Cascading Modular U-Nets for Document Image Binarization

Seokjun Kang*, Brian Kenji Iwana*, Seiichi Uchida*
*Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan
Email: {seokjun.kang, brian, uchida}@human.ait.kyushu-u.ac.jp

***Abstract*—In recent years, U-Net has achieved good results in various image processing tasks. However, conventional U-Nets need to be re-trained for individual tasks with enough amount of images with ground-truth. This requirement makes U-Net not applicable to tasks with small amounts of data. In this paper, we propose to use "modular" U-Nets, each of which is pre-trained to perform an existing image processing task, such as dilation, erosion, and histogram equalization. Then, to accomplish a specific image processing task, such as binarization of historical document images, the modular U-Nets are cascaded with inter-module skip connections and fine-tuned to the target task. We verified the proposed model using the Document Image Binarization Competition (DIBCO) 2017 dataset.**

***Keywords*-U-Net; binarization; convolutional neural network**

I. Introduction

U-Nets [1] was developed for biomedical image segmentation tasks by Ronneberger et al. U-Nets have been used for various image processing tasks, for instance, semantic segmentation. The strength of U-Nets is to capture the context of the input in the *contracting path* and to enable sophisticated localization in the *expanding path*. This forms a symmetrical structure.

However, U-Nets have a limitation that they are an end-to-end learning model that can only be used for specific purposes. For example, to divide a cell in a biomedical image using U-Nets, we need to use annotated images as labels which consider only segmentation. If an input image is exposed to various noise, it is necessary to remove the noise through pre-processing. This is against the concept of end-to-end learning that U-Net strives for. In addition, U-Nets are usually used to perform special purposes, such as dividing cells in a biomedical image or segmenting figures and text in document images. U-Nets specifically trained for the dataset and cannot be easily used for other tasks. Furthermore, if the size of the dataset used is too small to optimize the model, we cannot guarantee that the U-Net's performance is sufficient.

U-Net has been further improved to overcome its limitation by changing the structure of U-Nets or stacking several U-Nets. Research has been conducted to improve the segmentation performance of U-Nets by changing the internal structure of U-Net such as Residual U-Net (RUN) [2]. In addition, in order to obtain robust segmentation performance on small training datasets, stacked U-Net research has been researched [3]–[6]. However, these studies were also unable to escape the problem of ensuring sufficient data for U-Net's optimization.

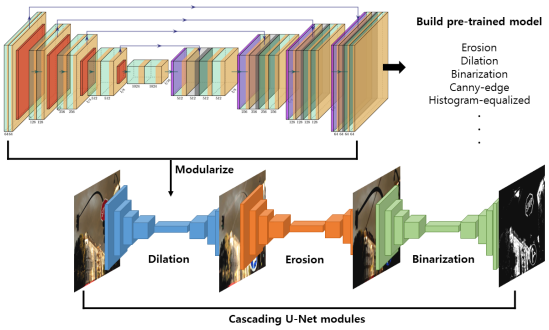


Figure 1: Concept of the proposed model. After cascading, fine-tuning is performed with the (often small) training image set of the target task. As shown later, the modules are further connected with inter-module skip-connections, which are omitted here for simplicity.

The purpose of the paper is to propose a cascaded U-Net model using modularized pre-trained U-Nets, as shown in Fig. 1, for overcoming this limitation by a more efficient strategy. Specifically, we first create U-Net modules, each of which realizes the conventional image processing techniques such as Canny edge, erosion, dilation, etc. These modules are fully trained using a general large dataset and therefore have no problem with data availability. Then, the modularized U-Nets are cascaded in an appropriate order with inter-module connections, which acts as the "glue" for connecting the modular U-Nets. Finally, we perform fine-tuning of the entire network with a (generally small) training dataset of the target task. Since each modular U-Net is already trained to play an image processing task, we can expect that no large dataset is necessary for this fine-tuning to accomplish the target task.

