Efficient Generative Adversarial Networks for Color Document Image Enhancement and Binarization Using Multi-scale Feature Extraction

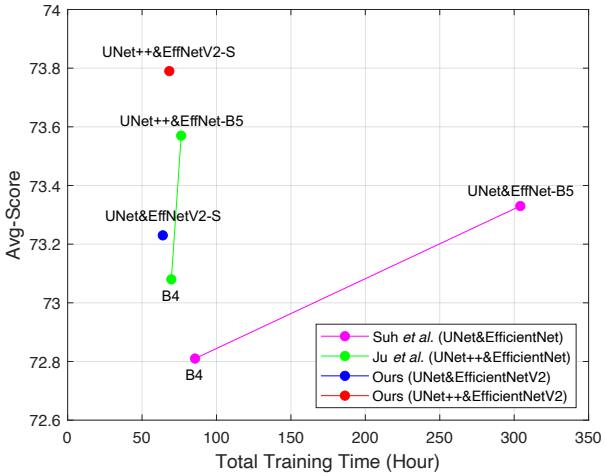
Rui-Yang Ju* and KokSheik Wong† and Jen-Shiun Chiang‡
* Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei City, Taiwan
† School of Information Technolog, Monash University Malaysia, Bandar Sunway, Malaysia
‡ Department of Electrical and Computer Engineering, Tamkang University, New Taipei City, Taiwan
E-mail: jryjry1094791442@gmail.com; wong.koksheik@monash.edu; jsken.chiang@gmail.com

***Abstract*—The outcome of text recognition for degraded color documents is often unsatisfactory due to interference from various contaminants. To extract information more efficiently for text recognition, document image enhancement and binarization are often employed as preliminary steps in document analysis. Training independent generative adversarial networks (GANs) for each color channel can generate images where shadows and noise are effectively removed, which subsequently allows for efficient text information extraction. However, employing multiple GANs for different color channels requires long training and inference times. To reduce both the training and inference times of these preliminary steps, we propose an efficient method based on multi-scale feature extraction, which incorporates Haar wavelet transformation and normalization to process document images before submitting them to GANs for training. Experiment results show that our proposed method significantly reduces both the training and inference times while maintaining comparable performances when benchmarked against the state-of-the-art methods. In the best case scenario, a reduction of 10% and 26% are observed for training and inference times, respectively, while maintaining the model performance at 73.79 of Average-Score metric. The implementation of this work is available at https://github.com/RuiyangJu/Efficient_Document_Image_Binarization.**

I. Introduction

Document image enhancement and binarization play important roles in document analysis, significantly impacting subsequent stages of the recognition process and layout analysis. For instance, color-degraded documents often suffer from various types of contaminants, such as paper yellowing, text fading, and page bleeding [1], [2]. These degradations seriously affect the accuracy of technologies such as Optical Character Recognition (OCR) and document image understanding.

Although existing state-of-the-art (SOTA) GAN-based methods [3], [4] have achieved excellent performance on benchmark datasets, they do not consider training and inference times, despite these two metrics are critical in practical applications. Our experiments reveal that these methods suffer from long training and inference times. To address this issue, we present an efficient method including novel generators, discriminators, and loss functions for document image binarization that significantly reduces both the training and inference times while maintaining the model performance, and the results are summarized in Fig. 1. Our contributions are as follows:



Method	Model	ASM↑	Train↓	Infer↓
Suh *et al.* [3]	UNet&EffNet-B5	73.33	304.12h	0.82h
Ju *et al.* [4]	UNet++&EffNet-B5	73.57	76.29h	1.04h
Ours	UNet&EffNetV2-S	73.23	63.91h	**0.68h**
Ours	UNet++&EffNetV2-S	**73.79**	**68.43h**	0.77h

Fig. 1. Graph of Avg-Score metric vs. Total Training Time (measured on (H)-DIBCO datasets using single NVIDIA GeForce RTX 4090 GPU).

