Diffusion Denoising Process with Gated U-Net for High-Quality Document Binarization

Sangkwon Han , Seungbin Ji , Jongtae Rhee *

Posted Date: 30 August 2023

doi: 10.20944/preprints202308.2048.v1

Keywords: document binarization; deep learning; gated convolution; generative model; latent diffusion models; text stroke

Article

Diffusion Denoising Process with Gated U-Net for High-Quality Document Binarization

Sangkwon Han, Seungbin Ji and Jongtae Rhee *

Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, Korea; hsk0314@dgu.ac.kr (S.H.); voiagerd@dgu.ac.kr (S.J.)

* Correspondence: jtrhee@dongguk.edu

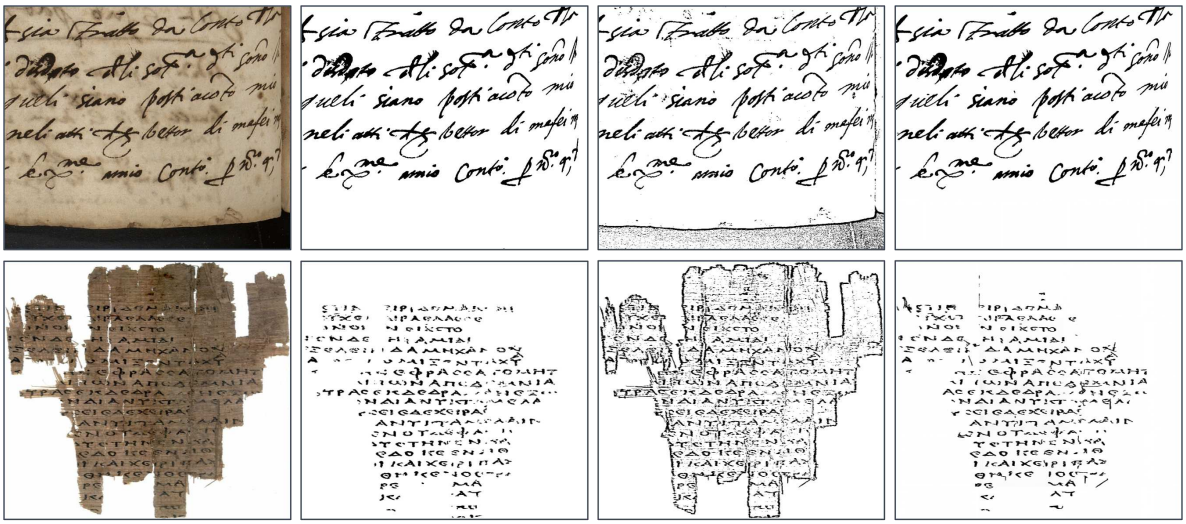
Abstract: Binarization of degraded documents is an important preprocessing task for various document analysis such as OCR and historical document analysis. Existing studies have applied various convolutional neural network (CNN) models and generative models for document binarization, but they do not show generalized performance for noise that the model has not seen and it suffers from extracting elaborate text strokes. In this paper, to overcome these challenges, we utilize latent diffusion model (LDM), which is known for high-quality image generation model, for the first time in document binarization. By utilizing the iterative diffusion-denoising process in latent space, it shows high-quality cleaned binarized image generation and high generalized performance through using both data distribution and time step while training. Additionally, we apply gated U-Net to the backbone network to preserve text strokes using trainable gating value. Gated convolution can extract elaborate text stroke by allowing the model to focus on text region by combining gating value and feature. Furthermore, we maximize the effectiveness of the proposed model by training it with a combination of LDM loss and pixel-level loss, which is suitable for the model structure. Experiments on H-DIBCO and DIBCO benchmark datasets show that the proposed model outperforms existing methods.

Keywords: document binarization; deep learning; gated convolution; generative model; latent diffusion models; text stroke

1. Introduction

Document images are essential data for digital document analysis such as OCR, historical document restoration and document classification [1]. However, document images obtained in real world are degraded document images due to various noise such as shadow, stain, ink smear, bleed-through, overwriting and variable background intensity [2]. These degraded document images have negative effect on various digital document analysis. Therefore, in order to effectively perform a digital document analysis, preprocessing of the degraded document image is essential. Document binarization is an important preprocessing task to obtain a clean binarized image by restoring a text region from a degraded document [3] . Document binarization does not simply remove specific noise from a degraded document image, but comprehensively processes various degradation factors [4]. However, converting a degraded document containing non-uniform noise into a clean binarized document is difficult to solve using traditional binarization algorithms as shown in Figure 1. Also, during this process, the difficulty of preserving elaborate text strokes is included. Due to these difficulties, research on document binarization of degraded document images is essential.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1



(a) (b) (c) (d)

Figure 1. Difficulties in preserving text stroke and solving severe degradation. (a) show the severe degraded document images. (b) show ground-truth images. (c) are the result images of [5], one of the traditional binarization algorithm-based methods. (d) are the result images of the proposed model, one of the deep learning-based methods. The proposed model is effective in preserving text strokes even in the case of severe degradation.

Research on document binarization has been ongoing for decades. Through decades of research, various image binarization algorithms [5–7] have been proposed, and they contributed to document binarization on normal and uniform images. However, it has limitations in distinguishing non-uniform and complex degradation and extracting elaborate text region in degraded document images [8]. To overcome these problems, deep learning-based computer vision methods have been applied to document binarization recently. Deep learning-based methods [9–11] such as segmentation and deep neural networks have been utilized. In [12], they proposed a model that redefined document binarization as a pixel classification problem by applying FCN (Fully Convolutional Network). In addition, in [8], the performance of document binarization was improved by training iterative neural network by combining the existing binary algorithm and deep learning. Recently, document binarization based on generative adversarial networks (GANs) has also been proposed. In [13–15], they successfully applied GAN as an image-to-image task to generate clean binarized document images from degraded document images. These deep learning-based methods overcame the limitations of traditional binarization algorithms and improved performance on text region preservation. However, these deep learning-based methods still have trouble extracting elaborate text strokes from non-uniform and complex degradation, and mode collapse [16], which degrades performance due to concentration on a specific data distribution, occurs.

To solve the above problems, we propose a document binarization model based on latent diffusion model [17], defining the document binarization as the iterative diffusion-denoising process as shown in Figure 2. Intuitively, this process generates a clean binarized document image, removing gaussian noise as each time step passes. To this end, we use the diffusion model [18], which has shown success in image generation tasks. Diffusion model is a model that performs image generation tasks by adding and removing gaussian noise in an image, and has the advantage of not relying heavily on training data by training to remove noise using both the distribution of data and time steps. Latent diffusion models showed improvement in time efficiency and precision of image generation by using the diffusion model in the latent space [17]. By applying this latent diffusion model to document binarization, the generalization effect on unseen noise is increased and the feature of text strokes is finely adjusted using the latent space. To the best of our knowledge, this is the first work introducing the diffusion model into document binarization. Additionally, a gated convolution is applied to the model backbone

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

network for elaborate text stroke extraction. Gated convolution demonstrated its performance in binary mask learning on the segmentation models [19]. In this work, gated convolution is effective in distinguishing between text and background region by training feature through original convolution and gating value, which means text region information. We utilize gated convolution in latent space, not in pixel space, so that the proposed model extracts more elaborate text strokes.

