UDBNET: Unsupervised Document Binarization Network *via* Adversarial Game

Amandeep Kumar[1*] Shuvozit Ghose[2*] Pinaki Nath Chowdhury[3] Partha Pratim Roy[4] Umapada Pal[5]
[1]Techno Main Salt Lake, Sector V, Kolkata, India. [2]Institute of Engineering and Management Kolkata, India .
[4]Indian Institute of Technology Roorkee, India . [3,5]Indian Statistical Institute Kolkata, India .
kumar.amandeep015@gmail.com[1], shuvozit.ghose@gmail.com[2]

Abstract— **Degraded document image binarization is one of the most challenging tasks in the domain of document image analysis. In this paper, we present a novel approach towards document image binarization by introducing three-player min-max adversarial game. We train the network in an unsupervised setup by assuming that we do not have any paired-training data. In our approach, an Adversarial Texture Augmentation Network (ATANet) first superimposes the texture of a degraded reference image over a clean image. Later, the clean image along with its generated degraded version constitute the pseudo paired-data which is used to train the Unsupervised Document Binarization Network (UDBNet). Following this approach, we have enlarged the document binarization datasets as it generates multiple images having same content feature but different textual feature. These generated noisy images are then fed into the UDBNet to get back the clean version. The joint discriminator which is the third-player of our three-player min-max adversarial game tries to couple both the ATANet and UDBNet. The three-player min-max adversarial game stops, when the distributions modelled by the ATANet and the UDBNet align to the same joint distribution over time. Thus, the joint discriminator enforces the UDBNet to perform better on real degraded image. The experimental results indicate the superior performance of the proposed model over existing state-of-the-art algorithm on widely used DIBCO datasets. The source code of the proposed system is publicly available at https://github.com/VIROBO-15/UDBNET.**

I. Introduction

Document image binarization is a rudimentary problem in the field of Document analysis. Binarization itself, is the prepossessing backbone of many document image processing systems (DIPSs) [1], [2]. The performance of the high level processing tasks, such as image segmentation [3], word recognition [4], [5], optical character recognition (OCR) [6], and document layout analysis (DLA) [7] is greatly dependent on the success of the binarization task. Technically, document image binarization is the technique of converting color document images or gray-level images into a binary representation, where the main objective is to classify each pixel as foreground(text/ink) or background(parchment/paper). In other words, it is the process of discarding the unnecessary noisy information while preserving the meaningful visual information.

Document image binarization can be considered as an easy task for images of uniform distribution. However, in real-world scenarios under significant image noise and uneven background, binarization is a quite challenging problem. Moreover, the document images suffer from various degradation due to faint characters, bleed-through background, clutter and artifacts, dark patches, creases, faded ink, non-uniform variation of intensity, inadequate maintenance, aging effect, ink stains, lighting conditions, warping effect during acquisition etc. Faded ink creates difficulty during distinguishing light text from background. Bleed through occurs when content from the back of a page becomes visible or ‘leaks’ through. It creates difficulty in labeling foreground and background during binarization process as it can misinterpret background as foreground. Uneven illumination happens when the image is suffering from shadow effect or inconsistent lighting during acquisition. In addition to the above, dark patches are quite difficult to remove for various reasons. Firstly, these patches are of varying sizes and intensities. Secondly, they appear as stains of arbitrary shapes. Thirdly, they are often present in areas containing characters. Therefore, the study on binarization for document images, specially in the context of degraded images, is highly essential.

In general, binarization methods [8] [9] [10] works for supervised setup. In the supervised setup, we need ground-truth binarized image along with the degraded image. But, it is difficult to get the corresponding ground truth binary image in many scenarios like in case of historical document image. To address these drawbacks, Bhunia *et al.* [11] first attempts to introduce unsupervised setup in the domain of document image binarization. For this purpose, they employ Texture Augmentation Network (TANet) that superimposes the noisy appearance of the degraded document on the clean binary image to generate multiple degraded image of same textual content with various noisy textures and later utilize Binarization Network (BiNet) to get back the clean version of the document image. Although this method has shown better results over the previous state-of-the-art methods, it has several limitations. Firstly, the TANet is completely unaware about the content at which it is conditioned on. Thus, the corresponding discriminator can not verify if the content of the generated noisy image remain consistent or not. Secondly, there exist no performance quantifier that validates the performance of the BiNet on real degraded noisy image. Finally, the Binarization Network (BiNet) has dataset bias towards generated noisy images. But, to adddress the dataset bias, BiNet does not use any kind of formulation or other techniques. In our observation, these limitations are due to the fact that the TANet

[*] Authors contributed equally

and BiNet both employ straight-forward two-player Generative Adversarial Network (GAN) [12] objectives and model two different uncorrelated conditional distributions. In this paper, we address these limitations by introducing adversarial min-max game in the domain of unsupervised document image binarization. Similar to the TANet and BiNet, we propose Adversarial Texture Augmentation Network (ATANet) and Unsupervised Documenet Binarization Network (UDBNet) which utilize three-player GAN objectives. The proposed third player is a joint discriminator tries to couple both the Adversarial Texture Augmentation Network (ATANet) and Unsupervised Document Binarization Network (UDBNet). Our three-player min-max adversarial game comes to an end, when the distribution modelled by the Adversarial Texture Augmentation Network (ATANet) and the Unsupervised Document Binarization Network (UDBNet) align to the same joint distribution over time. Therefore, the contributions of this paper are as follows:

- To be our best of knowledge, we are the first one to introduce adversarial game in the domain of document image binarization by proposing Adversarial Texture Augmentation Network (ATANet) and Unsupervised Document Binarization Network (UDBNet).
- We introduce a joint discriminator which tries to couple the ATANet and UDBNet so that it can tackle the dataset bias problem and perform well on the real degraded document image.
- Our approach shows a superior performance on widely used DIBCO datasets as compared to the existing state-of-the-arts methods.

