CT-Net

He, Sheng; Schomaker, Lambert

Published in:
Pattern recognition

DOI:
10.1016/j.patcog.2021.108010

Publication date:
2021

Pattern Recognition 118 (2021) 108010

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

CT-Net: Cascade T-shape deep fusion networks for document binarization

Sheng He [a,*], Lambert Schomaker [b]

[a] Department of Radiology, Boston Children's Hospital, Harvard Medical School, Harvard University, Boston MA 02215, USA
[b] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, PO Box 407, Groningen 9700 AK, The Netherlands

ARTICLE INFO

Article history:
Received 14 June 2020
Revised 19 October 2020
Accepted 27 April 2021
Available online 5 May 2021

Keywords:
Cascade T-Net
Binarization
Enhancement
DIBCO
Convolutional neural networks

ABSTRACT

Document binarization is a key step in most document analysis tasks. However, historical-document images usually suffer from various degradations, making this a very challenging processing stage. The performance of document image binarization has improved dramatically in recent years by the use of Convolutional Neural Networks (CNNs). In this paper, a dual-task, T-shaped neural network is proposed that has the main task of binarization and an auxiliary task of image enhancement. The neural network for enhancement learns the degradations in document images and the specific CNN-kernel features can be adapted towards the binarization task in the training process. In addition, the enhancement image can be considered as an improved version of the input image, which can be fed into the network for fine-tuning, making it possible to design a chained-cascade network (CT-Net). Experimental results on document binarization competition datasets (DIBCO datasets) and MCS dataset show that our proposed method outperforms competing state-of-the-art methods in most cases.

1. Introduction

Document binarization, which is a common pre-processing step, aims to classify each pixel in a degraded image into either text or background. Document binarization is a fundamental research topic for document processing approaches including layout analysis [1,2], historical document analysis [3–5] and word spotting [6,7]. The quality of document images and their binarization maps affect the performance of high-level processing tasks, especially for methods which extract features based on binarized images [3]. Therefore, a high-quality and accurate binarization of a degraded image can boost the performance of the final task in the workflow. The problem of binarization attracts interest in the field, leading to nine document image binarization contests (DIBCO) from the year 2009 and several document images with ground-truth are released every year (except 2015).

Document binarization is a challenging problem because most documents suffer from various degradations such as pitch-black or pure-white margins; feeble contrast between ink and parchment; smear; stain; uneven illumination and pen strokes; artifacts and bleed-through [8,9]. Examples are shown in Fig. 1. These degradations affect the results of edge detection, intensity distribution computation and stroke-width estimation, which are the basic steps of traditional binarization methods [10–13]. Therefore, the performance of these traditional binarization methods is limited and none of them can deal with all types of degradations.

Recently, the performance of document binarization has been greatly improved by using the convolutional neural network [8,14], which has a large information capacity and it can correct various degradations learned from the training set. Document binarization is similar to the semantic segmentation problem [15] and the neural networks used for semantic segmentation can also be used for binarization, such as the fully convolutional network [16], hierarchical deep-supervised network [14] inspired by holistic edge detection (HED) [17] and U-Net [18,19].

1.1. Motivation

There are two types of tasks related to document binarization: (1) **Binarization task**. Most methods consider the document image binarization as a single task and train the neural network to directly classify each pixel into either text or background [8,14,20]. (2) **Enhancement task**. The aim of an enhancement task is to train a model to improve the quality of the input image and output an enhanced version which is locally uniform [21,22]. A neural network for enhancement should learn the degradations which are present in the original image, yielding a clear and enhanced image [19]. In this paper, we train a neural network to jointly learn

* Corresponding author.
E-mail addresses: heshengxgd@gmail.com (S. He), L.Schomaker@ai.rug.nl (L. Schomaker).

https://doi.org/10.1016/j.patcog.2021.108010
0031-3203/

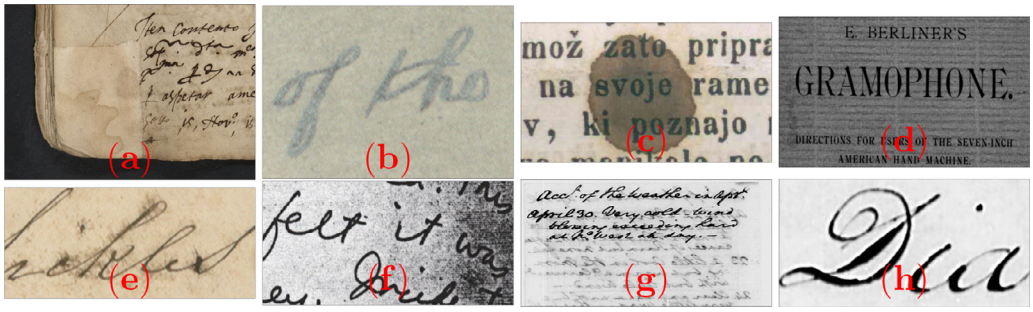


Fig. 1. Examples of typical image degradations from DIBCO datasets (a) black margin (b) feeble contrast between ink and parchment (c) smear (d) textural background (e) smudging of text (f) uneven distribution (g) bleed-through (h) uneven pen strokes.

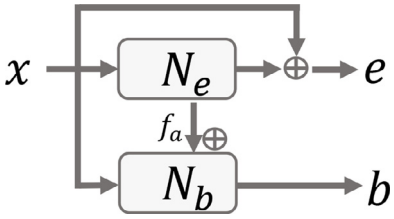


Fig. 2. The proposed framework for document binarization and enhancement. x is the input image, e is the enhanced image and b is the binarized image. N_e and N_b are networks for enhancement and binarization, respectively. f_a is the adaption function which transfers the deep features from N_e to N_b.

the binarization and enhancement task and transfer the learned degradations from the enhancement task to the binarization task to improve the performance. Training the neural network jointly for enhancement and binarization is expected to improve the performance of binarization because the learned features are shared between two tasks and the risk of over-fitting for binarization is reduced. In addition, features learned for enhancement contain the degradation information and adapting these features inspired by the deep-adaptive learning approach [23] to binarization can further improve the performance.

The proposed framework shown in Fig. 2 implicitly decomposes the binarization into two steps, corresponding to two neural networks: N_e is the residual network to learn the enhanced image e and N_b is the network to learn the binarized image b, given the input image x. The enhanced image e can be denoted by: $e = N_e + x$, thus the network N_e learns the degradations (the differences between the degraded and clear images) in the image $N_e = e - x$. We transfer the deep features learned from the network N_e to the binarized network N_b by the add operation, making the network N_b indirectly learning from $x + N_e = x + e - x = e$ which is the enhanced image. The network N_e learns the degradations which are transferred to the binarization network N_b to correct the deep features of the N_b. Thus, unlike the traditional neural network learning the binarization directly from the input image [14,20], our network N_b yields good performance by learning the binarization from the corrected or enhanced (degradation-free) deep features learned by the N_e.

1.2. Method

Although there are different ways to implement the framework shown in Fig. 2, in this paper, we propose a T-shape network (T-Net) which performs these two tasks within one neural network, similar to the multi-task learning problem [24] which aims to model related tasks jointly. The T-Net receives the degraded image as input and outputs two maps: one is the binarization map and another one is the enhancement map. The difference between document enhancement and binarization is that document enhancement aims to improve the perceptual quality by removing or correcting various degradations [19] and the output image of enhancement is an improved version of the input image, which it is not guaranteed to be binarized.

Since the enhanced image can be considered as an improved version of the input image, it can also be used as the input for fine-tuning, making it possible to design a cascade neural network to deal with various types of degradations iteratively. The cascade approach presents a powerful architecture for boosting performance iteratively and has been widely used in various tasks [25–27]. Cascade neural networks usually have multiple stages and the earlier stages can correct the easy degradations whereas the later stages can pay more attention to the hard degradations in the document images [28].

Inspired by the multiple-task and cascade-learning methods, we propose a Cascade T-shape neural network (CT-Net) for document binarization. The CT-Net consists of several T-Nets and each T-Net has three branches, one branch is the encoder which converts the input images into features maps in different levels and the other two branches are two decoders corresponding to the binarization task and the enhancement task. The encoder is shared between two tasks and two different decoders learn specific features for each task. In addition, as inspired by the deep adaptive learning method [23], we also transfer the features learned for enhancement to the features for binarization, named CT_{ada}-Net. As discussed above, the learned features in the decoder for enhancement contain the degradation information of the input image, which makes the decoder of binarization to learn indirectly from the enhanced features and thus improve the performance.

