Document Binarization using Recurrent Attention Generative Model

Shuchun Liu[1], *Feiyun Zhang*[1], *Jie Shao*[2], *Pan He*[3], *Mingxi Chen*[1], *Yufei Xie*[1], *Yining Lin*[1], *Yao Peng*[1]

1. Qiniu AtLab, Shanghai, China
2. Fudan University, Shanghai, China
3. University of Florida, Gainesville, USA

{liushuchun, zhangfeiyun, chenmingxi, xieyufei, pengyao}@qiniu.com, shaojie@fudan.edu.cn, pan.he@ufl.edu

Figure 1: Document binarization: combining non-local attention and spatial RNN with state-of-the-art algorithms [1] for document processing, improves performance. Top: original document images. Bottom: binary result images generated with our proposed method.

***Abstract*—Image binarization is an elementary pre-processing step in the document image analysis and recognition pipeline. It is well-known that contextual and semantic information is beneficial to the separation of foreground text from complex background. We develop a simple general deep learning approach, by introducing a recurrent attention generative model with adversarial training. The DB-RAM model comprises three contributions: First, to suppress the interference from complex background, non-local attention blocks are incorporated to capture spatial long-range dependencies. Second, we explore the use of Spatial Recurrent Neural Networks (SRNNs) to pass spatially varying contextual information across an image, which leverages the prior knowledge of text orientation and semantics. Third, to validate the effectiveness of our proposed method, we further synthetically generate two comprehensive subtitle datasets that cover various real-world conditions. Evaluated on various standard benchmarks, our proposed method significantly outperforms state-of-the-art binarization methods both quantitatively and qualitatively. Experiment results show that the proposed method can also improve the recognition rate. Moreover, the proposed method performs well in the task of image unshadowing, which evidently verifies its generality.**

I. Introduction

Image binarization is one of the key pre-processing steps in document image analysis and recognization pipeline. The resulting bi-level image information decreases the computation load and enables the utilization of the simplified analysis methods in the subsequent stages. However, it is challenging to infer an appropriate threshold for the correct binarization of a document from its color or grayscale representation, due to factors such as physical degradation of the document, adverse lighting, or imaging conditions, and limitations on resolution [2], [3].

Most traditional approaches compute the local or global thresholds based on image statistics such as color, gradient. To maximize the separability of the resultant classes in gray levels, Otsus method [4] use the global discriminating thresholding technique, based on a simple linear discriminant criterion. But their performance get worse when applied to scenarios with small size objects and complex background. Niblack's binarization algorithm [5] is developed to preserve minute details at a local level, where they introduced a local window to estimate a threshold value based on the calculation of local mean and standard deviation of pixels value. One disadvantage of these approaches is that they ignore permutations or spatial arrangements of image pixels *i.e.*, statistics ignore shape. Missing these related semantic content information leads to unsatisfied results. These approaches also introduced image dependent parameters according to the image foreground and background conditions. On the other hand, approaches such as the Markov Random Field (MRF) model [6], [7], [8] and Laplacian Energy [9] have been applied to consider image binarization as an optimization problem, where fewer parameters are involved. These approaches are more robust on dealing with different degraded documents. However, the results might be inferior if improper energy definition is introduced when directly applying the MRF model.

Recent work has demonstrated the feasibility of applying machine learning technique for document image analysis [10], [11], [12], [13], [14]. In particular, deep learning based approaches have been applied for binarization

of document images and achieved state-of-the-art performance [15], [16]. In comparison to methods with handcrafted heuristic rules, the advantage lies on their generalizability, only requiring labeled images for building the discriminative model [17]. The key idea is to train Convolutional neural networks (CNNs) to distinguish the category of each pixel of the image, thus separating different layers of the document accordingly.

Towards accurate image binarization, many previous approaches rely on prior information of image context. For example, the connectivity of an individual character must be maintained for optical character recognition and textual compression. Our observation is that current CNN based models are in lack of suitable strategies to utilize global visual clues (*i.e.*, text orientation, line spacing) and model the corresponding long-range dependency. Convolutional operations process a local neighborhood at a time, enabling to capture short-range dependency. Repeating these local operations to propagate information across the whole image results in a larger and deeper architecture, which is computationally inefficient and more difficult to optimize. In addition, current public benchmark training datasets are not sufficient to capture various noisy distributions (*i.e.*, ink spills, artifacts, stains) in real-life situation.

