

Complex Image Transformations by Neural Networks

姜, 碩竣

<https://hdl.handle.net/2324/4475152>

出版情報 : Kyushu University, 2020, 博士 (学術), 課程博士
バージョン :
権利関係 :



**Complex Image Transformations
by Neural Networks**

Sougjun Kang

Department of Advanced Information Technology
Doctor of Philosophy

KYUSHU UNIVERSITY

February 2021

Complex Image Transformations

by Neural Networks

Sougun Kang

Department of Advanced Information Technology
February 2021
Doctor of Philosophy

Abstract

U-Net, which was suggested for precise pixel-wise segmentation of biomedical images, is one of the excellent image-to-image transformation models. Due to its effective pixel-wise localization representation capability, U-Net is widely used in various research fields. There are lots of studies for improving the performance of U-Net based models or training U-Net based models with insufficient data. In this paper, two novel methods, which are based on U-Net, are introduced. First, cascading modular U-Nets (CMU-Nets) are suggested to train the model by using a limited number of training data. CMU-Nets are organized with modularized U-Nets trained by a sufficient amount of different training data similar to the target data. The modularized pre-trained U-Nets are arranged by the optimal order and cascaded as an end-to-end system. The cascaded modularized U-Nets are connected by the proposed novel inter-module skip-connections for transferring pixel-wise localization information between the modules. Consequently, CMU-Nets are trained by the target data as a fine-tuning process. Second, for improving the performance of CMU-Nets by attention mechanism, an effective improvement method that adopts Squeeze and Excitation network (SE-Net) is introduced in CMU-Nets. The performance of conventional CNN-based models can be improved by unconstrained SE-Net utilization. However, CMU-Nets comprise pre-trained U-Net modules and are connected by the inter-module skip-connections, a novel method of SE-Net utilization should be suggested. Therefore, the proposed method, which can be called Attention CMU-Nets (ACMU-Nets), is suggested by reflecting these characteristics of CMU-Nets. To demonstrate the performance of these two proposed methods, diverse quantitative and qualitative analysis results are introduced. Since these two proposed methods are designed to address complex image transformation issues, the methods are trained by using the Document Image Binarization Competition (DIBCO) and Handwritten DIBCO (H-DIBCO) dataset for complex historical document image binarization.

Thesis Supervisor: Seiichi Uchida
Title: Professor

Acknowledgments

It is a great honor for me to write this acknowledgment. Because of your love, I can finish my doctoral course at Kyushu University that made me a more valuable person. My Ph.D. life at Kyushu University is the most precious memories and will be unforgettable until the day of death. I really appreciate everyone who has supported me a lot.

Above all, I am grateful to Prof. Uchida, who was my father, friend, senior, and supervisor for me during my three years of challenging Ph.D. I never forget the day I first met you. I thought that my abilities were not enough to be with you, but I was so happy when you recognized my willingness to work with you and told me to do research together in your lab. Even though I have put in a lot of effort to live up to your expectations, I still do not seem to be able to meet your expectations. Even though my Ph.D. is finished, I will never forget your tears for students as well as your dedication to research. If someone asks me ‘what is the greatest luck in your life so far’, I will undoubtedly say ‘my doctoral life with you’. Thanks for your great teaching.

Next, thanks to my best friend and senior, Associate Prof. Iwana. You made me a completely different person. I have done nothing for you, but you have spent your precious time with me and shared your knowledge. I am convinced that I can finish my tough Ph.D. program since I was with you. I am grateful not only for your research support but also for your emotional exchange. Thanks to you, I was not lonely during my Ph.D. Although we have to part now, I look forward to the day I can meet you again. You are my eternal friend and brother.

Thank you to the members of the Human Interface Lab who have been with me for 3 years. Thanks to you, I learned and experienced a lot and felt happy. I would like to express my gratitude to Associate Prof. Bise, Assistant Prof. Suehiro, Assistant Prof. Hayashi, Prof. Ohyama, Mrs. Maeda, Mrs. Fukutomi, and all students. Thanks to you, I was able to focus on my research and finish my Ph.D. I’m leaving the lab now,

but I look forward to the day we meet again.

Finally, I would like to express my gratitude to my greatest families in the world. I have given small hope and big disappointment every time to my parents. However, my father and mother have always supported me and loved me. I cannot dare understand and repay the love you gave me, but I will live hard so that you can be happy for me. Also, I would like to express my gratitude to my wife for her constant support in my doctoral life. Since I know your hard work better than anyone else, I will never forget your dedication. I would like to express my gratitude to Joonho Lee, Jinho Lee, Jae-Kwan Lee, and my family who have always cheered for me.

I think our lives are living to find one happiness in tens of thousands of pains. Especially since 2020, because of COVID-19, we are living in pain. However, pain in our lives makes us stronger and more valuable the happiness that will surely come to us. I will live without forgetting today to become a person who can give humanity one small happiness. Thank you.

Contents

1	Introduction	15
1.1	Background	15
1.2	Motivation	17
1.2.1	Importance of Sufficient Training Data in ANN-based Systems	17
1.2.2	Robust Image Transformation of Image-to-image CNN at The Test Phase	19
1.3	Purpose	20
1.3.1	Complex Image Transformation with Insufficient Image Data by Cascading Pre-trained U-Net Modules	21
1.3.2	Performance Improvement with Attention Mechanism at The Test Phase	22
1.4	Contribution	23
1.5	Organization of Thesis	24
I	Complex Image Transformation with Insufficient Data	26
2	Background and Related Work	27
2.1	Image Analysis with CNNs	27
2.1.1	Deep CNNs	27
2.1.2	Image-to-image transformation by CNNs	28
2.1.3	U-Net for Precise Image Segmentation	31

2.1.4	Data Shortage Issues and Solutions in U-Net	32
2.2	Document Image Analysis	34
2.2.1	Rule-based Document Image Analysis Systems	34
2.2.2	Neural Network-based Document Image Analysis Systems	35
2.3	Related Work	36
2.3.1	U-Nets and Stacked U-Nets for Complex Image Transformation	36
2.3.2	Document Image Binarization Systems	37
2.4	Summary	40
3	Cascading Modular U-Nets (CMU-Nets)	43
3.1	Pre-trained U-Net Modules in CMU-Nets	46
3.2	Cascading Structure in CMU-Nets	46
3.3	Selective Inter-module Skip-connections in CMU-Nets	48
4	Experiment for Document Image Binarization	51
4.1	Experiment Setup	51
4.1.1	Dataset for Fine-tuning	51
4.1.2	Dataset for Pre-training	52
4.1.3	Experimental Environment	54
4.2	Evaluation Results	54
4.2.1	Evaluation Criteria of DIBCO	54
4.2.2	Quantitative Evaluations	57
4.2.3	Qualitative Evaluations	60
4.2.4	Results for Types of Noise	60
4.3	Discussion	63
4.3.1	Analysis of the Inter-module Skip-connections	63
4.3.2	Analysis of Pre-trained U-Net Modules Utilization in CMU-Nets	67
4.3.3	On The Order of U-Net Modules	67
4.4	Summary	68

<i>CONTENTS</i>	9
II CMU-Nets Incorporating SE-Nets	71
5 Background and Related Work	73
5.1 Attention Mechanism	73
5.1.1 Attention Mechanism in Natural Language Processing	73
5.1.2 Attention Mechanism in Computer Vision	75
5.2 Related work	78
5.2.1 U-Net based Models with Attention Mechanism	78
5.2.2 U-Net based Models with SE-blocks	78
5.3 Summary	79
6 Attention CMU-Nets (ACMU-Nets)	81
6.1 Modified Training Method of Individual U-Net	83
6.2 SE-Nets in Individual U-Net Modules	83
6.3 SE-Nets in Inter-module Skip-connections	85
7 Experiment of ACMU-Nets	87
7.1 Experimental Setup	87
7.1.1 Dataset	87
7.1.2 Experimental Environment	88
7.1.3 Evaluation Criteria	88
7.2 Experimental Results	88
7.2.1 Quantitative Evaluations	88
7.2.2 Qualitative Evaluations	89
7.2.3 Effectiveness of SE-Nets in ACMU-Nets	89
7.3 Summary	92
8 Conclusion	95
8.1 Conclusion	95
8.2 Future Work	97

List of Figures

1-1	The examples and correlation of image processing and computer vision tasks.	16
1-2	Examples of pixel-wise annotation insufficiency in public datasets. . .	18
1-3	Examples of input image variations that may degrade image-to-image transformation (binarization) performance.	19
1-4	The design process of CMU-Nets.	21
2-1	The architecture of LeNet5.	28
2-2	Examples of a difference between a CNN-based image-to-values model and image-to-image model.	29
2-3	The architecture of U-Net.	31
2-4	Overlap-tile strategy of U-Net	32
2-5	Examples of a threshold-based image processing technique's results. .	35
3-1	The difference between CMU-Nets and existing stacked U-Nets. . . .	44
3-2	Concept of CMU-Nets.	45
3-3	Examples of ordering pre-trained U-Net modules.	47
3-4	Visual examples of cascading pre-trained U-Net modules.	48
4-1	Examples of DIBCO and H-DIBCO datasets.	52
4-2	Examples of the MSCOCO images for the modularized U-Nets. . . .	53

4-3	Comparative results on DIBCO and H-DIBCO datasets from 2016 to 2018.	61
4-4	Comparative results on DIBCO and H-DIBCO datasets from 2009 to 2014.	62
4-5	Binarization results of the proposed model according to the type of noise.	64
4-6	Comparative results between the CMU-Nets without any inter-module skip-connections and a conventional single U-Nets.	65
4-7	Worse case of the CMU-Nets without inter-module skip-connections. .	65
4-8	The reason why the proposed method of the inter-module skip-connections should be used in CMU-Nets.	66
4-9	Results indicating why the pre-trained modules should be used. . . .	67
4-10	Output images according to the methods of using U-Net modules. . .	69
4-11	Result images of models arranged in a different order than the proposed method.	69
5-1	Example of seq2seq's sequence transformation process.	74
5-2	Example of attention mechanism process in NLP	74
5-3	Example of attention mechanism process in CNN-based models. . . .	76
5-4	The architecture of a SE-block.	77
6-1	Concept of ACMU-Nets.	82
7-1	Comparative results with the DIBCO 2017 winner's, CMU-Nets, and ACMU-Nets.	90
7-2	The output images of CMU-Nets with densely attached SE-blocks in the inter-module skip-connections.	91
7-3	Effectiveness of SE-blocks between CMU-Nets and randomly initialized CMU-Nets.	93

List of Tables

4.1	Comparative Results on DIBCO 2009 and H-DIBCO 2010	57
4.2	Comparative Results on DIBCO 2011	58
4.3	Comparative Results on H-DIBCO 2012 and DIBCO 2013	58
4.4	Comparative Results on H-DIBCO 2014	58
4.5	Comparative Results on H-DIBCO 2016, 2018 and DIBCO 2017 . . .	59
4.6	Comparative results according to differences of inter-module skip-connections by using the DIBCO 2018 dataset for evaluation.	66
7.1	Comparative Results on DIBCO 2017	89
7.2	Results of the proposed method with the revised training method on DIBCO 2017	89

Chapter 1

Introduction

1.1 Background

Image-informatics is a research area to analyze, generate, and transform images for various purposes and can be roughly divided into two subareas: image processing and computer vision. Image processing is a discipline that studies the mathematical signal processing techniques of images, such as image quality enhancement [1, 2], geometric transformation [3, 4], color transformation [5, 6], reconstruction [7, 8], compression [9, 10], and comprehension [11, 12]. In contrast, computer vision is research area to extract useful information into an image for conducting specific purposes such as object recognition [13, 14], detection [15, 16, 17], and tracking [18, 19].

Early systems for image processing and computer vision were designed based on specific rules, such as setting optimized threshold values or adjusting the proper size of filters [20, 21]. In rule-based systems, since setting optimized rules for proper results is mightily important, these systems require technical know-how to set these threshold values or filter size. Even though these heuristic rule-based systems can often operate image processing or computer vision tasks properly [23, 24], the performance of these rule-based systems can be changed easily by the conditions [22]. Furthermore, as shown in Fig. 1-1, since recent image-processing and computer vision tasks have been

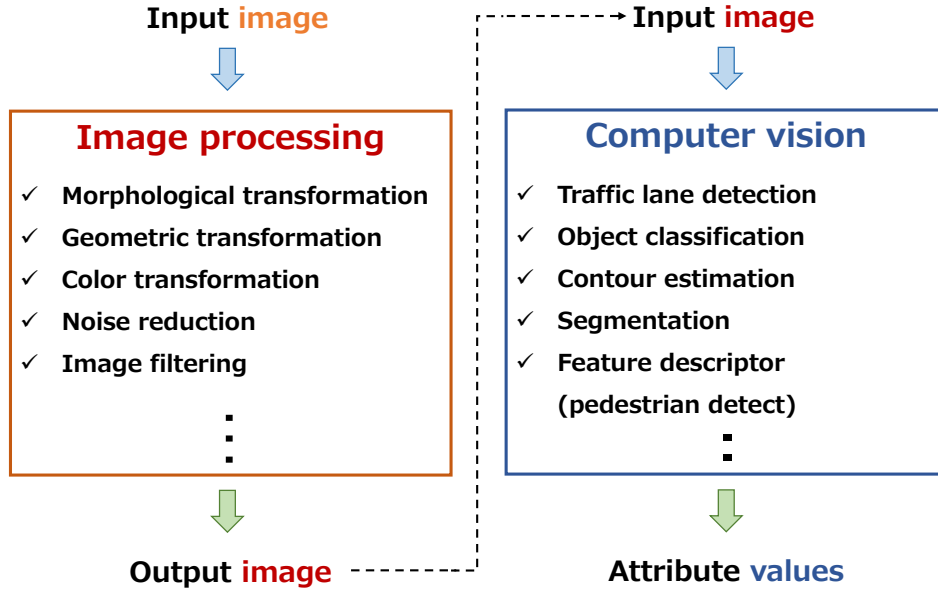


Figure 1-1: The examples and correlation of image processing and computer vision tasks. Early tasks had been conducted separately. However, since recent tasks have been more sophisticated, lots of their algorithms have required and been conducted simultaneously.

more sophisticated, the number of required algorithms for solving the complicated image-processing and computer vision tasks has increased. These requirements make rule-based systems more heuristic and a complicated structure.

After introduction of Artificial Neural Networks (ANNs), these heuristic rule-based systems have been replaced by various ANN-based systems [25, 26]. ANN-based systems do not need to set an optimized threshold value or finding the proper size of filters for generating the desired output since the output of ANN-based systems can be determined by sufficient training with image data. Moreover, since these ANN-based systems can be trained by using lots of training images in various conditions, these systems more adaptive compared with the existing rule-based systems [27].

1.2 Motivation

1.2.1 Importance of Sufficient Training Data in ANN-based Systems

Even though ANN-based systems are more powerful and adaptive than existing rule-based systems, a sufficient amount of high-quality annotated training data is required to train these ANN-based systems effectively [65]. For example, when we want to train Convolutional Neural Networks (CNNs) for vehicle detection, a large amount of vehicle images with annotation information like the position values of vehicles is required as well as input images. However, it is difficult for individuals to create a sufficient amount of annotations since generating proper annotations has several issues such as time-consuming, label-ambiguity, and requirement of expert knowledge.

Therefore, researchers have developed their CNN-based systems by using public datasets with accurate annotation, such as MNIST and ImageNet dataset [25], to avoid the shortage problem of training data [66, 47, 67]. These public datasets help researchers optimize the model training of their CNN models. However, since those datasets are prepared for classification tasks, they only help the classification systems and are not useful for different tasks. For example, to train CNN models for pixel-wise image-to-image transformation such as a semantic segmentation task, those datasets are useless. Instead, training image datasets with pixel-wise annotations are required.

Compared to the case of generating annotations for classification like ImageNet, generating pixel-wise annotations for image-to-image transformation is time-consuming much more. Fig. 1-2 shows examples of training data insufficiency for pixel-wise image-to-image transformation. Compared with the ImageNet dataset that contains over 14 million images [25], datasets with annotated images such as the Document Image Binarization Competition (DIBCO) dataset contain only dozens to hundreds of pixel-wise annotated images. Even though we can consider to generate these pixel-wise annotations, it is not easy for typical images with millions of pixels. In addition,

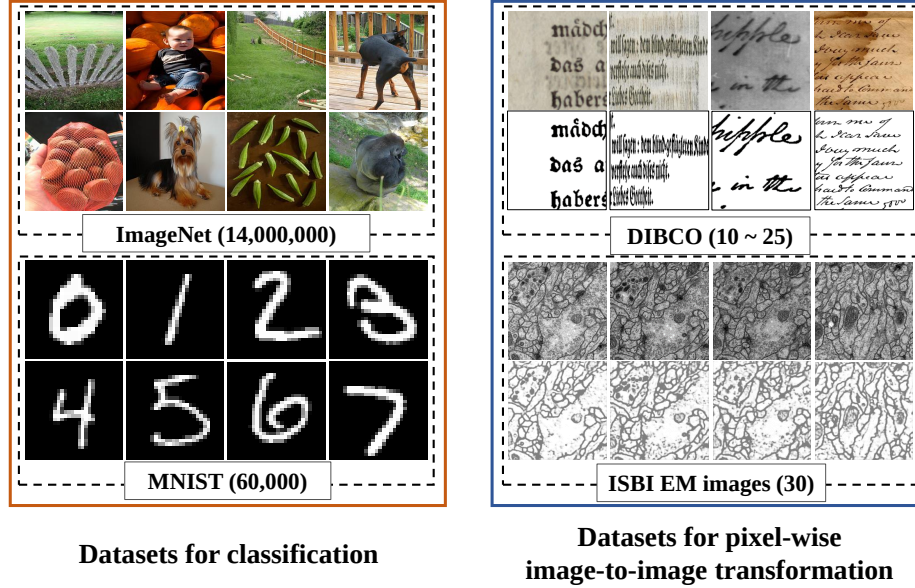


Figure 1-2: Examples of pixel-wise annotation insufficiency in public datasets. Datasets for classification like the ImageNet dataset contain a huge amount of annotations. However, datasets for pixel-wise image-to-image transformation like the DIBCO dataset contain only dozens of annotated images.

pixel-level annotation is equivalent to a pixel-level classification and thus it always encounters ambiguous cases where the correct class (e.g., foreground or background) is not clear. Moreover, some tasks, such as medical image binarization, require the participation of experts in the task and thus do not allow us to use crowdsourcing tasks.

