JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 1-18 (2023)
DOI: 10.6688/JISE.202301_39(1).0001

Weakly Supervised Semantic Segmentation for Headdress of Thangka Images

WENJIN HU[1,2,+], JIAHAO MENG[1,3], LI JIA[1,3], FULIANG ZENG[1,2] AND PANPAN XUE[1,2]
[1]*Key Laboratory of China's Ethnic Languages and Information Technology*
Ministry of Education
Lanzhou, 730000 P.R. China
[2]*School of Mathematics and Computer Science*
[3]*National Languages Information Technology, Boulder*
Northwest Minzu University
Lanzhou, 730000 P.R. China
E-mail: hwjforwork@126.com[+]; mjhforwork@163.com; 743507179@qq.com;
y181730416@stu.xbmu.edu.cn; y191730464@stu.xbmu.edu.cn

In order to overcome the limitations of the existing headdress segmentation methods for portraits Thangka images and the high cost of the pixel-level annotation is fully supervised semantic segmentation, we propose a weakly supervised semantic segmentation method with box-level annotations. Firstly, the proposed method uses the Canny algorithm to obtain the rough edge of the headdress. Then, the new method improves the EDLines algorithm to extract the key points of the headdress. Finally, we use Polygon's processing to generate feature masks according to the characteristics of the headdress. Experiments show that in the segmentation of the headdress of Buddha in portraits Thangka pictures, the index mIoU of the proposed method has 7.56% higher than SDI and 6.11% higher than WSIS_BBTP of the segmentation result, which are two state-of-the-art methods.

Keywords: Thangka image, semantic segmentation, weakly supervised, the headdress of Buddha, the CEDLines-polygons

1. INTRODUCTION

Thangkas are traditional Tibetan works of art that generally depict a Buddhist deity, scenario, or mandala. It is aimed to act as a reference for the introspective or saying a prayer experience. Thangka is regarded as the encyclopedia of Tibetan culture, with a history spanning over 1,300 years. Nowadays, new Thangkas decorated by contemporary Tibetan creatives are becoming popular around the world art investor and collection market. Therefore, the research of Thangka images has excellent cultural and historical value.

The portraits Thangka is one of the crucial components of Thangka images. As one of the essential elements in portraits Thangka images, the headdress contains an abundance of religious information and is a manifestation of the sacredness status of the Buddha. Different types of headdresses show various responsibilities and attributes of the Buddha. Semantic segmentation tries to classify pixels into one of the categories, which is essential in machine vision and understanding. Therefore, the semantic segmentation of the headdress of Buddha is a valuable research task, which could promote the understanding and description of Thangka images. Thangkas are powerful instructional tools that portray the

Received May 18, 2021; revised November 30, 2021; accepted February 3, 2022.
Communicated by Ching-Hsien Hsu.
[+] Corresponding author.

Buddha's life, the lifestyles of different influencing lamas, or other divinities and Hindu gods.

In recent years, the research of Thangka image segmentation mainly focuses on the damaged area of the image [1, 2]. Only a few scholars have been involved in the segmentation of the headdress of Buddha in portraits Thangka images. They proposed a headdress of the Thangka segmentation algorithm based on the characteristics of headlights of some Buddha [3]. This algorithm locates the headdress area by a threshold, then uses the visual attention model and area connectivity to detect the headlight area, and subsequently, the location range is being used as the headdress. A method for detecting RHT circles has been proposed [4]. The Randomized Hough transform is a stochastic modified version of the classical Hough transform that is frequently used to detect curves (straight line, circle, ellipse, *etc.*) Firstly, this method uses the characteristics of the headlights to locate the area of the headdress. Then, it combines the spatial distribution of color, the outer contour of the headdress, and the edge detection results to segment the headdress. These works have good segmentation effects in portraits Thangka images with round headlights. However, these methods do not combine the segmentation results with semantic information, and these methods cannot segment the headdress in portraits Thangka images without round headlights. The convolutional neural networks can solve these problems well due to its excellent performance in the area of semantic segmentation. Thangkas, also known as 'Tankas,' are huge items of works of art depicting the life of Buddhist, different influencing Lamas, as well as other 'Bodhisattva' deities. Thangka works of art are usually created on cloth garments accompanied by a silk picture. Devdis paintings are works of art formed on the wall.

Fig. 1. Sample Thangka image.

The fully supervised semantic segmentation network has made good advances in modern years [5-8, 23, 24] and attained outstanding results. Since the task of segmentation is to allocate a label from a label set to every pixel in an image, a significant number of pixel-level captions are needed for having trained the supervised segmentation network. However, the pixel-level annotation of the image is a labor-intensive work, which makes it expensive. Semantic segmentation is the process of assigning a classifier to each pixel in the image (such as flower, person, road, sky, ocean, or car). Each pixel in an image is assigned a label from a sticker set by semantic segmentation. In the partially controlled setting, moreover, the set of data contains images and correlating mappings that really are easily obtainable, including such markers of objects in the image.