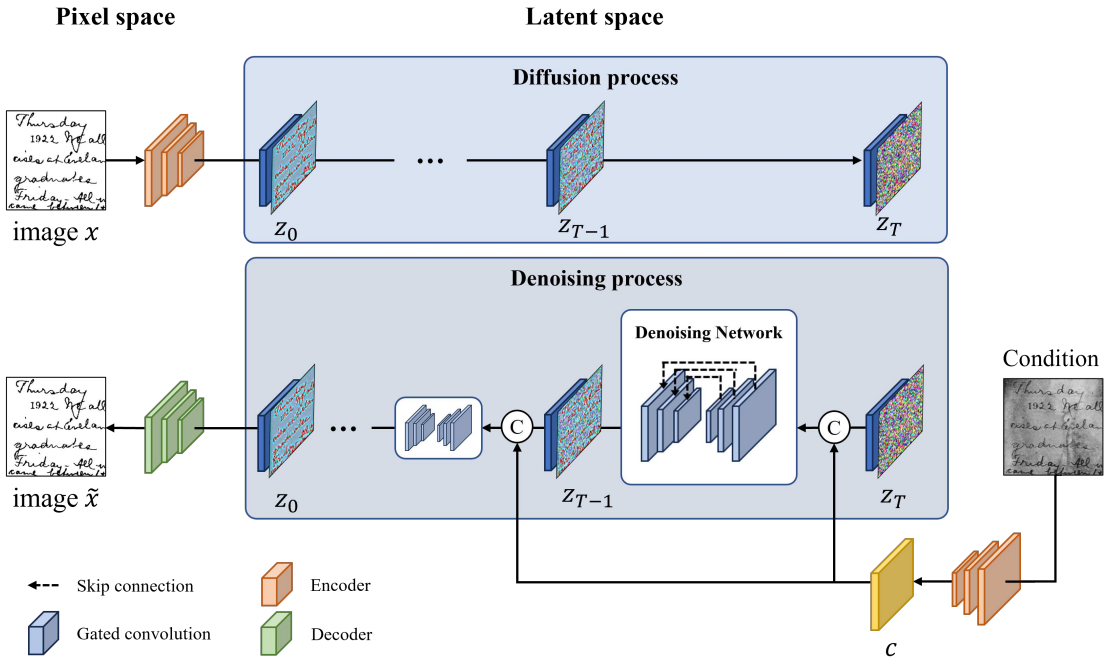


Figure 2. Architecture of the proposed model for document binarization. The network architecture can be divided into pixel space and latent space, and the diffusion-denoising process proceeds according to the time step in latent space..

To demonstrate the effectiveness of the proposed model, we conducted experiments on H-DIBCO, DIBCO (Document Image Binarization Contest) dataset [20–23]. For evaluation, we used four most commonly used metrics in document binarization and the proposed model outperformed the existing state-of-the-art methods in all metrics. In summary, this paper makes three significant contributions:

- We propose a novel approach by redefining document binarization as an iterative diffusion-denoising problem for degraded document images. This is the first work to introduce a latent diffusion model to document binarization to generate high-quality clean binarized document images.
- We use gated convolution in the latent space for elaborate text strokes extraction. This makes it easier to distinguish text from background by updating the gating value as a guide for the text region.
- The proposed model generates high-quality clean binarized document images and outperforms existing methods on several (H-)DIBCO datasets.

The rest of this paper is formulated as follows. In Section 2, related works are mentioned. Section 3 introduces the proposed model. Experimental results are discussed in Section 4. Conclusion are mentioned in Section 5.

 doi:10.20944/preprints202308.2048.v1

2. Related Work

2.1. Document Binarization

Document binarization aims to perform pixel-wise binary classification of the background and text in a document image. Methods for document binarization can be divided into traditional binarization algorithm-based methods and deep learning-based methods. Traditional binarization algorithms use various nonparametric-based threshold algorithm. These methods perform binary classification through a threshold for background and text. Otsu's method [6] uses a global threshold method that maximizes the distance between background and text and finds the maximum interclass variation for binary classification. These global threshold methods are effective for document images with uniform degradation, but are not suitable for document images with non-uniform degradation. To overcome this problem, pixel-wise local threshold methods have been proposed. Local threshold methods were used in [5,7], and more accurate classification was performed by measuring the relationship between nearby pixels based on local statistical information. These binarization algorithm-based methods are suitable at binarization of uniform noise, but still have difficulties in documents with various non-uniform noise such as overwriting [8].

To solve above problems, deep learning-based methods have been studied over the last few years. In [8], an iterative neural network was constructed to generate clean binarized images from non-uniformly degraded images, and Otsu's algorithm was combined to the enhanced images. In [24], they succeeded in obtaining a binarized image by training mid-level representation through an encoder-decoder architecture consisting of CNN. Encoder-decoder architecture was also used to construct a network for selectional output in [25]. In [26], they proposed a cascading U-Net architecture for complex document image processing tasks. Akbari et al. [27] utilizes convolutional neural networks to identify foreground pixels using input-generated multichannel images. Also, generative models based on generative adversarial networks (GANs), which regard document binarization as an image-to-image task [28], have been proposed. Zhao et al. [13] proposed a clean binarized image generation model through multi-scale information combination based on conditional-GANs (cGANs). Also, [29] utilized cGANs and showed performance improvements in watermark removal, deblurring and binarzation. Lin et al. [30] proposed a 3 stage method for binarization by combining discrete wavelet transform and GAN. In [15], color-independent adversarial networks are constructed in two stages to learn global and local features of document.

2.2. Diffusion Model for image-to-image Task

Diffusion models [18,31,32] showed great success in image generation and are used in various image generation tasks. Rombach et al. [17] used the diffusion method in latent space to finely adjust semantic features of images to improve the quality of inpainting, super-resolution and image-to-image tasks. Also, diffusion models have been applied for image segmentation tasks. In [33], a distribution of segmentation masks was generated through stochastic sampling process. Kim et al. [34] proposed a diffusion adversarial representation learning model through switchable sptially-adative denormalization for vessel segmentation. Chen et al. [35] used the diffusion process on detection box proposals for object detection, and [36] used the diffusion process for image depth estimation.

2.3. Gated Convolutions

Gated convolutions are used for various tasks such as segmentation [37,38], inpainting [19], and language modeling [39]. Li et al. [37] proposed Gated Fully Fusion (GFF), which selectively fuses multi-level features using gated convolution in a fully connected way for segmentation. In [40], they proposed Context-Gated Convolution (CGC) that adaptively modifies the weights of convolutional layers according to the global context. Zhang et al. [41] used gated convolution for vessel segmentation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

The network learns how to emphasize the edge of the vessel by utilizing gated convolution on the features extracted through the encoder-decoder architecture. In [19], a feature selection mechanism that can dynamically learn features for each spatial location is proposed to solve the problem of vanila convolution that uses the same filter for all input pixels. For the dynamic feature selection mechanism, gated convolution was used to effetively distinguish valid pixels from invalid pixels.

3. Method

We propose a binarization model of degraded document image through an iterative diffusion-denoising process. In this process, we use the latent diffusion model [17], which learns features of data distribution through the process of adding and removing gaussian noise to an image at each time step. The iterative diffusion-denoising process proceeds in the latent space through a pre-trained autoencoder and generates a clean binarized document image through the decoder of the autoencoder. In addition, by applying gated convolution in the denoising process, text stroke extraction performance is increased by updating the gating value corresponding to the text region.

In this section, first, the existing diffusion model is explained in the preliminaries, and then the architecture of the proposed model is explained. Next, the conditioned denoising process through the gated U-Net is described. Finally, the loss function of the proposed model is explained.