To overcome this limitation, various data augmentation methods such as rotating and shifting are considered [71]. However, data augmentation methods are not an optimal method to solve the problem of insufficient data, since data augmentation is conducted by using existing training images [72].

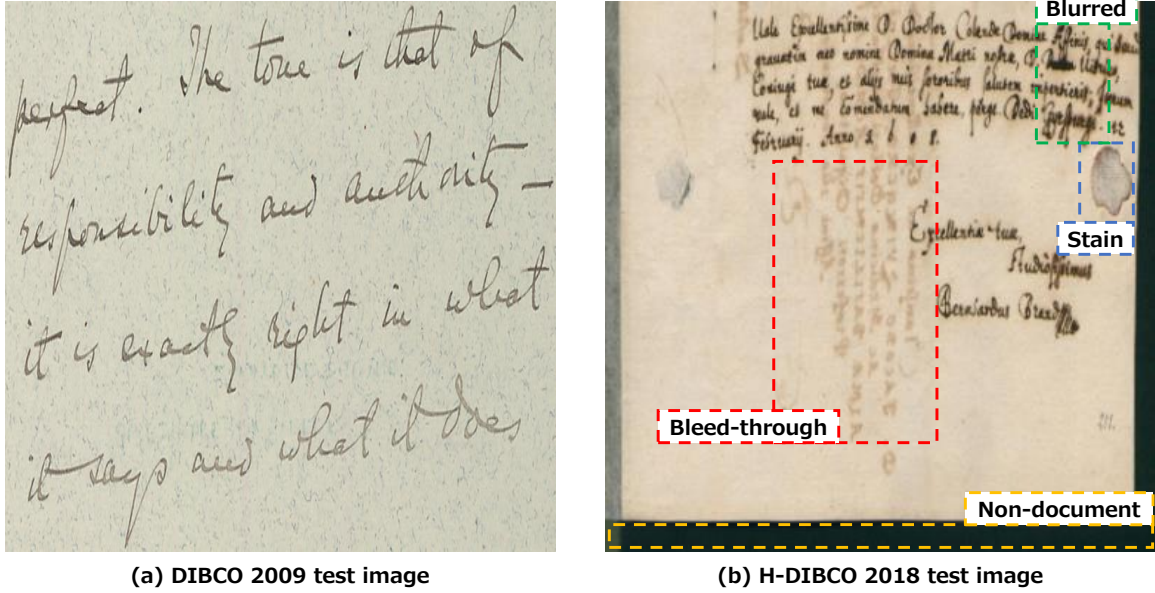


Figure 1-3: Examples of input image variations that may degrade image-to-image transformation (binarization) performance. Image (a) and (b) are DIBCO 2009 and H-DIBCO test images. Several noises like bleed-through and non-document noises are added compared with (a).

1.2.2 Robust Image Transformation of Image-to-image CNN at The Test Phase

Compared to image-to-values CNN systems, image-to-image CNN systems are more complicated to acquire training data and require more effort to complete their model training. Therefore, researchers expect their image-to-image CNN systems to show robust performance regardless of test data conditions. However, Fig. 1-3 demonstrates that these image-to-image CNN models should consider newly added noises in the recently published test images. Compared with Fig 1-3 (a), the H-DIBCO 2018 test image was exposed at various noises like bleed-through, stain, and non-document noises. These various noises can lead to the poor performance of well-trained image-to-image CNN systems at the test phase.

A straightforward remedy to make image-to-image CNN models robust to a large variation of input images is to prepare sufficient training images. However, as noted

in the previous section, the preparation of a sufficient number of annotated images is very difficult in practice. Data augmentation methods also have a limitation to deal with the unexpected variations in input images. In addition, these tasks are time-consuming and cannot be fundamental solutions for robust performance at the test phase.

To solve the problem effectively, researchers have come to realize that after model training is complete, fixed parameters must be effectively handled in the test phase [181]. Therefore, researchers have performed many studies to deal with the change of input image due to fixed parameters, among them, an attention mechanism has been applied to solve this problem [75].

1.3 Purpose

The purpose of this thesis is twofold: training an image-to-image transformation model with insufficient training data and improving the performance of the image-to-image transformation model with attention mechanisms. The two proposed models are based on U-Net [49] that was suggested for precise semantic segmentation of biomedical images since U-Net has been utilized in various image-to-image transformation tasks [174, 173, 55].

For the first purpose of this thesis, I proposed a novel U-Net based image-to-image CNN models called Cascading Modular U-Nets (CMU-Nets). CMU-Nets achieves robust image transformation with a limited number of training data. For the second purpose of this thesis, I proposed Attention CMU-Nets (ACMU-Nets), which introduces an attention mechanism to CMU-Nets. More specifically, Squeeze and Excitation network (SE-Net) [64] is introduced to ACMU-Nets to improve the complex image transformation performance.

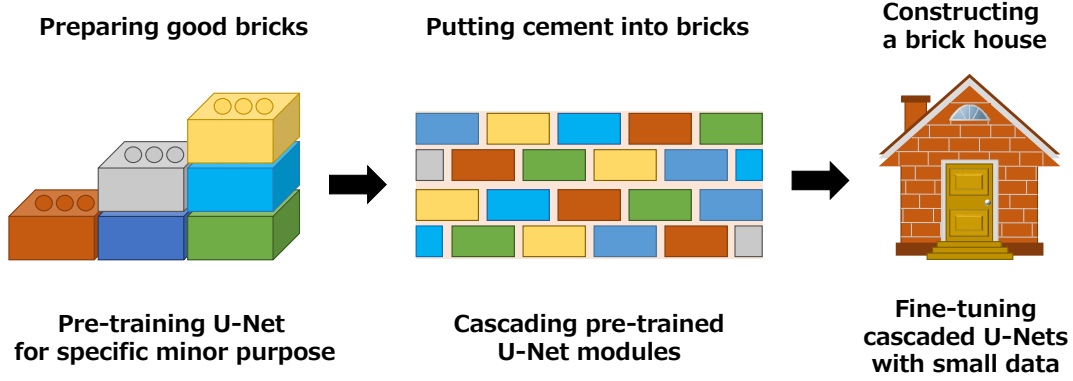


Figure 1-4: The design process of CMU-Nets can be illustrated as constructing a brick house.

1.3.1 Complex Image Transformation with Insufficient Image Data by Cascading Pre-trained U-Net Modules

For robust image-to-image transformation performance of U-Net based systems by using a limited number of training data, I proposed CMU-Nets in this thesis. If the training data are insufficient for effective model training, the direct end-to-end generation in image-to-image CNN models can affect performance degradation [114]. However, CMU-Nets adopt modularized pre-trained U-Nets for solving the data shortage problem and the cascading structure of the individual U-Net modules for conducting complex image transformation.

The design process of CMU-Nets can be illustrated as constructing a brick house like Fig. 1-4. In the first step of designing CMU-Nets, like preparing good bricks for constructing a brick house, individual U-Net modules are secured for specific minor purposes of a target task. CMU-Nets are verified the image transformation performance in historical document image binarization. Therefore, the individual U-Net modules are trained for the pre-processing of image binarization by using a sufficient amount of image data.

Next, the pre-trained modules are cascaded and connected like putting cement into the bricks. For conducting the target task of CMU-Nets, the pre-trained U-Net mod-

ules for minor purposes should be cascaded. Compared with the conventional deep-depth U-Net based models, additional connections between the used U-Net modules are required since CMU-Nets are organized with pre-trained U-Nets.

Finally, to optimize the cascaded model, fine-tuning is conducted with extremely small data. The cascaded U-Net modules should be fine-tuned by using a small amount of target data for conducting the target task.

The key ideas of CMU-Nets can be summarized as follows: 1) the decomposition of the complex target task into a cascade of simpler image-to-image transformation tasks and 2) the pre-trained of U-Nets for realizing each simple transformation task by a different (general) dataset with sufficient training data. Since these pre-trained U-Net modules can conduct image processing techniques like dilation and histogram-equalization, CMU-Nets, which are organized with these pre-trained U-Net modules, can conduct complex computer vision tasks that include these image processing techniques.

1.3.2 Performance Improvement with Attention Mechanism at The Test Phase

CNN-based models with attention mechanisms have shown significant performance improvement in various image processing [74]. Since CNN-based models with attention mechanisms can adjust (recalibrate) the parameters for effective feature extraction at the test phase, attention mechanisms can be utilized to design adaptive CNN models. Among these attention mechanisms, the powerful channel-wise recalibration ability of SE-Nets can improve the representation performance of the network by modeling the NN's interdependencies between the convolutional channel features [75]. Moreover, SE-Nets can be utilized easily since SE-Nets can be attached in a network unconstrainedly and do not burden parameter increase of a network largely.

Since SE-Nets have the strengths of effective performance improvement with simple attachments of SE-Nets, ACMU-Nets utilizes the SE-Nets for improving the per-

formance of CMU-Nets. As reviewed before, SE-Net is an attention mechanism and has a function to re-calibrate the important features adaptively and has given more accurate results in various image-to-image transformation tasks. Since CMU-Nets also aims to realize image-to-image transformation, we can expect a similar performance by the introduction of SE-Nets.

However, since CMU-Nets adopt two different strategies compared with the existing U-Net based models: pre-trained U-Net modules and connections between U-Net modules, conventional utilization of SE-Nets can be not proper to be adopted in CMU-Nets. Therefore, in ACMU-Nets, effective SE-Net utilization should be suggested for improving the model performance at the test phase.

1.4 Contribution

The focus of this thesis is to develop methods that are efficient model learning methods of image-to-image transformation model. According to the studies on CMU-Nets and ACMU-Nets, I confirmed that the studies make several contributions in complex image transformation.

The contributions of this thesis are as follow:

- To solve the problem of an insufficient amount of training samples, CMU-Nets were introduced. CMU-Nets adopted pre-trained U-Net modules and connections between U-Net modules to be trained with a limited number of target images. CMU-Nets achieved robust historical document image binarization as complex image transformation. Since CMU-Nets are organized with pre-trained U-Net modules for various image processing, CMU-Nets proved that a complex computer vision task like document image binarization can be conducted well. CMU-Nets proved that improving the model performance and training the model with an insufficient amount of data are possible simultaneously.
- For improving the performance of CMU-Nets, CMU-Nets incorporating SE-

Nets called ACMU-Nets are proposed. In the ACMU-Nets' research, the proper utilization of SE-Nets in CMU-Nets is suggested. Also, the modified training method of individual U-Net modules affects the performance improvement of ACMU-Nets.

- Qualitative and quantitative results of CMU-Nets are demonstrated in detail. I evaluated CMU-Nets by using all DIBCO and Handwritten DIBCO (H-DIBCO) datasets from 2009 to 2018 [76, 77]. Moreover, I evaluated CMU-Nets by various types of noise. By demonstrating various comparative results with several image binarization methods, I confirmed that CMU-Nets achieve the state-of-the-art performance of historical document image binarization.
- Qualitative and quantitative results of ACMU-Nets are also demonstrated. I experimentally demonstrate the value of SE-blocks in CMU-Nets and simply cascaded U-Nets to verify how SE-Nets can contribute to model performance. I evaluated ACMU-Nets by using DIBCO 2017 dataset [78]. I confirmed that ACMU-Nets achieve better performance in document image binarization compared with several image binarization systems including CMU-Nets.

1.5 Organization of Thesis

The remainder of this thesis is organized as follows:

Part I describes the complex image transformation of historical document image binarization. Chapter 2 provides the background and the related work of complex image transformation deep learning methods and document image processing. Chapter 3 demonstrates the detailed information of CMU-Nets that are novel cascaded pre-trained modularized U-Nets. Chapter 4 demonstrates experimental results with quantitative and qualitative results by using all DIBCO [79] and H-DIBCO [77] datasets.

Part II describes the complex image transformation of ACMU-Nets. Chapter 5 provides the background and related work. Chapter 6 introduces the detailed information of ACMU-Nets. Chapter 7 demonstrates evaluated quantitative and qualitative results by using DIBCO 2017 [78] dataset.

The conclusion of this thesis is drawn in Chapter 8.

Part I

Complex Image Transformation with Insufficient Data

Chapter 2

Background and Related Work

2.1 Image Analysis with CNNs

2.1.1 Deep CNNs

The first CNN model was introduced by LeCun et al. for handwritten zip code recognition [80], and then LeNet5 was suggested as a supplemented CNN model [28]. In Fig. 2-1 LeNet5 consists of two convolutional layers, two sub-sampling layers, and two fully-connected layers. After the introduction of LeNet5, with various studies for improving the performance of CNN [81, 82], Krizhevsky et al. suggested an improved architecture of a CNN model, called AlexNet [25]. AlexNet utilized various strategies used in recent years such as Rectified Linear Unit (ReLU), dropout, overlapping pooling, Local Response Normalization (LRN), and data augmentation to improve CNN performance.

After the introduction of AlexNet, CNN studies have been conducted to develop deeper models for better performance [48]. Among them, GoogleNet [83] and VGGNet [47] focused to improve the performance of CNN through designing a deeper-depth architecture of CNN. In particular, VGGNet reduced the kernel size of the convolutional layer from 7×7 or 5×5 to 3×3 to design a deep-depth CNN model. Ac-

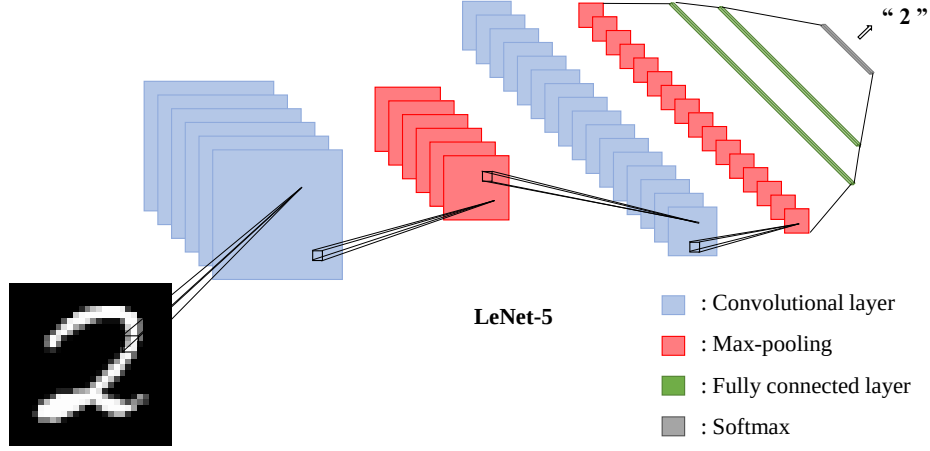


Figure 2-1: The architecture of LeNet5.

cording to the architectural changes, the network was made deeper and the amount of total weight used at the same time was reduced. Afterward, He et al. solved the performance deterioration caused by blindly increasing the depth of a CNN model by using the residual block [84]. These studies are still being used as base models for recent CNN-based models.

Since deep-depth CNNs have shown robust performance in object classification, numerous CNN-based models have been proposed for conducting diverse image processing tasks [29, 30]. CNN-based models have shown their powerful ability of image processing in various fields such as classification [31, 32], tracking [33, 34], image generation [35, 36], and semantic segmentation [37, 38]. Among them, CNN-based models for semantic segmentation enable to perform a pixel-wise image-to-image transformation since these models utilize the extracted features of an image to generate an image output [39, 40].

2.1.2 Image-to-image transformation by CNNs

Among various image processing studies, image-to-image transformation is a research field that is actively being performed from existing rule-based systems [41, 42] to recent CNN-based systems [43, 44]. These image-to-image transformation tasks include

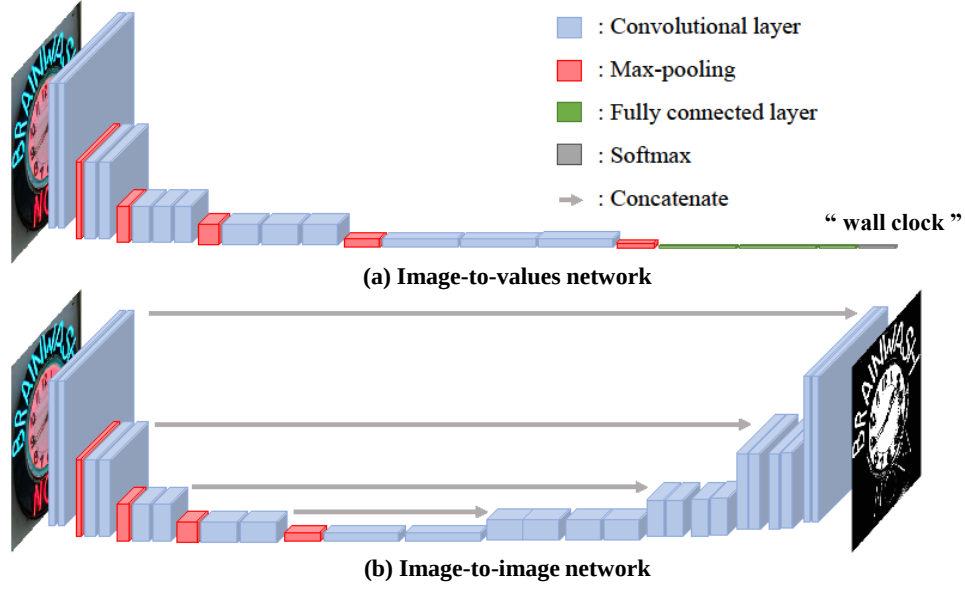


Figure 2-2: Examples of a difference between CNN-based image-to-values and image-to-image models. Fig. 2-2 (a) is an example of an image-to-values network, VGGNet [47]. Fig. 2-2 (b) is an example of an image-to-image network, U-Net. U-Net is based on VGGNet.

simple tasks such as morphological or geometrical image transformation as well as complex tasks such as instance or semantic segmentation. The existing rule-based systems can generate their output image by passive criteria through thresholds [45] or find their optimal threshold values for effective image transformation by calculating the pixel value of the image. In contrast, CNN-based models for image-to-image transformation can generate their output image by learning not only information on a single-pixel value but also global information such as attribute information. Therefore, CNN-based models for image-to-image transformation have shown robust image transformation, although their image transformation tasks include complicated issues [46].

