are generated for each pixel on the feature map provided for the FPN structure. The intersection between each anchor box and the actual box marked on the image is then calculated to derive multiple candidate regions of interest (ROI) with varying sizes; The regional proposal network (RPN) is ultimately utilized to execute binary classification and boundary regression operations on the candidate ROI. Subsequently, ROIs with low classification scores are pruned, while maintaining a positive-to-negative sample ratio of 1:3. This approach effectively mitigates the impact of category imbalance on network accuracy and reduces unnecessary computational overhead in the second stage. The second phase commences with the implementation of two alignment operations: (1) Realign the ROI provided in the initial stage and map the spatial coordinates of the ROI in the original image to the corresponding pixels in the feature map; (2) Convert the ROI of different sizes on the

feature map into a fixed size. Additionally, in order to minimize the error caused by quantization during the alignment process, the bilinear interpolation method is utilized to calculate the value of each pixel from adjacent grid points on the feature map. This enables extraction of essential feature information within ROI and facilitates accurate completion of classification, regression, and segmentation tasks. Ultimately, an additional segmentation branch is integrated into the fully connected layer to execute classification and regression on the pixels within each Region of Interest (ROI), thereby producing a mask for each individual object.

3 Improve Mask RCNN algorithm

3.1 Introducing the Swin Transformer for feature extraction network

Mask RCNN is an instance segmentation network based on convolutional neural network implementation. Nevertheless, CNN networks commonly experience the following challenges : (1) Due to the limitation of receptive field, the network lacks the coherent connection between features in the high-level semantic information. (2) The performance of the instance segmentation mask and the inference speed of the network are heavily reliant on the object detector, leading to poor network performance in complex scenes. To address the aforementioned issues, this paper employs Swin Transformer[12] as the feature extraction network for Mask RCNN in order to enhance the backbone network's capacity to incorporate non-local information and capture global semantic features from high-dimensional image data.

Based on the number of heads in the multi-head attention module, the Swin Transformer model is categorized into four structures: Swin-T, Swin-S, Swin-B, and Swin-L. Compared with other Swin Transformer models, Swin-T can use fewer parameters to achieve the same accuracy. Therefore, this paper conducts experimental research based on Swin-T. The network architecture of Swin-T consists primarily of four components: patch partition layer, linear embedding layer, Swin Transformer Block, and Patch Merging layer. The overall structure is illustrated in Fig. 2. During the process of model training, the image undergoes initial processing by the patch partition layer to partition and flatten it. Subsequently, the flattened feature map is subjected to linear transformation in the linear embedding layer to handle channel data for each pixel. The resulting transformed data then enters the Swin Transformer

Block for attention calculation. Finally, after three iterations of patch merging layers and Swin Transformer Block stacks, the computed feature map is generated.

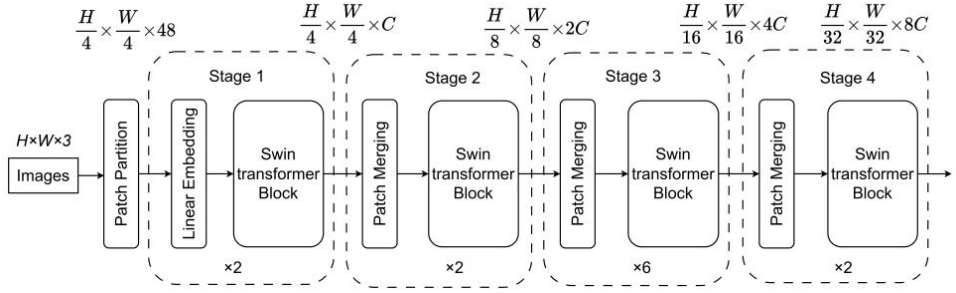


Fig. 2. Swin-T network structure

The patch partition layer primarily conducts down-sampling processing on the input feature image, as illustrated in Fig. 3. Initially, each 2×2 adjacent pixels of the input feature map are partitioned into a patch, and the pixels at corresponding positions within each patch are extracted and concatenated along the depth dimension. Subsequently, the concatenated feature map is normalized using LayerNorm layer. Finally, the normalized result is fed into the fully connected layer for linear transformation along the depth dimension to obtain the ultimate output. The width and height of the resulting feature map are halved compared to the input feature map, while the depth is doubled.