Different from these approaches, we aim at proposing a simple general approach to automatically output binarization results, by introducing a recurrent attention generative model with adversarial training. We utilize the recent Pix2Pix framework proposed by [1] as the backbone of the DB-RAM model. We present the first attempt to incorporate non-local operations for image binarization. These modules efficiently extract reliable correlation between image pixels or image patches, allowing to capture the long-range dependency. To leverages the prior knowledge of text orientation and semantics, we explore the use of Spatial Recurrent Neural Networks (SRNNs) to pass spatially varying contextual information across an image. We provide a comprehensive comparison with recent competitors (such as Pix2Pix [1] and cycleGAN [18]), in which our proposed method achieves state-of-the-art performance in image binarization over serveral benchmark datasets, demonstrating the superiority of the proposed components. We synthesize and release two comprehensive subtitle datasets for the purpose of training and evaluation, which includes both Chinese and English, complex backgrounds, various fonts with different sizes, colors and blurring settings. Moreover, the proposed method performs well in the task of image unshadowing, which evidently verifies its generality.

II. Related Work

A. Text Image Binarization

Most traditional text image binarization methods are based on global or local discriminating thresholds, such as the Otsus method [4]. They assume that images contains two classes of pixels following bi-modal histogram, thus calculating the optimum threshold to separate the two classes and maximize their inter-class variance. However, if object sizes are small or variances of object and background intensities are large, the performance of global thresholding techniques will be heavily degraded [19]. Niblacks method [5], a well-known local threshold binarization method, computes the mean and standard deviation of each block, by introducing the concept of local window. This algorithm shows more robustness compared to the global ones. Still, these approaches introduced image-dependent parameters, which is a non-trivial task of tuning these parameters to a satisfied model. The Markov Random Field (MRF)model [6], [7], [8] and Laplacian Energy are robust methods for image binarization. The limitation is that these methods are very sensitive to the selection of energy function. The results might be inferior with an improper energy definition.

Convolutional Text Binarizer (CTB) is proposed by [20], which is designed for complex color test image binarization. This system does not need any tunable parameter and considers both the color distribution and the geometrical properties of characters. It is more robust and greatly enhances text recognition rates of classical document OCRs on images with noise and contrast variations. Messaoud et al. [21] presents a new approach based on a combination between a preprocessing step and a localization step to apply binarization method on selected objects-of-interest. [22] proposes a new algorithm called BM (Bernsen and Mean), which overcomes the influence of uneven illumination and thereby has a great processing effect for images with rich details and a variety of shapes. However, the parameters of BM algorithm, such as the weights of local threshold and global threshold (λ_1, λ_2), need to be manually adjusted according to images condition. For video text recognition, a new fusion method based on wavelet sub-bands and gradient of different directions is proposed by [23]. This method uses k-means clustering algorithm in different row-wise and column-wise way to obtain text candidates. [24] develops a novel skeleton-based binarization method in order to separate text from complex backgrounds to make it processable for standard OCR software. [15] proposes Fully Convolutional Networks (FCN) trained with a combined Pseudo F-measure (P-FM) and F-measure FM loss, outperforms the competition winners for 4 of 7 DIBCO competitions and is competitive with the state-of-the-art methods on Palm Leaf Manuscripts. Different from these prior works, we utilize Generative Adversarial Nets for image binarization.

B. Generative Model

Generative Adversarial Nets (GANs) [25] is a novel way to train generative models. Many prior works have used GAN to solve various tasks with impressive results. [26] is able to control face attributes by modifying and injecting the conditional information to the model. [27] develops a novel deep architecture with GAN formulation to translate visual concepts from characters to pixels. [28] proposes an attentive generative adversarial network, whose generator

consists of an attentive-recurrent network and a contextual autoencoder with skip connections. Its discriminator is then formed by a series of convolution layers and guided by the attention map. [29] proposes a conditional version of GAN (cGAN), which can be constructed by simply feeding the data as additional input layer to condition both the generator and discriminator. By learning a structured loss, cGan can better penalize any differences between output and ground truth. Furthermore, [1] proposes a Pix2Pix architecture which is not application-specific. Their experiment results demonstrate that it is applicable in a wide variety of settings. [18] proposes cycle-consistent adversarial networks, by adding a cycle consistency loss to translate between domains without paired input-output examples.

Motivated by works mentioned above, we utilize Pix2Pix framework [1] as the backbone of our network. We further explore several techniques to leverage the prior knowledge of text orientation and semantics, with modeling a long-range dependency (attention). In Sec. IV, we will show some evaluations between the DB-RAM model and Pix2Pix.

C. Attention Model

Capturing long-range dependencies is of central importance in deep neural networks [30]. Deep neural networks automatically learning feature representation by stacking multiple end-to-end convolutional or recurrent modules, where each sub module processes correlation within a spatial or temporal local regions. Still, capturing the long-range dependencies requires repeatedly stacking multiple modules, which hinders the learning and inference efficiency.