Early CNN-based models took image-to-values transformation structures, such as image classification and detection [47, 48]. Since image-to-values CNN models capture and learn proper features in an image well, these image-to-values CNN models can output attribute information like class labels of a classification task as shown in

Fig 2-2. However, for an image-to-image transformation task such as segmentation and generation, the architecture of image-to-values CNN models should be revised. Therefore, robust image-to-image models have been introduced such as FCNs [39], SegNet [37], and U-Net [49].

Compared with image-to-values CNN models, image-to-image transformation models are designed to produce an image output with attribute information. As shown in Fig 2-2, an image-to-image network can output an image compared with the attribute outputs of image-to-values networks like class labels. Due to the strength of these image-to-image transformation models, image-to-image transformation models such as U-Net are being developed to perform complex image transformation, such as medical image processing [50, 51], document image binarization [52, 53], and autonomous driving cognitive systems [54, 55].

The purpose of semantic segmentation is to classify all pixels in an image into a pre-defined number of classes, which is called dense prediction because it makes predictions for all pixels. Existing CNN models lose detailed location information due to the number of parameters such as AlexNet and VGGNet, and the use of layers that reduce overtaking. To solve the problem of loss of location information, Long et al. proposed a Fully Convolutional Network (FCN) that uses convolutional layers instead of fully connected layers [39].

For effective semantic segmentation, FCN has applied two techniques different from existing CNN-based models. The first is 1×1 convolution process instead of a fully-connected layer and the second is a deconvolution process using skip-connection. FCN gained the ability to process the entire image at once while maintaining spatial information by using 1×1 convolution. In addition, FCN can perform effective pixel-wise prediction through an up-sampling process and a process of combining pooled features through skip-connection to restore the reduced feature map that has been subjected to several convolution and pooling processes to the original image size. These FCN studies have had a great influence on the semantic segmentation and

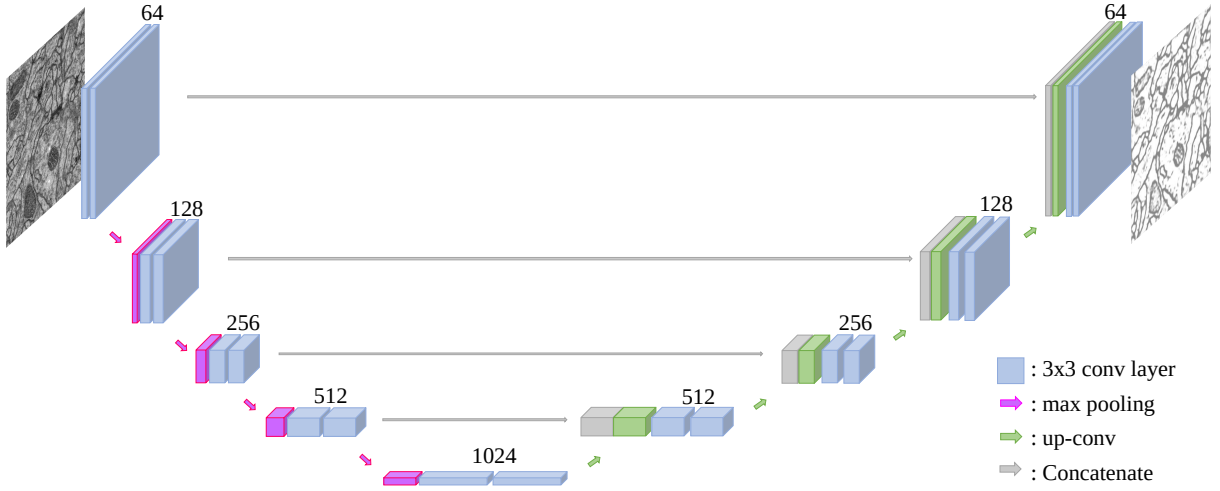


Figure 2-3: The architecture of U-Net.

image-to-image transformation task.

2.1.3 U-Net for Precise Image Segmentation

After introducing FCN, a CNN-based model for precision biomedical image segmentation called U-Net was proposed by Ronneberger et al [49]. In Fig. 2-3, the structure of U-Net is demonstrated. U-Net's contracting path is organized by using VGGNet that includes 3×3 convolution layers, 2×2 max-pooling operations, and ReLU as activation functions. That is, 3×3 Conv $\rightarrow$ ReLU $\rightarrow 2 \times 2$ max-pooling $\rightarrow 3 \times 3$ Conv $\rightarrow$ ReLU $\rightarrow 2 \times 2$ max-pooling. On the other hand, the expanding path of U-Net conducts up-sampling operations with concatenating between the upsampled and downsampled feature map. After that, the last part of U-Net performs a 1×1 convolution operation to map 64 component feature vectors to the desired number of classes.

U-Net has two different important points compared to FCN. The first is that the number of feature channels is large during the upsampling process. This means that context can be propagated to successive resolution layers. The second is that possible to conduct robust segmentation by using the overlap-tile technique. In Fig. 2-4, it describes the overlap-tile technique of U-Net, where tiles of different sizes overlap each

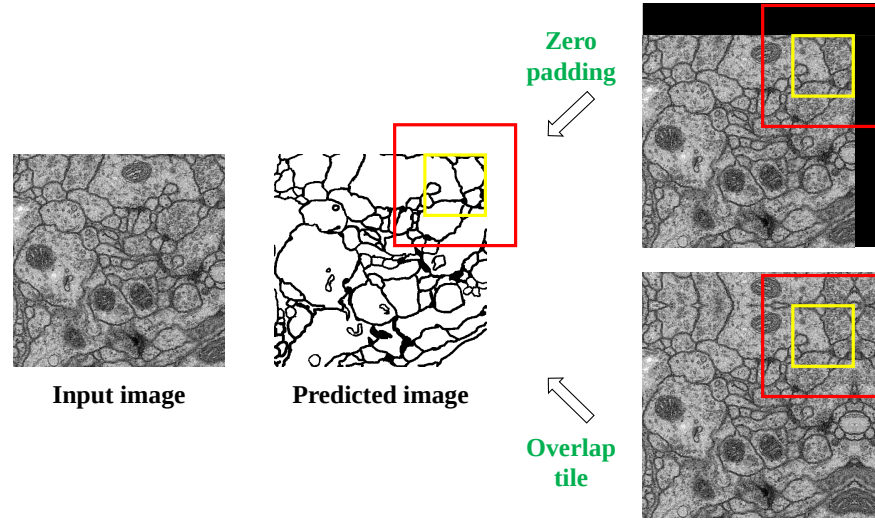


Figure 2-4: Overlap-tile strategy of U-Net. Red bounding box is patch as an input and yellow box is prediction of segmentation.

other to perform prediction. In this process, missing data is extrapolated through the mirroring method.

In the end, U-Net showed effective semantic segmentation performance in the biomedical field by applying the up-sampling architecture and the skip-connection of an extended concept than FCNs. These strengths of U-Net have led to various U-Net based models being proposed, and are used not only in the field of image segmentation but also in various image transformation tasks.

2.1.4 Data Shortage Issues and Solutions in U-Net

Since U-Net is also a CNN-based model, the model training of U-Net should be performed using image data. CNN-based models basically face the problems of underfitting or overfitting when researchers optimize model training by using insufficient training data. Even though these problems can be solved by using the validation process, dropout, etc., securing sufficient high-quality data and performing model training with using these sufficient data can solve these problems effectively. However, obtaining enough high-quality data is a very difficult task. For example, to

perform U-Net training for comprehensive document image analysis, various types of document images from printed document images to handwritten document images must be secured. In addition, time-consuming manual work is sometimes required to create ground truth images of the acquired images. In the end, the problem of securing sufficient data led researchers to develop new U-Net models, and new approaches to overcome the data problems.

Data augmentation is the method of increasing the amount of learning data through various physical changes [71]. It crops, rotates, flips, translates, and resizes the image to increase the amount of data by giving morphological changes, and it also increases the total amount of data by changing information inside the image such as color modification and noise addition. These data augmentation strategies have been still considered since it is an easy way to increase insufficient image data. Even though the data augmentation strategy is an effective data utilization method on CNN, the data augmentation method is limitedly used since it is performed using only the given data. Furthermore, excessive changes in the process of data augmentation may cause deterioration in learning performance.

Transfer learning is a method of using the weight of a pre-trained model as an initial value [73]. When researchers want to train CNN-based models for specific purposes, training a randomly-initialized CNN-based model is a hard and time-consuming task. Therefore, researchers can utilize a well-trained CNN model as a pre-training process and train their CNN-based models by using their specific data for the purposes. In addition to the effect of shortening the overall training time, transfer learning is also used for model training using insufficient data due to the advantage of being able to easily apply optimal weights to the model. However, to perform transfer learning, it is easy to have the same structure as the model in which weights are to be used. Moreover, if the training data of the base model for weight transfer and the data for fine-tuning are not similar, the performance of the newly trained model may be degraded.

Domain adaptation is a method that enables learning even when there is no information on the target domain or is very sparse [87, 88, 89]. Domain adaptation is a method when the domain of the task, such as object recognition, is changed to a different domain (target domain) that is somewhat related but not the same from the previously learned source domain, and then transfers the trained information. The domain adaptation method also utilizes the method of using the domain of external source data like transfer learning, but it also trains the target data domain so that model training can be performed even in extreme situations.

2.2 Document Image Analysis

2.2.1 Rule-based Document Image Analysis Systems

With the development of computers and various scanning devices, documents can be easily captured as images, and various image processing technologies have been utilized for document image analysis [90]. The purpose of document image analysis is to recognize the text and graphic components of document images. Document image analysis research has been conducted in a variety of ways not only to improve the quality of document images [91] but also to recognize the components of document images [92]. Document image analysis techniques have been developed along with various image processing techniques [93]. For effective and efficient document image processing, various pre-processing processes are required [94, 95]. Since these pre-processing methods are applied as part of a single integrated document analysis system, they should be lighter and less complex than the main algorithms. Therefore, many preprocessing processes adopt filtering or thresholding methods to simply provide the converted image.

Even though these rule-based pre-processing technologies have been efficiently used in document analysis systems, they are often sensitive to user intervention. In Fig. 2-5, threshold-based image binarization techniques produce very different re-



Figure 2-5: Examples of a threshold-based image processing technique’s results. Images from upper left to lower right correspond to the original image, the binarized result controlled by the threshold value 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9.

sults depending on the threshold value. Therefore, early threshold-based technologies showed a huge preformation according to the threshold value set by the user. To compensate for the problem of these global threshold-based technologies, locally adaptive threshold-based technologies and hysteresis threshold-based technologies have been suggested. Although these threshold-based technologies are designed with a relatively simple structure, they have been used in recent document analysis systems since they show robust performance.

2.2.2 Neural Network-based Document Image Analysis Systems

Compared with rule-based document image analysis systems, neural network-based document image analysis systems can generate appropriate results with minimal user intervention [28, 58]. For robust document image transformation performance, neural networks only require a sufficient amount of high-quality training images. In addition, since neural networks have been improved tremendously [96], the neural network-based systems have been also improved significantly [97]. These improved document analysis systems can be adopted in various document analysis tasks such as page analysis [98], text recognition [99], and non-textual information recognition [100].

Early neural network-based document image analysis systems used early neural networks such as MLP [28]. Since then, CNN-based document analysis systems have had great success, e.g. achieving 100% accuracy on the NIST tax forms dataset or higher accuracy than existing fine image processing based feature extractors [97, 101]. These successes stimulated the emergence of more diverse CNN-based document analysis systems [66, 102], among which document analysis systems using image-to-image transformation networks were proposed [43, 103]. These image-to-image transformation networks based systems showed superior performance compared to previous systems, such as the DIBCO 2017 winner system [78]. In DIBCO 2017, Bezmaternykh et al. won by suggesting a U-Net based document image binarization system [104], and H-DIBCO 2018 also showed better binarization performance than the image processing based H-DIBCO 2018 winner system [105]. That is, well-trained image-to-image transformation models can replace great traditional document analysis systems.

2.3 Related Work

In this chapter, U-Nets, stacked U-Nets, and document image binarization systems are described briefly.

2.3.1 U-Nets and Stacked U-Nets for Complex Image Transformation

Due to the significant achievements of CNN models [106, 107, 108], lots of variants of CNN have been studied for tackling complex tasks. Among these variants, image-to-image transformation CNN-based models have shown significant semantic segmentation [49, 39, 37].

U-Net was proposed to deal with biomedical image segmentation, has been shown to have significant pixel-wise segmented results [49]. Due to the strength of U-Net, variants of an U-Net have been studied to transform the shape of the traditional U-

Net [109], to segment liver and tumor with densely connected DenseNet [110] and U-Net, to expand 2-dimension operation to 3-dimension operations [111], and to generate outputs by conditional U-Net [112].

To improve the performance of U-Net, studies on the utilization of cascaded U-Nets has been implemented [113, 114, 115]. In the case of [114], stacked U-Net (SUNet) is suggested for complex natural image segmentation. SUNet applied customized U-Net modules included outer residual connections. Therefore, SUNet cascades a large number of U-Net modules that are downsized the conventional U-Net. Sevastopolsky et al. suggested a novel U-Net-based model that diagnoses glaucoma effectually [113]. Their model comprises several stacked U-Nets by using the conventional U-Net and residual U-Net (RUN) blocks. They adopted a method of copying and cropping features by simply cascading the features of the contracting path, which is referred to as skip-connections. They also utilized a method of elastic deformation and weighted cross-entropy strategies to augment the data. Isensee et al. proposed a cascaded 3D U-Net architecture in the case of large size images [116]. They utilized cascaded 3D U-Nets for minimizing global contextual information loss since conventional U-Net adopted local feature extraction. Moreover, Tang et al. proposed coupled U-Nets (CU-Nets) for robust human pose estimation [117]. CU-Nets coupled each U-Net pair with their proposed coupling connections. CU-Nets achieved state-of-the-art performance with fewer parameters than other approaches.

2.3.2 Document Image Binarization Systems

Image binarization is to classify all pixels of an image by black and white. Since the binarized image can be utilized to detect features of objects included in the image such as character recognition, image binarization is known as a basic work for utilization in various applications of document image analysis. In rule-based image banarization methods, setting proper threshold values is required for robust image binarization [118].

Rule-based image binarization can be defined by three types: global fixed thresholding, locally adaptive thresholding, and hysteresis thresholding. Otsu's binarization method [20] is one of global fixed thresholding methods to find an optimal threshold value that minimizes intra-class variance between two classes or maximizes inter-class variance when pixels of an image are classified into two classes based on an optimal threshold value. Since a global fixed thresholding method utilize a single value to classify pixels of an image, binarizing an image with drastic changes in brightness such as high-contrast is a difficult task. Therefore, locally adaptive thresholding methods that calculate threshold values applied to different small areas of an image [22, 119] and hysteresis thresholding methods that are techniques in which one's own classification result can vary depending on the classification result of surrounding pixels [120] have been applied instead of global fixed thresholding methods. These rule-based image binarization methods had shown robust binarization performance, and thus these methods had been applied in historical document image binarization. Among them, various rule-based document image binarization systems have been introduced in DIBCO and H-DIBCO.

In 2009, the 1st DIBCO was held [79] and has released about 20 images almost every year since. According to DIBCO and H-DIBCO, various and novel binarization methods have been suggested [77, 78, 105]. Early document binarization systems in DIBCO and H-DIBCO were implemented by applying traditional image processing techniques [79, 76]. These studies focused on explicitly defining the optimal thresholds for separating foreground and background areas. Alternatively, there were studies that tackle automatically setting the thresholds [20, 22, 121]. However, these rule-based approaches have limitations. They cannot always be generalized to every task. In historical document image binarization, the binarization systems would have to handle unexpected cases such as bleed-through, fading ink, and paper aging. Moreover, they sometimes rely on manually-tuned parameters.

To overcome the problems of traditional image processing techniques, several doc-

ument image binarization systems in recent DIBCO and H-DIBCO have been suggested that were based on ANNs. In DIBCO 2017, the winner team suggested a U-Net based document image binarization systems [104]. The winner established two main strategies in their image binarization system. First, they put effort into augmenting the data for model training due to the limitations of contest images provided by DIBCO and H-DIBCO. They acquired thousands of document images similar to contest images and used them for model training. Second, they implemented a single robust U-Net through optimization techniques, various reinforcement strategies, and a tiling coverage method instead of a sliding window method in conventional U-Nets.

