Recognize highly similar sewing gestures by the robot

Sijie Yang, Xiaohua Wang and Wenjie Wang

Abstract
The autonomous and efficient learning of sewing gestures by robots will bring great convenience to the garment industry. To improve the accuracy of robots in detecting sewing gestures with high similarity, three detection models based on deep learning are proposed in the paper. First, in order to improve the detection accuracy and detection speed of sewing gestures under complex backgrounds, we added a dense connection layer to the low-resolution network layer of YOLO-V3 to enhance the transmission and reuse rate of image features. Secondly, a deeper ResNet50 residual network is introduced to replace the VGG16 basic network in the original SSD model. The feature pyramid structure is used to fuse high-level semantic features and low-level semantic features, which can improve the detection accuracy of small-sized sewing gestures. Finally, the parallel spatial-temporal dual-stream network separately extracts the temporal feature and the spatial feature of sewing gestures. The fusion of time feature and space feature improves the detection accuracy of the coherent sewing gesture. The results show that the suggested three models can effectively detect four sewing gestures with high similarity. Among them, the spatial-temporal two-stream convolutional neural network has the highest detection accuracy. The improved SSD model has faster detection speed than the improved YOLO-V3 model and other mainstream algorithms.

Keywords
Sewing gestures, target detection, feature extraction, behavior recognition

Date received: 20 April 2021; accepted: 14 January 2022

Introduction
In the textile field, the fabric of the garments is soft and extensible, which makes the sewing process difficult. Many factors such as the human body and the environment need to be considered during the sewing process. Cooperating with robot technology and skilled worker technology can achieve better sewing performance. Collaborative robots accurate and rapid understanding of workers’ sewing skills is the key to human-machine collaboration.

Scholars have done a lot of research on how to solve the problem of robots’ learning of sewing skills. Takashi applied the variable gain learning control method to automatic sewing trajectory tracking of the robot arm. Panagiotis used fuzzy logic to define the expected environmental feedback, designed a hierarchical control system, and realized the robot’s adaptive sewing function to unknown sewing materials. Paraskevi combined visual servo networks and neural networks to design an adaptive reasoning learning system for curved fabric sewing robots. Huang et al. divided the worker’s sewing actions into segments, each segment was modeled by the GMM model, and GMR was used to obtain the entire sewing action to...
complete the learning of human sewing actions. The above methods can solve the perturbation of robot parameters and the uncertainty of the contact stiffness of the external working environment to a certain extent.

The sewing action recognition method based on computer vision detects various action categories by extracting detailed feature information in the sewing action. At this stage, human action recognition methods mainly include image samples based on a single keyframe and video samples. The image recognition method based on a single keyframe is fast and easy to implement. It can well detect some actions with representative keyframes. The recognition method based on video samples obtains the information of the action space and time dimension very well. This method is more flexible and scalable. It has a good recognition effect for complex and coherent actions. Generally, characteristics of the action-image are obtained by the method based on artificial design features and the method based on deep learning features. Yilmaz and Shah extracted the specific information of the action according to the change of the space-time volume of the moving target in the time series, which has good robustness against changes in the viewing angle. Gorelick et al. incorporated the structural positional relationship in the time and space of the action into the characteristics. Jiang et al. used the information of key points to construct action descriptors and clustered key points and analyzed actions by establishing judgment points to construct action descriptors and clustered key points and analyzed actions by establishing judgment points to construct action descriptors and clustered key points. Matikainen et al. found that the description of the outline of the action alone cannot well represent the texture features in the middle of the outline. Therefore, the representation method of local features is more important. Liu et al. used spatial-temporal interest point detection methods (STIPs) for target detection, which provides new ideas for local representation methods. Matikainen et al. divided the video samples into spatial frames, and the divided spatial frames were sequentially sent to the convolutional neural network. The method inputs the spatial information map of the decomposed video action sequence into the spatial stream convolutional network, the spatial stream convolutional neural network is used to extract the movement information of the limb trajectory in a series of time series. The fusion of temporal and spatial features greatly improves the accuracy of gestures recognition.

Four sewing gestures of inner overlock seam, hemmed seam, cladding seam, and cut fabric are very similar. Due to the complex environment of the sewing workshop, the robots cannot accurately and quickly understand and recognize the sewing movements of the workers’ hands. Therefore accurately identifying similar sewing gestures in complex backgrounds is the focus of this article.

The rest of the paper is organized as follows: In section 2, the improved YOLO-V3 convolutional neural model is used to detect sewing gestures with complex backgrounds. In section 3, the improved SSD method is used to solve the problem of low accuracy of YOLO for small target sewing gesture detection. In Section 4, the spatial-temporal convolutional neural network method is used to solve the problem of poor detection of sewing gestures.
with high similarity of sewing gestures and strong continuity of motion in independent single-frame images. Section 5 shows the experimental results. Conclusions and future research work are in Section 6.