The remaining of the paper is organized as follows: In section II, we discuss about the related works in the field of document image binarization. In section III, we describe the proposed framework. The datasets, implementation, baselines methods and performance analysis are discussed in section IV. Section V concludes the paper.

II. Related Works

Document image binarization is a classical research problem in computer-aided document analysis and has been studied extensively over the past few decades. Document image binarization aims at converting the document image into either foreground text or background. The most simple and widely used approach is thresholding, which sets the pixels under a threshold value to 0 and the rest to 1. Thresholding methods are primarily of three types : global, local and hybrid. In high quality images, global algorithms can effectively estimate a threshold based on the entire image. Global thresholds can be calculated using gray level histogram [3], circular statistics [13], error minimization [14], histogram entropy [15] and moment preserving principle [16]. Clustering models [17] can also learn mappings in an unsupervised manner based on global features to separate background and foreground. However performance degrades when they are applied to images having variations in background due to illumination, occlusion or degradation. For such cases, local adaptive methods perform better. Some of the common local thresholding approaches can be seen in the works of Bernsen *et al.* [18], Niblack *et al.* [19], Sauvota *et al.* [20]. In the work by Niblack *et al.* [19], a major drawback is that if the foreground text is sparse, a lot of background noise will remain in the binary image. Sauvota *et al.* [20] alleviates this by assuming the foreground pixels to be closer to background ones. On a similar note, Wolfe *et al.* [21] normalizes the contrast and the mean gray level of the neighbourhood to modify the threshold.

Apart from threshold based techniques, non-threshold based strategies have also been studied extensively in literature. Some notable approaches include Markov Random Field (MRF) modeling of an input image, which minimizes a cost function by regarding the target binarized image as a binary MRF. Howe *et al.* [22] proposed an algorithm where they defined the cost function based on combination of the Laplacian energy of image intensity for computing local likelihood of foreground-background pixels and Canny edge detection for detecting discontinuities. The cost function is minimized by graph cut computation. Howes method [22] is efficient and yields good results, however it is parameter dependent. Howe *et al.* [23] improvised on this method by adaptive tuning of two parameters to yield better performance. Howes technique formed the basis of the first winning algorithm proposed by Kliger and Tal in the DIBCO 2016 competition [24]. They combined Howes algorithm with a novel pre-processing step based on linear transformation of the image onto a spherical surface where concavities correspond to foreground in the original image.The concavities are estimated using the Hidden Point Removal Operator [25] which outputs a probability of a pixel belonging to a concavity.

All these proposed techniques perform well in the context they are applied to. But these methods fail to generalize in the context of binarizing any kind of document subjected to a varied degree of illumination, background noise and degradation.Recently pixel-wise binarization approaches have been proposed in literature where each pixel is classified as text or background. Pastor-Pellicer [26] proposed a CNN framework consisting of two groups of convolution layers and a fully connected layer. Each pixel is classified into text or background by using a sliding window centred at the classified pixel. Such an approach has also been used in binarizing musical documents by Calvo-Zaragoza *et al.* [27] . These pixel wise classification techniques have shown good performance, however their most conspicuous drawbacks include being computationally very expensive since they involve labelling each pixel in the document image and classifying each pixel independently without exploiting contextual information in any pixels neighbourhood. To incorporate this contextual information, Afzal *et al.* [28] propose a pixel wise classification method where they formulate the binarization procedure as a sequence learning problem. They use a 2D LSTM model which takes in a 2D sequence of pixels as input and classifies each pixel as foreground or background. This achieved better results but still suffered from huge computational complexity. To alleviate this, Tensmeyer *et al.* [29] proposed a novel multi-scale fully convolutional network for document image binarization. Recently Calvo-Zaragoz *et al.* [30] proposed a fully convolutional-selectional auto encoder model that has been trained to learn a patch-wise mapping of the document image to its corresponding binarized version. This performs a fine-grained categorization in which each pixel gets a different activation value depending on whether the target label of the pixel is text or background. Other approaches involving convolution networks include the winning algorithm of DIBCO 2017 competition [31], where the winning team used a U-Net encoder decoder architecture for accurate pixel classification. Vo *et al.* [8] introduced a hierarchical deep supervised network for document binarization which achieves state of the art performance on several benchmark datatsets. Westphal *et al.* [9] proposed a Grid LSTM network for binarization, yet it achieves lesser performance than Vos method [8]. To learn the document degradation, He *et al.* [10] proposed an iterative fine tuning technique to learn the mappings from a degraded input document image to the expected clean and uniform images followed by a classifier to output the binarized image.

In case of unsupervised image-to-image translation task, one of the first major works that uses deep network is CycleGAN [32]. Following this work, there have been numerous attempts to design unsupervised or semi-supervised framework for different computer vision tasks like depth estimation [33], image captioning [34], [35] etc. Most of these works use popular cycle-consistency loss to learn an unsupervised mapping between two different domains. In contrast to all such works, Bhunia *et al.* [11] employ Texture Augmentation Network (TANet) that superimposes the noisy appearance of the degraded document on the clean binary image to generate multiple degraded image of same textual content with various noisy textures and later utilize Binarization Network (BiNet) to get back the clean version of the document image.