CNN-based Document binarization methods have been also introduced in the various international conferences as well as DIBCO. Akbari et al. suggested two-channels convolutional networks are trained to classify foreground pixels [122]. Their two-channel models utilize original images and its binarized images by using the structural symmetric pixels (SSPs) [123]. Peng et al. suggested multi-resolution attention FCN [39] model [124]. Their method infers the probability of text areas by utilizing a multi-resolution attention model for overcoming conventional document binarization methods that adopt hand-craft and empirical processes. In addition, Mondal et al. suggested a 2D morphological network by concatenating feature maps that are generated by dilated and eroded layers [125]. Hollaus et al. suggested a multispectral document binarization system inspired by the U-Net [126]. Their model utilized pre-trained ResNet-50 [127] in the contracting path of the U-Net for reducing training time and enhancing segmentation performance. According to these studies, CNNs based document binarization systems have been demonstrated that they can replace the techniques based on image processing technologies.

However, the limited historical document binarization ground truth is a challenge for traditional deep learning architectures. Because of this difficulty, there are still lots of systems based on traditional image processing technologies in recent DIBCO. For example, Mondal et al. suggested an image binarization system based on fast

Fuzzy C-Means with image processing techniques [128]. Since their system adopts image processing techniques such as stroke width calculation, SSPs, and Niblack's thresholding [129], the system is a learning-free system by contrast with ANNs-based systems. Also, the winner of H-DIBCO 2018 [105] is composed of morphological transformations, stroke width transform (SWT) [130], and Howe's binarization [131] to divide foreground and background pixels. Their system adopted image post-processing to decrease noise effect and conserve text stroke connectivity. Even though the H-DIBCO 2018 winner team did not utilize an ANN scheme for building their system, they clinched the victory of H-DIBCO 2018.

2.4 Summary

Rule-based image processing systems have shown good performance in various image analysis, although these systems adopt relatively simple structures. These rule-based systems can be optimized by the user's know-how to set the optimum threshold value or the filter size. However, the outputs of these rule-based systems can vary with the set threshold values. Therefore, these rule-based systems aimed to find the optimal threshold values such as adaptive and hysteresis thresholding methods.

In contrast, since CNN-based models can easily produce great results with minimal user intervention, CNNs have been used for various image analysis. CNN started from a very simple structure like LeNet and has been actively utilized not only for fundamental and theoretical studies such as effective activation function and deeper architecture, but also for complex image analysis applications such as classification, tracking, and segmentation. Due to these significant achievements of CNN-based models, recent CNN-based models have shown cognitive abilities beyond humans in many fields.

CNN-based models not only have been improved their performance, but to perform tasks that were not possible before. Among them, U-Net has been proposed

for image-to-image transformation. Since U-Net have shown significant a pixel-wise prediction ability in biomedical image analysis, various improved U-Net variants have been suggested [109, 111]. Among various U-Net studies, studies of stacked U-Nets have been suggested for enhancing U-Net's performance [113, 114, 132].

In the document image binarization research, like general image processing systems, rule-based systems [20, 22] to U-Net-based systems [133] have been proposed. Among them, U-Net-based image binarization systems have achieved significant binarization performance in DIBCO and H-DIBCO [78, 104].

Chapter 3

Cascading Modular U-Nets (CMU-Nets)

CMU-Nets have several differences from existing stacked U-Net models. In Fig. 3-1, I illustrated the differences between the existing stacked models and the proposed model visually.

- Compared with the existing stacked U-Nets studies, CMU-Nets are designed for a different purpose. Existing stacked or cascaded U-Net research aims to enhance the model performance by organically connecting two or more U-Nets. However, the proposed model is designed to improve the performance with a limited number of training images. Therefore, CMU-Nets are comprised of pre-trained U-Net modules, whereas the existing stacked U-Net models are trained with sufficient samples from scratch.
- Compared with the individual U-Net module in the existing stacked U-Nets, the role of an individual U-Net is different. Each U-Net in the existing stacked U-Nets has no specific role by itself, whereas each U-Net in CMU-Nets is trained to solve a specific image processing task. Although, CMU-Nets are finally optimized with fine-tuning.

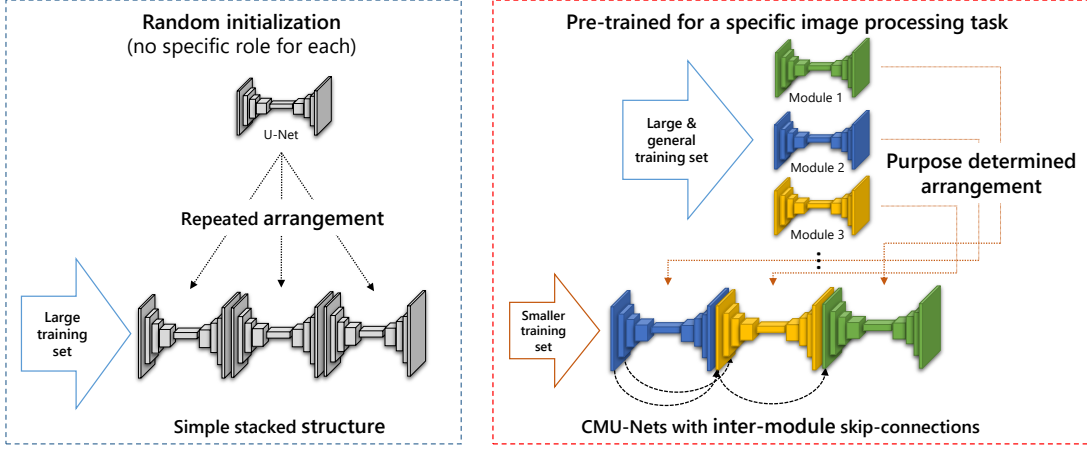


Figure 3-1: The difference between CMU-Nets and existing stacked U-Nets.

- The last and most important difference is their network structure. The proposed CMU-Nets has a novel structure between modular U-nets called inter-module skip-connections. The proposed inter-module skip-connections are important for achieving better performance as proved experimentally in the later section.

In Fig. 3-2, I illustrate the overall structure of CMU-Nets visually. Before explaining the detailed information of the proposed model, I explain U-Net's two primary characteristics. First, U-Net's structure consists of the contracting path and expanding path structures for pixel-to-pixel prediction. The contracting path in U-Net extracts context features through convolutions and down-sampling (max-pooling) and the expanding path contains up-sampling and de-convolutional layers. Second, U-Net adopts skip-connections for transmitting features from the contracting path to the expanding path. The skip-connections in U-Net induce more accurate localization in the expanding path. According to these characteristics, U-Net has shown robust image segmentation results. I was inspired by these characteristics of U-Net and I utilize them for designing CMU-Nets.

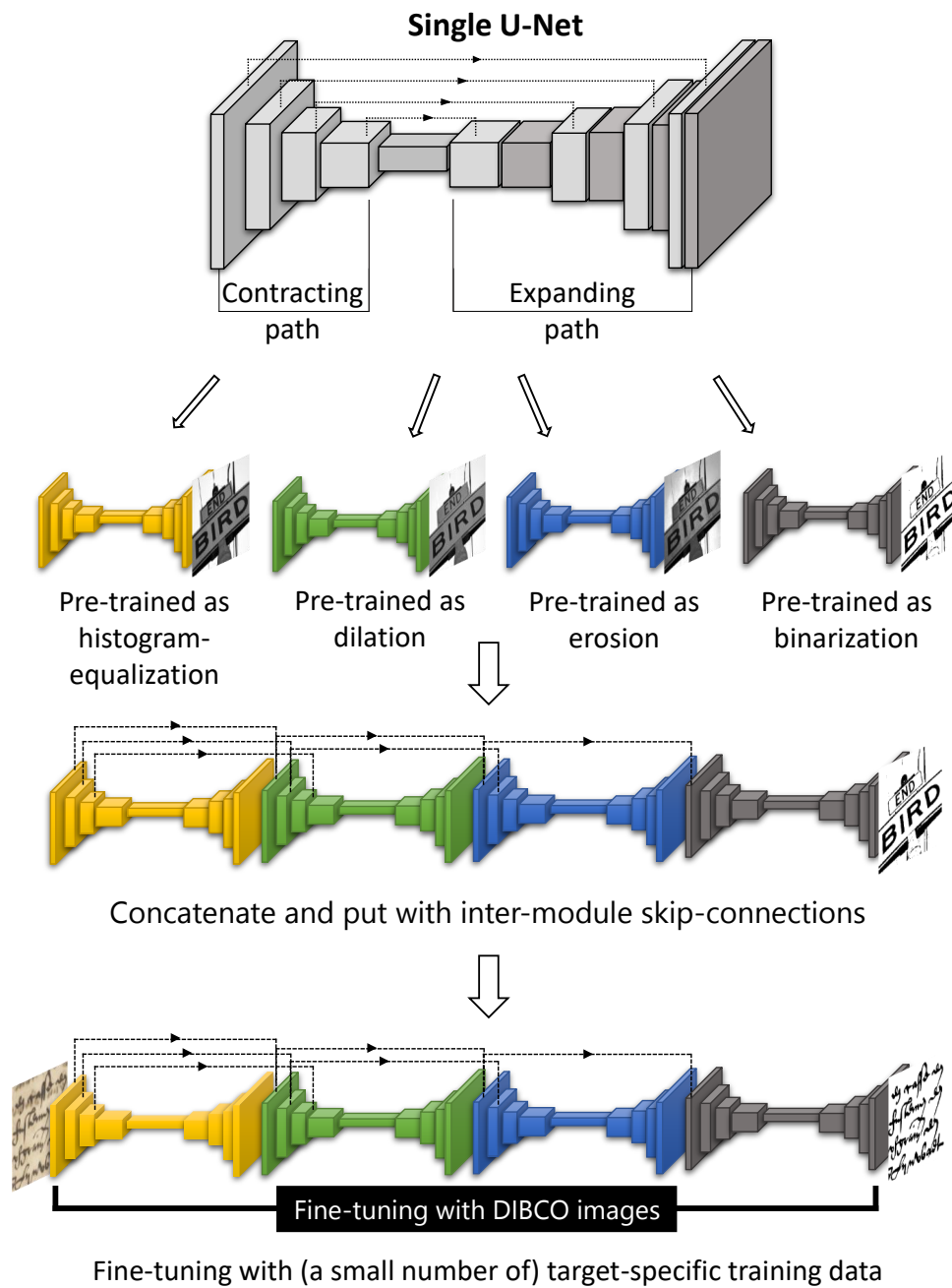


Figure 3-2: Concept of CMU-Nets. Individual U-Net modules are trained for specific image processing tasks. Then, they are cascaded and connected with inter-module skip-connections. Finally, the cascaded modular U-Nets are fine-tuned by using DIBCO datasets.

3.1 Pre-trained U-Net Modules in CMU-Nets

As shown in Fig. 3-2, CMU-Nets are comprised of pre-trained modular U-Net blocks. Each modular U-Net conducts a specific images-to-image transformation task. Since CMU-Nets are trained to conduct document image binarization, each modular U-Net is trained to conduct a specific document image processing. Therefore, I considered emulating traditional image processing techniques for procuring the pre-trained U-Net modules. According to the methods of the DIBCO participants, lots of document image binarization systems adopt pre-processing for decreasing obstacles such as unnecessary noise [79, 105]. Similar to the general image binarization systems, I designed the modular U-Net blocks to target the same image processing techniques.

Before procuring pre-trained modular U-Nets, in the first step, I should define which modules are needed. Since the goal is document image binarization and thus I prepare five image pre-processing modules, dilation, erosion, histogram-equalization, Canny-edge, and Otsu's binarization [20] in the beginning. Like Fig. 3-2, I build a standard U-Net for realizing modular U-Nets, and then I train these U-Nets using a large generic image dataset [134] that is external to DIBCO. According to this process, I secure five pre-trained U-Nets in the beginning.

3.2 Cascading Structure in CMU-Nets

In Fig. 3-2, after securing pre-trained U-Net modules, I stack these pre-trained U-Net modules for document image binarization. Compared with the conventional stacked U-Nets, since the individual U-Net module of CMU-Nets is pre-trained as transfer learning, an arbitrary cascading method can be improper. Therefore, One of the most important points to be considered for CMU-Nets is how to cascade the modular U-Nets.

In Fig. 3-3, I demonstrate how to order the pre-trained U-Net modules before connecting them. I can describe cascading U-Net modules as assembling bricks for

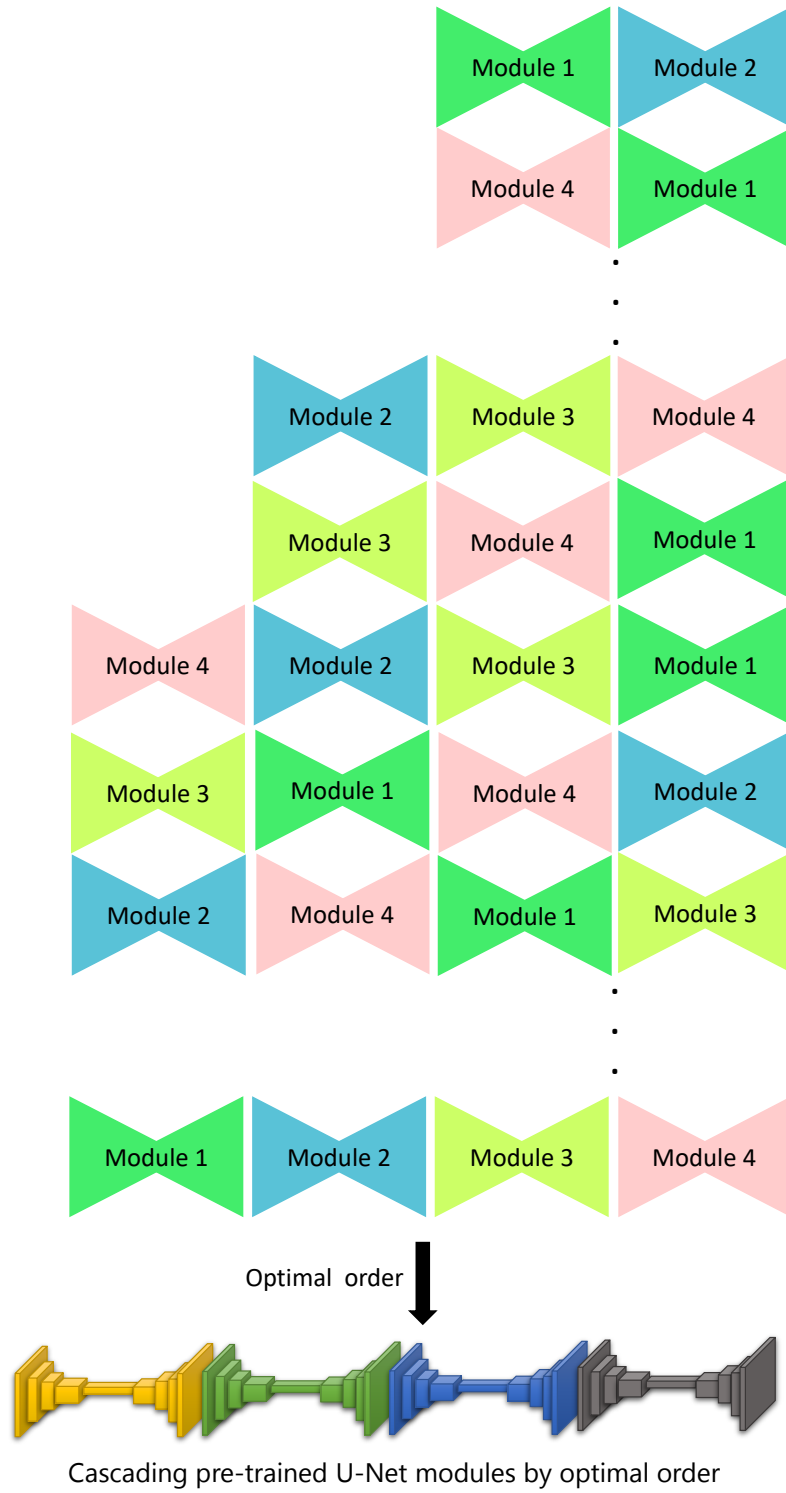


Figure 3-3: Examples of ordering pre-trained U-Net modules.

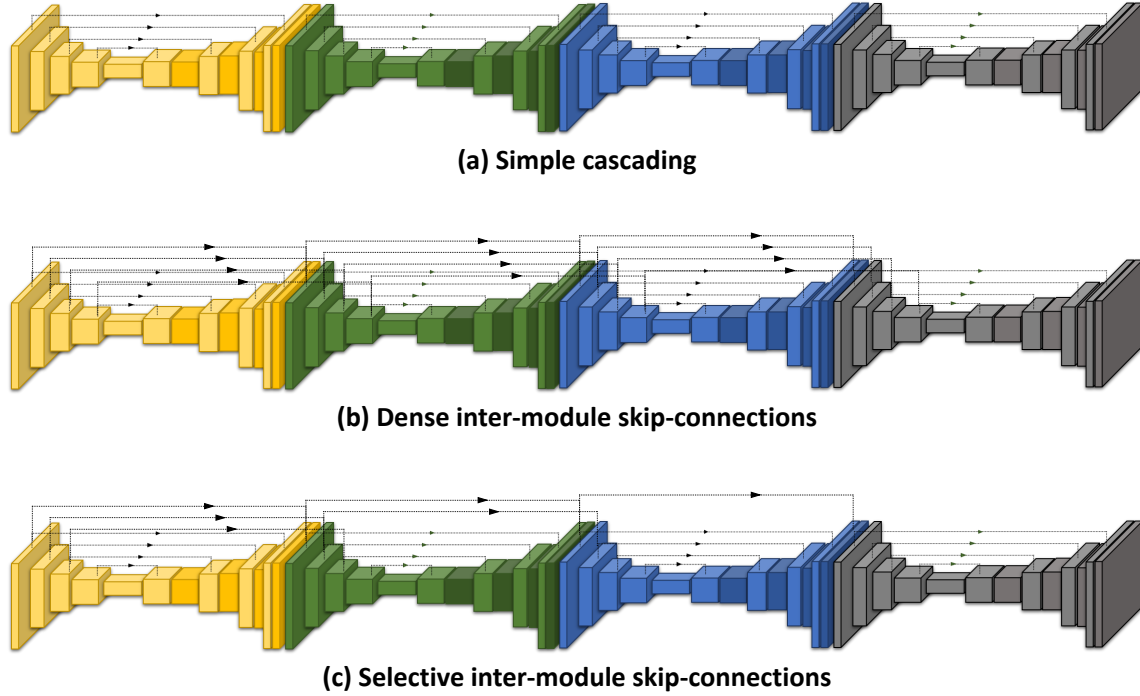


Figure 3-4: Visual examples of cascading pre-trained U-Net modules.

constructing a building. When constructing CMU-Nets, I should consider the order that the bricks should be used. In the proposed model, the optimal cascading order of the pre-trained modules is found through experimental results. After cascading the pre-trained U-Net modules, connecting them with the proposed module connection method and the fine-tuning process by using DIBCO data are conducted.