3.1. Preliminaries

Diffusion Models [18,31] are probabilistic models that aim to estimate a data distribution $p(x)$ by iteratively denoising noise from a normally distributed variable. The model performs various generative tasks through an iterative diffusion-denoising process. In the training stage, the diffusion model goes through a diffusion process that generates a noise vector x_t by gradually adding gaussian noise ϵ during the time step $t \in [0, T]$ from data x_0. The diffusion process $q(x_t \mid x_0)$ can be formulated as,

$$q(x_t \mid x_0) := \mathcal{N}(x_t \mid \sqrt{\alpha_t} x_0, (1 - \alpha_t) I), \tag{1}$$

where α_t is a parameter for the variance schedule and is related to the degree of noise addition.

The denoising process generates a denoising vector x_{t-1} from a random noise vector x_t through a denoising network. Through this iterative process, x_0 is generated as a result. The denoising process $p_\theta(x_{t-1} \mid x_t)$ is as follows,

$$p_\theta(x_{t-1} \mid x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \tag{2}$$

where $\mu_\theta(x_t, t)$ is a neural network to generate x_0 and learns iterative noise removal. σ_t^2 is a noise schedule and depends on α.

The diffusion model trains the denoising network $\epsilon_\theta(x_t, t)$ with equal weights, and the total loss L_{DM} is formulated as follows,

$$L_{DM} = \mathbb{E}_{x_0, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{3}$$

When sampling x_0, x_0 is generated using trained denoising network from the random noise vector x_T.

3.2. Network Architecture

Network architecture of the proposed model is based on the structure of latent diffusion models [17]. The overall architecture is shown in Figure 2. During the training process, a binarized document image given as input is converted into latent vector from the pixel space by the encoder of the pre-trained autoencoder. Utilization of the latent space makes it possible to finely adjust semantic features of latent vector [17,42,43]. Next, the diffusion process creates a gaussian noise vector by adding constant gaussian noise to the converted latent vector according to the time step. And, in the denoising process, the original latent vector is restored from the random noise vector through an iterative denoising network by injecting the damaged document image as condition. The completely denoised latent vector is converted into a cleaned binarized document image in pixel space through the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

decoder of the autoencoder. The proposed model is suitable for high-quality text region extraction as well as noise removal of damaged document images through an iterative diffusion-denoising process. It is demonstrated to achieve competitive performance through extensive experiments.

3.3. Document Diffusion-Denoising Network

3.3.1. Document Image Compression

The main process of the proposed model is performed in latent space, not in pixel space. To use the latent space, the document image in pixel space is converted into a latent vector through the autoencoder. In this process, we utilize $VQGAN$ [43] using vector quantization layers. This method compresses a high-dimensional image vector through an encoder, and restores the compressed latent vector through a decoder absorbed by a vector quantization (VQ) layer to prevent information loss. In this process, it learns the classification of codebook that determines discrete latent vectors through the VQ layer. That is, by learning $|Z|$, the number of codebooks, the latent space is normalized by the VQ layer. Specifically, given an image of $x \in R^{H\times W\times C}$, the encoder E converts x into the latent vector $z \in R^{h\times w\times c}$, and the decoder D learns the process of restoring z back to $\tilde{x}$. H, W, C mean height, width, and channel of the image vector, and h, w, c mean height, width, and channel of the compressed latent vector, respectively. $VQGAN$ can preserve a specific region of an original image in the latent space by normalizing it to the latent space using the VQ layer. The latent space compression process of the proposed model compresses image data of 256$\times$ 256 size with a single channel into a latent vector of 64$\times$ 64 size with 3 channels. As a result, even if real image vector is converted to a latent vector via $VQGAN$, it is possible to preserve the text, background, and text boundary region like Figure 3.

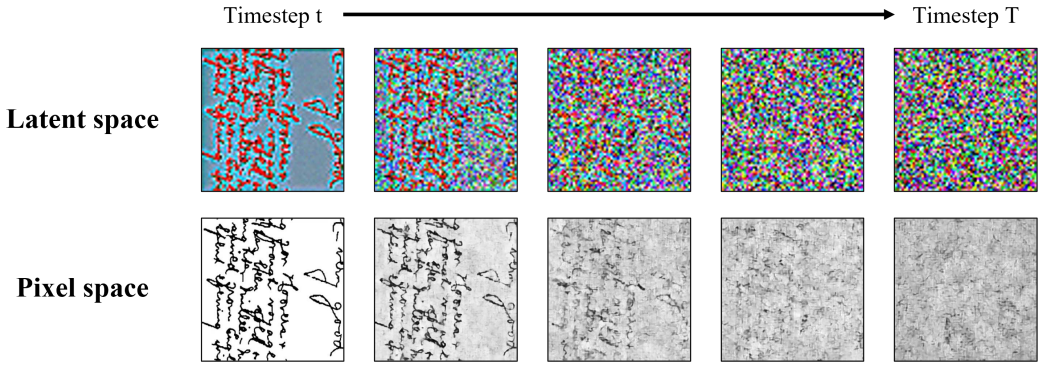


Figure 3. Example images of latent vector compressed through an autoencoder and image vector in pixel space. It shows the change of the vector in each space according to the time step.

3.3.2. Diffusion-Denoising Process

The iterative diffusion-denoising process of the proposed model is based on diffusion models [18,31] and proceeds according to equation 1 and equation 2 in latent space. The diffusion process of the network generates the noise vector z_t by gradually adding the gaussian noise ϵ to the latent vector $z = E(x)$ generated by the encoder. Then, in the denoising process, z_0 is restored by gradually removing noise from the noise vector z_t through the gated U-Net architecture, which is called a denoising network. To this end, training proceeds by calculating the difference in distribution of the outputs of each process at the same time step. Gated U-Net architecture and its process are shown in Figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

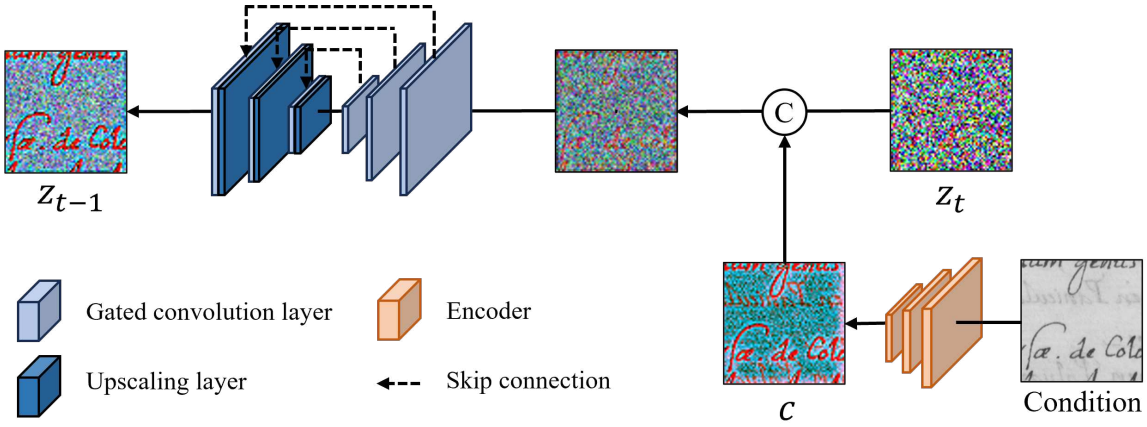


Figure 4. Architecture of denoising process. Our denoising network consists of gated U-Net, which is a fully gated convolution. Denoising process at a specific time step passes through gated U-Net by concatenating the latent vector of the degraded document with the noise vector of the previous time step.