(a) We are the first to introduce training and inference times as evaluation metrics, which have not been considered in existing SOTA GAN-based methods; (b) We adopt the Average-Score metric (ASM) to provide a more comprehensive assessment because we discover cases where PSNR value cannot correctly reflect a model's performance, and; (c) Our proposed method outperforms SOTA methods in terms of model performance (as measured by ASM), training, and inference times by designing novel generators, discriminators, and loss functions.

II. Related Work

Document image binarization has advanced with the introduction of fully convolutional networks (FCNs). Tensmeyer *et al.* [5] formulated binarization as a pixel classification learning task and utilized FCNs for this task. Inspired by

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

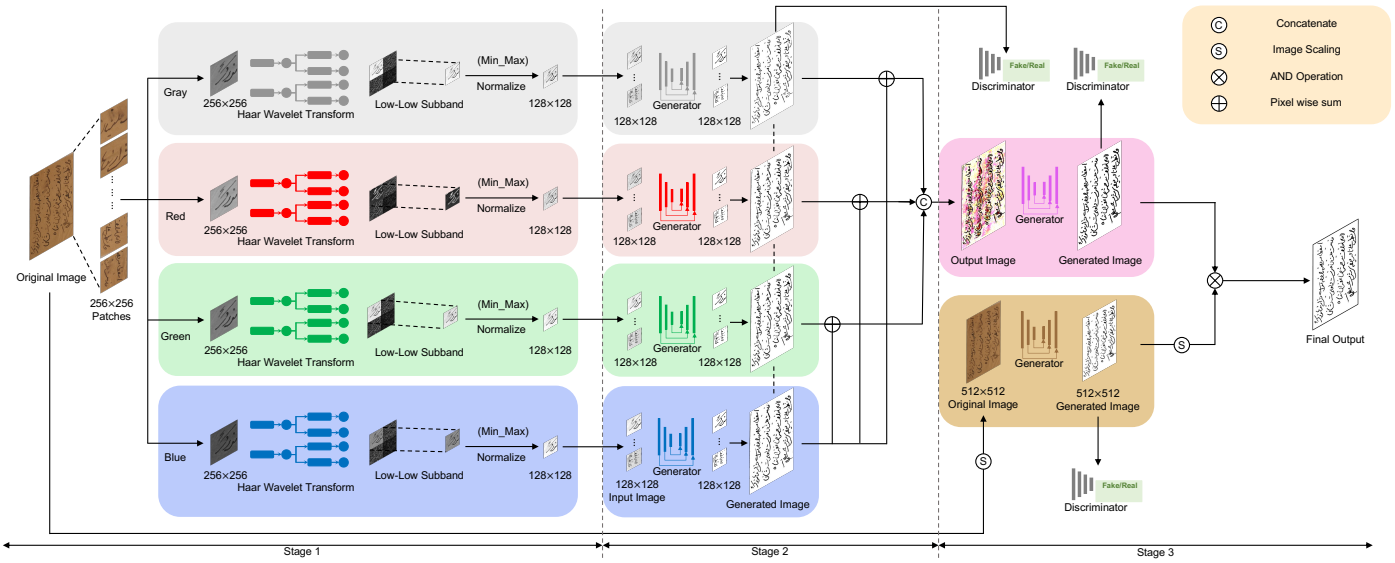


Fig. 2. The novel three-stage network architecture of the proposed method. Stage 1: document image processing, Stage 2: document image enhancement, and Stage 3: document image binarization.

UNet [6], Peng *et al.* [7] proposed a convolutional encoder-decoder model to perform binarization. He *et al.* [8] proposed DeepOtsu, which initially utilized convolutional neural networks (CNNs) for document image enhancement and subsequently applied Otsu's method for document image binarization. In addition, Zaragoza et al. [9] employed a selective autoencoder method to parse document images and subsequently binarizing them using global thresholding.