In this paper, we evaluate the performance of the proposed framework on the document image binarization task using the Document Image Binarization Competion (DIBCO) 2017 [7], while the proposed framework can be applicable to any image conversion task with a small training dataset. Through this, we demonstrate that we are able to get state-of-the-art results in image binarization using multiple evaluation measures. We also propose the possibility of using modularized U-Nets and show that the modularized U-Nets can be cascaded and fine-tuned using small training sets.

II. Related work

In this section, we briefly mention related work on U-Nets, stacked U-Nets, DIBCO, and related studies.

A. U-Nets and stacked U-Nets

In recent times, CNNs have achieved remarkable results in the field of image recognition [8]–[10] and document analysis [11]–[13]. This is possible due to the expansion of CNNs to pixel-to-pixel tasks [1], [14], [15]. U-Nets, in particular, have had impressive results in semantic segmentation [1].

Since 2017, the use of multiple U-Nets in research has been carried out in order to further enhance segmentation performance [3]–[6]. For example, Sevastopolsky et al. [4] diagnosed glaucoma effectively by using multiple stacked U-Nets. Sevastopolsky et al. suggested two main strategies to carry out accurate segmentation. First, they stacked RUN blocks to obtain good segmentation results in glaucoma images. Second, they stacked the U-Nets and the segmentation performance of the model is improved.

B. DIBCO and H-DIBCO

DIBCO [7] and H-DIBCO [16] are contests that started in 2009 and was held to objectively evaluate and compare binarization performance in document image analysis. Through this competition, there have been many advances in document binarization technologies every year. Document image binarization is a popular and still active topic as evidenced by the DIBCO and H-DIBCO. The primary challenge of binarization is determining the thresholds at which to binarize the pixels. In the past, the threshold was tackled directly [17]–[19]. However, document binarization systems were introduced at DIBCO which utilized image neural network models. Through the use of neural networks, binarization became possible without specific consideration to thresholds [12], [20]. The winner of H-DIBCO 2018 applied a traditional image processing technique [16]. Their binarization method organized the morphological transformations, stroke width transform (SWT) [21] and Howe's binarization [22], to separate foreground and background pixels. Afterwards, their system adopted image pre-processing to remove noise and preserve text stroke connectivity.

Interestingly, as mentioned in [16], the performance of the DIBCO 2017 winners team's system was slightly better than the 2018 winners team's system. So, we also need to analyze the DIBCO 2017 winner's algorithm. The DIBCO 2017 winners team developed a document binarization system using a U-Net [7]. In the case of the DIBCO 2017 winners, they analyzed the dataset and found the causes of performance degradation, such as luminance, document's boundary, etc. So, they ameliorated the dataset's problem through image processing methods such as elastic deformation, projection noise. While they were able to build a robust binarization system, the study has limitations. First, when using a single U-Net, performance degradation occurs when using a new dataset. Also, they have to use a huge amount of training dataset to optimize their model. However, in the case of document image binarization, the number of annotated document images is often insufficient. Secondly, the conventional image processing method is vulnerable to new noise and backgrounds. This is because it is not easy to deal with all of the obstacles. The systems introduced at DIBCO clearly demonstrated good binarization performance, but they also did not solve the problem of an insufficient sized dataset. Therefore, we obtain effective binarization performance by the following method.