Weakly supervised semantic segmentation has lately made significant progress. Those methods are explored to learn semantic segmentation with the supervision of image-level [10-12, 25, 26], scribble [13], and bounding boxes [14-17]. For example, given an image and its CAMs, AffinityNet [10] builds a neighborhood graph and calculates the semantic similarity between the adjacent coordinates. Computer-aided manufacturing, also known as Computer-aided Modeling or Computer-aided Machining, is the use of an operating system to manage industrial equipment and associated hardware in the manufacturing of workpieces. To infer object affinities, the proposed Deep Affinity Network (DAN) learns compressed and still comprehensive features of pre-detected items at several abstraction levels and needs to perform extensive and comprehensive pairing of possible combinations of those characteristics in any two frames. DAN also needs to take into account multiple objects which appear and disappear among video sequences. For reputable online tracking, we just use eventually results from efficient affinity calculations to correlate objects in the consecutive frames deep into the image sequences. Finally, the ScribbleSup modifies the CAM through semantic diffusion to obtain a more distinct object shape. ScribbleSup [13] is predicated on a graphical model that understands system parameters and reproduces data from scribbled notes to uncleared pixels. SDI [16] tried to start generating section proposals by incorporating the MCG [18] and GarbCut [19] methods. However, the main research objects of these methods are natural images that usually used to segment objects such as animals, vehicles, people, and landscapes. To the best of knowledge, no study on Thangka image segmentation using a convolutional neural network has indeed been published. Considering the labeling is cheaper for weakly supervised semantic segmentation. Labeling a pixel-level annotation, for example, costs about 15 times more than labelling a bounding box and 60 times more than labelling an image class [9]. Especially the cost of labeling could be getting larger for annotation because of the rich structure in Thangka image. It is a better choice to segment the headdress of Buddha by using the weakly supervised semantic segmentation method. Because of the complex structure of the Thangka image, as shown in Fig. 1, and the lack of spatial position information of the target object in the image-level labeling, the semantic segmentation methods based on image-level labeling are difficult to segment the headdress effectively. Compared with the image-level method, the semantic segmentation method based on scribble provides the location information of several pixels. The core of this method is to propagate pixel information from scribbles to unmarked pixels and generate a mask for training through the similarity between superpixels. Due to the production materials and the composition characteristics of Thangka, the headdress in some images is highly consistent with its surrounding background, as shown in the first row in Fig. 1. As a result, it is challenging for the semantic

segmentation procedure relies on scribble to retrieve the headdress mask of these Thangka images adequately for training. The box-level semantic segmentation methods are distinct from the two methods mentioned above. The box-level annotation in these methods can provide both object position and rough boundary information. These box-level weakly supervised semantic segmentation methods are normally applied for high-quality natural images, which have good segmentation effects while reducing the workload of annotation. However, Thangka images are rich in color and structure and without depth characteristics, which are quite distinguished from natural images. Thus, these box-level weakly supervised semantic segmentation methods have limited effectiveness in the headdress segmentation. In summary, the existing method of weakly supervised semantic segmentation is not good enough to segment the headdress of Buddha in portraits Thangka images. Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals. Being able to learn dense semantic representations of images without supervision is an essential problem in computer vision. With human supervision and insight, the process of accomplishing final segmentation accuracy (this is in contrast to automatic segmentation, where no human intervention and guidance is required). In this paper, in order to overcome the limitations of the existing headdress segmentation methods, we propose a weakly supervised semantic segmentation method for headdress with box-level annotations based on the structural features. This method can be flexibly merged with existing fully supervised semantic segmentation networks and convert it into a weakly supervised network for segmentation of the headdress. Experimental studies demonstrate the effectiveness outperforms other state-of-the-art softly semantic segmentation technique of box-level captions in the segmentation of Buddha's headdress. Our contributions to this report are as follows: We built the first dataset of box-level annotations for the headdress of Buddha for portraits Thangka images.

- We are the first to use convolutional neural networks for semantic segmentation of Thangka images.
- We proposed a weakly supervised semantic segmentation method with box-level annotations of the headdress of Buddha in portraits Thangka images.

The remaining portion of this article is arranged as follows. Part two goes over the specific details of the suggested protocol. Chapter 3 describes the dataset and assessment performance measures utilized, and also going to perform confirmation and comparative experiments. Section four must go over the proposed method. Finally, fifth section outlines our findings.

2. PROPOSED METHOD

2.1 Overview

Most of the existing semantic segmentation methods based on convolutional neural networks require pixel-level annotation data for training, and these annotations need to be costly. By analyzing the Thangka image, it is found that the shape of the headdress is more complicated than the natural image, which makes the labeling cost is higher. Therefore, considering the complexity of the Thangka image, we choose to carry out the research of weak supervised semantic segmentation based on the box-level annotation. This work aims to extract a more accurate edge of the headdress from the annotation area.

First of all, by changing the Threshold of Canny algorithm [20] and comparing the

experimental results, it was found that we could obtain different rough edge of the headdress in the image. The 'Canny' method uses two thresholds. For example, if the threshold is [0.1 0.15] then the edge pixels above the upper limit (0.15) are considered and edge pixels below the threshold (0.1) are discarded. Therefore, we processed the input box image in this step and named this process as Canny processing. Because of Thangka image has some problems in long-term preservation, such as weathering. There were many interferences in the rough edge obtained by the Canny algorithm, and it could not be used as the final feature mask. Therefore, in the next step, the Edline algorithm [21] was used to extract line segments based on the images obtained by the Canny algorithm. To keep the sharp edges, you need to use a more sophisticated filter than the Gaussian blur. Two easy options are the Bilateral filter or the Guided filter. These two filters are effortless to implement, and they provide good results in most cases: gaussian noise removal preserving edges. Then, we formulate the rules according to the position and the inclination angle of the line segment, which not only removed most of the noise but also obtained the set of key points. These key points can effectively cover the headdress area. We call this process EDLines processing. Finally, we divide the contour of the headdress into two parts, take the center point of the bottom edge of the image as the origin, then connect all the coordinate points, and take the longest and shortest two-line segments at a certain angle. The points corresponding to the obtained longest side are connected as the upper edge of the headdress, and the points corresponding to the shortest side are connected as the lower edge of the headdress. Then, we expand the upper edge while eliminating the protrusion of the lower edge, and ultimately obtain a smaller effective contour range relative to the bounding box. We define this process as Polygons processing.

