

IMPROVING DOCUMENT BINARIZATION VIA ADVERSARIAL NOISE-TEXTURE AUGMENTATION

Ankan Kumar Bhunia¹, Ayan Kumar Bhunia^{2*}, Aneeshan Sain³, Partha Pratim Roy⁴

¹Jadavpur University, India ²Nanyang Technological University, Singapore
³Cognizant Technology Solutions, India ⁴Indian Institute of Technology Roorkee, India
²ayanbhunia@ntu.edu.sg

ABSTRACT

Binarization of degraded document images is an elementary step in most of the problems in document image analysis domain. The paper re-visits the binarization problem by introducing an adversarial learning approach. We construct a Texture Augmentation Network that transfers the texture element of a degraded reference document image to a clean binary image. In this way, the network creates multiple versions of the same textual content with various noisy textures, thus enlarging the available document binarization datasets. At last, the newly generated images are passed through a Binarization network to get back the clean version. By jointly training the two networks we can increase the adversarial robustness of our system. Also, it is noteworthy that our model can learn from unpaired data. Experimental results suggest that the proposed method¹ achieves superior performance over widely used DIBCO datasets.

Index Terms— Document image binarization, Adversarial Learning, Augmentation, Style transfer, Unpaired data.

1. INTRODUCTION

Document image binarization is a fundamental problem in the field of Document analysis. Although binarization seems to be quite easy for images of uniform distribution, it can be challenging under real-world scenarios where the document images suffer from various degradations due to aging effect, inadequate maintenance, ink stains, faded ink, bleed-through background, wrinkles, warping effect, non-uniform variation of intensity and lighting conditions during document scanning.

Early document binarization methods [1, 2, 3, 4] for binarization used various thresholding techniques, which include finding a single (or multiple) appropriate threshold(s) for classifying pixels in the image as belonging to foreground or background. Recently, few deep learning frameworks [5,

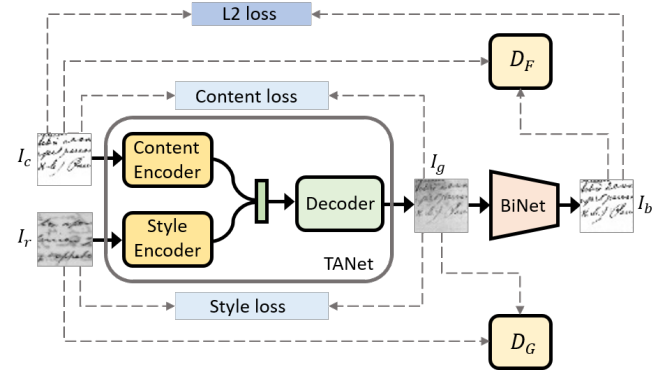


Fig. 1. Illustration of the Proposed framework: It consists of two networks TANet and BiNet. TANet takes a clean document image I_c and a degraded reference image I_r and tries to generate an image I_g with same textual content as I_c retaining the noisy texture of I_r . Next, BiNet tries to get back the clean image by de-noising the generated one.

6] have also been applied for binarization of document images. The objective here is not to predict a threshold but to directly output a binary mask segregating the foreground text and the background noise. These deep learning based models require a considerable amount of paired training data. The publicly available binarization datasets are not sufficient to learn various possible noise distributions (i.e., artifacts, stains, ink spills, etc.) that may occur in real-life situations.

In this paper, we intend to increase the utility of the available bounded datasets by proposing a novel adversarial learning technique. The basic idea is to generate a new set of augmented images of high perceptual quality that combine the semantic content of a clean binary image with the noisy appearance of a degraded document image. Thus, the low-level distribution of visual features in an image is modified while maintaining the semantic content. In this way, we can generate multiple degraded versions of same textual content with various noisy textures. For this purpose we propose a Texture Augmentation Network that superimposes the noisy appearance of the degraded document on the clean binary image. Then, the output image is passed through a Binarization Net-

*Corresponding Author

¹The full source code of the proposed system is publicly available at <https://github.com/ankanbhunia/AdverseBiNet>

work called BiNet to get back the clean version of the document image. Both the networks are jointly trained in an adversarial manner. The goal is to generate harder adversarial training samples with lots of variations. On the other hand, BiNet tries to learn from the hard augmentations for better performance. By jointly training the two networks, we can enhance the adversarial robustness of our binarization model. The advantage of this technique is that the system can learn from unpaired images. It becomes very useful in case of ancient historical documents as it is difficult to get the corresponding binary images for these documents. If the system supports unpaired setting then the data collection process becomes easier. We can easily get large number of unpaired images with less effort by collecting document images and clean images independently from different sources. However, no other previous work on document binarization have tried to utilize the unpaired dataset. In this paper, the proposed system can be trained with unpaired data.