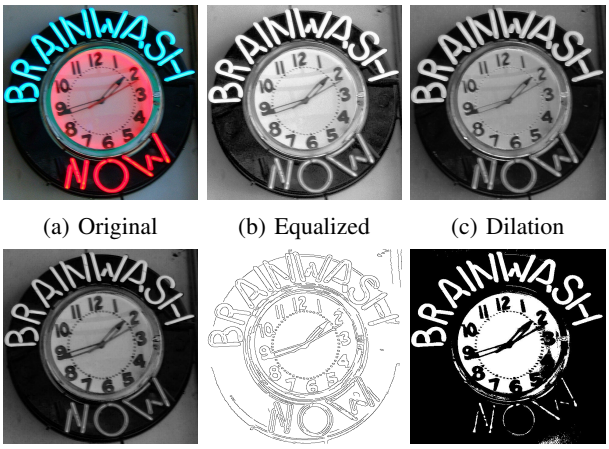
III. Proposed method

U-Nets can be described by two main characteristics. The first is the implementation of a Fully Convolutional Network (FCN) [14] model with a *contracting path* and *expanding path* structure to facilitate image-to-image processing. The second characteristic is copying and cropping of the *contracting path*'s features to make a more accurate localization in the *expanding path*. Sevastopolsky et al. [4] described the method of copying and cropping the features by simply concatenating the features of the *contracting path*, which is referred to as using skip-connections. In the original work [4], they adopted a method of elastic deformation and weighted cross entropy strategies to augment the data. Ultimately, using data augmentation, U-Net even performed quite well with small amounts of annotated images on the Electron Microscopic (EM) and the IEEE International Symposium on Biomedical Imaging (ISBI) challenge. However, for the purposes of the paper, we utilize a strategy for using pre-trained modular U-Nets using sufficient annotated images and then fine-tuning the cascaded model on the small amounts of the data in the DIBCO 2017 dataset.

A. Modular U-Nets

Document binarization using conventional image processing techniques has limitations and a document binarization system using U-Nets was introduced in DIBCO 2017 [7]. The U-Net of the DIBCO 2017's winner has shown considerable results in document binarization, but as document images become more complex and noise increases, performance degradation can occur. Therefore, we want to implement a system based on U-Net that realizes the conventional image processing technique required in document binarization. We refer to this as *modular U-Nets* because single image processing tasks are contained in U-Net based modules for the end goal of a stacked or cascaded U-Net.

In this paper, five modular U-Nets are prepared to simulate the following conventional image processing techniques, Otsu's binarization [17], erosion, dilation, Canny-edge, and histogram-equalization. Each U-Net is pre-trained individually using the MS COCO [23] dataset. The MS COCO-Text dataset contains a total of 63,686 images, of which



(a) Original (b) Equalized (c) Dilation

(d) Erosion (e) Canny edge (f) Binarizaion

Figure 2: Example of the COCO-Text images for training U-Nets.

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

(e) (f) (g)

Figure 3: The result images of the simply cascaded modular U-Nets. (a) Original image. (b) Ground truth. (c) Single U-Net for binarization. (d) Cascaded modular U-Nets with histogram-equalization and binarization modules. (e) Cascaded U-Net with erosion, dilation, and binarization modules. (f) Cascaded modular U-Nets with histogram-equalization, erosion, dilation, and binarization modules. (g) Cascaded modular U-Net with inter-module skip connections and using the histogram-equalization, erosion, dilation, and binarization modules.

10,000 were used as the training dataset. An example of annotated images by using COCO-Text dataset is shown in Fig. 2. In the case of the Canny-edge image in Fig. 2, the background is expressed in white and the edge area is expressed in black. For erosion and dilation, we set the kernel size to 2×2. This is to see that the noise is removed by performing a minimum transformation through morphological transformation. Histogram-equalization is an image processing technique that makes the brightness of the image uniform by redistributing the brightness values of the whole image. Contrast-limited Adaptive Histogram Equalization (CLAHE) [24] can be used to prevent excessive contrast correction by performing local equalization through mask setting. However, we adopt the simple histogram equalization to train our module. The reason for this is that the final goal of the proposed system is the document binarization and not histogram-equalization. The generated datasets from the conventional image processing techniques were used in the training of each of the modular U-Nets. Model training performed 10 epochs with 4,000 steps in total. Through this process, we are able to secure five pre-trained U-Nets.

B. Cascaded modular U-Nets

In order to cascade the model in an optimal way, we proceeded to verify the arrangement order of each module and find the requisite quantity of U-Nets to be used. Fig. 1 illustrates a concept of the proposed model which consists of pre-trained U-Net modules. We cascaded from two to four modular pre-trained U-Nets.

Unfortunately, a simple cascade does not work appropriately. The result of the binarization by the simple cascade is shown in Fig. 3, where four modular U-nets for histogram-equalization, erosion, dilation, and binarization are cascaded. In this figure, we can see that the results of the previous module are not transferred well to the next module. Especially, the localized features that are processed in the *contracting path* are not transmitted correctly. This results in partially noisy areas and untreated areas.

