Available online at https://jcupt.bupt.edu.cn

The Journal of China Universities of Posts and Telecommunications

April 2025, 32(2): 18 – 30

Traditional Chinese painting instance segmentation algorithm based on the integrating spatial structure characteristics

Hou Xiaogang[1,2], Zhao Haiying[2](✉), Li Huabiao[1], Liang Xiaoyue[3], Yang Jiaxin[2]

1. National Museum of China, Beijing 100006, China
2. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
3. Beijing Peony Electronic Group Co., Ltd, Beijing 100089, China

Abstract

Different objects in Chinese paintings contain rich cultural connotations. Segmenting and extracting different objects in Chinese paintings through technical methods is an effective way to enhance cultural added value and activate cultural resources. Although the existing deep learning methods can extract multi-level features for instance segmentation, the location relationship features of instances are not fully utilized, resulting in poor segmentation results for the traditional Chinese painting (TCP) instance segmentation. In this paper, a novel TCP image instance segmentation algorithm based on the integration of spatial structure characteristics (SSC) was proposed, and is called SSC-Net. Firstly, considering the characteristics of TCP images, such as the gradual color blending and discontinuous contour lines, an instance information entropy composed of color entropy, formed by regional variance, and contour entropy, formed by contour point regression is proposed. Then, aiming at the problem that the existing network structure is difficult to fully consider the location relationship features of instances in TCP images, based on the residual neural network (ResNet) structure, a Chinese painting instance segmentation network framework composed of mask branch and position branch that can integrate spatial structure features is proposed. Finally, the color entropy and contour entropy are input into the mask branch and position branch of the SSC-Net structure respectively, so as to realize the instance segmentation of TCP. The quantitative and qualitative experiments on the challenging TCP database show that, compared with the state-of-the-art algorithms in the same category, the SSC-Net achieves good experimental results with average precision (AP) of 53.89% and 25.8 frame per second (FPS). The segmentation results meet the practical application requirements.

Keywords instance segmentation, information entropy characterization, mask branch, position branch, spatial structure integration

1 Introduction

TCP is one of the important carriers of Chinese excellent traditional culture, containing rich cultural elements. Therefore, how to apply advanced technologies such as deep learning to segment the objects with different cultural meanings in the TCP image has important theoretical significance and practical value, and is also an important basis for the activation and utilization of cultural resources[1]. The segmentation results will provide necessary technical

Special Issue: The 27th Annual Meeting of The China Association for Science and Technology
Corresponding author: Zhao Haiying, E-mail: zhaohaiying@bupt.edu.cn
DOI: 10.19682/j.cnki.1005-8885.2025.0015

support for the interpretation of cultural genes, the utilization of cultural content digital and the construction of Chinese cultural material databases[2].

With the development of science and technology, image processing technology is widely used in various fields, such as the street scene segmentation task[3], the integration of communications and computing[4-5]. Although the existing instance segmentation models achieve significant results in natural scenes[6], limited by the lack of Chinese painting related datasets and instance segmentation models designed for TCP data, instance segmentation algorithms that perform well in public datasets are not suitable for TCP image segmentation[7]. In general, the reasons why existing segmentation models are unable to achieve ideal results on TCP datasets are as follows.

1) Problem of insufficient training samples

The existing datasets for deep learning model training are mainly concentrated in the fields of traffic control, medical detection, and remote sensing monitoring, while the pixel-level labeled training data for some special fields (such as Chinese painting object extraction) is still very scarce.

2) Specificity of TCP

Unlike natural images, the TCP is a kind of visual art created by artificial use of line, color, and structure on the flat surface. For example, the color has the characteristics of halo and gradual change, the content is more complex, and the position relationship between the instance and the whole in the TCP is more prominent and important. So the existing models trained on natural image data are not applicable to this kind of data.

3) The reason for insufficient attention

Although cultural resource data has received a certain degree of attention in recent years, it is still in its early stages and has not received sufficient attention from researchers, so there are relatively few models for Chinese painting segmentation.

Chinese painting is a visual art that uses lines, colors, and structures to shape the plane manually. Therefore, the location positional relationship between instances and the whole is more prominent and important. Existing deep learning methods can extract multi-level features for segmentation, but the positional relationship features between the Chinese painting instance and the whole are not extracted through the convolutional network, resulting in poor performance of the final segmentation. Based on the ResNet[8] structure, and combined with the proposed instance information entropy representation algorithm, a novel TCP instance segmentation algorithm based on the integration of spatial structure characteristics is proposed, and term it as SSC-Net. The main contributions of this paper are summarized as follows.

Firstly, a novel instance information entropy characterization algorithm consisting of color entropy formed by regional variance and contour entropy formed by contour point regression is proposed to effectively characterize TCP objects to be segmented.

Secondly, based on the ResNet network structure, a novel Chinese painting instance segmentation network framework composed of mask branch and position branch is proposed. The model can fully consider the positional relationships of instances in TCP by integrated the spatial structure features.

Finally, experimental evaluations on the challenging TCP database indicate that the SSC-Net is more suitable for TCP segmenting than the state-of-the-art instance segmentation algorithms in the same category.

2 Related work

Image segmentation method based on deep learning is one of the hot and difficult problems in computer vision research. According to the different labels of segmentation objects, it is mainly divided into two categories: the semantic segmentation and the instance segmentation. Instance segmentation, which provides different labels for separate instances of the same class of objects, is a technique that solves object detection and semantic segmentation simultaneously. According to research task, the existing research progress is analyzed from the following three aspects: Chinese painting image segmentation, feature representation, and instance segmentation framework.