Sewing gesture recognition based on improved model of YOLO-V3 (You Only Look Once)

YOLO-V3 improved model

The YOLO-V3 target detection network converts some detection problems into regression problems. The network uses the DarkNet53 network as a basic network to extract features. The network divides each picture into multiple grids. Each grid predicts a target bounding box, a confidence score, and a category conditional probability. In the detection process, the grid that falls into the center of a sewing gesture is used to identify the sewing gesture target. When multiple bounding boxes detect the same target at the same time, the non-maximum suppression (NMS) method is used to select the best bounding box. The accuracy of a predicted bounding box is reflected by the size of confidence. The YOLO-V3 network model detection process is shown in Figure 1.

Due to the limitations of the Darknet53 network, simply increasing the network level will cause the gradient to disappear and explore, which leads to low accuracy in detecting sewing gestures with complex backgrounds. The paper adds DenseNet densely connected network to the lower resolution transmission layer of the original YOLO-V3 network, which can enhance the transmission of sewing gesture feature information and promote feature reuse and integration. Each layer in the dense connection of DenseNet can directly access the gradient from the loss function and the original input signal, which greatly solves the problem of gradient disappearance when the number of network layers is too deep. The transfer function in the form of $\text{BN-ReLU-Conv}(1 \times 1) + \text{BN-ReLU-Conv}(3 \times 3)$ is used in paper. The $1 \times 1$ convolution kernel in the transfer function prevents the input feature map from being too large.

The improved YOLO-V3 network model is shown in Figure 2.

Figure 1. The detection process of YOLOV3 network.

YOLO-V3 improved model training parameter settings. The training of the YOLO-V3 improved model uses the official website pre-training weight file to initialize the network weight parameters, which can speed up the model convergence. Specific training parameter values are shown in Table 1.
The stochastic gradient descent (SGD) network is used to optimize the network model. The learning rate is adjusted according to the number of iterations. The initial learning rate is 0.001. When the model is iterated to 8000 times and 9000 times, the learning rate decays by 10 times. We set burn-in to 1000 at the beginning of training. When the number of updates is less than 1000, the learning rate strategy changes from small to large. When the number of updates is greater than 1000, the learning rate adopts an update strategy from large to small.

In the bounding boxes of all tested sample data sets, we use the K-means cluster analysis method to find 9 a priori box dimensions suitable for the sample data. The K-means distance measurement formula is shown in formula (1).

$$D(box, centerd) = 1 - IOU(box, centerd)$$  

(1)

$D(box, centerd)$ is the distance from each label box to the center point of each cluster. $IOU(box, centerd)$ is the Intersection-over-Union between each label box and each cluster center.

Nine sets of prior frame dimensions are used in the training process: (121, 82), (163, 98), (158, 154), (205, 119), (239, 150), (203, 218), (302, 179), (316, 253), (324, 351). Arranging the size of the prior frame from small to large, and it is evenly divided into three feature maps of different scales. Feature maps with larger scales use smaller prior boxes. Finally, these prior boxes are used to detect sewing gestures.

Training of YOLO-V3 improved model. The improved YOLO-V3 model uses 4000 sewing gesture images. The model is trained 5000 times. The dynamic process of training is observed by drawing the loss curve. The corresponding loss value change curve is shown in Figure 3.

It can be seen from the curve in Figure 3: the loss value of the model decreases rapidly in the previous iterations, which means that the model fits quickly; when iterative training is 2000 times, the loss value of the model slows down; when the iteration reaches 5000 times, the loss value converges to 0.0025.

Table 1. Parameter values of YOLO-V3 improved model.

| Parameter name | Value   |
|----------------|---------|
| Batch          | 64      |
| Subdivisions   | 8       |
| Iterations     | 4000    |
| Steps          | 4000, 4500 |
| Scales         | 0.1, 0.1|
| Momentum       | 0.9     |
| Decay          | 0.005   |
| NMS            | 0.25    |
| Burn-in        | 1000    |

Figure 2. Improved network model of YOLO-V3.

Figure 3. The loss curve of the improved YOLO-V3 model.
The detection performance of the model is further evaluated by calculating the Avg IOU between the predicted bounding box and the true bounding box. The change curve of the Intersection-over-Union is shown in Figure 4. As the number of model iterations increases, the Intersection-over-Union of the real frame and the predicted frame is constantly improving. When the number of iterations is 5000, the model Intersection-over-Union tends to 0.9.

We select the best prediction model from trained models. By calculating the accuracy rate, recall rate, $F_1$ value, and average Intersection-over-Union under different confidence thresholds, the best threshold model is screened out. The result is shown in Figure 5. In the threshold interval of $(0, 1)$, the value is calculated once every span of 0.05, and a total of 20 sets of data are calculated. We set the following rules on the priority: accuracy rate $>$ recall rate $>$ IOU. After the threshold reaches 0.6, the accuracy rate gradually stabilizes, and the optimal range is about 0.6–1.0. In this range, the best recall rate is 0.85, the corresponding confidence threshold is 0.6, and the IOU value is about 0.71. Therefore, we choose 0.6 as the optimal threshold.

The improved YOLO-V3 model has a better detection effect on large-size sewing gestures, but the accuracy will decrease when recognizing smaller sewing gestures. The reason is that the single convolutional network in YOLO-V3 easily ignores low-level feature information. The high-level features can only provide a small part of the feature information about a small target to the model, which is not good for sewing gesture recognition with a small size.

