ISSN:1738-6764, eISSN:2093-7504

Document Image Binarization by GAN with Unpaired Data Training

Quang-Vinh Dang [1] and Guee-Sang Lee [2,*]

[1] Chonnam National University; Ph.D. Student; quangvinh242003@yahoo.com
[2] Chonnam National University; Professor; gslee@chonnam.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2020.16.2.008

Manuscript Received 5 August 2019; Received 22 April 2020; Accepted 8 May 2020

Abstract: *Data is critical in deep learning but the scarcity of data often occurs in research, especially in the preparation of the paired training data. In this paper, document image binarization with unpaired data is studied by introducing adversarial learning, excluding the need for supervised or labeled datasets. However, the simple extension of the previous unpaired training to binarization inevitably leads to poor performance compared to paired data training. Thus, a new deep learning approach is proposed by introducing a multi-diversity of higher quality generated images. In this paper, a two-stage model is proposed that comprises the generative adversarial network (GAN) followed by the U-net network. In the first stage, the GAN uses the unpaired image data to create paired image data. With the second stage, the generated paired image data are passed through the U-net network for binarization. Thus, the trained U-net becomes the binarization model during the testing. The proposed model has been evaluated over the publicly available DIBCO dataset and it outperforms other techniques on unpaired training data. The paper shows the potential of using unpaired data for binarization, for the first time in the literature, which can be further improved to replace paired data training for binarization in the future.*

Keywords: document image binarization; unpaired training data; gan; unsupervised learning

1. Introduction

Document image binarization is a method of classifying image pixels for one of two following categories: foreground and background. It is a fundamental issue in the field of document analysis because it affects further stages of the recognition process in document images. Although binarization seems quite easy for uniform images, it is possible challenges in real situations in which document images are degraded differently due to insufficient maintenance conditions. The significant impact of the binarization process has made the community hold the Document Image Binarization Competition (DIBCO) in conjunction with ICDAR and ICFHR conferences started in 2009 [1]. Binary handling of documents uses techniques to form other fields, such as removing image noise, semantic segments, image recovery, and background deletion such as handwritten segmentation contest [2].

With the emerging of powerful deep convolutional neural network (CNN) models for object classification, detection, and segmentation, many methods have been developed for document image binarization such as in [3-6]. In [4], the authors formulated binarization as a learning task for pixel classification and applied a novel Fully Convolutional Network (FCN) model that utilizes images with different sizes. In [5], a hierarchical deep supervised network (DSN) model can predict text or background pixels at different feature levels. Text pixels are distinguished significantly from background noises at higher-level features. Besides, text pixels are presented shapely at lower-level features. Then, these features are combined to have better outputs. In [6], the authors used convolutional auto-encoders to learn how to map an input image to its arbitrary output, in which activations state pixels belonging to either foreground or background. Consequently, once trained, the model can analyze degraded document images and binarize it. In another paper [7], the U-net architecture consists of an encoder and a symmetric decoder. In addition to this, there are skip-connections to copy information from

www.kci.go.kr

the encoder to the decoder to prevent losing information during the downsampling of the encoder. This architecture is applied and has the best performance in DIBCO 2017 competition [3]. However, these deep learning models require a substantial amount of paired training data that is not sufficient to research.

Some papers work on unpaired data such as papers [8, 9] but they are for other tasks. In [8], the authors presented an approach to learning how to translate an image from the source domain X to the Y target domain with unpaired data. The goal is to understand G: X → Y map so that the distribution of images from G (X) cannot be distinguished from the Y distribution by using an adversarial loss. The authors combine it with an inverse mapping F: Y → X and introduce a cycle consistency loss to map again F(G(X)) ≈ X. In [9], the authors proposed that the model train on both paired and unpaired data. This model can generate fake paired images from unpaired data. Therefore, it increases the number of available data as one kind of augmentation. Therefore, we use the baseline model of [8, 9] to compare with our method.

In this paper, there is a new proposed approach by utilizing unpaired training data for binarization problem. The contributions of this study may be summarized as follows. (1) The main contribution is the proposal of architecture that learns from unpaired image data to binarize degraded document images. The use of unpaired data training for binarization has never been tried before and it the first attempt in the literature as far as we know of. The designed structure is more suitable to the binarization problems than existing deep network models, especially in cases with the lack of data. (2) The proposed method is evaluated on the DIBCO datasets and achieves results that are better than that of state-of-the-art models on unpaired data.

The rest of the paper is organized as follows. Section 2 describes related works. Section 3 presents the proposed method for document image binarization. Section 4 discusses our experiments and results with other techniques. Finally, the paper is concluded in section 5.

2. Related Work

Generative adversarial networks (GANs) [10] are generative models that are particularly designed for image generation tasks with impressive results [11, 12]. This is because they use adversarial loss function. It forces the generated images to be indistinguishable from real images. In GAN training, a generator is trained to generate fake images like real images. So, it can fool a discriminator that tries to distinguish between fake images and real images. To generate more meaningful images, in conditional generative adversarial nets (CGAN) [13], the authors proposed that the model utilizes prior conditioned information to integrate into generated images. There are many successful applications of CGAN models, such as image editing [14], text-to-image translation [15] and inpainting [16].

Image-to-image translation models learn to map image inputs to image outputs. By using a conditional framework, Isola et al. [17] design the image-to-image translation with conditional adversarial networks (pix2pix) to learn the mapping function. Based on pix2pix, Zhu et al. [18] presented multimodal image-to-image translation (BicycleGAN) which achieves multi-modal image-to-image translation using paired data. However, most of the models need paired data for the training process, which are usually costly to obtain.

