Applied Deep Learning Final Project Report

Rui-Yang Ju R12944059
I-Yuan Kuo R10922A23
Rong-Chi Liou R11528025
Yi-Kai Liao 40947038s

Abstract

Due to the imperfect preservation of ancient documents over a long period of time, various types of degradation including staining, yellowing, and ink bleeding causes serious interference with the image processing. This final project applied Generative Adversarial Networks (GAN), utilizing UNet and UNet++ architectures, with EfficientNetV2 as the encoder to generate images, respectively. The performances of models were evaluated on several Document Image Binarization Contest (DIBCO) datasets and we compared them with the baseline model using UNet architecture with EfficientNet as the encoder. The experimental results show that although the model performance of the method that using UNet++ architecture with EfficientNetV2 is not satisfactory, the model performance of the method that using UNet architecture with EfficientNetV2 is better than the performance of the baseline model. The implementation code is released at `https://github.com/RuiyangJu/Final_Project`.

1 Introduction

Document image binarization is an important research topic, with researchers employing various methods to extract text information from degraded document images. The methods of this research can be broadly categorized into two classes: one is the traditional methods and another is the deep learning-based methods.

The challenge of this research lies in the fact that ancient preserved documents often undergo various forms of interference and degradation, such as paper yellowing, text fading, and page contamination (Lee et al., 2009; Hedjam and Cheriet, 2013). The quality of degraded document images processed by the traditional methods are often unsatisfactory, with some unresolved issues including text degradation and ink bleeding (Kligler et al., 2018; Endo et al., 2020).

With the advancement of deep learning, models employed in semantic segmentation tasks have started to be applied in document segmentation, such as Fully Convolutional Networks (FCN) (Peng et al., 2013; Long et al., 2015) and UNet (Ma et al., 2018). (Suh et al., 2022) proposed a two-stage Generative Adversarial Network (GAN) method to binarize degraded document images. The research (Ju et al., 2022, 2023) utilized UNet (Ronneberger et al., 2015) for document segmentation and employed GAN (Goodfellow et al., 2020) for generating the binarization results. Based on the above research, this final project improved the network architecture to enhance the model performance.

To evaluate the performance of the proposed model in binarizing document images for this final project, we compared it with the baseline model (Ju et al., 2023) on the datasets from the Document Image Binarization Contest (DIBCO) 2011, 2013, 2014, 2016, 2017, and 2018 (Pratikakis et al., 2011, 2013; Ntirogiannis et al., 2014; Pratikakis et al., 2016, 2017, 2018).

The main contributions of this final project are as follows:

- EfficientNetv2 was employed to replace EfficientNet of the original method as the encoder of the generator. Compared to the baseline model, the newly proposed GAN models had better model performances and generated document binarization images more efficiently.
- To show the difference between UNet and UNet++ architecture of the generator, we separately applied both model architectures to perform document segmentation. A comparative analysis was conducted to assess the impact of these two model architectures on the performance of the model.
- We used Binary Cross Entropy and Focal Loss

function to train the generator respectively and performed ablation experiments to analyze the effect of the two loss functions.

2 Related Works

2.1 EfficienNet vs. EfficientNetV2

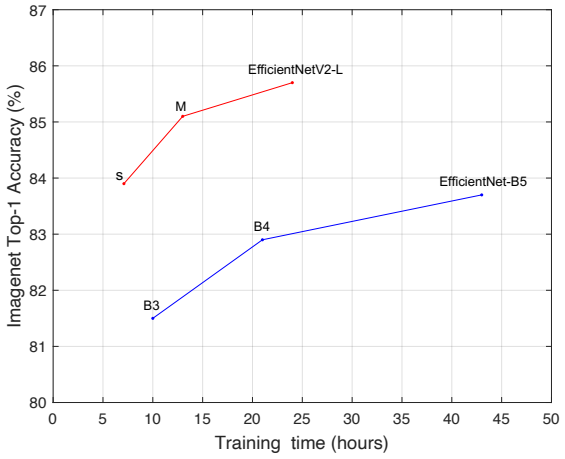
EfficientNet (Ronneberger et al., 2015) is a Convolutional Neural Network (CNN) architecture proposed by the researchers at Google, aiming to improve the efficiency and performance of the model by scaling in different dimensions using the composite scaling method. EfficientNetV2 (Tan and Le, 2021) is an improvement of EfficientNet (Ronneberger et al., 2015), aimed at improving the model flexibility, performance, and generalization capabilities. This improvement involves innovations in network design and architecture, with the goal of adapting to various application scenarios and hardware requirements.

To compare the model performance of EfficientNet and EfficientNetV2, we have reviewed the related papers and collated the results to produce Figure 1. From the table of Figure 1, we can see that on ImageNet dataset (Deng et al., 2009), EfficientNetV2-M spends less training time than EfficientNet-B4 and obtains better model performance, so we choose EfficientNetV2-M to replace EfficientNet-B4 of the original method as the encoder.