Inspired by classical non-local operator for image filtering, [30] proposes the non-local neural network that eases the problem, by directly modeling correlation between each positions in one single module. They relate non-local operation to recent self-attention [31] as a special case of non-local operations in the embedded Gaussian version [30]. Self-attention, also called intra-attention, is an attention mechanism relating different positions and helps modeling long-range, multi-level dependencies. It has been used in a variety of tasks. [32] proposes this mechanism for question encoding in their Factoid Question Answering model. For sentiment analysis and entailment, [33] proposes self-attention mechanism to extract different aspects of the sentence into multiple vector representations. [31] dispenses with complex recurrent and convolutional neural networks in dominant sequence transduction models and proposes their model solely based on attention mechanism. They introduce the self-attention mechanism to both their encoder and decoder model. Considering of using convolutional layers alone is computationally inefficient for modeling long-range dependencies in images, [34] introduces self-attention to the GAN framework. Inspiring by these, we adopt non-local attention mechanism to the Pix2Pix framework, since long text images need long-range dependencies to capture global information (*i.e.*, image textures, image styles, color statistics).

III. Methodology

Fig. 2 provides an overview of our network. There are two main parts in our network: the generative and discriminative networks. The DB-RAM is the abbreviation of our Document Binarization Recurrent Generative Attention Model. To be consistent with the idea of Pix2Pix framework [1], we provide pairs of images which contain original text images and corresponding binary results as ground truth. The sizes of input images are 32×280. The conditional GANs learn a mapping from observed image x and random noise vector z, to y, $G : \{x, z\} \rightarrow y$. The objective of a conditional GAN can be expressed as:

$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \tag{1}$$

where G represents the generative network, and D represents the discriminative network. Similar to [1], we use L1 distance which encourages less blurring:

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{x,y,Z}[\|Y - G(x, z)\|_1] \tag{2}$$

A. Generative Network

As shown in Fig. 2, our generative network mainly consists of three parts: non-local attention module, spatial RNN module and Residual blocks [35]. The purpose of utilizing non-local attention mechanism is to better extract long-range dependency information between text regions. To leverage the prior knowledge of text orientation and semantic, the spatial RNN module is further explored. The experiment results will show that these two modules are complementary that can jointly boost up the performance.

1) Non-local Attention Module: . The generative networks start with 3 convolution layers that help extract features from the input images. Then the non-local attention module is stacked to learning the weight function for each feature position. Considering the input consisting of queries and keys of dimension d_k , and values of dimension d_v, we follow the attention function proposed in [31], where the attention weight for each value is computed as the dot products of the query with all keys, divided by a temperature factor $\sqrt{d_k}$, followed with the softmax function:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Given the image features $\phi(X) \in R^{B \times C \times W \times H}$, we instantiate the attention function with the following linear projection:

$$Q = W_q * \phi(X) \tag{4}$$

$$K = W_k * \phi(X) \tag{5}$$

$$V = W_v * \phi(X) \tag{6}$$

where * represents the convolution operation, $W_q \in R^{C^{'} \times C}, W_k \in R^{C^{'} \times C}, W_v \in R^{C \times C}$ are the learned weighted matrices.

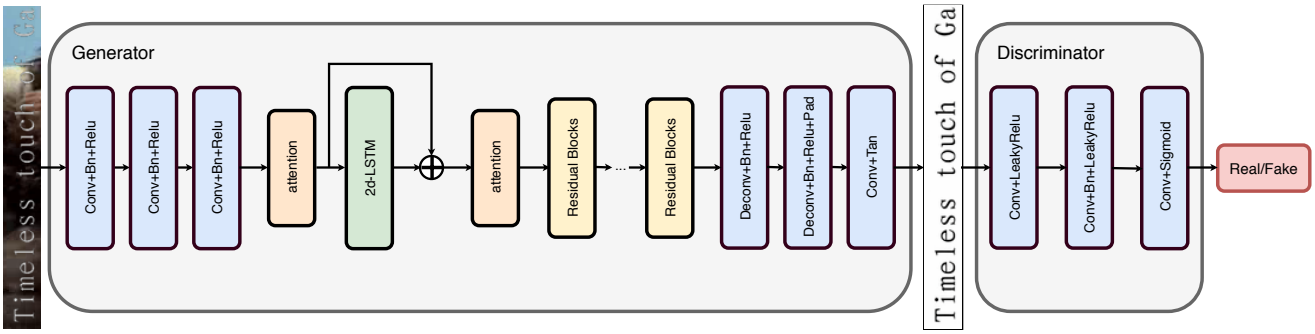


Figure 2: The architecture of our proposed network. The generator consists of Non-local Attention modules, Spatial RNN modules and Residual blocks. The discriminator is formed by a series of convolution layers.

