MFE-GAN: Efficient GAN-based Framework for Document Image Enhancement and Binarization with Multi-scale Feature Extraction

Rui-Yang Ju [a], KokSheik Wong [b], Yanlin Jin [c] and Jen-Shiun Chiang [d,*]

[a]Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
[b]School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, Subang Jaya, 47500, Malaysia
[c]Department of Electrical and Computer Engineering, Rice University, 6100 Main St, Houston, 77005, Texas, USA
[d]Department of Electrical and Computer Engineering, Tamkang University, No.151, Yingzhuan Rd., Tamsui Dist., New Taipei City, 251301, Taiwan

ARTICLE INFO

Keywords:
Image Generation
Document Image Processing
Document Image Enhancement
Document Image Binarization
Generative Adversarial Networks
Haar Wavelet Transformation

ABSTRACT

Document image enhancement and binarization are commonly performed before document analysis and recognition tasks to improve the efficiency and accuracy of techniques such as optical character recognition (OCR). This is because directly recognizing text in degraded documents, particularly in color images, often obtains unsatisfactory results. Training independent generative adversarial networks (GANs) for each color channel can generate images where shadows and noise are effectively removed, which in turn facilitates efficient text information extraction. However, employing multiple GANs for different color channels requires long training and inference times. To reduce both training and inference times of models for document image enhancement and binarization, we propose MFE-GAN, an efficient GAN-based framework with multi-scale feature extraction (MFE), which incorporates Haar wavelet transformation (HWT) and normalization to process document images before feeding them into GANs for training. In addition, we present novel generators, discriminators, and loss functions to improve the model's performance, and conduct ablation studies to demonstrate their effectiveness. Experimental results on the Benchmark, Nabuco, and CMATERdb datasets show that the proposed MFE-GAN significantly reduces both the total training and inference times while maintaining comparable performance in comparison to state-of-the-art methods. The implementation of this work is available at `https://ruiyangju.github.io/MFE-GAN`.

1. Introduction

Document image enhancement and binarization play important roles in document analysis, significantly impacting subsequent stages of the recognition process and layout analysis [1]. For instance, color-degraded documents often suffer from various types of degradation, such as paper yellowing, text fading, and page bleeding [2, 3]. These degradations can seriously affect the accuracy of techniques such as optical character recognition (OCR) [4, 5] and document image understanding [6, 7].

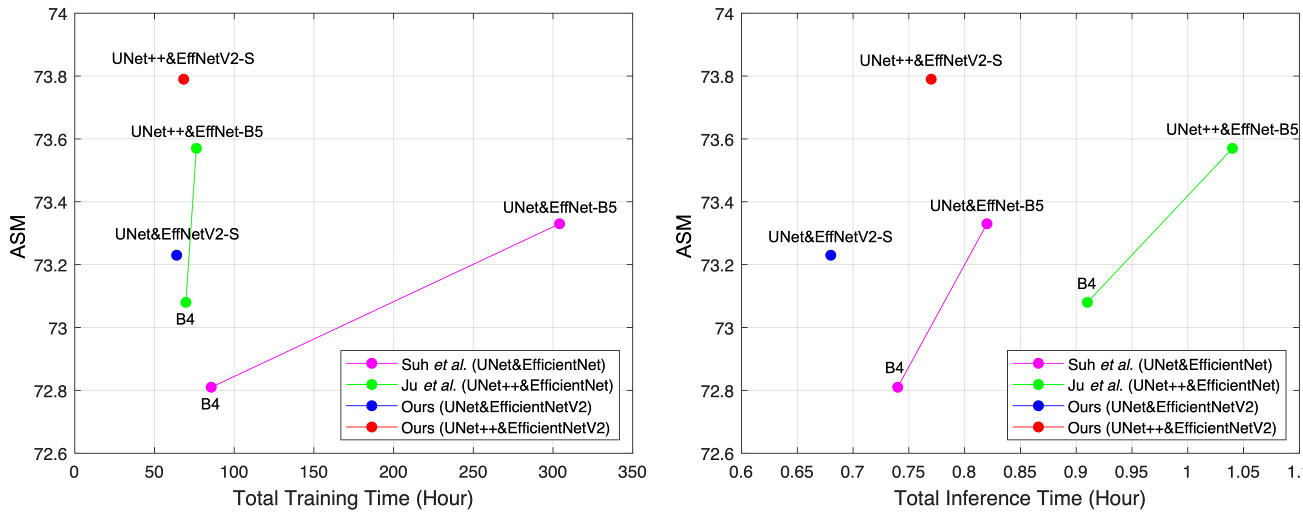
However, for color-degraded documents, traditional image processing methods [8, 9, 10] perform poorly in eliminating shadows and noise, sometimes even losing text information. Therefore, researchers have turned to deep learning-based methods, and many have achieved satisfactory results. For instance, Souibgui *et al.* [11] introduced a novel encoder-decoder architecture based on Vision Transformer (ViT), which achieved good performance on the H-DIBCO 2018 dataset [12]. Yang *et al.* [13] proposed an end-to-end gated convolutions-based network (GDB) to address the challenge of inaccurate stroke edge extraction in documents, and achieved state-of-the-art (SOTA) performance on the H-DIBCO 2014 and DIBCO 2017 datasets [14, 15]. For training and evaluation, these methods employ the "leave-one-out" strategy to construct the training set (viz., for the selected test set, all the remaining datasets are used to train the model). Considering the computing resources for model training, we believe that the strategy [16, 17, 18, 19, 20] of using fixed training and test sets as the Benchmark dataset is more efficient compared to the "leave-one-out" strategy.

Although the existing SOTA GAN-based methods [18, 20] achieve excellent performance on the Benchmark dataset, their total training and inference times are too long due to the use of six generative adversarial networks (GANs) [21]. As shown in Figure 1, these computational times are significantly high. To address this issue, we propose MFE-GAN, an efficient GAN-based framework that incorporates a novel multi-scale feature extraction (MFE) module, along with the generator, discriminator, and loss functions. Furthermore, we extend our previous conference version [22] by evaluating MFE-GAN on additional datasets.

The main contributions of this work are summarized as follows:

(a) Including both training and inference times as evaluation metrics, which are overlooked by previous methods.
(b) Discovering cases where PSNR does not always accurately reflect model performance, and introducing a new average score metric (ASM) for more comprehensive evaluation.
(c) Employing Haar wavelet transformation (HWT) with normalization for multi-scale feature extraction (MFE), effectively reducing training and inference time.

*Corresponding author: jsken.chiang@gmail.com
jryjry1094791442@gmail.com (Rui-Yang Ju); wong.koksheik@monash.edu (KokSheik Wong); neil.yl.jin@gmail.com (Yanlin Jin); jsken.chiang@gmail.com (Jen-Shiun Chiang)
ORCID(s): 0000-0003-2240-1377 (Rui-Yang Ju); 0000-0002-4893-2291 (KokSheik Wong); 0000-0001-8466-0660 (Yanlin Jin); 0000-0001-7536-8967 (Jen-Shiun Chiang)



Method	Generator	ASM↑	Total Training Time↓	Total Inference Time↓
Suh *et al.* [18]	U-Net & EfficientNet-B5	73.33	304.12h	0.82h
Ju *et al.* [20]	U-Net++ & EfficientNet-B5	73.57	76.29h	1.04h
MFE-GAN (Ours)	U-Net & EfficientNetV2-S	73.23	63.91h	0.68h
MFE-GAN (Ours)	U-Net++ & EfficientNetV2-S	73.79	68.43h	0.77h

Figure 1: Graphs (top) and table (bottom) compare the average-score metric (ASM) with respect to total training and inference times, measured on the Benchmark dataset using NVIDIA GeForce RTX 4090 GPUs. MFE-GAN, using U-Net & EfficientNetV2-S as the generator, trains 16%–79% faster than the compared methods, while inference time is reduced by 17%–35%.

(d) Outperforming SOTA GAN-based methods on three datasets in terms of model performance, training, and inference times through the incorporation of a novel MFE module, generator, discriminator, and loss functions.

The rest of this paper is organized as follows: Section 2 introduces the application of image generation networks in document image binarization, and reviews SOTA methods for color document image enhancement and binarization. Section 3 describes the proposed method, including the network architecture, multi-scale feature extraction, and loss functions. Section 4 analyzes the performance of the proposed method, quantitatively compares it with SOTA GAN-based methods on three datasets, and presents ablation studies to demonstrate the effectiveness of each component. Section 5 discusses the limitations of our method based on the analysis of visual results. Finally, Section 6 concludes the paper and highlights potential directions for future research.

2. Related Work

Document image binarization has advanced with the introduction of fully convolutional networks (FCNs) [23]. Tensmeyer *et al.* [24] formulated binarization as a pixel classification learning task and utilized FCNs for this purpose. Inspired by U-Net [25], Peng *et al.* [26] proposed a convolutional encoder–decoder model to perform binarization. He *et al.* [17] proposed DeepOtsu, which initially employed convolutional neural networks (CNNs) for document image enhancement and subsequently applied Otsu's method [8] for document image binarization. In addition, Zaragoza *et al.* [27] employed a selective autoencoder method to parse document images and subsequently binarize them using global thresholding.

The introduction of GANs [21] has enabled the generation of binarized document images. Zhao *et al.* [28] formulated binarization as an image-to-image translation task, employing conditional generative adversarial networks (cGANs) to address the challenge of combining multiscale information in binarization. Souibgui *et al.* [29] introduced an effective end-to-end framework based on cGANs (named the document enhancement generative adversarial network, DE-GAN) to restore degraded document images, achieving outstanding results on the DIBCO 2013, DIBCO 2017, and H-DIBCO 2018 datasets [30, 15, 12]. Deng *et al.* [31] proposed a method employing a dual discriminator generative adversarial network (DD-GAN) using focal loss as the generator loss function.

