

Article

# A Generative Approach for Document Enhancement with Small Unpaired Data

Mohammad Shahab Uddin <sup>1</sup>, Wael Khallouli <sup>2</sup>, Andres Sousa-Poza <sup>2</sup>, Samuel Kovacic <sup>2</sup> and Jiang Li <sup>1,\*</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529, USA; muddi003@odu.edu

<sup>2</sup> Department of Engineering Management & Systems Engineering, Old Dominion University, Norfolk, VA 23529, USA; wkhallou@odu.edu (W.K.); asousapo@odu.edu (A.S.-P.); skovacic@odu.edu (S.K.)

\* Correspondence: jli@odu.edu

**Abstract:** Shipbuilding drawings, crafted manually before the digital era, are vital for historical reference and technical insight. However, their digital versions, stored as scanned PDFs, often contain significant noise, making them unsuitable for use in modern CAD software like AutoCAD. Traditional denoising techniques struggle with the diverse and intense noise found in these documents, which also does not adhere to standard noise models. In this paper, we propose an innovative generative approach tailored for document enhancement, particularly focusing on shipbuilding drawings. For a small, unpaired dataset of clean and noisy shipbuilding drawing documents, we first learn to generate the noise in the dataset based on a CycleGAN model. We then generate multiple paired clean-noisy image pairs using the clean images in the dataset. Finally, we train a Pix2Pix GAN model with these generated image pairs to enhance shipbuilding drawings. Through empirical evaluation on a small Military Sealift Command (MSC) dataset, we demonstrated the superiority of our method in mitigating noise and preserving essential details, offering an effective solution for the restoration and utilization of historical shipbuilding drawings in contemporary digital environments.

**Keywords:** document enhancement; generative adversarial network; denoising; OCR; noise modeling



**Citation:** Uddin, M.S.; Khallouli, W.; Sousa-Poza, A.; Kovacic, S.; Li, J. A Generative Approach for Document Enhancement with Small Unpaired Data. *Electronics* **2024**, *13*, 3539. <https://doi.org/10.3390/electronics13173539>

Academic Editor: Abdussalam Elhanashi

Received: 29 July 2024

Revised: 28 August 2024

Accepted: 29 August 2024

Published: 6 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

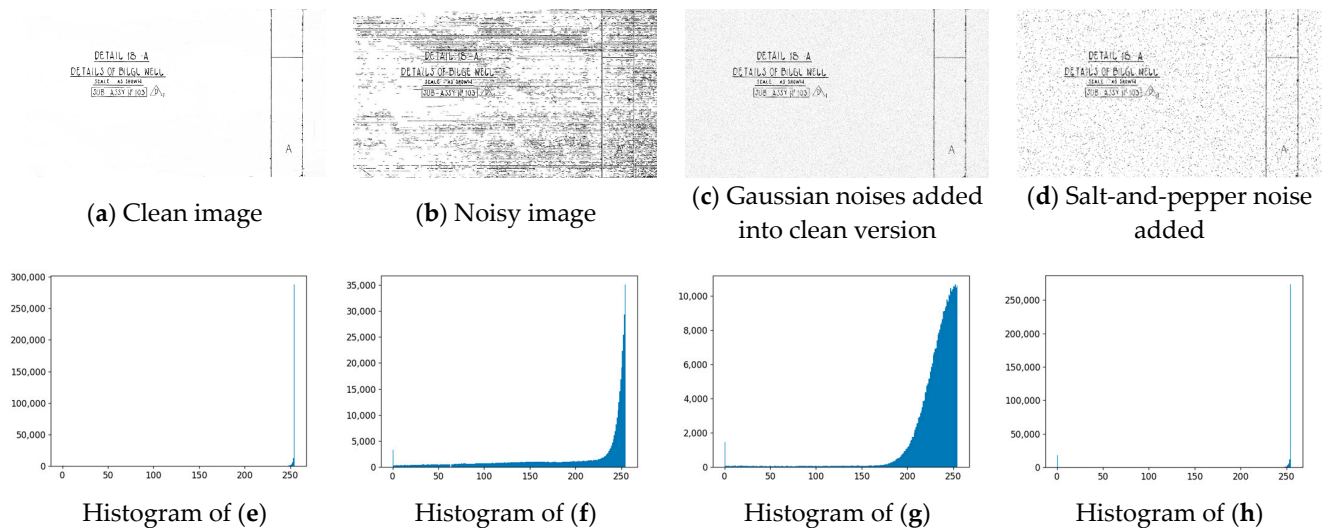
## 1. Introduction

Shipbuilding, one of the oldest industries known to humanity, traditionally relied on detailed drawings to guide construction processes before the advent of digital technology. Historically, these drawings were created by hand, making them not only invaluable for their technical accuracy but also as pieces of maritime heritage. However, with the rise of digital technologies, the shift from manual drafting to digital documents has become necessary, introducing several challenges, especially in the preservation and use of old drawings.

Scanned PDF documents, which serve as digital representations of these historical drawings, often suffer from inherent imperfections such as noise and artifacts introduced during the scanning process. These imperfections, compounded by the heavy and heterogeneous nature of the noise, render the documents unsuitable for seamless integration into modern computer-aided design (CAD) platforms like AutoCAD [1]. Traditional denoising methods, designed to mitigate noise in images adhering to well-defined noise models, prove ineffective when confronted with the complex noise structures present in scanned shipbuilding drawings [2–4].

The analysis of the noisy document images reveals that the noise is distinctly different from standard types like Gaussian and salt-and-pepper noise. In Figure 1, noisy image samples from the document show that the noise exhibits characteristics such as streaks, lines, or concentrated imperfections, unlike the random variations typical of Gaussian and salt-and-pepper noise. This suggests that the noise in Military Sealift Command (MSC) documents may stem from scanner artifacts, printer defects, or issues with the

original document. Histogram analysis (Figure 1) further supports this distinction: the original noisy image's histogram is skewed towards higher intensity values, unlike the more spread-out Gaussian noise or the distinct spikes of salt-and-pepper noise. These visual and statistical differences confirm that the noise in our document images is unique and does not follow common noise models.



**Figure 1.** Comparison of different noises. The clean version of the MSC document image shown in (a) was obtained by manually removing the noise in (b). We created the noisy images (c,d) from the clean image using gaussian noise and salt-and-pepper noise. We excluded the histogram for pixels of value “255” for better visualization in (e–h).

To address these challenges, we developed a specialized document enhancement strategy designed specifically for shipbuilding drawings stored as scanned PDF files. Initially, for a small dataset of clean and noisy shipbuilding drawings that were not paired, we trained a CycleGAN model [5] to simulate the noise patterns found in these documents. Subsequently, we created pairs of clean and noisy images from the clean samples in the dataset using the trained CycleGAN model. We then proceeded to train a Pix2Pix GAN model [6] using these image pairs to improve the quality of the shipbuilding drawings. Our method’s effectiveness was validated on a dataset from MSC, where it proved superior in reducing noise and retaining crucial details as compared to state-of-the-art methods.

## 2. Related Work

In this paper, we focus on document noise removal with the goal of preserving critical elements such as text, labels, and architectural details and eliminating scanning noise that cannot be modelled as common noise models such as Gaussian or salt-and-pepper noise. Regular image denoising is outside the scope of this paper. Our objective was to improve the readability and clarity of these documents without losing essential information. Document noise removal methods can be categorized into three groups: (1) traditional techniques that rely on basic image processing algorithms, (2) discriminative methods that employ machine learning models to classify and filter noise, and (3) generative approaches that use generative artificial intelligence (GAI) models to reconstruct clean images from noisy ones. Each of these methodologies offers unique advantages and challenges, which we summarize to highlight their contributions to the field of document image enhancement.