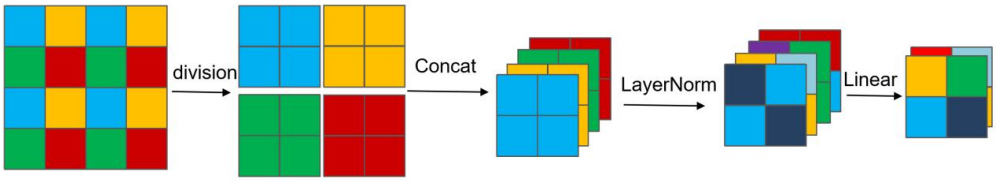


Fig. 3. Schematic diagram of the patch partition layer

The Swin Transformer Block consists primarily of LayerNorm (LN), windows multi-head self-attention (W-MSA), shifted windows multi-head self-attention (SW-MSA), and multilayer perceptron (MLP). The specific structure is illustrated in Fig. 4.

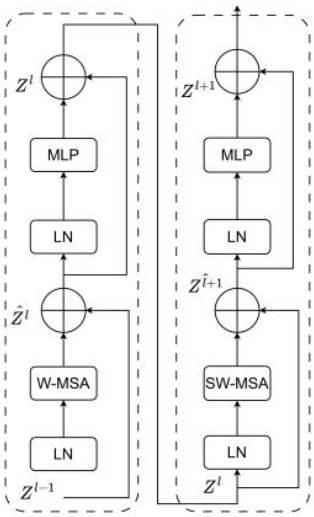


Fig. 4. Swin Transformer Block

The LayerNorm is responsible for normalizing the feature map of the input network; W-MSA conducts self-attention operations within each non-overlapping window. In comparison to MSA, it significantly reduces the computational load of the model, but the information transmission between windows is constrained. SW-MSA builds upon W-MSA by introducing shifted window and subsequently performing block processing. The window offset after block processing is recombined with the original W-MSA window shape, followed by attention mechanism operation within a single window. This approach addresses the challenge of information exchange between different Windows, and their integration can significantly enhance the model's feature extraction capability. The computational complexity of MAS and W-MSA can be determined using the following calculations:

$$\Omega(MAS) = 4hwC^2 + 2(hw)^2C \quad (1)$$

$$\Omega(W-MAS) = 4hwC^2 + 2M^2hwC \quad (2)$$

Where Ω denotes the computational complexity, h signifies the height of the feature map, w represents the width of the feature map, C indicates the depth of the feature map, and M stands for the size of each window.

The calculation principle diagram of W-MSA and SW-MSA is shown in Fig. 5.

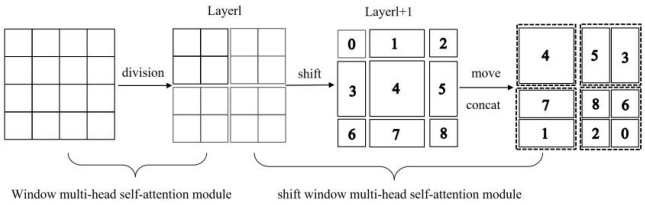


Fig. 5. Window multi-head self-attention module and shift window multi-head self-attention module

The self-attention mechanism is an important part of Swin Transformer Block, and its schematic diagram is shown in Fig. 6. Firstly, the input feature map undergoes linear transformation to convert it into two-dimensional sequence data. Subsequently, the Query array (Q), Key array (K), and Value array (V) are generated, followed by the addition of position encoding to capture positional information within the two-dimensional sequence. The Q, K, and V are then processed through the scaled dot product attention mechanism, and the resulting outputs are concatenated before being fed into a fully connected layer to obtain the final result as depicted in the following formula:

$$Attention(Q,\ K,\ V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Where d is the input dimension of the model; Q、K、V respectively is the linear transformation value of the feature map.