Unpaired Image-to-Image translation is to solve the issue of pairing training data, Zhu et al [8] proposed unpaired image-to-image translation using cycle-consistent adversarial networks (CycleGAN), which is unsupervised learning in the training process. It maps between two unpaired image domains with the aid of a cycle-consistency loss. It can produce image translations such as changing impressionism paintings to photorealistic images or creating images of zebras from images of horses. Because CycleGAN can translate an image to the target domain and back, we use it to translate degraded document images to the binary document domain with single unpaired data. So, we utilize CycleGAN as a baseline to compare our proposed model in section 4.

Learning disentangled representations aims at generating a wide diversity of data. In recent years, the unsupervised learning model has been paid attention. For example, interpretable representation learning by information maximizing generative adversarial nets (InfoGAN) [19] and learning basic visual concepts with a constrained variational framework (β-VAE) [20] have been proposed to learn disentangled representations without supervision. However, they failed to generate diverse outputs from a given source domain. To tackle this constraint, Xun Huang [21] proposed a multimodal unsupervised image-to-image translation (MUNIT) framework. This framework is utilized in our proposed model because it not only changes unpaired data to paired data but also generates diverse and realistic paired data.

In [22-25], the authors proposed neural style transfer approach to implement image-to-image translation. The model combines the content of one image with the style of another image to synthesizes a fake image. The

www.kci.go.kr

key idea of learning style representation is applying the gram matrix function to map features between a fake image and a real image. Recently, Ankan et al. [9] also adopt neural style transfer to use unpaired training data. However, the target of the [9] is to improve document binarization via adversarial noise-texture augmentation. In other words, it utilizes both paired and unpaired training data. So, we construct it as a baseline on unpaired training data to compare our proposed. model in section 4.

3. Proposed Method

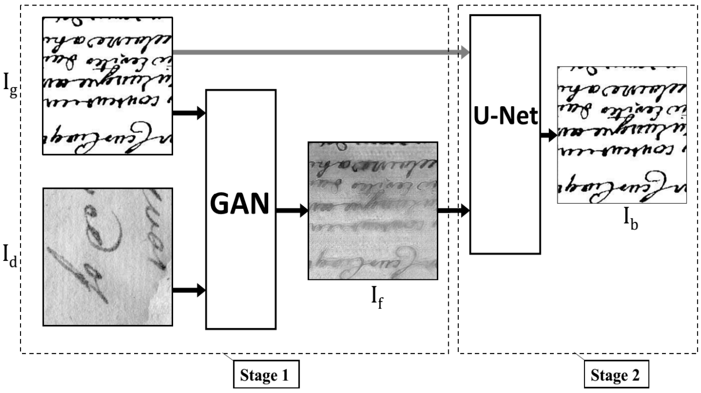


Figure 1. Illustration of the proposed framework in the training process: in stage 1, degraded image patches I_d and ground truth image patches I_g of another degraded image patches are fed into GAN model for generating fake degraded image patches I_f. In stage 2, U-net tries to get back clean binarized image patches I_b by denoising the generated ones I_f.

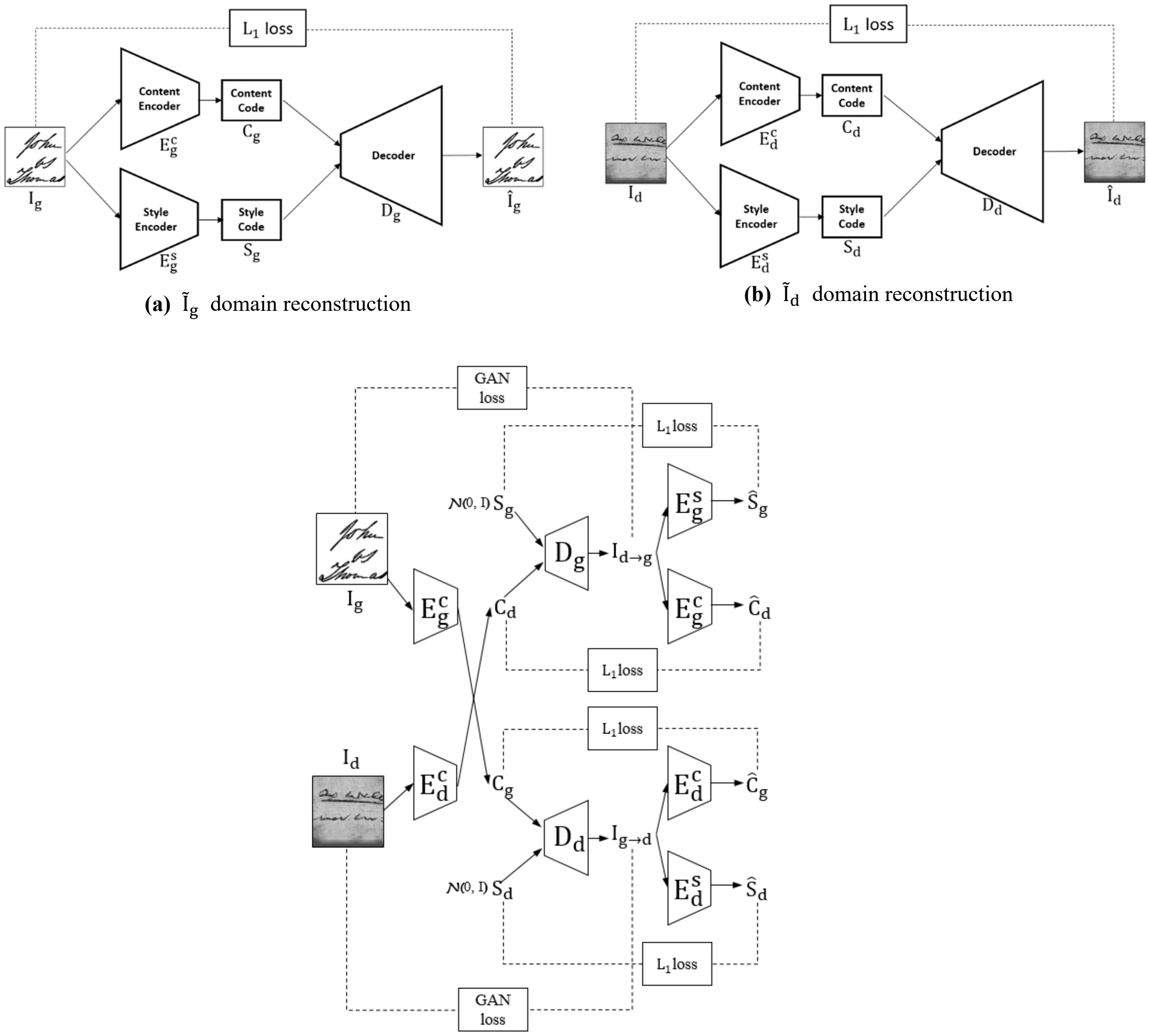
The details of our proposed model are given in this section. The model consists of two networks: a fake image generating network by utilizing GAN and U-net [7] for binarization network.