2.1 Chinese painting image segmentation

In the study of TCPs, Zou et al.[9] proposed a novel

computational approach to address the era in which a painting was drawn by using the appearance and shape features extracted from the paintings. Gao et al.[10] extracted features from two aspects, the first is color details and hue angle, and the other is angle features such as stroke, shape, and line. Then, scale-invariant feature transform (SIFT) feature detector and edge detector are fused to fully describe the visual differences in important areas. Sheng et al.[11] applied computer vision technology to study different applications of Chinese painting, such as the sentiment classification of Chinese paintings via feature recalibration of deep network aggregation. Fan et al.[12] discussed the visual complexity of Chinese paintings from the aspects of line, texture, light, and shade, and selected six color levels for statistical quantification, thus proposed to measure the visual complexity and richness of works. Cohen et al.[13] proposed an unsupervised semantic segmentation algorithm for paintings based on domain adaptation. In short, the current research on TCP image segmentation is mainly based on traditional machine learning algorithms, but the research on semantic segmentation methods based on deep learning is not sufficient.

2.2 Feature representation and fusion

Feature representation is the key step of the segmentation algorithm, and the quality of feature representation has a direct impact on segmentation results. Different researchers proposed different feature representation models. For example, the U-Net[14] realizes scale prediction by feature combination through skip connection. The single shot multibox detector (SSD)[15] method realizes multi-scale fusion by detecting features of different layers at the same time. A top-down architecture with lateral connections feature pyramid network (FPN)[16] is developed for building high-level semantic feature maps at all scales by leveraging the inherent multi-scale pyramid hierarchy of deep convolutional networks. Kirillov et al.[17] added boundary detection information in the process of clustering to improve the accuracy of the model. Visual geometry group network (VGGNet)[18] effectively enhances the abstract expression and refinement capture ability of the network for image features by introducing a deeper network hierarchical structure and adopting a smaller filter unit. DeepLab series[19-20], on the basis of retaining and developing the core technology of dilated convolution, optimize the ability to capture and utilize multi-scale features of images and refine the segmentation results gradually. Entropy[21], in the field of communication, is expressed in mathematical language as the expectation of the amount of information that can be obtained from all possible values of a variable, i.e., from all possible events. In recent years, many scholars have applied information entropy to the field of image processing. Such as, Celik[22] proposed a spatial entropy-based global and local image contrast enhancement algorithm by considering the distribution of spatial locations of gray levels of an image. Zhou et al.[23] proposed an image feature description method based on local entropy, which solved the problem that the traditional feature description method SIFT would lose some important information when describing image features at a single scale. Chen et al.[24] proposed an automatic salient object detection algorithm based on maximum entropy estimation. Considering that the color of Chinese painting images has the characteristics of halo and gradual change, the existing convolutional network is difficult to effectively extract the positional relationship features between Chinese painting instances and the whole, so the FPN[16] structure is used for multi-scale feature fusion, and the instance information entropy representation is formed for the feature representation of Chinese painting instances segmentation.

2.3 Instance segmentation algorithm

Instance segmentation, which assigns different labels to separate instances of the same class of objects, is the most challenging problem in the field of image segmentation. Semantic contour detecton (SCD)[25] can be regarded as the earliest instance segmentation algorithm, which combines the segmentation task and the detection task for the first time. Instance segmentation based on object detection first generates the detection region of each instance object, and then predicts the mask of each region by using the detector,

such as faster region-based convolutional neural network (Faster R-CNN)[26], etc. ResNet[8] is a method that regards instance segmentation as a task of first detection and then segmentation. The main steps of this method are to first extract features from images by ResNet + FPN, then scan these feature maps to generate proposals, and finally generate feature maps on regions of interest and output semantic segmentation masks. This approach can be traced back to deep mask[27], which uses a sliding window to generate instance masks. Liu et al.[28] proposed an adaptive pooling method based on the instance segmentation framework to improve information retention. Mask region-based convolutional neural network (Mask R-CNN)[29] is a simple and effective instance segmentation method with good segmentation performance on a single object. Segmenting objects by locations (SOLO)[30] views the task of instance segmentation from a completely new perspective by introducing the notion of instance categories, and achieves better accuracy on the single-shot instance segmentation. PolarMask[31] formulates the instance segmentation problem as predicting the contour of an instance through instance center classification and dense distance regression in a polar coordinate. BlendMask[32] can effectively predict dense per-pixel position-sensitive instance features with very few channels, and learn attention maps for each instance with merely one convolution layer, thus being fast in inference. In recent years, there have been a series of studies on the improvement of Mask R-CNN. For example, Cai et al.[33] proposed Cascade R-CNN, which improves the performance of the model by gradually improving the prediction and adaptive training. In the CenterMask[34] model, an anchor-free instance segmentation method is proposed to alleviate the saturation problem in the VoVNet model. Object scale inconsistency is one of the difficulties in instance segmentation, and multi-scale feature fusion is a good solution[35]. Although the above methods have achieved good results in their respective fields, these methods are difficult to take into account the large and small targets at the same time, and they all rely on the mask candidate regions generated by the original image, and do not make full use of the features extracted by the input network of a large number of images, so it is difficult to meet the accuracy requirements of TCP image segmentation.

3 Model principle and realization

SSC-Net architecture is mainly composed of the information entropy characterization module based on the FPN structure and the spatial structure integration module based on the ResNet structure. The overall architecture is shown in Fig. 1.