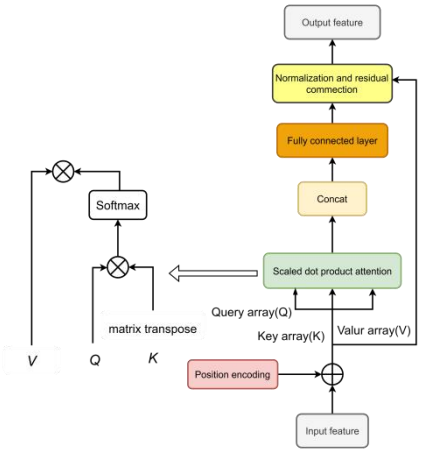


Fig. 6. Self-attentional mechanism

The multilayer perceptron partitions the input feature map based on category information, and its structural diagram is illustrated in Fig. 7. The multilayer perceptron is primarily composed of two fully connected layers, an activation function layer, and two dropout layers. Among these components, the Gaussian Error Linear Unit (GELU) serves as the primary activation function. In comparison to other activation functions, GELU enhances model integrity and accuracy while ensuring training stability. Dropout can prevent overfitting by randomly losing a subset of neurons.

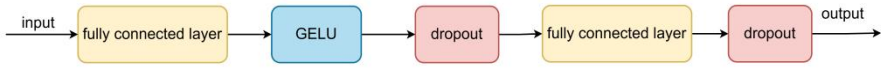


Fig. 7. Multilayer perceptron structure

3.2 Introduce the Convolutional Block Attention Module

In recent years, the attention mechanism has been widely employed in tasks such as image recognition and object detection[13-14]. Only a marginal increase in computational load is required to integrate this mechanism into the original network model, enabling automatic focus on pixel-rich areas during the training process. Mask RCNN incorporates a Fully Convolutional Network (FCN) to perform pixel-wise instance segmentation of the target within the candidate region. Due to the rough edge of wool fiber and the relatively smooth nature of cashmere fiber, FCN is unable to capture sufficient edge information from the cashmere fiber. Hence, a Convolutional Block Attention Module(CBAM) is incorporated into the Mask branch of the Mask RCNN network to direct the model's attention towards learning channel and spatial dimension dependencies.

The Convolutional Block Attention Module consists primarily of a Channel Attention Module and a Spatial Attention Module, as illustrated in Fig. 8. The attention mechanism is incorporated into the Mask RCNN network to enable the model to focus on pixels containing rich content, specifically those representing wool and cashmere. This effectively suppresses image noise and ultimately enhances the accuracy of segmenting wool and cashmere masks. The enhanced segmentation branch is delineated into four sequential steps, as depicted in Fig. 8 : (1) The 14×14 region of interest (ROI) obtained in the initial stage was fed into four convolutional layers with 3×3 convolution kernels to generate the feature map X, $X \in R^{W \times H \times C}$,where W、H、C denote the width, height, and channel number of the feature map respectively.

(2) The feature map X is input into the CAM module to amplify the model's focus on salient features within the channel, and the output channel directs attention to the feature map X_{cag}; (3) The input X_{cag} is fed into the SAM module to augment the model's attention towards significant features, and the resulting attention feature X_{sag} is guided in the output space; (4) The transposed convolution layer, equipped with a 2 ×2 convolution kernel, is employed to upsample X_{sag} to 28×28. Subsequently, the 1 ×1 convolution layer is utilized for pixel-wise category prediction on the feature map in order to derive the final wool and cashmere mask map. Fig. 9 illustrates the configuration of CAM and SAM.