3.3 Selective Inter-module Skip-connections in CMU-Nets

Simply cascading U-Nets can cause the information to be lost between modules. This phenomenon can cause deterioration or excessive processing of noise removal in a portion of the foreground in the image. Therefore, to address this problem, I adopted inner skip-connections of U-Net to the cascading U-Nets.

Inner skip-connections in U-Net are utilized to transfer the features of contracting path’s layers to expanding path’s layers. In the conventional U-Net, according to this process, precise image re-composition can be conducted. Due to the strength of the inner skip-connections, cascaded U-Nets with inter-module skip-connections between modules have been proposed [117]. The features of the previous U-Net modules also can be transferred by utilizing the inter-module skip-connections as inner skip-connections.

As shown in Fig. 3-4, there are three possible strategies of cascading modular U-Nets: (a) simple cascading, (b) cascading with dense inter-module skip-connections, and (c) cascading with selective inter-module skip connections.

The inter-module skip-connections transfer information that can be lost within individual U-Nets. However, using a dense inter-module skip-connection scheme might carry excessive information and can be difficult to train. Therefore, I use a selective inter-module skip-connection strategy for the proposed CMU-Nets. In the experimental results in Section 4, I compare the proposed selective inter-module skip-connection strategy with a dense one to confirm this.

Chapter 4

Experiment for Document Image Binarization

4.1 Experiment Setup

4.1.1 Dataset for Fine-tuning

Since the target task of CMU-Nets is document image binarization, I utilized DIBCO and H-DIBCO datasets for fine-tuning of CMU-Nets. In the International Conference on Document Analysis and Recognition (ICDAR), there are several competitions for various document image analyses. Among them, DIBCO and H-DIBCO have been held to share various robust document image binarization methodologies since document image binarization is an important step in the document image analysis and recognition pipeline. The DIBCO and H-DIBCO benchmarking datasets that are used in the contest contain machine-printed and handwritten document images that are representative of the potential problems which are challenging in the binarization process. As shown in Fig. 4-1, the foreground areas of original images are binarized (black) areas.

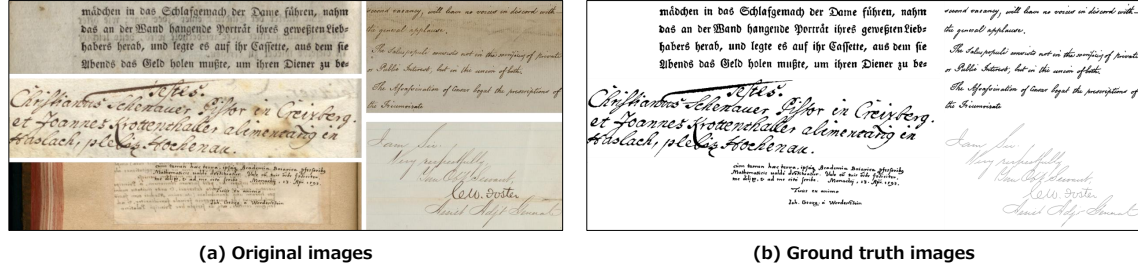


Figure 4-1: Examples of all DIBCO and H-DIBCO datasets. Images (a) and (b) are the original images and the ground truth images of DIBCO and H-DIBCO datasets.

The total number of DIBCO and H-DIBCO images is about 20 images per year, for a total of 116 images. Even though I utilize all DIBCO and H-DIBCO images for fine-tuning, the number of these images is insufficient for conducting a fine-tuning task. Therefore, I cropped and resized the DIBCO and H-DIBCO images to 256×256 . From this, 100 images from each DIBCO and H-DIBCO are extracted for the training sets, and I obtained a total of 900 images for fine-tuning. For a fair evaluation of each DIBCO and H-DIBCO competition, a leave-one-year-out strategy is taken for the fine-tuning. For example, when evaluating the H-DIBCO 2016 dataset, 100 images from each competition other than H-DIBCO 2016 are used as the training set. This is repeated for each competition evaluation. In other words, the fine-tuning is only done using images independent from the test set.

4.1.2 Dataset for Pre-training

Due to the limited amount of images, optimizing a binarization system by utilizing only the DIBCO images is a challenging task. For many machine learning methods, there is usually a focus on data augmentation. However, in the proposed method, I do not need to secure a large number of document binarization specific images, since I can train the U-Net modules on a dataset that is easier to secure.

Thus, before training CMU-Nets, I obtained images for training the pre-trained U-Net modules. An advantage of using the proposed modular U-Nets is the ability to

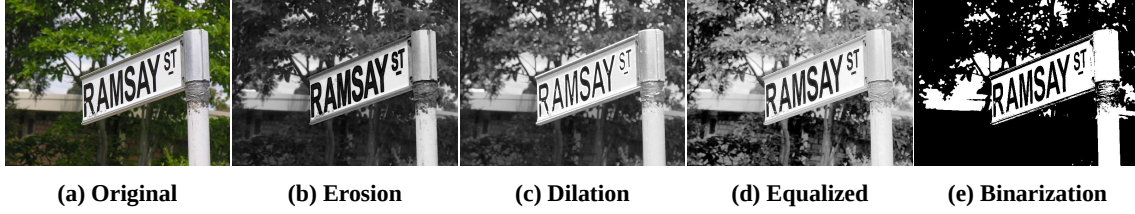


Figure 4-2: Examples of the MSCOCO images for the modularized U-Nets. Images from (a) to (e) correspond to the original, eroded, dilated, histogram-equalized, and binarized image.

utilize transfer learning using only a small number of DIBCO and H-DIBCO images to enhance the binarization performance. In this experiment, I used the COCO-Text [135] dataset for training the U-Net modules. Using the COCO-Text images, I created 10,000 images using Canny-edge, histogram-equalization, erosion, dilation, and the Otsu’s binarization methods for each module. An example of an image in the created dataset is shown in Fig. 4-2.

The ground truth for the specific image processing modules is as follows. I set the kernel size to 2×2 for the dilation and erosion data. This kernel size is selected so that the respective modular U-Net learns the minimum morphological transformation. Furthermore, I utilized a module for histogram-equalization, which makes the histogram of the values uniform. There are variants of histogram-equalization techniques, such as adaptive histogram-equalization, contrast limited adaptive histogram equalization (CLAHE) [1], multi-peak histogram equalization (MPHE) [136], and multi-purpose beta optimized bi-histogram equalization (MBOBHE) [137]. These variants show the robust redistributing performance of the brightness values. However, I decided to use a simple histogram-equalization method to train the modular U-Net. The reason is that the final purpose of the proposed system is the document image binarization and not histogram-equalization, specifically.

4.1.3 Experimental Environment

The experimental conditions for model training are as follows. For procuring pre-trained U-Net modules, each U-Net module is trained for 40,000 iterations using batches of 4. The fine-tuning for optimizing the cascading U-Nets is performed 10,000 steps using batches of 4. The size of all used images is 256×256 .

The basic architecture of the single modular U-Net is based on the original U-Net. In the contracting path, I adopted 3×3 convolution blocks, a ReLU activation function, and 2×2 max-pooling with stride 2. In the expanding path, these modular U-Nets operate the reverse process of the contracting path with concatenating inner skip-connections. The output of each modular U-Net is a grayscale image. The result of the binarization U-Net module is binary and thus the range of the output value is 0 or 1. The number of convolutional layers in each modular U-Net is 13. In the proposed method, I use four modular U-Nets with 6 inter-module skip-connections. I applied Adam optimizer [138] for model training with an initial learning rate of 0.0001. I used a sigmoid activation function in the output layer of CMU-Nets. The final results use a threshold of 0.5 on the output.

4.2 Evaluation Results

4.2.1 Evaluation Criteria of DIBCO

The DIBCO and H-DIBCO have standard metrics and evaluation criteria for each respective competition [77]. In recent competitions, DIBCO adopted four criteria, F-measure (FM), pseudo-F-measure (F_{ps}) [139], Peak Signal-to-Noise Ratio (PSNR), and Distance Reciprocal Distortion measure (DRD) [140]. I explain these criteria briefly for understanding the following quantitative results.

FM is the harmonic mean of precision and recall. According to a comparison between the output image and the ground truth image, I can calculate the FM value.

Namely, the FM is calculated by:

$$FM = 2 \frac{P \cdot R}{P + R}, \quad (4.1)$$

where precision P is:

$$P = \frac{TruePositive}{TruePositive + FalsePositive}, \quad (4.2)$$

and recall R is:

$$R = \frac{TruePositive}{TruePositive + FalseNegative}. \quad (4.3)$$

However, in [139], Ntirogiannis et al. mentioned that pixel-based evaluation measures such as FM can cause poorly binarized images with broken characters and false alarms to obtain higher precision and recall value than a better binarized image. The F_{ps} is suggested to overcome erroneous pixel-based evaluation method by weighing the distance between foreground and background boundary pixels. F_{ps} , which adopts pseudo-recall (R_{ps}) and pseudo-precision (P_{ps}) to solve this problem.

R_{ps} is calculated by:

$$R_{ps} = \frac{\sum_x \sum_y G(x, y) \cdot B(x, y)}{\sum_x \sum_y G(x, y) \cdot W_G(x, y)}, \quad (4.4)$$

and P_{ps} is calculated by:

$$P_{ps} = \frac{\sum_x \sum_y G(x, y) \cdot B(x, y)}{\sum_x \sum_y B(x, y) \cdot W_B(x, y)}, \quad (4.5)$$

where $G(x, y)$ and $B(x, y)$ are defined as the ground truth image and the binarized result image at pixel coordinate (x, y) , respectively. $W_G(x, y)$ and $W_B(x, y)$ are weight maps for the ground truth and binarized image, respectively. R_{ps} and P_{ps} weigh the ground truth images based on the distance of pixels to the text regions.

The PSNR evaluates the amount of noise in the result compared to a result image and a ground truth image. If the PSNR value is high, there is more similarity between the ground truth and the result image. PSNR can be expressed by:

$$\begin{aligned} PSNR &= 10 \log \left(\frac{MAX_{image}^2}{MSE} \right), \\ &= 20 \log \left(\frac{MAX_{image}}{\sqrt{MSE}} \right), \\ &= 20 \log (MAX_{image}) - 10 \log (MSE), \end{aligned} \quad (4.6)$$

where MAX_{image} represents the maximum value of an image (e.g. 8-bit grayscale image is 255). MSE is the mean squared error or:

$$MSE = \frac{\sum_x \sum_y (G(x, y) - B(x, y))^2}{XY}, \quad (4.7)$$

DRD can be used to measure the visual distortion specific for images. It is associated with human visual perception and measures distortion of every pixel. where X and Y are the number of pixels in the x and y directions. DRD can be used to measure the visual distortion specific for images. It is associated with human visual perception and measures distortion of every pixel. DRD is expressed by:

$$DRD = \frac{\sum_k DRD_k}{NUBN}, \quad (4.8)$$

where DRD_k is the distortion of the k -th pixel and is defined by:

$$DRD_k = \sum_{i=-2}^2 \sum_{j=-2}^2 + 4 |G_k(i, j) - B_k(i, j)| \times W_{norm}(i, j), \quad (4.9)$$

where $|G_k(i, j) - B_k(i, j)|$ is the difference between the ground truth and the binarized result at pixel (i, j) within a 5×5 block surrounding k . W_{norm} is a normalized weight matrix. In Eq. (4.8), NUBN is the number of non-uniform 8×8 blocks in the ground truth image.

Table 4.1: Comparative Results on DIBCO 2009 and H-DIBCO 2010

Year	Method	FM	PSNR	NRM	MPM
2009	Otsu [20]	82.17	15.06	5.63	13.86
	Sauvola and Pietikäinen [22]	87.26	16.69	6.61	3.39
	Guo et al. [141]	83.31	16.64	–	–
	Zhao et al. [142]	94.10	20.30	–	–
	Competition winner [79]	91.24	18.66	4.31	0.55
	Proposed	96.67	20.91	2.45	0.41
2010	Otsu [20]	85.27	17.51	9.77	1.35
	Sauvola and Pietikäinen [22]	75.30	15.96	16.31	1.96
	Guo et al. [141]	86.75	17.93	–	–
	Zhao et al. [142]	94.03	21.12	–	–
	Competition winner [76]	91.50	19.78	5.98	0.49
	Proposed	95.90	21.69	2.74	0.33

Negative Rate Metric (NRM) and Misclassification Penalty Metric (MPM) are depreciated measures that are no longer used by current DIBCO competitions. They are only used for reporting comparison results to older DIBCO winners.

4.2.2 Quantitative Evaluations

I compare the quantitative results of the proposed method with the previous DIBCO and H-DIBCO competition winners and other reported state-of-the-art results. Tables 4.1, 4.2, 4.3, 4.4 and 4.5 consist of comparative results for DIBCO and H-DIBCO separated by year. I divided the tables according to the evaluation criteria applied by DIBCO and H-DIBCO. In addition, Table 4.5 is separated to effectively compare the results of recent datasets. Note, the results of DIBCO 2015 is missing since DIBCO was not held in 2015.

In these tables, there are the results of Otsu’s method [20] and Sauvola and Pietikäinen [22] reported by DIBCO and H-DIBCO. The results of Howe’s method [131] are based on image processing techniques. The results from Vo et al. [52], Guo et al. [141], He and Schomaker [27], and Zhao et al. [142] method are based on ANNs.

The proposed method has the highest performance in 27 out of 36 evaluation crite-

Table 4.2: Comparative Results on DIBCO 2011

Year	Method	FM	PSNR	DRD	MPM
2011	Otsu [20]	82.1	15.7	9.0	–
	Sauvola and Pietikäinen [22]	82.1	15.6	8.5	–
	Howe [131]	91.7	19.3	3.5	–
	Vo et al. [52]	93.3	20.1	2.0	–
	Guo et al. [141]	83.82	16.50	6.55	–
	He and Schomaker [27]	93.4	19.9	1.9	–
	Zhao et al. [142]	93.8	20.3	1.8	–
	Competition winner [143]	80.9	16.1	104.4	64.4
	Proposed	95.5	19.9	1.8	0.5

Table 4.3: Comparative Results on H-DIBCO 2012 and DIBCO 2013

Year	Method	FM	F_{ps}	PSNR	DRD
2012	Otsu [20]	80.18	82.65	15.03	26.46
	Sauvola and Pietikäinen [22]	82.89	87.95	16.71	6.59
	Guo et al. [141]	86.40	89.00	17.86	4.67
	Zhao et al. [142]	94.96	96.15	21.91	1.55
	Competition winner [77]	89.47	90.18	21.80	3.44
	Proposed	95.16	96.44	21.37	1.13
2013	Otsu [20]	83.94	96.52	16.63	10.98
	Sauvola and Pietikäinen [22]	85.02	89.77	16.94	7.58
	Howe [131]	91.3	91.7	21.3	3.2
	Vo et al. [52]	94.4	96.0	21.4	1.8
	Guo et al. [141]	82.35	85.16	17.37	8.09
	Zhao et al. [142]	95.28	96.47	22.23	1.39
	Competition winner [144]	92.12	94.19	20.68	3.10
	Proposed	95.88	96.38	22.97	1.46

Table 4.4: Comparative Results on H-DIBCO 2014

Year	Method	FM	F_{ps}	PSNR	DRD
2014	Otsu [20]	91.78	95.74	18.72	2.65
	Sauvola and Pietikäinen [22]	86.83	91.80	17.63	4.90
	Howe [131]	91.7	19.3	3.5	1.08
	Vo et al. [52]	96.66	97.59	23.23	0.79
	Guo et al. [141]	92.30	94.86	19.17	2.37
	He and Schomaker [27]	95.9	97.2	22.1	0.9
	Zhao et al. [142]	96.41	97.55	22.12	1.07
	Competition winner [145]	96.88	97.65	22.66	0.902
	Proposed	97.05	97.91	22.37	0.91

Table 4.5: Comparative Results on H-DIBCO 2016, 2018 and DIBCO 2017

Year	Method	FM	F_{ps}	PSNR	DRD
2016	Otsu [20]	86.61	88.67	17.80	7.46
	Sauvola and Pietikäinen [22]	82.52	86.85	16.42	5.56
	Vo et al. [52]	90.10	93.57	19.01	3.58
	Guo et al. [141]	88.51	90.46	18.42	4.13
	He and Schomaker [27]	91.4	94.3	19.6	2.9
	Zhao et al. [142]	91.66	94.58	19.64	2.82
	Competition winner [146]	87.61	91.28	18.11	5.21
	Proposed	93.09	94.85	19.18	3.03
2017	Otsu [20]	77.73	77.89	13.85	15.54
	Sauvola and Pietikäinen [22]	77.11	84.1	14.25	8.85
	Zhao et al. [142]	90.73	92.58	17.83	3.58
	Competition winner [78]	91.04	92.86	18.28	3.40
	Proposed	91.57	93.55	15.85	2.92
2018	Otsu [20]	51.45	53.05	9.74	59.07
	Sauvola and Pietikäinen [22]	67.81	74.08	13.78	17.69
	Zhao et al. [142]	87.73	90.60	18.37	4.58
	U-Net [49]	82.08	84.79	15.91	7.15
	Competition winner [105]	88.34	90.24	19.11	4.92
	Proposed	89.71	91.62	19.39	2.51

ria and has significant performance improvements compared to the previous winners' results. According to the results from Table 4.1 to Table 4.5, I obtained an average FM of $94.50 \pm 4.79\%$ and an average F_{ps} of $95.12 \pm 3.50\%$. The largest performance improvement between the proposed and the previous winner's method was in H-DIBCO 2012. The performance difference between the proposed and the previous winner's method was 5.69 in FM. Also, the largest performance improvement in F_{ps} was 6.26. In Table 4.5, I added the results of the conventional single U-Net in DIBCO 2018. Also, I demonstrate the results of simply cascading U-Nets and CMU-Nets with dense inter-modules skip-connections in Section 4.3.1.