Because our method is divided into three main sections: Canny processing, EDLines processing, and Polygons processing, we named it CEDLines-Polygons. The results indicate that EDLines generates similar or better two lines than LSD while trying to run up to 11 seconds faster. The overall algorithm flow chart is shown in Table 1. This approach is flexible to merged with existing, fully supervised semantic segmentation networks. In this work, we chose Mask R-CNN [7] as the backbone network due to its effectiveness, as shown in Fig 2. We use the CEDLines-Polygons method to generate and update the GT

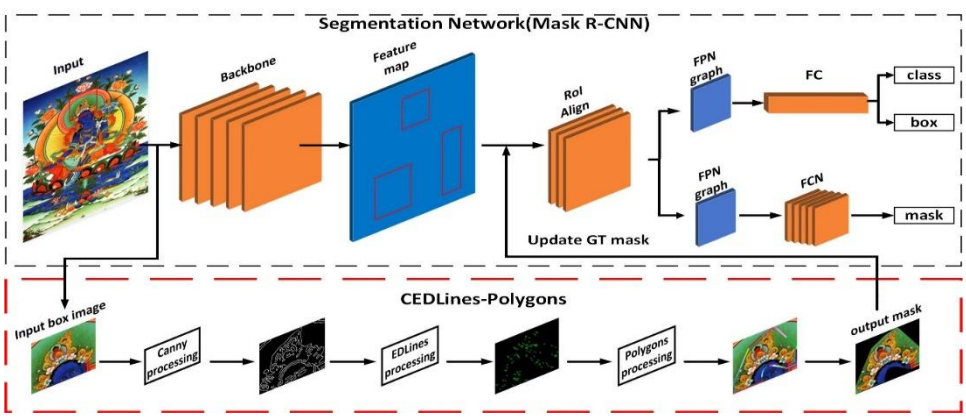


Fig. 2. The CEDLines-Polygons integrated into Mask R-CNN. The bottom part of the figure is an overview of the CEDLines-Polygons.

Mask values at a suitable place. We will introduce the three parts in detail in the following content of this section. CED – Critical Edge Detection – is proposed and evaluated. The methods take priority given the message or critical links inside the house.

Table 1. Overview of the CEDLines-polygons algorithm.

Proposed method: Overall process of the CEDLines-Polygons algorithm.

Input: I_{box} /* The original image of the box-level labeled area. */
Step 1: Canny processing
Input: I_{box}
$I_g \leftarrow I_{box}$ is processed by the Canny algorithm
Define P_a for the number of pixels of I_{box} and P_c for the number of pixels which value is 255 in I_g.
While $P > 0.1$ **do**
$I_g \leftarrow I_{box}$ changes the threshold
$P \leftarrow P_c / P_a$
Output: I_g

Step 2:
Input: I_g
$L_{org} \leftarrow I_g$ /*L_{org} is a directed line segment containing coordinates obtained by I_g processing through the EDLines algorithm.*/
$L_i \leftarrow$removes 10 percent of the shorter segment in L_{org}
Define P_{start} is the starting point of the directed line segment in L_i and P_{end} is the end point.
If P_{start} to the left of P_{end} **Then** reset the starting point of the line segment
Suppose P_{start} is the coordinate system's centre, the x-axis is horizontal, and the y-axis is vertical.
Define the slope of the line segment as P_{angle}, the set of key points is P_k
For Line in L_i **do**
If P_{end} in the left half I_g and $P_{angle} \subset (-45°, -15°)$ **or** P_{start} in the right half I_g and $P_{angle} \subset (15°, 45°)$ **or** $P_{angle} \subset [-15°, 15°]$ **then** remove the line segments from L_i
As shown in the Fig. 4 (e), divided the image equally into 12 parts.
If P_{start} in the green part on the left **or** Pend in the green part on the right **then** remove the line segments from L_i
As shown in the Fig 4 (f), divided the image equally into 4 parts.
For Line in L_i **do**
If P_{start} is in the upper-left **or** P_{start} is in the lower-right **then** add P_{end} to P_k
If P_{start} in the lower-left **or** upper-right **then** add P_{start} to P_k
Output: P_k

Step 3:
Input: P_k
As shown in the Figs 5 (a) and (b), divide the image into 12 regions, keep the longest and shortest line segments in each region.
Define the point corresponding to the longest line segment as $Ai(i = 1, 2, 3, \ldots)$, the point corresponding to the shortest line segment as $Bi(i = 1, 2, 3, \ldots)$.
Define rules 1 and rules2 /* as shown in the 2.4*/
While Ai in $Ai(i = 1, 2, 3, \ldots)$ Satisfy Rule 1 **do**
Rule 1
Connect Ai sequentially, we can get the upper contours
While Bi in $Bi(i = 1,2,3\ldots)$ Satisfy Rule 2 **do**
Rule 2
Connect Bi sequentially, we can get the lower contours
Finally, Connecting the upper and lower contours to get the feature mask of the headdress
Output: feature mask of the headdress
Output: feature mask of the headdress