C. Inter-module skip-connections

To properly transfer information from previous modules, we establish skip-connections among the entire modules. Much like the intra-module U-Net skip-connections, the inter-module skip-connections copy and crop the features from matching layers of the *contracting path* to layers of different U-Net modules. We considered that the inter-module skip-connections are useful for carrying lost information from layers of previous modules to layers of subsequent modules. However, through this work, we realized that excessive inter-module skip-connections leads to too many features transference and degradation in binarization performance. In addition, this can be useful for most image-to-image tasks, for binarization, it can also bring unwanted noise.

Therefore, we suggest the architecture of the modular cascading U-Nets with inter-module skip-connections that is illustrated in Fig. 4. Using the proposed inter-module skip-connections scheme, significant features are correctly carried over between modules. In addition, fewer skip-connections also aids the ease of training the network compared with the case of excessive skip-connects. Therefore, as shown in Fig. 4, we limit the number of skip-connections to the subsequent modules to $n-1$, where n is the total number of modules. This ensures that the localization features of each module are appropriately reflected in the final result while at the same time minimizing interference between modules. In Fig. 3 (g), we can see the result image of the

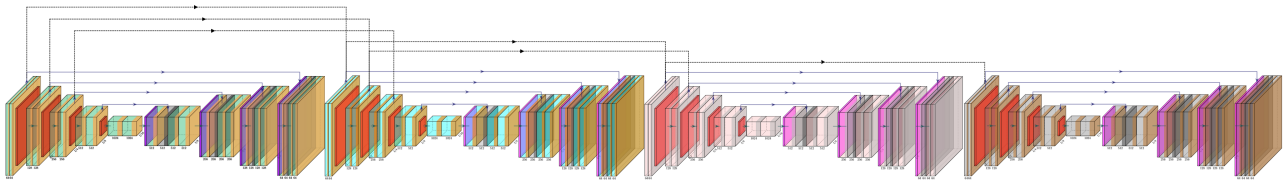


Figure 4: Architecture of the proposed model with inter-module skip-connections.

proposed method included the proposed inter-module skip-connections. Comparing Fig. 3 (f) and Fig. 3 (g) with GT image, we can see that background noise is less segmented in Fig. 3 (g) than that of Fig. 3 (f). This allows the inter-module skip-connections to ensure that the information from the previous module is properly propagated.

IV. Experimental results on historical document binarization

A. Dataset

In this experiment, The COCO-Text and DIBCO 2017 dataset are used. One advantage of using these modular U-Nets is the ability to utilize a transfer learning from the MS COCO-Text dataset in order to enhance the binarization results with the DIBCO 2017 dataset. The published DIBCO 2017 dataset is very small in number with only 85 training images and 20 test images. Because of insufficient DIBCO data, DIBCO allows the use of other datasets to train and test models. Therefore we were able to use COCO-Text images to train the proposed model. The number of COCO-Text images used to learn the initial modules is 10,000. Only the DIBCO 2017 dataset is used to train the final model after cascading the modules. A total of 503 images is used for training which consisted of the 85 images provided DIBCO and 418 cropped derivations of them. We used the 20 test images from DIBCO 2017 to verify the binarization performance of the final model.

B. Results and discussion

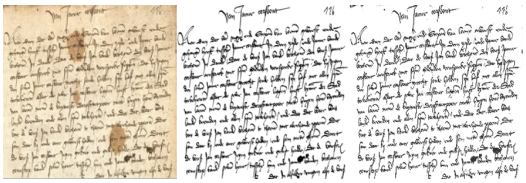
First, we evaluate the proposed model by comparing non-pre-trained model. We should elicit the importance of using pre-trained modular U-Nets in the proposed model. Also, U-Net itself has a robust image-to-image processing ability, thus we want to demonstrate the importance of the proposed method by comparing and verifying results. Fig. 5 shows differences among the proposed model, non-pre-trained cascaded model, and cascaded model without inter-module skip-connections. We find that using pre-trained modules with COCO-Text and using inter-module skip-connections among modules contribute greatly to the optimization of the entire model. Fig. 6 shows the binarization results of the proposed model according to the type of noise. In a single U-Net (DIBCO 2017 winners team's method), there is a problem that binarization cannot be performed at the portion where stains overlap with a character region. It is