In stage 1 of Figure 1, we apply MUNIT GAN [21] for generating diverse and realistic images. The image-to-image translation model consists of two auto-encoders, one for each domain. The latent code of each auto-encoder is composed of a content code C_x and a style code S_x (x is g or d). The model is trained with adversarial objectives that ensure the translated images I_f cannot be distinguished from real images I_d by a discriminator. Furthermore, we explain its operation through 2 processes.

In the GAN training process, the content encoder E_g^c and the style encoder E_g^s as shown in Figure 2(a) are trained to extract content code C_g and style code S_g from binary images, respectively. The content encoder E_d^c and the style encoder E_d^s as shown in Figure 2(b) are trained to extract content code C_d and style code S_d from degraded images, respectively. Trained decoders can rebuild images having content and style the same with corresponding input. The decoder D_g creates images having binary image style. The decoder D_d generates images having degraded image style. Because the objective for auto-encoder is to enforce the same between original image I_x and generated image $\hat{I}_x$ (x is g or d), we use the L1 loss function as the equations (1) and (2) for image reconstruction. Then, in Figure 2(c), the decoder D_d is fed by C_g and noise. It creates an image $I_{g\rightarrow d}$ with the same content as I_g and degraded image style taken randomly in style space $\tilde{S}_d$. Because the original image I_d and the fake image $I_{g\rightarrow d}$ are different in content but like degraded image style, we use the adversarial loss as the equation (8). The fake image $I_{g\rightarrow d}$ continues to feed into the encoder E_d^c and E_d^s to reconstruct content code $\hat{C}_g$ and style code $\hat{S}_d$, respectively. Because $\hat{C}_g$ and $\hat{S}_d$ should be the same as C_g and S_d, respectively, we apply the L1 loss function as the equation (5) for (S_d, $\hat{S}_d$) pair and the equation (3) for (C_g, $\hat{C}_g$) pair for latent code reconstruction (in a similar way for $I_{d\rightarrow g}$).

In the inference process of GAN, the trained encoders can disentangle content code C_x and style code S_x (x is g or d) from different domains as shown in Figure 3(a). The trained decoder D_d can generate fake images $I_{g\rightarrow d}$ (or I_f) from the content code C_g and degraded image style code taken randomly in style space $\tilde{S}_d$. In Figure 3(b), 5 generated images have the same content as the original image I_g and 5 different degraded image styles.

www.kci.go.kr



(a) $\tilde{I}_g$ domain reconstruction

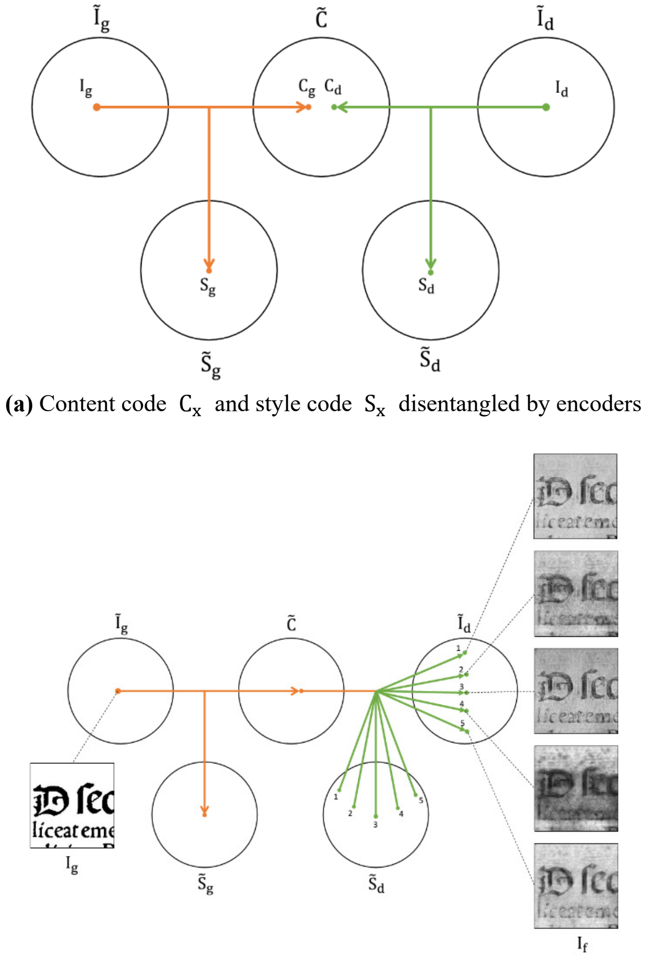
(b) $\tilde{I}_d$ domain reconstruction

(c) Cross-domain translation

Figure 2. GAN model overview in the training process. (a) and (b) is auto-encoder architecture. They extract latent code (C_x, S_x) for each domain. (c) Cross-domain translation model combines between content code and style code in another domain. To have generated fake images look like real samples, we train the model with GAN loss. $I_{g\to d}$ is a sample produced by translating I_g to domain $\tilde{I}_d$ (in a similar way for $I_{d\to g}$). To reconstruct both images and latent codes, we employ L_1 loss function. $\hat{I}_x$, $\hat{C}_x$, and $\hat{S}_x$ are reconstructed outputs having the same with I_x, C_x, and S_x (x is g or d), respectively.