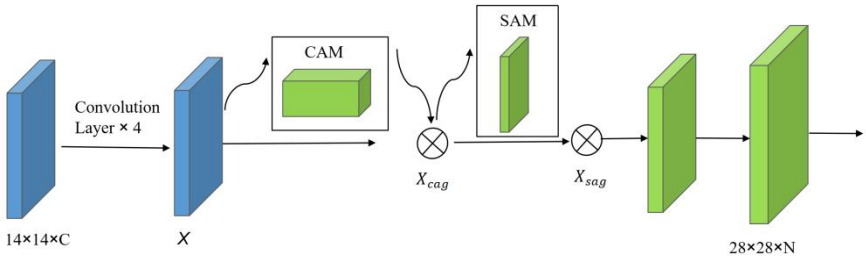


Fig. 8. Improved Mask RCNN segmentation branch

Within the CAM module, two channel descriptors X_{max} and X_{avg} are derived from the feature map X through parallel max pooling layers and global average pooling layers. Subsequently, these descriptors X_{max} and X_{avg} undergo element-wise summation via a shared Multilayer Perceptron (MLP) before being input into the sigmoid activation function to yield the weight coefficient (M_c), $M_c(X) = \sigma(MLP(X_{max}) + MLP(X_{avg}))$, where σ represents the sigmoid activation function. Finally, this obtained weight coefficient (M_c) is utilized to perform element-wise multiplication with the feature map in order to obtain the channel-guided attention feature map X_{cag}, $X_{cag} = M_c \otimes X$, Where ⊗ stands for element-wise multiplication.

Initially, the SAM module conducts average-pooling and max-pooling operations on the channel-guided attention feature map X_{cag} across their respective channels to obtain pooled feature maps P_{avg}、P_{max}. Subsequently, the corresponding elements of P_{avg} and P_{max} are added together and cascaded into the 3×3 convolutional layer, followed by application of the sigmoid activation function to derive weight coefficients (M_s), $M_s(X_{cag}) = \sigma(F_{3\times3}(P_{max} \circ P_{max}))$, Where $F_{3\times3}$ represents the 3

×3 convolutional layer, ∘ represents the cascade operation, and finally obtains the spatial attention guided feature map X_{sag},$X_{sag} = M_s(X_{cag})\otimes X_{cag}$.

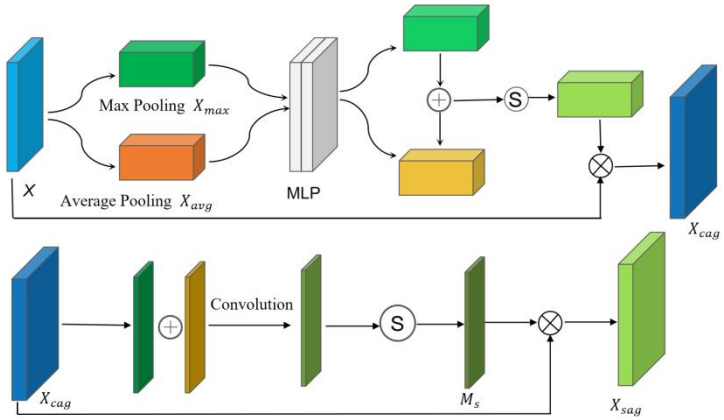


Fig. 9. CAM and SAM structure diagram

3.3 The optimization of activation functions.

The utilization of diverse activation functions enables a more effective capture of the non-linear relationships within the network, thereby enhancing the model's expressive capacity for complex data. This study replaces the detection head of Mask RCNN and the ReLu activation function in the RPN network structure with Mish[15], an alternative activation function that exhibited superior performance. The functional expression is as follows:

$$Mish = xtanh(ln(1 + e^x)) \quad (4)$$

The comparison of activation function curves is depicted in Fig. 10. In contrast to the Relu activation function, the Mish activation function demonstrates a higher degree of smoothness and exhibits continuous derivatives across both positive and negative intervals, thereby enhancing its ability to capture nuanced variations in input data. Furthermore, the Mish activation function yields non-zero gradients within the negative value region, effectively mitigating issues related to vanishing gradients during training. This characteristic facilitates improved feature extraction and contributes to enhanced training efficiency and model performance.