Recently, two SOTA methods employing multiple GANs for each color channel have been developed for document image enhancement and binarization tasks. Specifically, Suh *et al.* [18] proposed a two-stage GAN method using six improved CycleGANs [32] for color document image binarization. In Suh *et al.*'s method, the generator consists of U-Net [25] with EfficientNet [33], while the discriminator employs Pix2Pix GAN [34]. Ju *et al.* [20] introduced a three-stage GAN method based on the two-stage network architecture, employing six improved CycleGANs [32] with an enhanced generator using U-Net++ [35]. Although these

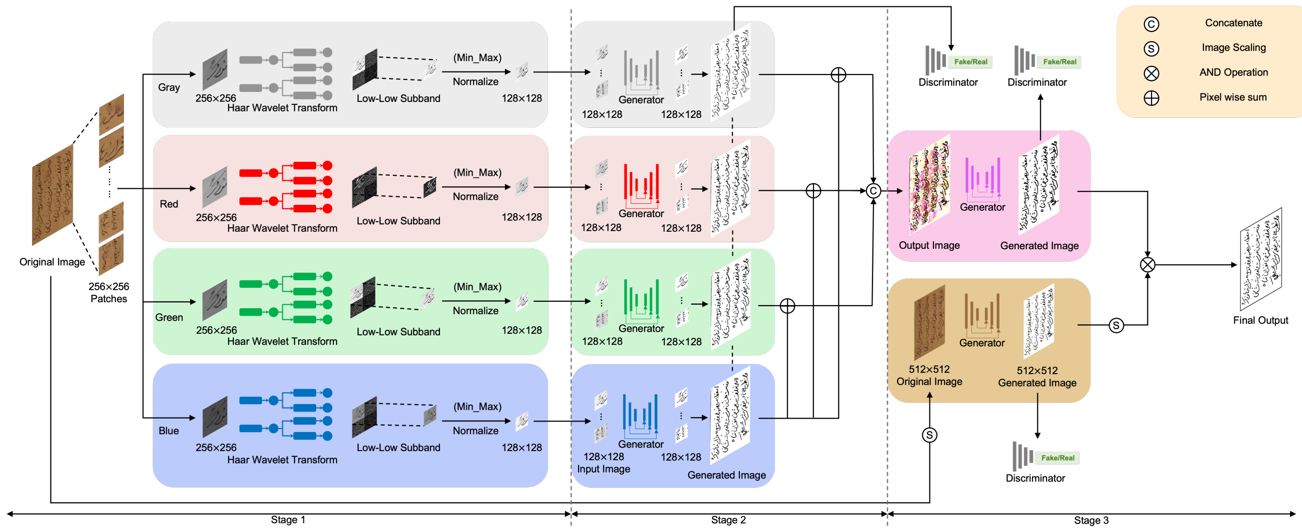


Figure 2: The overall architecture of three-stage GAN-based framework, named MFE-GAN. Stage 1: Document image processing; Stage 2: Document image enhancement; and Stage 3: Document image binarization.

methods consistently outperform other SOTA models on the DIBCO datasets, they suffer from unsatisfactory total training and inference times due to the use of multiple GANs.

3. Proposed Method

3.1. Network Architecture

We propose MFE-GAN, a novel three-stage GAN-based framework to perform document image processing, enhancement, and binarization, as shown in Figure 2.

In Stage 1, the original color document image is divided into several 256×256 pixel patches. Each patch is then split into four single-channel images (i.e., red, green, blue, and gray), because training on separate color channels tends to generate better results. For the MFE module, we apply HWT to each single-channel patch and extract the 128×128 pixel LL (low-low) sub-band. This sub-band is then normalized and serves as the input for Stage 2.

In Stage 2, MFE-GAN employs four independent generators using an encoder–decoder architecture based on U-Net++ [35] and an EfficientNetV2-S [36] backbone. Each 128×128 single-channel sub-band obtained from Stage 1 is fed into its corresponding generator, which outputs a 128×128 enhanced sub-image. As shown in Figure 2, the four enhanced sub-images are first combined using a pixel-wise summation and then concatenated to form the final output of Stage 2.

To standardize the generated outputs, a shared discriminator is employed for all independent generators. Specifically, we use the improved PatchGAN [32] as the discriminator, applying instance normalization to all layers except the first, because including instance normalization in the first layer would normalize the color information, which is not intended.

In Stage 3, multi-scale GANs are utilized for both local and global binarization to enhance the distinction between text and background. The output of Stage 2 is an image of the same size as the original input image and is fed into an independent generator that produces the local binarization output (B_{local}).

In addition, the original input image is scaled to 512×512 pixels using nearest-neighbor interpolation [37] and fed into another independent generator to produce the global binarization output (B_{global}). Each of these two branches (local and global) employs its own discriminator, forming two complete GANs. The final output (B_{final}) is obtained by an AND operation of the local and global binarization results ($B_{final} = B_{local} \otimes B_{global}$).

3.2. Multi-scale Feature Extraction

To reduce both total training and inference times, MFE-GAN employs its multi-scale feature extraction (MFE) module on 256×256 pixel patches in Stage 1. It is well known that reducing the input image size can significantly reduce the model training time, and decreasing the size of patches from 256×256 pixels to 128×128 pixels is consistent with this. However, directly reducing the image size using simple interpolation would negatively impact the model's performance. Instead of using interpolation for image size reduction, MFE-GAN employs Haar wavelet transformation (HWT) and normalization, which effectively preserve contour information and reduce noise interference while decreasing the image size. This approach is superior to interpolation methods that produce each output pixel based on its nearby pixels. We present the related experiments in Section 4.5.1, which demonstrate that the global binarization results of the images processed by HWT and normalization are closer to the ground-truth images than those processed by other interpolation methods.

During Stage 1 document image processing, HWT decomposes the input images into four sub-bands (LL, LH, HL, and HH). The low-frequency component (LL) encodes the contour information, and the high-frequency components (LH, HL, and HH) capture details and localized information. Therefore, we retain and normalize the low-low (LL) sub-band from HWT, effectively filtering out noise (which is often high-frequency) from color document images.

3.3. Loss Function

The convergence of the loss function is often unstable during the GANs training process [21]. To stabilize the loss function convergence in MFE-GAN, we we apply the objective function of the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) [38]. Since the goal of document image binarization is to classify each pixel into two categories (namely, text and background), we use binary cross-entropy (BCE) loss instead of the $L1$ loss used in the original method [34]. Experiments by Bartusiak *et al.* [39] also demonstrated that BCE loss outperforms $L1$ loss in binary classification tasks.

While BCE loss focuses on the accuracy of each individual pixel, Dice loss [40] emphasizes the accuracy of the entire region. Galdran *et al.* [41] demonstrated that combining BCE and Dice loss functions enhances segmentation performance at both the pixel-wise and region-wise levels. Since better segmentation performance at the regional level contributes to the greater completeness of the generated text, we use an improved WGAN-GP objective loss function, which includes both BCE loss $\mathbb{L}_{\text{BCE}}$ and Soft Dice loss $\mathbb{L}_{\text{Soft-DICE}}$ [42], expressed as follows:

$$\begin{aligned}\mathbb{L}_{\text{G}}(x, y; \theta_G) = &- \mathbb{E}_x[D(G(x), x)] + \lambda_1 \, \mathbb{L}_{\text{BCE}}(G(x), y) \\ &+ \lambda_2 \, \mathbb{L}_{\text{Soft-DICE}}(G(x), y),\end{aligned} \tag{1}$$

$$\begin{aligned}\mathbb{L}_{\text{D}} = &- \mathbb{E}_{x,y}[D(y, x)] + \mathbb{E}_x[D(G(x), x)] \\ &+ \alpha \mathbb{E}_{x, \hat{y} \sim P_{\hat{y}}}[(\| \nabla_{\hat{y}} D(\hat{y}, x) \|_2 - 1)^2],\end{aligned} \tag{2}$$

where x is the input image, $G(x)$ is the generated image, and y is the ground-truth image. λ_1 and λ_2 control the relative importance of different loss terms, while α denotes the gradient penalty coefficient. The discriminator D is trained for minimizing $\mathbb{L}_{\text{D}}$ to distinguish between ground-truth and generated images, while the generator G aims to minimize $\mathbb{L}_{\text{G}}$. The equations for BCE loss $\mathbb{L}_{\text{BCE}}$ and Soft Dice loss $\mathbb{L}_{\text{Soft-DICE}}$ are shown as follows:

$$\mathbb{L}_{\text{BCE}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbb{E}_{\hat{\mathbf{y}}, \mathbf{y}}[\mathbf{y} \log \hat{\mathbf{y}} + (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}})], \tag{3}$$

$$\mathbb{L}_{\text{Soft-DICE}}(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \frac{2|\hat{\mathbf{y}} \cap \mathbf{y}|}{|\hat{\mathbf{y}}| + |\mathbf{y}|} = 1 - \frac{2\langle \hat{\mathbf{y}}, \mathbf{y} \rangle}{\langle \hat{\mathbf{y}}, \hat{\mathbf{y}} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle}, \tag{4}$$

where $\mathbf{y}$ is the ground-truth, and $\hat{\mathbf{y}}$ is the predicted image.

4. Experiments

4.1. Datasets

4.1.1. Benchmark Dataset

DIBCO (Document Image Binarisation Contest) provides ten competition datasets, including DIBCO 2009 [43], H-DIBCO 2010 [44], DIBCO 2011 [45], H-DIBCO 2012 [46], DIBCO 2013 [30], H-DIBCO 2014 [14], H-DIBCO 2016 [47], DIBCO 2017 [15], H-DIBCO 2018 [12], and DIBCO 2019 [48]. These datasets include both machine-printed and handwritten images in grayscale and color, totaling 136 images.

BD (Bickley Diary) [49] was generously donated to the Singapore Methodist Archives by Mr. Erin Bickley. This dataset contains seven diary images, where factors such as lighting variations and fold damage make text recognition particularly challenging.

PHIBD (Persian Heritage Image Binarization Dataset) [50] comprises 15 historical manuscript images sourced from Mr. Mirza Mohammad Kazemaini's old manuscript collection in Yazd, Iran. The manuscripts within the images are affected by various degrees of degradation, including bleed-through, fading, and blurring.

SMADI (Synchromedia Multispectral Ancient Document Images) [51] was captured using a CROMA CX MSI camera, producing eight images for each document, resulting in a total of 240 images of authentic documents. The ancient documents in these images were written in iron-bile ink and date from the 17th to 20th centuries.

To ensure a fair comparison between the proposed MFE-GAN and the SOTA GAN-based methods [18, 20], we adopt the same strategy as in [18, 20] to construct the training set, as detailed in Table 1. The training set comprises images from DIBCO 2009 (10 images); H-DIBCO 2010 (10 images); H-DIBCO 2012 (14 images); Bickley Diary (7 images); PHIBD (15 images); and SMADI (87 images). The testing set consists of images from DIBCO 2011 (16 images); DIBCO 2013 (16 images); H-DIBCO 2014 (10 images); H-DIBCO 2016 (10 images); DIBCO 2017 (20 images); H-DIBCO 2018 (10 images); and DIBCO 2019 (20 images). Examples from the Benchmark dataset are shown in Figure 3.