III. Proposed Framework

In this section. we first briefly present the binarization model proposed by Bhunia *et al.* [11] as base model. Next, we describe the limitations of the base model. Finally, we introduce our novel unsupervised adversarial game and describe how Adversarial Texture Augmentation Network (ATANet) and Unsupervised Document Binarization Network (UDBNet) address these limitations efficiently.

A. Background: Base Models

The base model consists of two networks: Texture Augmentation Network(TANet) and Binarization Network(BiNet). Let C denotes the binarized clean image sampled from marginal distribution $P(C)$ and D denotes a degraded document image sampled from marginal distribution $P(D)$. TANet tries to model $P(D|C)$, i.e., given a clean image, it tries to generate a degraded version of it keeping the content same. On the other hand, BiNet tries to model $P(C|D)$, i.e., given a degraded image, it tries to generate the clean image. During inference, only BiNet is used.

Texture Augmentation Network : The TANet exploit a two-player GAN which consists of a generator network and a discriminator network. Conditional distribution $P(D|C)$ is approximately modelled by $P(G|C, D) \approx P(D|C)$, where the noisy texture of degraded image D is superimposed on the clean image C to generate noisy version of the clean image as G. Note that, the content of G and C remains similar and we do not use any paired-data here. On the other side, the discriminator tries to discriminate between output image G and degraded reference image D. The generator of the TANet use a content encoder and a style encoder to encode the semantic content of the C image and noisy texture of the D image explicitly. Next, the two encoded features are concatenated to obtain a mixed feature representation. Finally, this mixed representation is passed through a decoder network that outputs noisy generated image G. To ensure that the generated image G contains the same textual content as clean image C and the same texture element of the degraded image D, The TANet utilizes the following loss functions:

Adversarial loss: The objective of the adversarial loss is to constrain the output to make it similar to the degraded reference image D. The adversarial loss is defined as:

$$\begin{split}\mathcal{L}_T^{GAN}(T, D_T) = \mathbb{E}_{D\sim P(D)}[\log D_T(D)]+\\ \mathbb{E}_{C\sim P(C),D\sim P(D)}[\log(1 - D_T(T(C, D))]\end{split} \tag{1}$$

Where, the discriminator D_T tries to discriminate between the output image G from the degraded reference image D.

Style loss: While adversarial loss focuses on getting the overall structure of the generated image, an additional style loss $\mathcal{L}^s(T)$ ensures successful transfer of texture content from degraded reference image D to the input binarized clean image C. For this purpose, Gram matrices [36], [37] is used in *"conv1_1", "conv2_1","conv3_1","conv4_1","conv5_1"* layers of the encoder networks. Mathematically,

$$\mathcal{G}_{ij}^l = \sum_k \mathcal{F}_{ik}^l \mathcal{F}_{jk}^l \tag{2}$$

Where, F_{ik}^l is the activation of i^{th} filter at position k in layer l, gram matrix $\mathcal{G}_{ij}^l \in \mathbb{R}^{N_l \times N_l}$ is the inner product between vectorised feature maps i and j in layer l and N_l is the number of feature maps.

Content loss: To ensure the generated image G contains the same textual content as the clean binarized image C, a content loss is defined as follows:

$$\mathcal{L}^c(T) = ||M \odot C - M \odot G||_2 \tag{3}$$

Here, M denotes a binary mask that has value 0 in the background and 1 in the text region.

The overall objective of the TANet is defined as follows:

$$\mathcal{L}^{TANet} = \mathcal{L}_T^{GAN}(T, D_T) + \lambda_s \mathcal{L}^s(T) + \lambda_c \mathcal{L}^c(T) \tag{4}$$

Where, λ_s and λ_c are the tunable hyper-parameters to balance multiple objectives.

Binarization Network : Similar to TANet, BiNet exploits two-player GAN and employs an image-to-image translation framework consisting of a generator and a discriminator. While the generator of the BiNet tries to model $P(B|G) \approx P(D|C)$,

