Document Binarization using Recurrent Attention Generative Model

ShunChun Liu[1]
liuscgood@gmail.com
FeiYun Zhang[1]
zhangfeiyunzfy@gmail.com
MingXi Chen[2]
chenmingxi418@hotmail.com
YuFei Xie[3]
yufeixie@ica.stc.sh.cn
Pan He[4]
pan.he@ufl.edu
Jie Shao[5]
shaojie@fudan.edu.cn

[1] The AI Lab of ELEME Inc, China
[2] Computer Science Department, New York University, USA
[3] Computer Science Department, East China Normal University ,China
[4] Department of Computer and Information Science and Engineering, University of Florida,USA
[5] Fudan University,China
ByteDance AI Lab,China

Abstract

We develop a general deep learning approach, by introducing a recurrent attention generative model with adversarial training to do image binarization which is an elementary pre-processing step in the document image analysis and recognition pipeline. The document binarization using recurrent attention generative(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. The supplementary material also shows that the proposed method improves the recognition rate and also performs well in the task of image unshadowing, which evidently verifies its generality.

1 Introduction

It is well-known that contextual and semantic information is beneficial to the separation of foreground text from complex background. 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 grayscale or color representation, due to factors

such as physical degradation of the document, adverse lighting, or imaging conditions, and limitations on resolution [7, 27]. Most traditional approaches compute the local or global

Figure 1: Top: original document images. Bottom: binary result images generated with our proposed method.

thresholds based on image statistics such as color, gradient. To maximize the separability of the resultant classes in gray levels, Otsu's method [15] 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 [13] is developed to preserve minute details at a local level, where they introduced a local window to estimate a threshold value by calculating the local mean and standard deviation of pixels value. These approaches ignore permutations or spatial arrangements of image pixels which can leads to unsatisfied results, with missing these related semantic content information. On the other hand, the Markov Random Field (MRF) model [26, 31, 34] and Laplacian Energy [6] 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 [1, 23, 25, 30, 36]. In particular, deep learning based approaches have been applied for binarization of document images and achieved state-of-the-art performance [28, 32]. Comparing to methods with hand-crafted heuristic rules, the advantage lies on their generalizability, only requiring labeled images for building the discriminative model [11].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 [8] 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 leverage 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 [8] and cycleGAN [38]), 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.

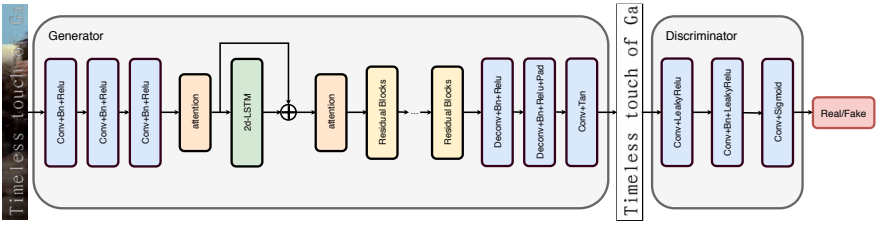


Figure 2: The architecture of our proposed network. The generator consists of Non-local Attention modules, Spatial RNN modules and Residual blocks. The discriminator is formed by a series of convolution layers.

2 Related work

2.1 Text Image Binarization

Most traditional methods of text image binarization are based on local or global discriminating thresholds, such as the Otsu's method [15], which calculate the optimum threshold to separate the two classes and maximize their inter-class variance. Niblack's method [13], a well-known local threshold binarization method, computes the mean and standard deviation of each block, by introducing the concept of local window. These approaches both introduced image-dependent parameters, which is a non-trivial task of tuning these parameters to a satisfied model. Convolutional Text Binarizer (CTB) proposed by [24], 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. [28] 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.

2.2 Generative Model

Generative model is the first part of Generative Adversarial Nets (GANs) [9],which is a novel way to train generative models. [22] develops a novel deep architecture by GAN formulation to translate visual concepts from characters to pixels. [21] proposes an attentive generative model, whose generator consists of a contextual autoencoder with skip connections and an attentive-recurrent network. [12] 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. Furthermore, [8] proposes a Pix2Pix architecture which is not application-specific. Their experiment results demonstrate that it is applicable in a wide variety of settings. [38] proposes cycle-consistent adversarial networks, by adding a cycle consistency loss to translate between domains without paired input-output examples. We utilize Pix2Pix framework [8] as the backbone of our network. We further explore several techniques to leverage the prior knowledge of text orientation and semantics by modeling a long-range dependency (attention). In Sec. 4, we will show some evaluations between the DB-RAM model and Pix2Pix.

