applied sciences

MDPI

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, Republic of Korea; hsk0314@dgu.ac.kr (S.H.); voiagerd@dgu.ac.kr (S.J.)
* Correspondence: jtrhee@dongguk.edu

Abstract: The binarization of degraded documents represents a crucial preprocessing task for various document analyses, including optical character recognition and historical document analysis. Various convolutional neural network models and generative models have been used for document binarization. However, these models often struggle to deliver generalized performance on noise types the model has not encountered during training and may have difficulty extracting intricate text strokes. We herein propose a novel approach to address these challenges by introducing the use of the latent diffusion model, a well-known high-quality image-generation model, into the realm of document binarization for the first time. By leveraging an iterative diffusion-denoising process within the latent space, our approach excels at producing high-quality, clean, binarized images and demonstrates excellent generalization using both data distribution and time steps during training. Furthermore, we enhance our model's ability to preserve text strokes by incorporating a gated U-Net into the backbone network. The gated convolution mechanism allows the model to focus on the text region by combining gating values and features, facilitating the extraction of intricate text strokes. To maximize the effectiveness of our proposed model, we use a combination of the latent diffusion model loss and pixel-level loss, which aligns with the model's structure. The experimental results on the Handwritten Document Image Binarization Contest and Document Image Binarization Contest benchmark datasets showcase the superior performance of our proposed model compared to existing methods.

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

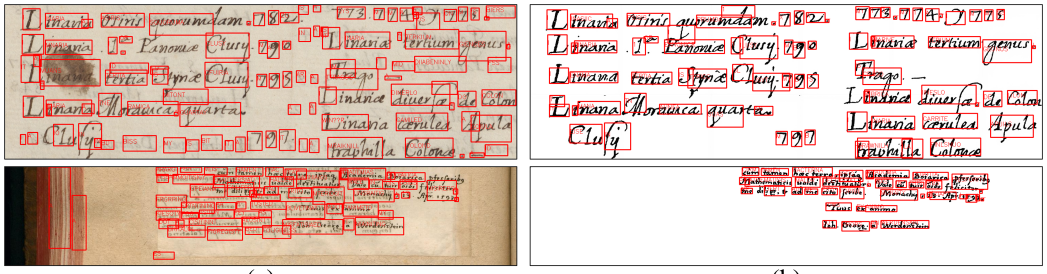
Citation: Han, S.; Ji, S.; Rhee, J. Diffusion-Denoising Process with Gated U-Net for High-Quality Document Binarization. *Appl. Sci.* **2023**, *13*, 11141. https://doi.org/10.3390/app132011141

Academic Editor: Jesús B. Alonso-Hernández

Received: 11 September 2023
Revised: 29 September 2023
Accepted: 8 October 2023
Published: 10 October 2023

1. Introduction

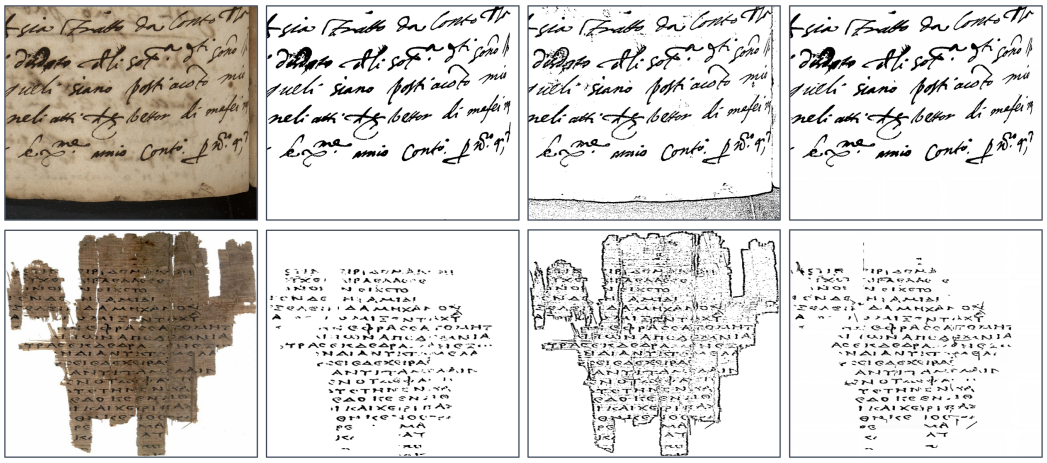
Document images play a crucial role in digital document analysis, encompassing tasks like optical character recognition, historical document restoration, and document classification [1]. However, real-world document images often suffer from degradation caused by a multitude of noise sources, including shadows, stains, ink smears, bleed-through, overwriting, and variable background intensity [2]. These instances of degraded document images significantly impede various aspects of digital document analysis. Therefore, to effectively perform digital document analysis, the preprocessing of degraded document images is essential. Document binarization stands out as a pivotal preprocessing task, aiming to produce pristine, binarized images by restoring text regions from degraded documents [3]. Document binarization goes beyond the removal of specific noise elements; it comprehensively addresses a variety of degradation factors [4]. As illustrated in Figure 1, document binarization plays a vital role in safeguarding the performance of document-related tasks such as optical character recognition against the detrimental impact of noise.



(a) (b)

Figure 1. OCR performance improvement as a result of document binarization. (**a**) OCR results for a document without binarization. (**b**) OCR results for a document with binarization using the proposed model. Document binarization is effective in resolving performance degradation in various document tasks.

However, converting a degraded document containing non-uniform noise into a clean, binarized document presents challenges that traditional binarization algorithms struggle to resolve, as shown in Figure 2. Furthermore, preserving intricate text strokes during this process comes with difficulties. Hence, research dedicated to the binarization of degraded document images is of paramount importance.



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

Figure 2. Images highlighting the difficulties of preserving text strokes and addressing significant degradation. (**a**) Significantly degraded document images. (**b**) Ground-truth images. (**c**) Resulting images of the model in [5], a traditional binarization algorithm-based method. (**d**) Resulting images of the proposed model, a deep learning-based method. The proposed model is effective in preserving text strokes, even in the case of significant degradation.

Research into document binarization spans several decades. During this time, various image binarization algorithms [5–7] have emerged, significantly contributing to the binarization of standard, uniform images. Nevertheless, they face limitations when confronted with the task of distinguishing non-uniform and complex degradation while extracting detailed text regions in degraded document images [8]. To address these challenges, recent efforts have turned to deep learning techniques within the realm of computer vision for document binarization. Methods rooted in deep learning [9–11], including segmentation and deep neural networks, have gained prominence. For instance, in one model, document binarization was redefined as a pixel classification problem by applying an FCN (Fully Convolutional Network) [12]. Additionally, the performance of document binarization was enhanced by training an iterative neural network, combining existing binary algorithms with deep learning [8]. Variations of neural network structures have also been proposed for document binarization. Peng et al. [13] performed document binarization by applying

multi-resolutional attention to an encoder–decoder network. In [14], a U-Net architecture with global–local branches was employed, which utilized both the global and local features of degraded documents. However, these deep neural network models face a challenge in that they cannot effectively extract both global and local features when applying the same convolutional filter across the entire image.

More recently, document binarization based on generative adversarial networks (GANs) has emerged. GANs have been successfully applied to image-to-image tasks to generate clean, binarized document images from their degraded counterparts [15–18]. Souibgu et al. [17] succeeded in removing watermarks and blur using a conditional GAN architecture. Also, a two-stage GAN architecture involving document image enhancement and binarization has been proposed [18]. These deep learning-based approaches have overcome the limitations of traditional binarization algorithms and exhibited improvements in text-region preservation. Nonetheless, these methods still face the challenge of extracting intricate text strokes from non-uniform and complex degradation, occasionally encountering mode collapse [19], which hampers performance due to a focus on specific data distributions. Specifically, these methods are robust against trained noise but do not perform well against untrained noise.