4.2.3 Qualitative Evaluations

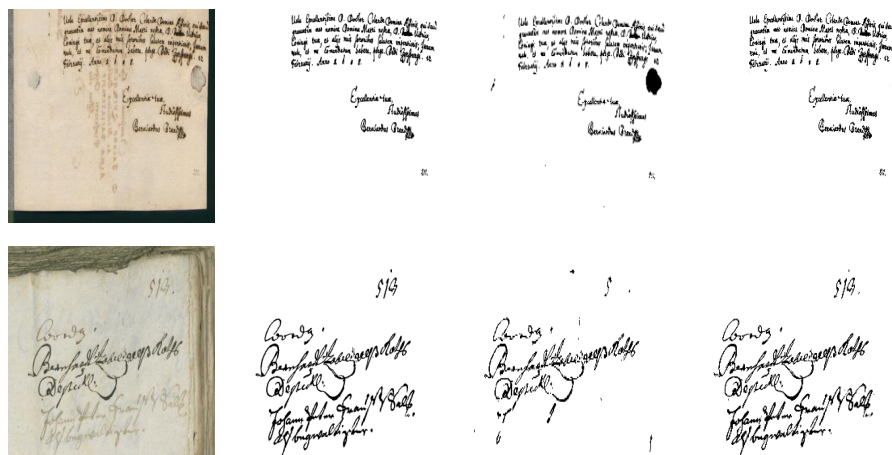
The qualitative results are illustrated in Figs. 4-3 and 4-4. The examples in Fig. 4-3 are from recent competitions in which the winners use modern methods and the examples in Fig. 4-4 are older competitions. The figures consist of the original images, the ground truth images, the competition winner's result images, and the proposed method's result images. According to the figures, I demonstrate that the proposed method performed better compared to the previous competition winners' methods. In Fig. 4-3, I can especially confirm that the proposed method has a significant performance improvement on binarization in the most recent DIBCO datasets. The reason for this is due to the recent datasets having more noisy images with more room for improvement when compared to the older datasets.

Furthermore, I can confirm that the proposed method can handle various kinds of noise in document images. In the case of Fig. 4-3 (a), the proposed method can erase the stain and extract the blurred character areas compared to the DIBCO 2018 winner's results. In the next example, Fig. 4-3 (b), the proposed method demonstrates robust binarization performance on bleed-through images. Finally, in Fig. 4-3 (c), the proposed method can erase the perpendicular line.

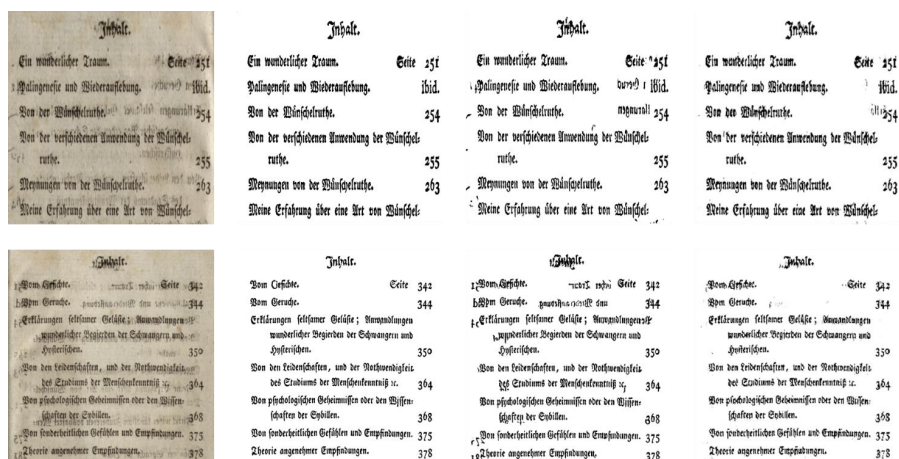
In the case of Fig. 4-4, the proposed method also demonstrates good binarized results. However, I could not find a large performance difference in the case of the older DIBCO and H-DIBCO datasets. The reason is due to the older DIBCO and H-DIBCO datasets consisting of document images with less noise compared to recent DIBCO and H-DIBCO datasets. In addition, the qualitative results of the older DIBCO and H-DIBCO were not introduced sufficiently.

4.2.4 Results for Types of Noise

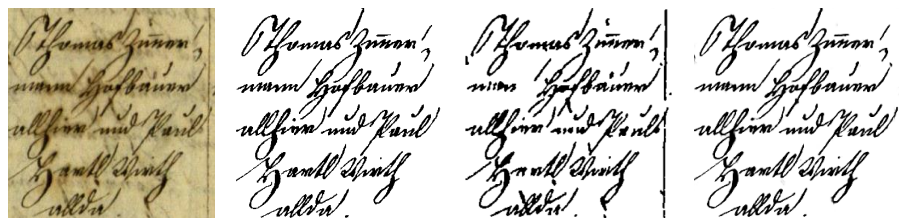
In this section, I demonstrate the binarized results on various types of noise. There are five types of noise in the DIBCO and H-DIBCO datasets, non-document regions,



(a) H-DIBCO 2018

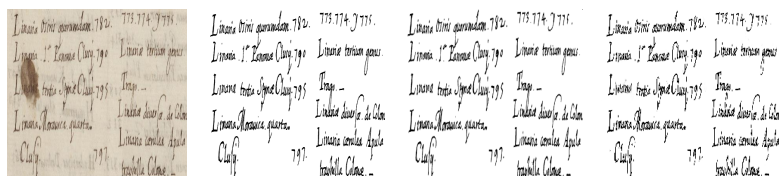


(b) DIBCO 2017

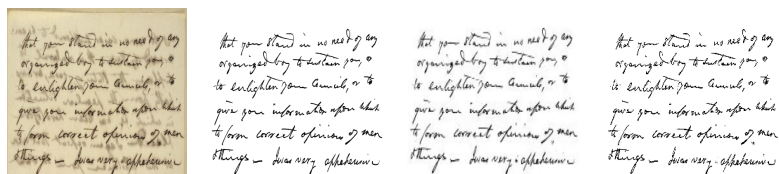


(c) H-DIBCO 2016

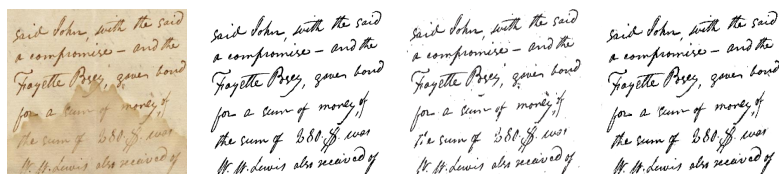
Figure 4-3: Result images compared with the previous DIBCO and H-DIBCO winners from 2016 to 2018. Images in the columns from left to right correspond to the original, the ground truth, the previous winner, and the proposed.



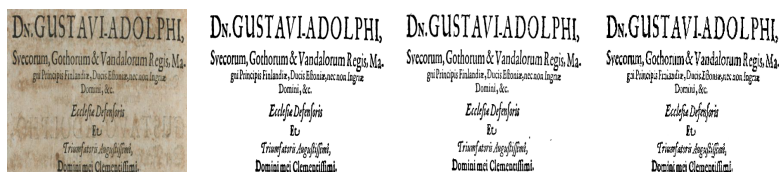
(a) H-DIBCO 2014



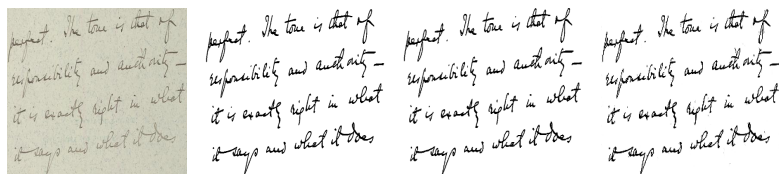
(b) DIBCO 2013



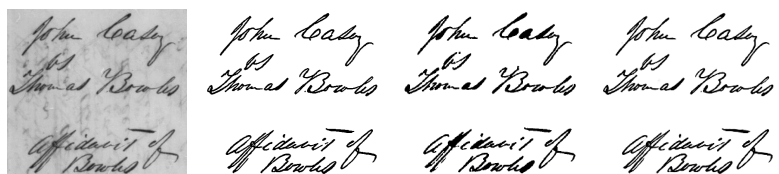
(c) H-DIBCO 2012



(d) DIBCO 2011



(e) H-DIBCO 2010



(f) DIBCO 2009

Figure 4-4: Result images compared with the previous DIBCO and H-DIBCO winners from 2009 to 2014. Images in the columns from left to right correspond to the original, the ground truth, the previous winner, and the proposed.

bleed-through, texture, luminance, and stains. Fig. 4-5 demonstrates the results of the proposed method on examples of documents with these types of noise. According to the Fig. 4-5, it can be observed that the proposed method is effective in removing the noise while preserving the text.

4.3 Discussion

4.3.1 Analysis of the Inter-module Skip-connections

In this section, I evaluate the proposed inter-module skip-connections and justify why they should be used in CMU-Nets. Before justifying the proposed inter-module skip-connections, I compared the result images between CMU-Nets without any inter-module skip-connections and a conventional single U-Net in Fig. 4-6. When I trained the CMU-Nets without any inter-module skip-connections, I pre-trained with the COCO-Text dataset and fine-tuning with DIBCO datasets. The pre-training process using the COCO-Text dataset is conducted using only binarized label images. I followed the same training conditions in this fine-tuning when I fine-tuned the proposed method.

According to Fig. 4-6, I confirmed that there is a tiny performance improvement between a conventional U-Net and the CMU-Nets without any inter-module skip-connections. The degraded case of the CMU-Nets without the inter-module skip-connections compared with a conventional U-Net is shown in Fig. 4-7. Furthermore, since the CMU-Nets contain four conventional U-Nets, training the CMU-Nets by using only a binarized COCO-Text dataset is difficult compared to a conventional U-Net. Therefore, I confirmed that training the CMU-Nets without inter-module skip-connections is not efficient compared with a conventional U-Net, and using separately pre-trained U-Net modules helps the model training of cascaded U-Nets models.

Also, I demonstrate the quantitative results compared with the results of using no inter-module skip-connections and dense inter-module skip-connections in Table 4.6.

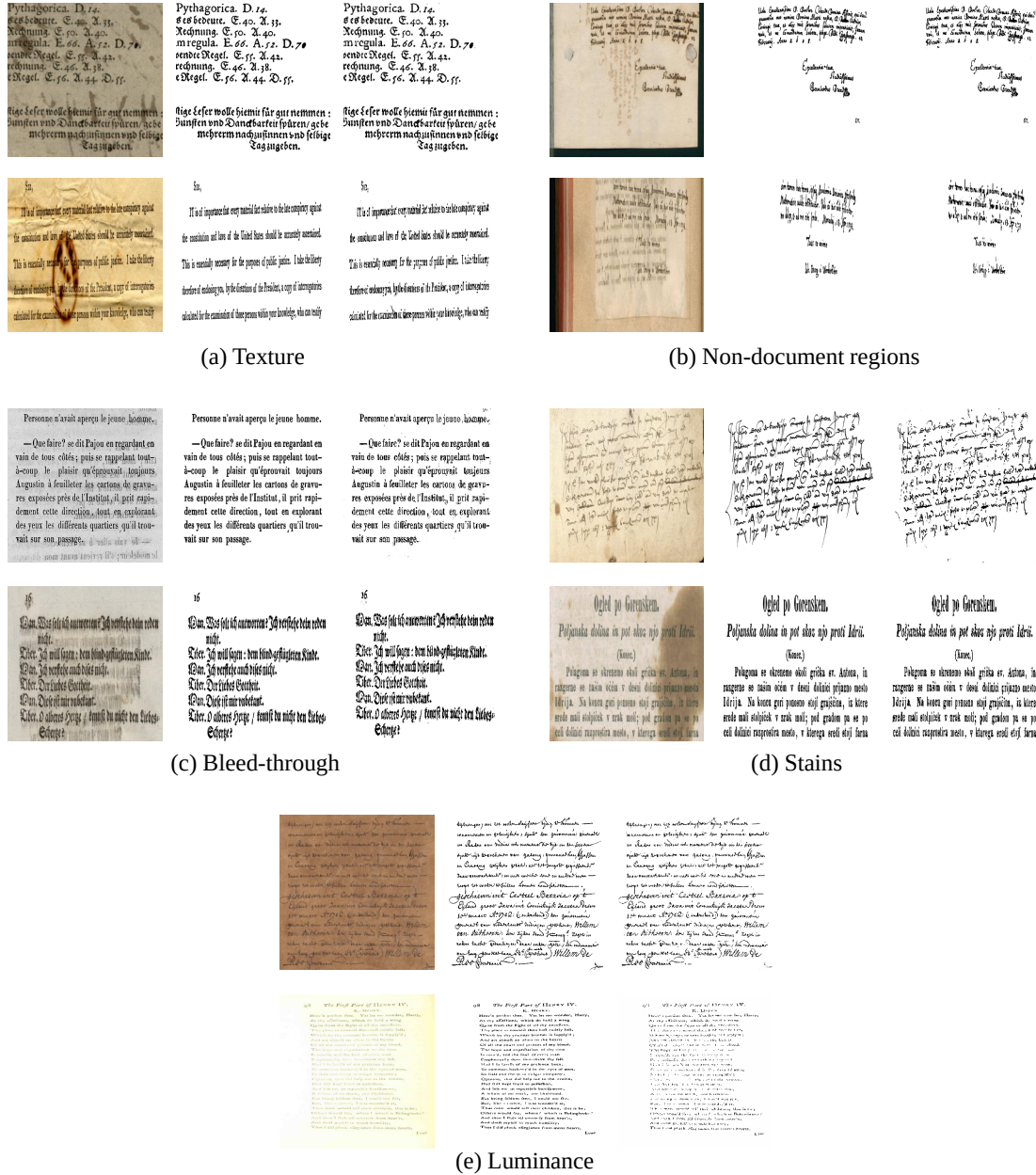


Figure 4-5: Binarization results of the proposed model according to the type of noise. Images in the columns from left to right correspond to the original, the ground truth, and the result images.

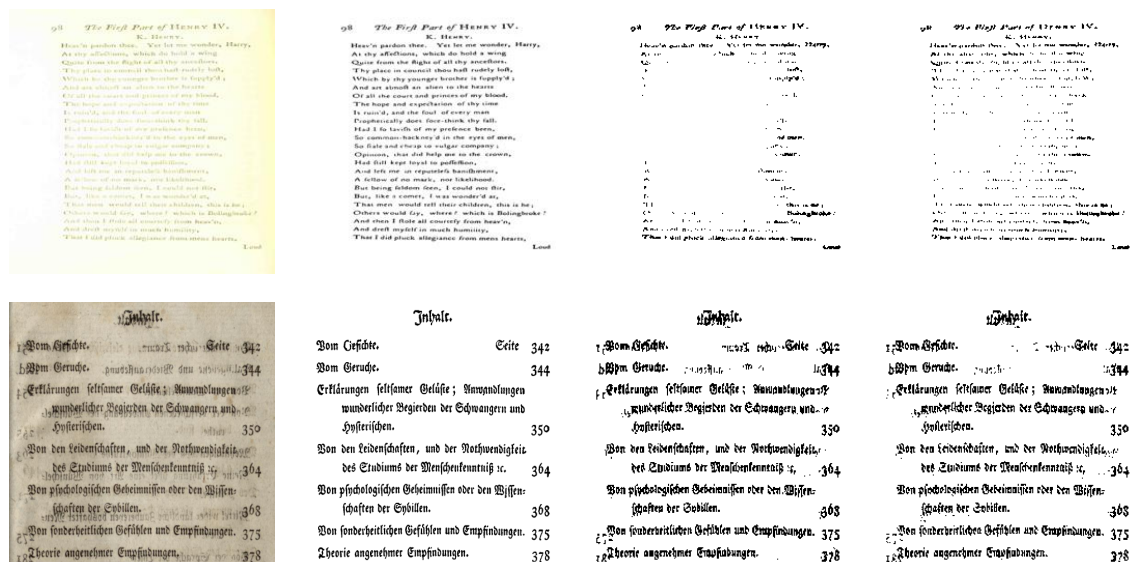


Figure 4-6: Comparative results between the CMU-Nets without any inter-module skip-connections and a conventional single U-Nets. Images in the columns from left to right correspond to the original image, the ground truth image, the result image of single U-Net, and the result image of CMU-Nets without any inter-module skip-connections.