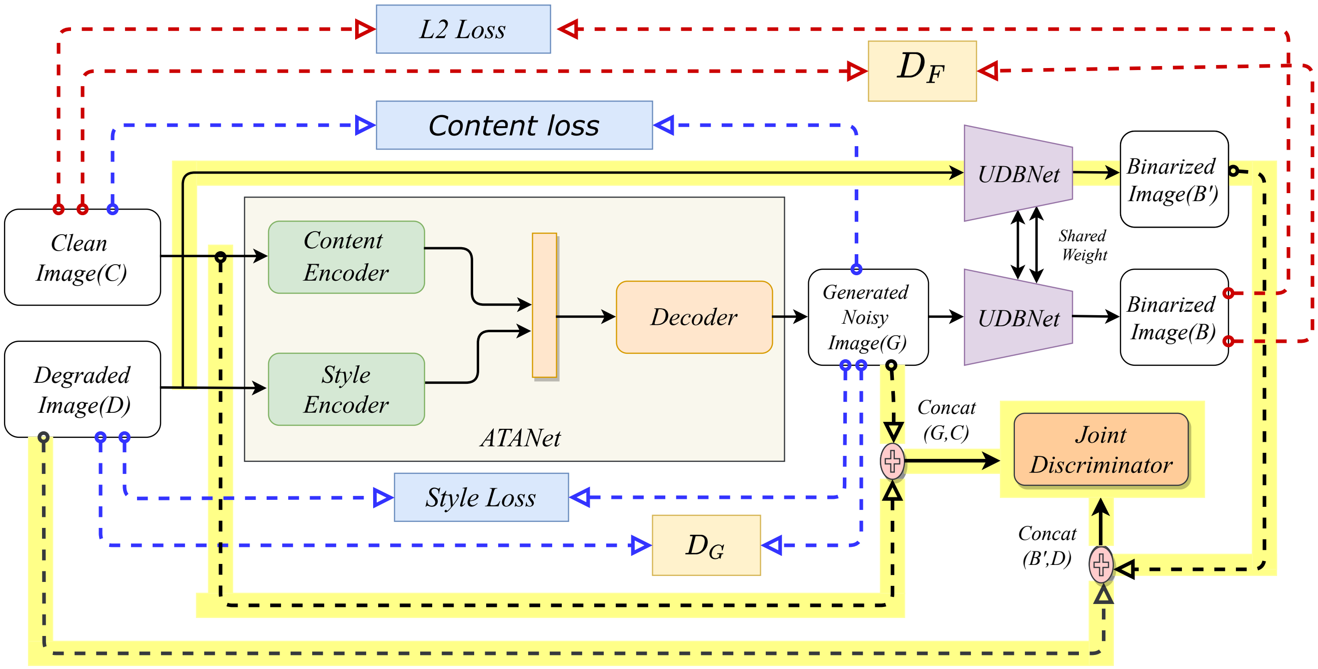

Fig. 1. Illustration of Our proposed Framework. The yellow highlighted region highlights our contribution over Bhunia *et al.* [11]. We have two networks: ATANet which takes Clean image C and Degraded image D as inputs and generates noisy degraded image G. On the other hand, UDBNet tries to get back the clean image from the generated noisy G. Thus, (C, G) acts as pseudo paired data to train the binarization network. Also, We feed degraded image D as an input to UDBNet and get the corresponding binarized image(B'). Then, we concat the image pairs (G, C) and (B', D) and feed into joint discriminator to couple both ATANet and UDBNet. This enforces the UDBNet to generalize better for real degraded images

where B is the binarized clean image of the newly generated noisy image, the discriminator determines how good the generator is in generating binarized images. The adversarial loss of the BiNet is:

$$\begin{aligned}\mathcal{L}_F^{GAN}(F, D_F) = \mathbb{E}_{C\sim P(C)}[logD_F(C)]+ \\ \mathbb{E}_{G\sim P(D|C)}[\log(1 - D_F(F(G)))]\end{aligned} \tag{5}$$

In times of training, for each input image G, there is corresponding ground truth image C. Thus, an additional L_2 loss is utilized to fully supervised the predicted binarization results along with the adversarial loss:

$$\mathcal{L}^{L2} = ||C - B||_2 \tag{6}$$

While the L_2 pixel loss helps to preserve the content, the adversarial loss guides to obtain sharper output image B by de-noising input noisy image G.

The overall objective of the BiNet is as follows:

$$\mathcal{L}^{BINet} = \mathcal{L}_F^{GAN}(F, D_F) + \lambda_{L2}\mathcal{L}^{L2}(F) \tag{7}$$

Where, λ_{L2} is a tunable hyper-parameter.

B. Limitations of Base Model

The limitations of the base model are given below:

Limitation 1. Although the texture augmentation network (TANet) tries to model $P(D|C)$ in the base model and generates noisy images, but it is completely unaware about the content at which it is conditioned on. Thus, the corresponding discriminator can not verify if the content of the generated noisy image remains consistent or not.

Limitation 2. For unpaired real degraded noisy image, the corresponding ground truth or binarized clean image is absent. The absence of ground truth image limits the scope of binarization network (BiNet) of the base model. Firstly, the BiNet can not be trained with real degraded noisy image as L_2 loss can not be utilized. Secondly, since the TANet and the BiNet model two different uncorrelated conditional distribution and are trained separately, these models are prone to overfitting. Finally, there exists no performance quantifier that validates the performance of the BiNet on real degraded noisy image.

Limitation 3. There exist a gap between generated noisy image distribution and real degraded noisy image distribution. As the Binarization network (BiNet) is completely trained on generated noisy image, the Binet has dataset bias towards generated noisy images. A model trained in the generated data can hardly perform well on the real data. This problem is quite similar to domain-shift [38] problem. But to minimize the generated-real domain shift in the context of document image binarization, the base model does not use any kind of formulation or other techniques.

C. Adversarial Game

In our unsupervised setup, we do not have any paired training data. Thus, we do not have any access to real joint distribution of clean and degraded image, $P_{real}(C, D)$. However, we can approximate this real distribution by texture augmentation network and binarization network. The joint distribution $P(C, D)$ can be factorized in two ways, namely

$$P(C, D) = \underbrace{P(D|C) * P(C)}_{P_T} = \underbrace{P(C|D) * P(D)}_{P_B} \tag{8}$$

Please note that, $P(D|C)$ is modelled by texture augmentation network and $P(C|D)$ is modelled by binarization network. $P_T(C, D)$ is obtained when the Generated noisy image G from Adversarial Texture Augmentation network is concatenated with the corresponding input clean image C. On the other side, when we pass a real degraded image D through binarization network to get a corresponding clean image B', we constitute $P_B(C, D)$ by concatenating them together. Thus, texture augmentation network plays role in modeling $P_T(C, D)$ and binarization network plays role in modeling $P_B(C, D)$. $P_T(C, D)$ and $P_B(C, D)$ are both approximated joint-distribution of real and degraded images. If we can properly align these two approximated joint distribution $P_B(C, D)$ and $P_T(C, D)$ together, it will closely get aligned with the real joint-distribution. In order to align $P_B(C, D)$ and $P_T(C, D)$, we propose joint discriminator that distinguishes whether a input sample is from the distribution $P_T(C, D)$ or the $P_B(C, D)$. The Adversarial Texture Augmentation Network(ATANet) and Unsupervised Document Binarization(UDBNet) network objective is to fool the discriminator such that it cannot distinguishes whether the input sample is from $P_T(C, D)$ or $P_B(C, D)$. Thus, the distributions of P_T and P_B gets aligned overtime.