The GAN model can generate diverse and multimodal degraded document image outputs by sampling different style codes. In Figure 3(b), the five outputs are the combination of five different style codes and one same content code. Therefore, the model not only changes unpaired data to paired data but also creates more paired data (I_g, I_f).

www.kci.go.kr



(a) Content code C_x and style code S_x disentangled by encoders

(b) Domain translation

Figure 3. The GAN model based on [21] in stage 1. (a) Encoders disentangle image information in each $\tilde{I}_x$ domain into the shared content space $\tilde{C}$ and the style space $\tilde{S}_x$. (b) To translate an image I_g in the domain $\tilde{I}_g$ to the domain $\tilde{I}_d$, the content code of I_g is recombined with a random style code in the style space $\tilde{S}_d$. The result is different fake degraded document images I_f with different style textures.

Bidirectional reconstruction loss function includes loss function for image reconstruction and loss function for latent code reconstruction. It ensures encoder is decoder's inverse and vice versa.

Loss function for image reconstruction ($J_{re}^{I_x}$) can restore an image (I_x) (x is g or d) sampled from the data distribution $p(I_x)$ ($I_x \sim p(I_x)$) following a direction (image → latent code → image) as shown in Figure 2(a) and 2(b).

$$J_{re}^{I_g} = E_{I_g \sim p(I_g)}\left[\left\|D_g\left(E_g^c(I_g), E_g^s(I_g)\right) - I_g\right\|_1\right] \tag{1}$$

$$J_{re}^{I_d} = E_{I_d \sim p(I_d)}\left[\left\|D_d\left(E_d^c(I_d), E_d^s(I_d)\right) - I_d\right\|_1\right] \tag{2}$$

in equation (1), E_g^c, E_g^s, and D_g are the content encoder, the style encoder, and decoder as shown in Figure 2(a), respectively. The result of $D_g\left(E_g^c(I_g), E_g^s(I_g)\right)$ is $\hat{I}_g$. So, $J_{re}^{I_g} = \left\|\hat{I}_g. - I_g\right\|_1$is L_1 loss function (in a similar way for $J_{re}^{I_d}$).

The loss functions for latent code reconstruction, ($J_{re}^{C_x}, J_{re}^{S_x}$) (x is g or d), are used to restore latent code sampled from the latent distribution ($C_x \sim p(C_x), S_x \sim q(S_x)$) as shown in Figure 2(c). The loss functions for reconstruction $J_{re}^{I_s}, J_{re}^{C_d}, J_{re}^{S_g}$ are as follows.

www.kci.go.kr

$$\mathrm{J}_{\mathrm{re}}^{\mathrm{C_g}} = \mathrm{E}_{\mathrm{C_g}\sim\ \mathrm{p}(\mathrm{C_g}),\mathrm{S_d}\sim\ \mathrm{q}(\mathrm{S_d})}\left[\left\|\mathrm{E}_{\mathrm{d}}^{\mathrm{c}}\left(\mathrm{D_d}(\mathrm{C_g},\mathrm{S_d})\right) - \mathrm{C_g}\right\|_1\right] \tag{3}$$

$$\mathrm{J}_{\mathrm{re}}^{\mathrm{C_d}} = \mathrm{E}_{\mathrm{C_d}\sim\ \mathrm{q}(\mathrm{C_d}),\mathrm{S_g}\sim\ \mathrm{p}(\mathrm{S_g})}\left[\left\|\mathrm{E}_{\mathrm{g}}^{\mathrm{c}}\left(\mathrm{D_g}(\mathrm{C_d},\mathrm{S_g})\right) - \mathrm{C_d}\right\|_1\right] \tag{4}$$

$$\mathrm{J}_{\mathrm{re}}^{\mathrm{S_d}} = \mathrm{E}_{\mathrm{C_g}\sim\ \mathrm{p}(\mathrm{C_g}),\mathrm{S_d}\sim\ \mathrm{q}(\mathrm{S_d})}\left[\left\|\mathrm{E}_{\mathrm{d}}^{\mathrm{s}}\left(\mathrm{D_d}(\mathrm{C_g},\mathrm{S_d})\right) - \mathrm{S_d}\right\|_1\right] \tag{5}$$

$$\mathrm{J}_{\mathrm{re}}^{\mathrm{S_g}} = \mathrm{E}_{\mathrm{C_d}\sim\ \mathrm{p}(\mathrm{C_d}),\mathrm{S_g}\sim\ \mathrm{q}(\mathrm{S_g})}\left[\left\|\mathrm{E}_{\mathrm{g}}^{\mathrm{s}}\left(\mathrm{D_g}(\mathrm{C_d},\mathrm{S_g})\right) - \mathrm{S_g}\right\|_1\right] \tag{6}$$