4.1.2. Nabuco Dataset

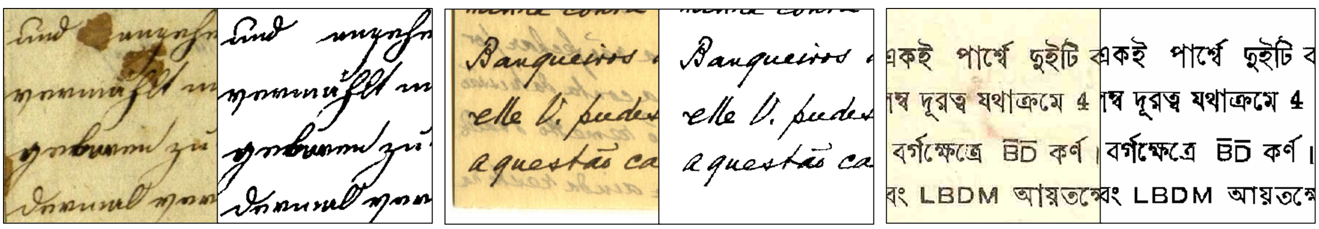
Nabuco [52] images were digitally compiled by Rafael Dueire Lins and historians from the Joaquim Nabuco Foundation between 1992 and 1994, using a true-color table scanner with a resolution of 200 dpi. The Nabuco bequest, consisting of 6,500 letters and postcards, both handwritten and typed, comprises approximately 30,000 pages. This bequest holds significant value for those studying the history of the Americas, as Joaquim Nabuco was one of the key figures in the abolition of slavery and the first Brazilian ambassador to the United States.

Due to the lack of corresponding ground-truth for most Nabuco images, we use only 35 images with available ground-truth, provided by the DIB team at CIn-UFPE, Brazil. Examples from these images are presented in Figure 3. This team offers two datasets, containing 15 and 20 color Nabuco images with their respective ground-truth. For evaluation, we perform a two-fold cross-validation procedure on the Nabuco dataset. The specific implementation details are shown in Table 1.

Table 1
Detailed usage statistics for the three datasets used in this work.

Dataset	Strategy	Training Set (Pages)			Test Set (Pages)	
		Original	Processed (256^2)	Processed (512^2)	Original	Processed (512^2)
Benchmark[1]	Following [18, 20]	143	120,174	804	102	582
Nabuco	Two-Fold Cross-Validation	15	32,038	90	20	120
		20	48,400	120	15	90
CMATERdb	Five-Fold Cross-Validation	4	5,308	24	1	6
		4	5,242	24	1	6
		4	4,942	24	1	6
		4	4,592	24	1	6
		4	2,140	24	1	6

[1] "Benchmark" refers to training sets containing DIBCO 2009, H-DIBCO 2010, H-DIBCO 2012, BD, PHIBD, and SMADI; and test sets containing DIBCO 2011, DIBCO 2013, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017, H-DIBCO 2018, and DIBCO 2019.



(a) Benchmark (b) Nabuco (c) CMATERdb

Figure 3: Examples from the three datasets used in this work: (a) Benchmark, (b) Nabuco, and (c) CMATERdb. Original images are shown on the left, and their corresponding binarized ground-truth on the right.

Table 2
Comparison of baseline and proposed model configurations.

Baseline [18]	
Generator	U-Net & EfficientNet-B5
Generator Loss	$-D(G(z)) + \lambda \mathbb{L}_{\text{BCE}}$
Discriminator	Similar Pix2pixGAN
Discriminator Loss	$D(x) - D(G(z)) + 10 \cdot \text{GP}$
MFE Module	✗

MFE-GAN	
Generator	U-Net++ & EfficientNetV2-S
Generator Loss	$-D(G(z)) + \lambda_1 \mathbb{L}_{\text{BCE}} + \lambda_2 \mathbb{L}_{\text{Soft-DICE}}$
Discriminator	Improved PatchGAN
Discriminator Loss	$D(x) - D(G(z)) + 10 \cdot \text{GP}$
MFE Module	✓

MFE: Multi-scale Feature Extraction; GP: Gradient Penalty; $\lambda = 50$, and $\lambda_1 = \lambda_2 = 25$.

4.1.3. CMATERdb Dataset

CMATERdb [53] is a dataset of Bengali and English manuscripts created by the Center for Microprocessor Applications for Training Education and Research (CMATER) at Jadavpur University, India. It comprises 5 images of color documents, including both camera-captured and scanned materials. These 5 images include a diverse range of document types, such as historical manuscripts and contemporary records, as well as degraded and well-preserved documents. Examples from this dataset are shown in Figure 3.

Since this dataset consists of only 5 images, each representing different conditions (i.e., blurred vs. clear, degraded vs. well-preserved), we perform a five-fold cross-validation procedure on the CMATERdb dataset. Specifically, we select four images for training and one for testing, the details are provided in Table 1.

4.2. Evaluation Metrics

For quantitative comparison, four classical metrics are employed, namely: f-measure (FM), pseudo-f-measure (p-FM), peak signal-to-noise ratio (PSNR), and distance reciprocal distortion (DRD). When comparing the performance of different methods, there are cases where the proposed MFE-GAN achieves SOTA-level FM and p-FM values, but its PSNR is lower than that of other methods. Inspired by Jemni *et al.* [54], we introduce the average-score metric (ASM) to evaluate the overall performance of each method more comprehensively:

$$ASM = \frac{FM + p\text{-}FM + PSNR + (100 - DRD)}{4}. \quad (5)$$

Note that in ASM, segmentation-quality metrics (FM, p-FM) are balanced against pixel-wise metrics (PSNR, DRD). We consider this reasonable, as it prevents a single metric, such as a low PSNR, from disproportionately penalizing an otherwise effective model. This is because, for methods

Table 3
PSNR (dB) of images resized using different methods: Interpolation/HWT/HWT&Normalization (Ours), for various datasets.

Method	DIBCO 2009	H-DIBCO 2010	H-DIBCO 2012	Bickley Diary	PHIBD	SMADI	Mean Values
Bicubic	71.45dB	72.22dB	71.67dB	64.29dB	69.58dB	69.88dB	69.85dB
Bilinear	70.94dB	72.16dB	71.46dB	64.07dB	69.71dB	69.86dB	69.70dB
Area	70.94dB	72.16dB	71.46dB	64.07dB	69.71dB	69.86dB	69.70dB
Nearest	70.95dB	72.04dB	71.59dB	64.20dB	69.69dB	69.83dB	69.72dB
Lanczos	71.42dB	72.22dB	71.69dB	64.30dB	69.58dB	69.89dB	69.85dB
HWT	62.65dB	67.11dB	59.67dB	53.76dB	58.00dB	59.48dB	60.11dB
Ours	71.77dB	72.74dB	72.85dB	64.44dB	70.76dB	69.44dB	70.34dB

Our method uses HWT&Normalization. The highest and second-highest PSNR values are highlighted in red and blue, respectively.

utilizing GANs to generate binarized images, the focus should be on the overall quality of the generated binarized image rather than on individual pixels. Furthermore, as we illustrate in Section 4.6, our proposed MFE-GAN can generate more complete images, although its PSNR is lower than that of the compared methods.

In addition, to demonstrate the efficiency of the proposed MFE-GAN relative to others when using the same computational resources, we calculate the total training and inference times for all models in hours (h). The total training time includes: the training time of models in Stage 2, the time required for Stage 2 models to generate all output images (Stage 2 Predict), the training time for the model using Stage 2's output images in Stage 3 (Stage 3 Top), and the training time for the model using resized original images (512×512) in Stage 3 (Stage 3 Bottom). The total inference time is defined as the time required to generate all images of the test set.

4.3. Baseline

We select the method [18], which has a similar network architecture, as the baseline. As shown in Table 2, the baseline method differs from MFE-GAN in terms of the generator and discriminator, their respective loss functions, and the multi-scale feature extraction (MFE) module.

4.4. Implementation Details

4.4.1. Data Preparation

To ensure a fair comparison, we employ the same dataset and data augmentation strategies for both MFE-GAN and the SOTA GAN-based methods [18, 20]. In Stage 1, the original input images are split into 256×256 pixel patches to match the input size of the ImageNet [55] dataset, as we utilize a pre-trained model based on this dataset. Data augmentation is applied to expand the training samples, with scaling factors of 0.75, 1, 1.25, and 1.5, as well as a rotation of 270°. For the Benchmark Dataset, this results in a total of 120,174 training image patches.

For global binarization in Stage 3, the input images are resized to 512×512 pixels. This set is further augmented through horizontal and vertical flipping, as well as rotations of 90°, 180°, and 270°, resulting in 804 training images of size 512×512 pixels for the Benchmark Dataset.

The Nabuco and CMATERdb datasets employ the same data augmentation strategies for their respective stages. The final number of processed training image patches and resized 512×512 pixel training images for these datasets are summarized in Table 1.

4.4.2. Pre-training and Training

All methods employ pre-trained weights from the ImageNet [55] dataset to improve training efficiency, due to constraints on data availability. Specifically, Suh *et al.* [18] and Ju *et al.* [20] use EfficientNet [33] as the encoder of their GANs, while MFE-GAN adopts EfficientNetV2 [36].

4.4.3. Training

To ensure a fair comparison of training and inference times, all models are trained on NVIDIA RTX 4090 GPUs using PyTorch as the implementation framework. The training parameters for Stage 2 and Stage 3 are largely similar, with the main exception being the number of epochs: 10 for Stage 2 and 150 for Stage 3. We choose the Adam optimizer to train the models and set the initial learning rate to 2×10^{-4}. In addition, the Adam optimizer parameters are set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for training both the generators and the discriminators.

4.5. Quantitative Comparisons

4.5.1. Multi-scale Feature Extraction

We explore other image-resizing techniques for multi-scale feature extraction, including interpolation-based algorithms such as bicubic, bilinear, area, nearest-neighbor, and Lanczos. We implement these techniques using the open-source computer vision library (OpenCV) to downscale all input images and the corresponding ground-truth images from 256×256 to 128×128. In addition, we employ the "HWT" and "HWT and normalization", i.e., our proposed MFE module. It should be noted that the resized images obtained from all these methods are non-binary.