The remainder of this paper is organized as follows. Section 2 covers a background overview of document binarization. In Section 3, we describe our proposed methods in detail. Experimental results are given in Section 4 and conclusions are presented in Section 5.

2. Related work

There are two groups of binarization methods proposed in the literature: traditional methods which are based on image processing techniques, usually at the single pixel-intensity level, and deep learning methods which apply convolutional neural networks to learn an end-to-end model for binarization while exploiting detailed regional image information. Traditional methods usually compute a threshold on each pixel while the trained deep models predict the label of each pixel in a document image.

2.1. Traditional threshold-based methods

The classical threshold-based method for binarization is the Otsu method [29], which computes a global intensity threshold

to minimize the intra-class variance on the whole document images and it provides a good performance if the histogram of the intensity of document images has bimodal distribution. The global threshold method does not work well on document images with high degradations and uneven distributions. Therefore, local threshold methods have been proposed, such as Niblack [30] and Sauvola [10], which compute the threshold on each pixel, based on statistical information in a local patch.

In [31], a local adaptive threshold method is proposed which relies on a background surface estimation for computing the threshold on each pixel. AdOtsu [12] introduces an estimated background map to determine whether to use the Otsu's threshold or not on a local patch because using Otsu directly on small patches is impossible due to the fact that it always separates pixels into two classes, which makes large errors on background patches without any text. T-transition pixels, which is a generalization of edge pixels, are defined and used for binarization in [32] since texts in document images usually have strong edges. The binarization methods proposed in [11,33,34] use the detected edge, combined with an adaptive contrast map computed by using local maximum and minimum [35] and ensemble strategy [36] which is tolerant to background and text variation caused by variant degradations, to determine the threshold on each pixel. The method proposed by Jia et al. [13] uses the structural symmetric pixels to compute the local threshold determined by the voting result of multiple thresholds on each pixel.

Howe [37] defines the binarization problem as an optimization problem based on a global-energy function whose data-fidelity term relies on the Laplacian of the image intensity and the smoothness term is related to edge discontinuities. Later, Howe [38] proposes a automatic parameter tuning method for binarization. Nafchi et al. [39] use phrase congruency-based features maps and Otsu's threshold for binarization. Recently, a game theory inspired binarization is proposed in [40], which uses a two-player, non-zero-sum, non-cooperative game to extract the local information, followed by a K-means method for text classification.

2.2. Deep learning-based methods

With the development of deep learning, convolutional neural networks have also been used for binarization, exploiting regional texture. A fully convolutional network, which is originally proposed for semantic segmentation in natural images [15], is adapted for the binarization task in [16] with a pseudo F-measure loss. Inspired by HED [17], a hierarchical deep-supervised network (DSN) architecture [14] is proposed to learn the labels of each pixel at different feature levels. A primal-dual network (PDNet) [8] which combines an energy minimization function and convolutional network learned features for binarization. In [18], a selectional auto-encoder network is used to predict a selection of values for each pixel's confidence value to determine whether the pixel belongs to text or background. Zhao et al. [20] formulate the binarization as an image-to-image translation task and use the conditional generative adversial networks (cGANs) to combine the multi-scale information for document image binarization.

Unlike the previous works which use a deep neural network to predict the label of each pixel directly, the DeepOtsu [19] method trains a network to output an enhanced image, which corrects the degradations in the input image in a gradual manner. The final binarization map can then be obtained from the enhancement map by applying the global Otsu method.

The proposed method in the current paper is also a deep learning-based method, based on a vanilla version of U-Net. However, the algorithm is modified to address the dual tasks of binarization and enhancement simultaneously, instead of requiring a separate final Otsu stage as in [19]. In addition, the proposed method here decomposes the binarization into two steps: enhancement and binarization. The enhancement step learns the degradations which is transferred to the binarization step, making the binarization based on the degradation-free deep features.

3. The proposed method

This section describes the proposed CT-Net and CT_{ada}-Net neural networks for document binarization and enhancement in a multi-task learning scenario.

3.1. Problem formulation

We used x to denote the input degraded image, b to denote the output binarized map and e to denote the enhancement version of x. Both the binarization and enhancement tasks can be regarded as an image-to-image translation problem, which can be solved by the typical U-Net [20,41].

In this paper, a T-shape neural network is proposed for both enhancement and binarization tasks (shown in Fig. 3), named as T-Net, which is defined as:

$$\begin{aligned} c, \vec{m} &= E(x) \\ e &= D_e(c, \vec{m}) + x \\ b &= D_b(c, \vec{m}) \end{aligned} \tag{1}$$

where E is the encoder which contains five convolutional layers and coverts the input image into feature maps $\vec{m}$ and c, c is the context and contains the semantic information of the input image and $\vec{m} = m_1, \cdots, m_5$ are the intermediate spatial feature maps from the encoder with different resolutions, D_b and D_e are the decoders for binarization and enhancement, respectively. More details can be found in Fig. 3. Our proposed T-Net can be considered as the generalized U-Net in the multi-task scenario.

Compared to the vanilla U-Net [41], our proposed T-Net has another branch D_e for enhancement, which also uses the context information c and feature maps $\vec{m}$ from the encoder branch. In other words, the two tasks share the same encoder E which learns the shared representation but use different decoders D_e and D_b which learn the task specific representation. Note that we added the original input x on the enhancement decoder D_e so it can be rewritten as $D_e(c, \vec{m}) = -(x - e)$, which means that the decoder D_e learns the negative degradations $d = -(x - e)$ which are the differences between the clear and enhanced image (e) and degraded image (x), similar to DeepOtsu [19]. As discussed in [23,42], sharing parameters among multi-tasks can greatly reduce the risk of overfitting on a specific task.

In T-Net, the decoders D_b and D_e learn specific features for each task without any interaction. However, the information learned by D_e for enhancement should be also helpful for binarization since the task of enhancement is to remove the degradations and obtain a clear and uniform image. In order to allow the network to learn the relationship between these two tasks, we also applied the shortcuts from the decoder D_e to the decoder D_b to improve the information flow, inspired by the multi-task learning [23,24]. The T-Net with the task adaption, named T_{ada}-Net, can be denoted as:

$$\begin{aligned} c, \vec{m} &= E(x) \\ e, \vec{n} &= D_e(c, \vec{m}) + x \\ b &= D_b(c, \vec{m} + \vec{n}) \end{aligned} \tag{2}$$

where $\vec{n}$ represents the spatial feature maps generated by the decoder D_e during recovering the enhancement image e. We added the intermediate spatial feature maps $\vec{m}$ and $\vec{n}$ in the decoder D_b since $\vec{n}$ is from the decoder D_e which contains the negative degradation information ($D_e(c, \vec{m}) = -(x - e)$). Thus, $\vec{m} + \vec{n}$ can remove

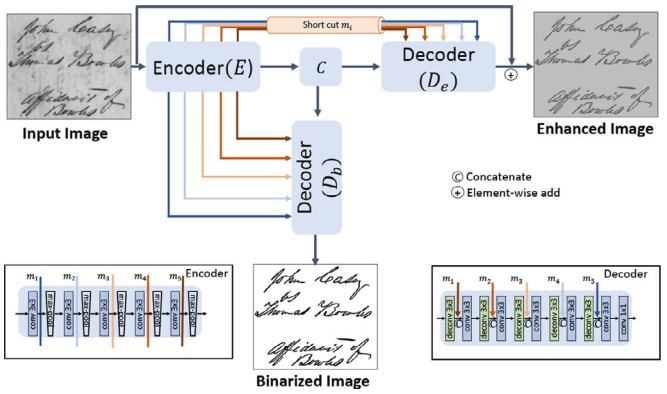


Fig. 3. The overall architecture of the proposed T-Net. It has three branches, encoder E and decoder D_e for enhancement task and D_b for binarization task. Short cuts are applied between encoder and decoders. In the encoder, we use max-pooling for downsampling and in decoder we use deconvolutional operations for upsampling.