To address the aforementioned issues, we propose an innovative approach to document binarization using a diffusion-denoising process combined with a gated U-Net. Our model transmits the probability of the text region as it passes through the layers of the gated U-Net, ultimately generating high-quality binarized document images through an iterative diffusion-denoising process.

We present a document binarization model based on the latent diffusion model (LDM) [20], conceptualizing document binarization as an iterative diffusion-denoising process, as shown in Figure 3. In essence, this process generates a clean, binarized document image by progressively eliminating Gaussian noise with each time step. To achieve this, we use the diffusion model in [21], which has proven effective in image-generation tasks. The model operates by introducing and removing Gaussian noise within an image, and it offers the advantage of reduced reliance on extensive training data, learning to remove noise by considering both data distribution and time steps. LDMs have further improved time efficiency and precision in image generation using the diffusion model within the latent space [20]. By applying an LDM to document binarization, we enhance the generalization capabilities against unseen noise and finely adjust the text-stroke features using the latent space. To the best of our knowledge, this is the first study to incorporate an LDM into document binarization.

Moreover, we integrate a gated convolution into the model's backbone network to facilitate intricate text-stroke extraction. Gated convolutions have previously shown their effectiveness in binary mask learning within segmentation models [22]. In our study, the gated convolution proves adept at distinguishing between text and background regions by jointly training features through original convolution and gating values, which encode text-region information. We use the gated convolution within the latent space rather than in the pixel space, allowing the proposed model to extract even more detailed text strokes.

To demonstrate the effectiveness of our proposed model, we conduct experiments using the Handwritten Document Image Binarization Contest (H-DIBCO) and Document Image Binarization Contest (DIBCO) datasets [23–26]. Our evaluation employs the four most commonly used metrics in document binarization, and the proposed model outperforms existing state-of-the-art methods in all metrics. In summary, our study 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 study to incorporate an LDM into document binarization to generate high-quality, clean, binarized document images.

- We use a gated convolution in the latent space for the precise extraction of text strokes. This approach simplifies the differentiation between the text and the background by continually updating the gating value as a guide for delineating the text region.
- The proposed model produces high-quality, clean, binarized document images and outperforms existing methods on multiple H-DIBCO datasets.

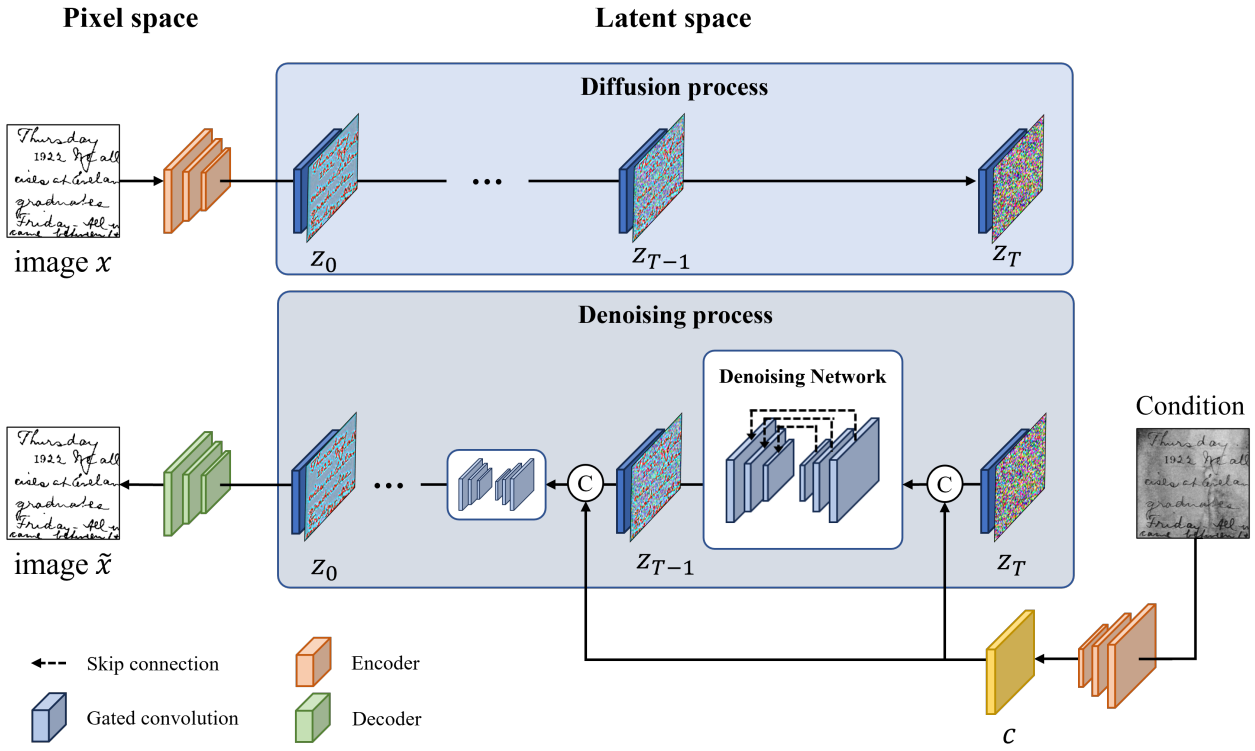


Figure 3. The architecture of the proposed model for document binarization. The network architecture can be divided into the pixel space and latent space, and the diffusion-denoising process proceeds according to the time step in the latent space.

The remainder of this paper is structured as follows. In Section 2, we review some related works. Section 3 presents the proposed model. Section 4 discusses the experimental results, and in Section 5, we conclude our study.

2. Related Works

2.1. Document Binarization

Document binarization aims to conduct pixel-wise binary classification of the background and text within a document image. The methods for document binarization can be categorized into two main groups: traditional binarization algorithm-based methods and deep learning-based methods. Traditional binarization algorithms rely on various nonparametric thresholding techniques. These methods achieve binary classification by applying a threshold to distinguish between the background and text. Otsu's method [6], for instance, uses a global thresholding approach that maximizes the separation between the background and text, seeking the highest interclass variation for binary classification. These global thresholding methods have proven effective for document images with uniform degradation but may not be well suited for document images with non-uniform degradation. To address this challenge, pixel-wise local thresholding methods have been introduced. Local thresholding methods enhance classification accuracy by considering the relationships between neighboring pixels based on local statistical information [5,7]. Although these algorithm-based binarization methods excel at handling uniform noise, they

still encounter challenges when dealing with documents containing diverse non-uniform noise such as overwriting [8].

To address these issues, deep learning-based methods have gained significant attention recently. An iterative neural network was developed to generate clean, binarized images from non-uniformly degraded images, and Otsu's algorithm was applied to enhance the resulting images [8]. Binarization was achieved by training mid-level representations through an encoder–decoder architecture composed of convolutional neural networks [27]. This encoder–decoder architecture was also used to construct a network for selective output [28]. A cascading U-Net architecture was proposed for tackling complex document image-processing tasks [29]. Akbari et al. [30] leveraged convolutional neural networks to identify foreground pixels using input-generated multichannel images. Furthermore, generative models based on GANs, which treat document binarization as an image-to-image task, have been introduced [31].