2.3 Attention Model

Capturing long-range dependencies is of central importance in deep neural networks [33]. 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, [33] 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 [29] as a special case of non-local operations in the embedded Gaussian version [33]. 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. [9] proposes this mechanism for question encoding in their Factoid Question Answering model. For sentiment analysis and entailment, [10] proposes self-attention mechanism to extract different aspects of the sentence into multiple vector representations. [29] introduces 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, [37] 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 (e.g. image textures, image styles, color statistics).

3 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 [8], 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 the target image 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 [8], we use L1 distance which encourages less blurring:

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

3.1 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 [5]. The purpose of utilizing non-local attention mechanism is to extract better 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.

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 [29], 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:

$$Q = W_q * \phi(X); K = W_k * \phi(X); V = W_v * \phi(X) \tag{4}$$

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.

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

$$Y = \alpha \times \text{Attention}(Q,K,V) + \phi(X) \tag{5}$$

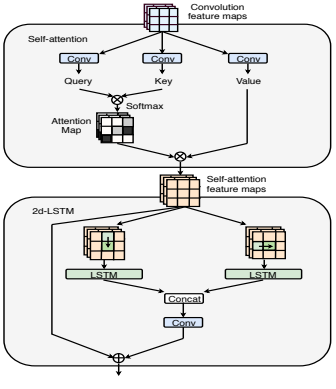


Figure 3: The architecture of non-local attention module and spatial RNN.

Layer Name	Output Size	Details
Cnn_1	$64\times32\times280$	7×7 64
Cnn_2	$256\times8\times70$	3×3 128 3×3 256
Attn_1	$256\times8\times70$	1×1 8 1×1 8 1×1 64
2d-LSTM	$512\times8\times70$	Hidden size = 256
Conv_1	$256\times8\times70$	1×1 256
Attn_2	$256\times8\times70$	1×1 32 1×1 32 1×1 256
Residual Block	$256\times8\times70$	3×3 256 3×3 256 $\times6$
Deconv_1	$128\times16\times140$	3×3 128
Deconv_2	$64\times32\times280$	3×3 64
Padd	$64\times38\times286$	Padding = 256
Conv_2	$3\times32\times280$	7×7 3

Table 1: Architecture of our proposed generative network

Spatial RNN module. 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. 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,

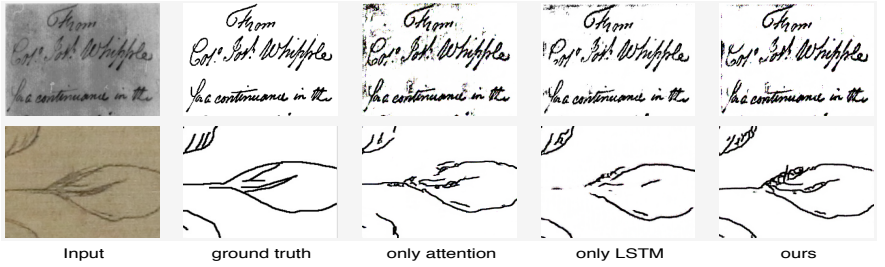


Figure 4: The results of ablation experiments.

the attention maps are fed into 6 Residual blocks [5] followed with 2 deconvolution layers and 1 convolution layer to generate corresponding binary image.

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. 2. 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)).

3.2 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) [35] 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:

$$SSIM(f,g) = \frac{(2\mu_f\mu_g + C_1)(2\sigma_{fg} + C_2)}{(\mu_f^2 + \mu_f^2 + C_1)(\mu_g^2 + \mu_g^2 + C_2)} \tag{6}$$

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

$$\mathcal{L} = \min_G \max_D \quad \mathcal{L}_{\text{cGAN}}(G,D) + \lambda \mathcal{L}_{L1}(G) + \beta(1 - SSIM(f,g)) \tag{7}$$

Dataset	Metrics	Method					
		Pix2Pix [8]	cycleGAN [38]	Non-local	only LSTM	Non-local +Bi-LSTM	DB-RAM
Sub_En	PSNR	34.64	28.94	36.33	36.35	**38.01**	37.87
	SSIM	0.9576	0.6080	0.9681	0.9678	**0.9898**	0.9886
Sub_Ch	PSNR	32.93	33.75	33.69	33.39	**34.51**	34.19
	SSIM	0.8915	0.5637	0.9010	0.8636	**0.9581**	0.9145
HDIBCO-16	PSNR	35.62	34.85	38.95	38.79	39.59	**39.70**
	SSIM	0.9258	0.8485	0.9312	0.9276	0.9340	**0.9365**
PLM	PSNR	40.44	42.18	43.11	40.64	43.20	**44.01**
	SSIM	0.8718	0.7764	0.8702	**0.8748**	0.8720	0.8735

Table 2: 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.