Sewing gesture recognition based on improved SSD (Single Shot Multi-Box Detector) model

SSD improved model

The SSD network uses multi-scale target features for target detection, which can improve the robustness of targets of different scales. Compared with the YOLO-V3 network, the SSD network has a higher recognition rate for small-sized targets. The number of deep network layers can reduce the constraints of gradient instability problems such as gradient disappearance and gradient explosion. The SSD model based on VGG16 only deepens the network depth, which leads to the aggravation of network degradation.

We introduce deep residual network (ResNet) and feature pyramid structure (FPN) into the basic SSD model. The deep residual network (ResNet) improves the multiplicative transfer between feature layers to additive transfer, which can enhance the connection between the front and back network layers. While ResNet effectively avoids gradient instability, it does not increase the parameters and complexity of the model. The calculation formula of the deep residual network (ResNet) is shown in formula (2).
In formula (2), $l$ represents the number of network layers. $x_l$ represents the output of the layer $l$. $H_l$ represents a non-linear transformation. For the ResNet network, the output of the layer $l$ is a non-linear transformation of the output of the layer $l - 1$.

The feature pyramid structure is used to realize the fusion of high-level and low-level feature information (the feature pyramid structure is shown in Figure 7). The down-sampling part calculates the feature level composed of multiple scale feature maps, and the deepest layer has the strongest feature information. The up-sampling part extracts higher-resolution features, and up-sampled features are enhanced by down-sampled features through the horizontal connection. Each horizontal connection part realizes the fusion of feature information of different sizes in high and low layers, which strengthens the detection ability of small-size sewing gestures.

The input image of the SSD improved model is a $512 \times 512$ RGB image. Its backbone superimposes five convolution modules based on ResNet50 to form a deep network. The structure of the improved model of SSD is shown in Figure 8.

The network of the SSD improved model contains 59 convolutional layers and five maximum pooling layers. The shallow and middle layers of the model extract feature information of sewing gestures, which are used to detect small-sized sewing gesture targets. The deep network extracts the sewing gesture features that contain the full image, and these features are used to detect large-sized sewing gesture targets. SSD network draws feature maps from Conv4, Conv5, three additional convolutional layers, and one pooling layer respectively. Each feature map will generate a priori frames of multiple scales, and generate prediction frames of multiple scales based on different frame ratios to judge sewing gestures of different scales to achieve multi-scale target position and category prediction. The detection result is produced by non-maximum suppression.

**Parameter setting of SSD improved model.** The SSD improved model uses a self-built sewing gesture data set to train the model. In the training process, 64 samples are used as the number of samples in a batch, and parameters are updated once for each batch of samples. The number of samples in each batch is divided eight times, and the divided samples are sent to the trainer. The learning rate adopts a dynamic adjustment strategy, and the initial learning rate is set to 0.001. The real frame and default frame matching strategy adopt the maximum Intersection-over-Union strategy. If the Intersection-over-Union is greater than the threshold value of 0.5, matching is performed. The parameter settings are shown in Table 2.

We use the idea of transfer learning to apply the feature extraction capabilities learned by the network on the largest
dataset ImageNet to small sample data and fine-tune it to further improve the detection effect of sewing gestures. The training is based on the ImageNet super-large image data set and the trained ResNet50 classification network model. The fully connected layer after the classification network is removed, and the convolution detection module is added to form the target detection network. We migrate the convolution model and parameters in the ResNet50 network, and we freeze the parameters of each layer in the middle of the model. Finally, a deep migration training model based on ResNet50 can be obtained for training and fine-tuning parameters for model category prediction and position regression.

**Training of SSD improved model.** The SSD improved model optimizes the loss function during training. Regression training is carried out through the location and target category, and the back-propagation mechanism is used to continuously update the model so that the Loss value is continuously reduced. We observe the dynamic training process of the model by drawing the loss function curve. The model loss function curve is shown in Figure 9. The loss function in the model is defined as shown in formula (3).

\[
L(z,c,l,g) = \frac{1}{N} (L_{conf}(z,c) + \alpha L_{loc}(z,l,g))
\]  

Formula (3), \(L(z,c,l,g)\) is the position loss. \(L_{loc}(z,l,g)\) represents the loss function of the real frame and the predicted frame of the function. \(L_{conf}(z,c)\) represents the loss judgment on the multi-category confidence level \(c\). \(N\) is the number of a priori boxes matching the object box. \(z\) is the matching value of the default box and real object boxes of different categories. \(\alpha\) is the weighting factor of the relationship between confidence loss and location loss (the default value is 1). \(c\) is the confidence of the predicted object frame. \(l\) is the position information of the predicted object frame. \(g\) is the position information of the real object frame.

It can be seen from Figure 9 that the loss value of the improved SSD model during the pre-training period drops rapidly, and the model fits quickly. After 3000 iterations of training, the model loss decreases slowly. In order to prevent over-fitting, early-stopping is used in training, and the training is automatically ended when the loss value does not decrease for four consecutive epochs.