Zhao et al. [15] proposed a model for generating clean, binarized images through multi-scale information fusion based on conditional generative adversarial networks (cGANs). Additionally, in [17], cGANs were employed, demonstrating performance improvements in tasks such as watermark removal, deblurring, and binarization. Lin et al. [32] introduced a three-stage method for binarization by integrating a discrete wavelet transform and GANs. Lastly, color-independent adversarial networks have been developed in two stages to capture both the global and local features of documents [18].

2.2. Diffusion Model for Image-to-Image Tasks

Diffusion models have shown remarkable success in image generation and have found applications in various image-generation tasks [19,21,33]. Rombach et al. [20] harnessed the power of the diffusion method within the latent space to finely adjust the semantic features of images, resulting in improved quality of tasks such as inpainting, super-resolution, and image-to-image transformations. Additionally, diffusion models have been used in image-segmentation tasks. A distribution of segmentation masks was generated using a stochastic sampling process [34]. Kim et al. [35] introduced a diffusion adversarial representation learning model, incorporating switchable spatially adaptive denormalization for vessel segmentation. Chen et al. [36] used the diffusion process to refine detection-box proposals in object detection, while another study used the diffusion process for image-depth estimation [37].

2.3. Gated Convolutions

Gated convolutions have found applications in various tasks, including segmentation [38,39], inpainting [22], and language modeling [40]. Li et al. [38] introduced the concept of gated full fusion, which selectively integrates multi-level features using gated convolutions in a fully connected manner for segmentation. Another study proposed context-gated convolutions, which dynamically adjust the weights of convolutional layers based on the global context [41]. Zhang et al. [42] leveraged gated convolutions for vessel segmentation. Their network learned to accentuate vessel edges using gated convolutions on features extracted via an encoder–decoder architecture. In [22], a feature selection mechanism capable of dynamically learning features for each spatial location was proposed to address the limitations of vanilla convolutions, which use the same filter for all input pixels. A gated convolution was used for this dynamic feature selection mechanism, effectively distinguishing between valid and invalid pixels.

3. Method

We introduce a binarization model for degraded document images through an iterative diffusion-denoising process. Within this process, we leverage an LDM [20], which learns data distribution features by iteratively introducing and removing Gaussian noise at each time step in an image. The iterative diffusion-denoising process operates within the latent space using a pre-trained autoencoder to generate a clean, binarized document image

through the autoencoder's decoder. Additionally, by incorporating a gated convolution into the denoising process, we enhance text-stroke extraction performance by dynamically updating the gating value associated with the text regions.

In this section, we first provide an overview of the existing diffusion model in the preliminaries, followed by an explanation of the architecture of our proposed model. We then delve into the conditioned denoising process facilitated by the gated U-Net. Finally, we elucidate the loss function used in our proposed model.

3.1. Preliminaries

Diffusion models [21,33] are probabilistic models that aim to estimate a data distribution $p(x)$ by iteratively denoising noise from a normally distributed variable. The model performs several generative tasks via 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 follows:

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

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 via a denoising network. Through this iterative process, x_0 is generated. 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 that generates 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 a trained denoising network from the random noise vector x_T.

3.2. Network Architecture

The network architecture of the proposed model is based on the structure of LDMs [20], as shown in Figure 3. Throughout the training process, an input binarized document image undergoes conversion into a latent vector within the pixel space, achieved by the encoder of a pretrained autoencoder. This use of the latent space enables precise adjustments to the semantic features of the latent vector [20,43,44]. Subsequently, the diffusion process generates a Gaussian noise vector by incrementally introducing constant Gaussian noise to the converted latent vector based on the time step. In the denoising phase, the original latent vector is restored from the random noise vector via an iterative denoising network, with the damaged document image serving as a conditioning factor. The fully denoised latent vector is then translated back into a clean, binarized document image in the pixel space, facilitated by the decoder of the autoencoder. The proposed model is suitable for high-quality text-region extraction and the removal of noise from damaged document images, which is achieved through an iterative diffusion-denoising process. Extensive experiments demonstrate its ability to deliver competitive performance.

3.3. Document Diffusion-Denoising Network

3.3.1. Document Image Compression

The primary process of the proposed model takes place in the latent space rather than in the pixel space. To use the latent space, the document image in the pixel space is transformed into a latent vector through the autoencoder. In this process, we use

$VQGAN$ [44] through vector quantization (VQ) layers. This method compresses a high-dimensional image vector using an encoder and subsequently reconstructs the compressed latent vector through a decoder, incorporating a VQ layer to prevent information loss. In this process, it learns the classification of a codebook that regulates 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.

In detail, 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 to $\tilde{x}$. H, W, and C refer to the height, width, and channel of the image vector, and h, w, and c refer to the height, width, and channel of the compressed latent vector, respectively. $VQGAN$ can preserve a particular 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 reduces image data with dimensions of 256 × 256 and a single channel into a latent vector with dimensions of 64 × 64 and three channels. Consequently, even if the real image vector is converted into a latent vector via $VQGAN$, it is possible to preserve the text, background, and text boundary region, as shown in Figure 4.

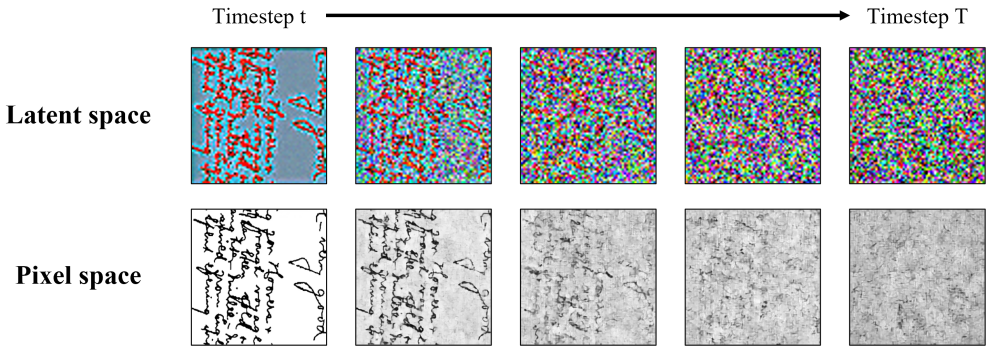


Figure 4. Exemplary images of a latent vector compressed through an autoencoder and image vector in the pixel space, showing the change in the vector in each space according to the time step.

3.3.2. Diffusion-Denoising Process

We use a diffusion-denoising process to create high-quality, clean, binarized document images. Initially, within the diffusion process, Gaussian noise is injected into the image's latent vector over multiple time steps to form a complete Gaussian noise vector. Subsequently, in the denoising process, the model undergoes training to progressively eliminate Gaussian noise from the complete Gaussian noise vector, ultimately producing a clean, binarized document image. Therefore, during the document binarization inference, the result is exclusively generated through the trained denoising process, as shown in Figure 5.