Adversarial Texture Augmentation Network: The ATANet Consists of three components: 1) a generator $\mathbf{T}$ that characterizes the conditional distribution $P_T(G|C, D)$ and generates noisy image G; 2) a discriminator $\mathbf{D_T}$ that discriminates the output image G from the degraded reference image D; 3) a joint discriminator $\mathbf{J_D}$ that distinguishes whether a pair of data (G, C) comes from $P_T(C, D)$ or $P_B(C, D)$.

Similar to our base model, the generator consists of content encoder and the style encoder in which we pass the clean image C and the degraded reference image D as the input, respectively. The latent representations after the encoding the images are simply concatenated and feed into the decoder. The architecture of decoder is symmetrical to the encoder, having the skip connection between the layers of content encoder and decoder as similar to our base model. The discriminator $\mathbf{D_T}$ tries to discriminate between the generated image G from the generator and the degraded reference image D. The pseudo generated-clean image pair (G, C) are fed into the joint discriminator $\mathbf{J_D}$ such that our joint discriminator tries to distinguish whether the input sample (G, C) is from distribution $P_T(C, D)$ or $P_B(C, D)$.

In this game, let a clean-degraded image pair (C, D) is sampled from distributions $P(C)$and $P(D)$, generator $\mathbf{T}$ produces a pseudo generated noisy image G given C following the conditional distribution $P_T(G|C)$. Hence, the pseudo clean-generated image pair is a sample from the joint distribution $P(C, D) = P(C)P_T(G|C)$.

To get the real world noisy, degraded document image having the textual appearance similar to degraded reference image D and textual content similar to the clean image C. we define adversarial loss of our ATANet as:

$$\begin{aligned} \underset{\mathbf{D_T}}{min}\,\underset{\mathbf{T},\mathbf{J_D}}{max}\,\mathcal{L}_T^{Adv}(\mathbf{D_T}, \mathbf{T}, \mathbf{J_D}) = \mathbb{E}_{D\sim P_D}[\log \mathbf{D_T}(D)] + \\ \mathbb{E}_{(C)\sim P(C),(D)\sim P(D)}[\log(1 - \mathbf{D_T}(\mathbf{T}(C, D))] + \\ \mathbb{E}_{(C)\sim P(C),(D)\sim P(D)}[\log(1 - \mathbf{J_D}(\mathbf{T}(C, D), C)] + \\ \mathbb{E}_{(D)\sim P(D)}[\log(\mathbf{J_D}(\mathbf{F}(D), D)] \end{aligned} \tag{9}$$

Where, F is Unsupervised Document Binarization Network. The adversarial loss is trained with flip flop fashion. The game will end when distribution $P_T(C, D)$ and distribution $P_B(C, D)$ will be in equilibrium and gets aligned over time.

Therefore, the overall objective of the ATANet is defined as:

$$\mathcal{L}^{ATANet} = \mathcal{L}_T^{Adv}(D_T, T, J_T) + \lambda_s \mathcal{L}^s(T) + \lambda_c \mathcal{L}^c(T) \tag{10}$$

Where, $\mathcal{L}^s(T)$ and $\mathcal{L}^c(T)$ are style loss and content loss similar to our base mode, λ_s and λ_c are the tunable hyper-parameters to balance multiple objectives.

Unsupervised Document Binarization Network: Similiar to ATANet, UDBNet consists of three components: 1) a generator $\mathbf{F}$ that characterizes the conditional distribution $P_B(B|G)$ and $P_B(B'|D)$ generates binarized clean image B and B' corresponding to G and D respectively; 2) a discriminator $\mathbf{D_F}$ determines how good the generator is in generating binarized images B; 3) a joint discriminator $\mathbf{J_D}$ that distinguishes whether a pair of data (B', D) comes from distribution $P_B(C, D)$ or $P_T(C, D)$.

We have used similar network architecture for the generator and the discriminator as in the base model. The generated noisy image from the ATANet G is fed into the generator of the UDBNet and generates the binarized image B. The discriminator tries to discriminate between the binarized image B and the original clean image C. The pseudo clean-degraded image pair (B', D) are fed into the joint discriminator $\mathbf{J_D}$ such that our joint discriminator tries to distinguish whether the input sample (B', D) is from distribution $P_T(C, D)$ or $P_B(C, D)$.

In this game, let a clean-degraded image pair (B', D) is sampled from distributions $P(C|D)$ and $P(D)$, generator $\mathbf{F}$ produces a pseudo binarized clean image B' given D following the conditional distribution $P_B(B'|D)$. Hence, the pseudo degraded-clean image pair is a sample from the joint distribution $P(C, D) = P(D)P_B(B'|D)$.