**Traditional document denoising methods.** Earlier work for document enhancement included global binarization, aiming to find a single threshold value for the entire document to eliminate those noise pixels, and local binarization, utilizing a dynamic threshold value for each pixel to classify image pixels into foreground (black) or background (white) [7,8]. Although thresholding methods continue to evolve, such as the global threshold selection

method based on fuzzy expert systems (FESs) that enhance image contrast and use a pixel-counting algorithm for threshold adjustment [9], they are sensitive to document condition and often fail to clean highly degraded images [10]. To address this challenge, energy-based methods were introduced, such as maximizing ink presence with an energy function while minimizing the degraded background [11] and using mathematical morphology to estimate background from the degraded image [12]. However, these handcrafted image processing algorithms often yielded unsatisfactory results.

**Discriminative methods.** Recently, discriminative deep learning has been introduced for document denoising. In [13], a 2D long short-term memory (LSTM) network was used to determine whether something belonged to text or background noise based on a sequence of its neighboring pixels. A vision-transformer-based encoder–decoder architecture was proposed in [14] to perform document enhancement through similar discriminative analysis for each pixel. However, a significant drawback of discriminative approaches is their dependency on paired datasets comprising noisy and clean images for effective training. Acquiring such paired datasets can be challenging, particularly for historical or rare documents, which limits the scalability and applicability of discriminative techniques for document denoising. This dependency often necessitates extensive manual annotation, which is both time-consuming and resource-intensive, constraining the practical implementation of discriminative models for real-world scenarios.

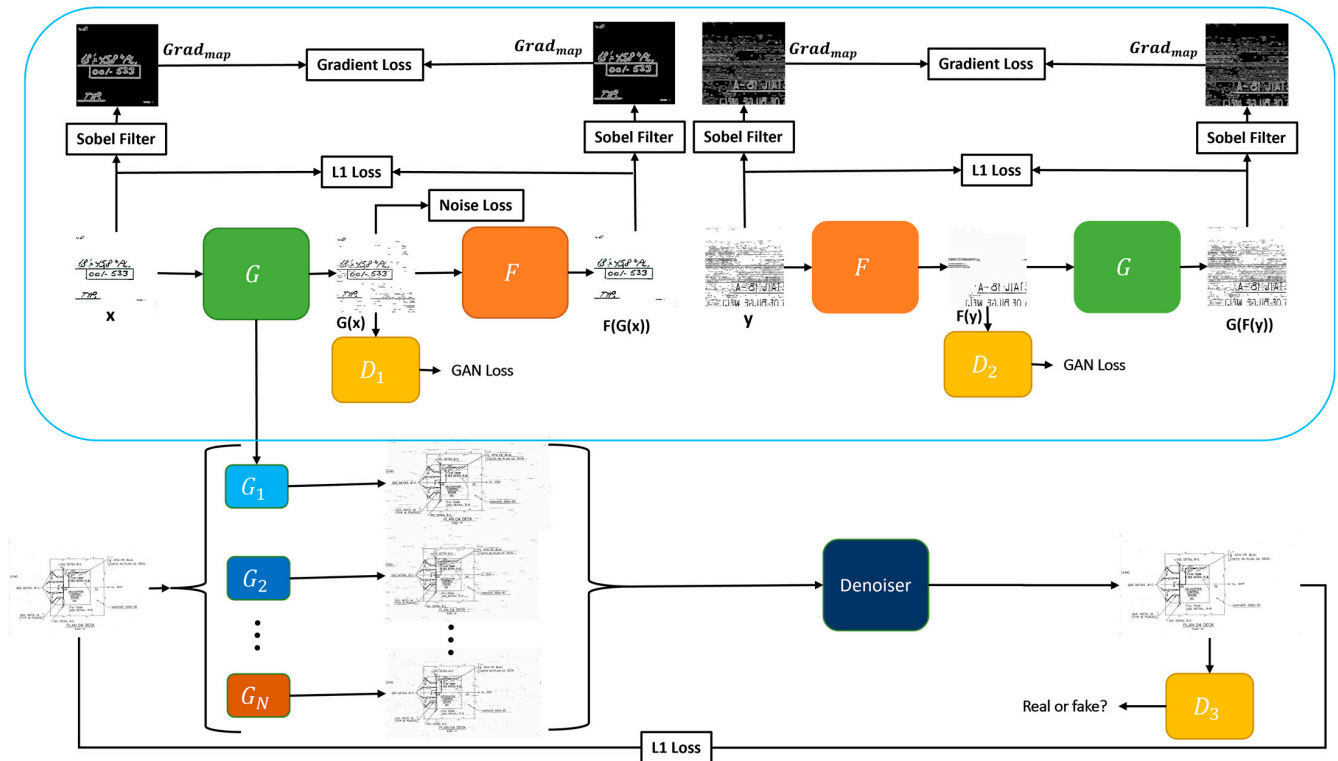
**Generative methods.** The realm of image denoising continues to evolve with the introduction of generative models like generative adversarial networks (GANs) [15] and diffusion models [16], and these can be grouped as those requiring paired noisy–clean data for training [17–19], not requiring [20–22], or hybrid [23]. Those generative methods requiring paired datasets for training typically employ the conditional GAN (cGAN) network [24] to learn a transformation function from noisy image domain to clean image domain, and the denoising task is converted as domain conversion. Recently, a diffusion-based framework [25] was proposed for document enhancement and it can be categorized in this group. The main drawback of these methods is that they require large, paired datasets to achieve competitive performance, which are not always possible in practice. For those approaches not requiring paired datasets for training, they typically employ the CycleGAN model to convert noisy image as clean image and vice versa under unsupervised learning with the guidance of the cycle-consistent GAN loss [5]. Therefore, they only require non-paired clean and noisy images for training. However, the performance of these approaches is degraded and not always satisfactory. The representative hybrid method [23] combines the unpaired learning capabilities of CycleGAN [5] with the paired learning advantages of Pix2Pix GAN [6] to enhance document images. This model still needs a paired image dataset for training, and it does not perform well when the available paired dataset is small like our situation.

In this paper, to address the challenges of the hybrid model for document enhancement, we propose novel loss functions to improve the training efficiency of the CycleGAN model. Additionally, we utilized multiple versions of the trained CycleGAN model to generate paired clean and noisy images for supervised training of a Pix2Pix GAN model to remove scanning noise from MSC documents.

### 3. Proposed Method

The overall architecture of the proposed model is shown in Figure 2. It builds upon CycleGAN and Pix2Pix GAN architectures and includes three stages of learning for document image denoising. In the first stage, we train a modified CycleGAN model to learn the noise model using unpaired clean and noisy images. CycleGAN consists of two generators,  $G$  and  $F$ , and two discriminators,  $D_1$  and  $D_2$ . Generator  $G$  converts clean images to noisy ones, while  $F$  does the opposite, and  $D_1$  and  $D_2$  perform adversarial learning in noisy and clean domains, respectively. During training, different versions of generator  $G$  are saved. In the second stage, we use the saved versions of  $G$  to generate different noisy images for each clean image in the dataset for data augmentation. These augmented data are paired

clean–noisy images. Finally, we train a Pix2Pix GAN model with the paired augmented dataset for effective denoising. Additionally, we propose novel loss functions to modify the CycleGAN training, including gradient loss and noise loss. After training, the trained Pix2Pix GAN model is used for denoising.



**Figure 2.** Overview of the proposed method. Blue boundary area shows our modified CycleGAN consisting of two generators, G and F, and two discriminators,  $D_1$  and  $D_2$ . G generates noisy images from clean inputs while F reconstructs clean images from noisy ones. The model is trained using a combination of L1 loss, gradient loss, noise loss, and GAN loss. Discriminators  $D_1$  and  $D_2$  are employed to differentiate between real and generated noisy images, as well as real and generated clean images, respectively. The lower part is the data augmentation process using G from the modified CycleGAN and the training of the Pix2Pix GAN model.

