Received January 24, 2021, accepted March 19, 2021, date of publication March 23, 2021, date of current version April 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068199

Thangka Mural Line Drawing Based on Cross Dense Residual Architecture and Hard Pixel Balancing

NIANYI WANG[1,2], **(Member, IEEE), WEILAN WANG**[2], **AND WENJIN HU**[1]

[1]School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, China
[2]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730030, China

Corresponding author: Nianyi Wang (livingsailor@qq.com)

This work was supported in part by NSFC under Grant 61862057 and Grant 62061042, and in part by the Program for Innovative Research Team of State Ethnic Affairs Commission (SEAC) under Grant [2018]98.

ABSTRACT Thangka murals are precious cultural heritage for Tibetan history, literature, and art. Digital line drawing of Thangka murals plays a vital role not only as an abstracted expression of Thangka for art appreciation but also as a fundamental digital resource for Thangka protection. Digital Thangka line drawing can be categorized as image edge detection, which as a fundamental problem for computer vision, aims to extract visually salient edges from images. Varieties of high-level computer vision tasks depend on edge detection. Although existing non-learning and learning-based edge detection methods have progressed, they failed to generate semantically plausible thin edges, especially thin in-object edges. We propose a novel deep supervised edge detection solution Richer In-object Thin Edge Network (RITE-Net) to generate line drawings of Thangka mural images. Compared to existing studies, firstly a new Cross Dense Residual architecture (CDR) is proposed to propagate abundant edge features effectively from shallow layers to deep layers of CNN using a long-range feature memory; Secondly, a new Hard Pixel Balancing (HPB) based loss function strategy is designed to focus on hard pixel distinguishment. Experiments and tests on different datasets show that the proposed RITE-Net is able to produce more visually plausible and richer thin edge maps comparing to the existing methods. Both objective and subjective evaluations validated the competitive performance of our method.

INDEX TERMS Edge detection, hard pixel balancing, line drawings, Thangka.

I. INTRODUCTION

As a kind of Tibetan encyclopedia [1], [2], Thangka murals have become a very important cultural Heritage to study Tibetan history, literature, and art. Digital line drawing of Thangka murals plays a vital role not only as an abstracted expression of Thangka for art appreciation but also as a fundamental digital resource for Thangka protection.

Digital Thangka line drawing can be classified as image edge detection, which as a fundamental problem for computer vision, aims to extract visually salient edges and object boundaries from images. Varieties of high-level tasks such as image recognition [3], object detection [4], [5], segmentation [6], [7], and image to image translation [8] depend on edge detection.

Although edge detection methods have progressed, there exist obvious limitations. 1) Traditional non-learning methods tend to either generate fake edges or lack important edges because of the absence of high-level image semantics (Fig.1(b)). This is because the traditional methods usually extract low-level local cues of brightness, colors, gradients, and textures, to classify edge and non-edge pixels. However, it is difficult to use these local low-level cues to represent object-level line information and image semantics [9]. This inherent disadvantage prevents non-learning methods from producing semantically plausible and visually satisfactory edges. 2) Learning-based methods face serious problems too although they have outperformed the non-learning methods in recent years due to the capability of extracting and learning of high-level image semantics [9]–[11]. The most common problems for learning-based methods such as incorrect edges and large-scale shadow areas that around edges also inevitably lead to unsatisfactory edge detection results.

Deep supervised learning has great potential for edge detection and has been used in existing learning-based edge detection methods such as HED [11], RCF [9], BDCN [12]

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

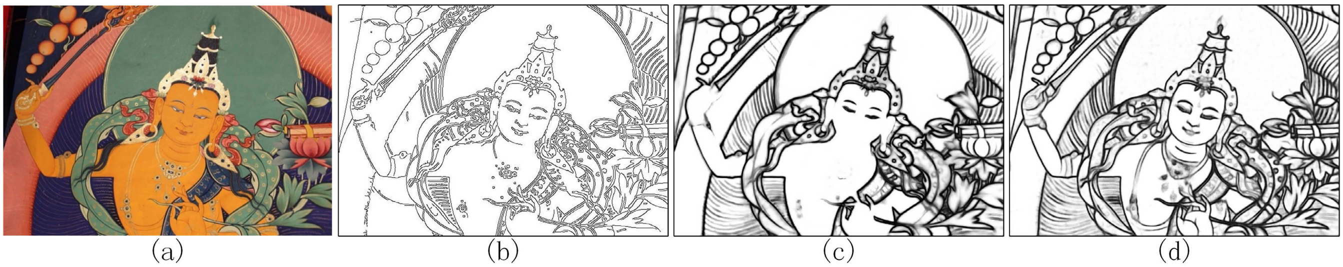


FIGURE 1. The proposed RITE-Net achieved the best edge-map with richer image semantics and more edge details (Fig.1(d)). The existing learning-based methods failed to generate vital edge details (eg: no face, hand and breast details in Fig.1(c)) while traditional non-learning method such as Canny tends to either generate fake edges or lacks important edges because of the absence of high-level image semantics (Fig.1(b)). Fig.1(a)) is a Thangka mural.

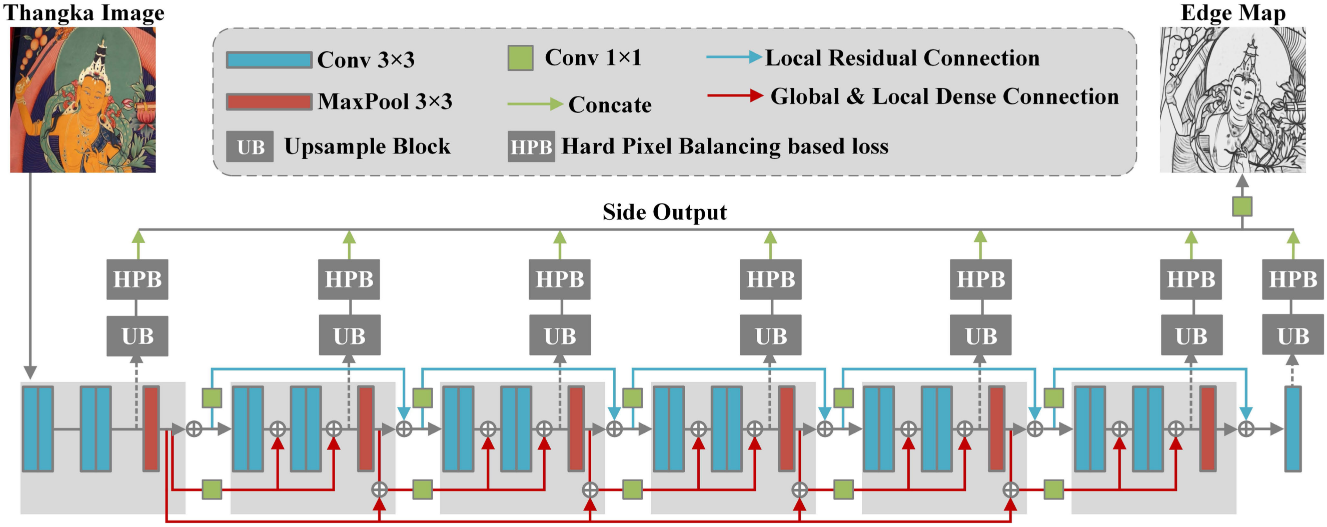

Thangka Image
Conv 3×3
MaxPool 3×3
UB Upsample Block
Conv 1×1
Concate
HPB Hard Pixel Balancing based loss
Local Residual Connection
Global & Local Dense Connection
Edge Map
Side Output
HPB
UB