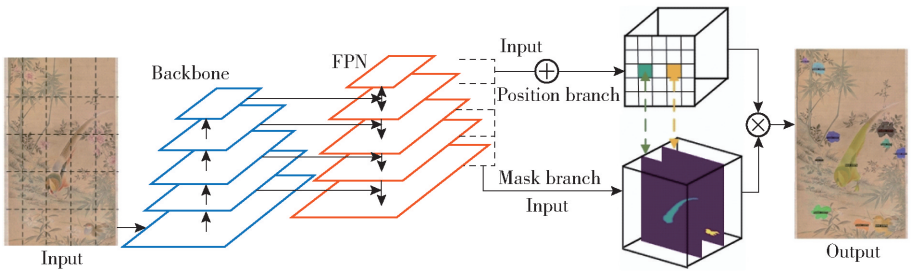

Fig. 1 Architecture of SSC-Net

3.1 Feature representation

TCP belongs to the visual art of artificial modeling on a plane by using lines, colors, and structures. Therefore, color and lines are two core elements for TCP creation. Considering the characteristics of intermittent discontinuous contours and haloed boundary colors in the TCP image, the regression direction and distance of each sampling point are calculated to get the contour entropy reflecting the boundary information of the instance. The three-dimensional (3D) color information and gradient of

each sampling point are calculated to obtain the color entropy reflecting the pixel color distribution of the instance. The color entropy and the contour entropy constitute the instance information entropy of the instance representation.

1) Contoure entropy

Considering the characteristics of intermittent discontinuity of contours and haloing of boundary colors in TCP images, a mask generation algorithm using contour entropy based on contour regression in the subsequent mask branch network is proposed.

The process starts with the input of the feature maps of each size obtained in the feature extraction phase to the subsequent mask branch to generate the heat map. Then, the center coordinates of the instance object are located by the target detector center-based object detection network (CenterNet)[36]. Finally, the initial contour is obtained by regressing the centroid bias, i.e., calculating the offset of the contour points with respect to the centroid coordinates, based on the feature maps and centroid features.

The adjustment of the initial contour requires an iterative process of random sampling, relative position judgment, and calculation of moving distance. First, a number of points $e_1, e_2, \ldots, e_n, \ldots, e_N$ are randomly selected from the initial contour, and for a particular point e_n, the classifier judges the relationship between this point and the instance position. That is, the classifier predicts whether the points are inside or outside the instance by randomly selecting pixels around the boundary of the real instance and extracting fine-grained features with respect to the boundary relative position input. Then, the high-resolution feature maps obtained from the feature extraction stage are further encoded through a convolutional layer to form the position vector $\boldsymbol{p}$ of the point e_n.

Finally, the vector $\boldsymbol{p}$ is input into the multi-layer classifier to iteratively adjust the initial contour, so as to finally obtain the contour entropy vector $\boldsymbol{p}_f$, which represents the position of the instance boundary points.

$$\boldsymbol{p}_f = \boldsymbol{p} + w\boldsymbol{d} \tag{1}$$

where $\boldsymbol{d}$ represents the moving direction and w represents the point moving distance. The move distance w is related to two factors, one is the instance size and the other is the confidence of the relative position of the point e_n. The larger the area of the instance object, the larger the moving distance and the faster the convergence rate. The confidence degree of relative position of the point e_n is obtained by the classifier. When the point e_n move once, the classifier will update the corresponding position of the point e_n accordingly, thus obtaining the next regression moving distance w is

$$w = \mathscr{k}\sqrt{A}\,|M(\boldsymbol{p}) - 0.5| \tag{2}$$

where $\mathscr{k}$ is the given coefficient, A is the shift amplitude, and $M(\cdot)$ is the multilayer classifier. The schematic illustration of the contour point regression is given in Fig. 2.

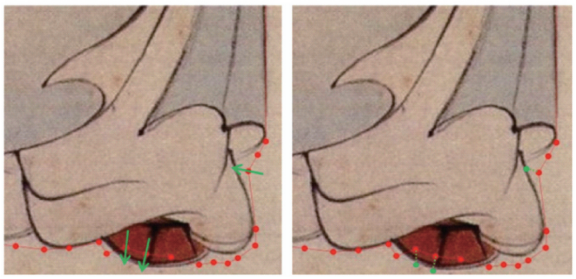

Fig. 2 Contour points regression process

2) Color entropy

The one-dimensional (1D) gray-scale entropy reflects the aggregation characteristics of the distribution of image pixel values, but the color space and contour information of an image contain more image information. Instances in different kinds of TCP have different tones and structures, such as flowers and birds in flower and bird paintings both have rich and vivid colors, but flowers for instance have similar internal colors, and bird instances have more varied internal colors, so the color features play an important role in identifying different instances.

In the Lab color space of digital image I, since two color channels a and b do not interfere with each other, the color entropy in color space information of image channels a and b is calcutated as

$$H(I) = -\sum_{-127}^{128} c(i)\lg c(i) \tag{3}$$

where the i represents the gray value of the color

channel [-127, 128], and $c(i)$ represents the proportion of pixels with gray value i in the a channel and b channel. For a picture, in addition to color richness, primary color is an important color feature. The dominant color ratio of the image and the information entropy of the image affect each other, the higher the dominant color ratio, the lower the color information entropy, and vice versa. In the Lab color space, the dominant color ratio can be computed as

$$R_{\mathrm{L,a,b}} = 2\frac{M_{\mathrm{L,a,b}}H_{\mathrm{L,a,b}}}{M_{\mathrm{L,a,b}} + H_{\mathrm{L,a,b}}} \tag{4}$$

where $R_{\mathrm{L,a,b}}$ denotes the color characterization of the image, $H_{\mathrm{L,a,b}}$ and $M_{\mathrm{L,a,b}}$ denote the information entropy and dominant color percentage of the image on the L, a, and b channels respectively.