Both the SSD improved model and YOLO-V3 improved model only use spatial dimension information of a single frame to detect sewing gestures. These methods lack time-series information between adjacent frame sequences, which have great limitations for the recognition of sewing actions.

### Sewing gesture recognition based on spatial-temporal dual-stream convolutional neural network

**Model of spatial-temporal dual-stream convolutional neural network**

The time flow information of sewing gestures is a powerful feature to distinguish similar gestures. Therefore, we use the spatial stream CNN network and time stream CNN network to extract the spatial and temporal features of sewing gestures to improve the effect of gesture recognition. The spatial flow network and time flow network are used to extract the features of the spatial position and temporal motion of sewing gestures respectively. Then, the spatial location characteristics and temporal motion characteristics that reflect sewing gestures are merged. The structure diagram of the dual-stream network model is shown in Figure 10.

The backbone of the spatial flow network model adopts the VGG model. The model takes a continuous RGB image of \(224 \times 224 \times 3\) as input, and the output is the probability distribution of different sewing gesture categories. The dual-stream network is composed of 13 convolutional layers and three fully connected layers. The convolution kernel has a size of \(3 \times 3\) and a step size of \(1 \times 1\), which is stacked into five blocks. The maximum pooling layer has a kernel size of \(2 \times 2\) and a step size of \(2\), which is connected after the convolutional layer. A fully connected

---

**Table 2. Parameter values of SSD improved model.**

| Parameter name | Value |
|----------------|-------|
| Batch          | 64    |
| Batch split    | 8     |
| Iterations     | 8000  |
| Decay factor   | 0.1, 0.1 |
| Momentum       | 0.9   |
| Decay rate of weight | 0.005 |
| Threshold      | 0.5   |
layer maps the sewing gesture feature to the feature vector. Finally, the Softmax classifier is used to output the probability distribution of different sewing gesture categories through four neurons.

The parameters of the temporal flow network model are consistent with the spatial flow convolutional neural network, and the size of the input optical flow graph is $224 \times 224 \times 2L$. To make better use of the motion timing characteristics in optical flow, we superimpose multiple optical flow diagrams, and the optical flow superimposed diagrams are sent to the time flow network model. The optical flow overlay is the “instantaneous velocity” of the pixels of the moving target in the continuous video frame in space. It can clearly and effectively characterize human body motion information, which greatly improves the performance of the dual-stream model. The dense optical flow method is used for the calculation of the optical flow diagram. After inter-frame segmentation is performed on sewing video samples, a set of dense optical flow vector fields between adjacent video frames $t$ and $t+1$ is calculated. The two-bit gray-scale optical flow diagram of the previous time is generated by tracking the frame at the next time, and it is decomposed into horizontal and vertical components.

The dense optical flow vector fields are superimposed to form multiple optical flow groups which are input to the time flow convolutional neural network for training.

Equation (4) calculates the horizontal component.

Equation (5) calculates the vertical component.

$$I_{x}(u,v,2k-1) = d_{x,k-1}^t(u,v) \tag{4}$$

$$I_{y}(u,v,2k) = d_{y,k-1}^t(u,v) \tag{5}$$

$$u = [1, w], v = [1, h], k = [1, L]$$

In formulas (4) and (5): $d_{x}^t(u,v)$ represents the displacement vector of point $(u, v)$ in the $t$ frame. $d_{x}^t$ and $d_{y}^t$ represent the vector fields of horizontal and vertical components. $w$, $h$, and $L$ represent the width, height, and length of sewing gesture video frame, respectively. $\tau$ is the current video frame.

Which layer merges the features extracted by temporal flow network and space flow network is a problem that we need to consider. In the following text, the fusion layer with the highest detection accuracy is taken as the best fusion position.

**Parameter setting of spatial-temporal dual-stream network.** The dual-stream network initializes the parameters by loading the pre-training weight file, which makes the network parameters in a better position to improve training speed. Some parameters in the training process are set according to experience, and the parameter values are shown in Table 3.

During training, the number of samples is 96 as a batch of samples, and the parameters are updated once for each batch of samples. Continuous 10 frames of RGB video frames are used as the input of spatial stream to achieve a balance between the computational complexity and amount of data information. The size of the input space map is $224 \times 224 \times 3$. The input of the time flow network is the superposition of horizontal and vertical optical flow graphs, and the size of the superposed optical flow graph is $224 \times 224 \times 2L$. According to previous experience, the

| Parameter name | Time flow network | Spatial flow network |
|----------------|-------------------|---------------------|
| Input          | $224 \times 224 \times 20$ | $224 \times 224 \times 3$ |
| Batch          | 96                | 96                  |
| Segmentation   | 10                | 10                  |
| Iterations     | 8000              | 8000                |
| Initial learning rate | 0.01               | 0.01                |
| Weight         | 0.90              | 0.90                |
effect of superimposing the optical flow diagram with $L = 10$ in the time domain is the best.

The small-batch stochastic gradient descent method is used to optimize the dual-stream network model. The initial learning rate is 0.001, which is reduced according to a fixed schedule. Changing the learning rate to $10^{-3}$ after 5000 iterations. After 7000 iterations, the learning rate is changed to $10^{-4}$.