The introduction of GANs [10] has enabled the generation of binarized document images. Two SOTA methods based on GANs have recently been developed for document image enhancement and binarization tasks. Specifically, Suh *et al.* [3] proposed a novel two-stage GAN method using six improved CycleGANs [11] for color document image binarization. In Suh *et al.*'s method, the generator consists of UNet [6] with EfficientNet [12], while the discriminator employs Pix2Pix GAN [13]. Ju *et al.* [4] introduced a novel three-stage GAN method based on the two-stage network architecture and employed six improved CycleGANs [11], with an enhanced generator using UNet++ [14]. Although these methods consistently outperform SOTA models on DIBCO datasets, both of them overlook training and inference times, which is important in practical applications.

III. Proposed Method

A. Network Architecture

Fig. 2 presents the proposed efficient GAN method, which has a three-stage network architecture. In Stage 1, the original color document image is divided into non-overlapping patches. Each patch is then split into four single-channel images (i.e., red, green, blue, and gray), because training models on different color channels tend to generate better results. To reduce the training time, we apply HWT and normalization to resize the image patch size from 256 $\times$ 256 to 128 $\times$ 128.

In Stage 2, we design four generators with the encoder-decoder architecture, using UNet++ [14] with EfficientNetV2-S [15] as the backbone. Each single-channel image is fed into an independent generator for individual training. To standardize the generated outputs, all independent generators share the same discriminator. Specifically, we use the improved PatchGAN [11] as the discriminator, applying instance normalization to all layers except the first layer, because including instance normalization in the first layer would normalize the color information, which is not what we aimed for.

In Stage 3, multi-scale GANs are utilized for both local and global binarization to enhance the distinction between text and background. The input of Stage 3 (i.e., 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 output images of local binarization (B_l). In addition, the original input image is scaled to 512 $\times$ 512 pixels using Nearest Neighbour Interpolation and fed into an independent generator, and the output images of global binarization (B_g) are generated. The final output B is the pixel-wise summation of the local and global binarization results ($B = B_l + B_g$).

B. Loss Function

Since the convergence of the loss function is unstable during the GANs training process [10], to stabilize the loss function convergence of GANs in the proposed method, we apply Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) [16] to the objective function for the model training. 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 $L1$ loss employed in the original method [13]. In addition, Galdran *et al.* [17] demonstrated that combining BCE and Dice loss functions enhances segmentation performance at

both the pixel and regional levels. Since better segmentation performance at the regional level leads to generated text of greater completeness, we use the improved WGAN-GP objective loss function, which includes both BCE loss and Soft Dice loss [18] expressed below:

$$\mathbb{L}_G = -\mathbb{E}_x[D(G(x), x)] + \lambda_1 \mathbb{L}_{BCE}(G(x), y) + \lambda_2 \mathbb{L}_{SoftDice}(G(x), y); \quad (1)$$

$$\mathbb{L}_D = -\mathbb{E}_{x,y}[D(y, x)] + \mathbb{E}_x[D(G(x), x)] + \alpha \mathbb{E}_{x,\hat{y}\sim P_{\hat{y}}}[(\|\nabla_{\hat{y}} D(\hat{y}, x)\|_2 - 1)^2]. \quad (2)$$

Here, 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 $\mathbb{L}_D$ to distinguish between ground-truth and generated images, while the generator G aims to minimize $\mathbb{L}_G$.

IV. Experiments

A. Datasets and Evaluation Metrics

To ensure a fair comparison between the proposed method and the SOTA methods [3], [4], we adopt the same strategy as in [3], [4] to construct the training set. The training set comprises images from DIBCO 2009, H-DIBCO 2010, H-DIBCO 2012, Bickley Diary (BD), Persian Heritage Image Binarisation Dataset (PHIBD), and Synchromedia Multispectral Ancient Document Images (SMADI) [19]–[23]. The testing set consists of images from DIBCO 2011, DIBCO 2013, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017, H-DIBCO 2018, and DIBCO 2019 [24]–[26].