However, in the case of the unconditional diffusion-denoising process, it is trained with data consisting only of binarized document images. Accordingly, a document image following a distribution similar to that of the binarized document image can be generated, but a degraded document image cannot be restored to a cleaned binarized document image. For document binarization, the model should not only generate binarized document image, but also generate clean binarized document image under the condition of degraded document image. Therefore, as shown in Figure 2, we configure the conditional diffusion-denoising process by adding degraded document images as condition. The loss function for training the denoising network ϵ_θ with the condition added is as follows,

$$L_{LDM} = \mathbb{E}_{E(x),c,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, E(c))\|_2^2\right] \tag{4}$$

To perform document binarization, in the denoising process, the latent vector $E(c)$ of the degraded document image c generated through the encoder is channel-wise concatenated as a condition. Also, we use skip connection to prevent information loss. When sampling a cleaned binarized document image, the latent vector of the degraded document image is concatenated with a random noise vector, and the noise is removed through a denoising network iteratively, trained as much as the time step.

3.4. Gated Convolution in Latent Space

We design the denoising network with gated U-Net architecture composed of fully gated convolution for elaborate extraction and seperation of text and background regions. Gated convolution proposed in [19] is a convolution that utilizes a dynamic feature selection mechanism that can learn features in each channel at each spatial location. Unlike the conventional convolution that multiplies the same filter on every spatial location, each location can be distinguished as a valid location and an invalid location by multiplying different spatial locations of the feature map with a weight called a gating value. The gating value is a value that allows more concentration on the text and text boundary region and has a value between 0 and 1 depending on each spatial location. Therefore, by passing the gating value of the feature map calculated in the previous layer to the next layer, it is possible to continuously give a guide on the valid location. In this work, valid location means text and text boundary region. In a specific region (x, y) of a specific channel, the gated convolution can be formulated as,

$$Gating_{x,y} = W_g \cdot I, \tag{5}$$

$$Feature_{x,y} = W_f \cdot I, \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

$$O_{x,y} = \phi(Feature_{x,y}) \odot \sigma(Gating_{x,y}), \tag{7}$$

where W_g and W_f are trainable convolution filters for extracting gating values and features, respectively, and I means an input feature map. σ is the sigmoid function and ϕ is the nonlinear activation functions. Therefore, the final output $O_{x,y}$ is calculated by element-wise multiplication of the value taking the sigmoid on the gating value and the value taking activation on the extracted feature value.

We use this gated convolution to construct a denoising network. In the denoising process, we design the network to extract elaborate text region by guiding information about the text and text boundary regions. The denoising network has a U-Net architecture and utilizes gated convolution in downsampling and upsampling processes. All downsampling consists of gated convolution, and upsampling performs upscaling through a scale factor and downsampling again. Therefore, the denoising network can be viewed as a fully gated convolution. Gated U-Net can effectively distinguish text region from background region in the latent space where each region is preserved by continuously injecting guides for each region through trainable gating values.

It is difficult to maximize the effect of gated U-Net, which consists of fully gated convolution, only with the loss of existing diffusion models that give training direction through the distribution of latent vector. Therefore, pixel-level loss L_{pix} is added for text region extraction and effective updates of gating values. This calculates the difference between the result of the decoder at a specific time step and the ground-truth in pixel space. We add two notable pixel-level losses to update the difference between $\hat{y}$, the output of the denoising network, and the ground-truth y, at a specific time step. The first is the binary cross entropy loss for binary value classification at the pixel level. The second is the dice loss, which computes the similarity between the predicted region of the model and the ground-truth region. Through this, L_{pix} between the predicted pixel $\hat{y}$ and the ground-truth pixel y is configured as follows, and L_{pix} can be formulated as equation 10,

$$L_{bce} = -ylog(\hat{y}) - (1-y)log(1-\hat{y}), \tag{8}$$

$$L_{dice} = 1 - \frac{2y\hat{y}}{y+\hat{y}}, \tag{9}$$

$$L_{pix} = L_{bce} + L_{dice} \tag{10}$$

The proposed model is trained to minimize L_{total}, weighted sum of L_{LDM} and L_{pix}.

$$L_{total} = L_{LDM} + \lambda L_{pix} \tag{11}$$

The network is updated according to equation 11. However, since pixel-level loss can be less effective at lower time step, λ is set as a penalty depending on the time step.

4. Experiments and Results

4.1. Dataset and Implementation details

We build a new training dataset through several existing document binarization datasets to evaluate and compare the performance of the proposed model with other methods. Training dataset consists of H-DIBCO [44–46], DIBCO [47–49], Bickley-diary dataset [50], Persian heritage image binarization dataset (PHIDB) [51], and Synchromedia Multispectral dataset (S-MS) [52]. Since most of the document images have a large size, we divided each image into 256 $\times$ 256 patches for train efficiency. Random rotation augmentation was performed on the divided patches to construct a dataset with a total of about 160k images. 90% of the built dataset was used as a training set, and 10% used as a validation set. Finally, for evaluation, H-DIBCO 2016 [20], DIBCO 2017 [21], H-DIBCO 2018 [22], DIBCO 2019-B [23] is used. The test dataset is not included in the training and validation set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

To train the proposed model, the time step of the diffusion-denoising process is set to 1000, and the AdamW optimizer [53] set to the initial learning rate $lr = 1 \times 10^{-6}$ is used. In addition, the autoencoder was pre-trained for 20 epochs for the train set, and was frozen during denoising network training. We use $LeakyReLU$ as the network's activation function. The denoising network was finally trained for about 1M steps with a batch size of 2. We use NVIDIA RTX 3090 GPU (24GB)×3 for training.

4.2. Evaluation Metrics

For quantitative evaluation of the proposed model, a total of four evaluation metrics [20–23] suitable for document binarization evaluation are selected. Metrics consist of F-measure (FM), pseudo-Fmeasure (pFM), Peak Signal-to-noise Ratio (PSNR), and Distance Reciprocal Distortion (DRD), commonly used in Document Image Binarization Contest (DIBCO).

F-measure is calculated using the precision and recall between the predicted pixel and the ground-truth pixel as,

$$FM = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

pseudo-Fmeasure was proposed in [54], and the stroke predicted through $pseudo\ Recall(pRecall)$ representing the percentage of the skeletonized ground-truth image and the stroke of the ground-truth are calculated as follows,

$$pFM = \frac{2 \times Precision \times pRecall}{Precision + pRecall} \tag{13}$$

PSNR is an image quality evaluation metric and is calculated for the similarity between the predicted image and the ground-truth. PSNR is calculated as follows, where C is the maximum value of an image pixel,

$$PSNR = 10 \log \frac{C^2}{MSE} \tag{14}$$

DRD is a metric proposed by [55] to measure visual distortion in binary document images. It measures the distortion for all the S flipped pixels as follows,

$$DRD = \frac{\sum_{k=1}^{S} DRD_k}{NUBN}, \tag{15}$$

where $NUBN$ is the number of the non-uniform (not all black or white pixels) 8×8 blocks in the ground-truth image, and DRD_k is the distortion of the k-th flipped pixel as defined in [55].