Training of spatial-temporal dual-stream network model. The training of the dual-stream network model uses 4000 images for a total of 12,000 training times. During the training process, we draw the loss value change curve and detection accuracy value change curve of the dual-stream network model to observe dynamic the training process. The loss value change curve is shown in Figure 11. The change curve of detection accuracy value is shown in Figure 12.

It can be seen from the loss value change curve that the loss value of the parallel spatiotemporal dual-stream convolutional neural network model is rapidly reduced in the previous iteration, and the model is quickly fitted. After iterative training 4000 times, the downward trend of model loss is slowing down. After about 8000 iterations, the model converged to 0.3. It can be seen from Figure 12 that the accuracy rate increases as the number of iterations increases. When iterated 3500 times, the detection accuracy rate increased rapidly. After 4000 iterations, the detection accuracy rate remains at 92.6%.

**Experiment**

**Processing of experimental data sets**

Collection of sewing gesture data set. The sewing gesture dataset samples mainly come from the sewing workshop. The data set contains more than 1000 sewing gesture motion videos, which include four sewing gestures with a high similarity: inner overlock seam, hemming seam, cladding seam, and cut fabric (as shown in Figure 13). It also includes action videos of gestures at different periods.

In the experiment, we frame 1000 self-built video samples at a frequency of $1 \times 10^7$ HZ. The total number of samples after framing is 6670, which includes four sewing gesture categories. Among the four sewing gestures, there are 1600 inner overlock sewing gesture targets, 1660 hemming sewing gesture targets, 1700 cutting fabric sewing gesture targets, and 1710 cladding sewing gesture targets. The data samples after framing are combined into data set 1 according to the sequence of framing. After that, the optical flow diagrams of all 6670 data samples in data set 1 are obtained, and the obtained optical flow diagrams are sequentially formed into data set 2 (as shown in Figure 14). Data set 1 is used as the input sample of the YOLO-V3 improved model, SSD improved model, and spatial stream CNN model. The optical flow graph in data set 2 is used as the input sample of the time flow CNN network model. The sample ratios of the training set, validation set, and test set in the two data sets are all 6:2:2.

Enhancement of data set. To improve the quality and richness of the experimental data set, we use traditional image enhancement methods to process the color, brightness, and direction of the segmented image. The 6670 video frames of sewing gestures in data set 1 are enhanced, and the obtained optical flow diagrams are sequentially formed into data set 2 (as shown in Figure 14). Data set 1 is used as the input sample of the YOLO-V3 improved model, SSD improved model, and spatial stream CNN model. The optical flow graph in data set 2 is used as the input sample of the time flow CNN network model. The sample ratios of the training set, validation set, and test set in the two data sets are all 6:2:2.

(a) Color correction: Under different lighting conditions, there is a certain deviation between the color of the collected image and true color. The gray world algorithm is used to eliminate the influence of light on the color rendering, and the processed pictures are added to the train data set.
Figure 13. Sewing gesture data set: (a) the first line of sewing gesture is for cutting fabric, (b) the second line of sewing gesture is for cladding seam, (c) the third line of sewing gesture is for hemming seam, and (d) the fourth line of sewing gesture is for inner overlock sewing.

Figure 14. Time sequence optical flow diagram: (a) inner overlock seam, (b) cladding seam, (c) cutting fabric, and (d) hemming seam.

(b) Image brightness: During 24 h production in a workshop, the impact of different light intensities on the image can be simulated through image brightness conversion. Too much change in brightness value affects the accuracy of manually labeling the bounding box, which inhibits the performance
Table 4. Number of images generated by data enhancement method.

| Sewing gesture | Serial number | Original images | Number of enhanced images |
|----------------|---------------|-----------------|--------------------------|
|                |               |                 | Color | Brightness | Spin | Blurry |
| Pique seam     | a             | 120             | 120   | 1100       | 160  | 100    |
| Hemmed seam    | b             | 120             | 120   | 1160       | 160  | 100    |
| Cut fabric     | c             | 120             | 120   | 1200       | 160  | 100    |
| Cladding seam  | d             | 120             | 120   | 1210       | 160  | 100    |

Figure 15. The sewing gesture data sample after image enhancement: (a) cutting fabric, (b) cladding seam, (c) hemming seam, and (d) inner overlock seam.

Experimental results

YOLO-V3 improved model's detection results of sewing gestures. To verify the performance of the improved YOLO-V3 model, we compare the proposed model with R-CNN, YOLO-V2, and the original YOLO-V3 on the collected sewing gesture dataset. The results are shown in Table 5. The detection result is shown in Figure 16.

Table 5. R-CNN, YOLO-V2, original YOLO-V3, and YOLO-V3 improved network for the detection results of sewing gestures.

| Detection network | AP (%) | Speed (s⁻¹) |
|-------------------|--------|-------------|
| R-CNN             | 92.56  | 0.6         |
| YOLO-V2           | 90.04  | 46.0        |
| YOLO-V3           | 92.16  | 44.0        |
| Improved YOLO-V3 network | 94.45  | 43.0        |

YOLO-V3 improved model’s detection results of sewing gestures. To verify the performance of the improved YOLO-V3 model, we compare the proposed model with R-CNN, YOLO-V2, and the original YOLO-V3 on the collected sewing gesture dataset. The results are shown in Table 5. The detection result is shown in Figure 16.