2.2 UNet vs. UNet++

UNet (Ronneberger et al., 2015) is a model architecture for medical image segmentation, especially for cell and organ segmentation. It is based on the encoder-decoder structure, where the encoder is responsible for capturing the contextual information of the image and the decoder is used to realize the localization of details. UNet++ (Zhou et al., 2018) improves on UNet by introducing a more complex connection structure (nested skip connections), which improves the ability of capturing multi-level features to achieve better performance in various segmentation tasks.

In order to compare the effects of UNet and UNet++ architectures on the segmentation results, we have reviewed the relevant papers and organized the results into Table 1. From Table 1, it can be seen that compared to UNet architecture model, UNet++ architecture model possesses higher accuracy on MonuSeg datset (Kumar et al., 2017).



	EfficientNet-B4 (Tan and Le, 2019)	EfficientNetV2-M (Tan and Le, 2021)
Top-1 Acc.	82.9%	85.1%
Parameters	19M	54M
Infer-time	30ms	57ms
Train-time	21hours	13hours

Figure 1: The comparison results of EfficientNet and EfficientNetV2 for evaluating model performance on ImageNet dataset.

Table 1: The comparison results of UNet and UNet++ for evaluating model performance on MonuSeg datset.

	UNet (Ronneberger et al., 2015)	UNet++ (Zhou et al., 2018)
Loss	0.429	0.159
Accuracy	0.867	0.914
Dice	0.803	0.762
F1-Score	0.716	0.745

3 Method

3.1 Network Architecture

The task of this final project is to perform image enhancement on degraded color documents and extract text information from the document images. We improve the generator based on the three-stage GAN method proposed by (Ju et al., 2023).

The purpose of the first stage of this method is image enhancement. As shown in Figure 2, we first split the input image into patches of 256×256 size and then split these patches into single channel images (red, green, blue, and gray). The obtained single channel images are processed with the haar wavelet transform (Stankovic and Falkowski, 2003) (which is one of the DWT) and the normalization respectively to achieve the effect of image enhance-

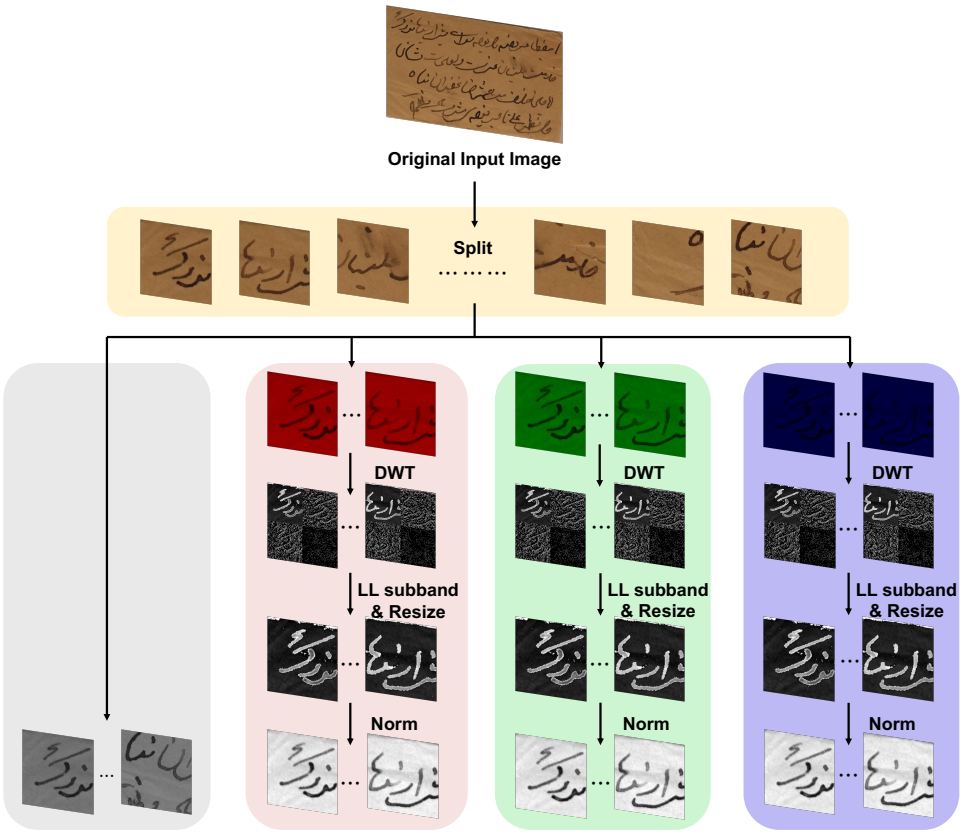


Figure 2: The architecture diagram of stage-1. The original input image is divided into patches of 256×256 size, and split into RGB three single-channel images and one grayscale image for the discrete wavelet transform (DWT) and the normalization.