4.3. Quantitative and Qualitative Comparison

To demonstrate the performance and effectiveness of the proposed model, we use a total of four benchmark datasets. DIBCO 2016, DIBCO 2017, and DIBCO 2018 datasets are made up of machine-printed and handwritten document images, and DIBCO 2019-B is a challenging benchmark dataset that includes papyrus-like materials and extreme degradation. We use four evaluation metrics of FM, pFM, PSNR, and DRD introduced in Section. 4.2. In addition, for fair comparison, a total of 8 models including the proposed model are reimplemented. We used publicly available source codes provided by the authors for reimplementation. Methods compared with the proposed model can be divided into traditional binarization algorithm-based methods and deep learning-based methods. Binarization algorithm-based methods include Otsu [6], and Sauvola [5], and deep learning-based methods include SAE [25], cGANs [13], Akbari et al. [27], Souibgui et al. [29], and Suh et al. [15]. Since the proposed model is a generative model, we compare with various generative models. Additionally, the performance of competition winners of each year [20–23] is added to the experimental results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

Table 1 shows the quantitative evaluation results of models. Compared to other methods, the proposed model achieved the best performance for the mean values on all datasets. The proposed model shows best or second best performance in terms of pFM and PSNR on all four datasets. Especially in DIBCO 2019-B dataset, which consists of the most complex and noise that the model has not experienced, the proposed model achieves the best performance in all metrics. In particular, existing methods show great decrement, while the proposed model shows impressive performance gain compared to others even in such severe noise. Although the quantitative differences between the proposed model and the others are marginal on other datasets, qualitative evaluation shows the proposed model performs better at document binarization. Therefore, it demonstrates that training through the iterative diffusion-denoising process is effective in removing complex noise and robust to various environments.

Table 1. Comparison of the proposed model with other methods in (H-)DIBCO dataset. The last row is the result of mean values on four datasets. Best performances are indicated in bold letters and second highest performances are underlined.

Dataset	Metric	Otsu [6]	Sauvola [5]	Competition winner	SAE [25]	cGANs [13]	Akbari [27]	Souibgui [29]	Suh [15]	Ours
2016	FM	86.64	79.57	88.72	88.11	**91.67**	90.48	84.45	91.11	88.95
	pFM	89.99	86.84	91.84	91.55	94.59	93.26	84.73	95.22	**95.45**
	PSNR	17.80	16.90	18.45	18.21	**19.64**	19.27	16.18	19.34	19.38
	DRD	5.52	6.76	3.86	4.51	**2.82**	3.94	7.25	3.25	3.74
2017	FM	80.63	73.86	**91.04**	85.72	90.73	85.59	80.63	89.33	89.36
	pFM	80.85	84.78	92.86	87.85	92.58	87.56	80.85	91.41	**94.02**
	PSNR	13.84	14.30	18.28	16.09	17.83	16.39	13.84	17.91	**18.33**
	DRD	9.85	8.30	**3.40**	6.53	3.58	7.99	9.85	3.83	3.83
2018	FM	51.56	64.04	88.34	75.77	87.73	76.51	77.59	**91.86**	88.43
	pFM	53.58	72.13	90.24	77.95	90.60	80.09	85.74	**96.25**	93.73
	PSNR	9.76	13.98	19.11	14.79	18.37	17.01	16.16	**20.03**	19.28
	DRD	59.07	13.96	4.92	13.30	4.58	8.11	7.93	**2.60**	3.95
2019-B	FM	22.47	50.57	67.99	47.57	61.64	47.00	49.83	66.83	**72.71**
	pFM	22.47	54.48	67.88	48.55	62.52	47.60	49.97	68.32	**75.24**
	PSNR	2.61	10.85	12.14	10.84	11.77	9.18	8.55	12.91	**14.37**
	DRD	213.58	33.73	26.87	32.00	24.11	70.50	53.18	19.80	**14.39**
Mean Values	FM	59.61	67.01	84.02	74.29	82.94	74.90	71.64	84.78	**84.86**
	pFM	61.53	74.56	85.71	76.47	85.07	77.13	71.85	87.80	**89.61**
	PSNR	11.01	14.01	17.00	14.98	16.90	15.46	12.86	17.55	**17.84**
	DRD	73.42	15.69	9.76	14.09	8.77	22.65	23.43	7.37	**6.48**

First, the proposed model for H-DIBCO 2016 dataset shows the highest performance in pFM and the second highest performance in PSNR. In DIBCO 2017 dataset, the proposed model shows the highest performance in pFM and PSNR. The high performance of the proposed model in pFM means that it extracts the text stroke best compared to other methods, and high performance in PSNR means that it produces high-quality results. In H-DIBCO 2018 dataset, the proposed model achieve the second highest performance in all four metrics. Results on DIBCO 2019-B dataset show significantly higher performance in all four metrics compared to other methods. DIBCO 2019-B dataset is a challenging dataset consisting of images with extremely severe degradation on materials such as papyrus and tree bark. Even in this challenging dataset, the proposed model shows the highest performance in all metrics. However, other models show extremely low performance on this dataset compared to other datasets. This demonstrates that the proposed model successfully extracts text strokes even in extremely degraded document images that have not been trained. Average results for DIBCO 2016, 2017, 2018, and 2019-B datasets show that the proposed model achieved the highest performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

in terms of all of the metrics. That is, we demonstrate the effectiveness of the proposed model for performing image generation through diffusion-denoising process and text stroke extraction through gated U-Net.

Qualitative evaluation on H-DIBCO 2016 dataset is shown in Figure 5. Shown image is the most challenging document image with complex degradation in H-DIBCO 2016 dataset. Figure 5g, the result image of [13], shows the highest quantitative result, but shows limitations in preserving elaborate text stroke. Figure 5i, the result image of [15], shows better results in preserving text strokes, but has limitations in removing noise. Figure 5j, the result image of the proposed model, shows that it not only shows high results in noise removal such as overwriting compared to other methods, but also performs best in preserving elaborate and accurate text strokes.

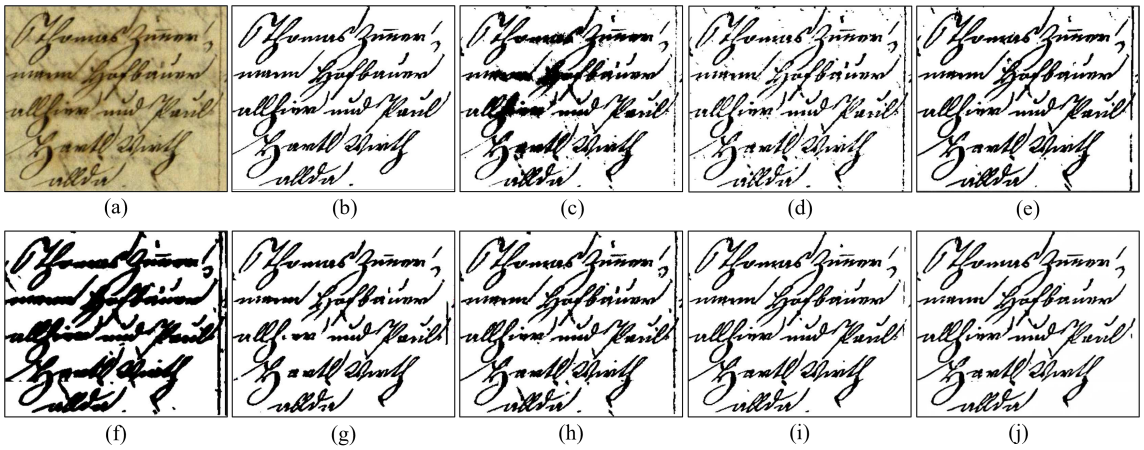


Figure 5. Binarization results of sample image in H-DIBCO 2016 dataset. (a) degraded image, (b) ground-truth, (c) Otsu [6], (d) Sauvola [5], (e) SAE [25], (f) Souibgui et al. [29], (g) cGANs [13], (h) Akbari et al. [27], (i) Suh et al. [15], (j) the proposed model.