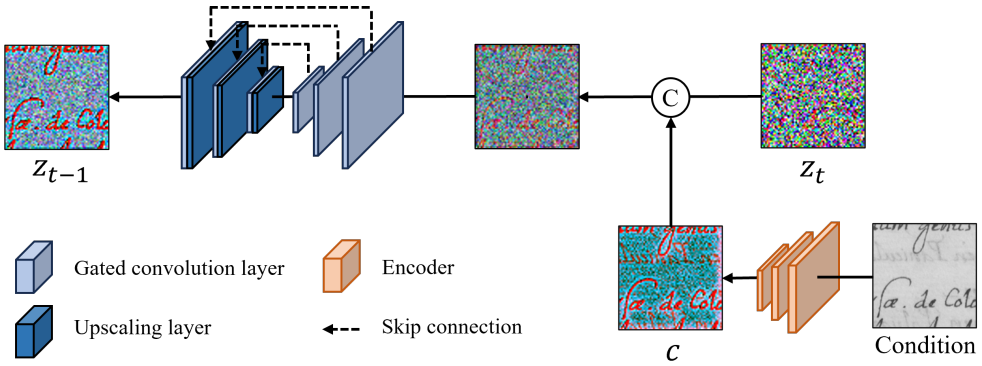


Figure 5. Illustration of the architecture of the denoising process. Our denoising network consists of a gated U-Net, which is a fully gated convolution. The denoising process at a specific time step passes through a gated U-Net by concatenating the latent vector of the degraded document with the noise vector of the previous time step.

The iterative diffusion-denoising process of the proposed model depends on the diffusion model [21,33] and proceeds according to Equations (1) and (2) in the latent space to enhance the computational cost and feature control ability, as described in Section 3.3.1. 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 known as the denoising network. To achieve this, the training process involves computing the difference in the distribution of the outputs from each process at the same time step. The gated U-Net architecture and its corresponding processes are illustrated in Figure 5.

However, in the case of the unconditional diffusion-denoising process, it is trained using data that solely comprise binarized document images. As a result, it can generate a document image following a distribution similar to that of a binarized document image; however, it cannot restore a degraded document image to a clean, binarized state. For document binarization, the model must generate not only a binarized document image but also a clean, binarized document image when presented with a degraded document image. Hence, as illustrated in Figure 3, we configure the conditional diffusion-denoising process by adding degraded document images as a condition. The loss function for training the denoising network ϵ_θ with the condition added is as follows:

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

In the document binarization process, during the denoising step, the latent vector $E(c)$ of the degraded document image c generated by the encoder is concatenated channel-wise as a condition. Additionally, we implement skip connections to mitigate information loss. When sampling a cleaned, binarized document image, the latent vector of the degraded document image is combined with a random noise vector. This noise is progressively removed through an iteratively trained denoising network, corresponding to the number of time steps.

3.4. Gated U-Net in the Latent Space

We have designed the denoising network using a gated U-Net architecture that incorporates a fully gated convolution to facilitate the detailed extraction and separation of the text and background regions. The U-Net architecture [10] leverages both the local and global features by combining the extraction of high-dimensional features with low-dimensional features through skip connections between the upsampling and downsampling stages. The gated convolution, as introduced in [22], represents a convolutional operation that integrates a dynamic feature selection mechanism capable of learning features in each channel at every spatial location.

In contrast to a traditional convolution, which applies the same filter to all spatial positions, a gated convolution multiplies each spatial position in the feature map by a distinct weight, referred to as a gating value. This gating value distinguishes each position as either valid or invalid. As it traverses through the layer, the gating value places more emphasis on the text and text boundary regions, assuming values 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 provide a guide on the valid location. In the present study, the valid location means the text and the text boundary regions. In a specific region (x, y) of a specific channel, the gated convolution can be formulated as follows:

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

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

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

where W_g and W_f are trainable convolution filters for extracting gating values and features, respectively, and I refers to an input feature map. σ is the sigmoid function, and ϕ is the nonlinear activation function. Therefore, the final output $O_{x,y}$ is determined through element-wise multiplication of the value obtained by applying the sigmoid function to the gating value and the value acquired by applying activation to the extracted feature value. By applying the sigmoid function to the gating value and subsequently multiplying it with the original feature, we ascertain the probability of a valid location.

We use this gated convolution method to build a denoising network. Within the denoising process, we have designed the network to meticulously extract text regions by incorporating information pertaining to both the text itself and the text boundary regions. The denoising network has a U-Net architecture and uses a gated convolution in the downsampling and upsampling processes. All downsampling processes involve a gated convolution, whereas the upsampling process entails upscaling via a scale factor followed by another downsampling step. Consequently, the denoising network can be conceptualized as a fully gated convolution. Moreover, we use skip connections between upsampling and downsampling to facilitate information sharing. In the latent space, the gated U-Net effectively discriminates between the text and background regions, preserving each region by continuously incorporating trainable gating values as guides.

Maximizing the effectiveness of a gated U-Net, comprising a fully gated convolution solely through the loss of existing diffusion models that provide training guidance based on the distribution of the latent vector, can be challenging. Hence, we introduce a pixel-level loss (L_{pix}) to enhance text-region extraction and facilitate effective updating of the gating values. L_{pix} calculates the disparity between the output of the decoder at a specific time step and the ground truth in the pixel space. We incorporate two noteworthy pixel-level losses to refine 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 measures the similarity between the model's predicted region 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 using 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}, a 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 the pixel-level loss can be less effective at a 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 built a new training dataset by combining several existing document binarization datasets to evaluate and compare the performance of our proposed model against other methods. The training dataset comprised images from H-DIBCO [45–47], DIBCO [48–50], the Bickley diary dataset [51], the Persian heritage image binarization (PHIDB) dataset [52], and the Synchromedia Multispectral (S-MS) dataset [53]. Given that most of the document

images were of substantial size, we patchified each image into 256 × 256 patches to enhance training efficiency. To diversify the dataset, we applied random rotation augmentation to the segmented patches, resulting in a dataset totaling approximately 160,000 images. For training purposes, 90% of the constructed dataset was allocated for training and the remaining 10% was reserved for validation. For evaluation, we used H-DIBCO 2016 [23], DIBCO 2017 [24], H-DIBCO 2018 [25], and DIBCO 2019-B [26]. These datasets comprise 10–20 document images, each subjected to various levels of general noise, with dimensions typically exceeding 1000 × 1000 pixels. Additionally, we incorporated DIBCO 2019-B, encompassing 10 images featuring novel noise characteristics that the model had not previously encountered. Notably, the test dataset was not included in the training and validation set.

To train the proposed model, the time step of the diffusion-denoising process was set to 1000, and the AdamW optimizer [54], with an initial learning rate of $lr = 1 \times 10^{-6}$, was used. Furthermore, the autoencoder was pre-trained with an 8192-sized codebook at a learning rate of $lr = 4.5 \times 10^{-6}$ over 20 epochs for the training set and was frozen during denoising network training. During the diffusion-denoising process, the latent space was utilized through the 8192-sized codebook of the pre-trained autoencoder. We used *LeakyReLU* as the network's activation function. The denoising network underwent training for about 1M steps with a batch size of 2. We used three NVIDIA RTX 3090 GPUs (24GB) for training. The network consisted of 590M parameters, and the model inference was carried out with 12GB of VRAM.

4.2. Evaluation Metrics

For the quantitative evaluation of the proposed model, four evaluation metrics [23–26] suitable for document binarization evaluation were selected. The metrics included the F-measure (FM), pseudo-F-measure (pFM), Peak Signal-to-Noise Ratio (PSNR), and distance Reciprocal Distortion (DRD), commonly used in DIBCO. The results of the proposed model and existing models were compared with the ground-truth binarized images to calculate the selected metrics.

The FM is calculated using the precision and recall between the predicted pixel and the ground-truth pixel as follows:

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

The pFM, proposed in [55], and the stroke predicted through the *pseudo-Recall* ($pRecall$) representing the percentage of the skeletonized ground-truth image and the stroke of the ground-truth image, is calculated as follows:

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

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

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

The DRD is a metric used to measure the visual distortion in binary document images [56]. 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 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 [56].

