






ColdBin: Cold Diffusion for Document Image Binarization

Saifullah Saifullah^{1,2}(✉) , Stefan Agne^{1,3} , Andreas Dengel^{1,2} ,
and Sheraz Ahmed^{1,3} 

¹ Smart Data and Knowledge Services (SDS), German Research Center for Artificial Intelligence GmbH (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany
{Saifullah.Saifullah, Stefan.Agne, Andreas.Dengel, Sheraz.Ahmed}@dfki.de

² Department of Computer Science, RPTU Kaiserslautern-Landau,
Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany

³ DeepReader GmbH, 67663 Kaiserslautern, Germany

Abstract. Document images, when captured in real-world settings, either modern or historical, frequently exhibit various forms of degradation such as ink stains, smudges, faded text, and uneven illumination, which can significantly impede the performance of deep learning-based approaches for document processing. In this paper, we propose a novel end-to-end framework for binarization of degraded document images based on cold diffusion. In particular, our approach involves training a diffusion model with the objective of generating a binarized document image directly from a degraded input image. To the best of the authors' knowledge, this is the first work that investigates diffusion models for the task of document binarization. In order to assess the effectiveness of our approach, we evaluate it on 9 different benchmark datasets for document binarization. The results of our experiments show that our proposed approach outperforms several existing state-of-the-art approaches, including complex approaches utilizing generative adversarial networks (GANs) and variational auto-encoders (VAEs), on 7 of the datasets, while achieving comparable performance on the remaining 2 datasets. Our findings suggest that diffusion models can be an effective tool for document binarization tasks and pave the way for future research on diffusion models for document image enhancement tasks. The implementation code of our framework is publicly available at: <https://github.com/saifullah3396/coldbin>.

Keywords: Document Binarization · Document Image Enhancement · Diffusion Models · Cold Diffusion · Document Image Analysis

1 Introduction

In the era of automation, accurate and efficient automated processing of documents is of the utmost importance for streamlining modern business workflows

This work was supported by the BMBF projects SensAI (BMBF Grant 01IW20007).

[1–3]. At the same time, it has vast applications in the preservation of historical scriptures [4–6] that contain valuable information about ancient cultural heritages and scientific contributions. Deep learning (DL) has recently emerged as a powerful tool for handling a wide variety of document processing tasks, showing remarkable results in areas such as document classification [1, 7], optical character recognition (OCR) [8], and named entity recognition (NER) [2, 9]. However, it remains challenging to apply DL-based models to real-world documents due to a variety of distortions and degradations that frequently occur in these documents. Document image enhancement (DIE) is a core research area in document analysis that focuses on recovering clean and improved images of documents from their degraded counterparts. Depending on the severity of the degradation, a document may display wrinkles, stains, smears, or bleed-through effects [10–12]. Additionally, distortions may result from scanning documents with a smartphone, which may introduce shadows [13], blurriness [14], or uneven illumination. Such degradations, which are particularly prevalent in historical documents, can significantly deteriorate the performance of deep learning models on downstream document processing tasks [15]. Therefore, it is essential that prior to applying these models, there be a pre-processing step that performs denoising and recovers a clean version of the degraded document image.

Over the past few decades, DIE has been the subject of several research efforts, including both classical [16, 17] and deep learning-based studies [6, 13, 18, 19]. Lately, generative models such as deep variational autoencoders (VAEs) [20] and generative adversarial networks (GANs) [21] have gained popularity in this domain, owing to their remarkable success in natural image generation [21, 22] and restoration tasks [23–25]. Generative models have attracted considerable attention due to their ability to accurately capture the underlying distribution of the training data, which allows them not only to generate highly realistic and diverse samples [22], but also to generate missing data when necessary [26]. As a result, a number of GAN and VAE based approaches have been recently proposed for DIE tasks, such as binarization [6, 18, 27], deblurring [6, 19], and watermark removal [6].

Diffusion models [28] are a new class of generative models inspired by the process of diffusion in non-equilibrium thermodynamics. In the context of image generation, the underlying mechanism of diffusion models involves a fixed forward process of gradually adding Gaussian noise to the image, and a learnable reverse process to denoise and recover the clean image, utilizing a Markov chain structure. Diffusion models have been shown to have several advantages over GANs and VAEs such as their high training stability [28–30], diverse and realistic image synthesis [31, 32], and better generalization to out-of-distribution data [33]. Additionally, conditional diffusion models have been employed to perform image synthesis with an additional input, such as class labels, text, or source image and have been successfully adapted for various natural image restoration tasks, including super-resolution [34], deblurring [35], and JPEG restoration [36]. Despite their growing popularity, however, there is no existing literature that has explored their potential in the context of document image enhancement.

In this study, we investigate the potential of diffusion models for the task of document image binarization, and introduce a novel approach for restoring clean binarized images from degraded document images using cold diffusion. We conduct a comprehensive evaluation of our proposed approach on multiple publicly available benchmark datasets for document binarization, demonstrating the effectiveness of our methodology in producing high-quality binarized images from degraded document images. The main contributions of this paper are two-fold:

- To the best of the authors’ knowledge, this is the first work that presents a flexible end-to-end document image binarization framework based on diffusion models.
- We evaluate the performance of our approach on 9 different benchmark datasets for document binarization which include DIBCO ’9 [37], H-DIBCO ’10 [11], DIBCO ’11 [38], H-DIBCO ’12 [39], DIBCO ’13 [12], H-DIBCO ’14 [40], H-DIBCO ’16 [41], DIBCO ’17 [42], and H-DIBCO ’18 [43].
- Through a comprehensive quantitative and qualitative evaluation, we demonstrate that our approach outperforms several classical approaches as well as the existing state-of-the-art on 7 of the datasets, while achieving competitive performance on the remaining 2 datasets.

2 Related Work

2.1 Document Image Enhancement

Document image enhancement (DIE) has been extensively studied in the literature over the past few decades [5, 16, 44–46]. Classical approaches to DIE were primarily based on global thresholding [16], local thresholding [44, 47] or their hybrids [48]. These approaches were based on determining threshold values to segment the image pixels of a document into foreground or background. In a different direction, energy-based segmentation approaches such as Markov random fields (MRFs) [49] and conditional random fields (CRFs) [50] and classical machine learning-based approaches such as support vector machines (SVMs) [17, 51] have also been widely explored in the past.