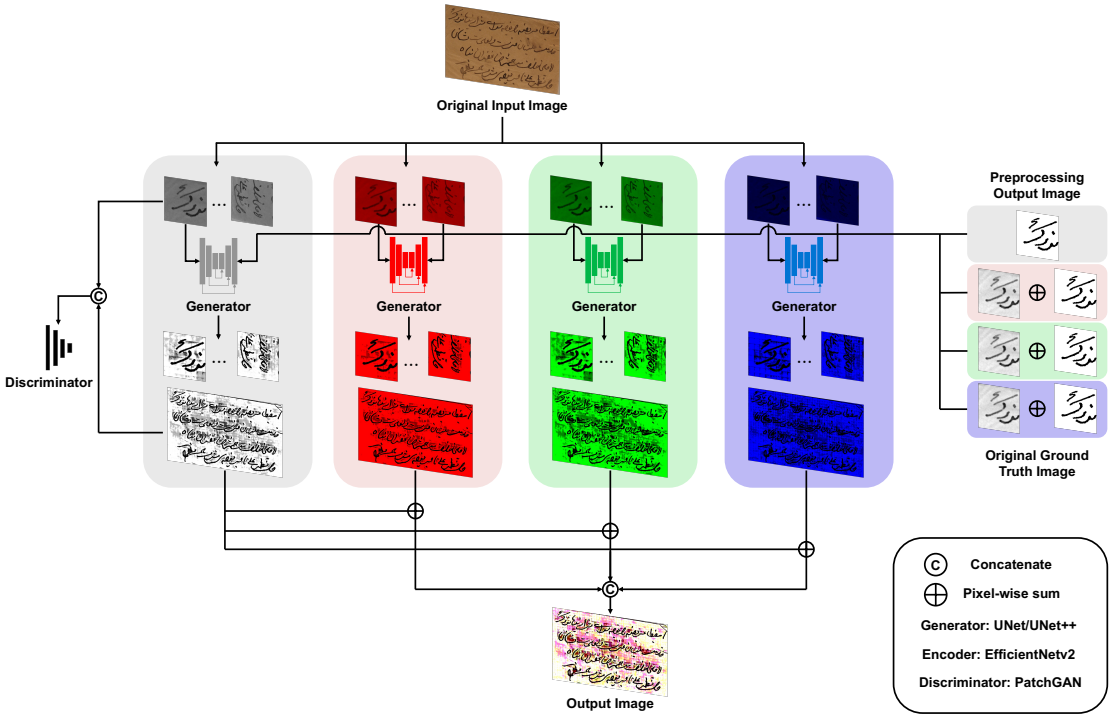


Figure 3: The architecture diagram of the stage-2. This method employs UNet/U-Net++ architecture with EfficientNetV2 as the generator and PatchGAN as the discriminator. The ground truth image during the training process is derived through pixel-wise summation of the dataset-provided image and the output from stage-1.

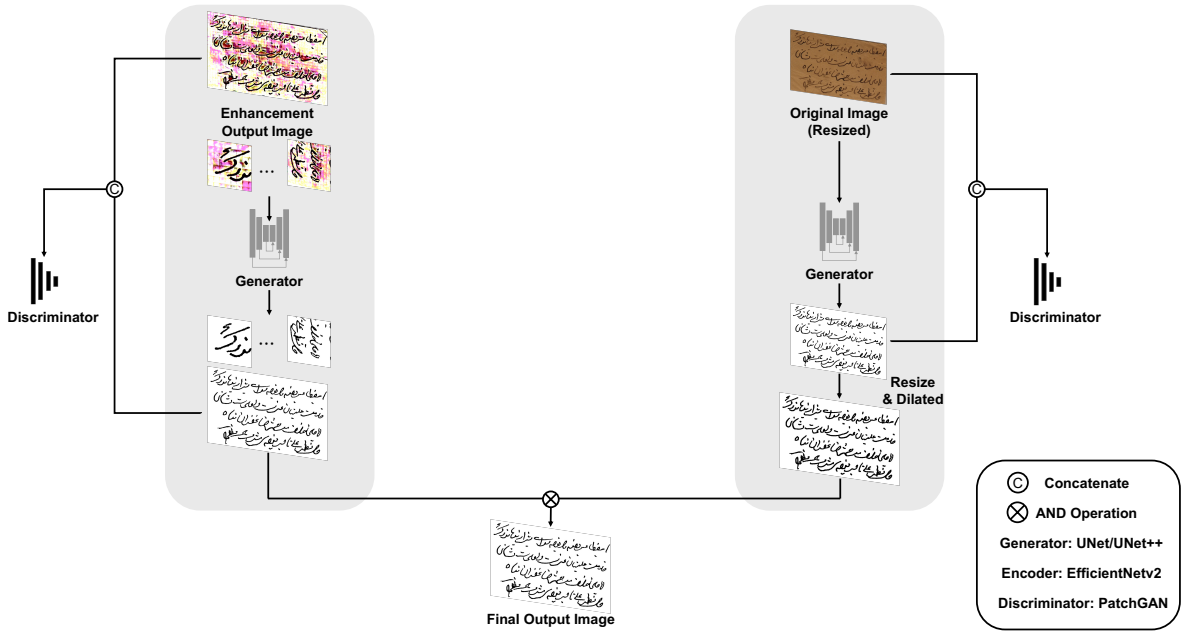


Figure 4: The architecture diagram of the stage-3. This stage comprises both the local and the global predictions. The local prediction involves training several 256×256 size patches, while the global prediction involves directly resizing the original image to 512×512 for training.

ment in preparation for the binarization process in the second stage.