4.3. Quantitative and Qualitative Comparison

To demonstrate the performance and effectiveness of the proposed model, we conducted evaluations using a total of four benchmark datasets. The DIBCO 2016, DIBCO 2017, and DIBCO 2018 datasets comprise machine-printed and handwritten document images, and DIBCO 2019-B presents a challenging benchmark dataset featuring papyrus-like materials and extreme degradation. For our evaluations, we used four established metrics—FM, pFM, PSNR, and DRD—as introduced in Section 4.2. Furthermore, to ensure a fair comparison, we reimplemented eight models, including our proposed model. These reimplementations were based on publicly available source code provided by the respective authors. The compared methods encompassed both traditional binarization algorithm-based approaches and deep learning-based methods. Among the binarization algorithm-based methods were those by Otsu [6] and Sauvola [5], whereas the deep learning-based methods included SAE [28], cGANs [15], and those by Akbari et al. [30], Souibgui et al. [17], and Suh et al. [18]. As our proposed model falls within the category of generative models, we compared it with various other generative models. Additionally, we incorporated the performance of the competition winners from each respective year [23–26] into our experimental results.

Table 1 shows the quantitative evaluation results of the models. Compared with other methods, the proposed model achieved the best performance for the mean values on all datasets. The proposed model demonstrated performance improvements of 0.08 in the FM, 1.81 in the pFM, 0.29 in the PSNR, and 0.89 in the DRD compared to the existing method [18], which demonstrated the second-highest performance. The performance improvements across all the metrics used to assess image quality and text-region preservation on DIBCO (the international competition on Document Image Binarization [26]) showed that the proposed method was effective in document binarization.

Table 1. Comparison of the proposed model with other methods on the (H-)DIBCO dataset. The last row shows the results of the mean values on the four datasets. The best performance is indicated in bold, and the second-highest performance is underlined.

Dataset	Metric	Otsu [6]	Sauvola [5]	Competition Winner	SAE [28]	cGANs [15]	Akbari [30]	Souibgui [17]	Suh [18]	Ours
2016	FM	86.64	79.57	88.72	88.11	**91.67**	90.48	84.45	<u>91.11</u>	88.95
	pFM	89.99	86.84	91.84	91.55	94.59	93.26	84.73	<u>95.22</u>	**95.45**
	PSNR	17.80	16.90	18.45	18.21	**19.64**	19.27	16.18	19.34	<u>19.38</u>
	DRD	5.52	6.76	3.86	4.51	**2.82**	3.94	7.25	<u>3.25</u>	3.74
2017	FM	80.63	73.86	**91.04**	85.72	<u>90.73</u>	85.59	80.63	89.33	89.36
	pFM	80.85	84.78	<u>92.86</u>	87.85	92.58	87.56	80.85	91.41	**94.02**
	PSNR	13.84	14.30	<u>18.28</u>	16.09	17.83	16.39	13.84	17.91	**18.33**
	DRD	9.85	8.30	**3.40**	6.53	<u>3.58</u>	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**	<u>88.43</u>
	pFM	53.58	72.13	90.24	77.95	90.60	80.09	85.74	**96.25**	<u>93.73</u>
	PSNR	9.76	13.98	19.11	14.79	18.37	17.01	16.16	**20.03**	<u>19.28</u>
	DRD	59.07	13.96	4.92	13.30	4.58	8.11	7.93	**2.60**	<u>3.95</u>
2019-B	FM	22.47	50.57	<u>67.99</u>	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	<u>68.32</u>	**75.24**
	PSNR	2.61	10.85	12.14	10.84	11.77	9.18	8.55	<u>12.91</u>	**14.37**
	DRD	213.58	33.73	26.87	32.00	24.11	70.50	53.18	<u>19.80</u>	**14.39**
Mean Values	FM	59.61	67.01	84.02	74.29	82.94	74.90	71.64	<u>84.78</u>	**84.86**
	pFM	61.53	74.56	85.71	76.47	85.07	77.13	71.85	<u>87.80</u>	**89.61**
	PSNR	11.01	14.01	17.00	14.98	16.90	15.46	12.86	<u>17.55</u>	**17.84**
	DRD	73.42	15.69	9.76	14.09	8.77	22.65	23.43	<u>7.37</u>	**6.48**

The proposed model achieved the best or second-best performance in terms of the pFM and PSNR on all four datasets. Particularly, on the DIBCO 2019-B dataset, which consists of the most complex noise that the model has not yet experienced, the proposed model achieved the best performance across all metrics. In particular, the existing methods showed great decrement, whereas the proposed model showed impressive performance gains compared to the other models, even when faced with such significant noise. Although the quantitative differences between the proposed model and the other models were marginal on the other datasets, the qualitative evaluation demonstrated that the proposed model performed better in document binarization compared to the other models. Therefore, it demonstrated that training through the iterative diffusion-denoising process was effective in removing complex noise and robust to various environments.

First, the proposed model delivered outstanding results across multiple benchmark datasets. In the H-DIBCO 2016 dataset, our model attained the highest performance in the pFM and the second-highest performance in the PSNR. The exceptional performance in the pFM underscores its superior ability to extract text strokes compared to the other methods, whereas the strong performance in the PSNR reflects the high quality of its results. Regarding the DIBCO 2017 dataset, our proposed model again took the lead, demonstrating the highest performance in both the pFM and PSNR. On the H-DIBCO 2018 dataset, our model achieved the second-highest performance across all four metrics. On the challenging DIBCO 2019-B dataset, which features images with extreme degradation on materials such as papyrus and tree bark, the proposed model outperformed all the others, achieving significantly superior results across all four metrics. This success is particularly noteworthy, as the other models struggled to perform well on this dataset, illustrating the effectiveness of our proposed model in extracting text strokes, even from extremely degraded document images it had not been explicitly trained on. When we consider the average results across the DIBCO 2016, 2017, 2018, and 2019-B datasets, it becomes evident that the proposed model consistently outperformed the other models across all of the metrics. This showcases the effectiveness of our model in terms of image generation through the diffusion-denoising process and text-stroke extraction via the gated U-Net.

The qualitative evaluation of the H-DIBCO 2016 dataset is shown in Figure 6. The figure depicts the most challenging document image with complex degradation from the H-DIBCO 2016 dataset. Figure 6g, the resulting image using the method in [15], shows the highest quantitative result, but with limitations in preserving elaborate text strokes. Figure 6i, the resulting image using the method in [18], shows better results in preserving text strokes, but with limitations in removing noise. Figure 6j, the resulting image using the proposed model, shows that it not only achieved better results in noise removal such as overwriting compared to the other methods, but also performed the best in preserving elaborate and accurate text strokes.

Figure 7 depicts an image from the DIBCO 2017 dataset, used for qualitative evaluation. This image contains noise, making it challenging to differentiate between the text and the background. Traditional algorithm-based methods often misclassify such noise as text. Similarly, deep learning-based methods face difficulties in scenarios with significant noise due to similarities in shape, size, and text stroke. However, in Figure 7j, we can see how the proposed model adeptly discerned between the background and text regions, showcasing its effectiveness in challenging scenarios.

Figure 6. Binarization results of a sample image from the H-DIBCO 2016 dataset. (**a**) Degraded image, (**b**) ground-truth image, (**c**) Otsu [6], (**d**) Sauvola [5], (**e**) SAE [28], (**f**) Souibgui et al. [17], (**g**) cGANs [15], (**h**) Akbari et al. [30], (**i**) Suh et al. [18], and (**j**) the proposed model.