In recent years, there has been a burgeoning interest in the application of deep learning-based techniques for the enhancement of document images [4, 52–54]. The earliest work in this area was majorly focused on utilizing convolutional neural networks (CNNs) [4, 5, 52, 55]. One notable example of this is the work of Pastor-Pellicer *et al.* [56], who proposed a CNN-based classifier in conjunction with a sliding window approach for segmenting images into foreground and background regions. Building upon this, Tensmeyer *et al.* [52] presented a more advanced methodology that entailed feeding raw grayscale images, along with relative darkness features, into a multi-scale CNN, and training the network using a pseudo F-measure loss. Another approach was proposed by Calvo-Zaragoza *et al.* [55], in which they utilized a CNN-based auto-encoder (AE) to train the model to map degraded images to clean ones in an end-to-end fashion. A similar

approach was presented by Kang *et al.* [5], who employed a pre-trained U-Net based auto-encoder model for binarization, with minimal training data requirements. Since then, a number of AE-based approaches have been proposed for DIE tasks [53, 54]. In a slightly different direction, Castellanos *et al.* [57] has also investigated domain adaptation in conjunction with deep neural networks for the task of document binarization.

Generative Adversarial Networks (GANs) have also been extensively explored in this field to generate clean images by conditioning on degraded versions [6, 19, 27, 46]. These methods typically consist of a generative model that generates a clean binarized version of the image, along with a discriminator that assesses the results of the binarization. Zhao *et al.* [27] proposed a cascaded GAN-based approach for the task of document image binarization and demonstrated excellent performance on a variety of benchmark datasets. Jemni *et al.* [58] recently presented a multi-task GAN-based approach which incorporates a text recognition network in combination with the discriminator to further improve text readability along with binarization. Similarly, Yu *et al.* [46] proposed a multi-stage GAN-based approach to document binarization that first applies discrete wavelet transform to the images to perform enhancement, and then trains a separate GAN for each channel of the document image. Besides GANs and CNN-based auto-encoders, the recent success of transformers in natural language processing (NLP) [9] and vision [59] has also sparked interest in transformers for the enhancement of document images. In a recent study, Souibgui *et al.* [45] proposed a transformer-based auto-encoder model that demonstrated state-of-the-art performance on several document binarization datasets.

3 ColDBin: The Proposed Approach

This section presents the details of our proposed approach and explains its relationship to standard diffusion [28]. The overall workflow of our approach is illustrated in Fig. 1. Primarily inspired by cold diffusion [60], our approach involves training a deep diffusion network for document binarization in two steps: a forward diffusion step and a reverse restoration step. As shown, in the forward diffusion step, a clean ground-truth document image is degraded to a specified severity level based on a given type of input degradation. In the reverse restoration step, a neural network is tasked with undoing the forward diffusion process in order to generate a clean ground-truth image from an intermediary degraded image. These forward and reverse steps are repeated in a cycle, and the neural network is trained for the binarization task by applying image reconstruction loss to its output. In the following sections, we provide a more detailed explanation of the forward and reverse steps of our approach.

3.1 Forward Diffusion

In the context of document binarization, let $P = \{(x, x_0) \sim (\mathcal{X}, \mathcal{X}_0)\}_{n=1}^N$ define a training set consisting of pairs of degraded document images x and their

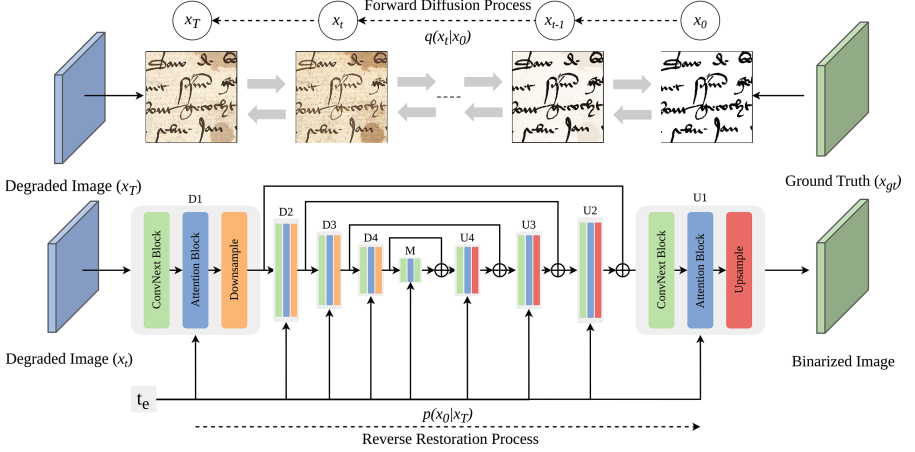


Fig. 1. Demonstration of the forward diffusion and reverse restoration processes of our approach. The forward diffusion process incrementally degrades a clean ground-truth image into its degraded counterpart. Whereas, the reverse restoration process, defined by a neural network, generates a clean binary image from a degraded input image

corresponding binarized ground-truth images x_0 . Let $\mathbb{D}(x_0, t)$ be a diffusion operator that adds degradation to a clean ground-truth image x_0 proportional to the severity $t \in \{0, 1, \dots, T\}$, T being the maximum severity permitted, then the degraded image at any given severity t can be derived as follows:

$$x_t = \mathbb{D}(x_0, t) \quad (1)$$

Consequently, the following constraint must be satisfied:

$$\mathbb{D}(x_0, 0) = x_0 \quad (2)$$

Generally in standard diffusion [28], this forward diffusion operator $\mathbb{D}(x_0, t)$ is defined as a fixed Markov process that gradually adds Gaussian noise ϵ to the image using a variance schedule specified by $\beta_1 \dots \beta_T$. In particular, it is defined as the posterior $q(x_1, \dots, x_T|x_0)$ that converts the data distribution $q(x_0)$ to the latent distribution $q(x_T)$ as follows:

$$q(x_1, \dots, x_T|x_0) := \prod_{t=1}^T q(x_t|x_{t-1})$$

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\beta_t}x_{t-1}, (1 - \beta_t)\mathbf{I})$$

where β_t is a hyper-parameter that defines the severity of degradation at each severity level t . An important property of the above forward process is that it allows sampling x_t at any arbitrary severity t in closed form: using the notation $\alpha_t := 1 - \beta_t$ and $\hat{\alpha}_t := \prod_{s=1}^t \alpha_s$, we have

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\hat{\alpha}_t}x_0, (1 - \hat{\alpha}_t)\mathbf{I}) \quad (3)$$

Which results in the following the diffusion operator $\mathbb{D}(x_0, t)$:

$$x_t = \mathbb{D}(x_0, t) = \sqrt{\hat{\alpha}_t}x_0 + \sqrt{1 - \hat{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (4)$$