FIGURE 2. The proposed Cross Dense Residual architecture (CDR) is capable of propagating abundant edge features effectively from shallow layers to deep layers of CNN through two different kinds of skip connections: 1) Local Residual Connection (blue lines in Fig.2), and 2) Global & Local Dense Connection (red lines in Fig.2). As inter-block connections, the Local Residual Connections are responsible for residual feature propagation between two neighboring blocks; the Global & Local Dense Connections propagate both local features from the previous block and global features from the output of the first block into each convolution layer of the current block. The proposed Hard Pixel Balancing (HPB) based losses are adopted to supervise six side outputs and the final output at the end of the network.

due to its outstanding feature extraction performance. However, two challenges prevent existing learning-based methods from generating satisfactory edge maps:

Challenge 1: No effective network architecture to propagate edge information properly from shallow layers to deep layers in a neural network. As a result, only some thick contours and lines appear in the final output while many vital edge details disappear (for example, in Fig.1(c), the face, fingers, and breast areas of the Buddha disappear).

Challenge 2: Lack of attention to the hard pixels near edges. The hard pixels refer to the pixels nearby edges in an image, which are more difficult to be classified into edge/non-edge pixels. Without paying attention to the hard pixels, both shadow areas and false edges could inevitably appear because of misjudging non-edge pixels as edge pixels by algorithms (we term this situation as false positive). Meanwhile, real edges also could disappear because of misjudging edge pixels as non-edge pixels (false negative) (for shadow areas see Fig.7 column 3-5). Owing to shadow areas, false edges, and missing real edges, this hard pixel problem inevitably degrades the quality of edge maps.

The objective of this work is to provide a better edge detection solution (not only boundary/contour detection between objects but also edge detection inside objects) to generate line drawings of Thangka mural images. In this paper, we propose a Richer In-object Thin Edge Network (RITE-Net, Fig.2) based on cross dense residual architecture and hard pixel balancing. The proposed method consists of two parts:

First (solution for Challenge 1), a novel Cross Dense Residual architecture (CDR) is proposed to propagate abundant edge features effectively from shallow layers to deep layers of CNN. CDR consists of densely connected blocks. CDR not only allows skip-connections between the outputs of previous and current blocks (short-range feature fusion) but also integrates features of shallow layers directly into the layers locate inside each deep block (long-range feature fusion). CDR stabilizes the training of deeper networks and propagates edge features correctly with a long-range feature memory.

Second (solution for Challenge 2), a new Hard Pixel Balancing (HPB) based loss function strategy is proposed to pay attention to hard pixel distinguishment. HPB reduces the relative loss for well-classified pixels, putting more focuses on hard, misclassified pixels. HPB enables training accurate edge detectors in the presence of a vast majority of easy background pixels by preventing the generation of false edges and the missing of true edges.

Experiments show that our RITE-Net provides a high-performance edge detector for generating accurate line drawings for Thangka murals. In summary, the main contributions of this work are:

1. A robust supervised deep CNN leaning framework for edge detection is proposed, termed as RITE-Net: Richer In-object Thin Edge Network.
2. A novel Cross Dense Residual architecture (CDR) is proposed to stabilize the training of deeper networks and propagate edge features correctly with a long-range feature memory.
3. A new Hard Pixel Balancing (HPB) based loss function strategy is proposed to focus on hard pixel distinguishment and training.
4. For the first time, we introduce an end-to-end learning-based method to generate line drawings for the traditional Thangka mural images. The proposed method produces more plausible and richer thin edge maps compared to the existing methods.

II. RELATED WORK

A. EDGE DETECTION

As one of the most fundamental problems in computer vision for several decades, there have emerged a large number of works on edge detection. For a detailed review see [13], [14]. Broadly speaking, most edge detection methods can be categorized into three groups: 1) traditional edge detectors; 2) classic learning based methods; and 3) recent deep learning methods.

1) TRADITIONAL NON-LEARNING EDGE DETECTORS

Early non-learning methods mainly focused on the utilization of texture, intensity, and color gradients. Sobel detector computes the gradient map and then generates edges by thresholding the gradient map [15]. The widely adopted Canny detector [16] uses Gaussian smoothing as a preprocessing step and then adopts a bi-threshold to produce edges. [17], [18] proposed zero-crossing theory based edge detectors. Among the early edge detectors, the Canny operator is more robust to noise. In fact, the Canny operator is still very popular across various tasks now because of its notable efficiency. Besides these early pioneering methods, mammal vision systems were associated with the processing of edge and contour [19]–[21]. However, these methods have poor accuracy.

For the traditional non-learning methods, the absence of high-level image semantics is the biggest issue, which inevitably leads to either generating fake edges or lacking real edges.

2) CLASSIC LEARNING-BASED EDGE DETECTION METHODS

Konishi *et al.* [22] proposed the first data-driven methods. [23] formulated changes in brightness, color, and texture as Pb features, and trained a classifier to combine the information from these features. Arbelaez *et al.* developed Pb into gPb [24]. In [25], Sketch token was proposed to represent mid-level information for contour and object detection. StructuredEdges [26] employed random decision forests to represent the structure of image patches and to produce edges by inputting color and gradient as features. Other methods such as BEL [27], Multi-scale [28], sparse representation [29], and dictionary learning [30], obtained acceptable results in most of the cases as well.

However, all these classic learning-based methods still have limitations in challenging scenarios since they tended to rely on hand-crafted features and had to employ sophisticated learning paradigms.

3) DEEP LEARNING METHODS

In recent years, deep convolutional neural networks (CNN) based methods achieved state-of-the-art performance for edge detection, such as N4-Fields [31], Deep-Contour [32], DeepEdge [33], and CSCNN [34]. Some newest methods have demonstrated promising F-score performance improvements. Xie and Tu [11] proposed an efficient and accurate edge detector, HED, by connecting side output layers with a holistically-nested architecture. Using the same architecture of HED, Liu *et al.* proposed an improved architecture RCF [9], to learn richer deep representations by extracting features from all convolutional layers. [12] proposed Bi-Directional Cascade Network (BDCN) structure, where each layer is supervised at its specific scale, rather than directly applying the same supervision to all CNN outputs. Inspired by both HED and Xception, [10] proposed a new method DexiNed to generate thin edge maps.

Although the aforementioned CNN-based methods have advanced the state-of-the-art significantly, all of them still inevitably generate fake edges or miss true edges since they fail to propagate rich edge details correctly through deep networks. Besides this, they paid no attention to hard pixel distinguishment, which leads to inaccurate edge maps as well.