Since TCP images usually have the characteristics of gradual halo in color, the extracted image features are fused to 1/4 size and then input them into the position branch. The specific process is to divide the image into $n \times n$ regions first, and then calculate the quantization of color features belonging to that location for each subregion, so as to fuse the color space features with the image spatial structure to obtain the feature vector of $n \times n \times [R_{\mathrm{L}}, R_{\mathrm{a}}, R_{\mathrm{b}}]$. The variance of the four neighboring regions above, below, left, and right of each region block is calculated for R_{L}, R_{a}, R_{b} of different regions respectively, to obtain the color entropy E, which reflects the correlation degree of the color space.

$$E = \alpha(R_{\mathrm{L}} - \overline{R}_{\mathrm{L}}) + \beta(R_{\mathrm{a}} - \overline{R}_{\mathrm{a}}) + \gamma(R_{\mathrm{b}} - \overline{R}_{\mathrm{b}}) \tag{5}$$

where R_{L}, R_{a} and R_{b} denote the color radius within different color channel regions, respectively. $\overline{R}_{\mathrm{L}}$, $\overline{R}_{\mathrm{a}}$, and $\overline{R}_{\mathrm{b}}$ denote the average values of the color radius in the corresponding regions, respectively. α, β and γ denote the corresponding weight coefficients.

3) Information entropy

In image processing, entropy can estimate the average information content of an image[37]. The information entropy formula is expressed as

$$H(x) = -\sum_{i'=1}^{I'} p(x_{i'}) \lg p(x_{i'}) \tag{6}$$

where $p(x_{i'})$ shows the probability of occurrence of event $x_{i'}$. From Eq. (6), it can be seen that information entropy has non-negativity, cumulativity, and monotonicity, which makes it widely used in probability theory, astrophysics, images, and other fields as a measure of effective information.

In this paper, inspired by the concept of information entropy[38], the information entropy theory is introduced to calculate the information entropy for each spatial element, which generated by the base segmentation network according to the contour entropy $\boldsymbol{p}_{\mathrm{f}}$ and the color entropy E. By calculating the information entropy for each spatial element of the feature map generated by the base segmentation network, we can determine regions with high certainty and regions with low certainty in an image, so as to establish connections between them and to enable information to flow from elements with low entropy to elements with high entropy to reduce the uncertainty of the elements to which information flows.

For an input TCP image I, let define $\{I(x,y) \mid 0 \leqslant x \leqslant H-1, 0 \leqslant y \leqslant W-1\}$, where H and W denote the length and width of image I respectively, and assume that I has a dynamic range of $[I_{\mathrm{d}}, I_{\mathrm{u}}]$, where $I(x,y) \in [I_{\mathrm{d}}, I_{\mathrm{u}}]$. The feature map $h_{\mathrm{c}} = \mathrm{softmax}(\boldsymbol{p}_{\mathrm{f}})$ and $h_{\mathrm{E}} = \mathrm{softmax}(E)$ denote the feature maps generated by the base segmentation network. Using the feature map $h_k(x,y)$, which generated by the base segmentation network according to the contour entropy vector $\boldsymbol{p}_{\mathrm{f}}$ and the color entropy E, the information entropy measure $S(h_k)$ is computed as

$$S(h_k) = -\sum_{x=1}^{X}\sum_{y=1}^{Y} h_k(x,y) \mathrm{lb}\ h_k(x,y) \tag{7}$$

The information entropy measure $S(h_k)$ is used to compute a discrete function f_k according to $f_k = S_k \Big/ \sum_{l=1,l\neq k}^{K} S_l$. The discrete function f_k is further normalized by $f_k \leftarrow f_k \Big/ \sum_{l=1}^{K} S_l$, and $F_k = \sum_{l=1}^{K} f_l$ denotes cumulative distribution function.

3.2 Network architecture integrating spatial structure characteristics

The mask branch and position branch constitute the core components of the SSC-Net architecture. The

color entropy and contour entropy are input into the mask branch and position branch of the network structure respectively, so as to realize TCP instance segmentation integrating spatial structure.

1) Mask branch

The mask branch uses contour regression to predict the mask for each instance. Sampling points are randomly selected on the edge during each iteration, and the regression direction and distance of each sampling point are calculated to achieve boundary prediction. It is responsible for predicting the mask containing instances at each position. The structure of the mask branch is shown in Fig. 3.

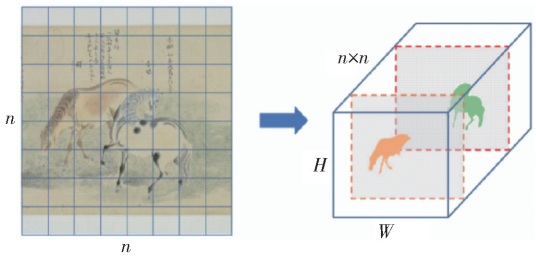

Fig. 3 Schematic illustration of mask branch

The mask branch divides the image into $n \times n = M$ grid regions, each of which can be understood as an instance position class. Starting from the top left, mark from left to right and from top to bottom. If an instance is in the (x, y) position class, the $nx + y$ channel will predict the mask belonging to that instance. If n is set too large, the predicted mask for an instance will increase, resulting in redundant computation. If n is set too small, a grid area may contain multiple objects, which means that one channel predicts multiple instance masks, losing the meaning of instance position mapping and affecting the utilization of position features. We set $n = 8$ according to the experimental comparison.

The 5-layer feature map is input into the mask branch, and the position coordinates of each instance (x, y) are input into the last two channels of the layer. The coordinates need to be normalized to between $[-1, 1]$, the number of channels becomes $256 + 2 = 258$ channels. Take the number of grids set at the current layer, and the output feature size is $W \times H \times n \times n$. Where H is the image height and W is the image width. By combining spatial structure features with multi-scale features, the problem of instance overlap is solved to a certain extent and the accuracy of the model is improved.