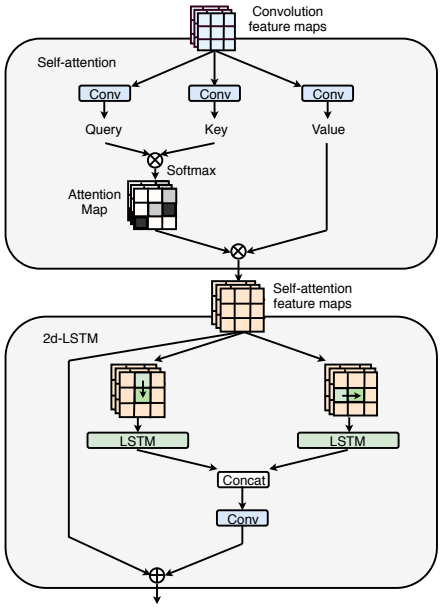


Figure 3: The architecture of non-local attention module and spatial RNN.

We multiply the output of the attention layer by a scale parameter α (is initialized as 0) and add back the input feature map. Therefore, the final output Y is given by:

$$Y = \alpha \times \text{Attention}(Q, K, V) + \phi(X) \tag{7}$$

2) Spatial RNN module: . Traditionally, at every step, an RNN model moves left-to-right along a sequence, consuming an input, updating its hidden state, and producing an output [36]. On the top of the first non-local attention layer, we instead extend this to two dimensions by moving the RNN along each row and along each column of the image. The mechanism naturally utilizes text orientation and line spacing information. We use LSTM for the RNN module. Still, there are other possible forms of recurrent neural networks that we could use. The output features of LSTM are followed with another non-local attention module to further guide the generator to extract the dependencies among pixels. By doing this, the utilization of two non-local attention layers which respectively focuses on global and local information that effectively guide our generative network to pay more attention to important text regions and eliminate interference of backgrounds. Finally, the attention maps are fed into 6 Residual blocks [35] followed with 2 deconvolution layers and 1 convolution layer to generate corresponding binary image.

3) Ablation Study: . To explore the contributions of the proposed components in our model, we analyzed the results of three extra experiments: model with only attention, model with only LSTM, model with attention plus Bi-LSTM. The quantitative comparisons are shown in Tab. III. As can be seen in Fig. 4, removing either non-local attention module or spatial RNN module causes loss on the performance, which proves that the combination of long-range dependency and semantic information is beneficial to suppress the interference. Moreover, traditional LSTM outperforms our spatial RNN module on two subtitle datasets, in contrast, it is inferior to the DB-RAM model on DIBCO and PLM. It reveals that Bi-LSTM, as a serialization module, is more suitable for processing single line text (Fig. 5(a), Fig. 5(b)). Conversely, spatial RNN module makes better use of line spacing information which is able to handle document images and natural scene images (Fig. 1, Fig. 5(c), Fig. 5(d)).

B. Discriminative Network

We apply 3 convolution layers for discriminative network with 1×1 filter size to differential fake images from real ones. Since the Structural Similarity (SSIM) [37] index is a method for measuring the similarity between two images. The higher the SSIM, the more similar the two images are. We introduce it as a part of our loss function. The SSIM index can be written as:

$$SSIM(f, g) = \frac{(2\mu_f\mu_g + C_1)(2\sigma_{fg} + C_2)}{(\mu_f^2 + \mu_f^2 + C_1)(\mu_g^2 + \mu_g^2 + C_2)} \tag{8}$$

where f is the fake image generated by the generative network, g is corresponding ground truth image, μ is the mean of image's pixels and σ is the standard deviation of

Figure 4: The results of ablation experiments.

Dataset	Subtitle_English	Subtitle_Chinese
Font Numbers	10	8
Font Types	N+I	N+I
Font colors	V	V
Gaussian Blur	True	True
Language	E	C
Training	100, 000	100, 000
Testing	10, 000	10, 000

Table I: The description of each synthetic dataset. N indicates non-italic. I indicates italic. V indicates a variety of colors. E indicates English. C indicates Chinese

Layer Name	Output Size	Details
Cnn_1	$64 \times 32 \times 280$	$7 \times 7 \quad 64$
Cnn_2	$256 \times 8 \times 70$	$3 \times 3 \quad 128$ $3 \times 3 \quad 256$
Attn_1	$256 \times 8 \times 70$	$1 \times 1 \quad 8$ $1 \times 1 \quad 8$ $1 \times 1 \quad 64$
2d-LSTM	$512 \times 8 \times 70$	Hidden size = 256
Conv_1	$256 \times 8 \times 70$	$1 \times 1 \quad 256$
Attn_2	$256 \times 8 \times 70$	$1 \times 1 \quad 32$ $1 \times 1 \quad 32$ $1 \times 1 \quad 256$
Residual Block	$256 \times 8 \times 70$	$3 \times 3 \quad 256$ $3 \times 3 \quad 256$ $\times 6$
Deconv_1	$128 \times 16 \times 140$	$3 \times 3 \quad 128$
Deconv_2	$64 \times 32 \times 280$	$3 \times 3 \quad 64$
Padd	$64 \times 38 \times 286$	Padding = 256
Conv_2	$3 \times 32 \times 280$	$7 \times 7 \quad 3$