To conduct a meaningful, apples-to-apples comparison against the binary ground-truth images, we must first binarize these intermediate outputs. Therefore, we apply a standard global thresholding algorithm (Otsu's method [8]) to these resized images, and then compute the PSNR values. We evaluate the impact of different image resizing techniques on six training sets by calculating the PSNR values (against the corresponding ground-truth images) and

Table 4
Training and inference time (hours) of the proposed and SOTA GAN-based methods on the Benchmark Dataset.

Method	Stage2 Train↓	Stage2 Predict↓	Stage3 Top↓	Stage3 Bottom↓	Training↓	Inference↓
U-Net & EfficientNet-B4 [18]	14.73h	3.75h	65.96h	1.17h	85.61h	0.74h
U-Net & EfficientNet-B5 [18]	16.30h	3.77h	282.80h	1.26h	304.12h	0.82h
U-Net++ & EfficientNet-B4 [20]	18.81h	3.95h	45.63h	1.29h	69.68h	0.91h
U-Net++ & EfficientNet-B5 [20]	21.23h	4.37h	49.23h	1.46h	76.29h	1.04h
U-Net & EfficientNetV2-S	11.60h	3.45h	47.47h	1.39h	63.91h	0.68h
U-Net++ & EfficientNetV2-S	14.12h	3.63h	49.29h	1.39h	68.43h	0.77h

The proposed MFE-GAN is highlighted in **bold**. The best and second-best performances are colored in red and blue, respectively.

Table 5
Quantitative comparison (ASM: FM/p-FM/PSNR/DRD, Total Training Time, Total Inference Time) of the proposed MFE-GAN and other methods for document image enhancement and binarization on the **Benchmark Dataset**.

Method	FM↑	p-FM↑	PSNR↑	DRD↓	ASM↑	Training↓	Inference↓
Otsu [8]	73.91	75.93	14.50dB	30.32	58.51	–	–
Sauvola [10]	75.83	80.72	15.62dB	9.65	65.63	–	–
U-Net & EfficientNet-B4 [18]	87.95	89.01	19.10dB	4.83	72.81	85.61h	0.74h
U-Net & EfficientNet-B5 [18]	88.56	89.90	19.31dB	4.46	73.33	304.12h	0.82h
U-Net++ & EfficientNet-B4 [20]	88.14	89.71	19.09dB	4.64	73.08	69.68h	0.91h
U-Net++ & EfficientNet-B5 [20]	89.13	90.35	19.30dB	4.49	73.57	76.29h	1.04h
U-Net & EfficientNetV2-S	88.83	89.87	19.07dB	4.86	73.23	63.91h	0.68h
U-Net++ & EfficientNetV2-S	89.69	90.78	19.15dB	4.45	73.79	68.43h	0.77h

The proposed MFE-GAN is highlighted in **bold**. The best and second-best performances are colored in red and blue, respectively.

computing the mean PSNR value for all images. The results are recorded in Table 3.

We observe that the mean PSNR value achieved by the "HWT" method is 60.11dB, indicating that images reduced directly using HWT have low similarity with the corresponding ground-truth images. In addition, the mean PSNR values for resized images produced by different interpolation methods are all below 70dB. However, the mean PSNR value for images processed by "HWT and normalization" reaches 70.34dB, which confirms that the images obtained by this method are closer to the corresponding ground-truth images at the pixel level. In conclusion, the results demonstrate that our "HWT and normalization" method is more effective than other interpolation-based image-resizing techniques for document image enhancement and binarization.

4.5.2. Benchmark Dataset

We compare MFE-GAN with the SOTA GAN-based methods [18, 20] for document image enhancement and binarization. As shown in Table 4, the compared methods using U-Net & EfficientNet-B5 or U-Net++ & EfficientNet-B5 [33] are already slower than our proposed MFE-GAN. Therefore, we do not further compare against methods using EfficientNet-B6, as this would further contradict our goal of reducing training and inference times.

Table 5 shows that MFE-GAN using U-Net++ [35] with EfficientNetV2-S [36] achieves the highest ASM of 73.79. It requires a total training time of 68.43h, which is also the second-shortest time. This is faster than the U-Net++ & EfficientNet-B5 method (76.29h), which yielded the second-highest ASM. Furthermore, the total inference time of MFE-GAN is 0.77h, notably lower than the 1.04h required by Ju *et al.*'s method [20] (using U-Net++ & EfficientNet-B5), representing a reduction of approximately 26%.

Moreover, MFE-GAN using U-Net & EfficientNetV2-S obtains the shortest total training and inference times (63.91h and 0.68h, respectively). Although the ASM value achieved by MFE-GAN of 73.23 is not the highest, when compared to Suh *et al.*'s method [18] (using U-Net with EfficientNet-B5) that yields 73.33 ASM, the training time is reduced from 304.12h to 63.91h, which is a remarkable decrease of approximately 78%. Overall, the experimental results demonstrate the efficiency and competitive performance of our proposed MFE-GAN.

Next, we compare the results achieved by all benchmark methods for each evaluation metric. MFE-GAN achieves the highest FM and p-FM values of 89.69 and 90.78, respectively, while maintaining lower total training and inference times than the method with the second highest FM and p-FM values. For the DRD metric, MFE-GAN achieves the second-highest value, but with a significantly reduced total training time of 68.43h compared to the 304.12h taken by the method with the highest DRD value. Although MFE-GAN does not achieve the highest PSNR, this metric does not directly reflect model performance in document image enhancement and binarization, as we discuss in detail in Section 4.6.

Table 6
Quantitative comparison (ASM: FM/p-FM/PSNR/DRD, Total Training Time, Total Inference Time) of the proposed MFE-GAN and SOTA GAN-based methods for document image enhancement and binarization on the **Nabuco dataset**.

Method **Generator**	Suh *et al.* [18] U-Net & B4	Suh *et al.* [18] U-Net & B5	Ju *et al.* [20] U-Net++ & B4	Ju *et al.* [20] U-Net++ & B5	MFE-GAN U-Net & V2-S	MFE-GAN U-Net++ & V2-S
FM↑	85.93	87.45	85.95	87.63	87.56	88.04
p-FM↑	86.57	88.16	86.37	88.27	88.22	88.72
PSNR↑	18.17dB	18.61dB	18.17dB	18.65dB	18.51dB	18.60dB
DRD↓	6.33	5.40	5.69	5.17	5.10	5.06
Stage2 Train Time↓	4.56h	5.71h	5.69h	6.62h	2.94h	3.02h
Stage2 Predict Time↓	2.04h	2.10h	2.22h	4.77h	2.19h	2.22h
Stage3 Top Time↓	20.36h	91.71h	22.20h	25.58h	18.30h	20.14h
Stage3 Bottom Time↓	0.51h	0.51h	0.52h	0.54h	0.55h	0.56h
ASM↑	71.08	72.21	71.20	72.34	72.30	72.58
Total Train Time↓	27.47h	100.03h	30.63h	37.51h	23.97h	25.93h
Total Inference Time↓	0.19h	0.22h	0.23h	0.26h	0.18h	0.21h

The best and second-best performances are highlighted in red and blue, respectively.

Table 7
Quantitative comparison (ASM: FM/p-FM/PSNR/DRD, Total Training Time, Total Inference Time) of the proposed MFE-GAN and other methods for document image enhancement and binarization on the **CMATERdb Dataset**.

Method **Generator**	Suh *et al.* [18] U-Net & B4	Suh *et al.* [18] U-Net & B5	Ju *et al.* [20] U-Net++ & B4	Ju *et al.* [20] U-Net++ & B5	MFE-GAN U-Net & V2-S	MFE-GAN U-Net++ & V2-S
FM↑	82.19	83.10	87.06	87.24	87.22	87.36
p-FM↑	88.17	89.44	91.49	92.31	91.66	92.46
PSNR↑	16.37dB	16.85dB	17.76dB	17.83dB	17.80dB	17.85dB
DRD↓	6.36	5.59	4.34	4.24	4.29	4.19
Stage2 Train Time↓	1543.30s	1878.08s	2023.23s	2276.39s	540.16s	609.36s
Stage2 Predict Time↓	455.98s	559.37s	541.80s	553.17s	554.10s	556.45s
Stage3 Top Time↓	7974.51s	42333.73s	6810.23s	7392.94s	7213.80s	7533.96s
Stage3 Bottom Time↓	593.68s	554.37s	573.99s	579.52s	585.98s	599.41s
ASM↑	70.09	70.95	72.99	73.28	73.10	73.37
Total Train Time↓	10567.47s	45325.55s	9949.25s	10802.02s	8894.04s	9299.18s
Total Inference Time↓	3.49s	3.87s	4.07s	4.65s	3.06s	3.42s

The best and second-best performances are highlighted in red and blue, respectively.

4.5.3. Nabuco Dataset

For the Nabuco dataset, we adopt a two-fold cross-validation strategy, where the final evaluation results are obtained by averaging the outcomes of both validations, as presented in Table 6.

MFE-GAN (U-Net++ & EfficientNetV2-S) achieves the highest FM and p-FM values of 88.04 and 88.72, respectively, as well as the lowest DRD value of 5.06. The highest PSNR of 18.65 dB is achieved by the model with U-Net++ & EfficientNet-B5 of Ju *et al.* [20], but MFE-GAN with U-Net++ & EfficientNetV2-S ranks second, achieving 18.60 dB. Notably, MFE-GAN also obtains the best ASM of 72.58, surpassing the second-best ASM of 72.34 achieved by the model of Ju *et al.* [20] with U-Net++ & EfficientNet-B5.

Furthermore, our two models rank first and second in terms of the shortest training time for Stage 2 and Stage 3 Top, respectively. The total training and inference times for MFE-GAN with U-Net & EfficientNetV2-S are the shortest, at 23.97h and 0.18h, respectively. Meanwhile, MFE-GAN with U-Net++ & EfficientNetV2-S achieves the second shortest total training time of 25.93h. In conclusion, the comparison outcomes on the Nabuco dataset are consistent with the results observed on the Benchmark dataset.

4.5.4. CMATERdb Dataset

To demonstrate the effectiveness of MFE-GAN in scenarios with limited data, we evaluate it on the CMATERdb dataset, which consists of only five representative images. We adopt a five-fold cross-validation strategy, indicating that the results in Table 7 are averaged over the five validation runs. Notably, due to the limited training and test data, we report the training and inference times in seconds (s).