The loss function of the mask branch is

$$L_{\mathrm{m}} = \frac{1}{N_{\mathrm{Pi}}} \sum_{n'=1}^{128} \| \boldsymbol{p}_{n'} - \boldsymbol{g}_{n'} \| \tag{8}$$

where N_{Pi} is the number of predicted instances, and $\boldsymbol{p}_{n'}$ and $\boldsymbol{g}_{n'}$ denote the contour point locations of predicted instances and ground truth, respectively. Considering the efficiency of the model, 128 sets of sampling points on each instance boundary are selected for regression.

2) Position branch

The position branch obtains the composition of example semantic categories and instance positions in the image by feeding the color entropy of the current region to the feature map channel and the classifier. Unlike prediction mask, semantic information and instance location information are less affected by feature map scale. Therefore, the position branch does not need to input all the 5-layer features of the backbone network, but through up-sampling and down-sampling, the features are unified to 1/4 size, and divided into $n \times n = M$ grid areas.

In position branch, there are multiple cases of different semantic categories in a position block, in order to reduce the computation, the instance is only kept with the highest confidence, i. e., a position region only belongs to one instance. When there are multiple instances in a region, the instance with the largest percentage of masks in the region will determine the semantic category and region composition determination of the location region. First, the color entropy of each neighboring region is calculated by Eq. (7), and the image feature maps are encoded using the same fully-connected layer and then jointly input to the classifier to construct the difference layer. Then, the Euclidean distance of the two regions is calculated using the feature encoding, color entropy, and semantic category information, and the confidence that the two regions belong to the same instance is output using the sigmoid activation function after the single-node fully connected layer. The instance

location prediction process is shown in Fig. 4.

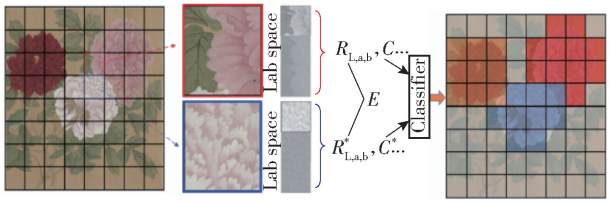

Fig. 4 Instance position prediction illustration of position branch

After completing the prediction of all positional semantic categories and instance compositions in the image, the loss function is calculated by comparing with the ground truth to determine the proportion of instances in the ground truth, so that a positional region only belongs to one instance. The position branch loss function L_p consists of two parts, the classification prediction and the position prediction, is shown as

$$L_p = \sum_{m=1}^{M} \lambda L_c + L_l;\ m = 1, 2, \ldots, M \tag{9}$$

where M is the number of location regions, L_c is the semantic category prediction loss, L_l is the location prediction loss, and λ is a given coefficient.

3) Combinatorial output

The ResNet + FPN network extracts 5-layer of features such as 1/4, 1/8, 1/16, 1/32, 1/64 of the image, and inputs them into the position branch and mask branch for prediction. The prediction marks the channel mask belonging to a certain instance as positive samples, and the others as negative samples. Then, the final instance mask prediction result is obtained by adding the weights. Finally, the prediction results are upsampled to the original image size to complete instance segmentation.

In the position branch, each spatial position class represents an area, and this branch predicts the composition of the area blocks for each instance. In the mask branch, each spatial position class maps to an instance mask prediction. By combining the outputs of the two branches through spatial position class correspondence channels that do not belong to instance predictions are filtered out, and masks belonging to the same instance are linearly combined by weighting and selected by threshold. The final segmentation prediction is obtained. Unlike the two-stage model that first performs object detection, the position branch is composed by predicting the instance location, improving the model's perception of instance shape and overall spatial structure of the image. It is also associated with the mask branch through spatial position classifications, and the linear addition of multiple channels of masks further improves segmentation accuracy. The final model is represented as

$$M_c = \sum_{n_r=1}^{N_r} \sum_{n_i=1}^{N_i} k_{n_r} m_{n_i} \tag{10}$$

where N_r is the number of location regions for an instance, N_i is the number of instances of the image, k_{n_r} is the weight obtained by quantizing the prediction confidence in the position branch with a specific step size, and m_{n_i} is the mask prediction belonging to the instance channel.

The loss function L_s of the model consists of two-branch loss, L_s is shown as

$$L_s = L_m + L_p \tag{11}$$

where L_m is the loss of the mask branch and L_p is the loss of the position branch.

4 Experiment and analysis

4.1 Experimental environment and dataset

The software environment for SSC-Net training is Ubuntu 18.04 operating system, Python 3.8 language environment, PyTorch 1.9.1 deep learning framework, OpenCV image processing library, CUDA 11.1 platform. The hardware environment is 12-core CPU 2.5 GHz, single NVIDIA RTX 2080Ti 11 GB card. Different models are trained for different iterations (18 000 or 36 000 or 90 000) with an initial learning rate of 0.01.

The experimental dataset mainly comes from various book scans and digital collections of different museums. The dataset consists of 1 000 images from three categories of traditional Chinese realistic painting: figure painting, flower and bird painting, and animal painting. Considering factors such as the color and layout integrity of TCP, while maintaining image semantics, data augmentation is carried out

using methods such as flipping, scaling, and cropping. After data augmentation, the dataset was increased to 3 000 images. All images in the dataset are more than 1 000 pixels of width and height, and an aspect ratio of 1.03. The images were divided into training and testing sets in an 8∶2 ratio. Use the EISeg annotation tool to annotate the dataset into 5 categories: person, horse, bird, flower, and dog. The annotated data is saved in JavaScript object notation (JSON) file format.