Figure 6 is the result image for DIBCO 2017 dataset for qualitative evaluation. The image contains noise that is difficult to distinguish between text and background. Traditional algorithm-based methods suffer from misclassifying such noise as text. In addition, deep learning-based methods also suffer from misclassification in the case of severe noise. It is difficult to distinguish this noise because it has a similar shape, size, and text stroke, but Figure 6j shows that the proposed model successfully distinguishes background and text region.

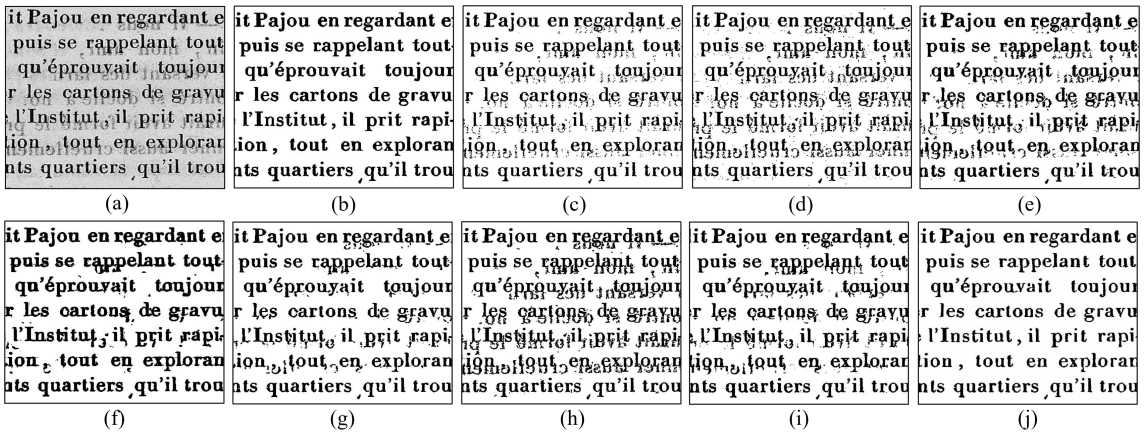


Figure 6. Binarization results of sample image in DIBCO 2017 dataset. (a) degraded image, (b) ground-truth, (c) Otsu [6], (d) Sauvola [5], (e) SAE [25], (f) Souibgui et al. [29], (g) cGANs [13], (h) Akbari et al. [27], (i) Suh et al. [15], (j) the proposed model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

The result image for H-DIBCO 2018 dataset is shown in Figure 7. H-DIBCO 2018 dataset consists of 10 handwritten document images with various degradations such as background intensity, shadow, and bleed-through. Figure 7 includes degradation of variable background intensity and ink smearing. Other methods have difficulty in removing these degradations. In the case of Figure 7g, which is the result image of [13], degradation is relatively resolved well, but it has difficulty in preserving precise text strokes such as small footnotes in the document. In the case of the proposed model in Figure 7i, it solves all types of degradation well and shows high performance in preserving elaborate text strokes.

Figure 7. Binarization results of sample image in H-DIBCO 2018 dataset. (a) degraded image, (b) ground-truth, (c) Otsu [6], (d) Sauvola [5], (e) SAE [25], (f) cGANs [13], (g) Akbari et al. [27], (h) Suh et al. [15], (i) the proposed model.

Figure 8 shows the result image for DIBCO 2019-B dataset. DIBCO 2019-B dataset consists of ancient document images that reflect various types of papyrus quality, ink and handwriting styles. In particular, it is challenging data that includes non-homogeneous properties such as resolution, lighting, and noise. Figure 8c and d, which are results of [13,15], cannot effectively distinguish text from background region according to the intensity of text and background. Because of this problem, it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

is extremely difficult to preserve text strokes. As shown in Figure 8e, the proposed model is effective for text and background classification and successfully preserves the text region. In addition, the types of degradation included in DIBCO 2019-B data are types that have not been experienced in the training data, and we demonstrate the proposed model to be effective even in severe degradation.

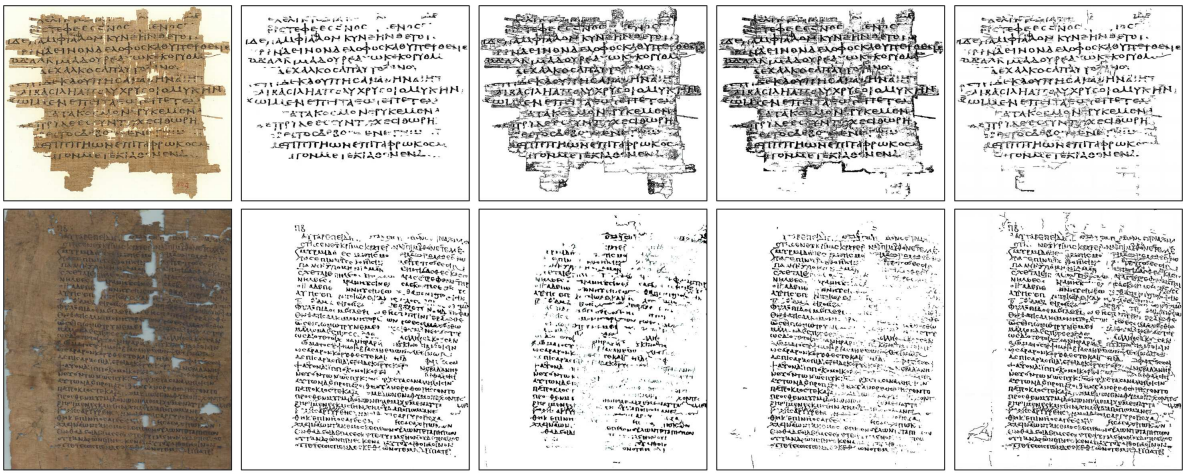


Figure 8. Binarization results of challenging images in DIBCO 2019-B dataset. (a) degraded image, (b) ground-truth, (c) cGANs [13], (d) Suh et al. [15], (e) the proposed model.

Additionally, Figure 9 shows that the proposed model is effective for elaborate region extraction. It is difficult to distinguish between background and text regions and extract accurate and elaborate text strokes from small parts like the example image. In this case, even in methods with relatively high quantitative evaluation [13,15,27] have difficulties in precisely extracting and preserving text regions, as shown in Figure 9c, d and e. However, in the case of the proposed model, as shown in Figure 9f, it successfully extracts and preserves precise text strokes despite these difficulties.