### 3.1. Modified CycleGAN for Noise Model Learning

We begin by training a modified CycleGAN (upper part in Figure 2) with a collection of unpaired clean and noisy document images. This CycleGAN architecture consists of two generators (G and F) and two discriminators ( $D_1$  and  $D_2$ ). Generator G transforms clean images into noisy versions, mimicking the observed noise patterns in our dataset. Conversely, generator F aims to recover clean images from noisy inputs. We used the ResNet [26] architecture with nine residual blocks for the generators proposed in CycleGAN, implementing patch-based architecture for our discriminators as proposed in [27].

The training process employs a combination of L1 loss to ensure pixel-level similarity between input and recovered clean images, and a gradient loss to encourage realistic noise patterns generated by G. During training,  $D_1$  differentiates between real noisy images and those generated by G, while  $D_2$  distinguishes real clean images from F’s outputs. We trained this part following the implementation of the CycleGAN with the GAN loss [6,15],

$$L_{GAN} = L_{GAN\_G}(G, D_1, X, Y) + L_{GAN\_F}(F, D_2, Y, X)$$

where  $X, Y$  are the domains of the clean and noisy images respectively. We also used the L1 loss between original unpaired clean and noisy images  $(x, y)$  and reconstructed images  $(F(G(x)), G(F(y)))$  as proposed in [5],

$$L_{L1Loss} = |F(G(x)) - x|_1 + |G(F(y)) - y|_1$$

We found that there is a positive correlation between noise level and gradient magnitude. For example, a noisy image has a larger gradient magnitude than its clean version, as shown in Figure 3. We proposed a novel gradient loss to encourage the generators to focus on the noise during training since our goal is to learn the noise model for paired data augmentation. The gradient loss is calculated using the Sobel filter to extract gradient magnitudes from images. While this approach uses gradient magnitude like the existing image processing methods [28–32], we calculated the gradient loss focusing on the noises. Our focus was on its integration with other loss functions to enhance the overall performance during training the modified CycleGAN. The key difference lies in how we combine the gradient loss with other loss functions to optimize both perceptual quality and structural preservation, rather than relying on a single aspect of image quality. The combination of gradient loss with other losses in our framework contributes to the improved performance demonstrated in our results.

$$L_{Gradient\_Loss} = |Grad(F(G(x))) - Grad(x)|_1 + |Grad(G(F(y))) - Grad(y)|_1$$

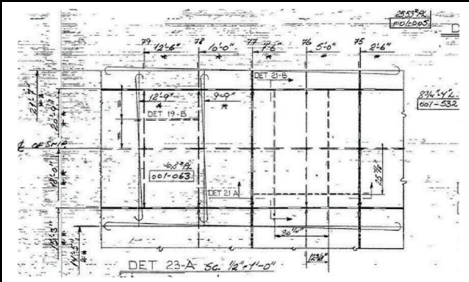
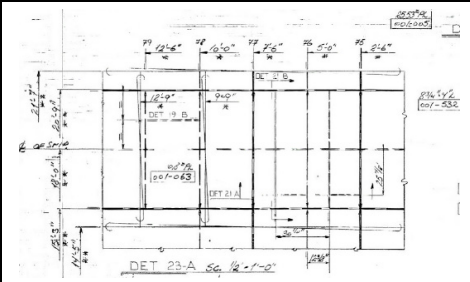
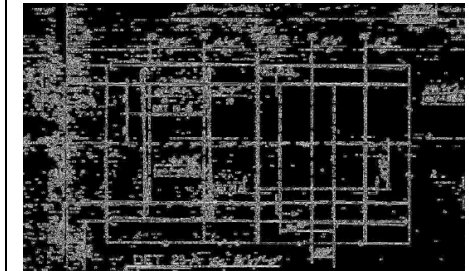
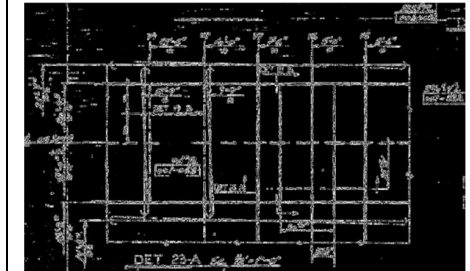
	Noisy	Clean
Image		
Gradient along x and y direction using Sobel operator		
Gradient Magnitude	19.56	9.28

Figure 3. Positive correlation between gradient magnitude and noise level.

Here,

$$Grad(.) = \frac{1}{N} \sum_N Grad_{map}(p, q)$$

$$Grad_{map}(p, q) = \sqrt{G_{x-axis}(p, q) + G_{y-axis}(p, q)}$$

$(p, q)$  is the location of a pixel in the gradient maps  $G_{x-axis}$  and  $G_{y-axis}$  generated by the Sobel filter from an image along the  $x$ -axis and  $y$ -axis, respectively.  $Grad_{map}(p, q)$  represents gradient magnitude at  $(p, q)$  pixel location.  $N$  represents the number of pixels with gradient magnitude below the threshold. We considered only the gradient magnitudes below a certain threshold. This threshold is chosen to distinguish between noise and edges (where

high gradients typically represent edges). During our experiments, we set 200 as the threshold value. So,  $Grad(\cdot)$  is the average of these low gradient magnitudes of an image.

As generator  $G$  is responsible for generating a noised version of a clean image, we impose a noise loss on the output of  $G$ . We calculated noise loss as the gradient magnitude of the output image,

$$L_{Noise_{loss}} = -Grad(G(x))$$

and the overall loss function for the modified CycleGAN model is

$$L_{mCycleGAN} = L_{GAN} + \alpha_1 * L_{L1Loss} + \alpha_2 * L_{Gradient_{Loss}} + \alpha_3 * L_{Noise_{loss}}$$

where  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  are hyperparameters combining the difference loss functions.

### 3.2. Paired Clean–Noisy Image Generation for Data Augmentation

To capture the variability of document noise, we checkpoint the model  $G$  at multiple stages during training, resulting in a collection of models  $G\_ensemble = \{G_1, G_2, \dots, G_{20}\}$ . Each  $G$  within the ensemble captures the noise characteristics at a specific point in training. After the training of the CycleGAN model, we utilize the  $G\_ensemble$  to generate a diverse set of noisy variations for each clean image  $x$  in  $X$  as  $X\_noise = \{G_1(x), G_2(x), \dots, G_N(x)\}$ , which are then used in the subsequent Pix2Pix GAN training.

### 3.3. Training of Pix2Pix GAN for Denoising

The final stage employs a Pix2Pix GAN architecture as a denoiser trained on the newly created paired dataset of clean and noisy images generated in stage 2. This Pix2Pix GAN utilizes a single generator denoiser that maps noisy document images to their clean counterparts. A discriminator ( $D_3$ ) guides the training process by distinguishing between real clean images and those produced by the denoiser. The architecture of the denoiser is similar to the architecture of  $G$  and  $F$ . Also, all the discriminators in our model have the same architecture of ResNet.

The GAN loss used for training  $D_3$  is

$$L_{cGAN}(Denoiser, D_3) = \mathbb{E}_{x,y}[\log D_3(x)] + \mathbb{E}_x[\log(1 - D_3(Denoiser(x\_noise)))]$$

We also used L1 loss between target and output from the denoiser:

$$L_{L1\_loss\_denoiser} = \mathbb{E}_{x,G(x)}[|x - Denoiser(x\_noise)|_1]$$

and the overall loss for training the Pix2Pix GAN is

$$L_{secondstep} = L_{cGAN}(Denoiser, D_3) + \beta * L_{L1\_loss\_denoiser}$$

where  $\beta$  is a hyperparameter combining the two loss functions.

### 3.4. Evaluation Metrics

**Natural image quality evaluator (NIQE):** NIQE [33] is a no-reference image quality assessment metric that measures the statistical naturalness of an image. It does so by comparing the visual characteristics of the denoised image to a model of natural images. Unlike traditional metrics that require a reference image for comparison, NIQE operates independently, making it ideal for evaluating referenceless denoised images. It provides a quantitative score that reflects the degree of distortion or unnaturalness in an image, helping to ensure that the denoising process maintains the intrinsic properties of natural scenes.