In order to reduce the interference of color in the processing of document image binarization, we used four independent generators in the second stage to generate images to extract the text information under different color backgrounds respectively. Each of the independent generators utilizes the encoder-decoder architecture, and their input images are the outputs of the first stage (four independent generators are input with single-channel images in red, green, blue, and gray, respectively). For the encoder in the generator, we used EfficientNetV2 (Tan and Le, 2021) instead of EfficientNet (Tan and Le, 2019) used in the original method. The original method used the UNet (Ronneberger et al., 2015) architecture as the generator, while we applied the UNet (Ronneberger et al., 2015) and UNet++ (Zhou et al., 2018) architectures to compare the performance of the different methods. As shown in Figure 3, four independent generators share a single discriminator. Since the discriminator in PatchGAN (Li and Wand, 2016; Zhu et al., 2017; Ledig et al., 2017) has a better generalization ability, we chose and slightly improved it as our discriminator.

The third stage of the network architecture is divided into the local prediction and the global prediction, as shown on the left and right sides of Figure 4, respectively. For the local prediction, we split the output of the second stage into patches of 256×256 size again and used a independent generator and a independent discriminator to generate the binarization results. To compensate for the loss of spatial context information of the document image due to the local prediction, the global prediction scales the original document image to the size of 512×512 and then inputs it directly to the independent generator to generate the image.

3.2 Loss Function

To enhance the convergence for the loss function, we implemented the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) (Gulrajani et al., 2017) on the objective function. The experimental results of (Bartusiak et al., 2019) have demonstrated that the binary cross-entropy (BCE) loss outperforms the $L1$ loss in binary classification tasks. Therefore, we chose to employ the BCE loss instead of the $L1$ loss used in the Pix2Pix GAN framework (Isola et al., 2017). The loss function of the WGAN-GP target, which incorporates the BCE loss, is expressed as follows:

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

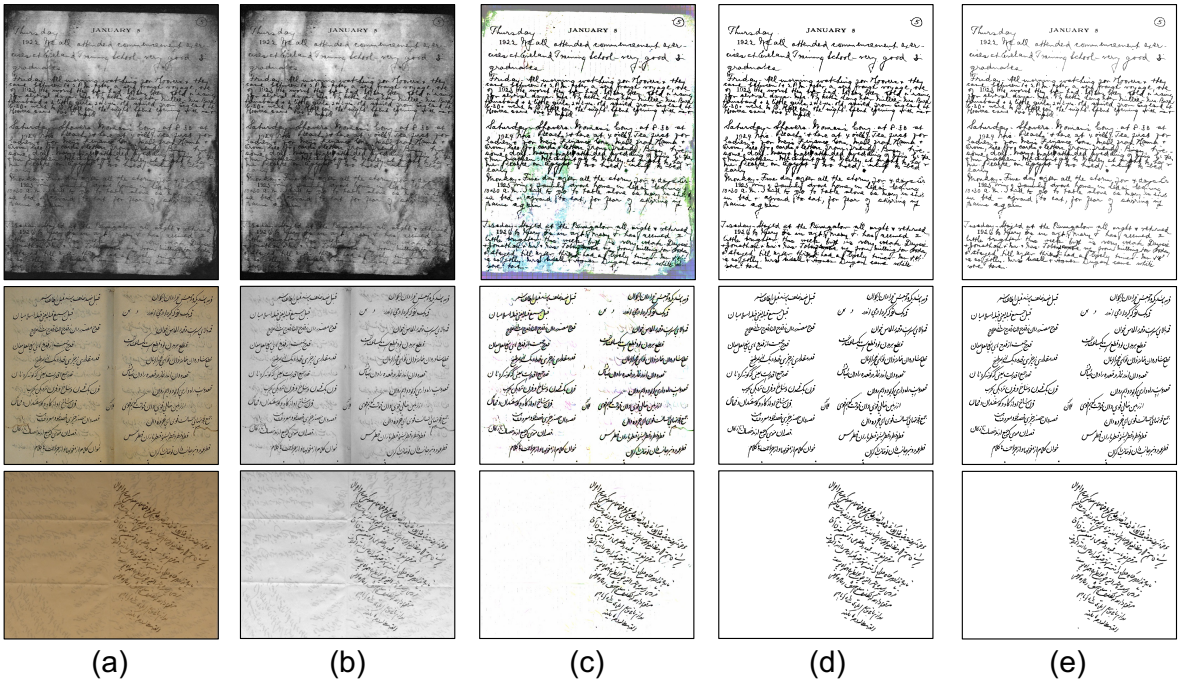


Figure 5: Examples of image enhancement and binarization on DIBCO dataset: (a) the original input image, (b) the LL subband image using discrete wavelet transformation and normalization (Stage-1), (c) the enhanced image using image enhancement method (Stage-2), (d) the binarization image using our method combining local and global features (Stage-3), (e) the ground truth image.