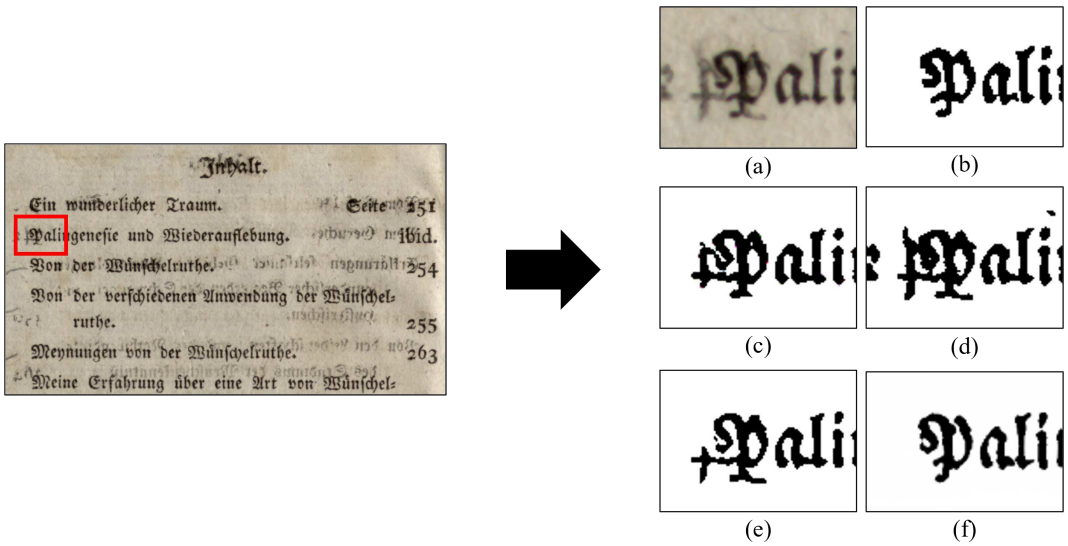


Figure 9. Binarization results of small text with degradation. (a) degraded image, (b) ground-truth, (c) cGANs [13], (d) Akbari et al. [27], (e) Suh et al. [15], (f) the proposed model. It shows that the proposed model effectively preserves text region precisely even in small text.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

4.4. Ablation study

We perform ablation experiments to evaluate each component of the proposed model. For evaluation, DIBCO 2016 benchmark dataset is used and the effectiveness of the proposed model is demonstrated using the four metrics in Section 4.2. The baseline is [17], and the diffusion-denoising process is performed in the latent space using an autoencoder. We compare the baseline with each of the proposed gated U-Net and pixel-level loss, and models with all of the components. A total of four models are compared, and experiments are performed in the same implementation setting except for each component of each model.

Table 2 shows results of the baseline, the proposed model without L_{pix}, the proposed model without gated U-Net, and the proposed model. When gated U-Net is added to the baseline, pFM (95.03) improves by 0.59 compared to the baseline (94.44). Text stroke preservation performance has been improved because filter training for gating values is additionally included. However, when pixel-level loss is not included, the feedback for gated convolution filter is not effective, so there is no significant performance improvement. When gated convolution was excluded, text and background regions are updated through pixel-level loss, resulting in improved performance in all metrics. Finally, in the case of the proposed model with pixel-level loss for training the gated convolution filter in gated U-Net, FM, pFM, PSNR, and DRD improved by 0.68, 1.00, 0.28, and 0.31, respectively, compared to the baseline. This demonstrates that the components of the proposed model are effective respectively, and all components are effectively connected for performance gain.

Table 2. Results of ablation study on H-DIBCO 2016 dataset. Best performances are indicated in bold letters.

H-DIBCO 2016	FM	pFM	PSNR	DRD
Baseline	88.25	94.44	19.09	4.05
Ours w/o L_{pix}	88.13	95.03	19.08	4.05
Ours w/o Gated U-Net	88.48	95.03	19.08	3.93
Ours	**88.93**	**95.44**	**19.37**	**3.74**

Additionally, to confirm the effect of each component, the results of model training after six epochs are shown in Table 3. We adopt FM which verifies the predicted text and background pixel-wise, and pFM which verifies each prediction region. When pixel-level loss is excluded from the proposed model, each metric is similar to or slightly increased from the baseline. The proposed model without gated U-Net indicates the model with the baseline's backbone network using LDM and pixel-level loss to update denoising network. This results in a significant improvement of FM by 8.00 and pFM by 7.94 compared to the baseline. In the case of the proposed model using gated U-Net for document binarization and pixel-level loss aimed at training gating values, it achieves the highest performance of each metric and shows a significant improvement with FM by 8.22 and pFM by 8.51 compared to the baseline. These results demonstrate that the components of the proposed model are effectively suitable for text stroke preservation and extraction.

Table 3. Results of ablation study on H-DIBCO 2016 dataset after training for 6 epochs. Best performances are indicated in bold letters.

H-DIBCO 2016 / epoch 6	FM	pFM
Baseline	77.93	85.66
Ours w/o L_{pix}	79.53	85.84
Ours w/o Gated U-Net	85.93	93.60
Ours	**86.15**	**94.17**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2023 doi:10.20944/preprints202308.2048.v1

Figure 10 shows the result images of the baseline, the proposed model without L_{pix}, the proposed model without gated U-Net, and the proposed model. We confirm that Figure 10f, a result of the proposed model with both L_{pix} and gated U-Net, is successful in distinguishing text and background of degraded document images compared to Figure 10c, a result of baseline.

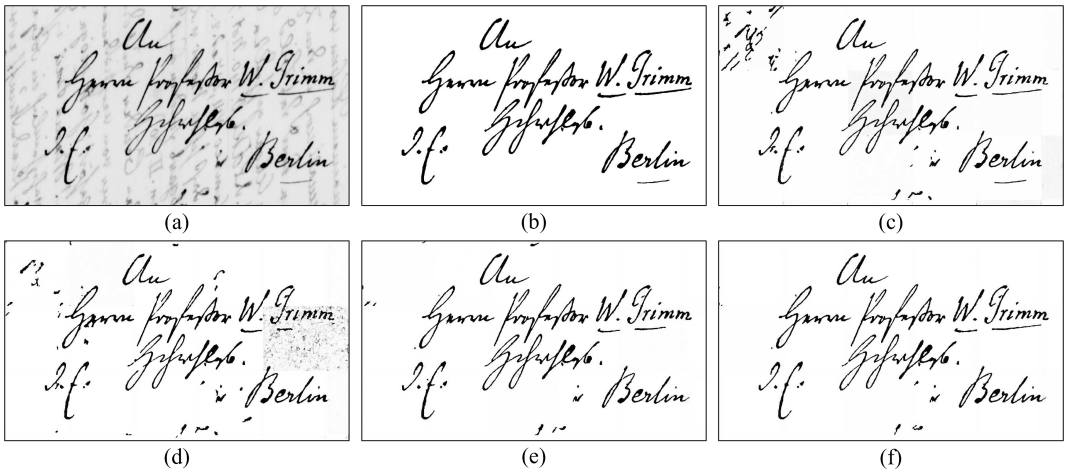


Figure 10. Binarization results of sample image in DIBCO 2016. (a) degraded image, (b) ground-truth, (c) baseline, (d) Ours w/o L_{pix}, (e) Ours w/o Gated U-Net, (f) Ours.

Additionally, in Figure 11, the elaborate text stroke preservation performance can be qualitatively confirmed. In Figure 11f, which is the result of the proposed model, noise removal and text stroke preservation are more accurate and precise than the baseline shown in Figure 11c. This proves that the diffusion-denoising process through gated U-Net and L_{pix} is effective in preserving more elaborate text strokes and generates high-quality results.

Figure 11. Binarization results of sample image to confirm the effect of preserving text region in DIBCO 2016. (a) degraded image, (b) ground-truth, (c) baseline, (d) Ours w/o L_{pix}, (e) Ours w/o Gated U-Net, (f) Ours.

5. Conclusion

In this paper, we propose a new document binarization model based on latent diffusion models that enables elaborate text stroke extraction and is robust against unexperienced degradation. The

diffusion-denoising process in latent space is utilized for the first time to generate high-quality cleaned binarized document image. In addition, denoising network is constructed with gated U-Net consisting of a fully gated convolution for text stroke preservation of severe degraded document images. In this process, we add a filter to update the gating value to preserve text region. Also, by using pixel-level loss to effectively update the filter for the gating value, text strokes are effectively preserved even in extremely degraded images. We confirmed this strength through extensive experiments on challenging benchmark datasets. According to the quantitative experimental results, the proposed model shows impressive performance gain in text stroke preservation and extraction compared to existing methods in (H-)DIBCO dataset. Through various qualitative evaluations, we confirm that the proposed model overcomes the limitations of existing methods and successfully performs text and background classification. In future research, we expect that the proposed model can be used for various tasks that utilize binary masks.