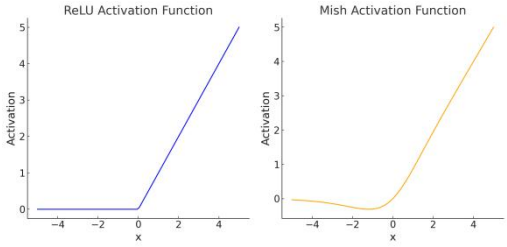


Fig. 10. Comparison of Relu and MIsh activation function curves

4 Experimental results and analysis

4.1 Experimental environment deployment and Parameter setting

The experimental operating system is Windows 10, and the development environment for the deep learning framework utilizes Python 3.9, CUDA 11.3, and PyTorch 1.10.1. The training process is carried out on an NVIDIA GeForce RTX 3080 Ti graphics card with a video memory of 12GB. The hyperparameters employed for training are detailed in Table 1.In this experiment, the initial learning rate is set to 0.0002, the batch size is set to 4, and the number of iterations per round is 3000.

4.2 Experimental data set preparation

The experimental dataset utilized in this study comprises 3125 images of various microfibers of wool and cashmere blend, sourced from the Ningbo Fiber Inspection Institute. Initially, Labelme is employed for annotating the dataset, utilizing a series of anchors to create polygons outlining the target fibers depicting wool and cashmere. Subsequently, labels for cashmere and wool are added to the annotated data, which is then saved in json format.

The dataset is partitioned randomly into training, validation, and test sets in a 7:2:1 ratio. The specific number of fibers in each set is detailed in Table 1.

Table 1. Fiber data statistics of the wool cashmere dataset

	Mixed image	wool	cashmere
training set	2188	3776	3281
validation	625	1135	957
test	312	612	473
Total	3125	5523	4648

4.3 Evaluation index

The instance segmentation algorithm primarily aims to segment distinct object instances of interest, fundamentally addressing the challenge of pixel classification and recognition. This study adopts average precision (AP) and mean average precision (mAP) as the evaluation metrics for assessing the wool and cashmere instance segmentation model.

4.4 Model training

Referring to the transfer learning method, the Swin Transformer backbone network is initialized by using the weights of the Swin Transformer model trained on the large dataset ImagNet during training, so that the parameters of the feature extraction network are relatively optimal and the model is avoided overfitting. Then, the Swin Transformer is fine-tuned on the data together with other structures to finally realize the instance segmentation of wool and cashmere.

4.5 Experimental results and analysis

Ablation experiments. To assess the impact of the newly introduced Swin-T model on network performance, various backbone network models are compared. This study selects ResNet50 and ResNet101 as the control group for the backbone network model based on the overall structure of the model. ResNet50 and ResNet101 consist of multiple convolution modules and residual modules, commonly utilized as backbones for Mask RCNN networks. Ablation experiments are conducted to verify the effect of integrating CBAM attention mechanism into the mask branch and replacing the Mish activation function on model performance, using Swin-T as the backbone network. The specific experimental results are presented in Table 2.

Table 2 Results of ablation experiments

Backbone network	CBAM	Mish	AP/%		mAP_{all}/%
			Wool	cashmere	
ResNet50	×	×	89.2	85.4	87.3
ResNet101	×	×	89.7	86.5	88.1
Swin-T	×	×	89.9	86.7	88.3
Swin-T	√	×	90.6	87.7	89.2
Swin-T	√	√	90.8	87.9	89.4

Note: × indicates that the factor is not included, and √ indicates that the factor is included. All AP values in the table are computed at IOU=0.5, and mAP_{all} denotes the mean average precision across all classes at IOU=0.5.