2.2 Canny Processing

We use Canny [20], EDLines [21], and other edge extraction and line detection algorithms for preliminary edge extraction experiments. Edge detection is an image analysis method used to detect points in a digital photo that have singularities, or pointed changes in brightness values. Because Thangka is manually drawn, there will be uneven color in the drawing process, and the color will be more uneven after long-term preservation. Therefore, the contours or edges extracted using the above algorithm have much noise and are disorganized, and cannot be directly used as the headdress mask. In image processing, an edge is characterized as a sequence of adjacent implications for the design where the density (grey or colour) values suddenly change. Edges indicate the separation of objects from one ‘s environment. During the experiment, it was found that the noise information contained in the images generated by the Canny algorithm under different thresholds will be different. Therefore, we use the Canny algorithm to process the input box image. Then, we introduce this process. Here we define P_a for the number of pixels of input box image and P_c for the number of pixels which value is 255 in the binary image after the edge detection by the Canny algorithm. Finally, P is obtained by the following definitions:

$$P(t) = \frac{P_c(t)}{P_a} \quad t \subset [0,200) \tag{1}$$

t is the threshold.

(a) (b) (c) (d)

Fig. 3. The results of different values of the Canny algorithm. The first row of pictures is the chignon, the second row is the coronet, and the third row is mitral; (a) Input box images, (b) $P \leq 0.2$, (c) $P \leq 0.1$; (d) $P \leq 0.05$.

As shown in Fig. 3, the contour information in the binary image can be expressed well through the Canny edge detection algorithm. Contours are simply a curve that helps con-

nect all points (along the boundary) that have the same color scheme or intensity. Contour lines are an effective tool for shape analysis and also object recognition. Canny edge detection is a technique of image analysis that detects edges in the image while repressing sound. The Canny edge detector is just an edge detection provider that is able to detect a broad range of edges in pictures that used a multi-stage method. By comparing the experimental results, we found that when the ratio of the pixel value of 255 to the total pixel number of the image processed by Canny algorithm was less or equal to 0.1, most of the noise could be removed and edge information could be effectively retained. The pixel value in the most fundamental form of two images is a 1-bit number indicating whether foreground or backstory. The byte picture is the most prevalent pixel template, where this amount is saved as an 8-bit integer with such a set of possible values ranging from 0 to 255. Therefore, suppose that the final output image of Canny processing is I_g, and the related definition of the function is as follows:

$$I_g = \{t|P(t) \leq 0.1, t \subset [0,200)\}. \quad (2)$$

After processing the Input box image through this process, most of the noise generated by the original Thangka image during long-term storage can be effectively removed, and a rough contour can be generated.

2.3 EDLines Processing

Because I_g generated by Canny processing is composed of curved segments and contains some noise, it still needs further processing. Considering that line segments are easier for subsequent image processing, the EDLines algorithm was used to extract line segments based on the images obtained by the Canny algorithm, and the curves in the image were converted into line segments. Through analyzing the label image of headdress, it is found that the headdress is located in the lower middle of the box labeled image. And the majority of the headdress lines trend from left to right, sloping up to the top and sloping down. At the same time, analyzing the line segment image, it was found that the Buddha's eyes and mouth, some scratches in the image became straight lines with low inclination after the line segment extraction. Therefore, we improved the EDLines algorithm [21] according to the characteristics above. Furthermore, the new EDLines algorithm is used to extract the key points of the headdress. Then, we will introduce EDLines processing. The name EDLines refers to the clean, contiguous (connected) chain of pixel intensities generated by our innovative detection algorithm, the Edge Drawing (ED) algorithm. EDLines, to their accurate result and charging speed, would be perfect for the next generation of true computer vision tasks.

Suppose that L_{org} is the result of line segment extraction by using the EDLines algorithm of I_g. In general, Thangka images may be damaged during long-term preservation. Even after the Canny algorithm processing, there may still be small noises. Thus, while the outcome of other detection algorithms requires extra processing to abilities and self-bonding boxes, that might or might not be feasible or result in factual errors, ED not just to produces excellently linked bounding boxes by default, but also does so at a breakneck pace once tried to compare to other detection algorithms. This causes the EDLines algorithm to generate some small noise line segments when extracting edge line segments.

Therefore, L_i is obtained by removing 10% of the shorter line segments in L_{org}. Then, L_i is processed according to the characteristics of the headdress to obtain the P_k. P_k is the key point set of contour. The detailed steps are given in Fig. 4.