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

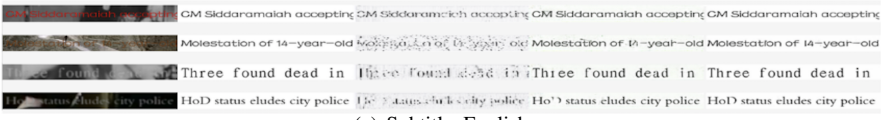
Table 3: The description of each synthetic dataset. 'N' indicates non-italic. 'I' indicates italic. 'V' indicates a variety of colors. 'E' indicates English. 'C' indicates Chinese

4 Experiment Results

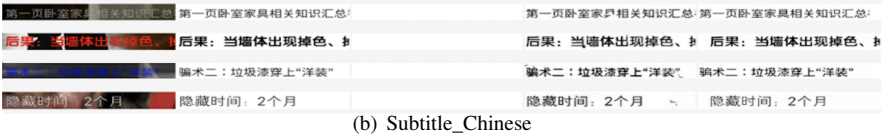
4.1 Image Binarization

4.1.1 Benchmark Datasets

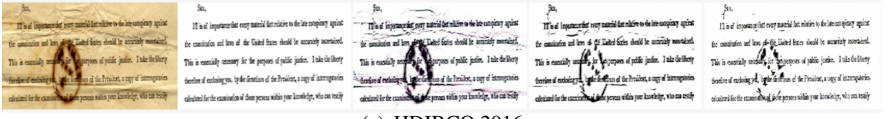
To comprehensively explore the performance of the DB-RAM model under different situations, we evaluate the DB-RAM model on DIBCOs [3, 14, 16, 17, 18, 19, 20], Palm Leaf Manuscripts (PLM) [2] and 2 synthesis datasets shown in Tab. 3. The DIBCOs contain images that range from gray scale to color, from 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.



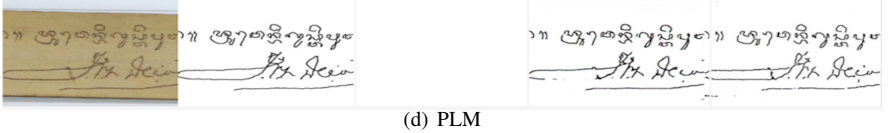
(a) Subtitle_English



(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), ground truth, cycleGAN [38], Pix2Pix [8] and DB-RAM. The cycleGAN [38] results in Fig.5 (b) and Fig.5 (d) are totally white.

4.1.2 Implementation Details

Tab. 1 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 [8] and our model, we train about 200 epochs. For cycleGAN [38], 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. (7), λ and β, are set to 5.0 and 5.0, respectively.

4.1.3 Quantitative Evaluation

Tab. 2 shows the quantitative comparisons on each dataset between our proposed method and other existing methods including GAN [4], Pix2Pix [8] and cycleGAN [38]. 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 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

Method	Metrics	Dataset						
		09	10	11	12	13	14	16
Winning System	P-FM	-	-	-	-	94.16	97.65	91.84
	FM	91.24	91.50	88.74	92.85	92.70	96.88	88.72
FCNs([28])	P-FM	92.59	**97.68**	**97.70**	**96.76**	**96.81**	94.78	93.76
	FM	89.76	**94.89**	93.64	92.53	93.17	91.96	89.52
Non-local+Bi-LSTM	P-FM	**94.68**	95.85	95.80	96.16	96.52	**98.18**	**94.84**
	FM	**93.90**	**94.89**	**94.59**	**94.93**	**96.04**	**97.58**	**91.13**

Table 4: Comparison of 'Non-local+Bi-LSTM' with DIBCO Competition Winners and [28]. The "-" symbols indicate unreported metrics.

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. Following [28], we used all 86 competition images from 2009-2016 for both training and test data. We adopted DIBCO 2013's Evaluation Tool and compared our results with every year's competition winning system and [28] using F-Measure (FM) and Pseudo-FM (P-FM).The experimental results are presented in Tab. 4. On DIBCO 2009, H-DIBCO 2014 and H-DIBCO 2016, 'Non-local+Bi-LSTM' outperforms both [28] and the best submitted entries of competitions in items of FM and P-FM.