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 our model's FM and p-FM values reach the SOTA level, but our PSNR value is lower than that of other methods. Inspired by Jemni *et al.* [27], we adopt the Average-Score metric (ASM) to evaluate the overall performance of each method more comprehensively:

$$ASM := \frac{FM + p\text{–}FM + PSNR + (100 - DRD)}{4}. \quad (3)$$

Note that in ASM, precision and recall have a greater impact on the value than PSNR, which we consider reasonable. This is because, for methods utilizing GANs to generate binarized images, the focus should be on the overall correctness of the generated image rather than on individual pixels.

B. Implementation Details

To ensure a fair comparison of performances, we utilize the same dataset and data augmentation techniques for our proposed method and the SOTA methods [3], [4]. In Stage 1, the original input images are split into 256 × 256 patches, to match the size of the images from the ImageNet [28] dataset, considering that we will use the pre-trained model based on this dataset. Data augmentation is then employed to expand the training samples, with sampling scales set at 0.75, 1, 1.25, 1.5, and rotation by 270°, resulting in a total of 120,174 training image patches. For global binarization (Stage 3), the input images are directly resized to 512 × 512 and subjected to horizontal and vertical flipping, as well as rotation by 90°, 180°, and 270°, resulting in 804 training images.

To avoid the influence of hardware differences on model performance, all methods are trained using a single NVIDIA RTX4090 GPU. In addition, all methods are implemented in Python using PyTorch as the framework. The backbone networks in all methods use weights pre-trained on the ImageNet [28] dataset to enhance efficiency in model training. The training parameter settings are largely similar for Stage 2 and Stage 3, except for the number of training epochs, i.e., 10 epochs for Stage 2 and 150 epochs for Stage 3. We choose Adam optimizer to train models and set the initial learning rate to 2×10^{-4}. In addition, we configure generators with $\beta_1 = 0.5$ and discriminators with $\beta_2 = 0.999$.

TABLE I
Training and inference times taken by the SOTA method (Baseline) and after apply HWT and normalization (Ours).

Method	Stage2 Train	Stage2 Predict	Stage3 Top	Stage3 Bottom	Total Train	Total Infer
Baseline	332.28h	3.56h	47.47h	1.63h	384.95h	1.12h
Ours	11.60h	3.45h	47.47h	1.39h	63.91h	0.68h
Baseline	465.28h	3.94h	52.88h	1.76h	523.86h	1.19h
Ours	14.12h	3.63h	49.29h	1.39h	68.43h	0.77h

Above two methods use Model A, and below two use Model B.

TABLE II
PSNR (dB) of images resized using different methods: Interpolation/HWT/HWT&Normalization (Ours).

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

The best and 1st runner up performances are in red and blue, respectively.