The data in the table demonstrates that utilizing Swin-T as the backbone model of the network leads to a significant enhancement in instance segmentation accuracy compared to using ResNet50 and ResNet101. Specifically, there is an increase of 1.6% and 0.2% in AP value for the model, and a respective increase of 0.7% and 0.2% for wool, as well as a rise of 1.3% and 0.2% for cashmere when Swin-T is employed as the backbone network. Furthermore, integrating CBAM into the mask branch results in a notable improvement with a 0.9% increase in AP value for cashmere and wool due to enhanced attention on learning space and channel attention, leading to more accurate prediction of mask edges for cashmere and wool instances. Additionally, replacing the Mish activation function enhances the model's AP value by 0.2%. These experiments collectively validate the effectiveness of our improved model.

Model comparison experiment. The STC-Mask RCNN model is contrasted with the instance segmentation models namely Mask RCNN, SOLOv2, and YOLACT. Mask RCNN, SOLOv2, and YOLACT all enlist ResNet50 as the backbone network. Additionally, all models assimilate the identical experimental milieu and training tactics as those in this investigation, and the experimental yields are depicted in Table 3

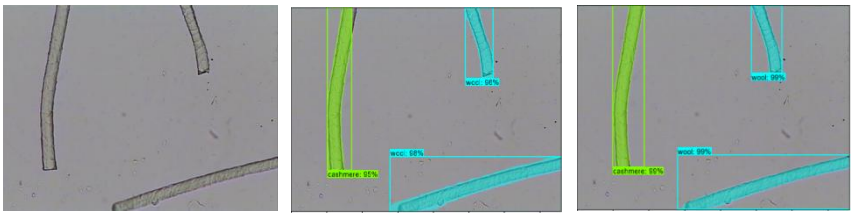
Table 3. Comparative experimental results of different instance segmentation models

Model	AP/%		mAP_{all}/%
	wool	cashmere	
SOLOv2	88.5	83.7	86.1
YOLACT	86.7	82.5	84.6
Mask RCNN	89.2	85.4	87.3
STC-Mask RCNN	90.8	87.9	89.4

From the data presented in the table, it is evident that the value of STC-Mask RCNN is 2.1 percentage points higher than that of the baseline model Mask RCNN. Specifically, the AP value of wool has increased by 2.5 percentage points, and that of cashmere has risen by 1.6 percentage points. In comparison with the SOLOv2 and YOLACT models, the accuracy has improved by 3.4 and 4.8 percentage points

respectively, thereby demonstrating the effectiveness of the improved model in this paper.

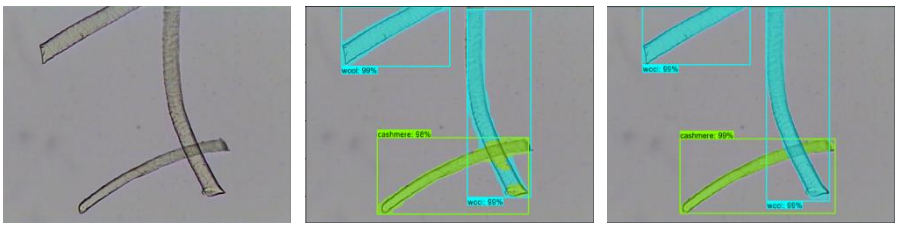
To further validate the efficacy of the STC-Mask RCNN model in enhancing the segmentation accuracy of wool and cashmere fiber instances, this paper randomly selected one image from each validation set for comparative analysis. Fig.11 presents the prediction results without crossed fibers. In contrast to the baseline model Mask-RCNN, the STC-Mask RCNN model can not only elevate the prediction score of the model, but also enhance the quality of the Mask, particularly improving the mask segmentation quality of the edges of wool and cashmere fibers.



(a) Original image (b) Mask RCNN (c) STC-Mask RCNN

Fig. 11. Effect of easily detected image prediction

Mixed fiber images of wool and cashmere frequently overlap, presenting a relatively high level of difficulty in detection. Such images are selected in the validation set for prediction and comparison, and the comparison results are presented in Fig. 12.