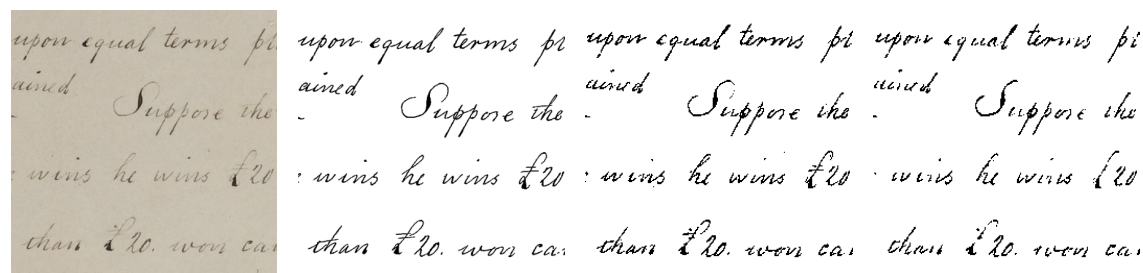


Figure 4-7: Worse case of the CMU-Nets without inter-module skip-connections. Images in the columns from left to right correspond to the original image, the ground truth image, the result image of single U-Net, and the result image of CMU-Nets without any inter-module skip-connections.

Table 4.6: Comparative results according to differences of inter-module skip-connections by using the DIBCO 2018 dataset for evaluation.

Method	FM	F_{ps}	PSNR	DRD
Without inter-module connections	83.71	85.94	14.29	7.19
With dense inter-module connections	85.19	86.59	13.26	6.37
Proposed method	89.71	91.62	19.39	2.51

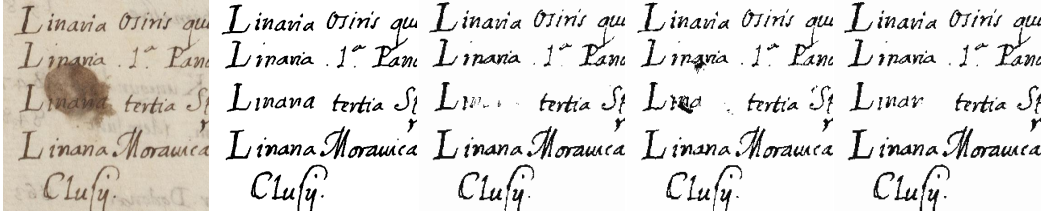


Figure 4-8: The reason why I adopt the proposed inter-module skip-connections in CMU-Nets. Images in the columns from left to right correspond to the original image, the ground truth image, the image without inter-module skip-connections, the image by adopting dense inter-module skip-connections and the proposed method.

The dense inter-module skip-connection approach uses skip-connections between each expanding path layer and each contracting path layer of the other modules. The model without inter-module skip-connections lacks skip-connections between modules, however, they still contain intra-module skip-connections.

The skip-connections in the original U-Net can transfer useful information that can be lost between layers. Due to this, I considered a dense approach for inter-module skip-connections. However, I found that using excessive (dense) inter-module skip-connections caused interference between modules. The reason for this is because most layers of the modular U-Nets are only relevant to their pre-trained task and might hinder the performance of unrelated intermediate tasks. However, some information is lost without the inter-module skip-connections, therefore, the more sparse approach from the proposed method is more effective. According to Fig. 4-8, I demonstrate why the proposed inter-module skip-connections are used in CMU-Nets.

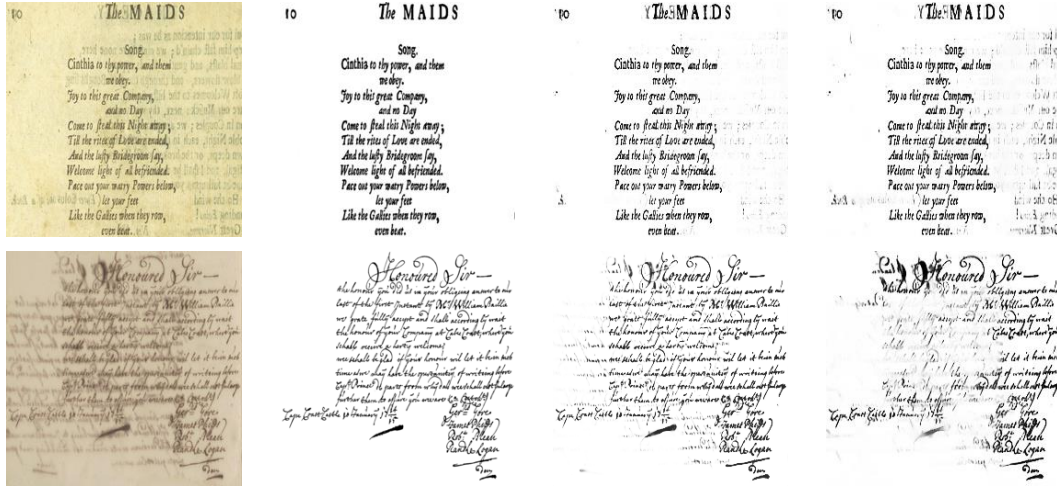


Figure 4-9: Results indicating why the pre-trained modules should be used. Images in the columns from left to right correspond to the original images, the ground truth images, the result images of the proposed model, and the result images of randomly-initialized cascaded model.

4.3.2 Analysis of Pre-trained U-Net Modules Utilization in CMU-Nets

In this section, I verify CMU-Nets that are randomly initialized cascading U-Nets. I demonstrate the evaluated results of the randomly initialized cascading U-Nets compared with the proposed model in Fig. 4-9. According to Fig. 4-9, I illustrate why these pre-trained U-Net modules should be utilized in CMU-Nets. Even though randomly initialized cascading U-Nets are the same architecture as CMU-Nets, the performance of the randomly initialized cascading U-Nets is worse than the proposed model. Due to the utilization of the pre-trained U-Net modules, CMU-Nets have shown significant performance by using an insufficient amount of training data.

4.3.3 On The Order of U-Net Modules

When I design CMU-Nets, I should consider the number of used pre-trained U-Net modules in CMU-Nets and the order of pre-trained U-Net modules since CMU-Nets

utilize these U-Net modules that are pre-trained for specific tasks. In the beginning, I demonstrate the result images of various cascaded U-Nets in Fig. 4-10. According to Fig. 4-10, I confirmed that the proposed model, which included four pre-trained U-Net modules, show significant performance compared with various cascaded U-Nets.

To find another good order to arrange the modules, I combined used four U-Net modules. The arranged order of CMU-Nets is as follows: histogram-equalization $\rightarrow$ dilation $\rightarrow$ erosion $\rightarrow$ binarization. Fig. 4-11 demonstrates some of the results of a different order of U-Net modules. In Fig. 4-11 (b), there is potential for utilization for CMU-Nets' order. However, since the histogram-equalization module is located behind erosion and dilation modules, the module's effect is not reflected well in CMU-Nets. Also, according to Figs. 4-11 (c), (d), and (e), I recognized that two things in CMU-Nets for document image binarization. First is that the erosion and dilation modules should be adjoined for effective noise removal. The second is that the binarization module should be placed at the last.

4.4 Summary

This chapter demonstrated the ability of CMU-Nets by various analyses and results. CMU-Nets are designed to conduct a historical document image binarization task. The proposed model adopts pre-trained modularized U-Nets trained by sufficient data. After then, the proposed model adopts a novel inter-module skip-connections in the cascaded pre-trained U-Nets. The proposed model proved that the utilization of pre-trained modular U-Nets and the proposed scheme of inter-module skip-connections in cascading U-Nets achieved robust binarization performance. The experiments were evaluated on all currently existing DIBCO and H-DIBCO datasets. It was able to maintain an average FM of $94.50 \pm 4.79\%$ and an average F_{ps} of $95.12 \pm 3.50\%$ for all DIBCO and H-DIBCO datasets.

Moreover, I justified some of the design choices, such as the proposed inter-module

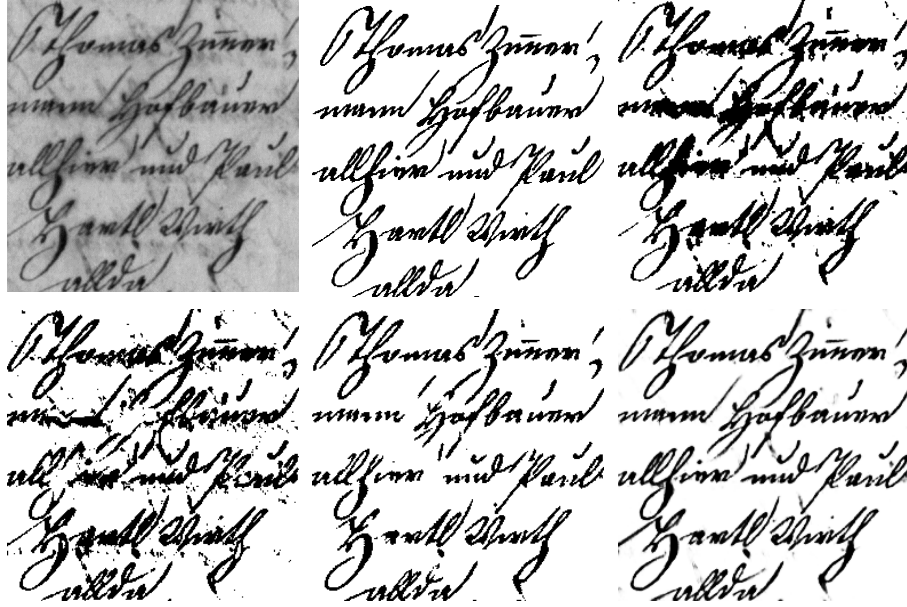


Figure 4-10: The result images of various cascaded U-Nets. Images in the columns from upper left to lower right correspond to the original images, the ground truth images, single U-Net for binarization, cascaded modular U-Nets with histogram-equalization and binarization modules, cascaded modular U-Nets with erosion, dilation, and binarization modules, and cascaded modular U-Net with the histogram-equalization, erosion, dilation, and binarization modules.

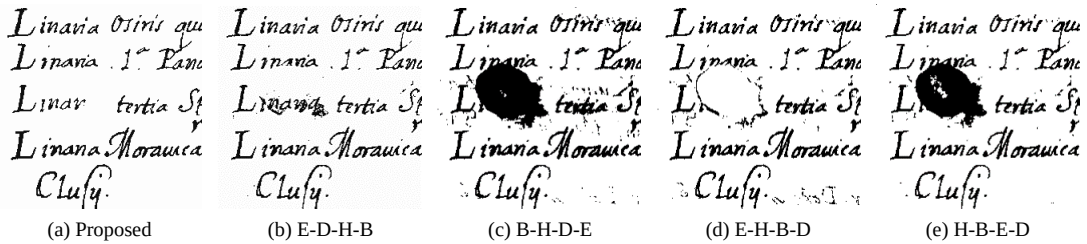


Figure 4-11: Result images of models arranged in a different order than the proposed method. For simplicity, I use abbreviations for the words as follows: B: binarization, D: dilation, E: erosion, and H: histogram-equalization.

skip-connection scheme. In the discussion section, I demonstrate detailed considerations when I design CMU-Nets such as the order of cascading U-Net, the necessity of the proposed inter-module skip-connections.

Part II

CMU-Nets Incorporating SE-Nets

Chapter 5

Background and Related Work

5.1 Attention Mechanism

In this section, I introduce the background knowledge of attention mechanisms in various research fields. Also, the related studies on the utilization of attention mechanisms and SE-Nets in U-Net is introduced briefly.

5.1.1 Attention Mechanism in Natural Language Processing

The attention mechanism was suggested to improve sequence to sequence model (seq2seq) [147] of Natural Language Processing (NLP) [56]. Seq2seq, which was proposed by Sutskever et al., is to solve a language model task. Seq2seq is designed for transforming an input sequence to a target sequence like a translation task. As shown in Fig. 5-1, seq2seq extracts a fixed-length context vector from an input sequence and generates an output by using the context vector. However, as the length of the sequence gets longer, there is a limit to remembering since seq2seq uses a fixed-length context vector.

For solving the limitation of seq2seq, an attention mechanism was utilized that

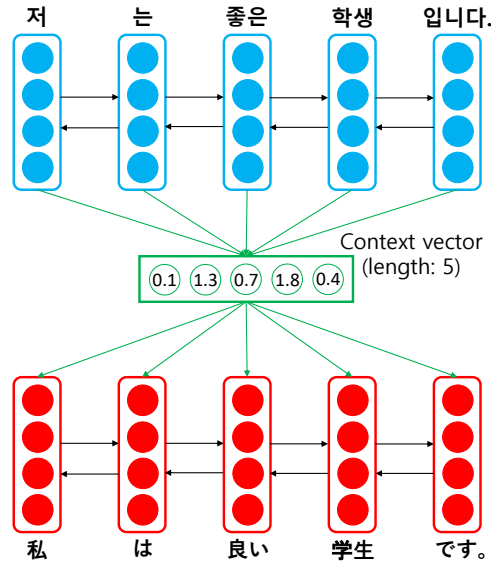


Figure 5-1: Example of seq2seq’s sequence transformation process. The encoder of seq2seq generates fixed-length context vectors and the decoder generates an output sequence by using the context vectors.

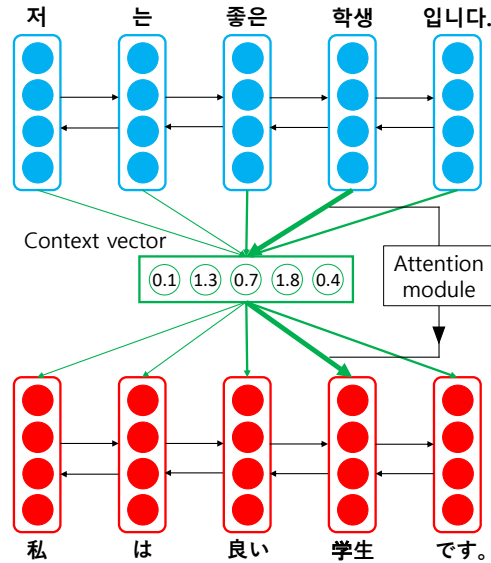


Figure 5-2: Example of attention mechanism process. Compared with the conventional seq2seq’s process, the attention mechanism extract context vectors regardless of its length.

was suggested by Bahdanau et al. The conventional seq2seq needs to compress all the necessary information of an input sentence into a fixed-length context vector. However, Bahdanau et al. proposed to encode an input sentence into a sequence of vectors and to use a subset of these vectors in the decoder called attention mechanism like as Fig. 5-2. By using the attention mechanism, the information of an input sentence does not need to be squashed into a fixed-length vector.

After the introduction of attention mechanisms, various attention mechanism studies have been conducted. Luong et al. suggested an improved attention mechanism [148] that simplifies the computation path of the conventional mechanism. Vaswani et al. suggested a non-recurrent sequence to sequence model with attention mechanism [149]. Since various attention mechanism-based models have shown robust performance in NLP [150, 151], attention mechanism is utilized in various research fields [152, 59].

5.1.2 Attention Mechanism in Computer Vision

Early CNN-based models with attention mechanism were proposed for an image captioning task [152, 153] and visual question answering [154]. Image captioning is desired to generate an explanation sentence by using an image as input data. Since image captioning can generate the description of an image, it can give the information of an image to a blind person. Among various image captioning research, Xu et al. suggested a novel visual attention network that generates an output sentence by looking at a certain part of an input image [152]. In the case of the conventional image captioning methods without attention mechanism, the architecture of these methods is to inject the output of a CNN's fully connected layer into a Long Short-Term Memory (LSTM) [155]. However, in [152], Xu et al. realized attention modules by using CNN's convolutional feature map and inject them into an LSTM. Since their proposed method utilizes the convolutional feature map, the spatial information of an input image is not lost. Due to the significant achievements of the

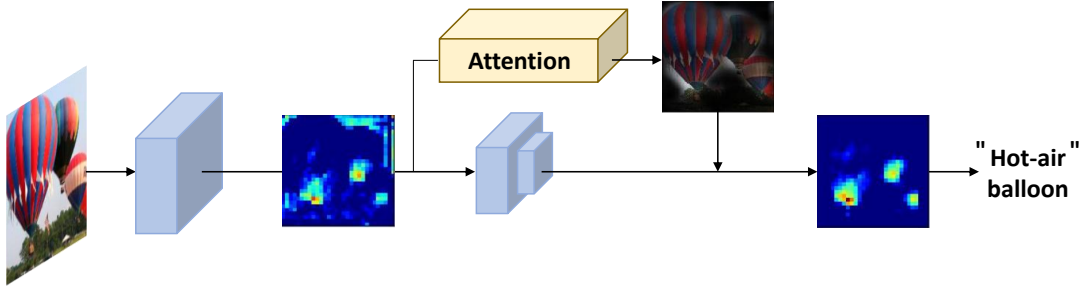


Figure 5-3: Example of attention mechanism process in CNN-based models. Due to an attention module, a CNN can generate a well-extracted feature mask. This attention module improves the accuracy of CNN's visual processing.

attention-based image captioning method, various image captioning methods have been suggested [156, 157, 158].

Visual question answering is a task that generates a proper answer when a question related to an input image is given [159]. Visual question answering requires a high-level image understanding and complex inference ability compared with image captioning [160]. Lots of visual question answering networks have been developed with various novel attention modules [161, 162, 163].