In terms of the FM, p-FM, PSNR, DRD, and ASM evaluation metrics, MFE-GAN with U-Net++ & EfficientNetV2-S achieves the best performance across all metrics, while Ju *et al.*'s [20] model with U-Net++ & EfficientNet-B5 ranks second. Furthermore, our top-performing model (U-Net++

Method	Generator	FM↑	p-FM↑	PSNR↑
Blank Image	–	–	–	10.90dB
Suh *et al.* [18]	U-Net & EfficientNet-B4	26.09	19.90	11.39dB
Ju *et al.* [20]	U-Net & EfficientNet-B4	60.39	56.35	12.19dB
MFE-GAN (Ours)	U-Net & EfficientNetV2-S	69.44	69.88	11.75dB

Figure 4: Representative visualized results from the test set (case 1): the first row from left to right shows input image, ground-truth image, and blank image; the second row from left to right shows Suh *et al.* [18], Ju *et al.* [20], and MFE-GAN.

& EfficientNetV2-S) also achieves the second-shortest training and inference times.

In contrast, Ju *et al.*'s [20]'s model, despite ranking second in performance, is significantly slower due to its longer Stage 2 training time. Moreover, MFE-GAN with U-Net & EfficientNetV2-S still achieves the shortest training and inference times of 8,894s and 3.06s, respectively, while achieving an ASM value of 73.10.

4.6. Visual Results

We randomly select three images from the test set of Benchmark Dataset for visual examination and to demonstrate that the PSNR metric does not directly reflect model performance.

As shown in Figure 4, MFE-GAN generates more complete foreground information. However, due to the high contamination of the document image, some noise is inevitable when generating additional foreground content. In contrast, Suh *et al.*'s method [18] and Ju *et al.*'s method [20] generate less foreground content. It should be noted that because the background is white, PSNR favors methods that generate less content, which leads to higher PSNR values. To put things into context, a blank image (i.e., all pixels set to white) yields a PSNR of 10.90dB. This value is deceptively high and not far from MFE-GAN's score of 11.75dB. However, it is obvious that the binarized image generated by MFE-GAN is closer to the ground-truth image than the blank image.

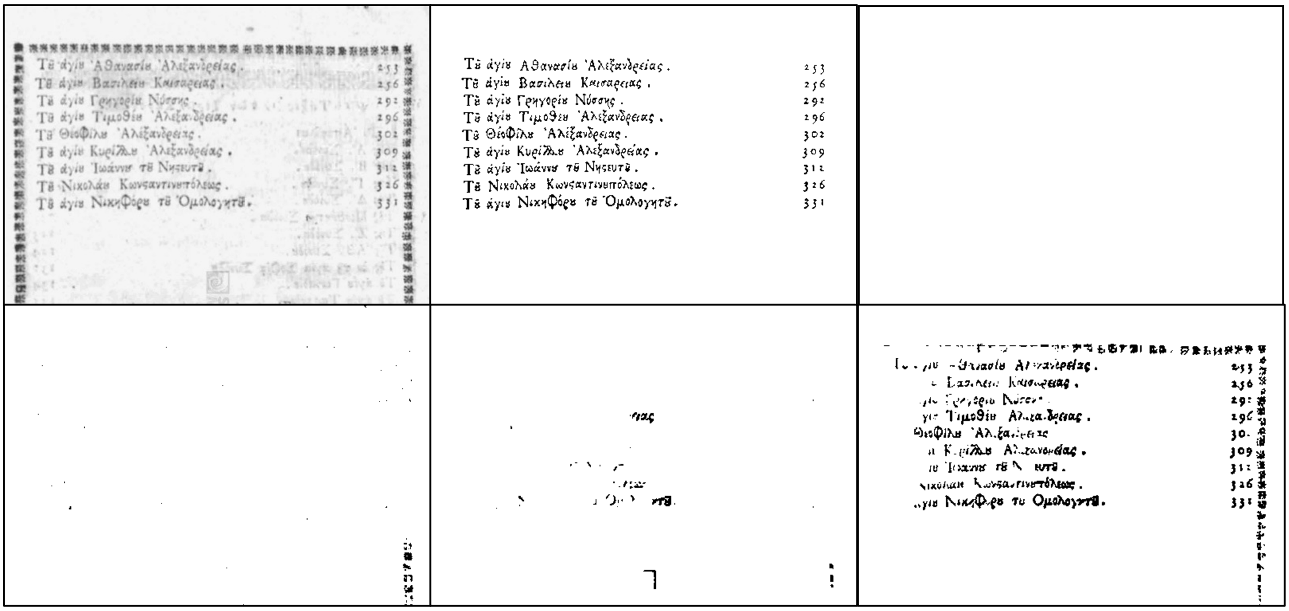
Figure 5 further confirms this conclusion. Here, a blank image yields a PSNR of 14.19dB, which is higher than that of MFE-GAN (14.08dB). However, it is obvious that our generated binarized image is closer to the ground-truth image than the blank one.

Figure 6 presents another case of a lower PSNR value, where MFE-GAN does not process background noise as effectively as the other two methods. However, MFE-GAN more faithfully generates the original text information compared to others. Furthermore, the PSNR metric is also unreliable in this case because the provided ground-truth itself is flawed, failing to capture the damaged edges of the page.

These observations support our claim that a higher PSNR value is not indicative of better model performance, and MFE-GAN can successfully generate more textual information.

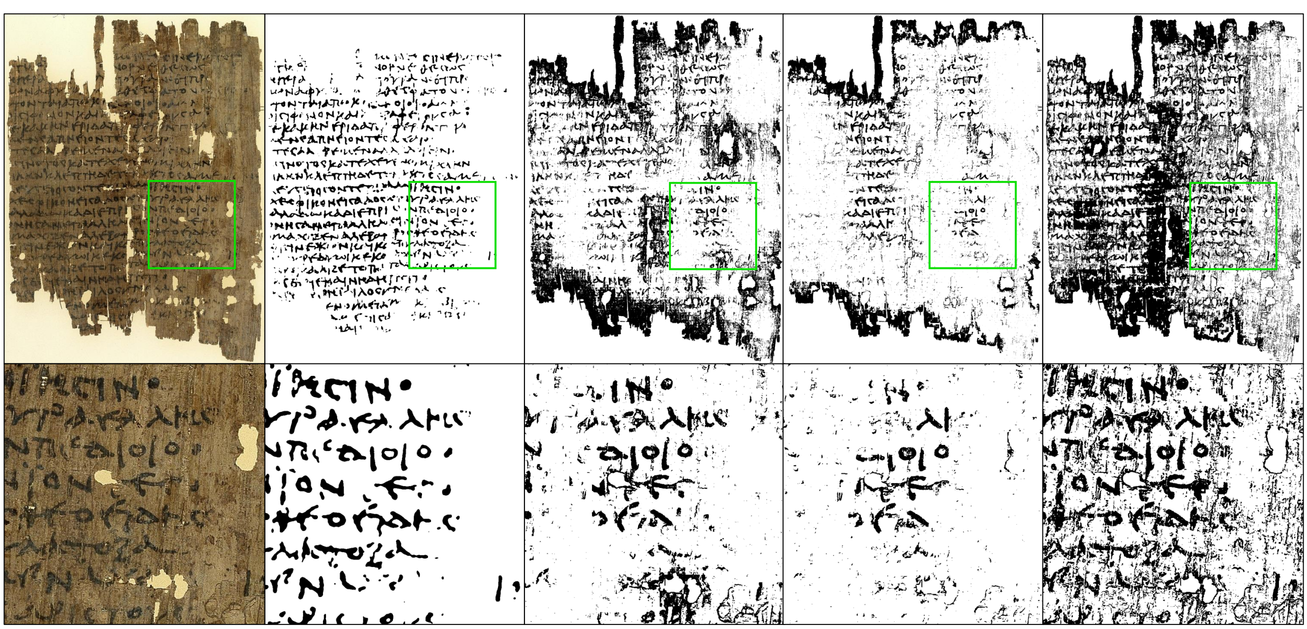
4.7. Ablation Study

To evaluate the contribution of each enhancement in MFE-GAN, we gradually replace or remove each component and observe the impact on performance. Table 8 summarizes the results under various configurations. To ensure a fair



Method	Generator	FM↑	p-FM↑	PSNR↑
Blank Image	–	–	–	14.19dB
Suh *et al.* [18]	U-Net & EfficientNet-B4	0.60	0.60	14.00dB
Ju *et al.* [20]	U-Net & EfficientNet-B4	10.05	9.32	14.23dB
MFE-GAN (Ours)	U-Net & EfficientNetV2-S	56.99	56.51	14.08dB

Figure 5: Representative visualized results from the test set (case 2): the first row from left to right shows input image, ground-truth image, and blank image; the second row from left to right shows Suh *et al.* [18], Ju *et al.* [20], and MFE-GAN.

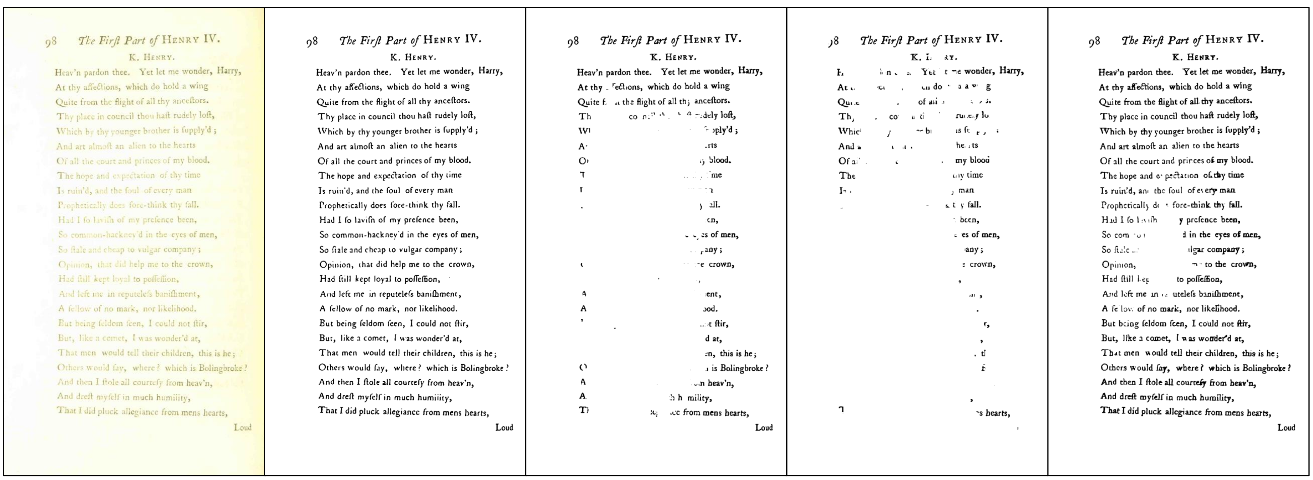


Method	Generator	FM↑	p-FM↑	PSNR↑
Suh *et al.* [18]	U-Net & EfficientNet-B4	38.46	38.68	6.75dB
Ju *et al.* [20]	U-Net & EfficientNet-B4	31.58	31.66	7.90dB
MFE-GAN (Ours)	U-Net & EfficientNetV2-S	45.49	45.58	6.14dB

Figure 6: Representative visualized results from the test set (case 3): from left to right shows input image, ground-truth image, Suh *et al.* [18], Ju *et al.* [20], and MFE-GAN.