B. HARD SAMPLE AND HARD PIXEL BALANCING

Hard sample training is not only a challenge for image classification because of class imbalance but also a potential problem for the task of edge detection since the essence of edge detection is classifying all pixels of an image to edge pixels and non-edge pixels. This imbalance problem makes the training of a deep CNN inefficient as most pixels are easy negatives that contribute no useful signal for learning [35]. A common solution to overcome class imbalance is to perform hard negative mining [36]–[38] that trains hard examples more during training [39]. In contrast, [35] proposed a novel solution, termed as Focal Loss, to naturally handle the class imbalance and to efficiently train on all examples equally without complex sampling strategy and also without easy negatives overwhelming the loss.

Similar to hard sample training, hard pixel training is also a challenge for edge detection. As mentioned before, hard pixels refer to the pixels that are near to edges and are more difficult to be classified into edge or non-edge pixels. The difference between hard sample and hard pixel is that the former is a problem of sample-level imbalance (hard samples and easy samples in a dataset) while the latter is a problem of pixel-level imbalance (hard pixels and easy pixels in an image).

Although the hard sample problem was partially resolved by the above-mentioned methods, they only focused on hard samples, but not on hard pixels. To the best of our knowledge, the hard pixel problem has not been discussed before in edge detection tasks.

III. METHODOLOGY

The proposed RITE-Net(Richer In-object Thin Edge Network) consists of two parts: 1) a novel Cross Dense Residual architecture (CDR) to propagate abundant edge features effectively from shallow layers to deep layers of CNN; and 2) a new Hard Pixel Balancing based loss function strategy (HPB) to pay attention to hard pixel distinguishment by reducing the relative loss for well-classified pixels, and putting more focuses on hard, misclassified pixels.

A. CROSS DENSE RESIDUAL ARCHITECTURE (CDR)

Although the existing deep learning methods such as HED [11], RCF [9], have achieved the state of art results for edge detection, they still fail to propagate abundant edge details correctly through deep networks, which inevitably leads to inaccurate edge maps. Fig.2 shows the proposed CDR architecture, which consists of six blocks. The CDR network not only produces an edge map by a side Upsampling Block (UB) at each block, but also generates a final edge map at the end of the network. All six side edge maps from the UBs and the final edge map are concatenated together to output a fused edge map.

There are two different kinds of skip connections in CDR: 1) Local Residual Connection (blue lines in Fig.2), and 2) Global & Local Dense Connection (red lines in Fig.2), which are inspired by ResNet [40] and DenseNet [41], respectively. As inter-block connections, the Local Residual Connections are responsible for residual feature propagation between two neighboring blocks; each Global & Local Dense Connection propagates both local features from the previous block and global features from the output of the first block into each convolution layer of the current block. Furthermore, 1×1 convolution (green square block in Fig.2) is adopted in the both connections to build correlation of channels.

In each CDR block, there are two convolution modules and one max-pooling layer. Each convolution has a 3×3 kernel size. For convolution modules, the first convolution is followed by batch normalization and ReLU activation while the second convolution is followed only by batch normalization. The max-pooling layer is set by 3×3 kernel size and stride 2.

Since the CDR is a multi-scale learning, it is necessary to restore each side output to the scale of ground truth image by Upsampling Block (UB). Inspired by HED [11] and DexiNed [10], we adopt a transpose convolution strategy in UB.

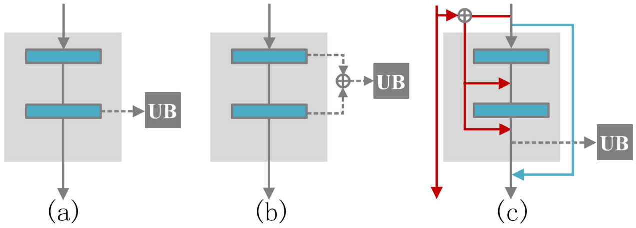


FIGURE 3. The proposed CDR improved the architecture of existing learning methods by adopting Local Residual Connection (blue line in Fig.3(c)) and Global & Local Dense Connection (red lines). HED [11] only utilizes CNN features from the last layer of each conv block (Fig.3(a)); RCF [9] improves HED by utilizing CNN features from all layers of each conv block (Fig.3(b)). Nevertheless, both of them have no global or local skip connections between blocks, which weakens the propagation ability of edge features through a deep neural network.

Summary of Advantages: Fig.3 illustrates the differences and improvements of our CDR to some existing representative deep learning methods. Powered by both the Local Residual Connection (blue lines in Fig.2 & Fig.3) and the Global & Local Dense Connection (red lines in Fig.2 & Fig.3), our CDR not only stabilizes the training of deeper networks, but also propagates richer edge details correctly with a long-range feature memory.

B. HARD PIXEL BALANCING (HPB)

Hard pixel training problem has a crucial impact on the performance of edge detection tasks because it is challenging to classify hard pixels into edge or non-edge pixels. Fig.4 is an illustration of hard pixels. We divide all pixels of an image into three types: 1) easy negative pixels that locate far from lines (upper right and down left areas of Fig.4(d) that pointed by 2 green arrowed lines); 2) easy positive pixels that locate inside of lines (pointed by 1 blue arrowed line); and 3) hard pixels that locate nearby lines (pointed by 2 orange arrowed lines). The hard pixels (both positive and negative) have more chance to be misclassified, which leads to false positive pixels and false negative pixels. As a result, there will not only appear false edges (false positive pixels) but also lack true edges (false negative pixels).

We propose a Hard Pixel Balancing (HPB) based loss function strategy to resolve this problem. HPB utilizes two different loss functions for two training stages respectively: 1) for pretrain stage, inspired by FocalLoss [35], we adopt a Pixel-level FocalLoss (PFL) to more focus on hard pixels and less focus on easy pixels; and 2) for finetune stage, a new piecewise cross entropy loss function is proposed to focus only on hard pixels.

1) PIXEL-LEVEL FocalLoss (PFL) FOR PRETRAIN STAGE

Let (i, j) be a pixel in an image I, $y(i, j)$ be ground truth pixel, $y'(i, j)$ be prediction $P(y = 1|x)$. For simplicity, we use y for

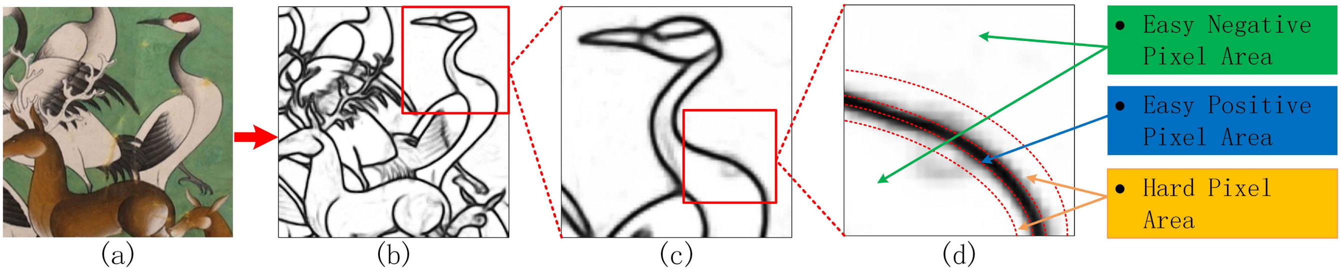


FIGURE 4. The illustration of hard pixel problem. It is challenging for learning based methods to classify the hard pixels that locate nearby lines (pointed by 2 orange arrowed lines in Fig.4(d)) into edge or non-edge pixels. As a result, there will not only appear false edges (false positive pixels) but also lack true edges (false negative pixels). Contrarily, easy pixels are much easier to be classified (both easy negative pixels that pointed by 2 green arrowed lines and easy positive pixels that pointed by 1 blue arrowed line; for more details, see Section 3.2.