in equation (3), $\mathrm{q}(\mathrm{S_d})$ is the prior normal distribution N(0,I), $\mathrm{p}(\mathrm{C_g})$ is given by $\mathrm{C_g} = \mathrm{E}_{\mathrm{g}}^{\mathrm{c}}(\mathrm{I_g})$. $\mathrm{E}_{\mathrm{d}}^{\mathrm{c}}$ and $\mathrm{D_d}$ are the content encoder and decoder as shown in Figure 2(c), respectively. The result of $\mathrm{E}_{\mathrm{d}}^{\mathrm{c}}\left(\mathrm{D_d}(\mathrm{C_g},\mathrm{S_d})\right)$ is $\hat{\mathrm{C}}_{\mathrm{g}}$. So, $\mathrm{J}_{\mathrm{re}}^{\mathrm{C_g}} = \left\|\hat{\mathrm{C}}_{\mathrm{g}} - \mathrm{C_g}\right\|_1$is L1 loss function (in a similar way for $\mathrm{J}_{\mathrm{re}}^{\mathrm{C_d}}, \mathrm{J}_{\mathrm{re}}^{\mathrm{S_d}}$ and $\mathrm{J}_{\mathrm{re}}^{\mathrm{S_g}}$). Because they are $\mathrm{L_1}$ loss functions, $\mathrm{J}_{\mathrm{re}}^{\mathrm{I}_x}, \mathrm{J}_{\mathrm{re}}^{\mathrm{C}_x}$ and $\mathrm{J}_{\mathrm{re}}^{\mathrm{S}_x}$ encourage sharp output images, preserving semantic content of the input images and diverse output images, respectively.

Adversarial loss ($\mathrm{J}_{\mathrm{GAN}}^{\mathrm{I_x}}$) ensures fake image outputs like real image inputs as shown in Figure 2(c). Because it maps fake image distribution to the distribution of real images, it should be:

$$\mathrm{J}_{\mathrm{GAN}}^{\mathrm{I_g}} = \mathrm{E}_{\mathrm{C_d}\sim\mathrm{p}(\mathrm{C_d}),\mathrm{S_g}\sim\mathrm{q}(\mathrm{S_g})}\left[\log\left(1 - \mathrm{D_1}\left(\mathrm{D_g}(\mathrm{C_d},\mathrm{S_g})\right)\right)\right] + \mathrm{E}_{\mathrm{I_g}\sim\mathrm{p}(\mathrm{I_g})}\left[\log \mathrm{D_1}(\mathrm{I_g})\right] \tag{7}$$

$$\mathrm{J}_{\mathrm{GAN}}^{\mathrm{I_d}} = \mathrm{E}_{\mathrm{C_g}\sim\mathrm{p}(\mathrm{C_g}),\mathrm{S_d}\sim\mathrm{q}(\mathrm{S_d})}\left[\log\left(1 - \mathrm{D_2}\left(\mathrm{D_d}(\mathrm{C_g},\mathrm{S_d})\right)\right)\right] + \mathrm{E}_{\mathrm{I_d}\sim\mathrm{p}(\mathrm{I_d})}\left[\log \mathrm{D_2}(\mathrm{I_d})\right] \tag{8}$$

where the discriminator $\mathrm{D_1}$ distinguishes between generated fake images $\mathrm{I}_{\mathrm{d}\to\mathrm{g}}$ and real input images $\mathrm{I_g}$. The discriminator $\mathrm{D_2}$ distinguishes between generated fake images $\mathrm{I}_{\mathrm{g}\to\mathrm{d}}$ and real input images $\mathrm{I_d}$.

So, total loss:

$$\begin{aligned}\min_{\mathrm{E_g},\mathrm{E_d},\mathrm{D_g},\mathrm{D_d}}\ \max_{\mathrm{D_1},\mathrm{D_2}}\ \mathrm{J}(\mathrm{E_g},\mathrm{E_d},\mathrm{D_g},\mathrm{D_d},\mathrm{D_1},\mathrm{D_2}) &= \mathrm{J}_{\mathrm{GAN}}^{\mathrm{I_g}} + \mathrm{J}_{\mathrm{GAN}}^{\mathrm{I_d}} + \\ &\beta_{\mathrm{I}}\left(\mathrm{J}_{\mathrm{re}}^{\mathrm{I_g}} + \mathrm{J}_{\mathrm{re}}^{\mathrm{I_d}}\right) + \beta_{\mathrm{c}}\left(\mathrm{J}_{\mathrm{re}}^{\mathrm{C_g}} + \mathrm{J}_{\mathrm{re}}^{\mathrm{C_d}}\right) + \beta_{\mathrm{s}}\left(\mathrm{J}_{\mathrm{re}}^{\mathrm{S_g}} + \mathrm{J}_{\mathrm{re}}^{\mathrm{S_d}}\right)\end{aligned} \tag{9}$$

where β_{I}, β_{c}, β_{s} are weights. They are chosen by experiment.

In stage 2 of Figure 1, the task of this stage is to binarize fake degraded document images. The paired data transformed from the unpaired data is fed into U-net network. U-net [7] is employed as the binarization network. The U-net architecture consists of an encoder and a symmetric decoder. The encoder is constituted by the general convolutional process. The decoder is constituted by transposed 2d convolutional layers. Furthermore, using skip-connections to copy information of feature from the encoder to the decoder help to prevent losing information during downsampling of the encoder. The pixel-wise softmax function is applied on the final feature map which is followed by the cross-entropy loss function. So U-net can classify each pixel into one of the background or foreground classes. U-net can localize and distinguish text by classifying every pixel, so the input and output share the same size. Therefore, we convert an original semantic segmentation task of U-net into a binarization task.

Moreover, U-net is chosen instead of pix2pix GAN [17] that is applied in [9]. This is because, in [17], the authors proved that U-net has a higher score than pix2pix in segmentation task. Furthermore, the winner in DIBCO 2017 competition [3] employed U-net for the document image binarization model.

4. Experiments and Results

We have used 9 publicly available document datasets. They include DIBCO 2009 [26], DIBCO 2011 [27], DIBCO 2013 [28], H-DIBCO 2010 [29], HDIBCO 2012 [30], H-DIBCO 2014 [31], Bickley diary [32], PHIDB [33], and S-MS [34] datasets. DIBCO 2013 dataset is selected for the testing set. The remaining datasets are used as training and evaluation sets. In the dataset for training, there are available paired data. However, with

www.kci.go.kr

the purpose of training on unpaired data, we change paired data to unpaired data. Then, we split images into patches with size 256 × 256. We perform augmentation with rotation: 0, 180, or 270 degrees. 90% of the obtained image patches are used as the training set, and the rest of the images are used as the validation set.