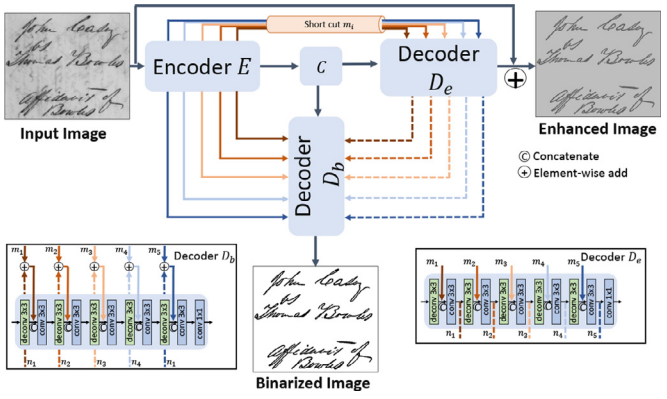


Fig. 4. The overall architecture of the proposed T_{ada}-Net. The structure of the encoder E is same to the one in Fig. 3. The dashed lines denote the short cuts from the decoder D_e to D_b.

the learned degradations from the input feature map $\vec{m}$, improving the performance of binarization. Fig. 4 shows the structure of T_{ada}-Net.

As mentioned in [19], the enhancement image e is the latent uniform and clear version of the input image and recovering it allows the network to learn the degradations and binarization iteratively. More precisely, the enhancement map e may still contain some slight degradations and it can be fed into the network for further processing, which makes it possible to design a successful architecture for binarization in a cascade way. In this paper, we extended the T-Net into cascade T-Net, named CT-Net, which is defined as:

$$e^{i+1}, b^{i+1} = T_i(e^i) \quad e^0 = x \tag{3}$$

where $T_i(x)$ can be either T-Net (CT-Net) or T_{ada}-Net (CT_{ada}-Net), e^i is the i-th enhancement map of T-Net-i and $e^0 = x$ is the original input image. Fig. 5 shows the framework of the proposed CT-Net.

3.2. Network configuration

Inspired by the U-Net structure [41], the encoder of the CT-Net contains five convolutional layers with channel size [32,64,128,256,512], followed by batch normalization, ReLU, and max-pooling (2×2, stride:2) layers. For decoders, the deconvolutional layers are used for upsampling feature maps which is concatenated with the corresponding spatial feature maps in $\vec{m}$ from the encoder branch, followed by a convolutional layer for feature combination. The size of all kernels used in convolutional and deconvolutional layers is 3×3, except the last layer which uses 1×1 kernel to compute the output map. The input color image size is set to $h \times w \times 3$, where h and w are height and width. The output layer of the enhancement has the same size as input. For binarization output, the size is $h \times w \times 1$, in which the value on each pixel denotes the probability computed by a sigmoid function.

The loss of the enhancement is defined as:

$$L_e = \frac{1}{n} \sum |e_t - e| \tag{4}$$

where e_t is the ground-truth of the enhancement image, e is the output of the network and n is the total number of pixels in the image. The loss of the binarization is the typical cross-entropy loss, given by:

$$L_b = -\frac{1}{n} \sum b_t \cdot log(p) + (1 - b_t) \cdot log(1 - p) \tag{5}$$

where b_t is the ground-truth of the binarization map and p is the probability from the neural network. In fact, each T-Net network in CT-Net can be trained separately or jointly. In this paper, we trained the network jointly and the final training loss is given by:

$$L_{train} = \sum_{i=1}^{N=3} \left(L_e^i + L_b^i\right) \tag{6}$$

where L_e^i and L_b^i are the enhancement and binarization losses in the i-th T-Net and $N = 3$ is the total number of the T-Net.

Table 1
The number of documents on each DIBCO dataset.

Data set	Handwritten	Printed	Total
DIBCO'09 [43]	5	5	10
H-DIBCO'10 [44]	10	-	10
DIBCO'11 [45]	8	8	16
H-DIBCO'12 [46]	14	-	14
DIBCO'13 [47]	8	8	16
H-DIBCO'14 [48]	10	-	10
H-DIBCO'16 [49]	10	-	10
DIBCO'17 [50]	20	-	20
H-DIBCO'18 [51]	10	-	10
Total	95	21	116

4. Experimental results

4.1. Datasets and evaluation metrics

We evaluated the proposed method on nine DIBCO benchmark datasets from the document binarization competitions, including DIBCO'09 [43], H-DIBCO'10 [44], DIBCO'11 [45], H-DIBCO'12 [46], DIBCO'13 [47], H-DIBCO'14 [48], H-DIBCO'16 [49], DIBCO'17 [50], H-DIBCO'18 [51]. The number of documents on each dataset is shown in Table 1. We used the leave-one-dataset-out strategy inspired by [8,20] for evaluation. More precisely, when evaluating on a particular DIBCO dataset, all rest of datasets are used for training. The training set also includes images from the Bickly-diary dataset [52], Persian Heritage Image Binarization dataset (PHIDB) [53] and Synchromedia Multispectral dataset [54], which were also used for training the networks [14,19].

Four metrics which were widely used in the contests [49–51] were used to assess the performance and quantitatively compared with state-of-the-art methods, including the F-measure (FM), pseudo F-measure (F_{ps}), peak signal-to-noise ratio (PSNR) and distance reciprocal distortion metric (DRD). For fair comparison, we used the evaluation tool provided by the DIBCO competitions to compute these metric scores in this paper.

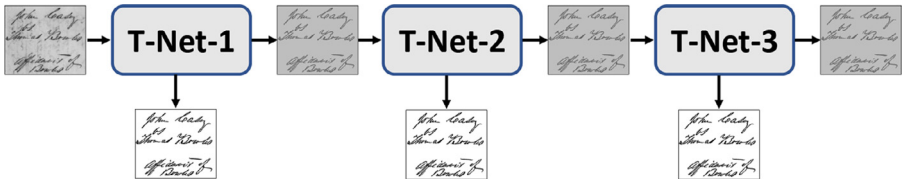


Fig. 5. The overall architecture of the proposed CT-Net with three T-Nets. The output of the enhancement map on T-Net-1 can be considered as the improved version of the original image which can be fed into T-Net-2 for fine-tune. This is the same for T-Net-3. T-Net-*n* are identical T-Nets with different parameters.

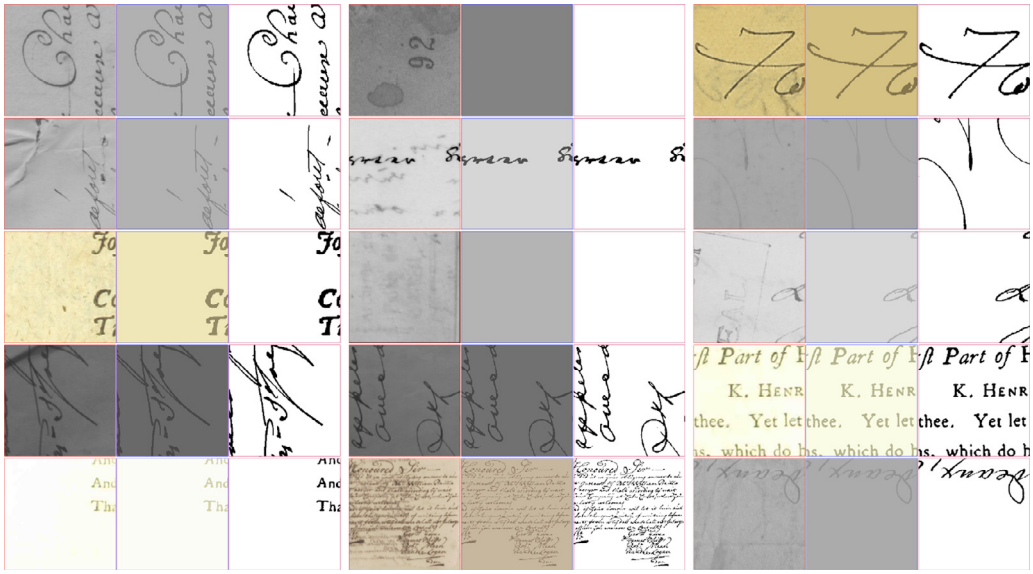


Fig. 6. Training samples (red box left column) with their corresponding enhancement ground-truth (middle column) and binarization ground-truth (right column). All patches have the same spatial size of 256 × 256. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Implementation details

Ideally, the entire image instead of patches can be processed by the proposed CT-Net. However, due to the limitation of GPU memory it is very hard to work with arbitrary-sized inputs. The most popular way is to crop small patches and resize them to a fixed width and height [18,20]. The most common size is 256×256, which is also adopted in the proposed CT-Net. For each image on DIBCO datasets, we cropped image patches with 256×256 with 10% overlap horizontally and vertically. For scale augmentation, we also sampled patches with the scale factor {0.75, 1.25.1.5} based on the input size of the network. In addition, the global patches whose sizes are equal to the minimum of the entire image width and height are cropped and resized to 256×256 for training and testing, inspired by Vo et al. [14]. The network is initialized by the Xavier method [55] and is trained by the Adam optimizer [56] with an initial learning rate 0.0001, which is decreased to half at every 50,000 iterations. The number of training iterations is 200,000 with the batch size 10 due to the limitation of the memory.