Figure 8 depicts an image from the H-DIBCO 2018 dataset. This dataset comprises 10 handwritten document images with diverse degradations, including variations in background intensity, shadows, and ink smearing. The figure illustrates the challenges posed by variable background intensities and ink smearing. Other methods struggled to effectively eliminate these degradations. Although Figure 8g, an outcome from a previous method [15], mitigates some of the degradations, the method encountered difficulties in preserving intricate text strokes, such as small footnotes, within the document. In contrast, in Figure 8i, the proposed model is shown to excel in resolving various degradation types and preserving elaborate text strokes.

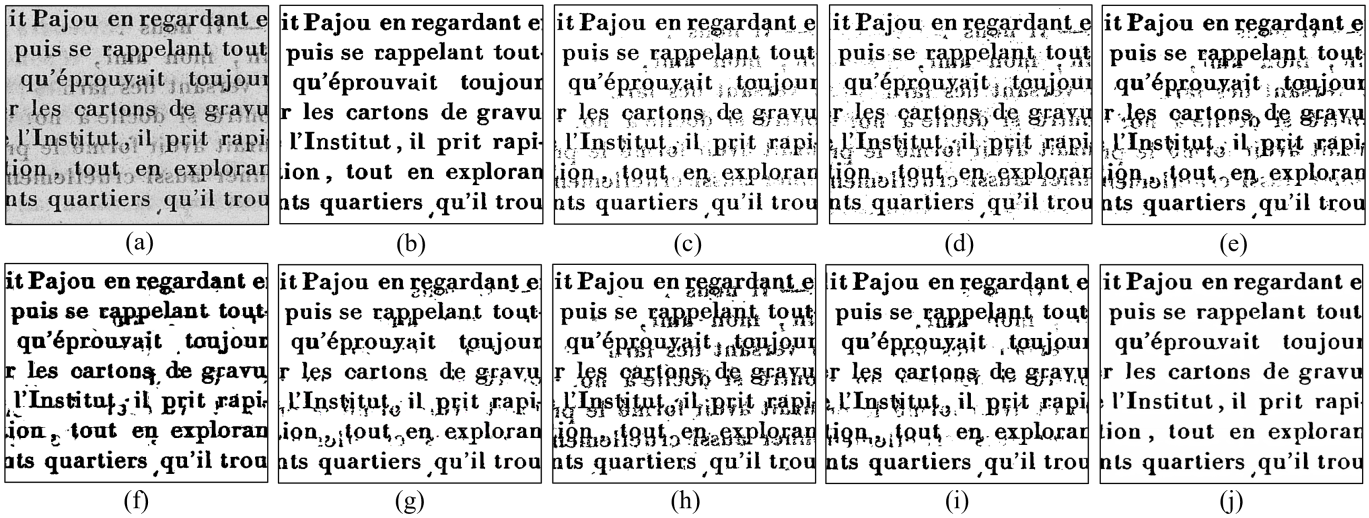


Figure 7. Binarization results of a sample image from the DIBCO 2017 dataset. (**a**) Degraded image, (**b**) ground-truth image, (**c**) Otsu [6], (**d**) Sauvola [5], (**e**) SAE [28], (**f**) Souibgui et al. [17], (**g**) cGANs [15], (**h**) Akbari et al. [30], (**i**) Suh et al. [18], and (**j**) the proposed model.

Figure 9 showcases an image from the DIBCO 2019-B dataset, which comprises ancient document images featuring a range of papyrus qualities, ink characteristics, and handwriting styles. This dataset presents a formidable challenge due to its non-homogeneous properties such as resolution, lighting, and noise variations. Figure 9c,d, correspond to the

results using the methods in [15,18], which encountered difficulties in effectively distinguishing the text from the background due to the intensity. Consequently, preserving text strokes was exceptionally challenging. In contrast, Figure 9e highlights the effectiveness of the proposed model in classifying the text and background regions while successfully preserving the text areas. Notably, the degradation types found in the DIBCO 2019-B dataset extended beyond those encountered in the training data, underscoring the proposed model's effectiveness in addressing even significant degradation.

Furthermore, Figure 10 illustrates the effectiveness of the proposed model in extracting detailed regions. Distinguishing between the background and text regions and accurately capturing intricate text strokes in small portions, as demonstrated in the sample image, can be challenging. In such cases, even methods with relatively high quantitative evaluations [15,18,30] encountered difficulties in precisely extracting and preserving text regions, as evident in Figure 10c–e. However, the proposed model, as shown in Figure 10f, successfully extracted and preserved precise text strokes despite these challenges.

Figure 8. Binarization results of a sample image from the H-DIBCO 2018 dataset. (**a**) Degraded image, (**b**) ground-truth image, (**c**) Otsu [6], (**d**) Sauvola [5], (**e**) SAE [28], (**f**) cGANs [15], (**g**) Akbari et al. [30], (**h**) Suh et al. [18], and (**i**) the proposed model.

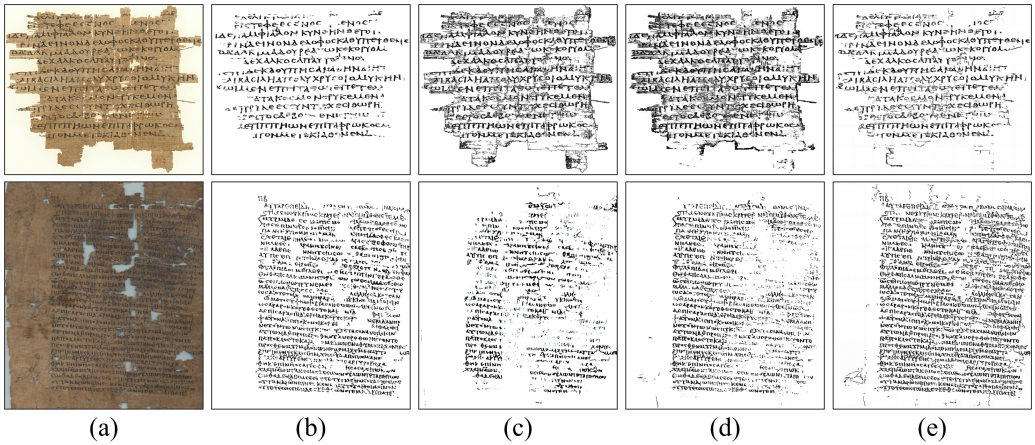


Figure 9. Binarization results of challenging images from the DIBCO 2019-B dataset. (**a**) Degraded image, (**b**) ground-truth image, (**c**) cGANs [15], (**d**) Suh et al. [18], and (**e**) the proposed model.