Table II: Architecture of our proposed generative network

image's pixels. Both C_1 and C_2 are the constants. Overall, the final loss is:

$$\mathcal{L} = \min_G \max_D \quad \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{L1}(G) + \beta(1 - SSIM(f, g)) \tag{9}$$

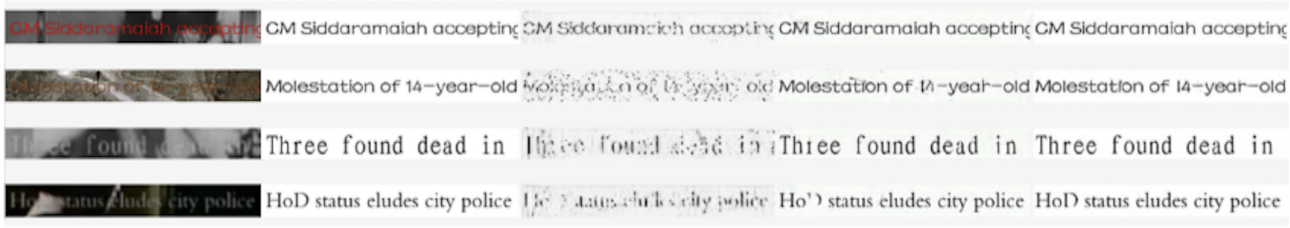
IV. Experiment Results

A. Image Binarization

1) Benchmark Datasets: To comprehensively explore the performance of the DB-RAM model under different situations, we evaluate the DB-RAM model on DIBCOs [38], [39], [40], [41], [42], [43], [44], Palm Leaf Manuscripts (PLM) [45] and 2 synthesis datasets shown in Tab. I. The DIBCOs contain images that range from gray scale to color, from machine printed to handwritten, and finally, from real to synthetic. The palm leaf manuscripts contain discolored parts and artefacts due to aging and low intensity variations or poor contrast, random noises, and fading. For DIBCOs experiments, a total of 9 datasets are used: DIBCO 2009, DIBCO 2011, DIBCO 2013, H-DIBCO 2010, HDIBCO 2012, H-DIBCO 2014, Bickley diary, PHIDB, and S-MS datasets. Out of these datasets, DIBCO 2013 dataset is selected for testing purposes. For the testing, the remaining datasets are used as a training set. We convert the images from these datasets to patches of size 256×256. For PLM experiment, we vertically cut every image into 10 equal parts, then randomly split the 500 images into 400 for training and 100 for testing. We apply a variety of fonts, colors, backgrounds and Gaussian blurring processing to ensure the diversity of data.

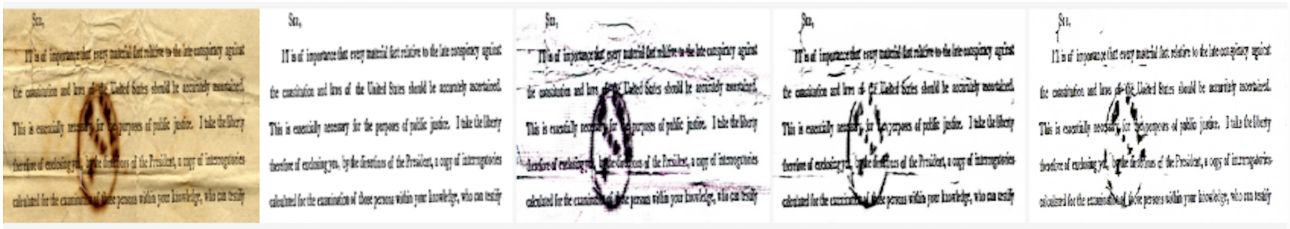
2) Implementation Details: We use Pytorch to implement our proposed model. Tab. II shows detailed architectures of our generative network. We train our network end-to-end with lambda learning rate which is initialized with 0.0002. For Pix2Pix [1] and our model, we train about 200 epochs. For cycleGAN [18], we train only 50 epochs since its training speed is too slow to converge. We train our model on a single Tesla-V100-PCIE graphics card with 16GB memory for each experiment and the batch size is 16 for 2 synthesis datasets, 1 for DIBCOs and PLM. As training, two weights of the loss function showed as Eq. (9), λ and β, are set to 5.0 and 5.0, respectively.