Our approach maintains the same forward process as standard diffusion, except that Gaussian noise ϵ is not used to define the diffusion operator $\mathbb{D}(x_0, t)$ (hot diffusion). Rather, we define it as a cold diffusion operation that interpolates between the binarized ground-truth image x_0 and its degraded counterpart image x based on the noise schedule $\beta_1 \dots \beta_T$. More formally, given a fully degraded input image x and its respective binarized ground-truth image x_0 , an intermediate degraded image x_t at severity t is then defined as follows:

$$x_t = \mathbb{D}(x_0, x, t) = \sqrt{\hat{\alpha}_t}x_0 + \sqrt{1 - \hat{\alpha}_t}x, \quad x_0 \sim \mathcal{X}_0, x \sim \mathcal{X} \quad (5)$$

Note that this procedure is essentially the same as adding Gaussian noise ϵ in standard diffusion, except that here we are adding a progressively higher weighted degraded image to the clean ground-truth image to generate an intermediary noisy image. In addition, our diffusion operator for binarization is slightly modified $\mathbb{D}(x_0, x, t)$ and requires both the ground-truth image x_0 and the target degraded image x for forward the process.

3.2 Reverse Restoration

Let $\mathbb{R}(x_t, t)$ define the reverse restoration operator that restores any degraded image x_t at severity t to its clean binarized form x_0 :

$$\mathbb{R}(x_t, t) \approx x_0 \quad (6)$$

In standard diffusion [28], generally this restoration operator $\mathbb{R}(x_t, t)$ is defined as a reverse Markov process $p(x_0, \dots, x_{T-1}|x_T)$ that transforms the data from the latent variable distribution $p_\theta(x_T)$ to the data distribution $p_\theta(x_0)$ parameterized by θ ; the process generally starting from $p(x_T) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$:

$$p(x_0, \dots, x_{T-1}|x_T) := \prod_{t=1}^T p_\theta(x_{t-1}|x_t)$$

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 \mathbf{I})$$

Our approach uses the same reverse restoration process as the standard diffusion [28], with the exception that it begins with a degraded input image $x_T \sim \mathcal{X}$ instead of Gaussian noise $x_T \sim \mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$. In practice, $\mathbb{R}(x_t, t)$ is generally implemented as a neural network $\mathbb{R}_\theta(x_t, t)$ parameterized by θ which is trained to perform the reverse restoration task. In our approach, the restoration network $\mathbb{R}_\theta(x_t, t)$ is trained by minimizing the following loss:

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{X}, x_0 \sim \mathcal{X}_0} \|\mathbb{R}_\theta(D(x_0, x, t), t) - x_0\| \quad (7)$$

Algorithm 1. Training

-
- 1: **Input:** Ground truth image x_0 and its corresponding degraded image x pairs $P = \{(x_0, x)\}_{k=1}^K$, and total diffusion steps T
 - 2: **Initialize:** Randomly initialize the restoration network $\mathbb{R}_\theta(x_t, t)$
 - 3: **repeat**
 - 4: Sample $(x_0, x) \sim P$, and $t \sim \text{Uniform}(\{1, \dots, T\})$
 - 5: Take the gradient step on
 - 6: $\nabla_{\theta} \|\mathbb{R}_\theta(x_t, t) - x_0\|, x_t = \sqrt{\hat{\alpha}_t}x_0 + \sqrt{1 - \hat{\alpha}_t}x$
 - 7: **until** converged
-

where $\|\cdot\|$ defines a norm, which we took as standard ℓ_2 norm in this work. The overall training process of the restoration network is given in Algorithm 1. As shown, the restoration network $\mathbb{R}_\theta(x_t, t)$ is initialized with a maximum severity level of T . In each training iteration, a mini-batch of degraded images x and their corresponding binarized ground-truth images x_0 is randomly sampled from the training set P , and the degradation severity is randomly sampled from the integer set $\{1, \dots, T\}$. The severity value t is then used in combination with the ground-truth x_0 and degraded image x pairs to compute the intermediate interpolated images x_t using Eq. 5 (line 6). The restoration network $\mathbb{R}_\theta(x_t, t)$ is then used to recover a binarized image from the interpolated image x_t . Finally, the network is optimized in each step by taking the gradient step on Eq. 7 (line 6).

3.3 Restoration Network

The complete architecture of the restoration network $\mathbb{R}_\theta(x_t, t)$ used in our approach is illustrated in Fig. 1. As shown, we used a U-Net [61] inspired architecture as the restoration network which takes as input the degraded image x_t and the diffusion severity $t \in 1, 2, \dots, T$ and generates a binarized image as the output. The input severity level t is transformed into a severity embedding t_e based on sinusoidal positional encoding as proposed in [62]. The embedded severity and the image are then passed through multiple downsampling blocks, a middle processing block and then multiple upsampling blocks to generate the output image. Each downsampling and upsampling block is characterized by two ConvNeXt [63] blocks, a residual block with a linear attention layer, and a downsampling layer. The middle block consists of a ConvNeXt block followed by an attention module and another ConvNeXt block and is inserted between the downsampling and upsampling phases.

3.4 Inference Strategies

We investigated two different inference strategies for restoring images from their degraded counterparts: direct restoration and cold diffusion sampling. Direct restoration simply applies the restoration operator $\mathbb{R}_\theta(x_t, t)$ to a degraded input image x with degradation severity t set to T . On the other hand, cold diffusion

sampling as proposed in [60] iteratively performs the reverse restoration process over T steps as described in Algorithm 2. Although a number of sampling strategies have been proposed previously for diffusion models [28, 64], Bansal *et al.* [60] demonstrated in their work that this sampling strategy performs better than standard sampling [28] for cold diffusion processes, and therefore it has been investigated in this study.

4 Experiments and Results

In this section, we first describe the experimental setup, including datasets, evaluation metrics, and the training process. Subsequently, we present a comprehensive quantitative and qualitative analysis of our results.

4.1 Experimental Setup

Datasets. 9 different DIBCO document image binarization datasets were used to assess the performance of our proposed approach. These datasets include DIBCO '9 [37], DIBCO '11 [38], DIBCO '13 [12], and DIBCO '17 [42], as well as H-DIBCO '10 [11], H-DIBCO '12 [39], H-DIBCO '14 [40], H-DIBCO '16 [41], and H-DIBCO '18 [43]. A variety of degraded printed and handwritten documents are included in these datasets, which exhibit various degradations such as ink bleed through, smudges, faded text strokes, stain marks, background texture, and artifacts.

Evaluation Metrics. Several evaluation methods have been commonly used in the literature for evaluating the binarization of document images, including FM (F-Measure), pFM (pseudo-F-Measure), PSNR (Peak Signal-to-Noise Ratio), and DRD (Distance Reciprocal Distortion), which have been adopted in this study. A higher value indicates better binarization performance for the first three metrics, while the opposite is true for DRD. Due to space constraints, detailed definitions of these metrics are omitted here and can be found in [11, 12].