The proposed framework is summarized in Figure 1. The main contributions of our study are as follows: (1) To the best of our knowledge, our work is the first attempt to use a Generative adversarial model in document binarization problem. (2) We propose a texture augmentation network to augment image datasets, by generating adversarial examples online. (3) We employ adversarial learning technique driven by a general GAN objective where the GAN loss plays a complementary role in training TANet and BiNet jointly. Also, it is noteworthy that the method is able to learn from unpaired datasets.

2. METHODOLOGY

In this section, we present the details of our proposed binarization model. The model consists of two networks: Texture Augmentation Network (TANet) and Binarization Network (BiNet). Given a clean document image, TANet tries to obtain a noisy version of that image by transferring the noisy texture of a reference document image comprising of various degradations. On the other hand, the BiNet tries to binarize the newly generated noisy image.

Texture Augmentation Network. To combine the semantic content of a clean document image and the noisy texture of a degraded document image, the first step would be to separate the content and texture representation explicitly. For this purpose, we employ a Content encoder and a Style encoder. Given a clean image I_c and reference noisy element I_r , the encoders learn to extract latent representations R_c and R_r , by leveraging the conditional dependence of the content and texture images. Both encoders have the same configuration with eight convolutional blocks of filter size 5×5 and stride 2. Each convolutional layers are associated with Leaky ReLU and Batch Normalization.

After extracting the content and texture representations, we perform simple concatenation to obtain a mixed representation. Then, it is passed through a Decoder network that

maps the combined representations to an output image that has the same textual content as the clean image and the same texture element as the noisy input. The Decoder architecture is symmetrical to the Encoders with series of deconvolution-BatchNorm-LeakyReLU up-sampling blocks with tanh activation for the final output. The output and the clean image differ in appearance, but they have the same textual element. However, due to the down-sampling process in the content encoder, only a part of the input is stored, resulting in a significant information loss in each layer which can not be used to generate the output image. To deal with this, we adopt skip-connection between the layers of the content encoder and the decoder. We concatenate the feature-map of each down-sampling block in the content encoder with the corresponding feature map of the up-sampling block in the decoder. We represent the TANet as G .

$$I_g = G(I_c, I_r; \theta_G) \quad (1)$$

where, I_g is the generated output and θ_G is the parameters of TANet. The image generated by the TANet should follow some constraints and objectives: (1) It should look real and can not be distinguished from the real-world noisy, degraded document images (2) It has similar texture appearance as the degraded reference document image I_r . (3) It has same textual content as the clean document image I_c . To incorporate above constraints in the training process of TANet, we adopt the following loss functions for TANet.

Adversarial loss. we use an adversarial objective to constrain the output to look similar to the reference document image. Assuming I_r is sampled according to some data distribution \mathcal{P}_r and I_c is sampled from distribution \mathcal{P}_c , the loss is defined as

$$L_G^{GAN}(G, D_G) = \mathbb{E}_{I_r \in \mathcal{P}_r} [\log D_G(I_r)] + \mathbb{E}_{I_c \in \mathcal{P}_c, I_r \in \mathcal{P}_r} [\log(1 - D_G(G(I_c, I_r)))] \quad (2)$$

where the discriminator D_G tries to discriminate between the output image from the degraded reference image.

Style loss. the adversarial loss focuses on getting the overall structure but sometimes it is not enough to capture the fine details of the texture. We use an additional style loss L^S to ensure the successful transfer of texture from the reference image to the clean document image. Following [7, 8], we have used the technique of matching the gram matrices. It captures the correlations between the different feature responses extracted from certain layers of a pre-trained VGG-19 network. Mathematically, gram matrix $\mathcal{G}_{ij}^l \in \mathcal{R}^{N_l \times N_l}$ is the inner product between the vectorised feature maps i and j in layer l :

$$\mathcal{G}_{ij}^l = \sum_k \mathcal{F}_{ik}^l \mathcal{F}_{jk}^l \quad (3)$$

where, N_l is the number of feature maps and \mathcal{F}_{ik}^l is the activation of i^{th} filter at position k in layer l . We use 5 lay-

ers of VGG-19 network ("conv1_1", "conv2_1", "conv3_1", "conv4_1", "conv5_1") to define our style loss.