(a) Original image (b) MasK RCNN (c) STC-Mask RCNN

Fig. 12. Prediction effect of hard-to-detect images

From the predicted figure, it can be observed that when the baseline model Mask RCNN is employed to predict the cross fibers of wool and cashmere the category prediction at the fiber intersection is indistinct, and the instance segmentation accuracy of the model at the edge of the fibers is not high. STC-Mask RCNN ameliorates the above two circumstances exceedingly well and generates a pixel-level mask for the cross fibers.

5 Conclusion

In response to the issue of low accuracy in wool and cashmere instance segmentation, this study presents an automated fiber recognition model for instance segmentation, named STC-Mask RCNN, based on Mask Region Convolutional Neural Network . This paper primarily enhances the model in the following aspects: (1) incorporating Swin Transformer as the backbone network to enhance feature extraction capability; (2) introducing Convolutional Block Attention Module into the mask branch to guide the model's focus on learning channel and spatial dimension; (3) replacing ReLu activation function in detection head and RPN network structure with Mish activation function to improve model accuracy.

The experimental results on the wool and cashmere dataset demonstrate that employing Swin Transformer as the backbone network of the model leads to a 1% increase in the AP value. Furthermore, integrating the Convolutional Block Attention Module into the mask branch results in a 0.9% improvement in AP value, and the AP value of the STC-Mask RCNN model has witnessed an augmentation of 2.1%. Compared with other models, our proposed model achieves superior accuracy and significantly enhances wool and cashmere instance segmentation precision.

The STC-Mask RCNN model is utilized for precise instance segmentation of wool and cashmere, generating corresponding mask images. While it can accurately detect the quantity of wool and cashmere fibers, specific parameters such as fiber diameter have not been investigated. In future research, the acquired mask image is employed to automatically measure the fiber diameter, thereby providing a reference for the measurement of the fiber diameter of wool and cashmere fibers.

References

1. [1] GU, H., Wang, M., Yan, Chang, et al. (2017). Review on the Detection Methods of Cashmere Fiber. *Chinese Journal of Cilia*, (01), 69-71.
2. Mao, X. F., Lin, S. J., Hu, J. M., et al. (2010). Discussion on the strategies and methods of identification of cashmere and wool fiber. *China Fiber Inspection*, (09), 50-55.
3. Li, J. (2016). A Summary of Cashmere Inspection. *Chinese Fiber Inspection*, (12), 83-85.
4. Li, Y. H., Ma, Y., & Li, Y. Q. (2012). A Summary of Detection Method of Cashmere, Wool Fiber Identification. *China Fiber Inspection*, (11), 58-61.

5. Han, Y. Y., Wang, G. P., Shi, G. Q., et al. (2022). Analysis of inspection and identification methods of cashmere and sheep wool. *Textile Report*, 41(10), 28-30.
6. Luo, J. L., Lu, K., Zhang, P. P., et al. (2021). Current situation and prospect of identification methods of cashmere and wool. *Wool Textile Science and Technology*, 49(10), 112-117.
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems* (pp. 1097-1105). Curran Associates Inc.
8. Girshick, R., Donahue, J., Darrell, T., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8).
9. He, K., Gkioxari, G., Dollar, P., et al. (2017). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 1-12.
10. Cong, M. F., Li, Z. Y., Lu, Y., et al. (2022). Cashwool fiber recognition technology based on Mask R-CNN deep learning. *Modern Textile Technology*, 30(02), 36-40+47. https://doi.org/10.19398/j.att.202104002
11. Ren, S., He, K., Girshick, R., et al. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137-1149.
12. Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022). Washington, D.C.
13. Hu, J., Shen, L., Albanie, S., et al. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 2011-2023.
14. Woo, S., Park, J., Lee, J. Y., et al. (2018). CBAM: Convolutional Block Attention Module. In D. M., L. A., & G. F. (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 3-19). Cham: Springer.
15. Misra, D. (2019). Mish: A self regularized non-monotonic neural activation function [EB/OL]. Retrieved February 6, 2022, from https://arxiv.org/abs/1908.08681.