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

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 5: 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.

4.1.4 Qualitative Evaluation

Fig. 5 shows the results of Pix2Pix [8] and cycleGAN [33] on each dataset in comparison to our results. As can be seen, when image background is relatively complex and similar to font color, cycleGAN [33] and Pix2Pix [8] 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 [33] 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 [33]. 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 [33] 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-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

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

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 5 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.

5 Conclusion

We have proposed a new binarization method. The method utilizes Pix2Pix framework, where the generative 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 model's ability in handling more image enhancement tasks.

References

[1] Showmik Bhowmik, Ram Sarkar, Bishwadeep Das, and David Doermann. Gib: a game theory inspired binarization technique for degraded document images. *IEEE Transactions on Image Processing*, 2018.

[2] Jean-Christophe Burie, Mickaël Coustaty, Setiawan Hadi, Made Windu Antara Kesiman, Jean-Marc Ogier, Erick Paulus, Kimheng Sok, I Made Gede Sunarya, and Dona 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*, pages 596–601. IEEE, 2016.

[3] Basilis Gatos, Konstantinos Ntirogiannis, and Ioannis Pratikakis. Icdar 2009 document image binarization contest (dibco 2009). In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 1375–1382. IEEE, 2009.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Nicholas R Howe. A laplacian energy for document binarization. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 6–10. IEEE, 2011.

[7] Nicholas R Howe. Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3):247–258, 2013.

[8] 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)*, pages 5967–5976, July 2017. doi: 10.1109/CVPR.2017.632.

[9] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*, 2016.

[10] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[11] Roland Memisevic. An introduction to structured discriminative learning. Technical report, Technical report, University of Toronto, Toronto, Canada, 2006.

[12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[13] Wayne Niblack. *An introduction to digital image processing*, volume 34.

[14] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813. IEEE, 2014.

[15] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[16] I. Pratikakis, B. Gatos, and K. Ntirogiannis. Icdar 2011 document image binarization contest (dibco 2011). In *2011 International Conference on Document Analysis and Recognition*, pages 1506–1510, Sept 2011. doi: 10.1109/ICDAR.2011.299.

[17] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. H-dibco 2010-handwritten document image binarization competition. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 727–732. IEEE, 2010.

[18] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 817–822. IEEE, 2012.

[19] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icdar 2013 document image binarization contest (dibco 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476. IEEE, 2013.

[20] Ioannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos. Icfhr2016 handwritten document image binarization contest (h-dibco 2016). In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 619–623. IEEE, 2016.

[21] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2018.

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[23] Parthajit Roy and Swati Adhikari. An entropy-based binarization method to separate foreground from background in document image processing. *IUP Journal of Telecommunications*, 10(2), 2018.

[24] Zohra Saidane and Christophe Garcia. Robust binarization for video text recognition. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 874–879. IEEE, 2007.

[25] Mathias Seuret, Michele Alberti, Marcus Liwicki, and Rolf Ingold. Pca-initialized deep neural networks applied to document image analysis. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 877–882. IEEE, 2017.

[26] Bolan Su, Shijian Lu, and Chew Lim Tan. A learning framework for degraded document image binarization using markov random field. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3200–3203. IEEE, 2012.

[27] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4): 1408–1417, 2013.

[28] Chris Tensmeyer and Tony Martinez. Document image binarization with fully convolutional neural networks. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 99–104. IEEE, 2017.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[30] Ekta Vats, Anders Hast, and Prashant Singh. Automatic document image binarization. *arXiv preprint arXiv:1709.01782*, 2017.

[31] Quang Nhat Vo, Soo Hyung Kim, Hyung Jeong Yang, and Gueesang Lee. An mrf model for binarization of music scores with complex background. *Pattern Recognition Letters*, 69:88–95, 2016.

[32] Quang Nhat Vo, Soo Hyung Kim, Hyung Jeong Yang, and Gueesang Lee. Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74:568–586, 2018.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[34] Yanna Wang, Cunzhao Shi, Baihua Xiao, and Chunheng Wang. Mrf based text binarization in complex images using stroke feature. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 821–825. IEEE, 2015.

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[36] Wei Xiong, Jingjing Xu, Zijie Xiong, Juan Wang, and Min Liu. Degraded historical document image binarization using local features and support vector machine (svm). *Optik*, 164:218–223, 2018.

[37] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.