**Ma score:** The Ma score [34] is designed for evaluating the quality of super-resolved images without requiring reference images. To calculate the Ma score for an image, we need to extract three groups of statistical features: local frequency features, global frequency features, and spatial features. These features encompass the distribution of discrete cosine transform coefficients, wavelet coefficients, and the spatial discontinuity properties of pixel

intensity. Ma score utilizes three regression forests to independently model each group of features. The outputs from these forests are then combined linearly to estimate the final perceptual quality score. Ma score demonstrates a strong correlation with subjective evaluations of visual quality in super-resolved images, providing an effective metric for assessing image enhancement results [35].

**Perceptual Index (PI):** PI [36] is a comprehensive metric used to evaluate the quality of denoised images without ground truth. It combines the NIQE and the Ma scores, offering a unified measure of perceived image quality. By integrating both objective and subjective assessments, PI provides a robust indicator of how natural and visually pleasing an image appears. This makes it particularly useful in scenarios where human visual perception is a critical factor in judging image quality, ensuring that denoised images meet the expected standards of visual appeal,

$$PI = \frac{1}{2}((10 - Ma) + NIQE)$$

**Character error rate (CER):** CER is an evaluation metric commonly used in optical character recognition (OCR) tasks, which can also be applied to denoised images. It measures the percentage of characters in the denoised image that are incorrectly recognized by an OCR system. This metric is crucial for assessing the functional quality of denoised images, especially in applications where text readability is important. A lower CER indicates better preservation of text information, signifying that the denoising process has effectively maintained the legibility of characters within the image.

$$CER = \frac{\text{Total number of substitutions} + \text{Total number of insertions} + \text{Total number of deletions}}{\text{Total number of characters in ground truth}}$$

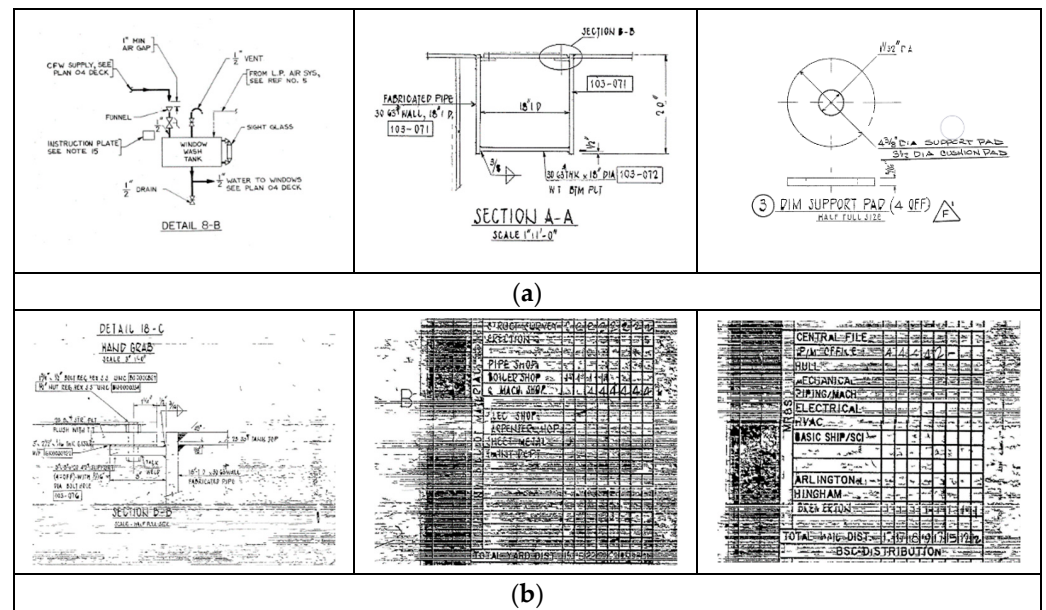
**Word error rate (WER):** WER is another OCR-based metric used to evaluate the quality of denoised images by measuring the accuracy of word recognition. It calculates the percentage of words that are incorrectly transcribed by an OCR system. Similar to CER, WER is essential for determining how well the denoising process preserves the readability of text in images. A lower WER means that more words are correctly recognized, indicating higher functional fidelity of the denoised image. This metric is particularly valuable in contexts where the accuracy of text extraction is critical, such as document scanning and archival applications.

$$WER = \frac{\text{Total number of substitutions} + \text{Total number of insertions} + \text{Total number of deletions}}{\text{Total number of words in ground truth}}$$

## 4. Experiment Setup

### 4.1. Dataset

The main challenge in training our models was the limited data availability. We had seven documents from MSC with just one document containing noise in specific areas. Six documents had 12 pages in total and the remaining document had 9 pages with noises at some specific locations. We created a dataset of 16 noisy and 117 clean images cropped from these documents for training and testing. Since we had only one document with noises in some locations, we cropped 16 noisy images from it. For training, we utilized 10 noisy and 67 clean images. Figure 4 shows examples of clean and noisy samples from the training set. Using this unpaired dataset, we trained the modified CycleGAN. We then employed the modified CycleGAN to generate 80 noisy images for each of the remaining 50 clean images of the dataset, resulting in a total of 4000 noisy–clean pairs. We used these pairs to train the Pix2Pix GAN model-based denoiser for denoising. Finally, the remaining 6 noisy images were used to test the denoiser.



**Figure 4.** Image samples from training dataset. (a) Clean samples from training dataset; (b) noisy samples from training dataset.

#### 4.2. Competing Methods

In our study, we compared our proposed method with three other prominent approaches for document noise removal: the CycleGAN model [5], the Otsu method [7], Sauvola’s method [8], and a hybrid method [23] combining CycleGAN and Pix2Pix GAN.

CycleGAN [5] is an unsupervised learning approach that aims to learn mappings between two domains without paired examples. In the context of document noise removal, CycleGAN is used to transform noisy document images into clean ones and vice versa. In our proposed method, the basic CycleGAN model is also utilized in the first step to convert noisy images to clean ones.

The Otsu method [7] is a traditional image processing technique used for global binarization. It works by calculating a single global threshold value that minimizes the intra-class variance between the foreground and background pixels. This method is computationally efficient and straightforward to implement. However, its main drawback lies in its sensitivity to document conditions. For highly degraded or unevenly illuminated documents, the Otsu method often fails to accurately distinguish between text and noise, leading to significant loss of information and poor denoising performance. Its reliance on a single global threshold makes it unsuitable for complex noise patterns that vary across the document.

Sauvola’s method [8] is an effective local thresholding technique for images with non-uniform backgrounds, particularly in text recognition applications. Rather than computing a single global threshold for the entire image, this approach calculates multiple thresholds for each pixel. These thresholds are determined using specific formulas that consider the mean and standard deviation within a local neighborhood, defined by a window centered on the pixel.

The hybrid method, mentioned in [23], combines CycleGAN and Pix2Pix GAN. In this approach, CycleGAN is first used to generate synthetic paired datasets from unpaired noisy and clean images. Each clean image just generates one clean–noisy image pair. Our proposed method builds on the hybrid approach, including CycleGAN and Pix2Pix GAN, by introducing novel loss functions and the generation of multiple clean–noisy image pairs for denoising.