The AP of a category and the mean AP (MAP) of all categories indicators are used to qualitatively evaluate the experimental results[39]. For the convenience of experimental explanation, different categories of APs are calculated for algorithm performance evaluation. In this paper, the AP_{50} and AP_{75} represent APs when intersection over union (IoU) is equal to 50% and 75%, respectively, AP_S represent targets with pixels not greater than 32×32 pixels in the segmentation area, AP_M represents the objects with pixels greater than 32×32 pixels but less than 96×96 pixels in the segmentation area, AP_L represents objects with pixels not less than 96×96 pixels in the segmentation area.

4.2 Parameter setting and ablation experiment

The procedure for checking the validity of the contour regression method of the mask branch is as follows. First, the feature map extracted by backbone is input into the mask generation module of Mask R-CNN to obtain $n \times n$ instance mask predictions based on pixel regression, and then multiply and add them with the position branch to obtain the output result. Finally, the contour regression method validity test of the mask branch is obtained. The verification results are shown in Table 1. It can be seen from the experimental that the contour-based regression has a significant improvement effect on the objects with large area, and the overall segmentation effect is improved by 2%.

Table 1 Effectiveness of differet regression methods %

Regression method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Contour	55.968	78.312	59.895	22.640	44.878	61.712
Pixel	53.134	76.019	58.195	20.478	42.385	60.256

Table 2 shows the experimental results of the effectiveness of the position branch. Instead of assisting the segmentation results through the instance position composition prediction, the final result is composed of the mask filter output of each channel directly in the mask branch. It can be seen from the experimental results that the position branch has a high improvement in accuracy, especially for the instance object with large area.

Table 2 Results of effectiveness of position branch %

Location prediction	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Yes	55.682	79.449	60.765	23.649	45.718	62.017
No	42.928	70.241	43.725	18.943	36.459	50.648

The number of contour regression points is a factor of the image experimental results. Too many sampling points will cause the model bloat and reduce the efficiency, while too few sampling points will not cover the image fine. Table 3 shows the experimental results of the number of contour regression points. According to the experimental results, the SSC-Net model has obvious improvement effect when the initial sampling points are increased, and then the growth is slow and approaches saturation. Therefore, 128 points are finally selected for contour regression in this paper.

Table 3 Effectiveness of different number points %

Number of points	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
64	46.662	71.216	53.282	17.143	39.721	52.945
96	51.532	75.752	57.318	18.349	43.164	58.425
128	55.682	79.449	60.765	19.649	45.718	62.017
160	55.026	78.733	59.961	19.624	44.885	61.556
192	54.103	77.624	59.034	19.584	45.631	62.013

ResNet[6] is used as the backbone. While keeping other conditions constant, experiments are conducted using the two most common network structures of ResNet, ResNet-50 and ResNet-101. Table 4 shows the experimental results of different backbone on different target categories. It can be seen that when choosing ResNet-101 as the backbone network, the AP reaches 55.682%, which is 1.786% higher than

ResNet-50 is used us the backbone network. Therefore, ResNet-101 is chosen as the backbone of the feature extraction network.

Table 4 Results of each backbone on different object categories

Category	AP/%	
	ResNet-50	ResNet-101
Person	49.942	52.648
Horse	38.412	40.612
Bird	53.691	54.617
Flower	37.982	40.165
Dog	89.453	90.368

In order to verify the influence of the introduction of FPN on the detection accuracy of different size targets under different scale information, Table 5 shows the influence of whether FPN is connected in the model on the output effect under the condition of other network structures unchanged. Among them, when removing the network with FPN structure, the feature map extracted by ResNet-101 is sampled to the same size as before and input to the position branch and mask branch. It can be seen from Table 5 that FPN structure fusion of multi-scale features has significantly improved the segmentation results, and AP has increased by 8.431% on average. From the APs and AP_L results, it can be seen that the access of FPN structure improves the detection accuracy of small and large objects at the same time, because the FPN structure realizes the prediction and fusion of multi-scale features. Therefore, the ResNet-101 + FPN combined structure is used as the backbone network of the model in this paper.

Table 5 Results with or without FPN access %

FPN	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Without	47.25	71.14	52.67	10.43	37.61	54.86
With	55.68	79.45	60.77	19.65	45.72	62.02

The number of instance location classes is an important parameter in the ResNet + FPN model. If n is set too large, it will not only cause an increase in calculation and training difficulty, but also make the final prediction effect worse because of the excessive number of channels. However, setting n too small will make insufficient utilization of spatial features and decrease segmentation accuracy. The experimental results are shown in Table 6. According to the experimental results, when n is less than 8, the accuracy of the model changes significantly with the increase of n. When n is greater than 8, the accuracy of the model grows slowly and there is a loss to fall back. Therefore, considering the accuracy and parameter quantity, $n = 8$ is selected in this paper, that is, the image is divided into 64 position regions for the experiment.

Table 6 Effect of n on AP %

n	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
4	50.88	74.27	56.68	13.49	42.36	59.69
6	53.57	77.33	59.32	17.06	44.31	60.93
8	55.68	79.45	60.77	19.65	45.72	62.02
12	56.46	79.12	62.46	21.25	46.67	62.65
16	55.4	79.64	59.93	20.74	45.82	60.87

Fig. 5 shows the location channel feature maps output by the SSC-Net model on each location channel, and it can be seen that the location mask of SSC-Net model performs well in capturing the spatial location of the instances. It shows that the location branch of SSC-Net achieves the acquisition of the instance location in the image by feeding the color entropy of the current region into the feature map channel and classifier.