$$\mathbb{L}_G = \mathbb{E}_x[D(G(x), x)] + \lambda\, \mathbb{E}_{G(x),y}[y \log G(x) + (1-y)\log(1-G(x))] \quad (2)$$

where the penalty coefficient is denoted as α, and uniform sampling occurs along a straight line connecting the ground truth distribution P_y, and the point pairs of the generated data distribution are represented as $P_{\hat{y}}$. The parameter λ is employed to regulate the relative significance of various loss terms, enhancing clarity regarding its role in determining the impact of different loss terms. The parameter of generator is denoted as θ_G, while the parameter of discriminator is denoted as θ_D, maintaining consistency in terminology and improving clarity. Within the discriminator, the target loss function $\mathbb{L}_D$ in Eq. (1) is employed to differentiate the generated image from the real image, specifying the role of the discriminator and the associated loss function more explicitly. In the generator, the target loss function $\mathbb{L}_G$ in Eq. (2) is utilized to minimize the distance between the generated image and the ground truth image in each color channel, enhancing understanding of the generator objective and the corresponding loss function.

4 Experiments

4.1 Datasets

This final project used 10 images from DIBCO 2009 (Gatos et al., 2009), 10 images from H-DIBCO 2010 (Pratikakis et al., 2010), 15 images from the Persian Heritage Image Binarization Dataset (PHIBD) (Nafchi et al., 2013), 14 images from H-DIBCO 2012 (Pratikakis et al., 2012), 87 images from the Synchromedia Multispectral Ancient Document Images Dataset (SMADI) (Hedjam and Cheriet, 2013), and 7 images from the Bickley Diary Dataset (Deng et al., 2010) as the training set. For the test set, this final project selected 16 images from DIBCO 2011 (Pratikakis et al., 2011), 16 images from DIBCO 2013 (Pratikakis et al., 2013), 10 images from H-DIBCO 2014 (Ntirogiannis et al., 2014), 10 images from H-DIBCO 2016 (Pratikakis et al., 2016), 20 images from DIBCO 2017 (Pratikakis et al., 2017), and 10 images from H-DIBCO2018 (Pratikakis et al., 2018). In total, there are 143 images in the training set, and 82 images in the test set.

4.2 Evaluation Metric

Four evaluation metrics were used for this final project, including F-measure (FM), pseudo-

Table 2: Compare the effectiveness of training the generator using two different loss functions: Binary Cross Entropy and Focal Loss. The best model performance are in **boldface**.

(a) Binary Cross Entropy

Datasets	FM↑	p-FM↑	PSNR↑	DRD↓
DIBCO 2011	88.19	88.99	19.08	3.35
DIBCO 2013	92.95	93.25	20.39	2.94
H-DIBCO 2014	95.35	95.69	20.79	1.39
H-DIBCO 2016	92.22	94.18	19.72	2.90
DIBCO 2017	89.14	90.19	18.02	3.61
H-DIBCO 2018	93.0	94.54	20.26	2.46
Mean	**91.82**	**92.81**	**19.71**	**2.78**
Standard Deviation↓	2.44	2.41	0.93	**1.01**

(b) Focal Loss (Alpha = 0.25, Gamma = 2.0)

Datasets	FM↑	p-FM↑	PSNR↑	DRD↓
DIBCO 2011	86.50	87.66	16.76	4.82
DIBCO 2013	87.81	87.93	17.28	5.15
H-DIBCO 2014	91.47	91.60	17.72	3.02
H-DIBCO 2016	89.14	89.90	17.79	4.65
DIBCO 2017	83.69	83.97	15.26	6.80
H-DIBCO 2018	87.49	88.20	17.59	5.23
Mean	87.68	88.21	17.07	4.95
Standard Deviation↓	**2.37**	**2.33**	**0.87**	1.11

F-measure (p-FM), peak signal-to-noise ratio (PSNR), and distance reciprocal distortion (DRD) to compare different model performances.

4.3 Training

This final project was implemented using Python in the framework of PyTorch, and we used NVIDIA RTX 3090 GPUs for training. We set the same hyperparameters of the training model used by (Ju et al., 2023). Specifically, we set 150 epochs for the global prediction training, while other model training were trained for 10 epochs. We selected the Adam (Kingma and Ba, 2014) optimizer, setting the learning rate to 2×10^{-4}. For the generator, we used $\beta_1 = 0.5$, and for the discriminator, we set $\beta_2 = 0.999$.