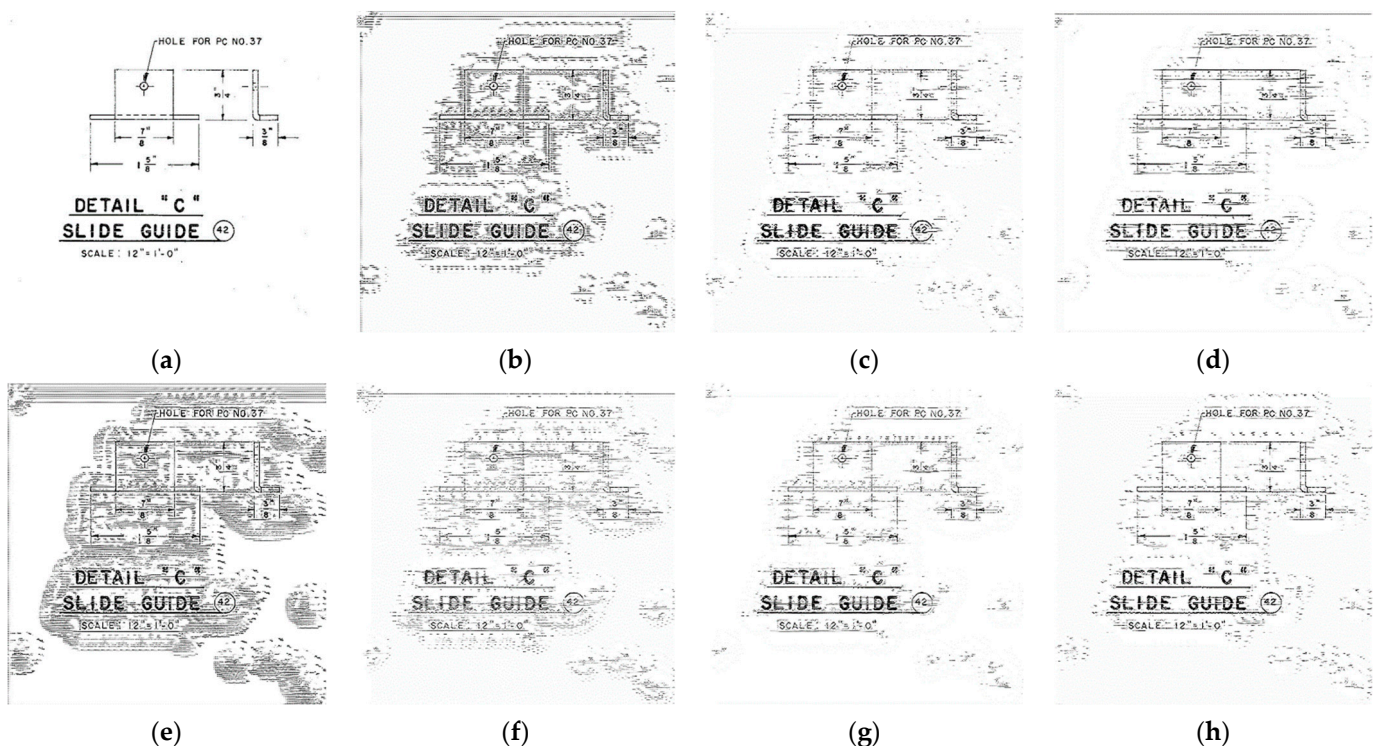
### 4.3. Implementation Details

We trained our model in two stages. We first trained the generators G and F for 800 epochs using a constant learning rate of 0.008 and the Adam optimizer. We trained the modified CycleGAN following the training steps mentioned in the original CycleGAN paper [16]. After this training phase, we obtained multiple versions of G, which were subsequently used to generate multiple noisy versions of clean images. In the second stage, we trained the Pix2Pix GAN model with those generated image pairs for 500 epochs with a learning rate of 0.0002. During each training iteration, we used randomly cropped patches of size  $256 \times 256$ . We implemented our proposed model using the PyTorch framework [37] and performed the experiments on an Nvidia V100 GPU.

## 5. Results

### 5.1. Results of Noisy Data Generation

The trained generator G successfully created 80 distinct synthetic noisy images for each of the 50 clean images collected from the MSC document. Figure 5 shows seven noisy images generated for one clean image, where each noisy image has different noise characteristics.



**Figure 5.** Synthetic noisy images. (a) Clean image. (b–h) Generated noisy images.

### 5.2. Visual Inspection of Denoised Images

Figure 6 shows four denoised images by the proposed method. It is evident that the denoised images exhibit remarkable improvements in visual clarity and text legibility. The enhanced documents demonstrate a notable reduction in noise levels, resulting in sharper text and clearer visual elements compared to their noisy counterparts. These qualitative observations highlight the efficacy of the proposed approach in achieving high-quality denoising results.

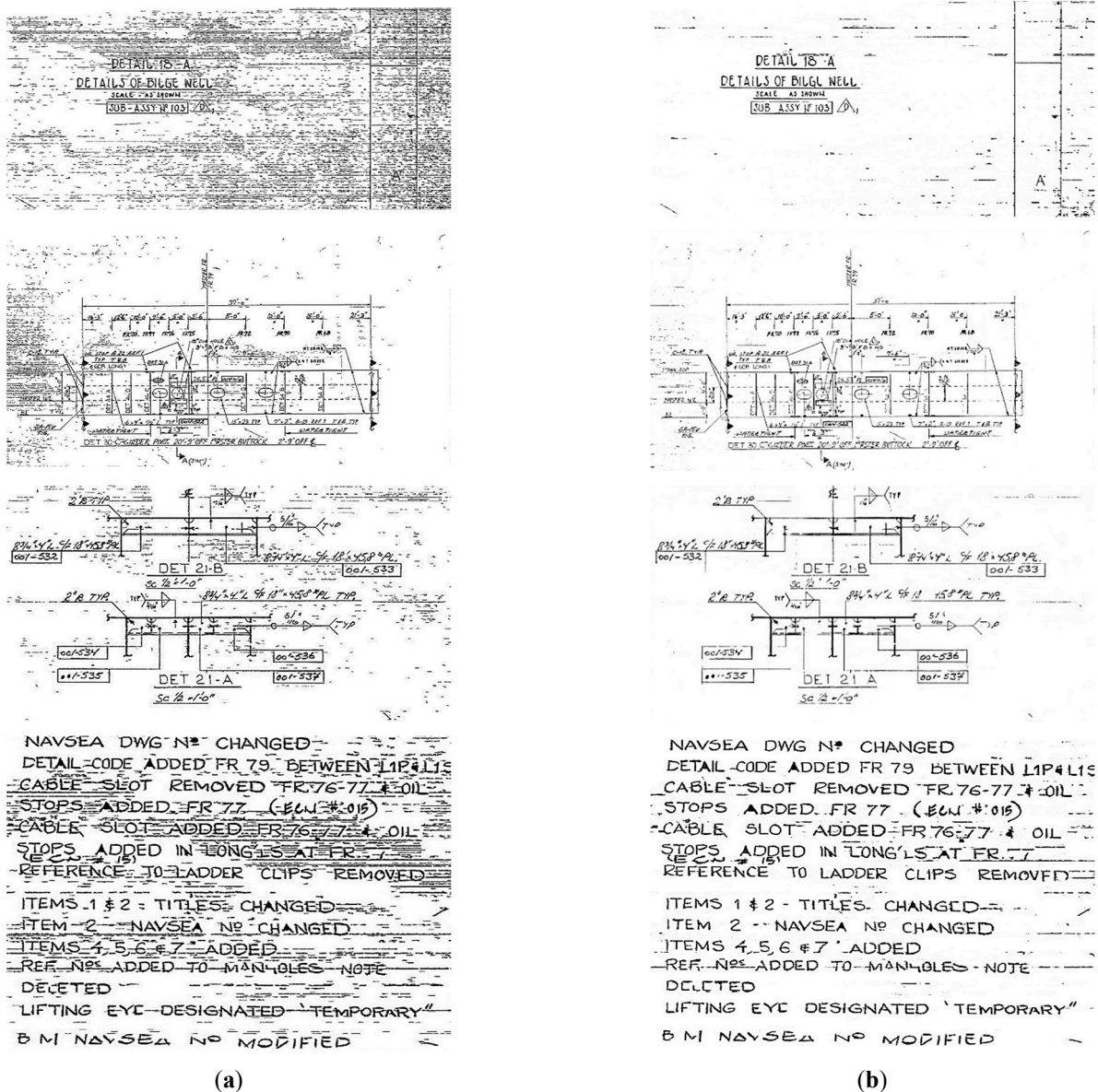
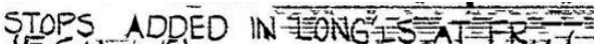
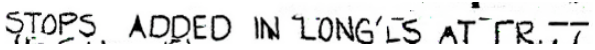




Figure 6. Denoised images by the proposed method. (a). Input noisy images. (b) Denoised images.