Table 5 shows that the improved YOLO-V3 model has a higher accuracy value than the other three networks. Since the improved YOLO-V3 model has a higher image feature reuse rate than the original YOLO-V3 model. The average accuracy is increased by 2.29% compared with the original YOLO-V3. In terms of detection speed, the improved model has a speed of 43.0 frames/s, which is
much higher than R-CNN, and the detection speed is the same as that of YOLO-V2 and YOLO-V3. The model checking effect is shown in Figure 16. It can be seen from the actual detection results that the improved YOLO-V3 model can accurately detect four different sewing gestures compared to the other three models. The improved YOLO-V3 network is more accurate in frame overlap.

The detection results of improved SSD model on sewing gestures. To test the recognition effect of the improved SSD model on sewing gestures, we compare it with Faster R-CNN, original SSD-VGG16, and SSD-Resnet50 networks. The detection results of the sewing gestures by Faster R-CNN, YOLO, original SSD, and SSD improved network are shown in Table 6. The detection result is shown in Figure 17.

Through the comparison of detection speed and accuracy results: In terms of detection speed, the improved SSD model and YOLO and SSD model are 46, 47, 51 frames/S, respectively. The three detection speeds are almost the same. In terms of detection accuracy, the final detection speed and accuracy of the improved SSD network are higher than the other three networks. Because the improved SSD model introduces the FPN structure, the original up-sampling structure is improved, and the high-level semantic features and low-level semantic features are better integrated, which leads to a greatly reduced probability of false detection and missed detection in the detection of small targets. Compared with Faster R-CNN network, YOLO network, and SSD network, the detection accuracy of small-area targets is increased by 4.81%, 10.59%, and 12.67%, respectively. Therefore, under the premise of ensuring a certain speed, the improved SSD model has significantly improved the detection accuracy of small-scale targets.

The detection results of improved SSD model by different fusion layers in spatial-temporal dual-stream convolutional neural network. To analyze the influence of different fusion positions in the spatial-temporal dual-stream convolutional neural network on the detection effect of sewing gestures, we combine the spatial features of sewing gestures and temporal features represented by dense optical flow graphs in convolutional layer and fully connected layer, respectively. We train on the public data sets UCF101 and HMDB51, and test on the self-built sewing gesture data set. The accuracy of the spatial-temporal dual-stream network is used to

---

**Table 6. Detection results of sewing gestures by Faster R-CNN, YOLO, original SSD, and SSD improved algorithm.**

| Detection algorithm | AP (%) | Small area target (%) | Speed (S⁻¹) |
|---------------------|--------|-----------------------|-------------|
| Faster R-CNN        | 85.91  | 63.32                 | 16          |
| YOLO                | 83.42  | 57.54                 | 46          |
| SSD                 | 85.69  | 55.46                 | 47          |
| Improved SSD algorithm | 88.79  | 68.13                 | 51          |
measure the recognition effect of complex and coherent sewing gestures. Table 7 shows the recognition results of sewing gestures in different feature layers of the spatial-temporal dual-stream network.

Table 7. Recognition accuracy values of different fusion positions (%).

| Feature fusion layer | UCF101 | HMDB51 | Sewing gesture data set |
|----------------------|--------|--------|------------------------|
| S-Conv5 T-Conv5      | 67.4   | 56.5   | 83.8                   |
| S-fc6 and T-fc6      | 73.7   | 65.4   | 87.1                   |
| S-fc7 and T-fc7      | 77.6   | 67.3   | 89.9                   |
| S-fc7 and T-fc6      | 77.1   | 68.4   | 89.1                   |
| S-fc6 and T-fc7      | 80.9   | 68.8   | 92.6                   |

It can be found from Table 7 that the recognition accuracy is closely related to the location of fusion. With the deeper the fusion location layer, the model can obtain higher semantic information, and the recognition accuracy has a certain improvement. By comparing in the fusion effect of the fully connected layer and convolutional layer, it is found that the dual-stream network can capture more efficient high-level semantic expressions when the fully connected layer is fused so that the recognition accuracy is increased by at least 3.3%. When fusing the space network layer fc7 layer and time network layer fc6 layer, the recognition accuracy rate drops by 0.8%. The reason for the decrease is that the deeper feature fusion cannot closely link the optical flow information with spatial information at the same time. The fusion model of the sixth fully connected layer of space and the seventh fully connected layer of time flow network can closely connect the spatial position information and temporal motion information of sewing gesture.

This layer fusion has very good detection results for sewing gestures and has the best performance on three data sets.

Comparison of detection accuracy of sewing gestures by YOLO-V3 improved network, SSD improved network, and dual-stream convolutional neural network

To verify the recognition effect of the three improved models proposed in the paper on sewing gestures, we use the improved YOLO-V3 detection method and the improved SSD detection network to train the model of video frame pictures with sewing gestures. On the other hand, a dual-stream network detection method is used to train the sewing gesture data set. The detection results of the three detection networks for sewing gestures are shown in Table 8.