For evaluation, we followed the measures of *F-measure*, *pseudo-F-measure*, *PSNR* and *DRD*, used in ICDAR competitions of DIBCO 2017 and DIBCO 2009 [3, 35].

The first one, or *F-measure*, is adequate for evaluation in binarization process because the distribution of foreground and background classes are often unbalanced. *F-measure FM* is computed as

$$FM = \frac{2TP}{2TP+FP+FN} \tag{10}$$

where *TP* denotes true positives where foreground pixels are predicted as foreground. *FP* stands for false positives where background pixels are predicted as foreground. *FN* refers to false negatives where foreground pixels are predicted as background.

Pseudo-F-measure (F_{ps}) is similar to *F-measure* but the relevance of each pixel is weighted based on its distance from stroke boundaries. It is computed as

$$F_{ps} = \frac{2R_{ps}P_{ps}}{R_{ps} + P_{ps}} \ , \tag{11}$$

where R_{ps} denotes pseudo-Recall [36]. It includes weights of the ground-truth text normalized according to the local stroke width. P_{ps} stands for pseudo-Precision [36]. It has weights constrained within an expanding to the ground-truth background area.

Peak signal-to-noise ratio ($PSNR$) measures how close a degraded document image is to its ground-truth image. It is computed as

$$PSNR = 10log\left(\frac{C^2}{MSE}\right) \tag{12}$$

where $MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\left(I(x, \ y)-I'(x, \ y)\right)^2}{MN}$ is the mean squared error. $I(x, \ y)$ and $I'(x, \ y)$ are the value of a pixel at the position $(x, \ y)$ of a ground-truth document image and a binarized document image, respectively. *M x N* is the dimension of the image. *C* is the difference between foreground and background.

Distance Reciprocal Distortion Metric (DRD) is the measure of visual distortion. We apply it in binary document images. It is computed as

$$DRD = \frac{\sum_{k=1}^{S} DRD_k}{NUBN} \tag{13}$$

where DRD_k is the distortion of the k-th flipped pixel and it is calculated using a 5×5 normalized weight matrix. $NUBN$ is the number of the non-uniform 8x8 blocks in the ground-truth image. A lower value for *DRD* indicates a better result while a higher value for *F-Measure*, *pseudo-F-Measure*, and *PSNR* demonstrates better performance.

We only implemented the following baseline models that can train on unpaired data. CycleGan [8] is implemented as the baseline model by utilizing the concept of cycle-consistent image translation frameworks. It can translate directly from the domain $\tilde{\mathrm{I}}_{\mathrm{d}}$ to the domain $\tilde{\mathrm{I}}_{\mathrm{b}}$ by using unpaired training data. The baseline model based on Ankan et al [9] is the combination of two GANs. One adopts the neural style transfer [37, 22] to generate fake images. Another being pix2pix GAN [17] is to binarize fake degraded document images. The original model trains on both paired and unpaired data. However, we trained it on unpaired data to compare with our model. Finally, we modified the model from [9]. It combines between GAN using the neural style transfer as [9] for creating fake images and U-net for binarization network. U-net is instead of pix2pix GAN [17] that is utilized in [9]. The only trained U-net is used to predict the result in the testing process.

www.kci.go.kr

Table 1. Results on DIBCO 2013 dataset with training on unpaired data

Methods	F-measure	Fps	DRD	PSNR
CycleGAN[8]	66.8	70.1	17.6	12.5
Ankan et al [9]	78.67	85.40	8.64	15.98
Modified model from [9]	80.29	**86.81**	7.82	16.30
Proposed method	**84.27**	85.85	**7.57**	**17.22**

In Figure 4, it shows the diverse and realistic generated image pairs are fed into U-net model. Particularly, the GAN model can generate five realistic degraded document images from only one ground-truth document image. So, each generated image pair includes a different fake degraded image and the same real ground-truth image. Then, the five image pairs are fed into U-net model to learn binarization way. Finally, the trained U-net model can convert degraded document images to binarized document images. In Figure 5, it displays the visual quality of the binary results of the full document image on the DIBCO 2013 dataset. Figure 5(c) displays the result from CycleGAN[8]. It contains a lot of noises and corrupted-text components. Figure 5(d) presents a better result than Figure 5(c) but text-stroke detail decreases. Figure 5(e) shows that the image result of the modified model from [9] contains weak noises but the character strokes are still thin. In Figure 5(f), the image output of the proposed model can preserve text-stroke information and remove background noises. This is because it takes advantage of the binarization ability of U-net and diverse and realistic image generation ability of MUNIT GAN.

We compared our proposed method with the baseline methods as shown in Table 1. In particular, the quantitative results show that the proposed method has the best regarding all four of the measurements except F_{ps} in the second-best place. The F-measure result of the proposed method is significantly higher than the second-place method. It shows the high efficiency of the proposed method because F-measure represents the general quality of the final result as well as the higher importance compared to other measurements in ICDAR competitions [3, 35]. In Ankan et al [9], the model works moderately. CycleGAN [8] obtains poor results compared to others. The modified model from [9] has a better result than the original model [9].