To attain the proper binarized image. We define adversarial loss of our UDBNet as:

$$\begin{aligned} \underset{\mathbf{D_F}}{min}\,\underset{\mathbf{F},\mathbf{J_D}}{max}\,\mathcal{L}_F^{Adv}(\mathbf{D_F}, \mathbf{F}, \mathbf{J_F}) = \mathbb{E}_{C\sim P_C}[\log \mathbf{D_F}(C)] + \\ \mathbb{E}_{G\sim P(D|C)}[\log(1 - \mathbf{D_F}(\mathbf{F}(G))] + \\ \mathbb{E}_{(D)\sim P(D)}[\log(1 - \mathbf{J_D}(\mathbf{F}(D), D)] + \\ \mathbb{E}_{(C)\sim P(C),(D)\sim P(D)}[\log(\mathbf{J_D}(\mathbf{T}(C, D), C)] \end{aligned} \tag{11}$$

Where, T is an adversarial texture augmentation network. The adversarial loss is trained with flip flop fashion. Therefore, the overall objective of the UDBNet is defined as:

$$\mathcal{L}^{UDBNet} = \mathcal{L}_T^{Adv}(D_T, T, J_T) + \lambda_{L2}\mathcal{L}^{L2}(F) \tag{12}$$

Where, $\mathcal{L}^{L2}(F)$ is L_2 loss, λ_{L2} is a tunable hyper-parameter.

D. Training Joint Discriminator via Flipped Label

Similar to the Texture Augmentation network (TANet) of the base model, we feed our clean image C and degraded image D into the content encoder and style encoder, respectively. The encoded representations are simply concatenated and passed through the decoder which outputs the noisy version of clean image G. The degraded reference image D is then fed into the generator of Unsupervised Document Binarization Network (UDBNet), which generates binarized version of the image B'. To feed into the joint discriminator $\mathbf{J_D}$, we perform simple concatenation of the input pairs (G, C) and (B', D). Both the Adversarial Texture Augmentation Network (ATANet) and Unsupervised Document Binarization Network (UDBNet) tries to fool the joint discriminator such that it cannot discriminate whether the input sample is from distribution $P_T(C, D)$ or the distribution $P_B(C, D)$. The joint discriminator $\mathbf{J_D}$ that distinguishes whether a pair of data (G, C) and (B', D) come from distribution $P_T(C, D)$ or $P_B(C, D)$. This enforces the UDBNet to generalize better for real degraded images although we are not using any paired-training data. The joint discriminator is trained using the flipped labels as utilized in [39].

IV. Experiment

A. Datasets

The experiments are conducted on the publicly available DIBCO datasets [40]. We train our model on DIBCO 2009 [41], DIBCO 2013 [42], H-DIBCO 2012 [43] and H-DIBCO 2014 [44] datasets. On the other hand, Challenging historical dataset like H-DIBCO 2016 [24] and DIBCO 2011 [45] are selected for evaluation purposes. we resizes the images from these datasets to patches of size 256 × 256 before feeding to our model. To evaluate the performance of our methods, we adopt four evaluation metrics. They are F-measure, pseudo F-measure (F_{ps}), distance reciprocal distortion metric (DRD), and the peak signal-to-noise ratio (PSNR). Similar to [11], we augment the training patches by rotating with an angle of 90, 180 and 270 degrees.

B. Implementation Details

We have implemented the entire model in Pytorch [46] and the experiments were done on a server having Nvidia Titan X GPU with 12 GB of memory. We have adapted step-wise training protocol for training our model. At first, we train ATANet for 15 epochs and generates the noisy version of the clean images. After that, we freeze the network such that the weights of the network does not alter. Next, we train UDBNet on generated noisy images for 20 epochs to generate its corresponding binary clean image. Then, We unfreeze the ATANet network. In next stage, we jointly train both the networks along with the joint discriminator for around 10 epochs in the flip flop fashion. At last, we fine-tune the model for 30 epochs. During the couple training, ATANet tries to generate more challenging adversarial samples that are used as a pseudo image pair for training the UDBNet. Thus, the training procedure helps the model to learn various degradation including aging effects, noises etc. During training, we use Adam optimizer with the learning rate of 0.0001. We take λ_S = 0.5, λ_C =10 and λ_{L2} = 100 throughout the experiment.

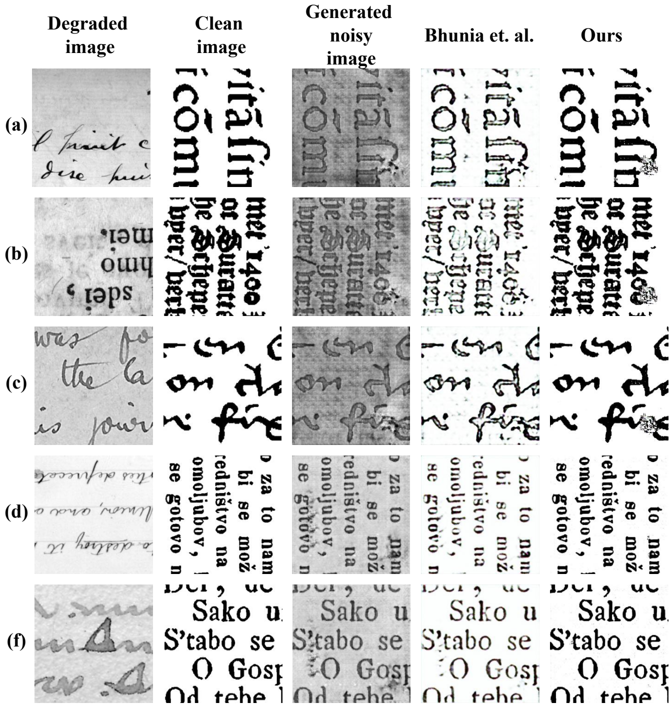

Fig. 2. Comparison of the qualitative results of predicted binarized images by Bhunia *et al.* [11] and our framework on the evaluation set

C. Baselines Methods

In this section, We present two alternative baselines to justify the effectiveness of our methods :

UDBNet-CL : The joint discriminator exploits domain confusion loss [47] to address the limitation described in the section III-B. The domain confusion loss gives equal importance to ATANet and UDBNet.