Figure 5: Results indicating why the pre-trained modules and inter-module skip-connections should be used. Images in the rows from top to bottom correspond to the original images, the ground truth images, the result images of the proposed model, the result images of the non-pre-trained cascaded model, and the result images of the modular cascaded model without inter-module skip-connections.

confirmed that the proposed model is successful in removing stains. The figure demonstrates that the proposed method is effective in handling the different types of noise.

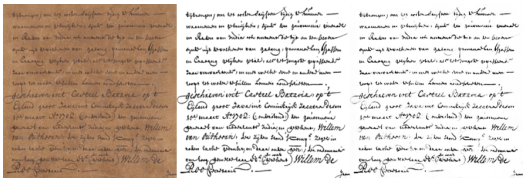
We use the evaluation tool provided by DIBCO to quantitatively verify the binarization performance of the proposed model. The evaluation tool compares the ground truth image with the binarization result using four measures. These measures consist of F-measure (FM), pseudo-FMeasure (F_{ps}), Peak Signal-to-Noise Ratio ($PSNR$), and Distance Reciprocal Distortion Metric (DRD).

FM is a harmonic mean value of precision and recall, which is a measure that can verify the precision and recall of the model to verify at once. F_{ps} has been proposed to overcome the problems of F-measure evaluation criteria for

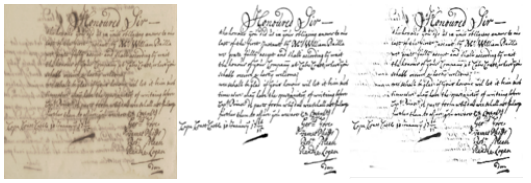


(a) Stained image

(b) Blob-noise image



(c) Low contrast image



(d) Bleed-through noise image

Figure 6: Binarization results of the proposed model according to the type of noise. Images in the columns from left to right correspond to the original images, the ground truth images, and the result images.

binarization [25]. F_{ps} is a version of FM with weighted $Precision$ and $Recall$ calculations designed to evaluate the performance according to the importance of the foreground pixels constituting the character in text detection, so that images having the same F-Measure value can also compare the binarization performance results.

$PSNR$ is a method of evaluating the difference between the detection result and the ground truth image. DRD has been used to measure the visual distortion of binary document images [26]. It correlates properly with human visual perception and measures a distortion of inverted every pixel. It is defined by:

$$DRD = \frac{\sum_{k=1}^{S} DRD_k}{NUBN}, \quad (1)$$

for S flipped pixels where $NUBN$ is the number of non-uniform 8×8 blocks in the ground truth image. DRD_k is the distortion of the k-th flipped pixel using a 5×5 normalized weight matrix as in [26].

We apply the results of the proposed model in the same way as the criteria to be evaluated by DIBCO. We compared the binarization system results published by DIBCO 2017 with the DIBCO evaluation method. Table I shows that comparison of the results of the proposed model with the results of DIBCO 2017 and 2018 winner teams. We show significant performance differences in the results of performance verification with DIBCO 2017 and 2018 participating teams. We also show that the results of the proposed model without pre-trained module and inter-module skip-connections. We obtain meaningful results for the need to use them. In addition, as the number of modules used in the model configuration increased, the binarization performance also improved. Finally, Fig. 7 shows the binarization result images of the proposed model and the binarization result images of the DIBCO 2017 winning team. Fig. 7 shows the result image according to the number of modules used in the cascaded model. This shows that the added modules were able to overcome the noise problems.