Table 8
Ablation study of each component for the proposed MFE-GAN on the **Benchmark Dataset**.

Component					Performance		
Backbone: EfficientNetV2-S	**Generator:** U-Net++	**Discriminator:** InstanceNorm	**Generator Loss Function:** $D(G(z))+\lambda_1\mathbb{L}_{\text{BCE}}+\lambda_2\mathbb{L}_{\text{Soft-DICE}}$	**MFE Module:** HWT&Norm	**ASM**	**Total Training**	**Total Inference**
	✓	✓	✓	✓	73.79	112.74h	1.21h
✓		✓	✓	✓	73.23	63.91h	0.68h
✓	✓		✓	✓	73.45	61.24h	0.89h
✓	✓	✓		✓	73.58	70.52h	0.91h
✓	✓	✓	✓		73.81	523.86h	1.19h
✓	✓	✓	✓	✓	73.79	68.43h	0.77h



Input | Ground-Truth | Suh *et al.* | Ju *et al.* | Ours

Figure 7: Results for a sample affected by overexposure, leading to lower contrast. The images from left to right show the input image, ground-truth image, Suh *et al.* [18] (U-Net & EfficientNet-B5), Ju *et al.* [20] (U-Net++ & EfficientNet-B5), and MFE-GAN (U-Net & EfficientNetV2-S).

comparison, all experiments were conducted using the same dataset and hyperparameter settings.

Specifically, replacing U-Net [25] with U-Net++ [35] in the generator and adding instance normalization to the discriminator improve model performance, with a slight increase in training time. Replacing EfficientNet-B5 [33] with EfficientNetV2-S [36] in the generator reduces both training and inference times, while the new loss function further improves performance. Finally, employing the MFE module (i.e., HWT and normalization) for multi-scale feature extraction significantly reduces training time from 523.86h to 68.43h, representing an 87% decrease, with only a 0.02 decrease in ASM.

Overall, each improvement objectively contributes to either performance gains or reductions in training and inference times.

5. Discussion

We select a low-contrast sample with overexposure from the Benchmark dataset to showcase the visual results of different methods, and to discuss the limitations of MFE-GAN. As shown in Figure 7, the input image is a yellowish manuscript, with the light source on the left side being excessively bright during capture. This results in a very low contrast between the text and background, making effective image binarization challenging.

Compared to Suh *et al.*'s [18] and Ju *et al.*'s [20] methods, MFE-GAN generates more detailed text. However, our final results remain unsatisfactory, with text still missing in the central region. We consider that, when the contrast between the text and background is too low, GANs trained independently on different color channels struggle to effectively distinguish the text from the background, which in turn affects the generation results.

Therefore, for documents affected by light pollution, we suggest that applying contrast enhancement techniques based on both global and local features could help GANs distinguish the text from the background more effectively. In addition, incorporating pre-processing methods based on exposure compensation may mitigate the impact of this issue. Moreover, we suspect that MFE-GAN may have limited generalization performance under extreme conditions, such as overexposure and low-contrast scenes, as the Benchmark Dataset contains few light-polluted samples. To address this, we suggest employing a data augmentation strategy to increase the proportion of such samples under extreme conditions in the training set and thereby enhance the model's generalization ability.

6. Conclusion and Future Work

Degraded color document image enhancement and binarization are important steps in document analysis. Current SOTA GAN-based methods can generate satisfactory document binarization results but suffer from long training and inference times.

To address this drawback, we propose MFE-GAN, an efficient three-stage GAN-based framework that incorporates an MFE module (i.e., HWT and normalization) for multi-scale feature extraction, which significantly reduces training and inference times. Furthermore, we introduce novel generators, discriminators, and a new loss function to further improve the performance of our proposed MFE-GAN. Experimental results on benchmark datasets demonstrate that MFE-GAN not only achieves superior model performance but also significantly reduces the total training and inference times in comparison to SOTA GAN-based methods.

For future work, we plan to combine document image binarization and document image understanding for practical applications, especially for ancient documents or historical artifacts. Such applications could include real-time translation, summarization, and related document retrieval.

CRediT authorship contribution statement

Rui-Yang Ju : Conceptualization, Methodology, Software, Validation, Data Curation, Writing – original draft, Writing - Review & Editing. **KokSheik Wong :** Project administration, Visualization, Writing - Review & Editing. **Yanlin Jin :** Validation, Writing – review & editing. **Jen-Shiun Chiang :** Resources, Funding acquisition, Supervision, Writing – review & editing.

Declarations

This paper is an expanded version of the paper [22] presented at the 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), held from October 22 to 24, 2025, in Singapore.

Acknowledgment

This research is supported by National Science and Technology Council of Taiwan, under Grant Number: NSTC 114-2221-E-032-011-.

Declaration of competing interest

The authors have no financial or proprietary interests in any material discussed in this article.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors only use generative artificial intelligence (AI) and AI-assisted technologies to improve language.

References

[1] Z. Yang, B. Liu, Y. Xxiong, L. Yi, G. Wu, X. Tang, Z. Liu, J. Zhou, X. Zhang, Docdiff: Document enhancement via residual diffusion models, in: ACM International Conference on Multimedia (ACM MM), 2023, pp. 2795–2806.

[2] B. Sun, S. Li, X.-P. Zhang, J. Sun, Blind bleed-through removal for scanned historical document image with conditional random fields, IEEE Transactions on Image Processing 25 (12) (2016) 5702–5712.

[3] N. Kligler, S. Katz, A. Tal, Document enhancement using visibility detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2374–2382.

[4] C. Duan, P. Fu, S. Guo, Q. Jiang, X. Wei, Odm: A text-image further alignment pre-training approach for scene text detection and spotting, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15587–15597.

[5] Z. Yang, J. Tang, Z. Li, P. Wang, J. Wan, H. Zhong, X. Liu, M. Yang, P. Wang, S. Bai, et al., Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, in: IEEE International Conference on Computer Vision (ICCV), 2025, pp. 21744–21754.

[6] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, S. Park, Ocr-free document understanding transformer, in: European Conference on Computer Vision (ECCV), 2022, pp. 498–517.

[7] D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, X. Liu, Docllm: A layout-aware generative language model for multimodal document understanding, in: Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 8529–8548.

[8] N. Otsu, A threshold selection method from gray-level histograms, IEEE transactions on systems, man, and cybernetics 9 (1) (1979) 62–66.

[9] W. Niblack, An introduction to digital image processing, Strandberg Publishing Company, 1985.

[10] J. Sauvola, M. Pietikäinen, Adaptive document image binarization, Pattern recognition 33 (2) (2000) 225–236.

[11] M. A. Souibgui, S. Biswas, S. K. Jemni, Y. Kessentini, A. Fornés, J. Lladós, U. Pal, Docentr: An end-to-end document image enhancement transformer, in: International Conference on Pattern Recognition (ICPR), 2022, pp. 1699–1705.

[12] I. Pratikakis, K. Zagori, P. Kaddas, B. Gatos, Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018), in: International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 489–493.

[13] Z. Yang, B. Liu, Y. Xiong, G. Wu, Gdb: gated convolutions-based document binarization, Pattern Recognition 146 (2024) 109989.

[14] K. Ntirogiannis, B. Gatos, I. Pratikakis, Icfhr2014 competition on handwritten document image binarization (h-dibco 2014), in: International Conference on Frontiers in Handwriting Recognition (ICFHR), 2014, pp. 809–813.

[15] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, Icdar2017 competition on document image binarization (dibco 2017), in: International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, 2017, pp. 1395–1403.

[16] Q. N. Vo, S. H. Kim, H. J. Yang, G. Lee, Binarization of degraded document images based on hierarchical deep supervised network, Pattern Recognition 74 (2018) 568–586.

[17] S. He, L. Schomaker, Deepotsu: Document enhancement and binarization using iterative deep learning, Pattern recognition 91 (2019) 379–390.