UDBNet-GRL : The Gradient Reversal layer [48] ensures that the adversarial discriminator views the two domains identically. Here, the joint discriminator utilize the Gradient Reversal layer.

D. Performance Analysis

From Table I, we observe that UDBNet-CL has achieved an improvement of 1.3% and 0.4% in F-Measure from DeepOtsu [10] and Bhunia [11] on H-DIBCO 2016 [24] dataset. On the other hand, UDBNet-GRL shows better performance than UDBNet-CL, an improvement of 1.8% and 0.9% in F-Measure from DeepOtsu [10] and Bhunia [11] H-DIBCO

TABLE I
COMPARSION OF OUR METHOD WITH BASELINE METHODS

Methods	F-Measure	F_{PS}	PSNR	DRD
UDBNet-CL	92.7	95.8	19.9	2.6
UDBNet-GRL	93.2	96.0	20.1	2.4
Ours	**93.4**	**96.2**	**20.1**	**2.2**

2016 [24]. Our method outperformed all the previous state-of-the-art methods because DeepOtsu [10] just uses stack refinement blocks and Bhunia [11] simply generates synthetic uncontrolled noisy image samples for training to improve the performance. In contrast, Our ATANet generates realistic degraded images including hard samples and also guides UDBNet to adopt to real noise distribution as depicted in Figures 2 and 3. However, out of three proposed approaches including baselines, the flipped label approach (ours) is found to be the best in all the four evaluation criteria because of its learning strategy. From Table II, it is obvious that our method has achieved significant improvement of 2.0%, 1.9%, 0.5% from DeepOtsu [10] and 1.1%, 0.6%, 0.2% from Bhunia [11] in F-Measure, F_{PS} and PSNR criteria on H-DIBCO 2016 [24] dataset. Also, our method has shown improved performance of 1.9%, 1.5% and 0.3% in F-Measure, F_{PS}, PSNR than Vo [8] on DIBCO 2011 dataset. In both the cases, the low DRD value of our method implies the robustness regarding visual distortion.

V. CONCLUSION

In this paper, we have proposed a novel approach towards document binarization by introducing three-player min-max adversarial game. We introduce a joint discriminator which tries to couple the Adversarial Texture Augmentation Network (ATANet) and Unsupervised Document Binarization Network (UDBNet) so that it can tackle the dataset bias problem and perform well on the real degraded document image. The proposed framework is simple and easy to implement. We demonstrate the effectiveness of our system by conducting experiments on publicly available DIBCO datasets. The results of the experiment show the superiority of our proposed model over the existing methods.

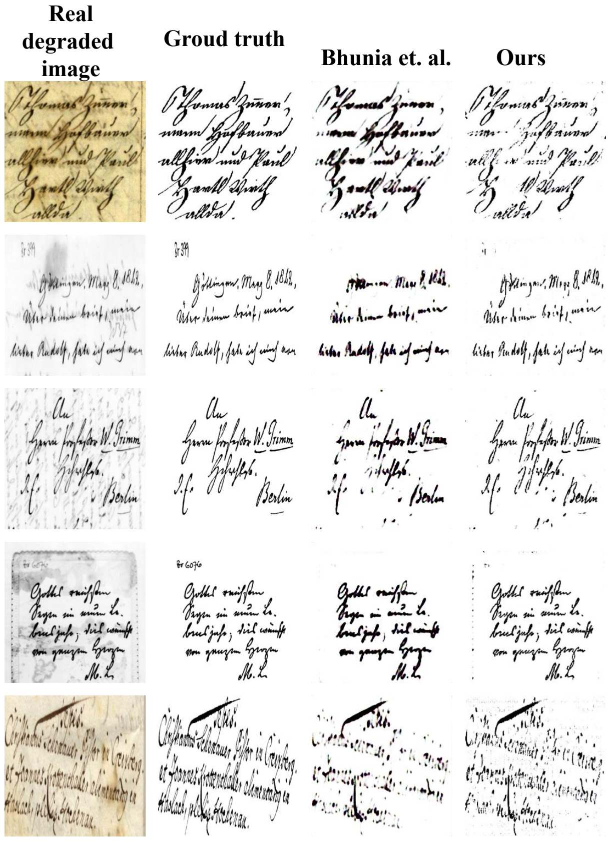

Fig. 3. Binarization results on real test images by passing through UDBNet.

REFERENCES

[1] Y. Chen and L. Wang, "Broken and degraded document images binarization," *Neurocomputing*, 2017.

[2] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, "Document image binarization using structural symmetry of strokes," in *ICFHR*, 2016.

[3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, 1979.

[4] S. Bhowmik, S. Polley, M. G. Roushan, S. Malakar, R. Sarkar, and M. Nasipuri, "A holistic word recognition technique for handwritten bangla words," *International Journal of Applied Pattern Recognition*, 2015.

[5] S. Bhowmik, S. Malakar, R. Sarkar, and M. Nasipuri, "Handwritten bangla word recognition using elliptical features," in *CICN*, 2014.