$y(i,j)$, y' for $y'(i,j)$. For a single image, we define PFL as:

$$L_{PFL} = \sum_{i,j}(-\alpha\, y\,(1-y')^{\gamma}\, log(y') - (1-\alpha)\,(1-y)\,(y')^{\gamma}\, log(1-y')) \quad (1)$$

where α is an adaptive parameter that can be configurated automatically for each image, γ is a hyperparameter. α is defined as:

$$\alpha = Y\_ \,/\, (Y_{+} + Y\_) \quad (2)$$

where Y_{+} and $Y\_$ denote the edge and non-edge ground truth label sets of an image, respectively. As for a typical image, the distribution of edge/non-edge pixels is heavily biased: 90% of the ground truth is non-edge, like in HED [11], α is a class-balancing weight for each image which can easily offset the imbalance between edge and non-edge pixels.

The hyperparameter γ is value greater than 0 that set manually, the role of $(1-y')^{\gamma}$ is to reduce the focus on easy positive pixels. Because the value of $(1-y')^{\gamma}$ has less impact to easy positive pixels ($y' > 0.5$) than to hard positive pixels ($y' < 0.5$). Similarly, the role of $(y')^{\gamma}$ is to reduce the focus on easy negative pixels.

Although the loss structure of our PFL is similar to FocalLoss [35], there are still key differences:

1) The proposed PFL is to focus on hard pixels of an image (pixel-level balancing) while FocalLoss is to focus on hard samples of an image dataset (sample-level balancing).
2) The parameter α is an adaptive parameter that can be automatically set in our PFL while it is a hyperparameter that needs to be set manually in FocalLoss.

2) PIECEWISE CROSS ENTROPY LOSS (PCEL) FOR FINETUNE STAGE

After the first stage training converges, a new piecewise cross entropy loss is designed for finetune stage.

We first define a function $g(y, y')$ as:

$$g(y, y') = 1 - f(1-m-y)\, f(1-m-y') - f(y-m)\, f(y'-m) \quad (3)$$

where y, y' are GT and prediction of a pixel respectively, m is a hyperparameter to determine upper threshold ($y{=}1 \wedge y'{<}m$) and lower threshold ($y = 0 \wedge y' > (1-m)$) of hard pixels. in this work we set $m = 0.6$. $f(x)$ is a piecewise function:

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & else \end{cases} \quad (4)$$

According to Eq.(3) and (4), we have:

$$g(y, y') = \begin{cases} 0, & y = 1 \wedge y' \geq m \\ 1, & y = 1 \wedge y' < m \\ 1, & y = 0 \wedge y' > (1-m) \\ 0, & y = 0 \wedge y' \leq (1-m) \end{cases} \quad (5)$$

Then we define the Piecewise Cross Entropy Loss as:

$$L_{PCEL} = \sum_{i,j} g(y, y')(-y\, log(y') - (1-y)\, log(1-y')) \quad (6)$$

According to Eq.(5) and (6), the loss $L_{PCEL} = 0$ for easy positive pixels ($y = 1 \wedge y' \geq m$) or easy negative pixels ($y = 0 \wedge y' \leq (1-m)$), that means the network no longer propagate the loss of easy pixels backward; as a result, it only trains the hard pixels.

Summary of Advantages: The proposed HPB strategy reduces the relative loss for well-classified pixels, putting more focuses on hard, misclassified pixels. HPB is capable of learning richer edge details and training accurate edge detectors by avoiding the generation of false edges and the missing of true edges greatly.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET AND EXPERIMENT SETUP

Although there are some popular datasets such as BSDS [42], NYUD [43], PASCAL [44], CID (contains only 40 contour images) [20], all of them are for training and evaluation of contour / boundary detection (here the words contour / boundary mean the edges between objects) or semantic segmentation tasks, but not for detailed edge detection (both in-object and inter-object edges). BIPED [10] is a new and better dataset for edge detection. Fig.5 shows the difference of datasets, we can easily find that the samples of BIPED (Fig.5(e)) contain rich in-object edges while BSDS (Fig.5(a)), NYUD (Fig.5(b)), PASCAL (Fig.5(c)), CID (Fig.5(d)) have no in-object edges. To the best of our knowledge, There are