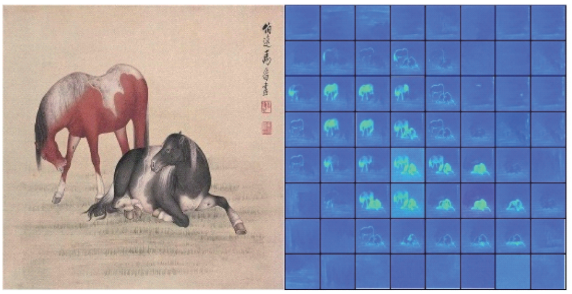

Fig. 5 Location channel feature map

4.3 Segmentation results and performance analysis

In order to verify the superiority of SSC-Net model on TCP dataset, ResNet-50 and ResNet-101 are used as backbone networks, respectively, SSC-Net with the state-of-the-art algorithms that perform instance image segmentation, including Mask R-CNN[29], SOLO[30], BlendMask[32], PolarMask[31]. The experimental result

is shown in Table 7. It can be seen that SSC-Net is 1.414% higher than the Mask R-CNN model with the best performance among the comparative algorithms, which illustrates the effectiveness of SSC-Net. Compared with the same algorithm, the performance of ResNet-101 as the backbone network is better than ResNet-50. In terms of algorithm efficiency, a light-weight version of SSC-Net outperforms the other comparison algorithms with of MAP 53.89% at 25.8 FPS evaluated on a single NVIDIA RTX 2080Ti card.

Table 7 Quantitative experimental results of different algorithms

Model	Backbone	MAP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%	Iteration/$\times 10^3$	Efficiency FPS
Mask R-CNN[29]	ResNet-50	52.937	73.584	58.374	20.936	42.224	59.219	18	15.8
	ResNet-101	54.268	78.386	59.456	21.243	44.149	61.394	18	13.6
SOLO[30]	ResNet-50	43.807	65.734	44.182	15.408	32.673	48.804	36	22.5
	ResNet-101	46.388	67.258	47.022	16.751	36.949	52.718	36	19.2
PolarMask[31]	ResNet-50	43.329	69.506	42.471	15.114	33.278	50.658	90	26.3
	ResNet-101	44.934	71.544	43.813	17.804	36.745	51.641	90	21.7
BlendMask[32]	ResNet-50	53.409	74.065	57.626	19.673	39.165	59.422	90	25.0
	ResNet-101	53.502	75.336	58.084	19.708	39.583	60.026	90	22.4
SSC-Net	ResNet-50	53.896	76.402	58.915	18.467	43.412	60.486	90	25.8
	ResNet-101	55.682	79.449	60.765	19.649	45.718	62.017	90	23.7

Fig. 6 shows the qualitative experimental results of SSC-Net model, which shows that the SSC-Net model can achieve effective instance segmentation of the specific objects of Chinese painting, especially it also has good robustness on Chinese paintings with large object scale changes.

(a) Original image (b) Result image

Fig. 6 Visual results of the SSC-Net

For example, the SSC-Net model can effectively segment the objects information such as zhinü, children, horses, and dogs in different original Chinese paintings. The qualitative segmentation results further verify the effectiveness of SSC-Net model for TCP data segmentation.

5 Conclusions

In this paper, SSC-Net instance segmentation algorithm was proposed for efficiently extracting cultural objects from TCP. By analyzing the spatial structure, color, and line features of the TCP data, an information entropy feature representation function composed of color entropy formed by regional variance and contour entropy formed by contour point regression is proposed, which is used to effectively represent cultural elements in TCP image. Based on the color entropy and contour entropy, a new network architecture composed of mask branch and position branch is designed, where the former completes the instance mask prediction by mapping the instance location class to the mask channel, and the latter realizes the instance semantic category prediction by adding the information entropy feature to the feature channel. Compared with the state-of-the-art instance

segmentation algorithms, SSC-Net achieves the best segmentation performance about the TCP image with AP of 53.89% and 25.8 FPS, and the segmentation results could meet the practical needs. In the future, we will consider introducing few-shot learning and incremental learning model, so as to alleviate the low segmentation accuracy problem of SSC-Net for TCP instance segmentation due to the lack of sufficient training samples.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFF0901700).

References

[1] ZHAO H Y, ZHU H, HOU X G. Traditional custume image semantic segmentation based on improved EMA unit. Journal of Beijing University of Posts and Telecommunications, 2022, 45(1): 69 – 74 (in Chinese).

[2] HOU X K, ZHAO H Y, MA Y, et al. Adaptive segmentation of traditional cultural pattern based on superpixel log-euclidean Gaussian metric. Applied Soft Computing Journal, 2020, 97: Article 106828.

[3] TIAN Z, ZHANG B W, CHEN H, et al. Instance and panoptic segmentation using conditional convolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 669 – 680.

[4] QI Y L, ZHOU Y Q, LIU Y F, et al. Traffic-aware task offloading based on convergence of communication and sensing in vehicular edge computing. IEEE Internet of Things Journal, 2021, 24(18): 17762 – 17777.

[5] ZHOU Y Q, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: a convergence of communication and computing. IEEE Communication Magazine, 2019, 57(5): 20 – 27.

[6] CAO J L, PANG Y M, ANWER R M, et al. SipMaskv2: enhanced fast image and video instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3798 – 3812.

[7] HOU X G, ZHAO H Y, MA Y. Fast image segmentation algorithm based on superpixel multi-feature fusion. Acta Electronica Sinica, 2019, 47(10): 2126 – 2133 (in Chinese).

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), 2016, Jun 27 – 30, Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE, 2016: 770 – 778.