Data Preprocessing. To train the restoration model on a specific DIBCO dataset, all the images from other DIBCO and H-DIBCO datasets as well as the Palm Leaf dataset [65] were used. The training set was prepared by splitting each degraded image and its corresponding ground truth image into overlapped patches

Algorithm 2. Cold Diffusion Sampling Strategy [60]

```

1: Input: A degraded sample  $x$ 
2: for  $s=t, t-1, \dots, 1$  do
3:    $\hat{x}_0 = R(x_s, s)$ 
4:    $x_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1)$ 
5: end for

```

Table 1. The size of the training and test sets for all DIBCO datasets is provided.

	DIBCO '9 [37]	H-DIBCO '10 [11]	DIBCO '11 [38]	H-DIBCO '12 [39]	DIBCO '13 [12]	H-DIBCO '14 [40]	H-DIBCO '16 [41]	DIBCO '17 [42]	H-DIBCO '18 [43]
Train	17716	17669	17448	16885	16231	17492	17300	15983	17202
Test (256)	135	150	217	356	542	182	256	610	266
Test (512)	40	46	66	104	166	55	74	184	74

of size $384 \times 384 \times 3$. Table 1 shows the total number of training set samples that were generated for each DIBCO dataset as a result of using the above strategy. During training, a random crop of size $256 \times 256 \times 3$ was extracted from each image and then fed to the model. Additionally, a number of data augmentations were used such as horizontal flipping, vertical flipping, color jitter, grayscale conversion, and Gaussian blur, all of which were randomly applied to the images. A specific augmentation we used in our approach was to randomly colorize the degraded image using the inverted ground truth image as a mask. This augmentation was necessary to prevent the models from overfitting to black-color text since most of the images in the DIBCO datasets consisted of black-color text on various backgrounds. Furthermore, we used ImageNet normalization with per-channel means of $\mu_{RGB} = \{0.485, 0.456, 0.406\}$ and standard deviations of $\sigma_{RGB} = \{0.229, 0.224, 0.225\}$ to normalize each image before feeding it to the model.

Training Hyperparameters. We initialized our restoration networks with maximum diffusion severity T set to 200 and severity embedding set to 64. For the forward diffusion process, we used a cosine beta noise schedule β_1, \dots, β_T as described in [66]. We trained our networks for 400k iterations with a batch size of 128, Adam optimizer, and a fixed learning rate of $2e - 5$ on 4–8 NVIDIA A100 GPUs.

Evaluation Hyperparameters. To evaluate our approach, we divided each image into patches of fixed input size, restored them using the inference strategies outlined in Sect. 3.4, and then reassembled them to produce the final binarized image. Depending on the size of the input patch, binarization performance can be greatly affected, since smaller patches provide less context for the model, whereas larger patches provide more context. In this work, we examined two different patch sizes at test time, which were 256×256 and 512×512 . It should be noted that we trained the models solely on 256×256 input images, and used images of size 512×512 only during evaluation.

Table 2. Comparison of different evaluation strategies on DIBCO '9 [37], H-DIBCO '12 [39], and DIBCO '17 [42] datasets. The top strategy for each metric is bolded.

Strategy / Patch size	DIBCO '9 [37]				H-DIBCO '12 [39]				DIBCO '17 [42]			
	FM \uparrow	p-FM \uparrow	PSNR \uparrow	DRD \downarrow	FM \uparrow	p-FM \uparrow	PSNR \uparrow	DRD \downarrow	FM \uparrow	p-FM \uparrow	PSNR \uparrow	DRD \downarrow
Direct Restoration / 256	93.83	96.25	20.34	2.75	96.09	97.16	23.07	1.42	92.21	94.33	18.93	2.80
Direct Restoration / 512	94.19	96.52	20.65	2.58	96.37	97.41	23.40	1.28	93.04	95.12	19.32	2.29
Cold Sampling / 256	93.55	96.05	20.03	2.71	95.70	96.68	22.69	1.46	89.57	91.66	18.18	3.61
Cold Sampling / 512	93.69	96.08	20.21	2.69	96.10	97.09	23.07	1.28	90.81	92.86	18.59	2.99

FM = F-Measure, p-FM = pseudo F-Measure PSNR = Peak Signal-to-Noise Ratio, DRD = Distance Reciprocal Distortion

Table 3. Performance evaluation of different methods for document binarization on all the DIBCO/H-DIBCO evaluation datasets. For each metric, the top **1st**, *2nd*, and 3rd methods are **bolded**, *italicized*, and underlined, respectively. The results presented here were generated using the **Direction Restoration / 512** evaluation strategy.