4.4 Ablation Study

4.4.1 Global Prediction

To answer the question raised by the professor in the questions and answers (Q&A) of In-person Presentation, we have showed the advantages of each stage, proving that the processing results without gobal prediction are not satisfactory.

As shown in Figure 5, (b) represents the result of preserving the LL-band image after DWT and performing normalization. This stage performs noise reduction on the original input image, which serves as the ground truth image for the generator. (c) The result of image enhancement using an adversarial network where the image has the background color removed and the text color highlighted. (d) is the final output image obtained using the proposed method and it can be seen that our output is close to the ground truth image (e).

4.4.2 Binary Cross Entropy v.s. Focal Loss

Taking into consideration the presence of artifacts or color biases across different image files, the image data encompasses a substantial number of pixels that are relatively easy to denoise, alongside a minority of pixels posing greater challenges for noise removal. Direct utilization of Cross Entropy may lead to the model disproportionately focusing on optimizing well-handled pixels during training. Consequently, we explored an alternative approach by training the Generator using Focal Loss instead of Binary Cross Entropy.

Despite experimentation, the empirical evidence indicates that the use of Binary Cross Entropy continues to yield superior results. Interestingly, employing Focal Loss with a gamma value of 2.0 during training results in enhanced stability. This observation suggests that the original methodology is adept at effectively addressing diverse pixel scenarios. Moreover, the incorporation of Focal Loss serves to intensify the learning of hard samples, thereby achieving comparable performance across disparate datasets.

Table 2 compares the mean scores and standard deviations of the generator trained using binary cross-entropy and focal loss. The results demonstrate that employing binary cross-entropy enables effective handling of diverse pixel conditions. Additionally, incorporating focal loss and accentuating challenging samples contributes to a reduction

Table 3: Quantitative comparison (FM/p-FM/PSNR/DRD) with the baseline model for degraded document image binarization on the several DIBCO datasets. The best model performance are in **boldface**.

(a) DIBCO 2011

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	90.21	91.52	**19.80**	2.88
UNet&EffNetV2	**92.65**	**93.79**	19.73	**2.55**
UNet++&EffNetV2	89.11	90.97	19.28	3.65

(b) DIBCO 2013

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	**94.41**	**95.02**	**21.36**	**2.15**
UNet&EffNetV2	94.10	94.63	21.05	2.32
UNet++&EffNetV2	93.90	94.38	21.00	2.48

(c) H-DIBCO 2014

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	96.29	96.82	21.66	1.11
UNet&EffNetV2	**96.30**	**96.94**	**21.67**	**1.10**
UNet++&EffNetV2	96.07	96.51	21.39	1.19

(d) H-DIBCO 2016

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	92.01	**95.17**	**19.68**	2.93
UNet&EffNetV2	91.86	95.00	19.64	2.94
UNet++&EffNetV2	**92.03**	95.16	19.65	**2.85**

(e) DIBCO 2017

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	88.87	89.98	**17.92**	**3.89**
UNet&EffNetV2	**89.66**	**90.81**	17.77	4.18
UNet++&EffNetV2	89.37	90.34	17.75	4.15

(f) H-DIBCO 2018

Methods	FM↑	p-FM↑	PSNR↑	DRD↓
(Ju et al., 2023)	**93.75**	**95.69**	**20.76**	**2.12**
UNet&EffNetV2	93.59	95.19	20.58	2.20
UNet++&EffNetV2	93.61	95.14	20.55	2.19

in standard deviation across the performance of the generator on various datasets.

4.5 Experimental Results

This final project evaluates different models on six DIBCO datasets (DIBCO 2011, DIBCO 2013, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 and H-DIBCO 2018). Since all above DIBCO datasets do not provide Optical Character Recognition (OCR) outputs, we used four evaluation metrics introduced in Section 4.2 to evaluate the baseline model and the improved models.

Table 3 shows the results of evaluating the baseline model and the improved models on each of six DIBCO datasets. On H-DIBCO 2014 dataset, we can see that the model of UNet architecture with EfficientNetV2 achieves the best performance in FM, p-FM, PSNR, and DRD. On DIBCO 2011 dataset, the model of UNet architecture with EfficientNetV2 improves the FM value of the baseline model from 90.21 to 92.56, and the p-FM value from 91.52 to 93.79. And on DIBCO 2017 dataset, this model also achieves higher FM and p-FM values than the baseline model. These experimental results demonstrate that the model UNet architecture with EfficientNetV2 outperforms the baseline model on a few DIBCO datasets.

However, the model performance of UNet++ architecture with EfficientNetV2 is not satisfactory, with the best performance only obtained on H-DIBCO 2016 dataset for the FM and DRD values. Since this model outperforms the baseline model only in terms of the two evaluation metrics, we cannot assess whether the binarization process is better or not, so we select an example on H-DIBCO 2014 dataset and the H-DIBCO 2016 dataset, respectively, for the presentation of the results.