[9] ZOU Q, CAO Y, LI Q Q, et al. Chronological classification of ancient paintings using appearance and shape features. Pattern Recognition Letters, 2014, 49: 146 – 154.

[10] GAO F, NIE J, HUANG L, et al. Traditional Chinese painting classification based on painting techniques. Chinese Journal of Computers, 2017, 40(12): 2871 – 2882 (in Chinese).

[11] SHENG J C, CHEN Y Q, HAN Y H. Sentiment classification of Chinese paintings via feature recalibration of deepnetwork aggregation. Journal of Computer-Aided Design and Computer Graphics, 2020, 32 (9): 1420 – 1429 (in Chinese).

[12] FAN Z B, LI Y N, ZHANG K, et al. Measuring and evaluating the visual complexity of Chinese ink paintings. The Computer Journal, 2022, 65(8): 1964 – 1976.

[13] COHEN N, NEWMAN Y, SHAMIR A. Semantic segmentation in art paintings. Computer Graphics Forum, 2022, 41 (2): 261 – 275.

[14] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'15): Part III, 2015, Oct 5 – 9, Munich, Germany. LNCS 9351. Berlin, Germany: Springer, 2015: 234 – 241.

[15] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision (ECCV'16): Part I, 2016, Oct 11 – 14, Amsterdam, Netherlands. LNCS 9905. Berlin, Germany: Springer, 2016: 21 – 37.

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, Jul 21 – 26, Honolulu, HI, USA. Piscataway, NJ, USA: IEEE, 2017: 936 – 944.

[17] KIRILLOV A, LEVINKOV E, ANDRES B, et al. InstanceCut: from edges to instances with MultiCut. Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, Jul 21 – 26, Honolulu, HI, USA. Piscataway, NJ, USA: IEEE, 2017: 7322 – 7331.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations (ICLR'15), 2015, May 7 – 9, San Diego, CA, USA. 2015.

[19] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision (ECCV'18): Part VII, 2018, Sep 8 – 14, Munich, Germany. LNCS 11211. Cham, Switzerland: Springer Nature Switzerland AG, 2018: 833 – 851.

[20] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834 – 848.

[21] SHANNON C E. A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(3): 379 – 423.

[22] CELIK T. Spatial entropy-based global and local image contrast enhancement. IEEE Transactions on Image Processing, 2014, 23(12): 5298 – 5308.

[23] ZHOU B Y, ZHAO H W, XIAO Y, et al. Image feature description method based on local entropy, Journal of Jilin University (Engineering and Technology Edition), 2017, 47(2): 601 – 608 (in Chinese).

[24] CHEN X, ZHAO H W, LIU P P, et al. Automatic salient object detection via maximum entropy estimation. Optics Letters, 2013,

38(10): 1727 – 1729.

[25] HARIHARAN B, ARBELAEZ P, BOURDEV L, et al. Semantic contours from inverse detectors. Proceedings of the 2011 International Conference on Computer Vision, 2011, Nov 6 – 13, Barcelona, Spain. Piscataway, NJ, USA: IEEE, 2011: 991 – 998.

[26] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137 – 1149.

[27] CHEN T S, LIN L, WU X, et al. Learning to segment object candidates via recursive neural networks. IEEE Transactions on Image Processing, 2018, 27(12): 5827 – 5839.

[28] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, Jun 18 – 23, Salt Lake, UT, USA. Piscataway, NJ, USA: IEEE, 2018: 8759 – 8768.

[29] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), 2017, Oct 22 – 29, Venice, Italy. Piscataway, NJ, USA: IEEE, 2017: 2980 – 2988.

[30] WANG X L, KONG T, SHEN C H, et al. SOLO: segmenting objects by locations. Proceeding of the 16th European Conference on Computer Vision (ECCV'20): Part XVIII, 2020, Aug 23 – 28, Glasgow, UK. LNCS 12370. Berlin, Germany: Springer, 2020: 649 – 665.

[31] XIE E Z, SUN P Z, SONG X G, et al. PolarMask: single shot instance segmentation with polar representation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), 2020, Jun 13 – 19, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 2020: 12190 – 12199.

[32] CHEN H, SUN K Y, TIAN Z, et al. BlendMask: top-down meets bottom-up for instance segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), 2020, Jun 13 – 19, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 2020: 8570 – 8578.

[33] CAI Z W, VASCONCELOS N. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483 – 1498.

[34] LEE Y, PARK J. Centermask: real-time anchor-free instance segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), 2020, Jun 13 – 19, Seattle, WA, USA. Piscataway, NJ, USA: IEEE, 2020: 13903 – 13912.

[35] LIN G C, LI S Y, CHEN Y F, et al. IDNet: information decomposition network for fast panoptic segmentation. IEEE Transactions on Image Processing, 2024, 33: 1487 – 1496.

[36] ZHOU B C, GILLES M, MENG Y Q. Structure SLAM with points, planes and objects. Advanced Robotics, 2022, 36(20): 1060 – 1075.

[37] OLIVA D, HINOJOSA S, OSUNA-ENCISO V, et al. Image segmentation by minimum cross entropy using evolutionary methods. Soft Computing: A Fusion of Foundations, Methodologies and Applications, 2019, 23(2): 431 – 450.

[38] BAI S, LIANG C, WANG Z, et al. Information entropy induced graph convolutional network for semantic segmentation. Journal of Visual Communication and Image Representation, 2024, 103: Article 104217.

[39] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context. Proceedings of the 13th European Conference on Computer Vision (ECCV'14): Part V, 2014, Sep 6 – 12, Zurich, Switzerland. LNCS 8693. Berlin, Germany: Springer, 2014: 740 – 755.

(Editor: Wang Xuying)