[6] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, and D. K. Basu, "A hierarchical approach to recognition of handwritten bangla characters," *Pattern Recognition*, 2009.

[7] T. A. Tran, K. Oh, I.-S. Na, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "A robust system for document layout analysis using multilevel homogeneity structure," *Expert Systems With Applications*, 2017.

[8] Q. Vo, S. Kim, H. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, 2018.

[9] F. Westphal, N. Lavesson, and H. Grahn, "Document image binarization using recurrent neural networks," in *DAS*, 2018.

[10] S. He and L. Schomaker, "Deepotsu: Document enhancement and binarization using iterative deep learning," *Pattern Recognition*, 2019.

[11] A. K. Bhunia, A. K. Bhunia, A. Sain, and P. P. Roy, "Improving document binarization via adversarial noise-texture augmentation," in *ICIP*, 2019.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[13] Y.-K. Lai and P. L. Rosin, "Efficient circular thresholding," *IEEE Transactions on Image Processing*, 2014.

[14] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern recognition*, 1986.

[15] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer vision, graphics, and image processing*, 1985.

TABLE II
QUANTATIVE RESULTS ON H-DIBCO 2016 AND DIBCO 2011 DATASET

Methods	**H-DIBCO 2016 Dataset**				**DIBCO 2011 Dataset**			
	F-Measure	F_{PS}	**PSNR**	**DRD**	**F-Measure**	F_{PS}	**PSNR**	**DRD**
Otsu [3]	86.6	89.9	17.8	5.6	82.1	84.8	15.7	9.0
Sauvola [20]	84.6	88.4	17.1	6.3	82.1	87.7	15.6	8.5
Howe [23]	87.5	92.3	18.1	5.4	91.7	92.0	19.3	3.4
Su [49]	84.8	88.9	17.6	5.6	87.8	90.0	17.6	4.8
Jia [50]	90.5	93.3	19.3	3.9	91.9	95.1	19.0	2.6
Vo [51]	87.3	90.5	17.5	4.4	88.2	90.3	20.1	2.9
Vo [8]	90.1	93.6	19.0	3.5	93.3	96.4	20.1	2.0
Westphal [9]	88.8	92.5	18.4	3.9	-	-	-	-
DeepOtsu [10]	91.4	94.3	19.6	2.9	93.4	95.8	19.9	1.9
Bhunia [11]	92.3	95.4	19.9	2.7	93.7	96.8	20.1	1.8
Ours	**93.4**	**96.2**	**20.1**	**2.2**	**95.2**	**97.9**	**20.4**	**1.5**

[16] W. Tsai *et al.*, "Moment-preserving thresholding-a new approach," *Computer Vision Graphics and Image Processing*, 1985.

[17] N. Papamarkos, "A technique for fuzzy document binarization," in *DocEng*, 2001.

[18] J. Bernsen, "Dynamic thresholding of gray-level images," in *ICPR*, 1986.

[19] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[20] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, 2000.

[21] C. Wolf, J.-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in *Object recognition supported by user interaction for service robots*, 2002.

[22] N. R. Howe, "A laplacian energy for document binarization," in *ICDAR*, 2011.

[23] N. Howe, "Document binarization with automatic parameter tuning," *IJDAR*, 2013.

[24] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icfhr2016 handwritten document image binarization contest (h-dibco 2016)," in *ICFHR*, 2016.

[25] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," in *TOG*, 2007.

[26] J. Pastor-Pellicer, S. España-Boquera, F. Zamora-Martínez, M. Z. Afzal, and M. J. Castro-Bleda, "Insights on the use of convolutional neural networks for document image binarization," in *ICANN*, 2015.

[27] J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Pixel-wise binarization of musical documents with convolutional neural networks," in *MVA*, 2017.

[28] M. Z. Afzal, J. Pastor-Pellicer, F. Shafait, T. M. Breuel, A. Dengel, and M. Liwicki, "Document image binarization using lstm: A sequence learning approach," in *HIP*, 2015.

[29] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *ICDAR*, 2017.

[30] J. Calvo-Zaragoza and A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, 2019.

[31] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icdar2017 competition on document image binarization (dibco 2017)," in *ICDAR*, 2017.

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

[33] C. Zheng, T.-J. Cham, and J. Cai, "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *ECCV*, 2018.

[34] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," *arXiv preprint arXiv:1903.10658*, 2019.

[35] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *CVPR*, 2019.

[36] L. Gatys, A. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *NIPS*, 2015.

[37] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016.

[38] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, "Covariate shift and local learning by distribution matching," 2008.

[39] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017.

[40] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "Icdar 2011 document image binarization contest (dibco 2011)," in *ICDAR*, 2011.

[41] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *ICDAR*, 2009.

[42] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "Icdar 2013 document image binarization contest (dibco 2013)," in *ICDAR*, 2013.

[43] I. Pratikakis, B. Gatos, and K.Ntirogiannis, "Icfhr 2012 competition on handwritten document image binarization," in *ICFHR*, 2012.

[44] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Icfhr2014 competition on handwritten document image binarization (h-dibco 2014)," in *ICFHR*, 2014.

[45] A. Shahab, F. Shafait, and A. Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," in *ICDAR*, 2011.

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[47] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, 2016.

[48] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[49] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *DAS*, 2010.

[50] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, "Degraded document image binarization using structural symmetry of strokes," *Pattern Recognition*, 2018.

[51] G. D. Vo and C. Park, "Robust regression for image binarization under heavy noise and nonuniform background," *Pattern Recognition*, 2018.