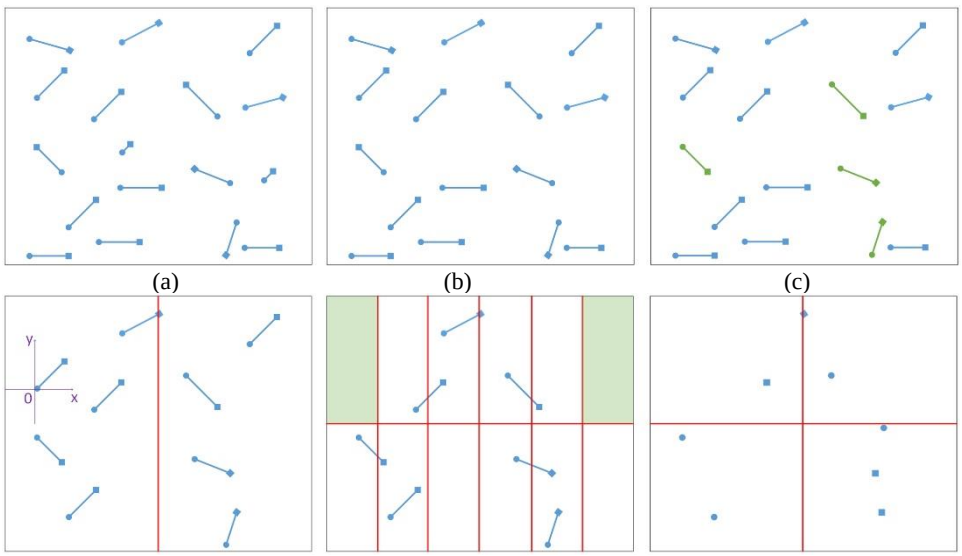


Fig. 4. (a)-(f) is the order of the EDLines processing.

(a) It is the corresponding graph of the line segment set L_{org}. The circular endpoints of the line segment in figure are the starting points and defined as P_{start}. The square endpoints of the line segment in the Fig are the ending points and defined as P_{end}.
(b) It is the graph after L_i removes 10 percent of the shorter segment.
(c) If P_{start} to the left of P_{end}, reset the starting point of the line segment.
(d) Suppose P_{start} is the center of the x-axis is horizontal, and the y-axis is vertical in this coordinate system.We define the slope of the line segment as P_{angle}. Removing the line segments where P_{end} is located in the left half and $P_{angle} \subset (-45°, -15°)$, the line segments where P_{start} is located in the right half and $P_{angle} \subset (15°, 45°)$, and the line segments of $P_{angle} \subset (-15°, 15°)$.
(e) As shown in the Fig, the graph is equally divided into 12 parts. Removing the line segment where P_{start} is located in the green part on the left and P_{end} is located in the green part on the right.
(f) As shown in the Fig, the graph is equally divided into 4 parts. Get the line segments whose P_{start} is in the upper-left and lower-right parts and add its P_{end} to P_k. At the same time, the P_{start} of the line segments whose P_{start} are located in the lower-left and upper-right sections are also added to P_k.

The image generated after the EDLines processing is the set of points of the headdress after eliminating the noise.

2.4 Polygons Processing

This process is to transform the points set P_k into a valid contour. These points are too numerous and scattered to directly used to generate a mask, so unnecessary points need to be removed. By analyzing the headdress structure, we designed the Polygons processing to obtain a small effective contour range relative to the bounding box. The detailed steps are given in Fig. 5.

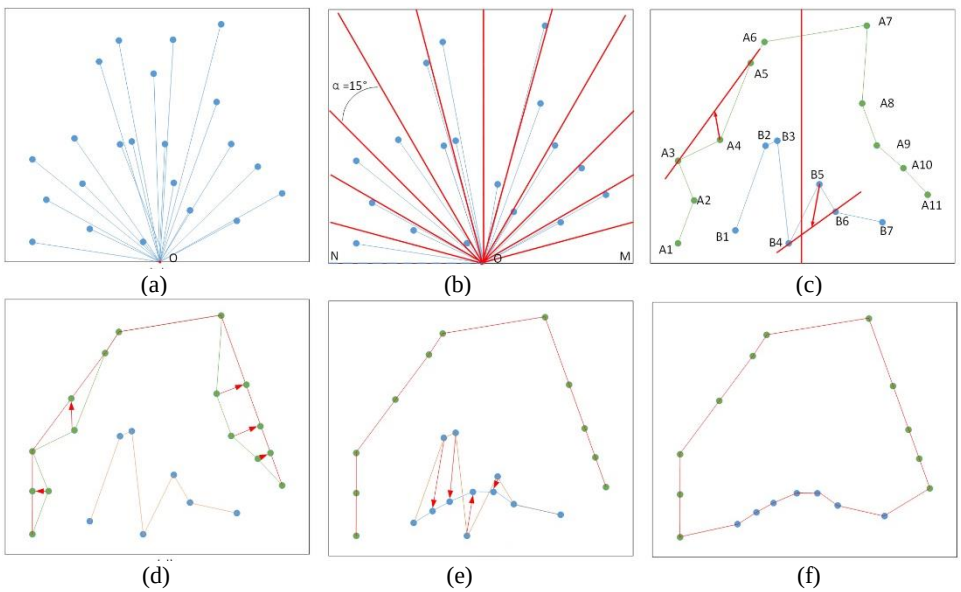


Fig. 5. (a)-(f) is the order of polygons processing.

1. Suppose that *O* is the midpoint at the bottom of the image. *O* is connected to each point in P_k, as shown in (a).
2. Then, as shown in (b), suppose that *O* is the origin of coordinates, the line segment *ON* represents 0° and the line segment *OM* represents 180°. A ray is generated every 15° to divide the image into 12 regions.
3. We keep the longest and shortest line segments in each region and define the line segment as the longest if there is only one in this region. Sort the longest and shortest line segments according to the angle of the line segments in ascending order. From that, we get *Ai*(*i* = 1, 2, 3, …) and *Bi*(*i* = 1, 2, 3, …), where *Ai*(*i* = 1, 2, 3, …) represents the point corresponding to the longest line segment, and *Bi*(*i* = 1, 2, 3, …) represents the point corresponding to the shortest line segment. Finally, *Ai*(*i* = 1, 2, 3, …) and *Bi*(*i* = 1, 2, 3, …) are connected in order, as shown in (c).