The ground-truth of the binarization is provided in DIBCO datasets while the ground-truth of the enhancement is built by using the same method in [19]. The value of each pixel on the enhancement ground-truth image is computed as the average pixel value with the same label within the patch. The label of each pixel can be obtained from the ground-truth of the binarization. Fig. 6 shows several training samples used in this paper for training the neural networks.

During testing, the patches are cropped from the test images with the same method used to crop patches for training. The predicted binarized maps from the trained networks are stitched to the original document size by averaging values over the overlapping patches. Therefore, the value of each pixel on the predicted maps is the probability in [0,1]. The final binarization map is computed with a global threshold T based on the predicted probability map. Unlike the work in [14] which uses a fix T learned from a separated training set for all documents, we used Otsu [29] to compute the threshold T for each document. It is very efficient to compute the Otsu threshold and each image has its own threshold depending on the quality of predicted probability maps from the network.

4.3. Performance evaluation

In this paper, we stacked three T-Nets due to the trade-off between performance and model complexity. It has been shown in DeepOtsu [19] that the performance can be improved by cascading at least three basic components (T-Net) in the network. We evaluated the performance of binarization results from each T-Net and T_{ada}-Net in the cascaded CT-Net and CT_{ada}-Net. Fig. 7 shows the average scores of F-measure and DRD overall 116 binarization maps with different n of CT-Net-n and CT_{ada}-Net-n. From the figure we can see that the performance in terms of F-measure and DRD increases as stacking more T-Net in the CT-Net. Practically, we also found the similar trends in terms of F_{ps} and PNSR. Thus, we can conclude that cascading different T-Nets can iteratively improve the performance. In addition, the CT_{ada}-Net provides better performance than CT-Net, which demonstrates the powerful of the adaptive shortcuts from the enhancement task to the binarization task.

We roughly divided all documents in DIBCO datasets into eight groups, according to main degradations contained in document images as shown in Fig. 1. The average F-Measure scores of the proposed CT-Net-3 and CT_{ada}-Net-3 is shown in Fig. 8, from which we

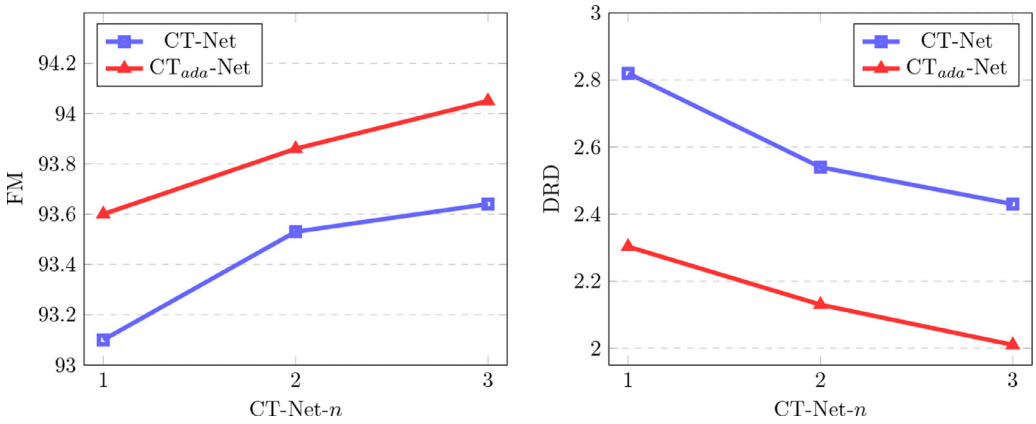


Fig. 7. The performance of the n-th T-Net and T_{ada}-Net in the cascaded network CT-Net and CT_{ada}-Net, respectively. The left figure shows the average F-measure (high values means good performance) and the right figure shows the average DRD scores (low values means good performance).

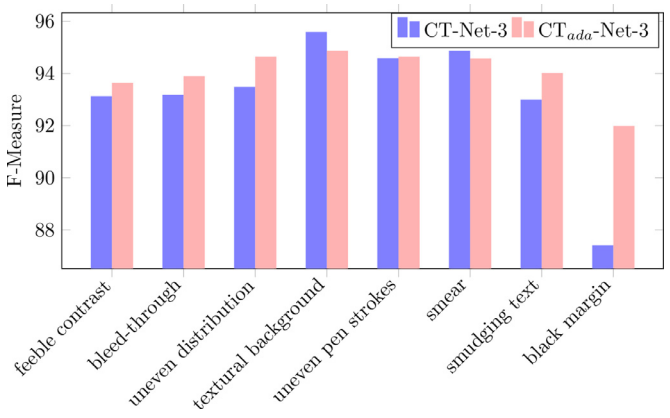


Fig. 8. The performance of the CT-Net-3 and CT_{ada}-Net-3 on images with different degradations.

can see that the performance of CT_{ada}-Net-3 outperforms CT-Net-3 in documents with most degradations except textural background and smear. Note that the CT_{ada}-Net-3 provides much better results than CT-Net-3 on documents with the degradation of black margins.

4.4. Comparison with winner methods submitted to DIBCO datasets

In this section, we compared the best score per metric for any submission on each year and the proposed CT-Net-3 or CT_{ada}-Net-3 methods. Table 2 shows the comparison between the proposed method and the winner among competition submissions on each DIBCO dataset. From the table we can see that our proposed methods outperform the winner methods for all of nine data sets in terms of all the evaluation metrics. The average scores of F-measure over all nine datasets are 91.33 and 94.28 for the winner and proposed CT-Net respectively. The results show that our proposed CT-Net achieved, on average, a 3.2% improvement over the winner results of DIBCO datasets.

4.5. Comparison with state-of-the-art methods

In this section, the proposed methods are qualitatively evaluated on the nine DIBCO datasets and compared with traditional binarization methods including Otsu [29], Sauvola [10], Su et al. [11], Howe [38], Lelore et al. [57], Jia et al. [13], Mitianoudis et al. [58], GiB [40] and deep learning-based methods including Hierarchical

Table 2
Comparison of the proposed methods with (H)DIBCO competition winners.

Database	Method	FM	F_{ps}	PSNR	DRD
DIBCO'09	Best Competition System	91.24	-	18.66	-
	Best of CT-Net	94.18	95.80	20.50	2.56
	Performance gains	2.94	-	1.84	-
H-DIBCO'10	Best Competition System	91.50	-	19.78	-
	Best of CT-Net	93.93	97.20	21.21	1.55
	Performance gains	2.43	-	1.34	-
DIBCO'11	Best Competition System	88.74	-	17.97	5.36
	Best of CT-Net	95.27	97.24	21.50	1.37
	Performance gains	6.53	-	3.53	3.99
H-DIBCO'12	Best Competition System	92.85	-	21.80	2.66
	Best of CT-Net	95.82	96.67	22.92	1.32
	Performance gains	2.97	-	1.12	1.34
DIBCO'13	Best Competition System	92.70	94.19	21.29	3.10
	Best of CT-Net	95.66	97.79	22.98	1.32
	Performance gains	2.96	3.60	1.69	1.78
H-DIBCO'14	Best Competition System	96.88	97.65	22.66	0.90
	Best of CT-Net	97.70	98.74	23.92	0.65
	Performance gains	0.82	1.09	1.26	0.25
H-DIBCO'16	Best Competition System	88.72	91.84	18.45	3.86
	Best of CT-Net	91.07	94.34	19.22	3.29
	Performance gains	2.35	2.50	0.77	0.57
DIBCO'17	Best Competition System	91.04	92.86	18.28	3.40
	Best of CT-Net	92.72	94.73	19.17	2.65
	Performance gains	1.68	1.87	0.89	0.75
H-DIBCO'18	Best Competition System	88.34	90.24	19.11	4.92
	Best of CT-Net	92.23	94.97	20.13	2.70
	Performance gains	3.89	4.73	1.02	2.22

DSN [14], PDNet [8], cGANs [20]. Tables 3, 4, 5 show the performance of the proposed method and the compared algorithms. Figs. 9-11 display several visual quality of binary results, highlighting the advantages and disadvantages of the proposed methods.