As shown in Figure 6 and Figure 7, we show the results of document images binarization using the baseline model, the model of UNet architecture with EfficientNetV2 and the model of UNet++ architecture with EfficientNetV2, respectively. Before obtaining the processed images, based on the experimental results in Table 3, we think that the best binarization result for image 008 of H-DIBCO 2014 dataset should be (d) the model of UNet architecture with EfficientNetV2, and the best binarization result for image 006 of H-DIBCO 2016 dataset should be (e) the model of UNet++ architecture with EfficientNetV2. As we can see from Figures 6 and 7, the results are the same as we expected.

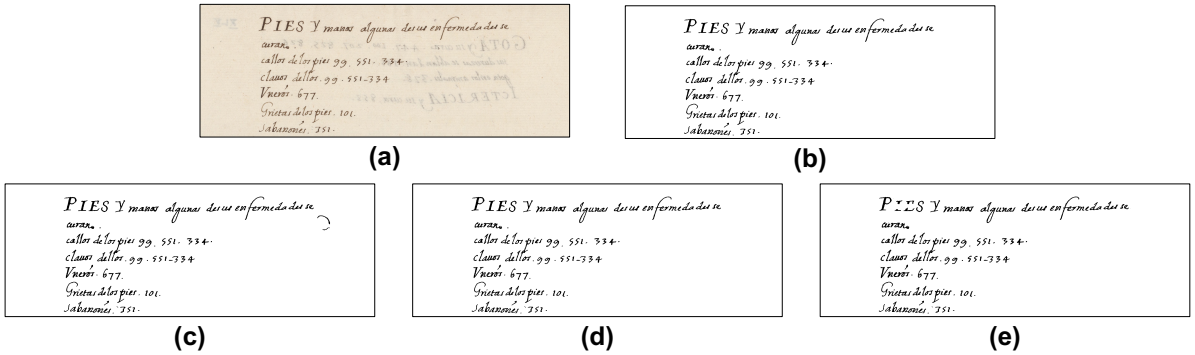


Figure 6: Example of document image binarization results from the input image 008 of H-DIBCO 2014 dataset by different methods: (a) original input images, (b) the ground truth, (c) baseline model (Ju et al., 2023), (d) UNet architecture with EfficientNetV2 model, (e) UNet++ architecture with EfficientNetV2 model.

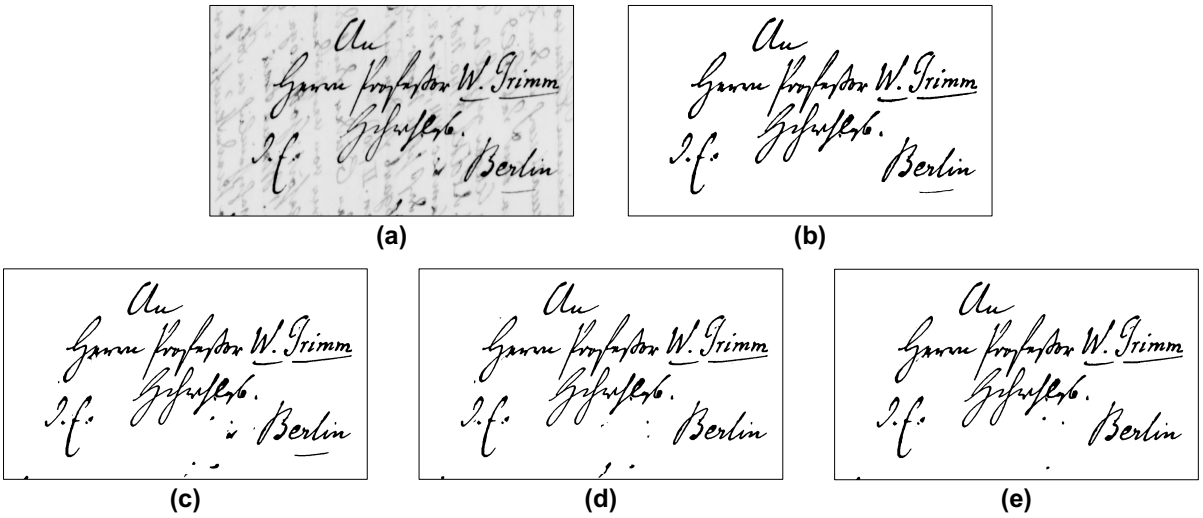


Figure 7: Example of document image binarization results from the input image 006 of H-DIBCO 2016 dataset by different methods: (a) original input images, (b) the ground truth, (c) baseline model (Ju et al., 2023), (d) UNet architecture with EfficientNetV2 model, (e) UNet++ architecture with EfficientNetV2 model.

5 Discussion

Before performing the experiments, we have reviewed many papers on UNet and UNet++ architectures, and all of them proved that UNet++ architecture is an improved version of UNet architecture and can obtain better segmentation results. Therefore, we hypothesize that UNet++ architecture can obtain better model performance. However, according to the experimental results, we found that the model performance of UNet++ architecture with EfficientNetV2 is not as good as that of UNet architecture with EfficientNet. We think that there may be two reasons for this result, the first is that there is an overfitting problem when using UNet++ architecture, and the second is that UNet++ architecture captures more features, including some interference information, so the result is worse than that.