5.3. Results of Optical Character Recognition

We conducted OCR on two document images before and after denoising with our proposed method and Table 1 lists the CER and WER performance. It is evident that denoising significantly enhances the readability and interpretability of the text within documents. The noisy images exhibit high levels of noise, which introduce distortions and make character recognition challenging. However, after denoising, the clarity of text is noticeably improved, with characters becoming more distinguishable and coherent. Quantitatively, the denoised images consistently demonstrate lower CER and WER values compared to their noisy counterparts, indicating improved accuracy in character and word recognition tasks. Overall, these findings emphasize the significant benefits of denoising in improving the performance of OCR systems.

**Table 1.** Effects of denoising on CER and WER. Our proposed method can remove the noise, keeping the text-related information intact.

Noisy	Clean
 <p>Ground Truth: STOPS ADDED IN LONG'S AT FR 77 Prediction: STOPS ADDED IN LONG'S AT FR. CER: 12.9% WER: 42.9%</p>	 <p>Ground Truth: STOPS ADDED IN LONG'S AT FR 77 Prediction: STOPS ADDED IN LONG'S AT FR IT CER: 9.7% WER: 28.6%</p>
 <p>Ground Truth: LIFTING EYE DESIGNATED 'TEMPORARY' Prediction: LIFTING EYE DESIGNATED TEMPORARY" CER: 5.6% WER: 25.00%</p>	 <p>Ground Truth: LIFTING EYE DESIGNATED 'TEMPORARY' Prediction: LIFTING EYE DESIGNATED 'TEMPORARY' CER: 0.0% WER: 0.0%</p>

5.4. Results of Comparative Study

We applied the proposed method and other competing models to enhance document images and Figure 7 shows two example denoised images. Visual inspection of the different denoised images reveals that our proposed approach consistently produced denoised documents with sharper text and clearer visual elements compared to other models. In contrast, documents processed by methods such as CycleGAN, CycleGAN + Pix2Pix, Otsu, and Sauvola’s method show varying degrees of residual noise, blurriness, and distortion, leading to reduced readability and visual clarity. Table 2 lists the four performance metrics of the four competing methods on the testing dataset. Our model consistently outperformed all the competing methods. With the lowest PI score, our model ensures superior perceptual image quality, maintaining fidelity and natural appearance in denoised documents. Additionally, its lower NIQE score indicates enhanced image quality compared to alternative approaches. Furthermore, it achieved the lowest CER and WER values, signifying superior accuracy in character and word recognition tasks.

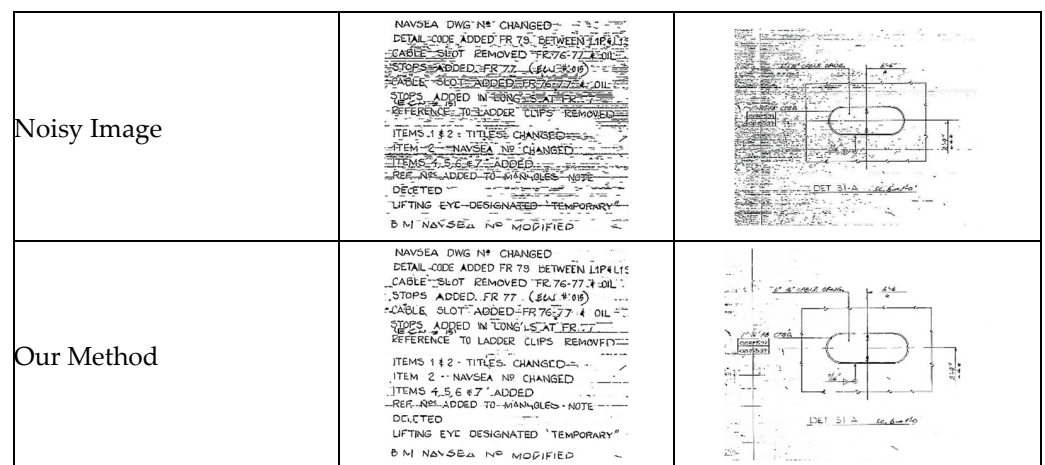


Figure 7. Cont.

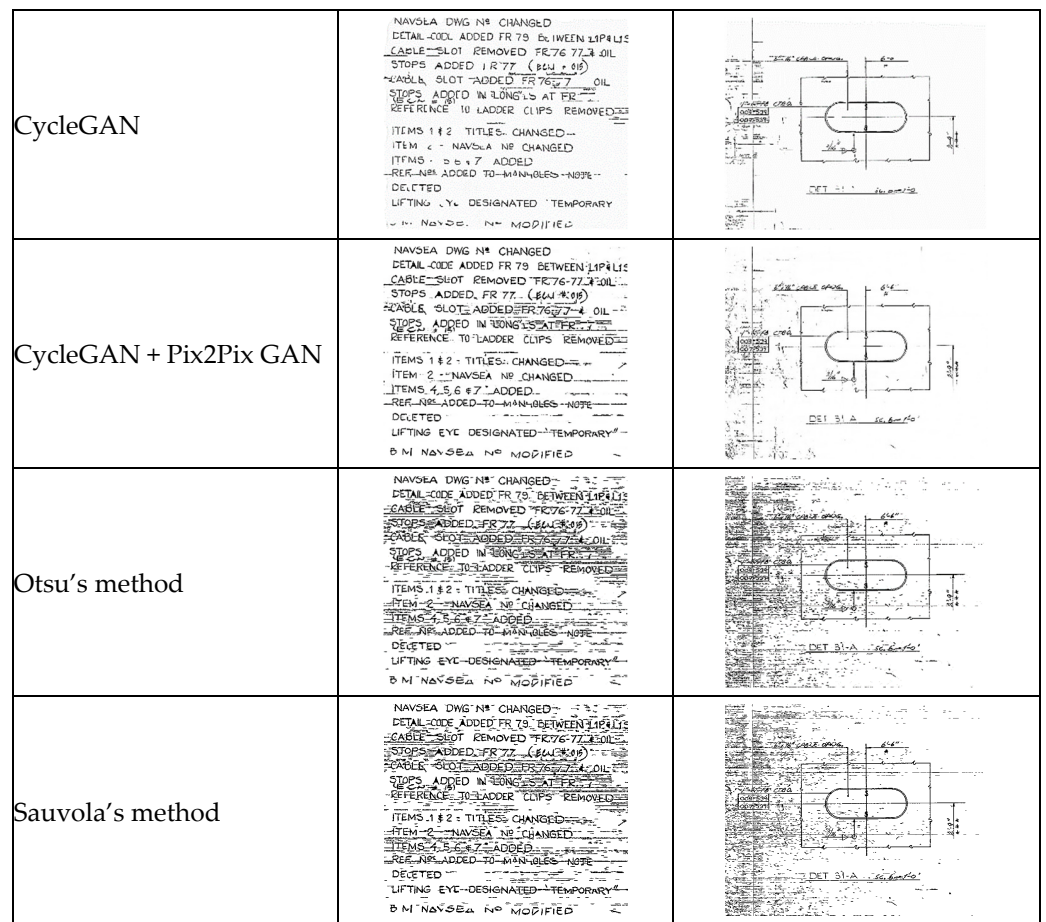


Figure 7. Denoising outputs from different methods. Our proposed method removed most of the noise without removing original structures and texts.