4.5.1. DIBCO'09 dataset

The dataset contains ten documents with various degradations such as bleed-through, smear or nonuniform background. From Table 3 we can see that our proposed CT_{ada}-Net-3 provides the best scores of F-measure, F_{ps} and PSNR. The cGANs [20] method which uses the cascaded conditional generative adversarial networks to generate the binarization image gives lowest visual distortion (best DRD score). Fig. 9 shows several examples of the proposed binary maps, revealing that our proposed methods can han-

Table 3
Performance comparison of different binarization methods on each DIBCO data set (best values are highlighted in **bold** and second best are highlighted in *italic*).

Method	DIBCO'09				H-DIBCO'10				DIBCO'11			
	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD
Otsu [29]	78.60	80.53	15.31	22.57	85.43	90.64	17.52	4.05	82.10	85.96	15.72	8.95
*Sauvola [10]	85.37	89.08	16.37	7.08	75.18	84.08	15.94	7.22	82.14	87.70	15.65	8.50
*Su et al. [11]	93.02	94.61	19.41	2.64	91.36	93.18	19.78	2.42	87.83	90.24	17.71	4.66
Howe [38]	94.04	95.06	*20.43*	*2.10*	93.59	94.81	21.08	1.72	90.79	92.28	19.01	4.46
*Lelore et al. [57]	93.93	95.10	20.21	2.17	93.82	94.27	21.09	1.79	92.48	94.11	19.37	2.97
Nafchi et al. [39]	93.36	-	19.55	-	91.78	95.08	19.80	-	92.57	-	19.29	2.28
Mitianoudis et al. [58]	90.27	92.69	18.08	3.71	88.97	91.16	18.32	3.38	89.13	93.79	17.90	3.47
Jia et al. [13]	93.05	94.60	19.29	2.40	89.46	93.94	18.86	2.93	91.92	95.09	18.98	2.64
GiB [40]	92.50	-	19.26	2.41	90.00	-	19.14	2.75	90.33	-	18.29	2.99
♠ DeepOtsu [19]	-	-	-	-	-	-	-	-	93.4	95.8	19.9	1.9
♠ Hierarchical DSN [14]	-	-	-	-	-	-	-	-	93.3	96.4	20.1	2.0
♠ *PDNet [8]	91.50	-	19.25	3.06	92.91	-	20.40	1.85	91.87	-	19.07	2.57
♠ cGANs [20]	*94.10*	*95.26*	20.30	**1.82**	**94.03**	95.39	*21.12*	*1.58*	93.81	95.70	20.26	1.81
♠ Proposed CT-Net-3	92.08	94.31	19.77	3.58	93.49	**97.20**	20.98	1.68	**95.27**	**97.24**	**21.50**	**1.37**
♠ Proposed CT_{ada}-Net-3	**94.18**	**95.80**	**20.50**	2.56	*93.93*	*96.74*	**21.21**	**1.55**	*94.17*	*96.92*	*20.76*	*1.69*

Note: * results are from Jia et al. [13] if possible. ♠ deep learning method. ♠* the best score per metric in [8].

Table 4
Performance comparison of different binarization methods on each DIBCO data set (best values are highlighted in **bold** and second best are highlighted in *italic*).

Method	H-DIBCO'12				DIBCO'13				H-DIBCO'14			
	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD
Otsu [29]	75.07	78.14	15.03	26.46	80.04	83.43	16.63	10.98	91.62	95.69	18.72	2.65
Sauvola [10]	81.56	87.35	16.88	6.46	82.71	87.74	17.02	7.64	84.70	87.88	17.81	4.77
Su et al. [11]	89.76	89.61	19.55	4.19	87.70	88.15	19.59	4.21	94.38	95.94	20.31	1.95
Howe [38]	93.73	94.24	21.85	2.10	91.34	91.79	21.29	3.18	96.49	97.38	22.24	1.08
Lelore et al. [57]	94.05	94.42	21.43	2.11	90.78	91.47	20.54	3.59	96.14	96.73	21.88	1.25
Jia et al. [13]	92.99	95.10	20.37	2.34	93.42	96.05	20.78	2.03	94.98	97.18	20.56	1.50
Mitianoudis et al. [58]	89.71	92.24	18.73	3.88	91.41	95.47	19.54	2.78	87.57	-	18.43	-
GiB [40]	90.99	-	19.34	3.09	91.14	-	19.58	2.77	94.00	-	19.93	1.79
♠ DeepOtsu [19]	-	-	-	-	-	-	-	-	95.9	97.2	22.1	0.9
♠ Hierarchical DSN [14]	-	-	-	-	94.4	96.0	21.4	1.8	96.66	97.59	*23.23*	*0.79*
♠ *PDNet [8]	93.04	-	20.50	2.92	93.97	-	21.30	1.83	89.99	-	20.52	7.42
♠ cGANs [20]	94.96	96.15	21.91	1.55	95.28	96.47	22.23	*1.39*	96.41	97.55	22.12	1.07
♠ Proposed CT-Net-3	*95.38*	**96.67**	*22.39*	*1.52*	**95.66**	**97.79**	**22.98**	**1.32**	**97.70**	**98.74**	**23.92**	**0.65**
♠ Proposed CT_{ada}-Net-3	**95.82**	*96.58*	**22.92**	**1.32**	*95.34*	*97.53*	*22.41*	1.57	*96.91*	*97.93*	22.62	0.88

Note: * results are from Jia et al. [13] if possible. ♠ deep learning method. ♠* the best score per metric in [8].

Table 5
Performance comparison of different binarization methods on each DIBCO data set (best values are highlighted in ***italic* bold** and second best are highlighted in *italic*).

Method	H-DIBCO'16				DIBCO'17				H-DIBCO'18			
	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD	FM	F_{ps}	PSNR	DRD
Otsu [29]	86.59	89.92	17.79	5.58	77.73	77.89	13.85	15.54	51.45	53.05	9.74	59.07
Sauvola [10]	84.64	88.39	17.09	6.27	77.11	84.10	14.25	8.85	67.81	74.08	13.78	17.69
Su et al. [11]	84.75	88.94	17.64	5.64	-	-	-	-	-	-	-	-
Howe [38]	87.47	92.28	18.05	5.35	90.10	90.95	*18.52*	5.12	80.84	82.85	16.67	11.96
Lelore et al. [57]	87.21	88.48	17.36	5.27	-	-	-	-	-	-	-	-
Jia et al. [13]	90.48	93.27	19.30	3.97	85.59	86.38	16.39	7.99	76.52	79.90	17.00	8.11
Mitianoudis et al. [58]	86.89	-	17.60	-	-	-	-	-	-	-	-	-
GiB [40]	91.15	-	19.18	3.20	-	-	-	-	-	-	-	-
♠ DeepOtsu [19]	*91.4*	94.3	*19.6*	*2.9*	-	-	-	-	-	-	-	-
♠ Hierarchical DSN [14]	90.10	93.57	19.01	3.58	-	-	-	-	-	-	-	-
♠ *PDNet [8]	90.18	-	18.99	3.61	-	-	-	-	-	-	-	-
♠ cGANs [20]	**91.66**	**94.58**	**19.64**	**2.82**	90.73	92.58	17.83	3.58	87.73	90.60	18.37	*4.58*
♠ Proposed CT-Net-3	89.62	91.60	18.63	4.70	**92.72**	*94.31*	**19.17**	*2.79*	*88.90*	*91.45*	*18.84*	5.58
♠ Proposed CT_{ada}-Net-3	91.07	*94.34*	19.22	3.29	*92.65*	**94.73**	**19.17**	**2.65**	**92.23**	**94.97**	**20.13**	**2.70**

Note: * results are from Jia et al. [13] if possible. ♠ deep learning method. ♠* the best score per metric in [8].

dle the various degradations such as bleed-through and feeble contrast between ink and parchment. However, Fig. 9 also shows a problematic sample involving large character sizes with fat strokes in a deviant ink color. The main reason is that there was no similar training material in the training dataset.