It was found that when these edges were connected, there would be concave or convex. To solve this problem, we define Rules 1 and 2. Suppose that *A3*, *A4*, and *A5* are three adjacent points, L_{A35} is a straight line generated by *A3* and *A5*. *B4*, *B5* and *B6* are three adjacent points, L_{B46} is a straight line generated by *B4* and *B6*.

Rule 1: Moving *A4* to the center of L_{A35} when *A4* is at the right of L_{A35} and on the left half of the image or *A4* is at the left of L_{A35} and on the right half of the image.
Rule 2: Moving *B5* to the center of L_{B46} when *B5* is above or below both *B4* and *B6*.

According to Rule 1, iterate through $Ai(i = 1, 2, 3, \ldots)$ sequentially from *A1* until no point satisfies Rule 1. It can get the upper contour of the headdress, and the result is shown in the red line in (d). According to Rule 2, $Bi(i = 1, 2, 3, \ldots)$ are traversed in sequence from *B1* until no point satisfies it. We can get the lower outline of the headdress, and the blue line is shown in (e). Finally, Connecting the upper and lower contours to get the feature mask of the headdress, which is (f) in this section.

3. EXPERIMENTS

First, we describe the dataset and evaluation metrics used in the experiments. The method's effectiveness is then substantiated through experiments and especially in comparison to state-of-the-art methodologies. Eventually, we went over the experiment results.

3.1 Dataset and Evaluation Metrics

Dataset. The Thangka image's headdress is categorized into three parts: coronet, chignon, and mitral. There is currently no standardized dataset for the segmentation of Thangka images. To verify the performance of the proposed method, we formed a dataset of 883 Thangka images to box-level labels. There are 345 coronets, 285 chignons, 247 mitrals, and 6 unknowns in the dataset. In order to perform the experiment better, we divide the data set according to the ratio of 7:2:1. During the experiment, 621 pictures have been used for training, 178 for verification, and 86 for test results. These images have been scaled to 1024x1024 pixels. Each image contains a category in coronet, chignon, mitral, and unknown. The sample image was shown in Fig. 1.

Evaluation Metrics. The standard evaluation metric, mean pixel Intersection-over-Union (mIoU) and Mean-Pixel-Accuracy (MPA) [22], by adopting. Mean Intersection-Over-Union is a quality assessment measures for semantic segmentation that calculates the Bounding box for each conceptual class before calculating the mean across classrooms. The related definition of the function are as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \tag{2}$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}. \tag{3}$$

$k+1$ is the total number of classes, and P_{ij} is the pixels from class I which are deduced to belong to class *j*. In those other words, P_{ii} signifies the true positive rate, whereas P_{ij} as well as P_{ji} signify the number of false negatives, respectively [22].

To prove the performance of our proposed algorithm, we choose the algorithm of

GrabCut [19], Mask R-CNN [7], and start-of-the-art methods (*i.e.*, SDI [16] and WSIS_ BBTP [17]) for comparison.

3.2 Method Verification

Implementation details. We use the widely used Mask R-CNN [7] model as the network structure to implement the proposed method. This work makes use of the Mask R-CNN base code from Matterport on Github (https://github.com/matterport/Mask RCNN). The parameters momentum, learning_rate, weight decay, images_per_gpu, setps_per_epoch and epochs are set to 0.9, 10^{-3}, 10^{-4}, 2, 150, 300. The heads layer is used as the training strategy for this experiment. All experimental studies in this article have been trained on a computer integrated with an Intel(R) Xeon(TM) E5-2690 CPU @ 2.90 GHz, SAMSUNG DDR3 1866MHz 16GB*2 Memory, and an Nvidia TitanX GPU with 12 GB of memory.

Baseline models. A convolutional neural network is a type of neural network that is designed primarily to handle image data in pattern recognition systems. We use two models as the baseline methods to compares. One method is GrabCut [19], and the other is the Mask R-CNN[7]. Both methods are trained using the box-level annotation Thangka dataset. GrabCut is the established technique to estimate an object segment from its bounding box. The backbone network we use is Mask R-CNN. Mask R-CNN is the most sophisticated supervised anomaly detection, case, and feature extraction method. Table 2 summarizes the test results. GrabCut is a graph-cut-based image segmentation method. The method assumes the color information of the target image and the professional experience using a Gaussian mixture method, beginning with a consumer frame around the object to be fragmented. Furthermore, the sample images of the result are shown in Fig. 6. As shown in Table 2, using the Tangka image dataset, our proposed method is 13.11% higher than the GrabCut segmentation result and 10.03% higher than the Mask R-CNN (box-level annotation training) segmentation result. A bounding box is a purely theoretical rectangle that acts as a reference point for object recognition and creates a collision package for that entity.