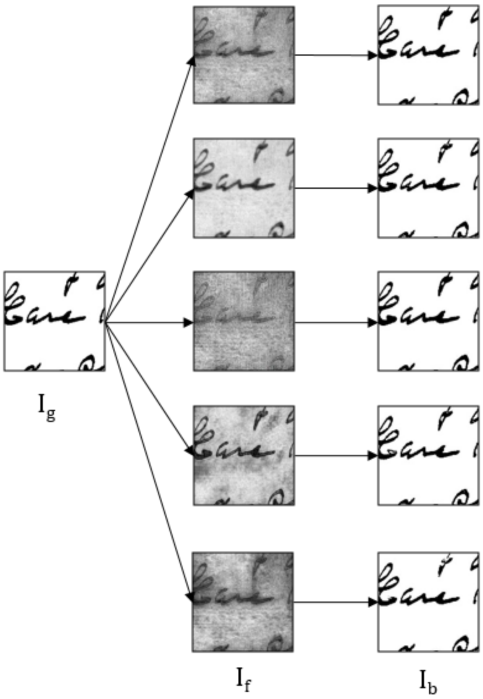
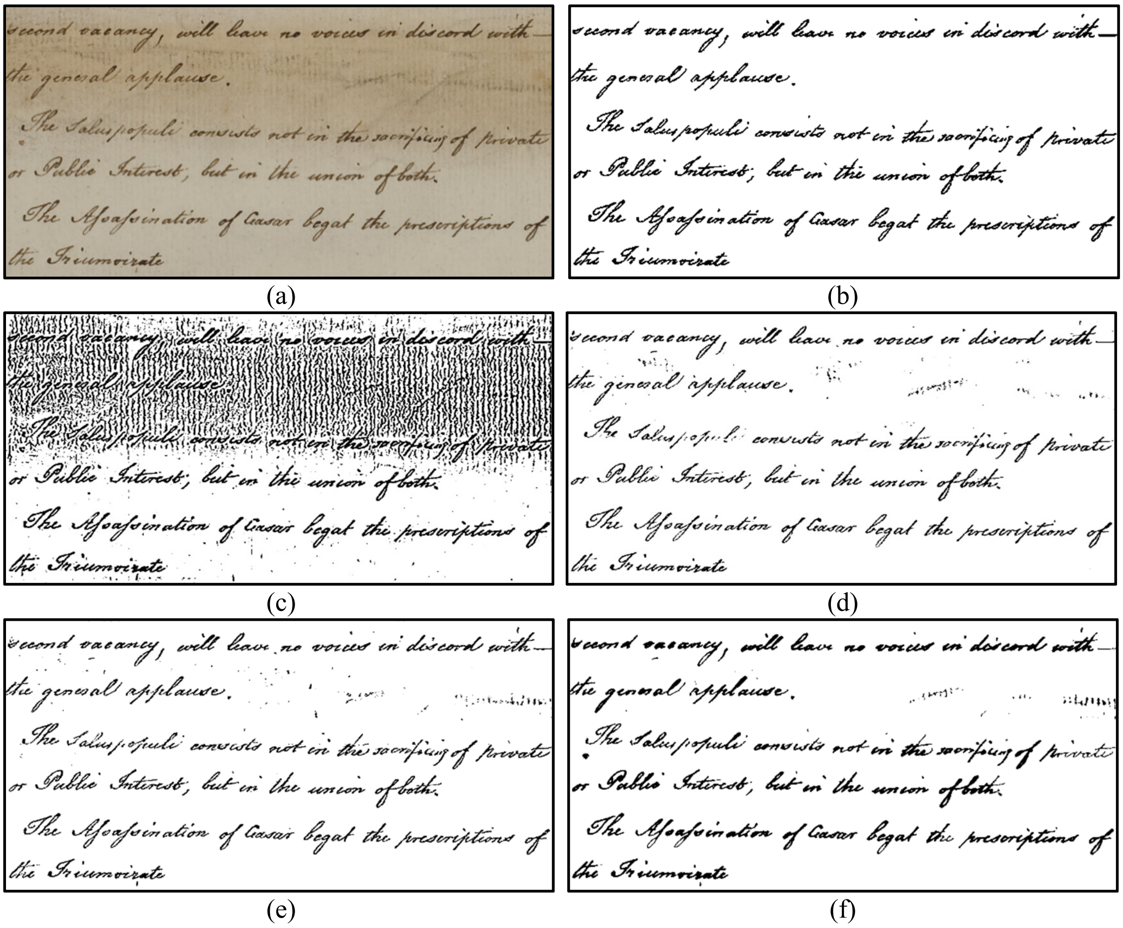


Figure 4. Binarization results on the validation set. Image patches are in the order: 1 ground truth image patch (I_g), 5 fake degraded image patches (I_f), and 5 clean binarized image patches (I_b) from left to right for each sample result

www.kci.go.kr



(a) (b) (c) (d) (e) (f)

Figure 5. Binarization results of the full document image on the DIBCO 2013 dataset produced by different methods: (a) original full image, (b) ground truth image, (c) CycleGan [8], (d) Ankan et al [9], (e) the modified model from [9], (f) the proposed method.

5. Conclusions

In this paper, with the combination of GAN and U-net, the model can train on unpaired datasets. It plays an important role in cases with a lack of data. This method opens the new approach for image binarization and other fields by utilizing unpaired data. In my knowledge, the paper is the first approach using unsupervised learning for document image binarization. Besides, our experimental results show that it outperforms existing state-of-the-art techniques for unpaired training data for document image binarization.

Acknowledgments: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 Robust Reading Competitions: Entries, Results, and Future Directions," in Int. J. Doc. Anal. Recognit, vol. 7, pp. 105-122, 2005, doi: https://doi.org/10.1007/s10032-004-0134-3.

[2] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei, "ICDAR 2013 Handwriting Segmentation Contest," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1402-1406, 2013, doi: https://doi.org/10.1109/ICDAR.2013.283.

www.kci.go.kr

[3] Loannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos, ”ICDAR 2017 Competition on Document Image Binarization (DIBCO 2017),” in *14th IAPR International Conference on Document Analysis and Recognition*, 2017, doi: https://doi.org/10.1109/ICDAR.2017.228.