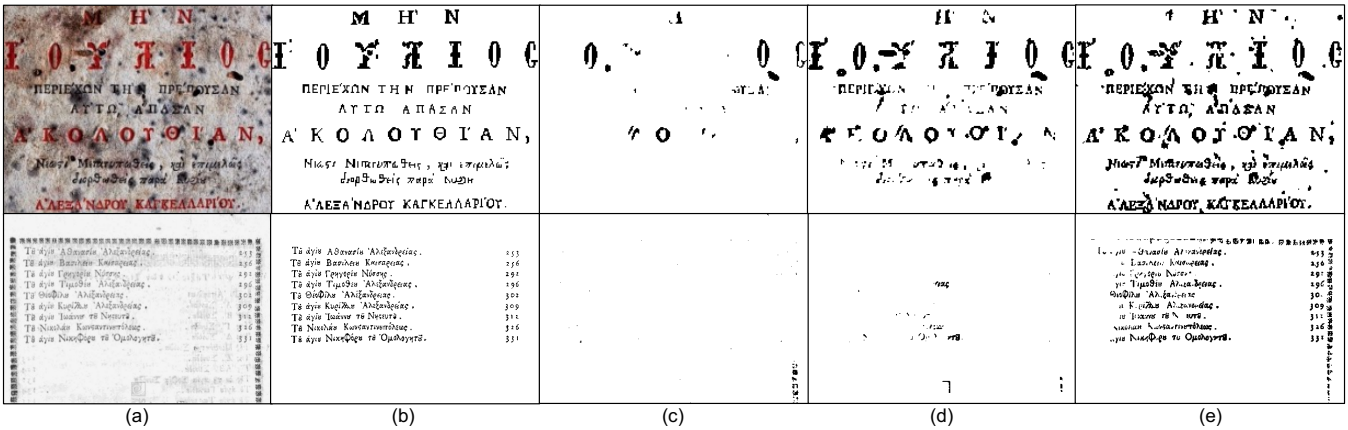
C. Multi-scale Feature Extraction

To reduce total training and inference times, our work proposes to resize both input and the corresponding ground-truth images for GANs training by half. To illustrate the effectiveness of HWT and normalization in Stage 1, we consider two GAN models, namely: Model A: UNet [6] with EfficientNetV2-S [15], and Model B: UNet++ [14] with EfficientNetV2-S [15]. Table I records the time taken for each stage, as well as the total training and inference times of different methods. Two configurations are compared: "with the application of HWT and normalization in Stage 1" and

TABLE III
QUANTITATIVE COMPARISON (ASM: FM/P-FM/PSNR/DRD, TOTAL TRAINING TIME, TOTAL INFERENCE TIME) OF THE PROPOSED METHOD AND SOTA METHODS FOR DOCUMENT IMAGE ENHANCEMENT AND BINARIZATION ON DIBCO DATASETS.

Method	Model	FM↑	p-FM↑	PSNR↑	DRD↓	ASM↑	Total Train↓	Total Inference↓
Suh *et al.* [3]	UNet&EfficientNet-B4	87.95	89.01	19.10dB	4.83	72.81	85.61h	0.74h
Suh *et al.* [3]	UNet&EfficientNet-B5	88.56	89.90	19.31dB	4.46	73.33	304.12h	0.82h
Ju *et al.* [4]	UNet++&EfficientNet-B4	88.14	89.71	19.09dB	4.64	73.08	69.68h	0.91h
Ju *et al.* [4]	UNet++&EfficientNet-B5	89.13	90.35	19.30dB	4.49	73.57	76.29h	1.04h
Ours	UNet&EfficientNetV2-S	88.83	89.87	19.07dB	4.86	73.23	63.91h	0.68h
Ours	UNet++&EfficientNetV2-S	89.69	90.78	19.15dB	4.45	73.79	68.43h	0.77h

The best and 1st runner up performances are in red and blue, respectively.



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

Method	Model	The first row images: FM↑	p-FM↑	PSNR↑	The second row images: FM↑	p-FM↑	PSNR↑
Blank Image	–	–	–	10.90dB	–	–	14.19dB
Suh *et al.* [3]	UNet&EfficientNet-B4	26.09	19.90	11.39dB	0.60	0.60	14.00dB
Ju *et al.* [4]	UNet&EfficientNet-B4	60.39	56.35	12.19dB	10.05	9.32	14.23dB
Ours	UNet&EfficientNetV2-S	69.44	69.88	11.75dB	56.99	56.51	14.08dB

Fig. 3. Representative visualized results from the test set: (a) input image, (b) ground-truth, (c) Suh *et al.* [3], (d) Ju *et al.* [4], and (e) ours.

"without (i.e., the *baseline* where split patches are directly supplied to the GANs)". Here, the total training time is the sum of the time taken for each stage, and the total inference time is the total time taken to generate images for all test sets. It can be seen that, for both models, the total training time is reduced when HWT and normalization are applied. Specifically, when HWT and normalization are applied, the training time is reduced from 384.95h to 63.91h for Model A, and from 523.86h to 68.43h for Model B. Similarly, the total inference time is reduced from 1.12h to 0.68h for Model A, and from 1.19h to 0.77h for Model B. This demonstrates that the use of HWT and normalization can significantly reduce the total training and inference times.