3) Quantitative Evaluation: Tab. III shows the quantitative comparisons on each dataset between our proposed method and other existing methods including GAN [25], Pix2Pix [1] and cycleGAN [18]. As shown, the DB-RAM model improves both the PSNR and SSIM values compared to these state-of-the-art methods. In Subtitle_Chinese dataset, the DB-RAM model outperforms the Pix2Pix method by 3% for SSIM index and achieve 0.9886 of SSIM in Subtitle_English dataset. As for DIBCOs and PLM datasets, the SSIM and PSNR index are also improved by the DB-RAM model. We also compare our whole network with some parts of it: 'Non-local' denotes baseline generator with only non-local attention; 'Only-LSTM' denotes

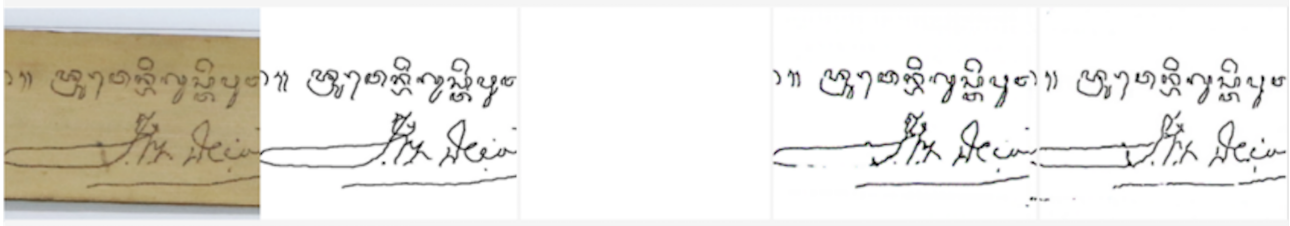


(a) Subtitle_English

(b) Subtitle_Chinese



(c) HDIBCO 2016



(d) PLM

Figure 5: Qualitative results of comparing a few state-of-the-art methods on several datasets. From left to right: text image (input), Pix2Pix [1], cycleGAN [18] and DB-RAM. The cycleGAN [18] results in Fig.4.2 and Fig.4.4 are totally white.

baseline generator with only 1d-LSTM; 'Non-local+Bi-LSTM' means the generator with self-attention and Bi-LSTM; 'DB-RAM' means our complete generator with non-local attention and 2D Spatial RNN. As shown in the evaluation table, 'DB-RAM' performs better than the other possible configurations. The results validate that Spatial RNN and non-local attention mechanism are beneficial on boosting up the performance of the generator. Moreover, due to the spatial structure and ability to capture long-range correlations, simultaneously processing line and column traversal is better than processing sequentially.

4) Qualitative Evaluation: Fig. 5 shows the results of Pix2Pix [1] and cycleGAN [18] on each dataset in comparison to our results. As can be seen, when image background is relatively complex and similar to font color, cycleGAN [18] and Pix2Pix [1] both tend to misrecognize text as background or turn the texture of background as a part of word. It needs to be mentioned that the cycleGAN [18] results on Subtitle_Chinese in Fig. 5(b) and PLM in Fig. 5(d) are totally white. The extremely poor performance can be explained by the mechanism of cycleGAN [18]. This method is proposed to learn an image-to-image translation in the absence of pairs of training data, which means the mapping from input images to output images has to be learned by the network itself during the training process. In our synthesis datasets, the variety of fonts and background can obviously increase the difficulty. As for PLM, the size of training set is only 40 which is too small for cycleGAN [18] to learn the mapping. By contrast, the DB-RAM model is considerably more effective in handling different fonts with a variety of sizes, colors, blurring setting and complex backgrounds. The DB-

Dataset	Method	Metrics	
		PSNR	SSIM
Sub_En	Pix2Pix [1]	34.64	0.9576
	cycleGAN [18]	28.94	0.6080
	Non-local	36.33	0.9681
	Only LSTM	36.35	0.9678
	Non-local+Bi-LSTM	**38.01**	**0.9898**
	DB-RAM	37.87	0.9886
Sub_Ch	Pix2Pix [1]	32.93	0.8915
	cycleGAN [18]	33.75	0.5637
	Non-local	33.69	0.9010
	Only LSTM	33.39	0.8636
	Non-local+Bi-LSTM	**34.51**	**0.9581**
	DB-RAM	34.19	0.9145
HDIBCO-16	Pix2Pix [1]	35.62	0.9258
	cycleGAN [18]	34.85	0.8485
	Non-local	38.95	0.9312
	Only LSTM	38.79	0.9276
	Non-local+Bi-LSTM	39.59	0.9340
	DB-RAM	**39.70**	**0.9365**
PLM	Pix2Pix [1]	40.44	0.8718
	cycleGAN [18]	42.18	0.7764
	Non-local	43.11	0.8702
	Only LSTM	40.64	**0.8748**
	Non-local+Bi-LSTM	44.01	0.8720
	DB-RAM	**44.60**	0.8735