It can be seen from Table 8 that the recognition accuracy of the dual-stream network method is better than that of the improved YOLO-V3 model and the improved SSD
Due to the preprocessing of the time series light flow chart with a large amount of data, the detection speed is slower than the YOLO-V3 improved model and the SSD improved model. The SSD improved model has the fastest detection speed among the three networks. It can be seen that the calculation of optical flow information and fusion of spatial-temporal features have a good reading of action semantics in the time dimension, which improves recognition accuracy. In practical applications, when there is no special requirement for recognition speed, the dual-stream convolutional neural network is the best choice for recognizing sewing gestures. When there is a higher requirement for speed, the improved SSD network has better performance.

**Recognition results of sewing gestures by TDD, C3D, TWO-Stream, P-CNN, and three improved models**

To further compare the advantages of the model proposed in the paper, we compare the dual-stream network model, the YOLO-V3 improved model, the SSD improved model, and five mainstream models. Five mainstream networks include IDT behavior detection methods that perform best in traditional methods, deep learning-based Time Division Duplexing (TDD), Convolutional 3D (C3D), Two-Stream, and Pose Convolutional Neural Network (P-CNN) methods. The test is carried out on the self-built data set, in which the ratio of the training set to the test set is 8:2. The final results are shown in Table 9.

From the results in Table 9, the method based on the deep neural network method can learn the semantic expression of high-level sewing gestures more easily, and the effect is better than the traditional IDT method. The dual-stream network can effectively use time dimension information, and it has better detection accuracy than other detection models. Compared with P-CNN, C3D, and TDD methods, the improved method of YOLO-V3 has improved detection accuracy by 8.2%, 4.1%, and 2.1%, respectively. Compared with IDT, Two-Stream, P-CNN, TDD, and C3D methods, the improved SSD network has improved detection accuracy by 5.9%, 5.2%, 8.5%, 2.4%, and 4.4%, respectively. Compared with the original dual-stream, P-CNN, C3D, and TDD methods, the spatiotemporal dual-stream network has improved recognition accuracy by 5.9%, 9.2%, 5.1%, and 3.1%, respectively. The improved SSD model has the fastest detection speed at 51 frames/s. The improved gesture recognition algorithm can ensure the detection speed while improving the recognition accuracy. Through the analysis of experimental results, the improved YOLO-V3 model, the improved SSD model, and the dual-stream convolutional neural network model are better than other networks in the recognition of sewing gestures. The effect of recognizing sewing gestures is better.

**Conclusion**

To recognize sewing gestures quickly and accurately, we recognize four gestures such as inner overlock seam, hemming seam, cladding seam, and cutting fabric. The improved YOLO-V3 model embeds a densely connected network in the original transmission layer of YOLO-V3 with a lower resolution to promote feature reuse and fusion. Compared with the original YOLO-V3 model, the recognition accuracy of four similar sewing gestures is increased by 2.29%. The improved SSD model uses the residual network to replace the original SSD backbone network. By combining the idea of a feature pyramid, the original up-sampling structure is improved, which better combines high-level semantic features and low-level semantic features. The overall recognition accuracy of the improved SSD model is 88.79%. The real-time detection speed reaches 51 frames/ s⁻¹. Compared with the SSD model, the YOLO model, and the Faster R-CNN model, the improved SSD model has increased the recognition accuracy of small targets by 4.81%, 10.59%, and 12.67%, respectively. The space-time dual-stream convolutional neural network can make full use of the temporal and spatial characteristics of sewing gestures, which greatly improves the utilization of feature information. Therefore, it has a higher recognition effect for sewing gestures. Experiments verify that the spatial-temporal dual-stream
network has the highest recognition accuracy rate for sewing gestures, with a recognition accuracy rate of 92.6%. But its detection speed is slow. The real-time detection speed is 0.6 frames/s. The three networks meet the accuracy and real-time requirements of sewing gesture recognition under different environments and application conditions.

The goal of future work is to recognize more and more complex sewing gestures. Besides, we will study how to further improve the recognition speed, and finally realize the real application in the field of human-machine collaborative sewing.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Natural Science Foundation of China [51905405]; Key Research and Development plan of Shaanxi province China[2019ZDLGY01-08]; Ministr of Education Engineering Science and Technology Talent Training Research Project of China [18JDG029]. Innovation Capability Support Program of Shaanxi[2021TD-29]. Key Research and Development program of Shaanxi province [2022GY-276].