4.5.2. H-DIBCO'10 dataset

Document images in this dataset are from historical collections and ambiguity exists in the text boundary locations. It is hard, even for humans, to label pixels near the boundary [9]. Therefore, most errors indeed are from pixels near text boundaries and weak ink traces. The best scores on this data set, for different metrics, are

Fig. 9. Binarization results on DIBCO'09. Left column shows the original images, the middle column shows the results of CT-Net-3 and the right column shows the results of CT_{ada}-Net-3. Black pixels are correctly classified text pixels and white pixels are correctly recognized background. Text pixels classified as background are marked in red while background pixels classified as text are show in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

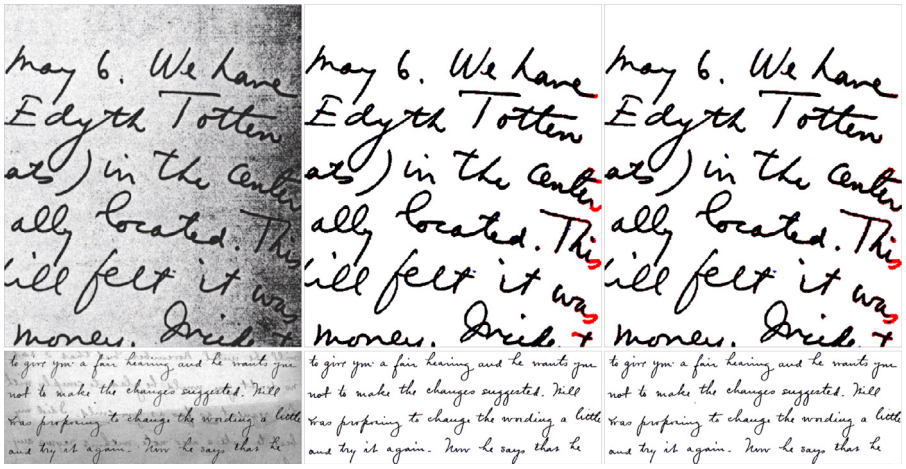


Fig. 10. Binarization results on DIBCO'11 and H-DIBCO'12. Left column shows the original images, the middle column shows the results of CT-Net-3 and the right column shows the results of CT_{ada}-Net-3. Black pixels are correctly classified text pixels and white pixels are correctly recognized background. Text pixels classified as background are marked in red while background pixels classified as text are show in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from different methods. The highest F-measure score is given by sGANs [20] while the proposed CT-Net-3 results in the best F_{ps}, which has a better correlation with OCR [9]. The best performance in terms of PSNR and DRD is given by CT_{ada}-Net-3 which indicates that our proposed CT_{ada}-Net-3 method gives the lowest visual distortion.

4.5.3. DIBCO'11 and H-DIBCO'12 datasets

Both of our proposed CT-Net-3 and CT_{ada}-Net-3 provide the better results than other state-of-the-art methods on all four metric scores on these two data sets. Fig. 10 shows two images which contain variation of contrast. The binarization maps of our proposed methods have a clear background but lose some strokes on the very noisy regions.

Most document images in DIBCO'11 dataset have a strong textural background. Thus, as discussed above, the CT-Net-3 provides better results than CT_{ada}-Net-3. However, document images in the H-DIBCO'12 dataset suffer from uneven distribution and weak ink traces, which can be solved efficiently by CT_{ada}-Net-3.

4.5.4. DIBCO'13 and H-DIBCO'14 datasets

Document images in DIBCO'13 contains large smear or show-through degradations. Although our proposed methods provide the best results over other methods, some strokes inside the smear are missed and noisy patterns from smear are retained in the final bi-

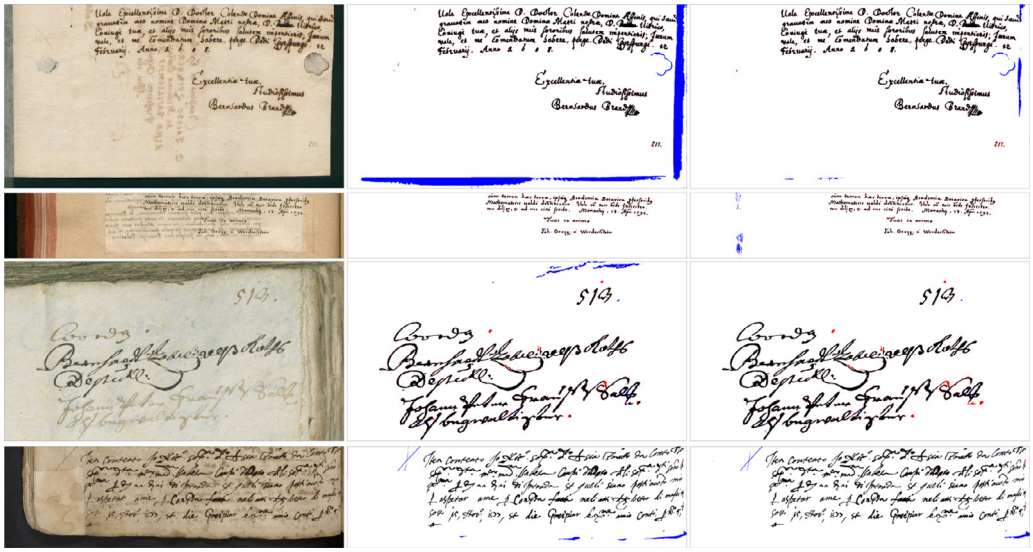


Fig. 11. Binarization results on H-DIBCO'18. Left column shows the original images, the middle column shows the results of CT-Net-3 and the right column shows the results of CT_{ada}-Net-3. Black pixels are correctly classified text pixels and white pixels are correctly recognized background. Text pixels classified as background are marked in red while background pixels classified as text are show in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

narization map. In these two datasets, the CT-Net-3 provides better performance than CT_{ada}-Net-3 since the document images suffer from the degradations of smear or weak ink traces which are hard to be corrected in the enhancement images, thus noises are introduced through the adaptive shortcuts in CT_{ada}-Net.

4.5.5. H-DIBCO'16 and DIBCO'17 datasets

Our proposed method provides the second best performance in terms of F_{ps} on the H-DIBCO'16 data set, which is lower than sGANs [20]. H-DIBCO'16 is a challenging dataset for binarization, compared to other DIBCO datasets which contains handwritten documents with various degradations such as bleed-through, uneven pen strokes and background.

On the DIBCO'17 dataset, our proposed methods provide the best results than other methods. Document images in this dataset have a large image size and most of them suffer from the bleed-through degradations.

4.5.6. H-DIBCO'18 dataset

The special degradation on H-DIBCO'18 is dark margins near the page boundaries. The proposed CT_{ada}-Net-3 provides best results than traditional methods, such as Jia's [13], Howe's [38] and the deep learning-based cGANs [20]. Fig. 11 shows several examples, indicating that our proposed CT_{ada}-Net-3 can deal with the dark margin and uneven distribution of text pixels in document images.

4.5.7. Summary over nine DIBCO datasets

Table 6 presents the mean values of the four evaluation metrics over nine DIBCO datasets. We compared our methods with five traditional methods and two deep learning-based methods. From the table we can see that the proposed CT_{ada}-Net-3 provides the best scores of four metrics. The CT-Net-3 without adaption achieves second best results in terms of F-measure and PSNR. Fig. 12 shows the visual results of one document on the H-DIBCO'16 dataset from different methods. From the figure we can see that our proposed methods provide good results on weak ink traces.

4.6. Generalization to The MCS dataset

In this section, we evaluated the generalization of the proposed methods on the Monk Cuper Set (MCS) dataset, which was firstly introduced in our previous work [19] and images in this dataset were taken using iPhone. It contains 25 pages sampled from a historical collection with heavy bleed-through degradations and textural background. This data set is public available on the author's website[1].