Content loss. It is required that the generated images have the same textual content as the clean document image. To incorporate the same in our training process, we define a masked mean square loss function. The loss penalizes the differences between the pixels of the content image and output image in the text region only. It can be defined as

$$L^c(G) = \|M \odot I_c - M \odot I_g\|_2 \quad (4)$$

where, M is a binary mask that has value 1 in the text region and 0 in the background. Thus, the total objective function to train TANet can be written as

$$L^{TANet} = L_G^{GAN}(G, D_G) + \lambda_s L^s(G) + \lambda_c L^c(G) \quad (5)$$

where, λ_s and λ_c are the weights to balance the multiple objectives.

Binarization Network. Given that we have generated the noisy version of a clean document image, our system tries to get back the clean binarized image from the generated one through another network called BiNet. The network employs an image-to-image translation framework consisting of a generator and a discriminator. The objective is to train a generator network F that takes the output of TANet I_g and obtains a binarized version of that image I_b .

$$I_b = F(I_g; \theta_F) \quad (6)$$

where, θ_F is the parameters of the network F . A discriminator network D_F is used to determine how good the generator is in generating binarized images. We have used similar network architecture for the generator and the discriminator as mentioned in [9]. During training, both the networks compete against each other in a min-max game. The training objective can be defined as

$$L_F^{GAN}(F, D_F) = \mathbb{E}_{I_c \in \mathcal{P}_c} [\log D_F(I_c)] + \mathbb{E}_{I_g \in \mathcal{P}_g} [\log(1 - D_F(F(I_g)))] \quad (7)$$

It is noted that the training, in this case, follows the "paired" setting. For each input image to the network F , there is corresponding ground truth image. Thus, we can employ full supervision on the predicted binarization results by leveraging the advantage of L_2 pixel loss along with the adversarial loss.

$$L^{L2}(F) = \|I_c - I_b\|_2 \quad (8)$$

The adversarial loss helps to obtain sharper output image by de-noising the noisy input whereas the L_2 loss helps to preserve the content. The final objective of BiNet is

$$L^{BiNet} = L_F^{GAN}(F, D_F) + \lambda_{L2} L^{L2}(F) \quad (9)$$

where, λ_{L2} is a weight parameter. In the next section, we will provide some salient details of the training process and discuss the appropriate weights.

3. EXPERIMENTS

In this section we discuss about the datasets, training details, baseline methods and experimental results regarding the evaluation of our proposed binarization model.

Datasets. For training and evaluating our model, we have used some publicly available document datasets. A total of 9 datasets are used in this work: DIBCO 2009 [10], DIBCO 2011 [11], DIBCO 2013 [12], H-DIBCO 2010 [13], H-DIBCO 2012 [14], H-DIBCO 2014 [15], Bickley diary [16], PHIDB [17], and S-MS [18] 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. At first, we convert the images from these datasets to patches of size 256×256 . To increase the number of patches, we augment the training patches by rotating with an angle of 90, 180, or 270. A small part (10%) of the obtained image patches is used as an evaluation set, and the rest of the images are used to train the model. In the training set, we have two set of images patches: degraded document set and as well as their binarized ground truths. The clean images are sampled from the binarized set, and reference images are sampled from the degraded document set in an unpaired manner.

Training. To train the model, we follow a particular stage-wise training protocol. At first, TANet is trained for 10 epochs. After the training, it should be able to generate noisy version of the clean images. In the next stage, BiNet is trained using the generated noisy images for another 10 epochs. At last, TANet and BiNet are fine-tuned together for around 30 epochs. We note that during the couple training, TANet tends to generate more challenging adversarial samples that are relatively hard to be detected by BiNet. This training strategy enforces the model to learn a various type of degradations including noises and artifacts. Figure 3 illustrates the texture transfer process qualitatively. At the time of testing, BiNet is used to get the binarized output of a given document image. Experiments are conducted on a server with 12 GB memory and single Nvidia Tesla K80 GPU. The model is implemented using TensorFlow library. Adam optimizer with learning rate 0.0001 is used to train the model. We take $\lambda_s = 0.5$, $\lambda_c = 10$ and $\lambda_{L2} = 100$ throughout the experiments. We used the following metrics to quantitatively measure the performance of our proposed model with those of the state-of-the-art algorithms and some baselines: F-measure, pseudo-F-measure (F_{ps}), Distance reciprocal distortion metric (DRD), and the peak signal-to-noise ratio (PSNR) metrics [13].