	Metrics	Methods													Ours	
		Otan [16]	Sauvola [44]	Lu [67]	Su [68]	Tenemeyer [52]	Vo [69]	He [70]	Zhao [27]	Sub [71]	Xiong [72]	Sonibgiu [45]	Jemni [58]	Yu [46]		
		1979	2000	2010	2013	2017	2018	2019	2019	2020	2021	2022	2022	2022		2023
		Thres.	Thres.	CV	CV	CNN-AE	CNN-AE	CNN-AE	GAN	GAN	SVM	Tr-VAE	Multitask-GAN	Multiple GANs	Diffusion	
Datasets	Metrics															
DIBCO '9 [37]	FM†	78.72	85.41	91.24	<u>93.50</u>	89.76	-	-	<i>94.10</i>	-	93.46	-	-	-	-	94.19
	p-FM†	-	-	-	-	<u>92.59</u>	-	-	95.26	-	-	-	-	-	-	96.52
	PSNR†	15.34	16.39	18.66	19.65	18.43	-	-	<u>20.30</u>	-	20.01	-	-	-	-	20.65
	DRD†	-	-	-	-	<u>4.82</u>	-	-	1.82	-	-	-	-	-	-	<u>2.58</u>
H-DIBCO '10 [11]	FM†	85.27	75.30	86.41	92.03	<i>94.89</i>	-	-	<u>94.03</u>	-	93.73	-	-	-	-	95.29
	p-FM†	90.83	84.22	88.25	94.85	97.65	-	-	<u>95.39</u>	-	95.18	-	-	-	-	<u>96.67</u>
	PSNR†	17.51	15.96	18.14	20.12	<i>21.84</i>	-	-	<u>21.12</u>	-	20.97	-	-	-	-	22.06
	DRD†	-	-	-	-	1.26	-	-	<u>1.58</u>	-	-	-	-	-	-	<u>1.36</u>
DIBCO '11 [38]	FM†	82.10	82.35	81.67	87.80	93.60	92.58	91.92	92.62	93.57	90.72	<i>94.37</i>	-	-	-	<u>94.08</u>
	p-FM†	85.96	88.63	-	-	97.70	94.67	95.82	95.38	95.93	-	96.15	-	-	-	<u>97.08</u>
	PSNR†	15.72	15.75	15.59	17.56	20.11	19.16	19.49	19.58	20.22	18.85	<i>20.81</i>	-	-	-	<u>20.51</u>
	DRD†	8.95	7.86	11.24	4.84	1.85	2.38	2.37	2.55	1.99	4.47	<i>1.63</i>	-	-	-	<u>1.25</u>
H-DIBCO '12 [39]	FM†	80.18	82.89	-	-	<u>92.53</u>	-	-	<i>94.96</i>	-	94.26	<i>95.31</i>	<u>95.18</u>	-	-	96.37
	p-FM†	82.65	87.95	-	-	96.67	-	-	96.15	-	95.16	<u>96.29</u>	94.63	-	-	97.41
	PSNR†	15.03	16.71	-	-	20.60	-	-	21.91	-	21.68	<i>22.29</i>	<u>22.00</u>	-	-	23.40
	DRD†	26.46	6.59	-	-	2.48	-	-	1.55	-	2.08	<u>1.6</u>	1.62	-	-	<u>1.28</u>
DIBCO '13 [12]	FM†	80.04	82.73	-	-	<u>93.17</u>	93.43	93.36	93.86	<u>95.01</u>	93.51	-	-	-	-	<i>95.34</i>
	p-FM†	83.43	88.37	-	-	96.81	95.34	<u>96.20</u>	96.47	96.49	94.54	-	-	-	-	97.51
	PSNR†	16.63	16.98	-	-	20.71	20.82	20.88	21.53	<u>21.69</u>	21.32	-	-	-	-	<i>22.27</i>
	DRD†	10.98	7.34	-	-	2.21	2.26	2.15	2.32	<u>1.76</u>	2.77	-	-	-	-	<i>1.59</i>
H-DIBCO '14 [40]	FM†	91.62	83.72	-	-	91.96	95.97	95.95	96.09	96.36	96.77	-	-	-	-	<u>96.65</u>
	p-FM†	95.69	87.49	-	-	94.78	97.42	98.76	98.25	97.87	97.73	-	-	-	-	<u>98.10</u>
	PSNR†	18.72	17.48	-	-	20.76	21.49	21.60	21.88	21.96	22.47	-	-	-	-	<u>22.47</u>
	DRD†	2.65	5.05	-	-	2.72	1.09	1.12	1.20	1.07	0.95	-	-	-	-	<u>0.96</u>
H-DIBCO '16 [41]	FM†	86.59	84.27	-	-	89.52	90.01	91.19	<u>91.66</u>	92.24	89.64	-	-	-	-	94.95
	p-FM†	89.92	89.10	-	-	93.76	93.44	<u>95.74</u>	94.58	95.95	93.56	-	-	-	-	94.55
	PSNR†	17.79	17.15	-	-	18.67	18.74	19.51	19.64	<i>19.93</i>	18.69	-	-	-	-	21.85
	DRD†	5.58	6.09	-	-	3.76	3.91	3.02	2.82	<u>2.77</u>	4.03	-	-	-	-	1.56
DIBCO '17 [42]	FM†	77.73	77.11	-	-	-	-	-	90.73	-	89.37	92.53	89.80	<u>90.05</u>	-	93.04
	p-FM†	77.89	84.10	-	-	-	-	-	92.58	-	90.8	95.15	89.95	<u>93.79</u>	-	<i>95.12</i>
	PSNR†	13.85	14.25	-	-	-	-	-	17.83	-	17.99	<i>19.11</i>	17.45	<u>18.27</u>	-	19.32
	DRD†	15.54	8.85	-	-	-	-	-	3.58	-	5.51	<i>2.37</i>	4.03	<u>2.94</u>	-	2.29
H-DIBCO '18 [43]	FM†	51.45	67.81	-	-	-	-	-	87.73	-	88.34	89.21	92.41	-	-	<i>91.66</i>
	p-FM†	53.05	74.08	-	-	-	-	-	90.60	-	90.37	92.54	<i>92.35</i>	-	-	95.53
	PSNR†	9.74	13.78	-	-	-	-	-	18.37	-	19.11	19.47	20.18	-	-	<i>20.02</i>
	DRD†	59.07	17.69	-	-	-	-	-	4.58	-	4.93	3.96	2.60	-	-	<i>2.81</i>

FM=F-Measure, p-FM=pseudo F-Measure PSNR=Peak Signal-to-Noise Ratio, DRD=Distance Reciprocal Distortion

Transformer-based autoencoders [45]. The results of our evaluation are summarized in Table 3, where FM, p-FM, PSNR, and DRD of each method are compared for different DIBCO/H-DIBCO datasets, with the top three approaches for each dataset **bolded**, *italicized* and underlined, respectively. As shown, our approach outperforms existing classical and state-of-the-art (SotA) approaches on 7 datasets, including DIBCO '9 [37], H-DIBCO '10 [11], DIBCO '11 [38], H-DIBCO '12 [39], DIBCO '13 [12], H-DIBCO '14 [40], and DIBCO '17 [42], ranking first on the majority of metrics, while performing competitively on the remaining 2 datasets H-DIBCO '16 [41] and H-DIBCO '18 [43]. It is worth mentioning that a number of recent SotA binarization techniques, including those presented by Yu *et al.* [46] and Jemni *et al.* [58], utilize several training stages, networks, or target objectives in order to achieve the reported results. Comparatively, our approach employs only a single diffusion network in an end-to-end fashion, and is able to outperform these methods across multiple datasets.

On DIBCO '9 [37] dataset, our approach scored the highest on all metrics except DRD, on which it ranked second. Furthermore, it demonstrated signifi-



Fig. 3. Qualitative results of our proposed method for the restoration of a few samples from the DIBCO and H-DIBCO datasets. These images are arranged in columns as follows: Left: original image, Middle: ground truth image, Right: binarized image using our proposed method

cant improvements in FM and PSNR on the H-DIBCO '10 [11] and DIBCO '11 [38] datasets in comparison to existing methods. We also observed a particularly noticeable improvement in PSNR with our approach on the H-DIBCO '12 [39], DIBCO '13 [12], and H-DIBCO '14 [40] datasets, with increases of 1.11, 1.71, and 1.91 compared to the previous state-of-the-art method, respectively. Similarly, despite lower DRD values on some datasets, it was significantly improved for these three datasets, with values of 1.28, 1.20, and 0.66, respectively. Similar performance improvements were observed on the DIBCO '17 [42] dataset as well,

where our approach ranked first on FM, PSNR, and DRD, and ranked second on p-FM. On H-DIBCO '18 [43], our approach placed third; however, it is evident from the results that our model demonstrated comparable performance to the top approaches.