[4] C.Tensmeyer and T. Martine, “Document Image Binarization with Fully Convolutional Neural networks,” in *14th IAPR International Conference on Document Analysis and Recognition*, 2017, doi: https://doi.org/10.1109/ICDAR.2017.25.

[5] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, “Binarization Of Degraded Document Images Based On Hierarchical Deep Supervised Network,” Journal Pattern Recognition, vol. 74, issue. C, pp. 568-586, Feb. 2018, doi: https://doi.org/10.1016/j.patcog.2017.08.025.

[6] Jorge Calvo-Zaragoza and Antonio-Javier Gallego, "A Selectional Auto-encoder Approach for Document Image Binarization," Journal Pattern Recognition, vol. 86, Jun. 2017, doi: https://doi.org/10.1016/j.patcog.2018.08.011.

[7] 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, pp. 234-241, 2015, doi: https://doi.org/10.1007/978-3-319.

[8] J. Zhu, T. Park, P. Isola, and A. Efros, “Unpaired Image-to-image Translation Using Cycle-Consistent Adversarial Networks,” in *ICCV*, 2017, doi: https://doi.org/10.1109/ICCV.2017.244.

[9] A. K. Bhunia, A. K. Bhunia, and P. P. Roy, “Improving Document Binarization Via Adversarial Noise-Texture Augmentation,” in *ICIP*, 2019, doi: https://doi.org/10.1109/ICIP.2019.8803348.

[10] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems,* vol. 2, pp. 2672-2680, Dec. 08-13, 2014.

[11] E. L. Denton, S. Chintala, and R. Fergus, “Deep generative image models using a laplacian pyramid of adversarial networks,” in *NIPS*, 2015.

[12] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *ICLR*, 2016.

[13] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[14] G. Perarnau, J. van de Weijer, B. Raducanu and J.M. Alvarez, “Invertible conditional gans for´image editing,” in *NIPS Workshop*, 2016.

[15] E. Mansimov, E, Parisotto, J. L. Ba, and R. Salakhutdinov, “Generating images from captions with attention,” in *ICLR*, 2015.

[16] H. Liu, B. Jiang, Y. Xiao, and C. Yang, ‘‘Coherent semantic attention for image inpainting,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 4170-4179, Oct. 2019.

[17] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017, doi: https://doi.org/10.1109/CVPR.2017.632.

[18] J. Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, “Toward multimodal image-to-image translation,” in *NIPS*, 2017.

[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *NIPS*, 2016.

[20] I. Higgins, L. Matthey, A. Pal, C. Burgess X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” in *ICLR*, 2017.

[21] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *ECCV*, 2018.

[22] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *CVPR*, pp. 2414-2423, 2016, doi: https://doi.org/10.1109/CVPR.2016.265.

[23] J. Johnson, A. Alahi, and L. Fei-Fei “Perceptual losses for real- time style transfer and super-resolution,” in *ECCV*, 2016.

[24] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, “Texture networks: Feed-forward synthesis of textures and stylized images,” in *ICML*, 2016.

[25] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman, “Preserving color in neural artistic style transfer,” in preprint arXiv:1606.05897, 2016.

[26] B. Gatos, K. Ntirogiannis, and I. Pratikakis, “ICDAR 2009 document image binarization contest (DIBCO 2009),” in *ICDAR*, pp. 1375-1382, 2009, doi: https://doi.org/10.1109/ICDAR.2009.246.

[27] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR 2011 document image binarization contest (DIBCO 2011),” in *ICDAR*, pp. 1506-1510, 2011, doi: https://doi.org/10.1109/ICDAR.2011.299.

www.kci.go.kr

[28] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR 2013 document image binarization contest (DIBCO 2013),” in *ICDAR*, pp. 1471-1476, 2013, doi: https://doi.org/10.1109/ICDAR.2013.219.

[29] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “H-DIBCO 2010 - handwritten document image binarization competition,” in *ICFHR*, pp. 727-732, 2010, doi: https://doi.org/10.1109/ICFHR.2010.118.

[30] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012),” in *ICFHR*, pp. 817-822, 2012, doi: https://doi.org/10.1109/ICFHR.2012.216.

[31] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICFHR 2014 competition on handwritten document image binarization (H-DIBCO 2014),” in *ICFHR*, pp. 809-813, 2014, doi: https://doi.org/10.1109/ICFHR.2014.141.

[32] F. Deng, Z. Wu, Z. Lu, and M.S. Brown, “BinarizationShop: a user-assisted software suite for converting old documents to black-and-white,” in *Proc. Annu. Joint Conf. Digit. Libraries*, pp. 255-258, 2010, doi: https://doi.org/10.1145/1816123.1816161.

[33] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, and M. Cheriet, “An efficient ground-truthing tool for binarization of historical manuscripts,” in *ICDAR*, pp. 807-811, 2013, doi: https://doi.org/10.1109/ICDAR.2013.165.

[34] R. Hedjam, H. Z. Nafchi, R. F. Moghaddam, M. Kalacska, and M. Cheriet, “ICDAR 2015 multispectral text extraction contest (MS-TEx 2015),” in *ICDAR*, pp. 1181-1185, 2015, doi: https://doi.org/10.1109/ICDAR.2015.7333947.

[35] B. Gatos, K. Ntirogiannis, and I. Pratikakis, “ICDAR 2009 document image binarization contest (DIBCO 2009),” *in Document Analysis and Recognition, 2009, ICDAR’09. 10th International Conference, IEEE*, pp. 1375-1382, 2009.

[36] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “Performance Evaluation Methodology for Historical Document Image Binarization,” IEEE Transactions on Image Processing, vol. 22, no. 2, pp. 595-609, 2013.

[37] L. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in *NIPS*, pp. 262-270, 2015.

© 2020 by the authors. Copyrights of all published papers are owned by the IJOC. They also follow the Creative Commons Attribution License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

www.kci.go.kr