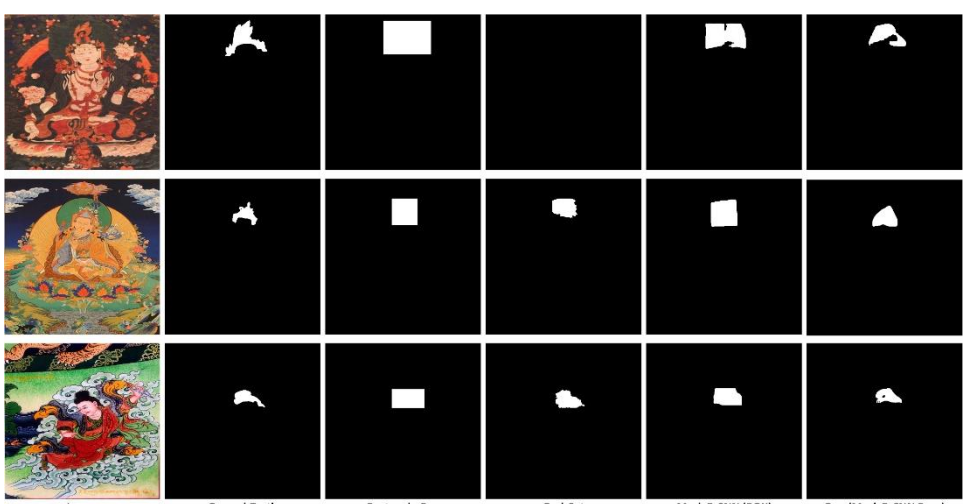


Fig. 6. Examples of the segmentation results of our approach and baseline methods. The first column is the original image. The second column is the ground-truth image. The third columns are the rectangle-box image. The following three columns show the results of GrabCut, Mask R-CNN (Box-level annotation), and proposed method (base on Mask R-CNN).

Table 2. Evaluation of Semantic segmentation results from different methods. B for box-level labels, M for mask-level labels.

METHOD	Publication	Supervision	mIoU(%)	MPA(%)
GT BOXES	–	–	46.87	47.16
GRABCUT [19]	–	B	44.71	47.47
MASK R-CNN [7]	ICCV'17	B	47.79	48.32
SDI [16]	CVPR'17	B	50.26	51.95
WSIS_BBTP [17]	NIPS'19	B	51.71	52.89
OURS	–	B	**57.82**	**61.04**

3.3 Comparison with the State-of-the-Art Methods

In order to verify the effectiveness of the proposed method, we compare the proposed method with SDI [16] and WSIS_BBTP [17]. The method of WSIS_BBTP and SDI are the state-of-the-art box-level supervision instances and semantic segmentation methods. Table 2 displays the results. Box-level annotation is used to train SDI, WSIS BBTP, and our method. The results show that our proposed method outperforms SDI and WSIS_ BBTP in the semantic segmentation of Thangka pictures. In our experiments, the index mIoU of the proposed method outperforms SDI around 7.56% and WISIS around 6.11%. Meanwhile, MPA is 9.09% higher than SDI and 8.15% higher than WISIS. As shown in Table 2. We provide examples of comparative experimental results in Fig. 7.

Fig. 7. Segmentation results of our approach and state-of-the-art methods. The original image is in the first column. The picture in the second column is the ground-truth image. The outcomes of SDI, WSIS BBTP, and the proposed approach are shown in the following 3 columns. The first row is a coronet, the second set is a chignon, and the last two rows are two mitral of varying shapes.

Simultaneously, we discovered that combining the model proposed in this work with the Mask R-CNN model can substantially increase learning rate. In Table 3, we show the training time for each model. In Table 3, the training time by the original Mask R-CNN model is shown in the first line, and the training time under the same condition is shown in the fourth line when the model proposed in this paper is combined with the Mask R-CNN model. It can be seen that the training speed is accelerated by 10.7%.

Table 3. Comparison of model performance.

METHOD	Supervision	Epoch	Traning time
MASK R-CNN	B	300	24h 14m 5s
SDI	B	300	32h 24m 20s
WSIS_BBTP	B	300	22h 47m 26s
MASK R-CNN+**OURS**	B	300	**21h 38m 20s**

4. DISCUSSION

The experimental results show that an improved algorithm is completely feasible, and it can effectively improve the segmentation effect. For a more comprehensive analysis, we provide the mIoU for the pre-class in the experiment in Table 4. From the data, this method has better performance. Experiments on the Thangka dataset with headdress segmentation display that our model outperformed state-of-the-art methodologies in the region of semi-supervised segmentation of the headdress. Table 4 shows that our method has a good effect on the segmentation of the coronet and chignon in the headdress, but there are still some shortcomings in the segmentation of the mitral. These good effects are due to the relatively concentrated pixel distribution of coronets and chignons in the Thangka image, so the contours extracted by our method contain less invalid pixel information, which is conducive to better training. However, the pixel distribution of some mitral is relatively loose, as shown in the second row and the third column in Fig. 8, this results in our approach inclu-

Fig. 8. Examples of the proposed method. The first row is the successful results, and the second row is the failure results.

Table 4. Per-class mean IoU (%).

Methods	background	coronet	chignon	mitral	mean
SDI	97.53	32.48	38.45	32.60	50.26
WSIS_BBTP	97.88	35.65	38.79	34.53	51.71
OURS	98.50	**47.03**	**48.97**	**36.81**	**57.82**

ding more invalid pixels in mask generation, making the poor training effect. In addition, the long-term preservation caused some Thangka images to be of low quality. This makes it noisier, which will also affect the segmentation results. This problem is worth being deeply studied in future work. Some examples of success and failure are shown in Fig. 8.