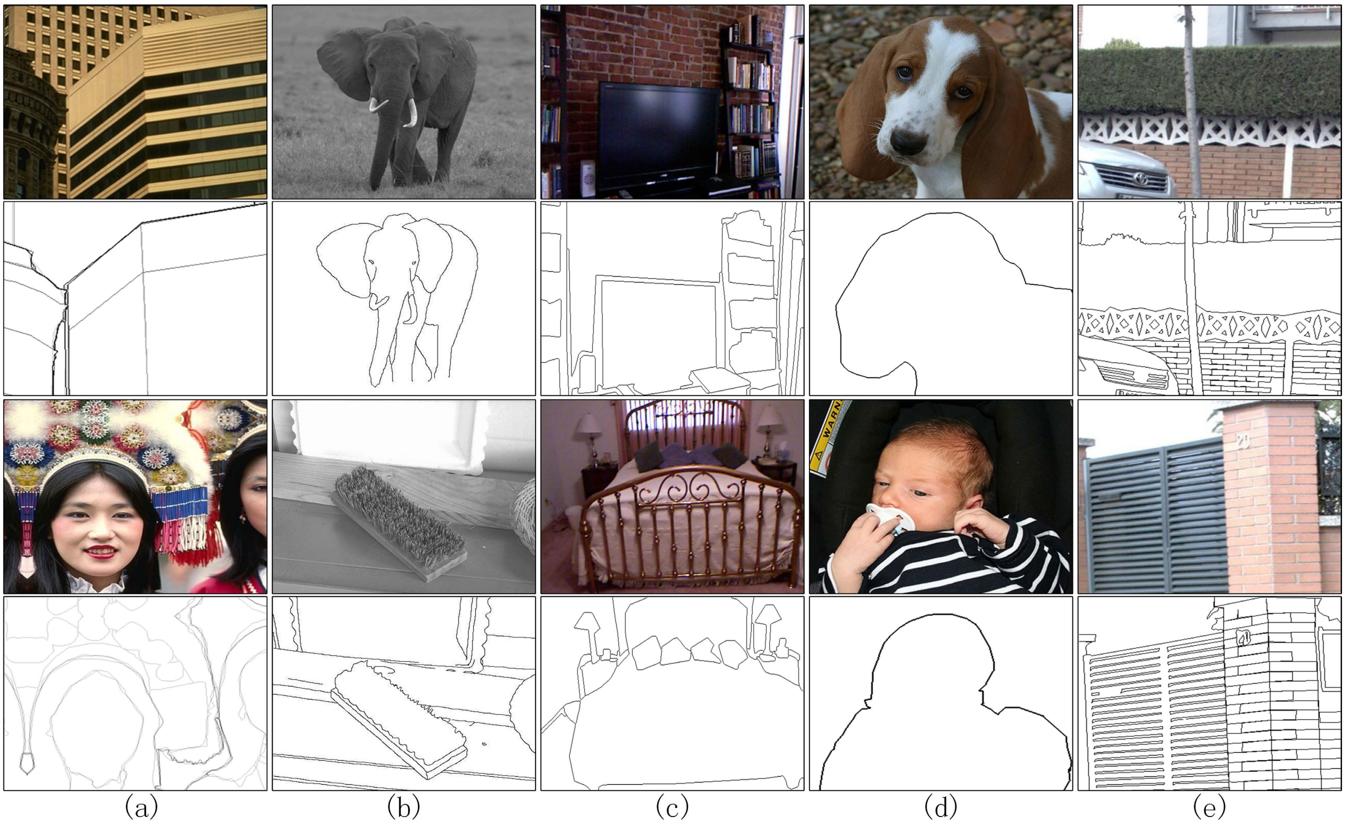


FIGURE 5. BIPED [10] (Fig.5(e)) is the most suitable public dataset for our edge detection task although there are some more popular datasets such as BSDS [42], NYUD [43], PASCAL [44], CID (contains only 40 contour images) [20]; in most cases, these datasets are suitable for contour / boundary detection (here the words contour / boundary mean the edges between objects) or semantic segmentation tasks, but not for detailed edge detection (both in-object and inter-object edge details). We can easily find that the samples of BIPED (Fig.5(e)) contain rich in-object edges while BSDS (Fig.5(a)), NYUD (Fig.5(b)), PASCAL (Fig.5(c)), CID (Fig.5(d)) have no in-object edges.

only two publicly available datasets intended for edge detection MDBD [45], BIPED [10]. However, edges in MDBD dataset have not been cross validated and contain wrong annotations in some cases [10]. Hence, edge detector algorithms trained by these incorrectly annotated edges would be misguided.

In this work, we adopt BIPED as a training dataset since there is no edge map ground truth for Thangka murals. BIPED contains 250 1280 × 720 images that have been carefully annotated and cross-checked. The dataset was randomly split into a training set (65%), a validation set (15%), and a testset (20%). Like HED, data augmentation process has been performed on the training set and validation set by random splitting, rotation, cropping, and flipping. For testing, we use both the above-mentioned BIPED testset and our Thangka images.

We trained our model on NVIDIA P100 GPU. The model converges after 50K iterations with a batch size of 8 at the pretrain stage and converges after 30K iterations at the finetune stage. In order to validate and compare the performance of our proposed edge detector, three state-of-the-art methods HED [11], RCF [9], and BDCN [12] were also trained with the BIPED dataset; in addition, the classic non-learning detector CANNY [16] was also evaluated.

B. PERFORMANCE ON BIPED TESTSET

1) QUANTITATIVE EVALUATION

We adopt the BIPED testset to evaluate edge detection accuracy using three standard measures: Average Precision (AP), F-score at both Optimal Dataset Scale (ODS) and Optimal Image Scale (OIS), where $F = \frac{2\times Precision\times Recall}{Presicion+Recall}$. Like previous works [9]–[12], a standard non-maximal suppression (NMS) [26] technique is adopted to obtain the final edge maps. Like HED and DexiNed [10], we provide two final prediction results RITE-f (the concatenation and fusion of all 7 outputs) and RITE-a (the average of all 7 outputs) for comparisons.

The quantitative evaluation results are shown in Table.1 and Fig.6. In Table.1, the RITE-a achieved the best AP while the RITE-f achieved the highest F-score both on OIS and ODS. This validated the effectiveness of our methods. In Fig.6, although the precision rates of RITE-f and RITE-a are not the highest under the condition of low recall rate, they are much higher than HED, RCF, and BDCN while the recall rate increases. This should be due to the proposed long-range feature memory of our CDR and the hard pixel balancing mechanism of our HPB. An interesting observation can be found in Fig.6: RITE-a has the longest curve, which demonstrates that RITE-a achieves the best performance than

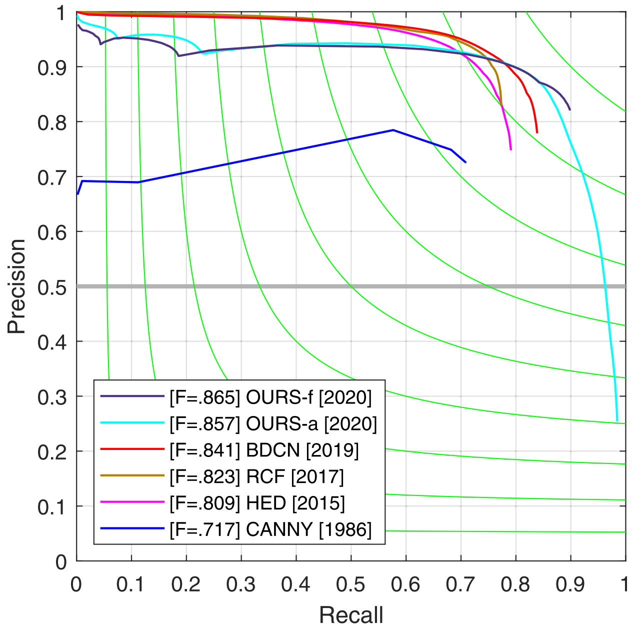


FIGURE 6. Our RITE-a and RITE-f are more robust than the other methods. Although the precision rates of RITE-f and RITE-a are not the highest under the condition of low recall rate, they are much higher than HED, RCF, and BDCN while the recall rate increases. This should be due to the proposed long-range feature memory of CDR and the hard pixel balancing mechanism of our HPB. An interesting observation can be found in Fig.6: RITE-a has the longest curve, which demonstrates that RITE-a achieves the best performance than the others including RITE-f under high recall rate. In this work, we prefer to use RITE-a. Whenever the term RITE is used it corresponds to RITE-a.

TABLE 1. The quantitative evaluation results show that the proposed RITE-Net outperformed the traditional edge detector Canny and three learning-based methods HED, RCF and BDCN. We provide two prediction results RITE-f (the concatenation and fusion of all 7 outputs) and RITE-a (the average of all 7 outputs) for comparisons. the RITE-a achieved the best AP while the RITE-f achieved the highest F-score both on OIS and ODS.