Table III: Quantity evaluation results. 'Sub_En' and 'Sub_Cn' ref to two synthetic datasets, respectively. 'Non-local' denotes baseline generator with only non-local attention; 'Only-LSTM' denotes baseline generator with only 1d-LSTM; 'Non-local+Bi-LSTM' means the generator with self-attention and Bi-LSTM; 'DB-RAM' means our complete generator with non-local attention and 2D spatial RNN.

Input	Metrics	
	ED	normalized ED
Original Chinese Images	2.65	0.23
Binary Chinese Images	**1.08**	**0.09**
Original English Images	3.39	0.14
Binary English Images	**0.53**	**0.02**

Table IV: The result of employing Google Vision API on original text images and the DB-RAM model's outputs. ED means the edit distance. the DB-RAM model decrease the edit distance as well as normalized distance.

RAM model can significantly improve the recognition rate of existing recognition pipeline. We utilize Google Vision API [1] to compare the results between original text images and the corresponding binary outputs given by the DB-RAM model. We test on 500 images for each subtitle dataset. Binarized images provided by the DB-RAM model can effectively ease these problems, for both English or Chinese datasets. Tab IV provides further quantitative result that evaluations on the edit distance and normalized edit distance are both significantly improved by ours which prove that the DB-RAM model is able to improve the recognition rate.

Fig. 6 shows some failure samples. However, these only happen when the background is extremely similar to the foreground, even too hard to human eyes.

[1]https://cloud.google.com/vision/

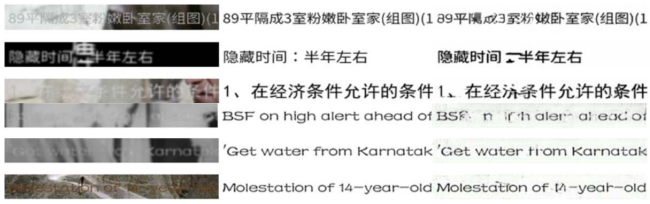


Figure 6: Some failure samples. From left to right are input image, ground truth and output.

Dataset	Method	Metrics	
		PSNR	SSIM
ISTD	Pix2Pix [1]	29.20	0.7854
	Only LSTM	29.34	0.8454
	Non-local	29.96	0.8699
	DB-RAM	**30.03**	**0.8719**

Table V: Quantity evaluation results of shadow removal task.

B. Shadow Removal

Similar to complicated backgrounds, shadow is another kind of interference that could hamper the performance of many computer vision applications. Therefore, shadow detection and removal plays an important role in pre-processing of outdoor images. Most of the previous work [46], [47] neglect high level semantic information. *i.e.*, under each shadow region, the lightness shares similar image statistics. The boundaries between shadowed and unshadowed regions usually have strong edges). [48] introduced CGAN [29] in shadow detection which efficiently reasons about the global scene structure and illumination conditions. Furthermore, [49] proposes STacked Conditional Generative Adversarial Network (ST-CGAN) for jointly learning shadow detection and shadow removal. This method is able to preserve the global scene characteristics hierarchically.

The DB-RAM model focuses on the combination of spatial long-range dependencies and textual information, this philosophy might also work on the shadow removal task since the correlations between shadow regions and global scenes also need attention and contain rich semantic information. To further explore the generalization of the DB-RAM model, we evaluated our proposed method on the large-scale Dataset with Image Shadow Triplets (ISTD) constructed by [49]. It contains 1870 triplets image under 135 different scenarios, in which 1330 is assigned for training while 540 is for testing. We mainly explore the shadow removal task, so only the shadow and shadow-free images are used in the experiment. Fig. 7 shows the results of shadow removal experiment. As shown in Tab. V, the DB-RAM model outperforms Pix2Pix [1] on both PSNR and SSIM index. It demonstrates that the DB-RAM model has a general capability to enhance images from different kinds of interference.

V. Conclusion

We have proposed a new binarization method. The method utilizes Pix2Pix framework, where the generative

Figure 7: results of shadow removal experiment. From left to right are input image, ground truth and output.

network produces the attention map combined with spatial RNN and the discrimination network uses SSIM loss to quantitative evaluation of generative models. This method is fully automatic and it can achieve the state-of-the-art performance in binarization on different datasets. As for shadow removal experiment, the DB-RAM model also performed well on ISTD dataset. It further validate the generality of it. Moreover, as an additional contribution, we synthesized and will release two datasets contain both English and Chinese subtitle text images to the public. In future work, we plan to investigate the DB-RAM models ability in handling more image enhancement tasks.