Author Contributions: Conceptualization, S.H. and J.R.; methodology S.H.; software, S.H.; validation, S.H., S.J.; investigation, S.H.; writing—original draft preparation, S.H., S.J.; writing—review and editing, S.H., S.J.; visualization, S.H. and S.J.; supervision, J.R.; project administration, J.R.; funding acquisition, J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative R&D program. (Project No. P0016096)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sulaiman, A.; Omar, K.; Nasrudin, M.F. Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *Journal of imaging* **2019**, *5*, 48.
2. Farahmand, A.; Sarrafzadeh, H.; Shanbehzadeh, J. Document image noises and removal methods **2013**.
3. Mustafa, W.A.; Kader, M.M.M.A. Binarization of document images: A comprehensive review. Journal of Physics: Conference Series. IOP Publishing, 2018, Vol. 1019, p. 012023.
4. Chauhan, S.; Sharma, E.; Doegar, A.; others. Binarization techniques for degraded document images—A review. 2016 5th international conference on reliability, infocom technologies and optimization (Trends and Future Directions)(ICRITO). IEEE, 2016, pp. 163–166.
5. Sauvola, J.; Seppanen, T.; Haapakoski, S.; Pietikainen, M. Adaptive document binarization. Proceedings of the fourth international conference on document analysis and recognition. IEEE, 1997, Vol. 1, pp. 147–152.
6. Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* **1979**, *9*, 62–66.
7. Niblack, W. *An introduction to digital image processing*; Strandberg Publishing Company, 1985.
8. He, S.; Schomaker, L. DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern recognition* **2019**, *91*, 379–390.
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
11. Westphal, F.; Lavesson, N.; Grahn, H. Document image binarization using recurrent neural networks. 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). IEEE, 2018, pp. 263–268.
12. Tensmeyer, C.; Martinez, T. Document image binarization with fully convolutional neural networks. 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE, 2017, Vol. 1, pp. 99–104.
13. Zhao, J.; Shi, C.; Jia, F.; Wang, Y.; Xiao, B. Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition* **2019**, *96*, 106968.

14. De, R.; Chakraborty, A.; Sarkar, R. Document image binarization using dual discriminator generative adversarial networks. *IEEE Signal Processing Letters* **2020**, *27*, 1090–1094.
15. Suh, S.; Kim, J.; Lukowicz, P.; Lee, Y.O. Two-stage generative adversarial networks for binarization of color document images. *Pattern Recognition* **2022**, *130*, 108810.
16. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.
17. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
18. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.
19. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4471–4480.
20. Pratikakis, I.; Zagoris, K.; Barlas, G.; Gatos, B. ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2016, pp. 619–623.
21. Pratikakis, I.; Zagoris, K.; Barlas, G.; Gatos. ICDAR2017 competition on document image binarization (DIBCO 2017). 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, Vol. 1, pp. 1395–1403.
22. Pratikakis, I.; Zagori, K.; Kaddas, P.; Gatos, B. ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 489–493. doi:10.1109/ICFHR-2018.2018.00091.
23. Pratikakis, I.; Zagoris, K.; Karagiannis, X.; Tsochatzidis, L.; Mondal, T.; Marthot-Santaniello, I. ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1547–1556. doi:10.1109/ICDAR.2019.00249.
24. Peng, X.; Cao, H.; Natarajan, P. Using convolutional encoder-decoder for document image binarization. 2017 14th IAPR international conference on document analysis and recognition (ICDAR). IEEE, 2017, Vol. 1, pp. 708–713.
25. Calvo-Zaragoza, J.; Gallego, A.J. A selectional auto-encoder approach for document image binarization. *Pattern Recognition* **2019**, *86*, 37–47.
26. Kang, S.; Iwana, B.K.; Uchida, S. Complex image processing with less data—Document image binarization by integrating multiple pre-trained U-Net modules. *Pattern Recognition* **2021**, *109*, 107577.
27. Akbari, Y.; Al-Maadeed, S.; Adam, K. Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images. *IEEE Access* **2020**, *8*, 153517–153534.
28. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
29. Souibgui, M.A.; Kessentini, Y. De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *44*, 1180–1191.
30. Lin, Y.S.; Ju, R.Y.; Chen, C.C.; Lin, T.Y.; Chiang, J.S. Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks. *arXiv preprint arXiv:2211.16098* **2022**.
31. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* **2020**.
32. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.
33. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. International Conference on Medical Imaging with Deep Learning. PMLR, 2022, pp. 1336–1348.
34. Kim, B.; Oh, Y.; Ye, J.C. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566* **2022**.
35. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788* **2022**.
36. Duan, Y.; Guo, X.; Zhu, Z. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021* **2023**.

37. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; Yang, K. Gated fully fusion for semantic segmentation. Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 11418–11425.
38. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing* **2017**, *9*, 446.
39. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. International conference on machine learning. PMLR, 2017, pp. 933–941.
40. Lin, X.; Ma, L.; Liu, W.; Chang, S.F. Context-gated convolution. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, 2020, pp. 701–718.
41. Zhang, Y.; Fang, J.; Chen, Y.; Jia, L. Edge-aware U-net with gated convolution for retinal vessel segmentation. *Biomedical Signal Processing and Control* **2022**, *73*, 103472.
42. Kwon, M.; Jeong, J.; Uh, Y. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960* **2022**.
43. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.
44. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. H-DIBCO 2010-handwritten document image binarization competition. 2010 12th International Conference on Frontiers in Handwriting Recognition. IEEE, 2010, pp. 727–732.
45. Pratikakis, I.; Gatos, B.; Ntirogiannis. ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). 2012 international conference on frontiers in handwriting recognition. IEEE, 2012, pp. 817–822.
46. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). 2014 14th International conference on frontiers in handwriting recognition. IEEE, 2014, pp. 809–813.
47. Gatos, B.; Ntirogiannis, K.; Pratikakis, I. ICDAR 2009 document image binarization contest (DIBCO 2009). 2009 10th International conference on document analysis and recognition. IEEE, 2009, pp. 1375–1382.
48. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1506–1510. doi:10.1109/ICDAR.2011.299.
49. Pratikakis, I.; Gatos, B.; Ntirogiannis. ICDAR 2013 document image binarization contest (DIBCO 2013). 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013, pp. 1471–1476.
50. Deng, F.; Wu, Z.; Lu, Z.; Brown, M.S. Binarizationshop: a user-assisted software suite for converting old documents to black-and-white. Proceedings of the 10th annual joint conference on Digital libraries, 2010, pp. 255–258.
51. Nafchi, H.Z.; Ayatollahi, S.M.; Moghaddam, R.F.; Cheriet, M. An efficient ground truthing tool for binarization of historical manuscripts. 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013, pp. 807–811.
52. Hedjam, R.; Cheriet, M. Historical document image restoration using multispectral imaging system. *Pattern Recognition* **2013**, *46*, 2297–2312.
53. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**.
54. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing* **2012**, *22*, 595–609.
55. Lu, H.; Kot, A.C.; Shi, Y.Q. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters* **2004**, *11*, 228–231.