Method	ODS	OIS	AP
Canny [16]	0.717	0.722	0.514
HED [11]	0.809	0.817	0.764
RCF [9]	0.823	0.826	0.754
BDCN [12]	0.841	0.844	0.806
RITE-a	0.857	0.863	**0.883**
RITE-f	**0.865**	**0.872**	0.826

the others including RITE-f under high recall rate. In this work, we prefer to use RITE-a. Whenever the term RITE is used it corresponds to RITE-a.

2) QUALITATIVE EVALUATION

Fig.7 shows some visual comparisons on the BIPED testset. We can easily find the visual differences between different edge detection methods. The proposed RITE-Net is capable of generating visually clear thinner edge maps compared to the classic non-learning method Canny and the existing representative learning methods. Canny detector either produces fake edges or lacks real edges; what's more, it contains no high-level image semantics. Although HED [11], RCF [9] and BDCN [12] can generate better edge maps than Canny, their results still not only contain plenty of unnecessary shadow areas around edges but also generate fake edges.

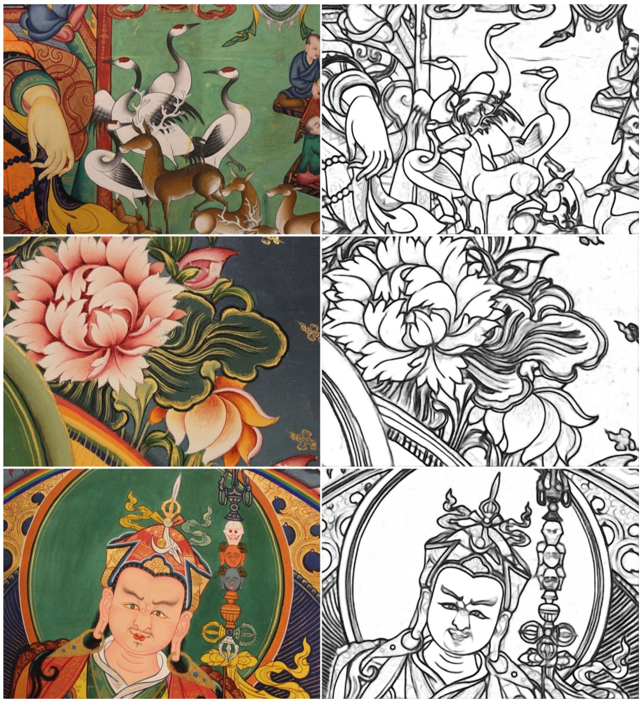


FIGURE 9. More edge examples of Thangka murals generated by our method. Our RITE-Net achieved state-of-art edge detection performance by generating visually satisfactory and semantically plausible edge maps for Thangka mural images.

C. PERFORMANCE ON THANGKA TESTSET

Fig.8 shows the visual comparison on the Thangka testset. Our RITE-net achieved the best edge maps with clear edges of both in-objects and inter-objects. Canny detector inevitably either produces fake meaningless edges or lacks semantically important edges because of its inherent characteristics we mentioned before. HED [11], RCF [9] and BDCN [12] failed to generate some important edges (especially those in-object edges). For the edge results of HED, RCF, and BDCN, nearly all lines in the red area are totally missing in the first row of Fig.8; in the second row, edge details of the Buddha's breast and the instrument are missing in HED and RCF while the edge details of lines of halos disappear in all 3 methods; in the third row, the inner edge details of the Buddha's ribbon are also totally missing in HED, RCF, and BDCN. On the contrary, our RITE-net is capable of generating nearly every important edge and represents the edge semantics of the original mural images well.

From the above test results we can find that although HED, RCF and BDCN are all trained with the same dataset BIPED like our RITE-Net, they still cannot achieve satisfactory edge maps. By combining our CDR architecture and HPB loss functions, the edge detection performance of the RITE-Net outperformed the other non-learning and learning methods. Fig.9 shows more edge maps of Thangka murals generated by our RITE-Net.

We performed a user study on the Thangka testset. 50 professional evaluators (experts and students of Thangka art) and 150 general evaluators are invited to evaluate Thangka line drawings generated by the above 5 methods. We set

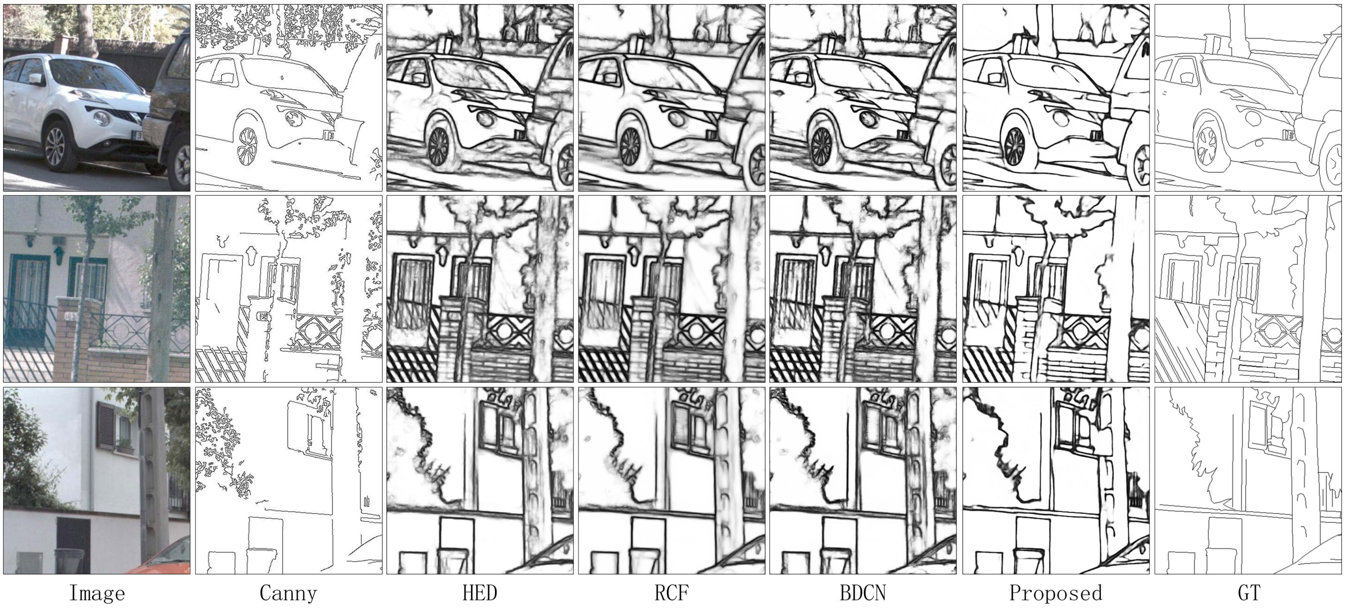


FIGURE 7. The proposed RITE-Net is capable of generating visually clear thinner edge maps comparing to the classic non-learning method Canny and the existing representative learning methods. Canny detector either produces fake edges or lacks real edges; what's more, it contains no high-level image semantics. Although HED [11], RCF [9] and BDCN [12] can generate better edge maps than Canny, their results still not only contain plenty of unnecessary shadow areas around edges but also generate fake edges.