Table 6
Comparison of the mean values for different methods over all nine DIBCO data sets. The best performance is shown in **bold** and the second-best is shown in *italic*.

Method	FM	F_{ps}	PSNR	DRD
1st rank of contest	89.83	-	19.79	-
*Otsu [29]	82.60	86.17	16.62	11.35
Sauvola [10]	82.33	87.46	16.68	6.85
Jia et al. [13]	92.68	95.15	19.87	2.44
Howe [38]	92.53	93.98	20.06	2.86
Mitianoudis [58]	89.14	-	16.90	-
♠ Tensmeyer [16]	92.20	*95.71*	20.16	2.74
♠ cGANs [20]	93.19	94.92	20.42	*2.24*
♠ Proposed CT-Net-3	*93.47*	95.47	*20.90*	2.57
♠ Proposed CT_{ada}-Net-3	**94.03**	**96.17**	**20.99**	**2.00**

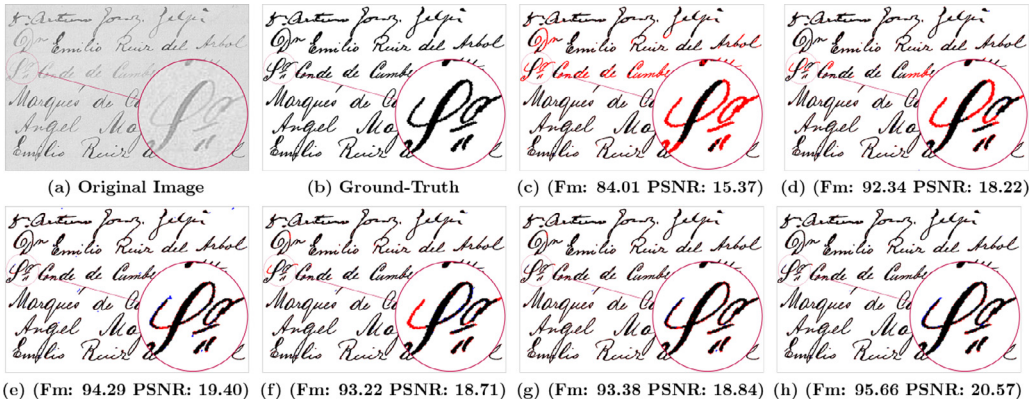
Note: * results are from Zhao et al. [20]. ♠ deep learning method.

Table 7 shows the average results overall 25 document images from nine models trained using the DIBCO data sets. Our proposed CT-Net achieves better results than other methods, which demonstrates that the proposed methods generalize very well on other degraded historical documents.

4.7. Computing time analysis

The number of parameters of the proposed T-Net is the 1.5 times that of the standard U-Net [41] since our T-Net shares the same encoder and has different decoders for different tasks. The networks were trained using a single GPU card (16 GB memory), taking about 24 h. For testing, the time of inference of each patch is less one second and patches from different images can be processing in parallel with multiple GPUs.

[1] https://www.ai.rug.nl/~sheng/



(a) Original Image (b) Ground-Truth (c) (Fm: 84.01 PSNR: 15.37) (d) (Fm: 92.34 PSNR: 18.22)

(e) (Fm: 94.29 PSNR: 19.40) (f) (Fm: 93.22 PSNR: 18.71) (g) (Fm: 93.38 PSNR: 18.84) (h) (Fm: 95.66 PSNR: 20.57)

Fig. 12. Visual example with the F-measure (**Fm**) and **PNSR** metric values of the binarization results of one document applied to H-DIBCO'16 and produced by different methods: (a) original image, (b) ground truth, (c) Su et al. [11], (d) Howe [38], (e) Jia et al. [13], (f) cGANs [20], (g) CT-Net-3, (h) CT_{ada}-Net-3. (Red pixel means text pixel recognized to background while blue pixel means background pixel recognized to text). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Comparison of the mean values for different methods on the MCS data set. The best performance is highlighted in **bold**.

Method	FM	F_{ps}	PSNR	DRD
Otsu [29]	69.3	70.5	11.8	34.0
Sauvola [10]	75.8	76.9	13.1	21.5
Su et al. [11]	82.8	87.4	15.2	16.8
Jia et al. [13]	85.4	88.7	15.8	7.1
Howe [38]	85.6	89.1	15.8	6.4
Deep-Sauvola [19]	87.0	89.9	16.2	6.1
CT-Net-1	88.3	92.2	16.7	4.8
CT-Net-2	88.7	92.8	16.9	4.4
CT-Net-3	88.7	92.7	16.9	4.4
CT_{ada}-Net-1	88.8	92.9	16.9	4.4
CT_{ada}-Net-2	88.9	93.0	16.9	4.3
CT_{ada}-Net-3	**89.1**	**93.2**	**17.0**	**4.2**

Note: * results are from He and Schomaker [19].

5. Conclusion

In this paper we have proposed a cascade T-shape network architecture for document binarization, which ha two versions: CT-Net and CT_{ada}-Net. The T-Net has two outputs corresponding to a binarization map and an enhancement image, constituting a multi-task learning framework. The CT-Net model is a network with three cascade T-Nets, and CT_{ada}-Net adapts the features learned for enhancement to the binarization task, which can improve the performance of binarization.

We have evaluated the proposed methods over nine widely-used competition DIBCO datasets and our proposed methods outperform all winner methods submitted to the competitions. In addition, our methods are superior to traditional and deep learning methods for binarization reported in the literature. An external historical document dataset is used to evaluate the generalization and the results show that our proposed methods generalize very well on these document images.

Future works include exploiting more complex structure of network and training different networks on documents with different degradations. Image enhancement, including traditional image operations and the trainable generative adversarial network, appears to be very important for boosting the performance of document binarization. It would be interesting to evaluate the effect of image enhancement on text recognition, in comparison to plain document binarization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Thanks the organizers of the DIBCO competitions. In addition, the authors would like to thank Zhenwei Shi to label the MCS dataset.