Baselines: We define following baselines.

U-Net: It is a trivial encoder-decoder network with skip-connections [19]. We take its same architecture as our generator unit of BiNet. It is noted that the network is trained in

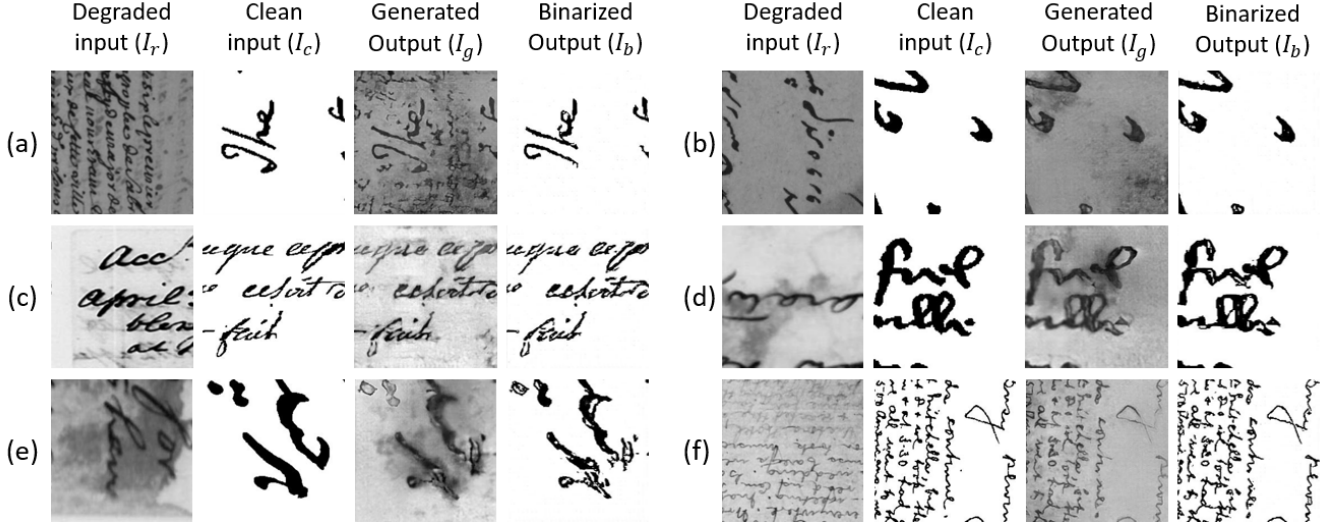


Fig. 2. Qualitative evaluation of texture transfer process and binarization technique on the evaluation set.

Methods	F-measure	F_{ps}	DRD	PSNR
BERN [22]	52.6	52.8	62.2	10.1
Sauvola [2]	85.0	89.8	7.6	16.9
Niblack [23]	72.8	72.2	24.9	13.6
Gatos [24]	83.4	87.0	9.5	17.1
Otsu [1]	83.9	86.5	11.0	16.6
Su [25]	87.7	88.3	4.2	19.6
Howe [26]	91.3	91.7	3.2	21.3
DSN [5]	94.4	96.0	1.8	21.4
U-Net [19]	89.6	91.8	5.9	17.6
Pix2pix [20]	94.8	97.0	2.7	20.8
CycleGAN [21]	66.8	70.1	17.6	12.5
Ours	97.8	98.7	1.1	24.3

Table 1. Quantitative results on DIBCO 2013 dataset

a paired setting. For each input image, there is corresponding ground truth. L2 pixel loss is used to train the complete model.

Pix2pix: It is an image-to-image translation framework inspired from [20]. The network resembles the BiNet part of our system. It is trained using adversarial loss and L2 loss using paired data.

CycleGAN: We employ this baseline by using the concept of cycle-consistent image translation frameworks [21]. The network utilizes unpaired data to train the model.