We also explore other image-resizing techniques, 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. Furthermore, we employ the "HWT", and "HWT and normalization (HWT&Norm)" methods. It is noteworthy that the resized images from all these methods are not binarized, which cannot be used to calculate the PSNR values directly with the corresponding ground-truth (binary) images. Therefore, we first apply global binarization 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 we compute the mean PSNR values for all images. The results are recorded in Table II. We observe that the mean PSNR value achieved by "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.

TABLE IV
ABLATION STUDY ON EACH IMPROVEMENT STEP. THE CHECKMARK ($checkmark$) INDICATES THAT A SPECIFIC CONFIGURATION IS IN USE. THE LAST ROW RECORDS THE RESULTS FOR THE PROPOSED METHOD.

Generator: EfficientNetV2-S	Generator: UNet++	Discriminator: InstanceNorm	Generator Loss Function: $D(G(z))+\lambda_1$BCE+λ_2DICE	Image Processing: HWT&Norm	ASM	Total Train	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

D. Comparison with SOTA Methods

We compare our proposed methods with the SOTA methods using GANs [3], [4] for document image enhancement and binarization, where the results are shown in Table III. Considering that the total training time for methods using UNet [6] or UNet++ [14] with EfficientNet-B5 [12] is already longer than that of our proposed method, we do not further compare the methods using EfficientNet-B6, as it is against the goal of reducing training and inference times.

We can see that our proposed method using UNet++ [14] with EfficientNetV2-S [15] achieves the highest ASM of 73.79. It requires a total training time of 68.43h, which is also thesecond shortest time. It is faster than UNet++&EfficientNet-B5 (76.29h) that yields the second highest ASM. Furthermore, the total inference time of our method is 0.77h, which is notably lower than 1.04h as required by Ju *et al.*'s method [4] using UNet++ with EfficientNet-B5, representing a reduction of approximately 26%. Moreover, the proposed method using UNet with EfficientNetV2-S obtains the shortest total training time and inference time of 63.91h and 0.68h, respectively. Although our achieved ASM value of 73.23 is not the highest, when compared to Suh *et al.*'s method [3] (using UNet 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 experiment results demonstrate the efficiency and competitive performance of our proposed method.

Next, we compare the results achieved by all benchmark methods for each evaluation metric. Our method achieves the highest FM and p-FM values of 89.69 and 90.79, 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, our method achieves the second highest value, but with a significantly reduced total training time of 68.43h compared to 304.12h taken by the method with the highest DRD value. Although our method does not achieve the highest PSNR, this metric does not directly reflect the model performance in document image enhancement and binarization. To validate this statement, we randomly select two images from the test set for visual inspection. As shown in Fig. 3, our method generates more complete foreground information. However, due to the high contamination of the document image, some noise is inevitable while generating more content. In contrast, Suh *et al.*'s method [3] and Ju *et al.*'s method [4] generate less content. It is such because the background is white, and PSNR favors methods generating less content, i.e., they will have higher PSNR values. To put things into context, a blank image (i.e., all pixels set to white) yields a PSNR of 14.19dB, which is higher than that of our proposed method (14.08dB). However, it is obvious that the binarized image generated by our method is closer to the ground-truth image than the blank image. These observations support our claim that a higher PSNR value is non-indicative of better model performance, and our method can successfully generate more textual information.

E. Ablation Study

To evaluate the contribution of each enhancement in our proposed method, we gradually replace or remove each improvement step and observe the impact on performance. Table IV summarizes the results under various configurations. To ensure experimental fairness, all experiments were conducted using the same dataset and hyperparameter settings.

Specifically, replacing UNet [6] with UNet++ [14] in the generator and adding instance normalization to the discriminator improve model performance, with a slight increase in training time. Replacing EfficientNet-B5 [12] with EfficientNetV2-S [15] in the generator reduces both training and inference times, while the new loss function further improves performances. Finally, employing HWT and normalization for multi-scale feature extraction significantly reduces training time from 523.86h to 68.43h, representing an 87% decrease, with a drop of only 0.02 in ASM. Overall, each improvement objectively contributes to either performance gains or reductions in training and inference times.