References

[1] K. Chen, H. Wei, J. Hennebert, R. Ingold, M. Liwicki, Page segmentation for historical handwritten document images using color and texture features, in: International Conference on Frontiers in Handwriting Recognition, 2014, pp. 488–493.
[2] G.M. Binmakhashen, S.A. Mahmoud, Document layout analysis: a comprehensive survey, ACM Comput. Surv. (CSUR) 52 (6) (2019) 1–36.
[3] S. He, P. Samara, J. Burgers, L. Schomaker, A multiple-label guided clustering algorithm for historical document dating and localization, IEEE Trans. Image Process. 25 (11) (2016) 5252–5265.
[4] J.A. Sánchez, V. Romero, A.H. Toselli, M. Villegas, E. Vidal, A set of benchmarks for handwritten text recognition on historical documents, Pattern Recognit. 94 (2019) 122–134.
[5] W. Sihang, W. Jiapeng, M. Weihong, J. Lianwen, Precise detection of chinese characters in historical documents with deep reinforcement learning, Pattern Recognit. 107 (2020) 107503.
[6] M. Stauffer, A. Fischer, K. Riesen, Keyword spotting in historical handwritten documents based on graph matching, Pattern Recognit. 81 (2018) 240–253.
[7] J.I. Toledo, M. Carbonell, A. Fornés, J. Lladós, Information extraction from historical handwritten document images with a context-aware neural model, Pattern Recognit. 86 (2019) 27–36.
[8] K.R. Ayyalasomayajula, F. Malmberg, A. Brun, PDNet: semantic segmentation integrated with a primal-dual network for document binarization, Pattern Recognit. Lett. 121 (2019) 52–60.
[9] K. Ntirogiannis, B. Gatos, I. Pratikakis, Performance evaluation methodology for historical document image binarization, IEEE Trans. Image Process. 22 (2) (2012) 595–609.
[10] J. Sauvola, M. Pietikäinen, Adaptive document image binarization, Pattern Recognit. 33 (2) (2000) 225–236.
[11] B. Su, S. Lu, C.L. Tan, Robust document image binarization technique for degraded document images, IEEE Trans. Image Process. 22 (4) (2012) 1408–1417.
[12] R.F. Moghaddam, M. Cheriet, AdOtsu: an adaptive and parameterless generalization of Otsu's method for document image binarization, Pattern Recognit. 45 (6) (2012) 2419–2431.
[13] F. Jia, C. Shi, K. He, C. Wang, B. Xiao, Degraded document image binarization using structural symmetry of strokes, Pattern Recognit. 74 (2018) 225–240.
[14] Q.N. Vo, S.H. Kim, H.J. Yang, G. Lee, Binarization of degraded document images based on hierarchical deep supervised network, Pattern Recognit. 74 (2018) 568–586.
[15] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[16] C. Tensmeyer, T. Martinez, Document image binarization with fully convolutional neural networks, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, IEEE, 2017, pp. 99–104.
[17] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[18] J. Calvo-Zaragoza, A.-J. Gallego, A selectional auto-encoder approach for document image binarization, Pattern Recognit. 86 (2019) 37–47.
[19] S. He, L. Schomaker, DeepOtsu: document enhancement and binarization using iterative deep learning, Pattern Recognit. 91 (2019) 379–390.
[20] J. Zhao, C. Shi, F. Jia, Y. Wang, B. Xiao, Document image binarization with cascaded generators of conditional generative adversarial networks, Pattern Recognit. (2019) 106968.
[21] R.F. Moghaddam, M. Cheriet, Beyond pixels and regions: a non-local patch means (NLPM) method for content-level restoration, enhancement, and reconstruction of degraded document images, Pattern Recognit. 44 (2) (2011) 363–374.
[22] K. Zagoris, I. Pratikakis, Bio-inspired modeling for the enhancement of historical handwritten documents, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, IEEE, 2017, pp. 287–292.
[23] S. He, L. Schomaker, Deep adaptive learning for writer identification based on single handwritten word images, Pattern Recognit. 88 (2019) 64–74.
[24] I. Misra, A. Shrivastava, A. Gupta, M. Hebert, Cross-stitch networks for multi–task learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3994–4003.
[25] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.
[26] B. Yu, D. Tao, Anchor cascade for efficient face detection, IEEE Trans. Image Process. 28 (5) (2018) 2490–2501.
[27] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al., Hybrid task cascade for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4974–4983.
[28] X. Li, Z. Liu, P. Luo, C. Change Loy, X. Tang, Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3193–3202.
[29] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[30] W. Niblack, An Introduction to Digital Image Processing, Strandberg Publishing Company, 1985.
[31] B. Gatos, I. Pratikakis, S.J. Perantonis, Adaptive degraded document image binarization, Pattern Recognit. 39 (3) (2006) 317–327.
[32] M.A. Ramírez-Ortegón, E. Tapia, L.L. Ramírez-Ramírez, R. Rojas, E. Cuevas, Transition pixel: a concept for binarization based on edge detection and gray-intensity histograms, Pattern Recognit. 43 (4) (2010) 1233–1243.
[33] S. Lu, B. Su, C.L. Tan, Document image binarization using background estimation and stroke edges, Int. J. Doc. Anal.Recognit. (IJDAR) 13 (4) (2010) 303–314.
[34] S. Lu, C.L. Tan, Binarization of badly illuminated document images through shading estimation and compensation, in: International Conference on Document Analysis and Recognition, vol. 1, 2007, pp. 312–316.
[35] B. Su, S. Lu, C.L. Tan, Binarization of historical document images using the local maximum and minimum, in: International Workshop on Document Analysis Systems, 2010, pp. 159–166.
[36] B. Su, S. Lu, C.L. Tan, Combination of document image binarization techniques, in: 2011 International Conference on Document Analysis and Recognition, IEEE, 2011, pp. 22–26.
[37] N.R. Howe, A Laplacian energy for document binarization, in: 2011 International Conference on Document Analysis and Recognition, IEEE, 2011, pp. 6–10.
[38] N.R. Howe, Document binarization with automatic parameter tuning, Int. J. Doc. Anal. Recognit. (IJDAR) 16 (3) (2013) 247–258.
[39] H.Z. Nafchi, R.F. Moghaddam, M. Cheriet, Phase-based binarization of ancient document images: Model and applications, IEEE Trans. Image Process. 23 (7) (2014) 2916–2930.
[40] S. Bhowmik, R. Sarkar, B. Das, D. Doermann, Gib: A {*G*} ame theory {*I*} nspired {*B*} inarization technique for degraded document images, IEEE Trans. Image Process. 28 (3) (2018) 1443–1455.
[41] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted intervention, Springer, 2015, pp. 234–241.
[42] J. Baxter, A Bayesian/information theoretic model of learning to learn via multiple task sampling, Mach. Learn. 28 (1) (1997) 7–39.
[43] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), in: 2009 10th International Conference on Document Analysis and Recognition, IEEE, 2009, pp. 1375–1382.
[44] I. Pratikakis, B. Gatos, K. Ntirogiannis, H-DIBCO 2010-handwritten document image binarization competition, in: 2010 12th International Conference on Frontiers in Handwriting Recognition, IEEE, 2010, pp. 727–732.
[45] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICDAR 2011 Document Image Binarization Contest (DIBCO 2011).
[46] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012), in: 2012 International Conference On Frontiers in Handwriting Recognition, IEEE, 2012, pp. 817–822.
[47] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICDAR 2013 document image binarization contest (DIBCO 2013), in: 2013 12th International Conference on Document Analysis and Recognition, IEEE, 2013, pp. 1471–1476.
[48] K. Ntirogiannis, B. Gatos, I. Pratikakis, ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014), in: 2014 14th International Conference on Frontiers in Handwriting Recognition, IEEE, 2014, pp. 809–813.
[49] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016), in: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2016, pp. 619–623.
[50] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, ICDAR2017 competition on document image binarization (DIBCO 2017), in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, IEEE, 2017, pp. 1395–1403.
[51] I. Pratikakis, K. Zagori, P. Kaddas, B. Gatos, ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018), in: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 489–493.
[52] F. Deng, Z. Wu, Z. Lu, M.S. Brown, Binarizationshop: a user-assisted software suite for converting old documents to black-and-white, in: Proceedings of the 10th Annual Joint Conference on Digital Libraries, ACM, 2010, pp. 255–258.
[53] H.Z. Nafchi, S.M. Ayatollahi, R.F. Moghaddam, M. Cheriet, An efficient ground truthing tool for binarization of historical manuscripts, in: 2013 12th International Conference on Document Analysis and Recognition, IEEE, 2013, pp. 807–811.
[54] R. Hedjam, H.Z. Nafchi, R.F. Moghaddam, M. Kalacska, M. Cheriet, ICDAR 2015 contest on multispectral text extraction (ms-tex 2015), in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 1181–1185.
[55] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
[56] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980(2014).
[57] T. Lelore, F. Bouchara, FAIR: a fast algorithm for document image restoration, IEEE Trans. Pattern Anal. Mach.Intell. 35 (8) (2013) 2039–2048.
[58] N. Mitianoudis, N. Papamarkos, Document image binarization using local features and gaussian mixture modeling, Image Vis. Comput. 38 (2015) 33–51.

Sheng He gained a cum laude Ph.D. degree in artificial intelligence from the University of Groningen, the Netherlands, in 2017. From 2017 to 2018, he was a postdoctoral fellow at the University of Groningen. In 2018, he joined Harvard Medical School as a research fellow. He received the Chinese government award for outstanding self-financed students abroad (2016) from the Chinese Scholarship Council and the Charles A. King Trust Postdoctoral Research Fellowship (2020, in Massachusetts, USA). His research interests include handwritten document analysis, deep learning, and medical image analysis.

Lambert Schomaker is Professor in Artificial Intelligence at the University of Groningen and was the director of its AI institute ALICE from 2001 to 2018. His main interests are pattern recognition and machine learning problems, with applications in handwriting recognition problems. He has contributed to over 200 peer-reviewed publications in journals and books (h=21/ISI, h=47/Google Citations). His work is cited in 23 patents. In recent years, his focus has been on continuous-learning systems and bootstrapping problems, where learning starts using very few examples. Prof. Schomaker is a senior member of IEEE, a member of the IAPR, and a member of Dutch research program committees in e-Science (at the NWO), Computational science and energy (Shell/NWO/FOM). He received two IBM Faculty Awards (in 2011 and 2012) for the Monk word-retrieval system in historical manuscript collections using high-performance computing.