Table 2. Quantitative comparison.

	PI Score	NIQE	Ma Score	CER	WER
EnsembleDoc	6.92	10.77	6.93	0.1176	0.4327
CycleGAN	10.33	18.02	7.36	0.3492	0.7851
CycleGAN + Pix2Pix	7.49	11.64	6.66	0.1405	0.5506
Otsu's method	13.40	23.48	6.68	0.1673	0.5205
Sauvola's method	13.35	23.36	6.65	0.205	0.672
Noisy	8.10	13.01	6.81	0.1287	0.4789

### 5.5. Results of Ablation Study

Our proposed approach consists of several components including CycleGAN, two novel loss functions, Pix2Pix GAN, and ensemble data augmentation. In this ablation study, we investigated the contribution of each component in the ablation study and the results are listed in Table 3. The results showed that the ensemble augmentation is an important component and significantly improved the performance of the proposed model. Figure 8 shows some intermediate results obtained in the ablation study, and the ensemble data augmentation component improved the denoising results significantly.

Table 3. Performance of ablation study.

CycleGAN	Loss Functions	Pix2Pix GAN	Ensemble Augmentation	PI	NIQE	CER (%)	WER (%)
✓				10.33	18.02	34.92	78.51
✓		✓		7.49	11.64	14.05	55.06
✓	✓			8.44	13.85	27.66	57.41
✓	✓	✓		8.10	12.88	11.44	44.94
✓	✓	✓	✓	6.92	10.77	11.76	43.27

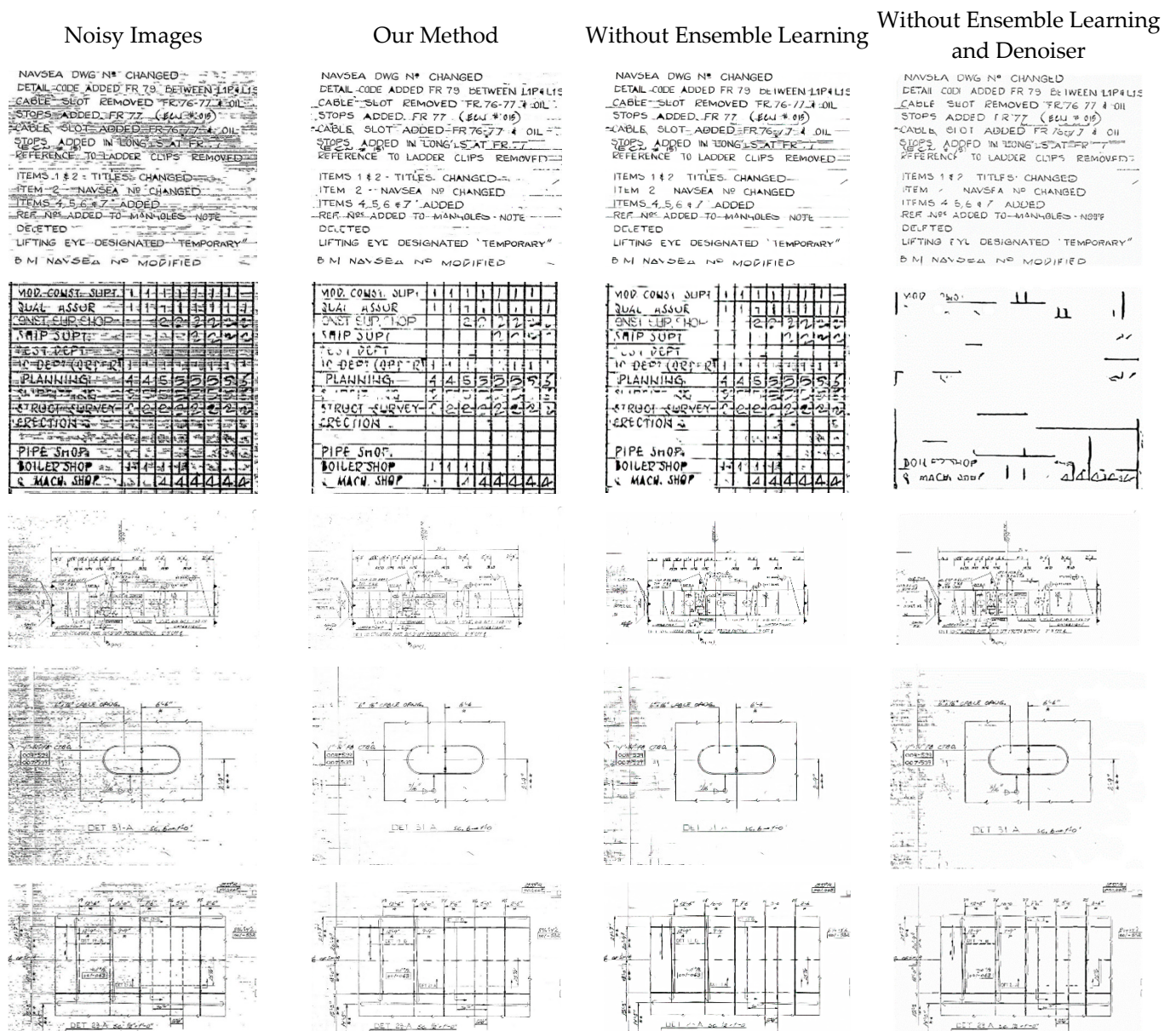


Figure 8. Results of ablation study of the proposed approach.

6. Discussion

The proposed approach combines the unpaired learning capabilities of CycleGAN with the paired learning strengths of Pix2Pix GAN using ensemble data augmentation. We compared the proposed model with a traditional document denoising algorithm, Otsu, a

generative approach, CycleGAN, and a hybrid method, CycleGAN + Pix2Pix GAN, for document noise removal. CycleGAN relies solely on small unpaired data and performed poorly, with a PI Score of 10.33, NIQE of 18.02, Ma Score of 7.36, CER of 0.3492, and WER of 0.7851. The hybrid method showed improvements in denoising performance with a PI Score of 7.49, NIQE of 11.64, Ma Score of 6.66, CER of 0.1405, and WER of 0.5506. It was still worse than the proposed method due to the limited pairs of images generated for denoising. Otsu failed to remove the scanning noise in our document images, resulting in a PI Score of 13.40, NIQE of 23.48, Ma Score of 6.68, CER of 0.1673, and WER of 0.5205. Our proposed method achieved the best performance except for the Ma score.

The ensemble data augmentation strategy in the proposed model harnessed the collective knowledge of multiple noise models, leading to more robust and reliable denoising outcomes. For example, the PI score reduced from 7.49 to 6.92, NIQE reduced from 11.64 to 10.77, CER reduced from 0.1405 to 0.1176, and WER reduced from 0.5506 to 0.4327, as shown in Table 2. In practice, obtaining paired clean and noisy images is typically challenging, especially in our application, where historical or rare documents are involved. The proposed ensemble data augmentation technique is an innovative way to generate large-sized paired datasets for learning.

The proposed loss function also improved the denoising performance over the CycleGAN model alone as shown in the ablation study results shown in Table 2. For example, PI improved from 10.33 to 8.44, NIQE from 18.02 to 13.85, CER from 34.92% to 27.66%, and WER from 78.51% to 57.41%, with the new loss function being added for training. Visual inspection also demonstrated crisper text and improved overall visual clarity with the loss function.

Our study has limitations. First, the proposed model still requires unpaired image data for training, which may not always be available in sufficient quantities. Additionally, while our novel loss functions and data generation strategies improve training efficiency, there is still room for enhancing the model's adaptability to extremely diverse noise patterns and document conditions. Future work will focus on further refining the model, exploring more sophisticated generative architectures, and developing techniques to minimize the dependency on unpaired datasets, aiming for more robust and generalizable document noise removal solutions.