[18] S. Suh, J. Kim, P. Lukowicz, Y. O. Lee, Two-stage generative adversarial networks for binarization of color document images, Pattern Recognition 130 (2022) 108810.
[19] R.-Y. Ju, Y.-S. Lin, J.-S. Chiang, C.-C. Chen, W.-H. Chen, C.-T. Chien, Ccdwt-gan: Generative adversarial networks based on color channel using discrete wavelet transform for document image binarization, in: Pacific Rim International Conference on Artificial Intelligence (PRICAI), 2023, pp. 186–198.
[20] R.-Y. Ju, Y.-S. Lin, Y. Jin, C.-C. Chen, C.-T. Chien, J.-S. Chiang, Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks, Knowledge-Based Systems (2024) 112542.
[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.
[22] R.-Y. Ju, K. Wong, J.-S. Chiang, Efficient generative adversarial networks for color document image enhancement and binarization using multi-scale feature extraction, Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (2025).
[23] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
[24] C. Tensmeyer, T. Martinez, Document image binarization with fully convolutional neural networks, in: International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, 2017, pp. 99–104.
[25] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015, pp. 234–241.
[26] X. Peng, H. Cao, P. Natarajan, Using convolutional encoder-decoder for document image binarization, in: International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, 2017, pp. 708–713.
[27] J. Calvo-Zaragoza, A.-J. Gallego, A selectional auto-encoder approach for document image binarization, Pattern Recognition 86 (2019) 37–47.
[28] J. Zhao, C. Shi, F. Jia, Y. Wang, B. Xiao, Document image binarization with cascaded generators of conditional generative adversarial networks, Pattern Recognition 96 (2019) 106968.
[29] M. A. Souibgui, Y. Kessentini, De-gan: A conditional generative adversarial network for document enhancement, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (3) (2020) 1180–1191.
[30] I. Pratikakis, B. Gatos, K. Ntirogiannis, Icdar 2013 document image binarization contest (dibco 2013), in: International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1471–1476.
[31] R. De, A. Chakraborty, R. Sarkar, Document image binarization using dual discriminator generative adversarial networks, IEEE Signal Processing Letters 27 (2020) 1090–1094.
[32] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223–2232.
[33] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
[34] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134.
[35] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Transactions on Medical Imaging 39 (6) (2019) 1856–1867.
[36] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: International Conference on Machine Learning (ICML), 2021, pp. 10096–10106.
[37] B. Jähne, Digital image processing, Springer, 2005.
[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, Advances in Neural Information Processing Systems (NeurIPS) 30 (2017).
[39] E. R. Bartusiak, S. K. Yarlagadda, D. Güera, P. Bestagini, S. Tubaro, F. M. Zhu, E. J. Delp, Splicing detection and localization in satellite imagery using conditional gans, in: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 91–96.
[40] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: International Workshop on Deep Learning in Medical Image Analysis (DLMIA), 2017, pp. 240–248.
[41] A. Galdran, G. Carneiro, M. A. G. Ballester, On the optimal combination of cross-entropy and soft dice losses for lesion segmentation with out-of-distribution robustness, in: Diabetic Foot Ulcers Grand Challenge (DFUC), 2022, pp. 40–51.
[42] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: International Conference on 3D Vision (3DV), 2016, pp. 565–571.
[43] B. Gatos, K. Ntirogiannis, I. Pratikakis, Icdar 2009 document image binarization contest (dibco 2009), in: International Conference on Document Analysis and Recognition (ICDAR), 2009, pp. 1375–1382.
[44] I. Pratikakis, B. Gatos, K. Ntirogiannis, H-dibco 2010-handwritten document image binarization competition, in: International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2010, pp. 727–732.
[45] I. Pratikakis, B. Gatos, K. Ntirogiannis, Icdar 2011 document image binarization contest (dibco 2011), in: International Conference on Document Analysis and Recognition (ICDAR), 2011, pp. 1506–1510.
[46] I. Pratikakis, B. Gatos, K. Ntirogiannis, Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012), in: International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2012, pp. 817–822.
[47] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, Icfhr2016 handwritten document image binarization contest (h-dibco 2016), in: International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 619–623.
[48] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, I. Marthot-Santaniello, Icdar 2019 competition on document image binarization (dibco 2019), in: International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1547–1556.
[49] F. Deng, Z. Wu, Z. Lu, M. S. Brown, Binarizationshop: a user-assisted software suite for converting old documents to black-and-white, in: Joint Conference on Digital Libraries (JCDL), 2010, pp. 255–258.
[50] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, M. Cheriet, An efficient ground truthing tool for binarization of historical manuscripts, in: International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 807–811.
[51] R. Hedjam, M. Cheriet, Historical document image restoration using multispectral imaging system, Pattern Recognition 46 (8) (2013) 2297–2312.
[52] R. Lins, Nabuco—two decades of processing historical documents in latin america, Journal of Universal Computer Science 17 (2011) 151–161.
[53] A. F. Mollah, S. Basu, M. Nasipuri, Computationally efficient implementation of convolution-based locally adaptive binarization techniques, in: International Conference on Information Processing (ICInPro), 2012, pp. 159–168.
[54] S. K. Jemni, M. A. Souibgui, Y. Kessentini, A. Fornés, Enhance to read better: a multi-task adversarial network for handwritten document image enhancement, Pattern Recognition 123 (2022) 108370.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[56] S. Suh, H. Lee, P. Lukowicz, Y. O. Lee, Cegan: Classification enhancement generative adversarial networks for unraveling data imbalance problems, Neural Networks 133 (2021) 69–86.

Table 9
Training and inference times of the original models before (Baseline) and after applying the Multi-scale Feature Extraction (MFE) module (Ours).

Method	Stage 2 Train	Stage 2 Predict	Stage 3 Top	Stage 3 Bottom	Total Training	Total Inference
U-Net & EfficientNetV2-S (Baseline)	332.28h	3.56h	**47.47h**	1.63h	384.95h	1.12h
U-Net & EfficientNetV2-S (MFE-GAN)	**11.60h**	**3.45h**	**47.47h**	**1.39h**	**63.91h**	**0.68h**
U-Net++ & EfficientNetV2-S (Baseline)	465.28h	3.94h	52.88h	1.76h	523.86h	1.19h
U-Net++ & EfficientNetV2-S (MFE-GAN)	14.12h	3.63h	49.29h	**1.39h**	68.43h	0.77h

The best value in each group is highlighted in **bold**.

Table 10
Quantitative comparison of MFE-GAN using different architectures and backbone models to construct the generator.

Generator	FM↑	p-FM↑	PSNR↑	DRD↓	ASM↑	Total Training↓	Total Inference↓
U-Net & EfficientNetV2-S	88.83	**89.87**	**19.07dB**	4.86	**73.23**	**63.91h**	**0.68h**
U-Net & EfficientNet-B4	87.87	88.57	18.82dB	5.17	72.52	69.38h	0.75h
U-Net & EfficientNet-B5	**88.88**	89.65	18.93dB	**4.85**	73.15	81.63h	0.83h
U-Net++ & EfficientNetV2-S	89.69	**90.78**	**19.15dB**	**4.45**	**73.79**	**68.43h**	**0.77h**
U-Net++ & EfficientNet-B4	89.40	90.38	19.01dB	4.87	73.48	85.45h	0.91h
U-Net++ & EfficientNet-B5	**89.76**	90.75	**19.15dB**	4.51	**73.79**	112.74h	1.21h

The proposed MFE-GAN employs U-Net & EfficientNetV2-S and U-Net++ & EfficientNetV2-S as generators. The best value in each group is highlighted in **bold**.

A. Effect on the MFE Module

To demonstrate the effectiveness of the MFE module (i.e., applying HWT and normalization) in Stage 1, we evaluate two configurations of MFE-GAN: Model A: U-Net [25] with EfficientNetV2-S [36], and Model B: U-Net++ [35] with EfficientNetV2-S [36].

Table 9 summarizes the time taken for each stage, as well as the total training and inference times of the compared methods. Two configurations are compared: one with the application of the MFE module in Stage 1 (processing 128 × 128 sub-bands), and one without (i.e., the *baseline*, where the original 256 × 256 patches are directly fed into the GANs).

Here, the total training time refers to the sum of the durations of all stages, while the total inference time denotes the time required to generate images for all test sets. It can be seen that for both models, the total training time decreases when the MFE module is applied.

Specifically, the training time decreases from 384.95h to 63.91h for Model A and from 523.86h to 68.43h for Model B when the MFE module is applied. Similarly, the total inference time decreases from 1.12h to 0.68h for Model A and from 1.19h to 0.77h for Model B. These results demonstrate that incorporating the MFE module significantly reduces both training and inference times.

B. Effect on the Generator Architecture

One significant contribution of this work is the design of novel generators for the proposed three-stage GAN framework. This work aims to enable the trained model to generate more foreground text information using these novel generators. To demonstrate that our proposed generator backbone (EfficientNetV2 [36]) is superior to the original backbone (EfficientNet [33]) when used with U-Net [25] or U-Net++ [35], we conduct a series of experiments.

As shown in Table 10, we compare the model performance as well as the total training and inference times across different generator architectures. For the encoder of the generators, we utilize EfficientNet-B4, EfficientNet-B5, and EfficientNetV2-S. As shown in Table 10, for different GAN encoders, the proposed MFE-GAN achieves shorter total training and inference times than the original ones, while maintaining comparable or higher ASM values.

Specifically, the original method using U-Net++ [35] with EfficientNet-B5 [33] achieves an ASM value of 73.79, with training and inference times of 112.74h and 1.21h, respectively. In contrast, MFE-GAN obtains the same ASM value but with a total training time of 68.43h and a total inference time of 0.77h, which represents a decrease of 39% and 36%, respectively. These experimental results demonstrate that the proposed MFE-GAN can greatly reduce the training and inference times while maintaining, or even improving, model performance.

C. Effect on the Loss Functions

To validate that combining both BCE loss and Soft Dice loss in the generator's loss function can improve model performance, we conduct a comparative experiment. Specifically, we adopt the loss function $D(G(z)) + 0.5 \times \mathbb{L}_{\mathrm{BCE}}$, as used in [18, 20], as our baseline (see Table 2). In addition, we introduce two additional configurations: (1) replacing BCE

Table 11
Quantitative comparison of our method with different loss functions.

Generator Loss Function	FM↑	p-FM↑	PSNR↑	DRD↓	ASM↑	Total Train↓	Total Inference↓
$D(G(z)) + \lambda \mathbb{L}_{\text{BCE}}$	89.41	90.42	19.10dB	4.61	73.58	70.52h	0.91h
$D(G(z)) + \lambda \mathbb{L}_{\text{Soft-DICE}}$	88.81	90.00	19.09dB	**4.31**	73.40	71.55h	**0.75h**
$D(G(z)) + \lambda_1 \mathbb{L}_{\text{BCE}} + \lambda_2 \mathbb{L}_{\text{Soft-DICE}}$	**89.69**	**90.78**	**19.15dB**	4.45	**73.79**	**68.43h**	0.77h

We set $\lambda = 50$ and $\lambda_1 = \lambda_2 = 25$. The best value in each group is highlighted in **bold**.

Table 12
Quantitative comparison (ASM: FM/p-FM/PSNR/DRD) of the proposed MFE-GAN and other methods on each DIBCO dataset.