Despite the high performance achieved on other datasets, our approach failed to achieve satisfactory results on the H-DIBCO '16 [41] dataset. Interestingly, upon inspecting the binarization outputs, we found that our approach was, in fact, quite capable of producing high quality binarization results for this dataset. The approach, however, had the tendency to generate slightly thicker text strokes compared to the ground truth images, which may explain why it did not produce the best quantitative results on this dataset. Figure 2 illustrates this effect by presenting two samples from the H-DIBCO '16 [41] dataset along with their corresponding ground truth images, binarized images derived from our method, and their difference. As can be seen from the difference image, our proposed approach produces binarized outputs very similar to the ground truth but with slightly thicker strokes in comparison. Overall, we observed that our approach demonstrated relatively consistent performance across the majority of DIBCO datasets and provided the highest FM and PSNR.

4.4 Qualitative Evaluation

This section presents a qualitative analysis of the binarization performance of our approach. In Fig. 3, we compare the binarization results of our approach with the ground truth for a few randomly selected samples from the different DIBCO and H-DIBCO datasets. As evident from the figure, our approach was highly effective at removing various types of noise, such as stains, smears, faded

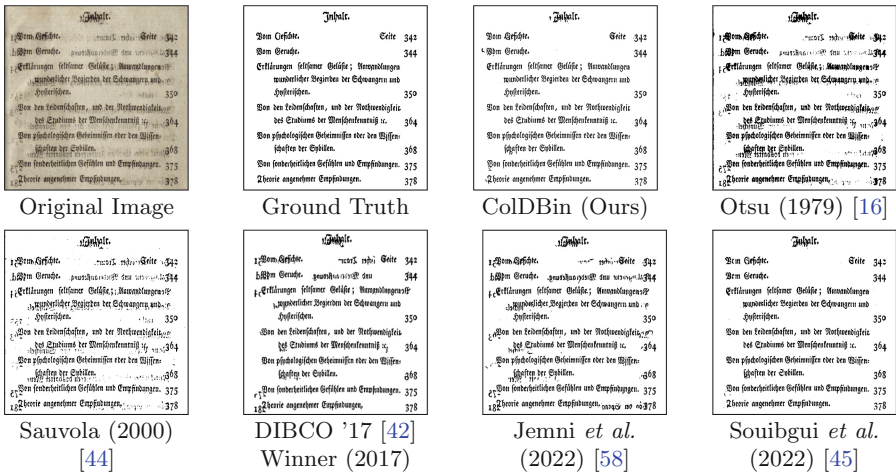


Fig. 4. Document binarization results for the input image 12 of DIBCO '17 [42] by different methods

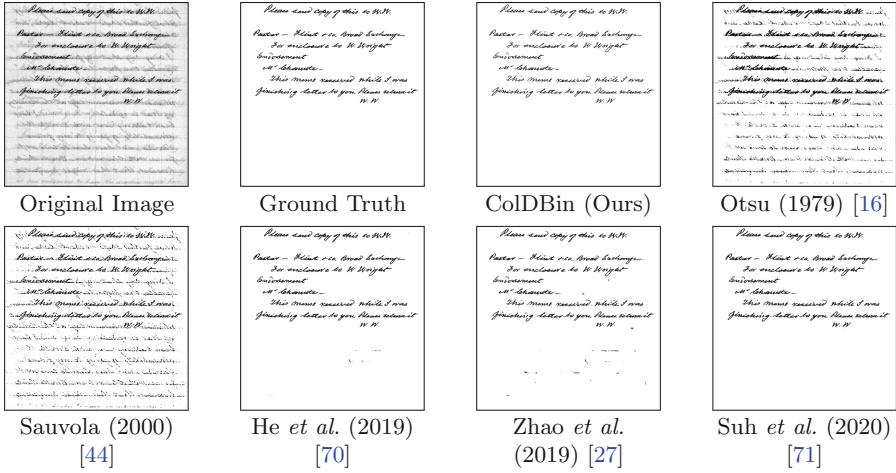


Fig. 5. Document binarization results for the input image HW5 of DIBCO '13 [12] by different methods

text, and background texture from a number of degraded document images. Moreover, it was able to produce high-quality binarized images that were visually comparable to the corresponding ground truth images, reflecting the exceptional quantitative performance discussed in the previous section.

Aside from comparisons with ground truth, we also compare the results of our approach to both classical and existing state-of-the-art (SotA) approaches. Figure 4 illustrates the binarization performance of various approaches, including ours, on sample 12 of the DIBCO '17 [42] dataset. The results demonstrate that our approach was successful in restoring a highly degraded document sample that many other approaches, including the multi-task GAN approach by Jenni *et al.* [58], failed to sufficiently restore. Interestingly, our results for this sample were visually similar to those obtained by Souibgui *et al.* [45], who used an encoder-decoder Transformer architecture for binarization. In Fig. 5, we compare the binarization performance of various approaches on another sample, namely, the HW5 from the DIBCO '13 [12] dataset. As can be seen, our approach was successful in restoring the image entirely, with the resulting image looking strikingly identical to the ground truth image. Additionally, we observed that our results for this sample were similar but slightly better than those of Suh *et al.* [71] and Yu *et al.* [46], who employed two-stage and three-stage GAN-based approaches for binarization, respectively.

4.5 Runtime Evaluation

In this section, we briefly analyze the runtime of our approach and compare it with other approaches. Since binarization speed depends on the size of input images, we evaluate the runtime in terms of secs/megapixel (MP) as used in

Table 4. Average runtimes for different binarization methods.