## 7. Conclusions

We proposed a generative AI model to remove scanning noise in engineering documents. The proposed model consists of two steps. In the first step, the CycleGAN model was utilized to train a set of models to convert clean images to different versions of noisy images using a set of unpaired clean and noisy images for training. In the second step, the generated image pairs were used to train a Pix2Pix GAN for noise removal. Additionally, a new loss function was developed to increase the performance of the model. Experimental results on engineering documents collected by MSC demonstrated remarkable performance in effectively suppressing noise while preserving crucial document details. These findings highlighted the potential of our proposed method to significantly improve document processing tasks, making it a valuable tool for various applications requiring high-quality document images.

**Author Contributions:** Conceptualization, J.L., A.S.-P. and S.K.; methodology, M.S.U. and J.L.; software, M.S.U. and W.K.; validation, M.S.U., J.L. and W.K.; formal analysis, M.S.U., J.L. and W.K.; investigation, M.S.U., J.L., W.K. and S.K.; resources, S.K. and A.S.-P.; data curation, M.S.U. and J.L.; writing—original draft preparation, M.S.U.; writing—review and editing, M.S.U. and J.L.; visualization, M.S.U. and J.L.; supervision, A.S.-P., S.K. and J.L.; project administration, S.K. and J.L.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research is based upon work supported, in whole or in part, by the U.S. Navy's Military Sealift Command through CACI under sub-contract P000143798-3, project 500481-003.

**Data Availability Statement:** Data is publicly unavailable due to privacy restrictions.

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

## References

1. *AutoCAD*, version 2022; Autodesk: San Rafael, CA, USA, 2022.
2. Khallouli, W.; Pamie-George, R.; Kovacic, S.; Sousa-Poza, A.; Canan, M.; Li, J. Leveraging Transfer Learning and GAN Models for OCR from Engineering Documents. In Proceedings of the 2022 IEEE World AI IoT Congress (AIoT), Seattle, WA, USA, 6–9 June 2022; pp. 15–21. [\[CrossRef\]](#)
3. Uddin, M.S.; Pamie-George, R.; Wilkins, D.; Sousa-Poza, A.; Canan, M.; Kovacic, S.; Li, J. Ship Deck Segmentation in Engineering Document Using Generative Adversarial Networks. In Proceedings of the 2022 IEEE World AI IoT Congress (AIoT), Seattle, WA, USA, 6–9 June 2022; pp. 207–212. [\[CrossRef\]](#)
4. Sadri, N.; Desir, J.; Khallouli, W.; Uddin, M.S.; Kovacic, S.; Sousa-Poza, A.; Cannan, M.; Li, J. Image Enhancement for Improved OCR and Deck Segmentation in Shipbuilding Document Images. In Proceedings of the 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 26–29 October 2022; pp. 45–51.
5. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232.
6. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
7. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. SMC* **1979**, *9*, 62–66. [\[CrossRef\]](#)
8. Sauvola, J.; Pietikäinen, M. Adaptive document image binarization. *Pattern Recognit.* **2000**, *33*, 225–236. [\[CrossRef\]](#)
9. Annabestani, M.; Saadatmand-Tarzan, M. A new threshold selection method based on fuzzy expert systems for separating text from the background of document images. *Iran. J. Sci. Technol. Trans. Electr. Eng.* **2019**, *43*, 219–231. [\[CrossRef\]](#)
10. 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. [\[CrossRef\]](#)
11. Hedjam, R.; Cheriet, M.; Kalacska, M. Constrained energy maximization and self-referencing method for invisible ink detection from multispectral historical document images. In Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24–28 August 2014; pp. 3026–3031.
12. Xiong, W.; Jia, X.; Xu, J.; Xiong, Z.; Liu, M.; Wang, J. Historical document image binarization using background estimation and energy minimization. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3716–3721.
13. Afzal, M.Z.; Pastor-Pellicer, J.; Shafait, F.; Breuel, T.M.; Dengel, A.; Liwicki, M. Document image binarization using lstm: A sequence learning approach. In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, Gammarth, Tunisia, 22 August 2015; pp. 79–84.
14. Souibgui, M.A.; Biswas, S.; Jemni, S.K.; Kessentini, Y.; Fornés, A.; Lladós, J.; Pal, U. Docentr: An end-to-end document image enhancement transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montréal, QC, Canada, 21–25 August 2022; pp. 1699–1705.
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
16. Croitoru, F.A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10850–10869. [\[CrossRef\]](#) [\[PubMed\]](#)
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. [\[CrossRef\]](#) [\[PubMed\]](#)
18. 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. [\[CrossRef\]](#)
19. Dang, Q.-V.; Lee, G.-S. Document image binarization with stroke boundary feature guided network. *IEEE Access* **2021**, *9*, 36924–36936. [\[CrossRef\]](#)
20. Bhunia, A.K.; Bhunia, A.K.; Sain, A.; Roy, P.P. Improving document binarization via adversarial noise-texture augmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2721–2725.
21. Tamrin, M.O.; Ech-Cherif, M.E.-A.; Cheriet, M. A two-stage unsupervised deep learning framework for degradation removal in ancient documents. In *Pattern Recognition. ICPR International Workshops and Challenges, Proceedings of the ICPR International Workshops and Challenges, Virtual, 10–15 January 2021*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 292–303.
22. Sharma, M.; Verma, A.; Vig, L. Learning to clean: A GAN perspective. In *Computer Vision—ACCV 2018, Proceedings of the 14th Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 174–185. [\[CrossRef\]](#)

23. Dutta, B.; Root, K.; Ullmann, I.; Wagner, F.; Mayr, M.; Seuret, M.; Thies, M.; Stromer, D.; Christlein, V.; Schür, J.; et al. Deep learning for terahertz image denoising in nondestructive historical document analysis. *Sci. Rep.* **2022**, *12*, 22554. [[CrossRef](#)] [[PubMed](#)]
24. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
25. Yang, Z.; Liu, B.; Xiong, Y.; Yi, L.; Wu, G.; Tang, X.; Liu, Z.; Zhou, J.; Zhang, X. DocDiff: Document enhancement via residual diffusion models. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 2795–2806.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
28. Yan, Q.; Xu, Y.; Yang, X.; Nguyen, T.Q. Single image superresolution based on gradient profile sharpness. *IEEE Trans. Image Process.* **2015**, *24*, 3187–3202. [[PubMed](#)]
29. Liu, L.; Hua, Y.; Zhao, Q.; Huang, H.; Bovik, A.C. Blind image quality assessment by relative gradient statistics and adaboosting neural network. *Signal Process. Image Commun.* **2016**, *40*, 1–5. [[CrossRef](#)]
30. Zhu, J.; Wang, N. Image quality assessment by visual gradient similarity. *IEEE Trans. Image Process.* **2011**, *21*, 919–933. [[PubMed](#)]
31. Zhang, B.; Sander, P.V.; Bermak, A. Gradient magnitude similarity deviation on multiple scales for color image quality assessment. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 1253–1257.
32. Xue, W.; Zhang, L.; Mou, X.; Bovik, A.C. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.* **2013**, *23*, 684–695. [[CrossRef](#)] [[PubMed](#)]
33. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “completely blind” image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [[CrossRef](#)]
34. Ma, C.; Yang, C.-Y.; Yang, X.; Yang, M.-H. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.* **2017**, *158*, 1–16. [[CrossRef](#)]
35. Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R.E.; Zhu, C. Real-world single image super-resolution: A brief review. *Inf. Fusion* **2022**, *79*, 124–145. [[CrossRef](#)]
36. Blau, Y.; Mechrez, R.; Timofe, R.; Michaeli, T.; Zelnik-Manor, L. The 2018 PIRM challenge on perceptual image super-resolution. In *Computer Vision—ECCV 2018 Workshops, Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2019.
37. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.