REFERENCES

[1] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5967–5976.

[2] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 16, no. 3, pp. 247–258, 2013.

[3] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," *IEEE transactions on image processing*, vol. 22, no. 4, pp. 1408–1417, 2013.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[5] W. Niblack, *An introduction to digital image processing*, vol. 34.

[6] B. Su, S. Lu, and C. L. Tan, "A learning framework for degraded document image binarization using markov random field," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3200–3203.

[7] Y. Wang, C. Shi, B. Xiao, and C. Wang, "Mrf based text binarization in complex images using stroke feature," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 821–825.

[8] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "An mrf model for binarization of music scores with complex background," *Pattern Recognition Letters*, vol. 69, pp. 88–95, 2016.

[9] N. R. Howe, "A laplacian energy for document binarization," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 6–10.

[10] M. Seuret, M. Alberti, M. Liwicki, and R. Ingold, "Pca-initialized deep neural networks applied to document image analysis," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 877–882.

[11] E. Vats, A. Hast, and P. Singh, "Automatic document image binarization," *arXiv preprint arXiv:1709.01782*, 2017.

[12] P. Roy and S. Adhikari, "An entropy-based binarization method to separate foreground from background in document image processing." *IUP Journal of Telecommunications*, vol. 10, no. 2, 2018.

[13] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, "Gib: a game theory inspired binarization technique for degraded document images," *IEEE Transactions on Image Processing*, 2018.

[14] W. Xiong, J. Xu, Z. Xiong, J. Wang, and M. Liu, "Degraded historical document image binarization using local features and support vector machine (svm)," *Optik*, vol. 164, pp. 218–223, 2018.

[15] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 99–104.

[16] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, pp. 568–586, 2018.

[17] R. Memisevic, "An introduction to structured discriminative learning," Technical report, University of Toronto, Toronto, Canada, Tech. Rep., 2006.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[19] S. U. Lee, S. Y. Chung, and R. H. Park, "A comparative performance study of several global thresholding techniques for segmentation," *Computer Vision, Graphics, and Image Processing*, vol. 52, no. 2, pp. 171–190, 1990.

[20] Z. Saidane and C. Garcia, "Robust binarization for video text recognition," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 874–879.

[21] I. B. Messaoud, H. Amiri, H. El Abed, and V. Margner, "New binarization approach based on text block extraction," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1205–1209.

[22] J. WANG, Z. HUANG, and A. TALAB, "A. new binarization method called bm aim to optimize detail of image [j]," *Journal of Wuhan University of Technology*, vol. 36, no. 8, pp. 127–132, 2014.

[23] S. Roy, P. Shivakumara, P. P. Roy, and C. L. Tan, "Wavelet-gradient-fusion for video text binarization," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 3300–3303.

[24] H. Yang, B. Quehl, and H. Sack, "A framework for improved video text detection and recognition," *Multimedia Tools and Applications*, vol. 69, no. 1, pp. 217–245, 2014.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[26] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[28] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2482–2491.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CVPR*, 2018.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[32] P. Li, W. Li, Z. He, X. Wang, Y. Cao, J. Zhou, and W. Xu, "Dataset and neural recurrent sequence labeling model for open-domain factoid question answering," *arXiv preprint arXiv:1607.06275*, 2016.

[33] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[34] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2874–2883.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[38] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 1375–1382.

[39] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-dibco 2010-handwritten document image binarization competition," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. IEEE, 2010, pp. 727–732.

[40] ——, "Icdar 2011 document image binarization contest (dibco 2011)," in *2011 International Conference on Document Analysis and Recognition*, Sept 2011, pp. 1506–1510.

[41] ——, "Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012)," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*. IEEE, 2012, pp. 817–822.

[42] ——, "Icdar 2013 document image binarization contest (dibco 2013)," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1471–1476.

[43] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Icfhr2014 competition on handwritten document image binarization (h-dibco 2014)," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 809–813.

[44] I. Pratikakis, K. Zagoris, G. Barlas, and 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.

[45] J.-C. Burie, M. Coustaty, S. Hadi, M. W. A. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I. M. G. Sunarya, and D. Valy, "Icfhr2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 596–601.

[46] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*. University of Bath, 2014.

[47] E. Arbel and H. Hel-Or, "Shadow removal using intensity surfaces and texture anchor points," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 6, pp. 1202–1216, 2011.

[48] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, and D. Samaras, "Shadow detection with conditional generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4520–4528.

[49] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1788–1797.