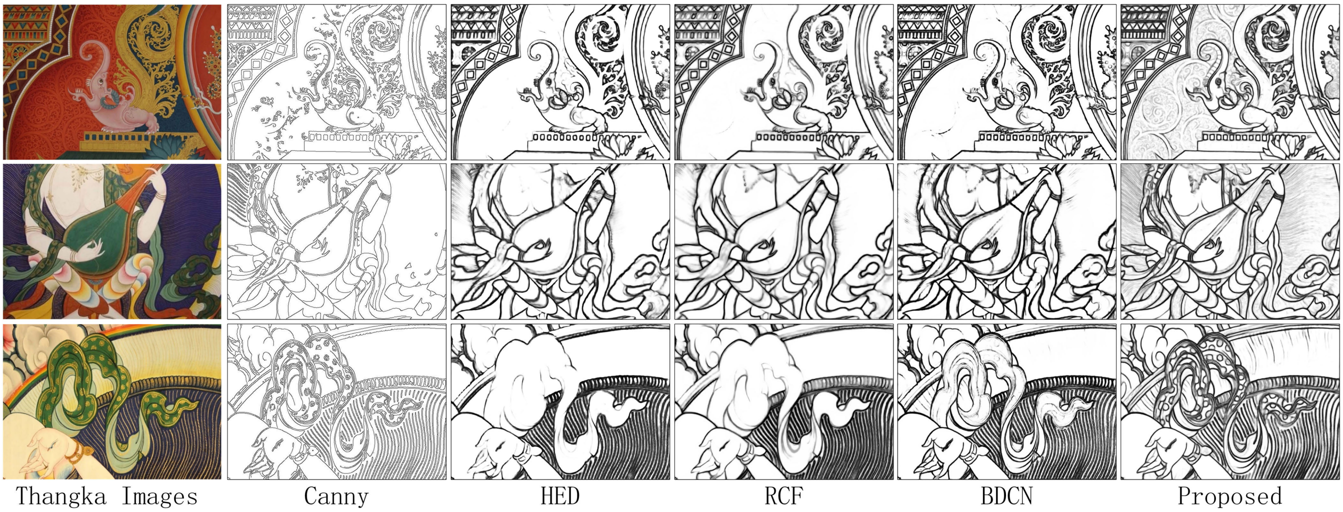


FIGURE 8. Our RITE-net achieved the best edge maps with clear edges of both in-objects and inter-objects on Thangka testset. Canny detector inevitably either produces fake meaningless edges or lacks semantically important edges because of its inherent characteristics we mentioned before. HED [11], RCF [9] and BDCN [12] failed to generate some important edges (especially those in-object edges). For the edge results of HED, RCF, and BDCN, nearly all lines in the red area are totally missing in the first row of Fig.8; in the second row, edge details of the Buddha's breast and the instrument are missing in HED and RCF while the edge details of lines of halo disappear in all 3 methods; in the third row, the inner edge details of the Buddha's ribbon are also totally missing in HED, RCF, and BDCN. On the contrary, our RITE-net is capable of generating nearly every important edge and represents the edge semantics of the original mural images well.

up 2 criteria for subjective evaluation: 1) line correctness and 2) artistic quality. We randomly chose 10 mural images from the Thangka testset for the invited participants to evaluate. Fig.10 shows the results. In Fig.10, the numbers are the averaged percentages of the votes on the edge maps that are regarded to be the best ones. In our user study, 57% of the 150 general evaluators believed that our results have the best visual performance while 48% of the 50 professional evaluators chose our edge maps as the best ones. The user study shown that our method outperformed the other methods and achieved the best performance.

D. ABLATION STUDY: EFFECTIVENESS OF HPB

To verify the effectiveness of the proposed HPB strategy, we replaced our HPB-based loss functions of stage1 (pretrain) and stage2 (finetune) with the Weighted Cross Entropy loss (more details, see [11]), respectively; and trained the model till converge. Table.2 and Fig.11 show a comparison of the ablation study of HPB-based loss functions.

In Table.2, we can see that the model trained by HPB strategies both in stage1 and stage2 (row 3 of Table.2) achieved the best scores of OIS, ODS, and AP. Meanwhile, we can also observe that the model trained only by the stage1 HPB

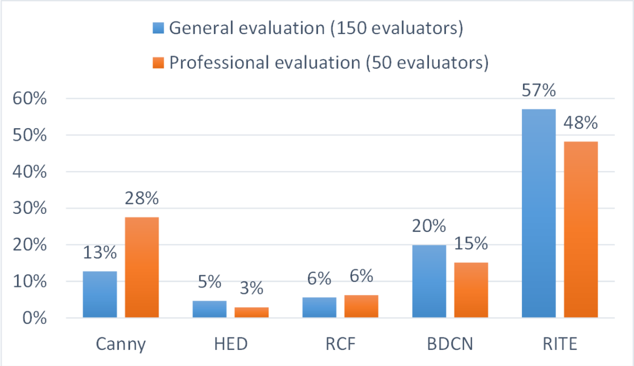


FIGURE 10. The user study shown that our method outperformed the other methods and achieved the best performance. 50 professional evaluators (experts and students of Thangka art) and 150 general evaluators are invited to evaluate Thangka line drawings generated by the 5 methods. 10 Thanka mural images are randomly chosen from the Thangka testset for the invited participants to evaluate. The numbers in the figure are the averaged percentages of the votes on the edge maps that are regarded to be the best ones. 57% of the 150 general evaluators believed that our results have the best visual performance while 48% of the 50 professional evaluators chose our edge maps as the best ones.

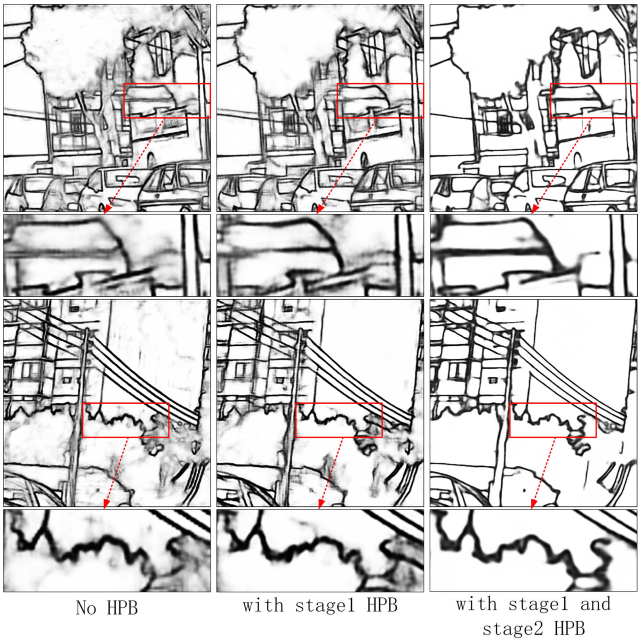


FIGURE 11. The proposed 2-stage HPB loss functions enabled our RITE-Net to produce visually clearer and thinner edges by focusing on those hard pixels near edges (see the third column), while there are grey shadow areas appear in both the first column (without any HPB strategy) and the second column (only with stage1 HPB strategy). These grey shadow areas are caused by the lack of focuses on hard pixels. While comparing the first and the second columns, it's not hard to observe there are some visual improvements in the edge maps of the second columns as well due to the stage1 HPB strategy we deployed.

strategy (row 2) outperformed the model deploying no HPB strategy (row 1).