6 Conclusion

This final project improves on the original three-stage GAN method for image enhancement and binarization of color document images. For the encoder in the generator, we applied EfficientNetV2 instead of EfficientNet of the original method. Based on this, we performed the experiments using the UNet and UNet++ architectures respectively. An additional attempt was made to compare the experimental results of training Generator through Focal Loss instead of Binary Cross Entropy. We compared our models with the baseline model on several Document Image Binarization Contest (DIBCO) datasets. The experimental results show that although the model performance of UNet++ architecture with EfficientNetV2 is not satisfactory, the model performance of UNet ar-

chitecture with EfficientNetV2 is better than the baseline model.

7 Baseline Model GitHub

GitHub of the baseline model is https://github.com/abcpp12383/ThreeStageBinarization, which we have modified to complete this final project report.

References

Emily R Bartusiak, Sri Kalyan Yarlagadda, David Güera, Paolo Bestagini, Stefano Tubaro, Fengqing M Zhu, and Edward J Delp. 2019. Splicing detection and localization in satellite imagery using conditional gans. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 91–96. IEEE.

Fanbo Deng, Zheng Wu, Zheng Lu, and Michael S Brown. 2010. Binarizationshop: a user-assisted software suite for converting old documents to black-and-white. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 255–258.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Kazuki Endo, Masayuki Tanaka, and Masatoshi Okutomi. 2020. Cnn-based classification of degraded images with awareness of degradation levels. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4046–4057.

Basilis Gatos, Konstantinos Ntirogiannis, and Ioannis Pratikakis. 2009. Icdar 2009 document image binarization contest (dibco 2009). In *2009 10th International conference on document analysis and recognition*, pages 1375–1382. IEEE.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Rachid Hedjam and Mohamed Cheriet. 2013. Historical document image restoration using multispectral imaging system. *Pattern Recognition*, 46(8):2297–2312.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Rui-Yang Ju, Yu-Shian Lin, Jen-Shiun Chiang, Chih-Chia Chen, Wei-Han Chen, and Chun-Tse Chien. 2023. Ccdwt-gan: Generative adversarial networks based on color channel using discrete wavelet transform for document image binarization. In *20th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2023, Jakarta, Indonesia, November 15–19*, pages 186–198. Springer Nature Singapore.

Rui-Yang Ju, Yu-Shian Lin, Yanlin Jin, Chih-Chia Chen, Chun-Tse Chien, and Jen-Shiun Chiang. 2022. Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks. *arXiv preprint arXiv:2211.16098*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Netanel Kligler, Sagi Katz, and Ayellet Tal. 2018. Document enhancement using visibility detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2374–2382.

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

Jung-San Lee, Chin-Hao Chen, and Chin-Chen Chang. 2009. A novel illumination-balance technique for improving the quality of degraded text-photo images. *IEEE Transactions on circuits and systems for video technology*, 19(6):900–905.

Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. 2018. Docunet: Document image unwarping via a stacked u-net. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4709.

Hossein Ziaei Nafchi, Seyed Morteza Ayatollahi, Reza Farrahi Moghaddam, and Mohamed Cheriet. 2013. An efficient ground truthing tool for binarization of historical manuscripts. In *2013 12th International Conference on Document Analysis and Recognition*, pages 807–811. IEEE.

Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. 2014. Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *2014 14th International conference on frontiers in handwriting recognition*, pages 809–813. IEEE.

Xujun Peng, Huaigu Cao, Krishna Subramanian, Rohit Prasad, and Prem Natarajan. 2013. Exploiting stroke orientation for crf based binarization of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1034–1038. IEEE.

Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. 2010. H-dibco 2010-handwritten document image binarization competition. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, pages 727–732. IEEE.

Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. 2011. Icdar 2011 document image binarization contest (dibco 2011). In *2011 International Conference on Document Analysis and Recognition*, pages 1506–1510. IEEE.

Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. 2012. Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In *2012 international conference on frontiers in handwriting recognition*, pages 817–822. IEEE.

Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. 2013. Icdar 2013 document image binarization contest (dibco 2013). In *2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476.

Ioannis Pratikakis, Konstantinos Zagori, Panagiotis Kaddas, and Basilis Gatos. 2018. Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 489–493. IEEE.

Ioannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos. 2016. Icfhr2016 handwritten document image binarization contest (h-dibco 2016). In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 619–623. IEEE.

Ioannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos. 2017. Icdar2017 competition on document image binarization (dibco 2017). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1395–1403. IEEE.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Radomir S Stankovic and Bogdan J Falkowski. 2003. The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44.

Sungho Suh, Jihun Kim, Paul Lukowicz, and Yong Oh Lee. 2022. Two-stage generative adversarial networks for binarization of color document images. *Pattern Recognition*, 130:108810.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Mingxing Tan and Quoc Le. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.