Runtime of different methods (secs/megapixel (MP))													
Otsu [16]	Sauvola [44]	Niblack [47]	Lu [67]	Su [68]	Bhowmik [51]	Tenssemeyer [52]	Vo [69]	Zhao [27]	Xiong [72]	Ours (D-256)	Ours (D-512)	Ours (S-256)	Ours (S-512)
Thres.	Thres.	Thres.	CV	CV	Game Theory	CNN-AE	CNN-AE	GAN	SVM	Diffusion	Diffusion	Diffusion	Diffusion
0.042	0.092	0.106	12.839	7.372	80.845	6.436	3.043	0.9819	19.306	0.9679	0.9918	135.47	193.49

D-256= Direct Reconstruction (256×256), **D-512** = Direct Reconstruction (512×512), **S-256** = Cold Sampling with T=200 (256×256), **S-512** = Cold Sampling with T=200 (512×512)

prior works [27, 72]. Both direct reconstruction and cold sampling were evaluated using a single NVIDIA GTX 1080Ti GPU with batch sizes of 4 and 32 for 512×512 and 256×256 image resolutions, respectively. The evaluation runtimes for other approaches were obtained directly from two papers [27, 72], which may have used different resources for evaluation and therefore we are only able to make a rough comparison. As shown in Table 4, with direct reconstruction, our approach had a runtime of ~ 1 sec/MP for both input image resolutions, which is comparable to the approach developed Zhao *et al.* [27], and is much lower than other computer vision methods [67, 68] and deep learning approaches [52, 69]. In contrast, the runtime for cold sampling scaled proportionally with the number of diffusion steps T. With T=200 in our experiments, for a 256×256 input resolution, sampling took $\sim 135\times$ more time than direct reconstruction, and for a 512×512 input resolution, it took $\sim 193\times$ more time. Thus, direct reconstruction was not only effective quantitatively and qualitatively, but also time-efficient in comparison with sampled reconstruction. It is worth noting that the problem of unreasonably high sampling times in diffusion models is well-known, and different sampling strategies [64, 73] have been proposed recently to overcome this problem.

5 Conclusion

This paper presents an end-to-end approach for binarization of document images using cold diffusion, which involves gradually transforming clean images into their degraded counterparts and then training a diffusion model that learns to reverse that process. The proposed approach was evaluated on 9 different DIBCO document benchmark datasets, and our results demonstrate that it outperforms traditional and state-of-the-art methods on a majority of datasets and does equally well on others. Despite its promising potential for document binarization, we believe it is also pertinent to discuss its limitations. As is the case with deep networks generally, the reliability of our models was quite dependent on the availability of data. While training datasets (DIBCO and Palm Leaf combined) have quite a lot of diversity in terms of sample distribution, the intra-class variance of samples was rather low, which necessitated training the models for a large number of iterations with various data augmentations in order to achieve the reported results. Therefore, to further enhance the performance of deep network-based approaches in the future, it may be worthwhile to invest resources in the creation of a large independent and diverse training dataset (whether synthetic or not) for binarization. We also observed a significant correlation between patch size and binarization performance with our approach. To

address this issue in the future, it may be worthwhile to investigate the possibility of conditioning the output of our model on the surrounding context of each image patch.

References

1. Afzal, M.Z., Kolsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, vol. 1, pp. 883–888 (2017)
2. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: pre-training of text and layout for document image understanding. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 20, pp. 1192–1200 (2020)
3. Li, P., et al.: SelfDoc: self-supervised document representation learning (2021). <https://arxiv.org/abs/2106.03331>
4. Hradiš, M., Kotera, J., Zemčík, P., Šroubek, F.: Convolutional neural networks for direct text deblurring. In: Proceedings of BMVC, vol. 10, no. 2 (2015)
5. 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 Recogn.* **109**, 107577 (2021)
6. Souibgui, M.A., Kessentini, Y.: DE-GAN: a conditional generative adversarial network for document enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020)
7. Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: DocXClassifier: towards an interpretable deep convolutional neural network for document image classification **9** (2022). <https://doi.org/10.36227/techrxiv.19310489.v4>
8. Subramani, N., Matton, A., Greaves, M., Lam, A.: A survey of deep learning approaches for OCR and document understanding. *ArXiv*, abs/2011.13534 (2020)
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186 (2018). <https://arxiv.org/abs/1810.04805v2>
10. Sulaiman, A., Omar, K., Nasrudin, M.F.: Degraded historical document binarization: a review on issues, challenges, techniques, and future directions. *J. Imag.* **5**(4) (2019). <https://www.mdpi.com/2313-433X/5/4/48>
11. Pratikakis, I., Gatos, B., Ntirogiannis, K.: H-DIBCO 2010 - handwritten document image binarization competition. In: 2010 12th International Conference on Frontiers in Handwriting Recognition, pp. 727–732 (2010)
12. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2013 document image binarization contest (DIBCO 2013). In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1471–1476 (2013)
13. Bako, S., Darabi, S., Shechtman, E., Wang, J., Sunkavalli, K., Sen, P.: Removing shadows from images of documents. In: Asian Conference on Computer Vision (ACCV 2016) (2016)
14. Chen, X., He, X., Yang, J., Wu, Q.: An effective document image deblurring algorithm. In: CVPR 2011, pp. 369–376 (2011)

15. Saifullah, S., Siddiqui, S.A., Agne, S., Dengel, A., Ahmed, S.: Are deep models robust against real distortions? A case study on document image classification. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1628–1635 (2022)
16. Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979)
17. Xiong, W., Xu, J., Xiong, Z., Wang, J., Liu, M.: Degraded historical document image binarization using local features and support vector machine (SVM). *Optik* **164**, 218–223 (2018)
18. Bhunia, A.K., Bhunia, A.K., Sain, A., Roy, P.P.: Improving document binarization via adversarial noise-texture augmentation. In: IEEE International Conference on Image Processing (ICIP) 2019, pp. 2721–2725 (2019)
19. Neji, H., Halima, M.B., Hamdani, T.M., Noguera-Iso, J., Alimi, A.M.: Blur2Sharp: a GAN-based model for document image deblurring. *Int. J. Comput. Intell. Syst.* **14**, 1315–1321 (2021). <https://doi.org/10.2991/ijcis.d.210407.001>
20. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. *Foundations Trends® Mach. Learn.* **12**(4), 307–392 (2019). <https://doi.org/10.15612F2200000056>
21. Goodfellow, I., et al.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020)
22. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation (2017). <https://arxiv.org/abs/1710.10196>
23. Mao, X., Shen, C., Yang, Y.-B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates Inc. (2016). <https://proceedings.neurips.cc/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf>
24. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks (2015). <https://arxiv.org/abs/1501.00092>
25. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks (2016). <https://arxiv.org/abs/1611.07004>
26. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention (2018). <https://arxiv.org/abs/1801.07892>
27. Zhao, J., Shi, C., Jia, F., Wang, Y., Xiao, B.: Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recogn.* **96**, 106968 (2019). <https://www.sciencedirect.com/science/article/pii/S0031320319302717>
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851. Curran Associates Inc. (2020). <https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>
29. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis (2021). <https://arxiv.org/abs/2105.05233>
30. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models (2022). <https://arxiv.org/abs/2206.00364>
31. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022). <https://arxiv.org/abs/2205.11487>
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP Latents (2022). <https://arxiv.org/abs/2204.06125>

33. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models (2022). <https://arxiv.org/abs/2201.11793>
34. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–14 (2022)
35. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16 272–16 282 (2022)
36. Kawar, B., Song, J., Ermon, S., Elad, M.: Jpeg artifact correction using denoising diffusion restoration models (2022). <https://arxiv.org/abs/2209.11888>
37. Gatos, B., Ntirogiannis, K., Pratikakis, I.: DIBCO 2009: document image binarization contest. *IJDAR* **14**, 35–44 (2011)
38. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 document image binarization contest (DIBCO 2011). In: International Conference on Document Analysis and Recognition 2011, pp. 1506–1510 (2011)
39. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In: International Conference on Frontiers in Handwriting Recognition 2012, pp. 817–822 (2012)
40. Ntirogiannis, K., Gatos, B., Pratikakis, I.: ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 809–813 (2014)
41. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 619–623 (2016)
42. Pratikakis, I., Zagoris, K., Barlas, G., Gatos, B.: ICDAR2017 competition on document image binarization (DIBCO 2017). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1395–1403 (2017)
43. Pratikakis, I., Zagori, K., Kaddas, P., Gatos, B.: ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018). In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 489–493 (2018)
44. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. *Pattern Recogn.* **33**(2), 225–236 (2000). <https://www.sciencedirect.com/science/article/pii/S0031320399000552>
45. Souibgui, M.A.: DocEnTr: an end-to-end document image enhancement transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR) (2022)
46. 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 (2022). <https://arxiv.org/abs/2211.16098>
47. Niblack, W.: An Introduction to Digital Image Processing. Strandberg Publishing Company, DNK (1985)
48. Ntirogiannis, K., Gatos, B., Pratikakis, I.: A combined approach for the binarization of handwritten document images. *Pattern Recogn. Lett.* **35**, 3–15 (2014). *Frontiers in Handwriting Processing*. <https://www.sciencedirect.com/science/article/pii/S016786551200311X>
49. Pinto, T., Rebelo, A., Giraldo, G.A., Cardoso, J.S.: Music score binarization based on domain knowledge. In: Iberian Conference on Pattern Recognition and Image Analysis (2011)

50. Ahmadi, E., Azimifar, Z., Shams, M., Famouri, M., Shafiee, M.J.: Document image binarization using a discriminative structural classifier. *Pattern Recogn. Lett.* **63**(C), 36–42 (2015). <https://doi.org/10.1016/j.patrec.2015.06.008>
51. Bhowmik, S., Sarkar, R., Das, B., Doermann, D.S.: GIB: a game theory inspired binarization technique for degraded document images. *IEEE Trans. Image Process.* (2019)
52. Tensmeyer, C., Martinez, T.: Document image binarization with fully convolutional neural networks (2017). <https://arxiv.org/abs/1708.03276>
53. Akbari, Y., Al-Maadeed, S., Adam, K.: Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images. *IEEE Access* **8**, 153 517–153 534 (2020)
54. Lore, K.G., Akintayo, A., Sarkar, S.: LLNet: a deep autoencoder approach to natural low-light image enhancement (2015). <https://arxiv.org/abs/1511.03995>
55. Calvo-Zaragoza, J., Gallego, A.-J.: A selectional auto-encoder approach for document image binarization. *Pattern Recogn.* **86**, 37–47 (2019). <https://www.sciencedirect.com/science/article/pii/S0031320318303091>
56. Pastor-Pellicer, J., Boquera, S.E., Zamora-Martínez, F., Afzal, M.Z., Bleda, M.J.C.: Insights on the use of convolutional neural networks for document image binarization. In: *International Work-Conference on Artificial and Natural Neural Networks* (2015)
57. Castellanos, F.J., Gallego, A.-J., Calvo-Zaragoza, J.: Unsupervised neural domain adaptation for document image binarization. *Pattern Recogn.* **119**, 108099 (2021)
58. Jemni, S.K., Souibgui, M.A., Kessentini, Y., Fornés, A.: Enhance to read better: a multi-task adversarial network for handwritten document image enhancement. *Pattern Recogn.* **123**, 108370 (2022). <https://doi.org/10.1016%2Fj.patcog.2021.108370>
59. Dosovitskiy, A.: An image is worth 16x16 words: transformers for image recognition at scale (2020). <https://arxiv.org/abs/2010.11929>
60. Bansal, A., et al.: Cold diffusion: inverting arbitrary image transforms without noise (2022). <https://arxiv.org/abs/2208.09392>
61. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation (2015). <https://arxiv.org/abs/1505.04597>
62. Vaswani, A.: Attention is all you need (2017). <https://arxiv.org/abs/1706.03762>
63. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022). <https://arxiv.org/abs/2201.03545>
64. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2020). <https://arxiv.org/abs/2010.02502>
65. Suryani, M., Paulus, E., Hadi, S., Darsa, U.A., Burie, J.-C.: The handwritten Sundanese palm leaf manuscript dataset from 15th century. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, pp. 796–800 (2017)
66. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, ser. *Proceedings of Machine Learning Research*, vol. 139. PMLR, 18–24 July 2021, pp. 8162–8171 (2021). <https://proceedings.mlr.press/v139/nichol21a.html>
67. Lu, S., Su, B., Tan, C.L.: Document image binarization using background estimation and stroke edges. *Int. J. Doc. Anal. Recogn. (IJ DAR)* **13**(4), 303–314 (2010)
68. Su, B., Lu, S., Tan, C.L.: Robust document image binarization technique for degraded document images. *IEEE Trans. Image Process.* **22**(4), 1408–1417 (2013)

69. Vo, Q.N., Kim, S., Yang, H.-J., Lee, G.: Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognit.* **74**, 568–586 (2018)
70. He, S., Schomaker, L.: DeepOtsu: document enhancement and binarization using iterative deep learning. *Pattern Recogn.* **91**, 379–390 (2019). <https://doi.org/10.1016%2Fj.patcog.2019.01.025>
71. Suh, S., Kim, J., Lukowicz, P., Lee, Y.O.: Two-stage generative adversarial networks for document image binarization with color noise and background removal (2020). <https://arxiv.org/abs/2010.10103>
72. Xiong, W., Zhou, L., Yue, L., Li, L., Wang, S.: An enhanced binarization framework for degraded historical document images. *J. Image Video Process.* **2021**(1) (2021). <https://doi.org/10.1186/s13640-021-00556-4>
73. Kong, Z., Ping, W.: On fast sampling of diffusion probabilistic models (2021)