V. CONCLUSION AND FUTURE WORK

Degraded color document image enhancement and binarization are important steps in document analysis. The current SOTA methods based on GANs can generate satisfactory document binarization results, but suffer from long training and inference times. To address this drawback, we propose a three-stage GAN method using HWT and normalization for multi-scale feature extraction, which greatly reduces the total training and inference times. Furthermore, novel generators, discriminators, and a loss function are designed to further improve the performance of our proposed method. Experiment

results on benchmark datasets demonstrate that the proposed method not only achieves superior model performance but also significantly reduces the total training and inference times in comparison to SOTA methods.

As future exploration, we can combine document image binarization and document image understanding for practical applications, especially for ancient documents or historical artifacts. The applications could include real-time translation, summarization, and retrieval of related documents/materials.

ACKNOWLEDGMENT

This work is supported by National Science and Technology Council of Taiwan, under Grant Number: NSTC 112-2221-E-032-037-MY2.

REFERENCES

[1] B. Sun, S. Li, X.-P. Zhang, and J. Sun, "Blind bleed-through removal for scanned historical document image with conditional random fields," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5702–5712, 2016.

[2] N. Kligler, S. Katz, and A. Tal, "Document enhancement using visibility detection," in *CVPR*, 2018, pp. 2374–2382.

[3] S. Suh, J. Kim, P. Lukowicz, and Y. O. Lee, "Two-stage generative adversarial networks for binarization of color document images," *Pattern Recognition*, vol. 130, p. 108 810, 2022.

[4] R.-Y. Ju, Y.-S. Lin, Y. Jin, C.-C. Chen, C.-T. Chien, and J.-S. Chiang, "Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks," *Knowledge-Based Systems*, vol. 304, p. 112 542, 2024.

[5] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *ICDAR*, vol. 1, 2017, pp. 99–104.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[7] X. Peng, H. Cao, and P. Natarajan, "Using convolutional encoder-decoder for document image binarization," in *ICDAR*, vol. 1, 2017, pp. 708–713.

[8] S. He and L. Schomaker, "Deepotsu: Document enhancement and binarization using iterative deep learning," *Pattern recognition*, vol. 91, pp. 379–390, 2019.

[9] J. Calvo-Zaragoza and A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86, pp. 37–47, 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[12] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[15] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*, 2021, pp. 10 096–10 106.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *NeurIPS 2022*, vol. 30, 2017.

[17] A. Galdran, G. Carneiro, and 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*, 2022, pp. 40–51.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016, pp. 565–571.

[19] R. Hedjam and M. Cheriet, "Historical document image restoration using multispectral imaging system," *Pattern Recognition*, vol. 46, no. 8, pp. 2297–2312, 2013.

[20] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *ICDAR*, 2009, pp. 1375–1382.

[21] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-dibco 2010-handwritten document image binarization competition," in *ICFHR*, 2010, pp. 727–732.

[22] F. Deng, Z. Wu, Z. Lu, and M. S. Brown, "Binarizationshop: A user-assisted software suite for converting old documents to black-and-white," in *JCDL*, 2010, pp. 255–258.

[23] S. M. Ayatollahi and H. Z. Nafchi, "Persian heritage image binarization competition (phibc 2012)," in *PRIA*, 2013, pp. 1–4.

[24] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Icfhr2014 competition on handwritten document image binarization (h-dibco 2014)," in *ICFHR*, 2014, pp. 809–813.

[25] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018)," in *ICFHR*, 2018, pp. 489–493.

[26] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, and I. Marthot-Santaniello, "Icdar 2019 competition on document image binarization (dibco 2019)," in *ICDAR*, 2019, pp. 1547–1556.

[27] S. K. Jemni, M. A. Souibgui, Y. Kessentini, and A. Fornés, "Enhance to read better: A multi-task adversarial network for handwritten document image enhancement," *Pattern Recognition*, vol. 123, p. 108 370, 2022.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.