5. CONCLUSIONS

This paper mainly studies the weak supervised semantic segmentation of headdress in the portraits Thangka images and proposes a box-level weakly supervised semantic segmentation method for the headdress of Buddha in the portraits of Thangka images. According to the structure and color characteristics of the Thangka image, the method first uses the Canny processing to extract the line drawing, then it uses the EDLines processing to extract the key points from the line drawing, and finally, the method uses Polygons processing to generate the mask area., This approach can be flexibly merged with the existing fully supervised semantic segmentation method to transform it into a box-level weakly supervised semantic segmentation method. Experiment results show that the proposed method is effective. And it outperforms existing weakly supervised semantic segmentation methods of box-level annotations in the area of Thangka image. In the future, we will consider further improving our method through semi-supervised and small sample learning to achieve better segmentation results. Moreover, we will explore the segmentation of other semantic objects in Thangka images, such as implements and sitting tables, so as to further promote the study of thangka images.

ACKNOWLEDGMENT

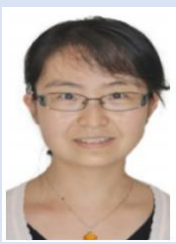
This work was supported in part by The Nature Science Foundation of China under Grant No. 62061042, 61561042, 61862057, Fundamental Research Funds for the Central Universities under Grant No.31920200066, program for innovative research for innovative research team of SEAC (No.2018 [98]) and Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education and by the special fund for talent introduction of northwestern nationalities university.

REFERENCES

1. W. Hu, W. Wang, and Z. Liu, "Damaged region segmentation of Thangka based on domain knowledge and multi-feature," *Journal of Central South University* (*Science and Technology)*, Vol. 47, 2016, pp. 2326-2333.
2. G. Amudha, "Dilated transaction access and retrieval: Improving the information

retrieval of blockchain-assimilated internet of things transactions," *Wireless Personal Communications*, 2021, pp. 1-21.
3. S. Dhote, C. Vichoray, R. Pais, S. Baskar, and P. M. Shakeel, "Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce," *Electronic Commerce Research*, Vol. 20, 2020, pp. 259-274.
4. J. Gao, H. Wang, and H. Shen, "Task failure prediction in cloud data centers using deep learning," *IEEE Transactions on Services Computing*, 2020, Vol. 15, pp. 1411-1422.
5. T. D. Ngo, T. T. Bui, T. M. Pham, H. T. Thai, G. L. Nguyen, and T. N. Nguyen, "Image deconvolution for optical small satellite with deep learning and real-time GPU acceleration," *Journal of Real-Time Image Processing*, Vol. 18, 2021, pp. 1697-1710.
6. G. Manogaran, P. M. Shakeel, H. Fouad, Y. Nam, S. Baskar, N. Chilamkurti, and R. Sundarasekar, "Wearable IoT smart-log patch: An edge computing-based Bayesian deep learning network system for multi access physical monitoring system," *Sensors*, Vol. 19, 2019, p. 3030.
7. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2980-2988.
8. G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925-1934.
9. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, and D. Ramanan, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 740-755.
10. J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4981-4990.
11. T. Shen, G. Lin, C. Shen, and I. Reid, "Bootstrapping the performance of webly supervised semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1363-1371.
12. X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1354-1362.
13. D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159-3167.
14. J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1635-1643.
15. L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4545-4554.
16. A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 876-885.
17. C. Hsu, K. J. Hsu, C. C. Tsai, Y. Y. Lin, and Y. Y. Chuang, "Weakly supervised

instance segmentation using the bounding box tightness prior," *Advances in Neural Information Processing Systems*, 2019, pp. 6582-6593.
18. J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, 2016, pp. 128-140.
19. C. Rother, V. Kolmogorov, and A. Blake, "Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, Vol. 23, 2012, p. 3.
20. J. A. Canny, "Computational Approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, 1986, pp. 679-698.
21. C. Akinlar and C. Topal, "EDLines: A real-time line segment detector with a false detection control," *Pattern Recognition Letters*, Vol. 32, 2011, pp. 1633-1642.
22. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv Preprint*, 2017, arXiv:1704.06857.
23. Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, Vol. 7, 2019, pp. 44247-44257.
24. X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proceedings of IEEE International Conference on Image Processing*, 2019, pp. 1440-1444.
25. L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Transactions on Image Processing*, Vol. 29, 2020, pp. 225-236.
26. X. Li, H. Ma, and X. Luo, "Weaklier supervised semantic segmentation with only one image level annotation per category," *IEEE Transactions on Image Processing*, Vol. 29, 2020, pp. 128-141.
27. https://github.com/WHUT-DCRC/Thangka.



Wenjin Hu received bachelor degree and master degree from Lanzhou University of Technology in 2003 and 2008. She received doctoral degree from Lanzhou University of Technology of China. She is an Associate Professor in Northwest University for Nationalities. Her main research interest is pattern recognition and image processing.

Jiahao Meng received bachelor degree from Zhengzhou University of Light Industry in 2016. Now he is studying for a master's degree at the Northwest Minzu University. His main research interests include image segmentation and image processing.

Li Jia received bachelor degree from Shangqiu Normal University in 2016. Now she is studying for a master's degree at the Northwest Minzu University. Her main research interests include NLP and data mining.

Fuliang Zeng received bachelor degree from Long Dong University of Computer Science and Technology in 2018. Now he is studying for a master's degree at the Northwest Minzu University. His main research interest is image recognition.

Panpan Xue received bachelor degree from Gansu Agricultural University in 2019. Now she is studying for a master's degree at the Northwest Minzu University. Her main research interest is few-shot object detection.