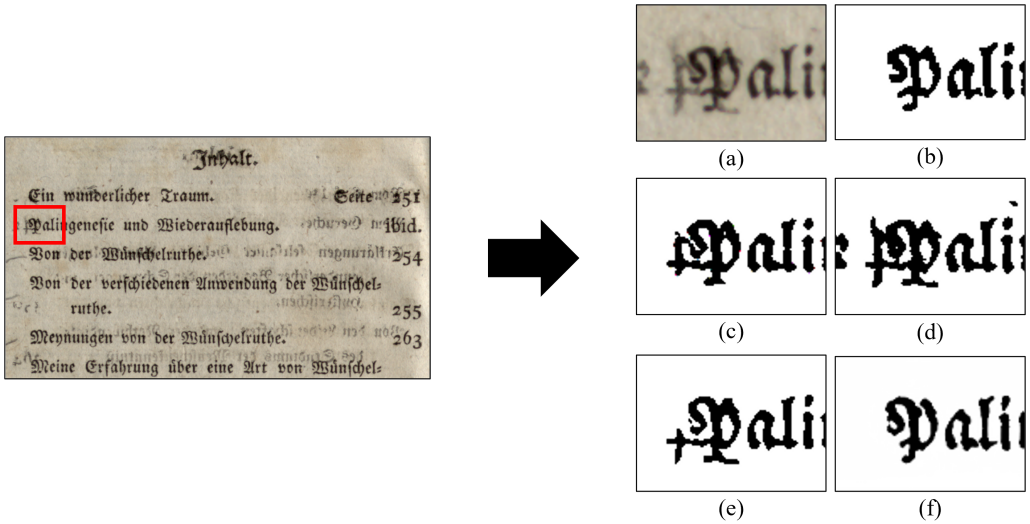


Figure 10. Binarization results of small text with degradation. (**a**) Degraded image, (**b**) ground-truth image, (**c**) cGANs [15], (**d**) Akbari et al. [30], (**e**) Suh et al. [18], (**f**) the proposed model. The proposed model could effectively preserve the text region, even for small text.

4.4. Ablation Study

We conducted ablation experiments to assess each component of the proposed model. For evaluation purposes, we employed the DIBCO 2016 benchmark dataset and demonstrated the effectiveness of the proposed model using the four metrics outlined in Section 4.2. The baseline reference was [20], and the diffusion-denoising process was executed in the latent space using an autoencoder. We compared the baseline with the proposed gated U-Net and pixel-level loss individually, as well as models incorporating all of these components. In total, we compared four models, conducting experiments under identical implementation settings except for the inclusion or exclusion of specific components within each model.

Table 2 shows the results of the baseline, proposed model without L_{pix}, proposed model without gated U-Net, and proposed model. When the gated U-Net was added to the baseline, the pFM (95.03) improved by 0.59 compared to the baseline (94.44). Text-stroke preservation performance was improved because filter training for the gating values was additionally included. However, when the pixel-level loss was not included, the feedback for the gated convolution filter was not effective; thus, there were no significant performance improvements. When the gated convolution was omitted, both the text and background

regions were updated solely through the pixel-level loss, leading to an improvement in performance across all metrics. Specifically, in the case of the proposed model with the pixel-level loss incorporated to train the gated convolution filter in the gated U-Net, we observed improvements in the FM, pFM, PSNR, and DRD of 0.68, 1.00, 0.28, and 0.31, respectively, compared to the baseline. This highlights the effectiveness of each individual component in the proposed model, and it also underscores the synergy of all components working together to achieve performance gains.

Furthermore, to assess the impact of each component, we present the results of model training after six epochs in Table 3. We used the FM to evaluate the accuracy of the pixel-wise predictions for the text and background, along with the pFM to assess predictions at a regional level. When excluding the pixel-level loss from the proposed model, each metric remained comparable to or slightly improved upon the baseline performance. For the proposed model without the gated U-Net, which used the baseline's backbone network with an LDM and pixel-level loss to update the denoising network, we observed significant improvements in the FM of 8.00 and the pFM of 7.94 compared to the baseline. Conversely, the proposed model using the gated U-Net for document binarization and the pixel-level loss to facilitate the training of the gating values achieved the highest performance across all metrics, where the FM increased by 8.22 and the pFM increased by 8.51 compared to the baseline. These outcomes underscore the effectiveness of the individual components in the proposed model, particularly in terms of preserving and extracting text strokes.

Table 2. The results of the ablation study on the H-DIBCO 2016 dataset. Best performance is indicated in bold.

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**

Table 3. Results of the ablation study on the H-DIBCO 2016 dataset after training for 6 epochs. The best performance is indicated in bold.

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**

Figure 11 shows the resulting images using the baseline, the proposed model without L_{pix}, the proposed model without the gated U-Net, and the proposed model. As shown in Figure 11f, the proposed model with both L_{pix} and the gated U-Net was successful in distinguishing between the text and background of degraded document images compared to the baseline shown in Figure 11c.

Furthermore, in Figure 12, the excellent text-stroke preservation performance can be qualitatively confirmed. In Figure 12f, which demonstrates the results of the proposed model, noise removal and text-stroke preservation are more accurate and precise compared to the baseline shown in Figure 12c. This proves that the diffusion-denoising process through the gated U-Net and L_{pix} was effective in preserving more elaborate text strokes and generating high-quality results.

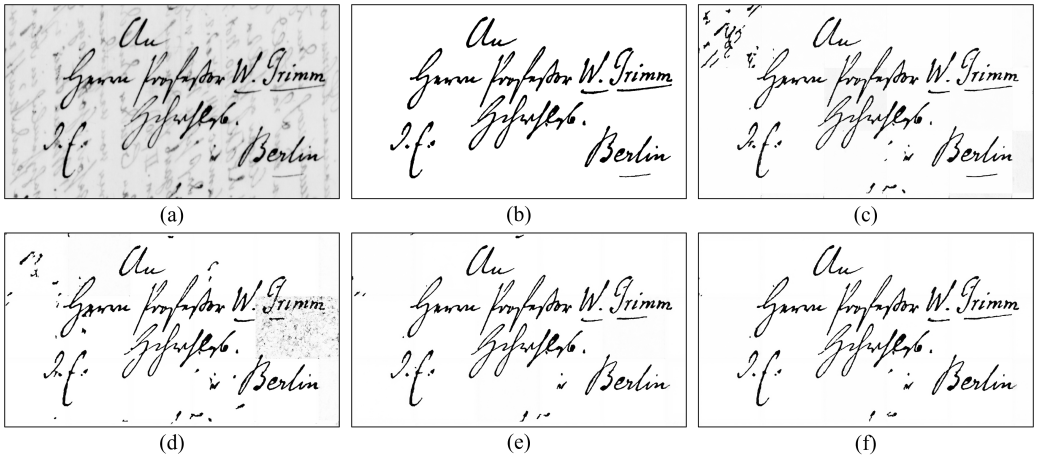


Figure 11. Binarization results of a sample image from DIBCO 2016. (**a**) Degraded image, (**b**) ground-truth image, (**c**) baseline, (**d**) our model w/o L_{pix}, (**e**) our model w/o gated U-Net, and (**f**) our model.

Figure 12. Binarization results of a sample image to confirm the effect of preserving the text region from DIBCO 2016. (**a**) Degraded image, (**b**) ground-truth image, (**c**) baseline, (**d**) our model w/o L_{pix}, (**e**) our model w/o gated U-Net, and (**f**) our model.

5. Conclusions

In this paper, we introduce a novel document binarization model rooted in LDMs that offers intricate text-stroke extraction and robustness against unanticipated degradation. For the first time, we use the diffusion-denoising process within the latent space to produce high-quality, clean, binarized document images. Furthermore, our denoising network takes the form of a gated U-Net, incorporating a fully gated convolution to safeguard text strokes in significantly degraded document images. Within this framework, we introduce a filter to update the gating values, ensuring the preservation of text regions. Using a pixel-level loss to dynamically update the filter for the gating values, we effectively safeguard text strokes, even in highly degraded images. Our extensive experiments on challenging benchmark datasets validate the effectiveness of our proposed model. Quantitative experimental results affirm the remarkable performance enhancements achieved by the proposed model in text-stroke preservation and extraction, particularly when compared to existing methods on the H-DIBCO datasets. Through diverse qualitative evaluations, we demonstrate how our model effectively overcomes the limitations of existing approaches and adeptly accomplishes text and background classification tasks. However, as illustrated in Figure 9,

we acknowledge that our model cannot achieve flawless binarization performance in scenarios featuring untrained noise. In future work, we anticipate that research focusing on data generalization will address these challenges, ultimately rendering our model adaptable to various tasks that rely on binary masks.