Table I: Comparative Results on DIBCO 2017

Method	FM	F_{ps}	$PSNR$	DRD
Proposed	**91.57**	**93.55**	15.85	**2.92**
Proposed without pre-training	81.15	82.91	13.40	8.23
Proposed without skip-connect.	87.07	89.29	15.05	4.98
H-DIBCO 2018 winner [16]	89.37	90.17	17.99	5.51
DIBCO 2017 winner [7]	91.04	92.86	**18.28**	3.40
Sauvola and Pietikäien [19]	86.61	88.67	17.80	5.56
Otsu [17]	83.52	86.85	16.42	7.49

V. Conclusion

This paper proposes a document binarization system using cascaded modular U-Nets. The proposed model is made up of four modular U-Nets trained for different image processing tasks and shows robust binarization performance even in environments with few annotated images. Through the use of effective inter and intra skip-connections, the U-Nets are able to transmit information between each other to construct a state-of-the-art model. We hope that this study of cascaded modular U-Nets will be applied not only to document binarization but also to other domains such as semantic segmentation and other image-to-image applications.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP17K19402 and JP17H06100.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Comput. Sci.* Springer, 2015, pp. 234–241.

[2] T. Lan, Y. Li, J. K. Murugi, Y. Ding, and Z. Qin, "RUN: residual U-Net for computer-aided detection of pulmonary nodules without candidate selection," *arXiv preprint arXiv:1805.11856*, 2018.

[3] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

Figure 7: Comparison with the proposed and the DIBCO 2017 winner's results. Images in the columns from left to right correspond to the original images, the ground truth images, the result images of the existing image processing methods applied to the proposed model, the result images of the proposed model by using three modular U-Nets, the result images of the DIBCO 2017 winners team, and the result images of the proposed method.

[4] A. Sevastopolsky, S. Drapak, K. Kiselev, B. M. Snyder, and A. Georgievskaya, "Stack-u-net: Refinement network for image segmentation on the example of optic disc and cup," *arXiv preprint arXiv:1804.11294*, 2018.

[5] S. Shah, P. Ghosh, L. S. Davis, and T. Goldstein, "Stacked u-nets: a no-frills approach to natural image segmentation," *arXiv preprint arXiv:1804.10343*, 2018.

[6] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Chellappa, "Stacked u-nets for ground material segmentation in remote sensing imagery," in *Conf. Comput. Vision and Pattern Recogn. Workshops*, 2018.

[7] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 competition on document image binarization (DIBCO 2017)," in *Int. Conf. Document Analysis and Recog.*, 2017.

[8] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Int. Conf. Mach. Learning*, 2013, pp. 1058–1066.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.

[10] S. Uchida, S. Ide, B. K. Iwana, and A. Zhu, "A further step to perfect accuracy by training CNN with larger data," in *Int. Conf. Frontiers in Handwriting Recog.*, 2016, pp. 405–410.

[11] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *Int. Conf. Pattern Recog.*, 2014.

[12] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *Int. Conf. Document Analysis and Recog.*, 2017.

[13] C. Wick and F. Puppe, "Fully convolutional neural networks for page segmentation of historical document images," in *Int. Workshop Document Analysis Sys.*, 2018.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Conf. Comput. Vision and Pattern Recog.*, 2015, pp. 3431–3440.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018)," in *Int. Conf. Frontiers in Handwriting Recog.*, 2018.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Sys., Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, jan 1979.

[18] S. Lu and C. Tan, "Binarization of badly illuminated document images through shading estimation and compensation," in *Int. Conf. Document Analysis and Recog.*, 2007.

[19] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recog.*, vol. 33, no. 2, pp. 225–236, 2000.

[20] X. Peng, H. Cao, and P. Natarajan, "Using convolutional encoder-decoder for document image binarization," in *Int. Conf. Document Analysis and Recog.*, 2017.

[21] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Conf. Comput. Vision and Pattern Recog.*, 2010.

[22] N. R. Howe, "Document binarization with automatic parameter tuning," *Int. J. Document Analysis and Recog.*, vol. 16, no. 3, pp. 247–258, 2012.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conf. Comput. Vision*, 2014, pp. 740–755.

[24] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems*. Elsevier, 1994, pp. 474–485.

[25] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance evaluation methodology for historical document image binarization," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 595–609, 2013.

[26] H. Lu, A. Kot, and Y. Shi, "Distance-reciprocal distortion measure for binary document images," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 228–231, 2004.