Results: We compare our proposed system with the baseline methods and some state-of-the-art binarization algorithms in Table 3. Some qualitative results are shown in Figure 3. From Table 3, we can see that our proposed method delivers the best quality results regarding all the four evaluation metrics. Also, we have obtained a low DRD score which implies that our method is also superior regarding the visual distortion. U-Net and Pix2pix work moderately, but

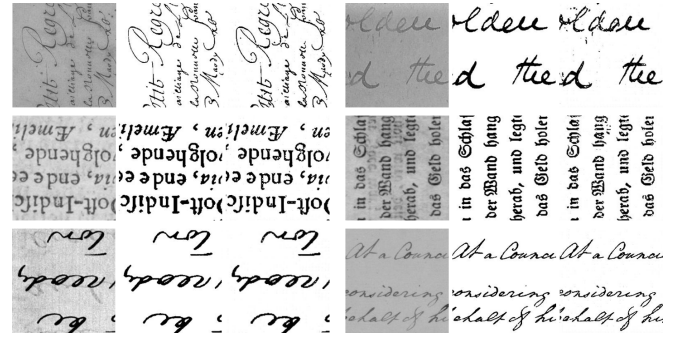


Fig. 3. Binarization results using the trained BiNet model on test set. Images are in the order: input, predicted and ground truth from left to right for each sample result.

CycleGAN obtains poor results as compared to others. Similar to CycleGAN our method also utilizes unpaired data, but the main binarization network (BiNet) of our model learns from paired samples which are created internally in our system. Thus, we can impose full supervision in the BiNet part that helps to generate high-quality results. However, in CycleGAN method, there is no scope to impose the full supervision.

4. CONCLUSION

In this paper, we re-visited the problem of document binarization by introducing a new adversarial learning technique that intends to increase the utility of the available bounded datasets. The noisy data augmentation is an integral part of our network that enforces the model to learn robust representation of various types of document degradations from unpaired data. Also, the experimental results suggest that our method is superior to existing state-of-the-art frameworks.

5. REFERENCES

- [1] 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.
- [2] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.
- [3] N. Phansalkar, S. More, A. Sabale, and M. Joshi, "Adaptive local thresholding for detection of nuclei in diversity stained cytology images," in *ICCSPP*. IEEE, 2011, pp. 218–220.
- [4] B. Gatos, I. Pratikakis, and S.J. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in *ICPR*. IEEE, 2008, pp. 1–4.
- [5] 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.
- [6] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *ICDAR*. IEEE, 2017, vol. 1, pp. 99–104.
- [7] L. Gatys, A.S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *NIPS*, 2015, pp. 262–270.
- [8] L.A. Gatys, A.S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016, pp. 2414–2423.
- [9] A. Konwer, A.K. Bhunia, A. Bhowmick, A.K. Bhunia, P. Banerjee, P.P. Roy, and U. Pal, "Staff line removal using generative adversarial networks," *arXiv preprint arXiv:1801.07141*, 2018.
- [10] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *ICDAR*. IEEE, 2009, pp. 1375–1382.
- [11] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "Icdar 2011 document image binarization contest (dibco 2011)," in *ICDAR*. IEEE, 2011, pp. 1506–1510.
- [12] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "Icdar 2013 document image binarization contest (dibco 2013)," in *ICDAR*. IEEE, 2013, pp. 1471–1476.
- [13] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-dibco 2010-handwritten document image binarization competition," in *ICFHR*. IEEE, 2010, pp. 727–732.
- [14] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012)," in *ICFHR*. IEEE, 2012, pp. 817–822.
- [15] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Icfhr2014 competition on handwritten document image binarization (h-dibco 2014)," in *ICFHR*. IEEE, 2014, pp. 809–813.
- [16] F. Deng, Z. Wu, Z. Lu, and M.S. Brown, "Binarization-shop: 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.
- [17] H.Z. Nafchi, S.M. Ayatollahi, R.F. Moghaddam, and M. Cheriet, "An efficient ground truthing tool for binarization of historical manuscripts," in *ICDAR*. IEEE, 2013, pp. 807–811.
- [18] R. Hedjam, H.Z. Nafchi, R.F. Moghaddam, M. Kalacska, and M. Cheriet, "Icdar 2015 contest on multispectral text extraction (ms-tex 2015)," in *ICDAR*. IEEE, 2015, pp. 1181–1185.
- [19] O. Ronneberger, P. Fischer, and 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.
- [20] P. Isola, J. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.
- [21] J. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.
- [22] J. Bernsen, "Dynamic thresholding of gray-level images," in *Proc. Eighth Int'l conf. Pattern Recognition, Paris, 1986*, 1986.
- [23] W. Niblack, *An introduction to digital image processing*, vol. 34, Prentice-Hall Englewood Cliffs, 1986.
- [24] B. Gatos, I. Pratikakis, and S.J. Perantonis, "An adaptive binarization technique for low quality historical documents," in *International Workshop on Document Analysis Systems*. Springer, 2004, pp. 102–113.
- [25] 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.
- [26] 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.