Author Contributions: Conceptualization, S.H. and J.R.; methodology S.H.; software, S.H.; validation, S.H. and S.J.; investigation, S.H.; writing—original draft preparation, S.H. and S.J.; writing—review and editing, S.H. and 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 the Korea Institute for the 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.

Data Availability Statement: Not applicable.

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

References

1. Sulaiman, A.; Omar, K.; Nasrudin, M.F. Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *J. Imaging* **2019**, *5*, 48. [CrossRef] [PubMed]
2. Farahmand, A.; Sarrafzadeh, H.; Shanbehzadeh, J. Document image noises and removal methods. In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, 21–23 March 2013.
3. Mustafa, W.A.; Kader, M.M.M.A. Binarization of document images: A comprehensive review. *J. Phys. Conf. Ser.* **2018**, *1019*, 012023. [CrossRef]
4. Chauhan, S.; Sharma, E.; Doegar, A. Binarization techniques for degraded document images—A review. In Proceedings of the 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 7–9 September 2016; pp. 163–166.
5. Sauvola, J.; Seppanen, T.; Haapakoski, S.; Pietikainen, M. Adaptive document binarization. In Proceedings of the Fourth International Conference on Document Analysis and Recognition, Ulm, Germany, 18–20 August 1997; Volume 1, pp. 147–152.
6. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]
7. Niblack, W. *An Introduction to Digital Image Processing*; Strandberg Publishing Company: København, Denmark, 1985.
8. He, S.; Schomaker, L. DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern Recognit.* **2019**, *91*, 379–390. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
11. Westphal, F.; Lavesson, N.; Grahn, H. Document image binarization using recurrent neural networks. In Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria, 24–27 April 2018; pp. 263–268.
12. Tensmeyer, C.; Martinez, T. Document image binarization with fully convolutional neural networks. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 99–104.
13. Peng, X.; Wang, C.; Cao, H. Document Binarization via Multi-resolutional Attention Model with DRD Loss. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 45–50. . [CrossRef]
14. Huang, X.; Li, L.; Liu, R.; Xu, C.; Ye, M. Binarization of degraded document images with global-local U-Nets. *Optik* **2020**, *203*, 164025. .: 10.1016/j.ijleo.2019.164025. [CrossRef]
15. Zhao, J.; Shi, C.; Jia, F.; Wang, Y.; Xiao, B. Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognit.* **2019**, *96*, 106968. [CrossRef]
16. De, R.; Chakraborty, A.; Sarkar, R. Document image binarization using dual discriminator generative adversarial networks. *IEEE Signal Process. Lett.* **2020**, *27*, 1090–1094. [CrossRef]
17. Souibgui, M.A.; Kessentini, Y. De-gan: A conditional generative adversarial network for document enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1180–1191. [CrossRef]

18. Suh, S.; Kim, J.; Lukowicz, P.; Lee, Y.O. Two-stage generative adversarial networks for binarization of color document images. *Pattern Recognit.* **2022**, *130*, 108810. [CrossRef]
19. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
20. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695.
21. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
22. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4471–4480.
23. Pratikakis, I.; Zagoris, K.; Barlas, G.; Gatos, B. ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). In Proceedings of the 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016; pp. 619–623.
24. Pratikakis, I.; Zagoris, K.; Barlas, G.; Gatos, B. ICDAR2017 competition on document image binarization (DIBCO 2017). In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 1395–1403.
25. Pratikakis, I.; Zagori, K.; Kaddas, P.; Gatos, B. ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). In Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; pp. 489–493. [CrossRef]
26. Pratikakis, I.; Zagoris, K.; Karagiannis, X.; Tsochatzidis, L.; Mondal, T.; Marthot-Santaniello, I. ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1547–1556. [CrossRef]
27. Peng, X.; Cao, H.; Natarajan, P. Using convolutional encoder-decoder for document image binarization. In Proceedings of the 2017 14th IAPR international conference on document analysis and recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 708–713.
28. Calvo-Zaragoza, J.; Gallego, A.J. A selectional auto-encoder approach for document image binarization. *Pattern Recognit.* **2019**, *86*, 37–47. [CrossRef]
29. 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 Recognit.* **2021**, *109*, 107577. [CrossRef]
30. 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. [CrossRef]
31. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
32. 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* **2022**, arXiv:2211.16098.
33. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
34. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. In Proceedings of the International Conference on Medical Imaging with Deep Learning, PMLR, Zurich, Switzerland, 6–8 July 2022; pp. 1336–1348.
35. Kim, B.; Oh, Y.; Ye, J.C. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv* **2022**, arXiv:2209.14566.
36. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. *arXiv* **2022**, arXiv:2211.09788.
37. Duan, Y.; Guo, X.; Zhu, Z. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv* **2023**, arXiv:2303.05021.
38. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; Yang, K. Gated fully fusion for semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11418–11425.
39. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
40. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 933–941.
41. Lin, X.; Ma, L.; Liu, W.; Chang, S.F. Context-gated convolution. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVIII 16; Springer: Cham, Switzerland, 2020; pp. 701–718.
42. Zhang, Y.; Fang, J.; Chen, Y.; Jia, L. Edge-aware U-net with gated convolution for retinal vessel segmentation. *Biomed. Signal Process. Control* **2022**, *73*, 103472. [CrossRef]
43. Kwon, M.; Jeong, J.; Uh, Y. Diffusion models already have a semantic latent space. *arXiv* **2022**, arXiv:2210.10960.
44. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12873–12883.

45. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. H-DIBCO 2010-handwritten document image binarization competition. In Proceedings of the 2010 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, India, 16–18 November 2010; pp. 727–732.
46. Pratikakis, I.; Gatos, B.; Ntirogiannis. ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; pp. 817–822.
47. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). In Proceedings of the 2014 14th International Conference on Frontiers in Handwriting Recognition, Hersonissos, Greece, 1–4 September 2014; pp. 809–813.
48. Gatos, B.; Ntirogiannis, K.; Pratikakis, I. ICDAR 2009 document image binarization contest (DIBCO 2009). In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 1375–1382.
49. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18-21 September 2011; pp. 1506–1510. [CrossRef]
50. Pratikakis, I.; Gatos, B.; Ntirogiannis. ICDAR 2013 document image binarization contest (DIBCO 2013). In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1471–1476.
51. Deng, F.; Wu, Z.; Lu, Z.; Brown, M.S. Binarizationshop: A user-assisted software suite for converting old documents to black-and-white. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, 21 June 2010 ; pp. 255–258.
52. Nafchi, H.Z.; Ayatollahi, S.M.; Moghaddam, R.F.; Cheriet, M. An efficient ground truthing tool for binarization of historical manuscripts. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 807–811.
53. Hedjam, R.; Cheriet, M. Historical document image restoration using multispectral imaging system. *Pattern Recognit.* **2013**, *46*, 2297–2312. [CrossRef]
54. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
55. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. Performance evaluation methodology for historical document image binarization. *IEEE Trans. Image Process.* **2012**, *22*, 595–609. [CrossRef]
56. Lu, H.; Kot, A.C.; Shi, Y.Q. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Process. Lett.* **2004**, *11*, 228–231. [CrossRef]