(a) DIBCO 2011

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	82.10	85.96	15.72dB	8.95
Sauvola [10]	82.14	87.70	15.65dB	8.50
U-Net & B4 [18]	89.38	90.44	19.71dB	3.25
U-Net & B5 [18]	89.64	91.24	19.76dB	3.02
U-Net++ & B4 [20]	89.02	89.96	19.67dB	3.02
U-Net++ & B5 [20]	91.89	93.58	19.73dB	2.95
U-Net & V2-S	92.47	93.14	19.77dB	2.81
U-Net++ & V2-S	92.83	93.50	19.92dB	2.58

(b) DIBCO 2013

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	80.04	83.43	16.63dB	10.98
Sauvola [10]	82.71	87.74	17.02dB	7.64
U-Net & B4 [18]	93.23	93.30	20.81dB	2.88
U-Net & B5 [18]	94.23	94.68	21.48dB	2.46
U-Net++ & B4 [20]	93.81	94.19	21.06dB	2.68
U-Net++ & B5 [20]	94.34	94.72	21.28dB	2.17
U-Net & V2-S	92.80	93.18	20.98dB	3.19
U-Net++ & V2-S	93.23	93.57	21.09dB	2.77

(c) H-DIBCO 2014

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	91.62	95.69	18.72dB	2.65
Sauvola [10]	84.70	87.88	17.81dB	4.77
U-Net & B4 [18]	96.19	96.71	21.58dB	1.15
U-Net & B5 [18]	96.37	96.90	21.78dB	1.09
U-Net++ & B4 [20]	95.96	96.33	21.31dB	1.22
U-Net++ & B5 [20]	96.38	96.96	21.85dB	1.08
U-Net & V2-S	96.28	96.77	21.78dB	1.13
U-Net++ & V2-S	96.36	97.72	21.91dB	1.08

(d) H-DIBCO 2016

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	86.59	89.92	17.79dB	5.58
Sauvola [10]	84.64	88.39	17.09dB	6.27
U-Net & B4 [18]	91.91	95.00	19.67dB	2.99
U-Net & B5 [18]	91.97	95.23	19.69dB	2.94
U-Net++ & B4 [20]	92.31	94.86	19.83dB	2.80
U-Net++ & B5 [20]	92.42	95.03	19.87dB	2.79
U-Net & V2-S	91.82	94.17	19.53dB	2.97
U-Net++ & V2-S	92.05	94.26	19.62dB	2.85

(e) DIBCO 2017

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	77.73	77.89	13.85dB	15.54
Sauvola [10]	77.11	84.10	14.25dB	8.85
U-Net & B4 [18]	89.65	90.42	17.76dB	3.82
U-Net & B5 [18]	89.89	90.97	17.95dB	3.61
U-Net++ & B4 [20]	87.58	88.39	17.58dB	4.67
U-Net++ & B5 [20]	88.72	89.81	17.97dB	3.77
U-Net & V2-S	88.43	89.36	17.60dB	4.04
U-Net++ & V2-S	89.34	90.14	17.64dB	3.80

(f) H-DIBCO 2018

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	51.45	53.05	9.74dB	59.07
Sauvola [10]	67.81	74.08	13.78dB	17.69
U-Net & B4 [18]	93.53	95.17	20.59dB	2.23
U-Net & B5 [18]	93.69	95.58	20.74dB	2.16
U-Net++ & B4 [20]	93.58	94.73	20.50dB	2.20
U-Net++ & B5 [20]	93.75	95.54	20.71dB	2.11
U-Net & V2-S	92.85	94.65	20.14dB	2.71
U-Net++ & V2-S	93.61	95.28	20.07dB	2.60

(g) DIBCO 2019

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	47.83	45.59	9.08dB	109.46
Sauvola [10]	51.73	55.15	13.72dB	13.83
U-Net & B4 [18]	61.76	62.00	13.58dB	17.46
U-Net & B5 [18]	64.11	64.74	13.77dB	15.97
U-Net++ & B4 [20]	64.75	69.49	13.65dB	15.87
U-Net++ & B5 [20]	66.42	66.80	13.70dB	16.53
U-Net & V2-S	67.17	67.83	13.70dB	17.17
U-Net++ & V2-S	70.41	70.96	13.79dB	15.49

(h) Mean Values

Method	FM↑	p-FM↑	PSNR↑	DRD↓
Otsu [8]	73.91	75.93	14.50dB	30.32
Sauvola [10]	75.83	80.72	15.62dB	9.65
U-Net & B4 [18]	87.95	89.01	19.10dB	4.83
U-Net & B5 [18]	88.56	89.90	19.31dB	4.46
U-Net++ & B4 [20]	88.14	89.71	19.09dB	4.64
U-Net++ & B5 [20]	89.13	90.35	19.30dB	4.49
U-Net & V2-S	88.83	89.87	19.07dB	4.86
U-Net++ & V2-S	89.69	90.78	19.15dB	4.45

The proposed MFE-GAN is highlighted in **bold**. The best and second-best performances are highlighted in red and blue, respectively. Since papers [18, 20] did not provide experimental results on all datasets for the configuration shown in the table, we have independently trained these models ourselves.

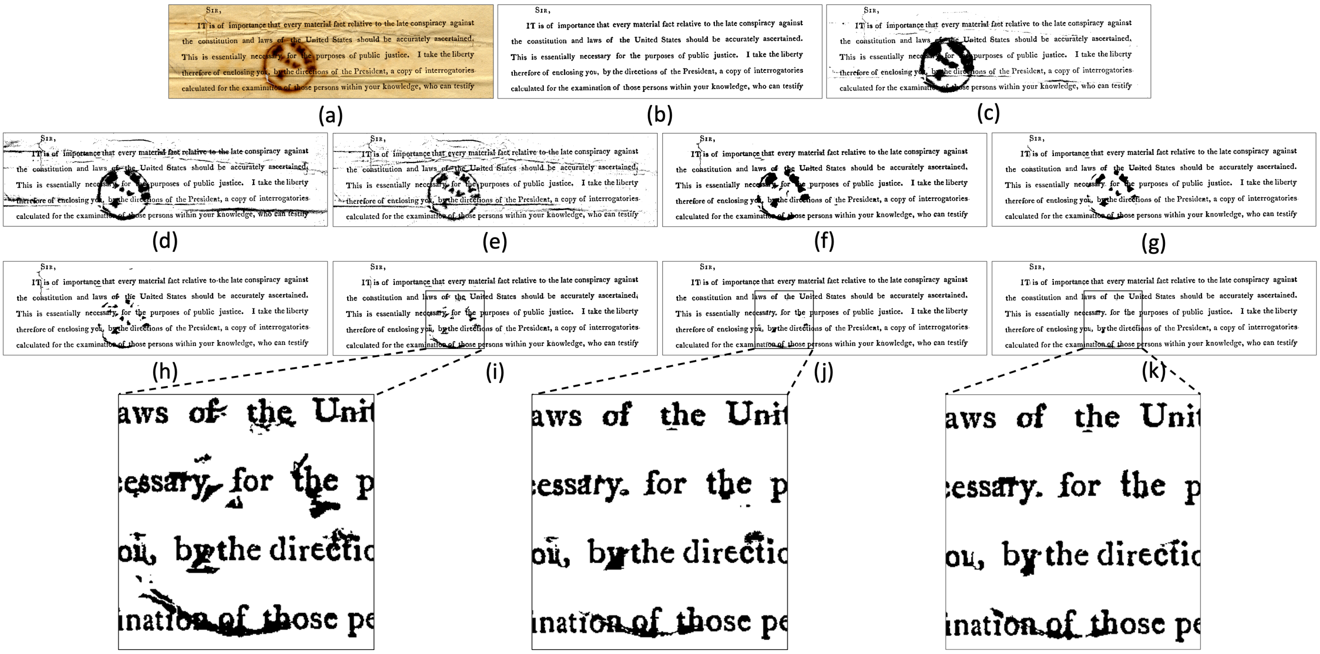


Figure 8: Qualitative comparison of binarization methods on a sample from the DIBCO 2013 dataset: (a) Input, (b) Ground-Truth, (c) Otsu [8], (d) Niblack [9], (e) Sauvola [10], (f) Vo [16], (g) He [17], (h) Zhao [28], (i) Suh [56], (j) Ju [20], (k) MFE-GAN.

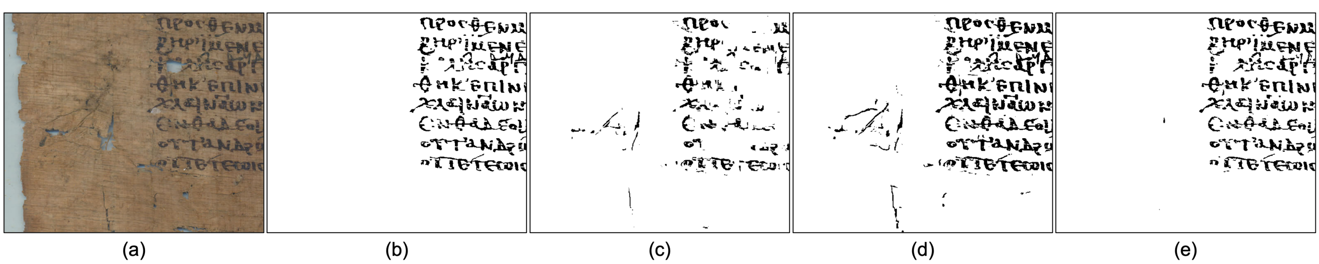


Figure 9: Qualitative comparison of binarization methods on a sample from the DIBCO 2019 dataset: (a) Input Image, (b) Ground-Truth, (c) Suh [18], (d) Ju [20], (e) MFE-GAN.

loss with Soft Dice loss: $D(G(z)) + 0.5 \times \mathbb{L}_{\text{Soft-DICE}}$; and (2) combining BCE loss and Soft Dice loss: $D(G(z)) + 0.25 \times \mathbb{L}_{\text{BCE}} + 0.25 \times \mathbb{L}_{\text{Soft-DICE}}$.

As presented in Table 11, the experimental results show that directly replacing BCE loss with Soft Dice loss leads to a decline in model performance. This demonstrates that BCE loss contributes positively to model performance in binary classification (text vs. background). In contrast, MFE-GAN, which combines both BCE loss and Soft Dice loss, achieves the best performance across FM, p-FM, and PSNR evaluation metrics. In addition, it outperforms the baseline model, achieving the shortest total training time and a reduced total inference time.

D. Results on Each DIBCO Dataset

Since the test set of the Benchmark dataset consists of several DIBCO datasets, the results for each DIBCO dataset are shown in Table 12. MFE-GAN (using U-Net++ & EfficientNetV2-S) achieves the best or second-best results on several datasets, including DIBCO 2011, H-DIBCO 2014, and DIBCO 2019, demonstrating its robustness to various types of document degradation. Overall, MFE-GAN achieves the highest mean FM and p-FM values (89.69% and 90.78%), along with a competitive DRD value of 4.45, outperforming the previous model (U-Net++ & EfficientNet-B5) proposed in [20].

E. Qualitative Comparison

Beyond the quantitative comparisons on different datasets, Figures 8 and 9 provide qualitative comparisons on the DIBCO 2013 and 2019 datasets between different methods. These figures demonstrate that SOTA GAN-based methods outperform traditional binarization methods in shadow and noise elimination. Furthermore, MFE-GAN excels in preserving textual content while effectively mitigating shadows and noise.