ORCID iD
Sijie Yang https://orcid.org/0000-0001-7747-2063

References
1. Liu R, Fung E and Abida Y. Evaluation of perceived comfort and functional performance of activewear. In: Yip J (ed.) Latest material and technological developments for activewear. Duxford: Woodhead Publishing, 2020, pp.89–118.
2. Pyka W, Jedrzejowski M, Chudy M, et al. On the use of textile materials in robotics. J Eng Fiber Fabr 2020; 15(2020): 1–8.
3. Kim M, Sul IH and Kim S. Development of a sewing machine controller for seam pucker reduction using online measurement feedback system. J Eng Fiber Fabr 2017; 12: 67–71.
4. Yoshimi T, Takezawa K and Hirayama M. An improvement of trajectory tracking accuracy of automatic sewing robot system by variable gain learning control. IFAC-PapersOnLine 2018; 51(22): 1–6.
5. Koustoumpardis PN and Aspragathos NA. Fuzzy logic decision mechanism combined with a neuro-controller for fabric tension in robotized sewing process. J Intell Robot Syst 2003; 36(1): 65–88.
6. Zacharia PT. An adaptive neuro-fuzzy inference system for robot handling fabrics with curved edges towards sewing. J Intell Robot Syst 2010; 58(3–4): 193–209.
7. Huang B, Ye M, Hu Y, et al. A multi-robot cooperation framework for sewing personalized stent grafts. IEEE Trans Ind Inform 2018; 14: 1776–1785.
8. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE 1998; 86(11): 2278–2324.
9. Weihua He. Research on key technologies of human behavior recognition. Chongqing University, 2012, pp.1–135, https://kns.cnki.net/kcms/detail/detail.aspx?FileName=1012047169.nh&DbName=CDFD2013
10. Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 2013; 103(1): 60–79.
11. Zeng YR, Zeng Y, Choi B, et al. Multifactor-influenced energy consumption forecasting using enhanced back-propagation neural network. Energy 2017; 127: 381–396.
12. Yilmaz A and Shah M. Actions sketch: a novel action representation. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, San Diego, CA, 20–25 June 2005, pp.984–989. Washington, DC: IEEE Computer Society.
13. Garelick L, Blank M, Shechtman E, et al. Actions as spacetime shapes. IEEE Trans Pattern Anal Mach Intell 2007; 29(12): 2247–2253.
14. Jiang YG, Yang J, Ngo CW, et al. Representations of key-point-based semantic concept detection: a comprehensive study. IEEE Trans Multimedia 2010; 12(1): 42–53.
15. Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatial-temporal features. In: Proceedings of joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, Beijing, China, 15–16 October 2005, pp.65–72. Piscataway, NJ: IEEE Press.
16. Matikainen P, Hebert M and Sukthankar R. Trajectory: action recognition through the motion analysis of tracked features. In: Proceedings of the 12th IEEE international conference on computer vision, Kyoto, Japan, 27 September–4 October 2009, pp.514–521. Piscataway, NJ: IEEE Press.
17. Liu J, Luo J and Shah M. Recognizing realistic actions from videos in the wild. In: Proceedings of IEEE conference on computer vision and pattern recognition, Miami, FL, 20–25 June 2009, pp.1996–2003. Piscataway, NJ: IEEE Press.
18. Jiang Y, Dai Q, Xue X, et al. Trajectory based modeling of human actions with reference points. In: Proceedings of European conference on computer vision, Florence, 2012, pp.425–438. Berlin, Heidelberg: Springer Verlag.
19. Wang H and Schmid C. (2014). Action recognition with improved trajectories. In: Proceedings of IEEE international conference on computer vision, Sydney, Australia, 2014, pp.3551–3558. Washington, DC: IEEE Computer Society.
20. Asadi-Aghbolaghi M, Clapes A, Bellantonio M, et al. Deep learning for action and gesture recognition in image sequences: a survey. In: Escalera S, Guyon I, Athitsos V, et al. (eds) Gesture recognition. Cham: Springer Verlag, 2017, pp.539–578.
21. Meng B, Liu XJ and Wang XL. Human behavior recognition based on quaternion spatiotemporal convolutional neural network. J Instrum 2017; 12(1): 2643–2650.
22. Wu X, Zhang J and Xu X. Hand gesture recognition algorithm based on faster R-CNN. J Comput Aided Des Comput Graph 2018; 30(3): 468–476.
23. He K, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision, Venice, Italy, 2017, pp.2921–2930. Washington, DC: IEEE Computer Society.
on computer vision, Venice, Italy, 2017, pp.2980–2988. New York: IEEE.
24. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: Computer vision and pattern recognition, Las Vegas, NV, 2016, pp.779–788. New York: IEEE.
25. Redmon J and Farhadi A. YOLO9000: better, faster, stronger. In: IEEE conference on computer vision and pattern, Honolulu, HI, 21–26 July 2017.
26. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector. Cham: Springer, 2016, pp.1–17.
27. Narayana P, Beveridge J and Draper B. Gesture recognition: focus on the hands. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, 18–23 June 2018, pp.5235–5244.
28. Zhang J, Dai Y, Porikli F, et al. Multi-scale salient object detection with pyramid spatial pooling. In: 2017 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017, pp.1286–1291.
29. Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 2013; 35(1): 221–231.
30. Wang X, Gao L, Song J, et al. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. IEEE Signal Process Lett 2017; 24(4): 510–514.
31. Wu Z, Wang X, Jiang YG, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM international conference on multimedia, Brisbane, Australia, 2015, pp.461–470. New York: ACM.
32. Simonyan K and Zisserman A. Two-stream convolutional networks for action recognition in videos. Adv Neural Inf Process Syst 2014; 1(4): 568–576.