In Fig.11, the proposed 2-stage HPB loss functions enabled our RITE-Net to produce visually clearer and thinner edges by focusing on those hard pixels near edges (see the third column of Fig.11), while there are grey shadow areas appear in both the first column (without any HPB strategy) and the second column (only with stage1 HPB strategy). These grey shadow areas are caused by the lack of focuses on hard pixels. While comparing the first and the second columns, it can be obviously observed that there are some visual improvements in the edge maps of the second column as well due to the stage1 HPB strategy we deployed.

TABLE 2. Ablation study validated the effectiveness of the proposed Hard Pixel Balancing strategies. The model trained by HPB strategies both in stage1 and stage2 (row 3 of Table.2) achieved the best scores of OIS, ODS, and AP. We can also see that the model trained only by the stage1 HPB strategy (row 2) outperformed the model deploying no HPB strategy (row 1).

Method	ODS	OIS	AP
No HPB	.828	.833	.823
With stage1 HPB	.838	.846	.800
With stage1 and stage2 HPB	.857	.863	.883

V. CONCLUSION

We proposed an edge detection method RITE-Net based on cross dense residual architecture and hard pixel balancing for generating accurate line drawings of Thangka murals. The proposed method consists of two parts: 1) a novel Cross Dense Residual architecture (CDR) to stabilize the training of deeper networks and propagate edge features correctly with a long-range feature memory; 2) a new Hard Pixel Balancing (HPB) based loss function strategy to focus on hard pixel distinguishment and training. For the first time, line drawings of the traditional Thangka mural images are successfully produced by an end-to-end learning-based method. Experiments on different datasets show that our method is able to produce more visually plausible and richer thin edge maps comparing to the existing methods. Both objective and subjective evaluations validated the performance of our method.

REFERENCES

[1] *Thanka Introduction*. Accessed: Jan. 2021. [Online]. Available: https://en.wikipedia.org/wiki

[2] W. Wang, J. Qian, and X. Lu, "Research outline and progress of digital protection on thangka," *Adv. Topics Multimedia Res.*, vol. 2, p. 67, Feb. 2012.

[3] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1270–1281, Jul. 2008.

[4] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 36–51, Jan. 2008.

[5] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[6] K. Zhang, L. Zhang, K.-M. Lam, and D. Zhang, "A level set approach to image segmentation with intensity inhomogeneity," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 546–557, Feb. 2016.

[7] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[9] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, Aug. 2019.

[10] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1923–1932.

[11] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[12] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3828–3837.

[13] X.-Y. Gong, H. Su, D. Xu, Z.-T. Zhang, F. Shen, and H.-B. Yang, "An overview of contour detection approaches," *Int. J. Autom. Comput.*, vol. 15, no. 6, pp. 656–672, Dec. 2018.

[14] D. Ziou and S. Tabbone, "Edge detection techniques-an overview," *Pattern Recognit. Image Anal.*, vol. 8, pp. 537–559, Mar. 1998.

[15] J. Kittler, "On the accuracy of the sobel edge detector," *Image Vis. Comput.*, vol. 1, no. 1, pp. 37–42, Feb. 1983.

[16] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[17] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. Roy. Soc. London. B, Biol. Sci.*, vol. 207, pp. 187–217, Feb. 1980.

[18] V. Torre and T. A. Poggio, "On edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 2, pp. 147–163, Mar. 1986.

[19] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 7, pp. 1160–1169, 1985.

[20] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 729–739, Jul. 2003.

[21] K.-F. Yang, S.-B. Gao, C.-F. Guo, C.-Y. Li, and Y.-J. Li, "Boundary detection using double-opponency and spatial sparseness constraint," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2565–2578, Aug. 2015.

[22] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. Chun Zhu, "Statistical edge detection: Learning and evaluating edge cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 57–74, Jan. 2003.

[23] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

[24] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[25] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3158–3165.

[26] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.

[27] P. Dollár, Z. Tu, and S. Belongie, "Supervised learning of edges and object boundaries," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Mar. 2006, pp. 1964–1971.

[28] X. Ren, "Multi-scale improves boundary detection in natural images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 533–545.

[29] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 43–56.

[30] R. Xiaofeng and L. Bo, "Discriminatively trained sparse code gradients for contour detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 584–592.

[31] Y. Ganin and V. Lempitsky, "N^4-fields: Neural network nearest neighbor fields for image transforms," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 536–551.

[32] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3982–3991.

[33] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4380–4389.

[34] J.-J. Hwang and T.-L. Liu, "Pixel-wise deep learning for contour detection," 2015, *arXiv:1504.01989*. [Online]. Available: http://arxiv.org/abs/1504.01989

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[36] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2241–2248.

[37] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 21–37.

[39] S. R. Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7082–7091.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[42] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 416–423.

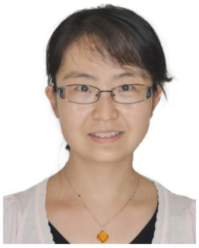
[43] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 746–760.

[44] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[45] D. A. Mély, J. Kim, M. McGill, Y. Guo, and T. Serre, "A systematic comparison between visual cues for boundary detection," *Vis. Res.*, vol. 120, pp. 93–107, Mar. 2016.

NIANYI WANG (Member, IEEE) received the B.S. degree in computer science and the Ph.D. degree in radio physics from Lanzhou University, in 2002 and 2014, respectively. He is currently an Associate Professor with the School of Mathematics and Computer Science, Northwest Minzu University, China. He is also a Visiting Scholar with the Department of Medical Biophysics and Medical Imaging, University of Western Ontario, Canada. His research interests include machine learning, artificial neural networks, image processing, and computer vision.

WEILAN WANG received the B.S. degree in mathematics from Northwest Normal University, Lanzhou, China, in 1983. In 1987, she was a Visiting Scholar with Sun Yat-sen University, Guangzhou, China; Tsinghua University, Beijing, China, from 2001 to 2002; Indiana University, Bloomington, USA, from 2006 to 2007. She is currently a Professor and a Doctoral Supervisor with the Key Laboratory of China's Ethnic Languages and Information Technology, Ministry of Education, Northwest Minzu University, Lanzhou, China. She has published more than 70 papers in major journals and international conferences in image processing field. Her research interests include image processing, pattern recognition, Tibetan information processing, and computer vision.



WENJIN HU received the B.S. degree in computer science and the Ph.D. degree in control theory and control engineering from the Lanzhou University of Technology, in 2003 and 2015, respectively. She is currently an Associate Professor with the School of Mathematics and Computer Science, Northwest Minzu University, China. Her current research interests include image processing, image quality evaluation, and pattern recognition.