Attention mechanism has been also applied in CNN-based models for improving CNN's image processing ability such as classification [164]. As shown in Fig. 5-3, Wang et al. suggested a residual attention network that is based on ResNet [127] with attention mechanism [74]. Sarafianos et al. suggested an effective visual attention mask for imbalanced attribute classification [165]. Wang et al. a novel residual attentional Siamese network for online visual tracking [166]. Sun et al. proposed a novel attention-based CNN which controls multiple object parts between different input images for fine-grained image recognition [167]. Among these attention mechanisms, novel attention mechanisms have been suggested to re-calibrate features of CNN's feature maps called SE-Nets [64].

Since SE-Nets won the ImageNet Large Scale Visual Recognition Competition

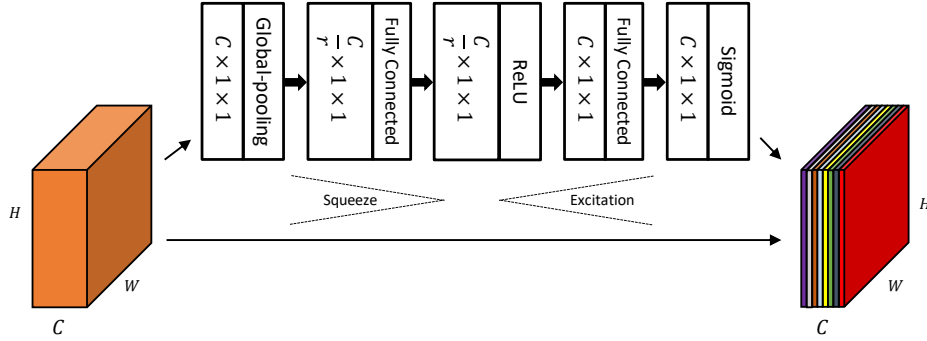


Figure 5-4: The architecture of a SE-block. According to the process of squeezing and exciting features, a SE-block can control the channel-wise feature response properly.

(ILSVRC) 2017 [64], various advanced studies on SE-Nets have been suggested [168]. In the squeeze operation of SE-block, important features of each channel are extracted by Global Average Pooling (GAP). After the squeeze operation of the SE-block, the excitation operation is conducted to perform feature re-calibration. In the excitation operation, channel-wise dependencies are calculated by tuning FC layers and the ReLU activation function. They adopted SE-blocks in VGGNet, GoogleNet, and ResNet evaluated them. SE-Net has two strengths: first is that attaching SE-blocks in anywhere of various NN models and second is the significant improvement with less parameter increase. In Fig. 5-4, visual architecture of SE-block is illustrated.

After introducing SE-Nets, Park et al. suggested a novel attention mechanism, called Bottleneck Attention Module (BAM), that is localized in the spatial pooling as a bottleneck part of a network [61]. BAM conducts channel and spatial attention operations parallelly. The refined convolutional features by these attention operations are transformed by a 3-dimensional attention map as the same size of an input tensor. They also suggested Convolutional Block Attention Module (CBAM) that adopts the revised connection method between spatial and channel attention [169]. These self-attention mechanisms have been utilized for improving various CNN-based models [170, 171].

5.2 Related work

5.2.1 U-Net based Models with Attention Mechanism

Due to the significant achievements of CNNs with attention mechanisms, studies of U-Net with attention mechanisms have been proposed. Oktay et al. proposed a novel attention U-Net that applied the stacked residual attention mechanism in U-Net [172]. They demonstrated robust semantic segmentation performance by attaching attention gates to all expanding paths of U-Net. N. Abraham and N. Khan suggested improved attention U-Net for lesion segmentation [173]. Jin et al. proposed to extract liver and tumor in CT scan images by a 2-dimension residual attention-aware U-Net [174]. Fan et al. suggested a novel multi-scale attention U-Net [175]. Li et al. proposed a novel attention U-Net for accurate retinal vessel segmentation [176].

Various FCN based models with attention mechanism also suggested as well as U-Net models with attention modules for various image transformation tasks [177, 178, 179, 180]. These attention mechanism based image-to-image transformation models have shown great achievements in various image transformation tasks.

5.2.2 U-Net based Models with SE-blocks

Since SE-blocks can attach in anywhere of the existing CNNs for improving the CNNs' performance, several CNN models with SE-Nets also have been proposed as well as CNN-based attention networks. For example, Roy et al. utilized SE-blocks for channel re-calibration for image classification with FCNs [181].

Since CNN-based models with SE-blocks have shown significant performance improvement, U-Net based models also have adopted SE-blocks for robust image transformation. Rundo et al. demonstrated robust semantic segmentation performance by applying SE-blocks inside U-Net [75]. They added SE-blocks to the existing U-Net contracting path and used SE-blocks as a way to concatenating by adding them in the upsampling process of the expanding path. Zhu et al. proposed 3D SE U-Net

for fast and fully automated anatomical segmentation [182]. They used SE-blocks to replace some of the max-pooling and convolution filters in U-Net to transfer features between layers. Cao et al. suggested a novel non-local network with SE-blocks-based global context attention blocks [183]. Lee et al. suggested a novel U-Net with SE-blocks that is designed to segment brain ischemic and white matter hyperintensities lesions [184].

5.3 Summary

In NLP, an attention mechanism was suggested to concentrate on specific features of input data. Since NLP systems included attention mechanism have shown significant performance, various attention-based networks for NLP have been studied. Moreover, beyond the conventional tasks of NLP, various attention-based networks have been suggested in various research fields such as computer vision.

Early CNN-based attention networks aim to solve a visual question answering and image captioning task [152]. After then, the attention mechanism is adopted in CNN models for improving image classification ability. SE-Net, BAM, and CBAM are novel attention architectures that improve CNN models' performance with a simple application. Due to the significant achievements of CNN-based attention networks, these attention methods have been applied in various research fields [171].

CNN-based models with attention mechanism have been studied for solving various image processing tasks such as classification [169] and segmentation [172]. Since these attention modules in CNN models can extract the context information of an input image, the modules improve the performance of CNN models.

Moreover, SE-Nets have been applied in CNN models to enhance the performance of various image processing tasks. SE-blocks in these CNN models have shown significant improvement of segmentation [181]. Due to these achievements of SE-Nets, researchers have still considered designing U-Net based models when they want to

improve these U-Net models with SE-Nets [75].

Since the existing SE-Net utilizations on U-Net aims to improve the performance of U-Net drastically, these U-Net models with SE-Nets are injected lots of SE-blocks for precise recalibration of channel-wise features. However, CMU-Nets should adopt pre-trained U-Net modules and the inter-module skip-connections for insufficient training data problem. Furthermore, CMU-Nets are a deep-depth network since CMU-Nets are organized with four U-Net modules. Therefore, the existing SE-Nets utilization can be not proper to adopt in CMU-Nets.

Chapter 6

Attention CMU-Nets (ACMU-Nets)

CMU-Nets, which are the base model of ACMU-Nets, have two major differences from existing stacked U-Nets. The first is that CMU-Nets adopt skip-connections between connected modules for robust feature transfer. In CMU-Nets, the proposed connection method between modules is applied called inter-module skip-connections. Inter-module skip-connections can transfer pixels' localization information to the next modules and prevent excessive interference from the previously used U-Net modules. Second, to solve the problem of insufficient document image quantity, CMU-Nets adopted pre-trained U-Net modules by using external datasets for training. Furthermore, since CMU-Nets are organized with four pre-trained U-Net modules, CMU-Nets are a deep-depth network compared with the existing U-Net models with SE-Nets. Therefore, if CMU-Nets adopt full attachment of SE-blocks like the existing U-Net models with SE-Nets, unconstrained SE-block utilization can cause excessive feature recalibration in CMU-Nets. As a result, ACMU-Nets should consider the utilization of SE-Nets.

In ACMU-Nets, as shown in Fig. 6-1, three strategies are adopted to design the proposed model compared with CMU-Nets. The first strategy is to adopt modified training methods of individual U-Net modules compared with the existing training method of CMU-Nets. The second strategy is to design the proper utilization of SE-

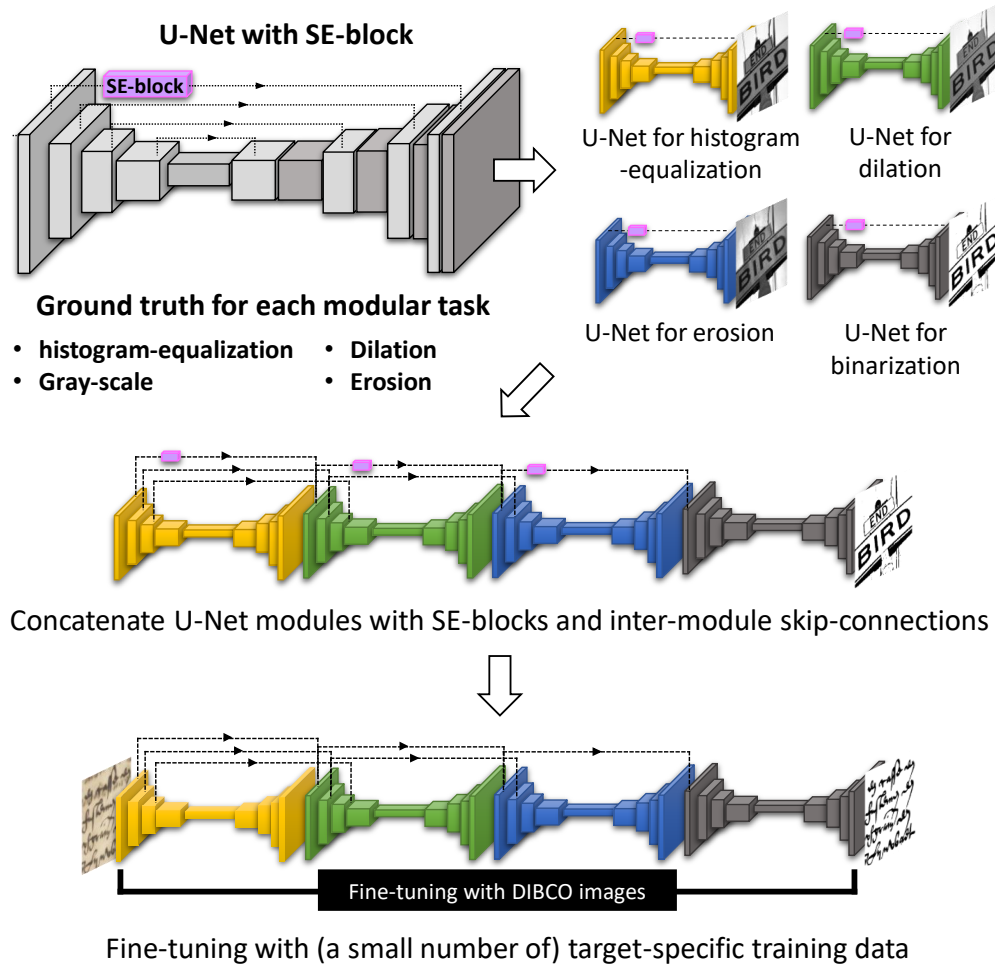


Figure 6-1: Concept of ACMU-Nets. Compared with CMU-Nets, the individual U-Net of ACMU-Nets adopt a SE-block in the skip-connection. After then, SE-blocks are injected in the first layers of the inter-module skip-connections.

Nets in the individual U-Net that is organized with ACMU-Nets. The third strategy is to design the inter-module skip-connections incorporating SE-Nets. In the later sections, detailed information on the three strategies will be introduced.

6.1 Modified Training Method of Individual U-Net

In CMU-Nets, when I trained an individual U-Net module, only gray-scaled images were used as the input data. The individual U-Net modules of CMU-Nets are trained by four types: histogram-equalization, erosion, dilation, and binarization. When these four U-Net modules are trained, the input image is a gray-scaled image since these four techniques can be conducted by using a gray-scaled image.

However, after cascading the individual U-Net module, the actual inputted image into the individual U-Net module is not a gray-scaled image. Even though the first order of the cascaded individual U-Net module utilizes a gray-scaled image as an input, the other modules utilize the output generated by the previously connected U-Net module. Therefore, gray-scaled, dilated, eroded, and histogram-equalized images are allocated as training images for each U-Net module. By adopting these four types of input images, each cascaded U-Net modules of ACMU-Nets can generate robust outputs.

Furthermore, the cascading U-Net modules with the modified training method can conduct robust output generation regardless of the connected order of the pre-trained U-Net modules. In the CMU-Net research, the connected order of individual U-Net modules was fixed since I knew the optimal connection order. However, it is difficult to set an optimal order of individual U-Net modules on CMU-Nets for conducting the other tasks compared with a document image binarization task. Due to this simple change, I can create pre-trained U-Net modules that are less affected by the order in which the modules are arranged. The effect of the modified training method is demonstrated in the later section.

6.2 SE-Nets in Individual U-Net Modules

According to various studies of CNNs with SE-Nets [168, 182, 75], I expected that SE-Nets will improve the performance of CMU-Nets. Especially, in [75], nine SE-

Nets are injected in U-Net for improving the performance of U-Net. Even though the total parameters of a CNN-based model with a single SE-Net are 10% more than that of a CNN-based model without a SE-Net [64], the CNN-based model with a SE-Net achieves significant performance improvement. Therefore, when individual U-Net modules of ACMU-Nets are designed in the beginning, nine SE-Nets are adopted in an individual U-Net like [75].

However, the individual U-Net with nine SE-Nets is revised as the U-Net with a single SE-Net like Fig. 6-1 consequently. Before utilizing SE-blocks in CMU-Nets, there are two major considerations to design ACMU-Nets. The first is that CMU-Nets is a model created by combining pre-trained U-Net modules. SE-blocks must be trained from the initial model training process to learn which features of the input image are required for accurate image-to-image transformation tasks. However, since CMU-Nets use pre-trained U-Net modules, the training process of the SE-blocks must be completed during the fine-tuning process. If I overload the training process of the SE-blocks intentionally enough, I can face the risk of over-fitting. The second is that each U-Net module used in CMU-Nets has its role (specific task). Among the used U-Net modules in CMU-Nets, there is just one module that performs image binarization. The others are designed to remove noise and improve image quality. Therefore, the features of the previous module extracted by the SE-block are not unconditionally required by the next module or image binarization module. Due to these two major considerations, the modified U-Net module with a single SE-Net is designed by two reasons.

The first reason to adopt a single SE-Net in an individual U-Net is excessive parameter increase of ACMU-Nets. Since CMU-Nets are organized four U-Net modules, 36 SE-Nets should be utilized at least. Furthermore, CMU-Nets should adopt the strategy of minimum parameter utilization for improving their performance since the individual U-Net was designed by using the original architecture of U-Net. Therefore, the most effective position of a SE-Net attachment should be found to minimize

the used parameters of ACMU-Nets and improve their performance simultaneously.

The second reason is that the degree of the performance improvement increase is negligible compared to the increase in usage of SE-Nets. In several studies of SE-Net utilization in CNN-based model [182, 75], these models are designed by using a single U-Net. However, ACMU-Nets are organized with four U-Net modules. Furthermore, the each individual U-Net module of CMU-Nets is a pre-trained U-Net module. Therefore, how each U-Net is trained or how each U-Net is connected is more important rather than how many SE-Nets are used when CMU-Nets are designed.

6.3 SE-Nets in Inter-module Skip-connections

Inter-module skip-connections of ACMU-Nets adopt three SE-Nets for improving the performance of ACMU-Nets like Fig. 6-1. Since Global Average Pooling (GAP) within a SE-Net compress global information into channel descriptors, the position of extracting global features in CMU-Nets is a proper position of SE-Net attachment. Therefore, a SE-Net is attached in the first layer of each individual U-Net.

In the inter-module skip-connections of ACMU-Nets, a SE-Net is attached in the first layer of the inter-module skip-connections like Fig. 6-1. Since global information of an image is transmitted in the first layer of the inter-module skip-connections, three SE-Net should be attached at least in ACMU-Nets. Even though SE-Nets can be attached in the other layers of the inter-module skip-connections, only three SE-Nets are injected for minimizing the total parameters of ACMU-Nets. After cascading pre-trained U-Net with the proposed methods of SE-Net utilization, ACMU-Nets are fine-tuned.

Chapter 7

Experiment of ACMU-Nets

7.1 Experimental Setup

7.1.1 Dataset

In this experiment, two datasets are utilized for model training and fine-tuning. To compare CMU-Nets and the proposed method properly, I train the proposed method on the same conditions. First, I use The COCO-Text dataset for pre-training the U-Net modules. However, according to Section 6.1, I mention that dilated, eroded, and histogram-equalized images are used with gray-scaled images for training each U-Net module. Therefore, a total of 10,000 images were acquired, with 2,500 images in each class. If input and output images are the same results (i.e. input image: dilated, output image: dilated), relevant images substitute gray-scaled images.

Also, for fine-tuning, I use the previous DIBCO and H-DIBCO datasets. The total number of all DIBCO and H-DIBCO images is 116 images. Since these images are insufficient for fine-tuning, 100 images from each DIBCO and H-DIBCO are made by cropping the original images. Therefore, I can procure a total of 800 DIBCO and H-DIBCO images for fine-tuning. The dataset of DIBCO 2017 is used for the test. The size of all images is resized to 256×256 .

7.1.2 Experimental Environment

In this experiment, each U-Net module is trained by 40,000 iterations using batches of 4 for pre-trained U-Net modules. The fine-tuning is conducted 10,000 iterations. Since all conditions should be equaled with conventional CMU-Nets, the used U-Net module is based on the original U-Net. I applied 3×3 convolution filters and 2×2 max-pooling filters. I fixed the value of reduction ratio (r) in SE-block by 16. The number of used convolutional layers in each U-Net module is 13. I used Adam optimizer with an initial learning rate of 0.0001 and a sigmoid activation function in the output layer.

7.1.3 Evaluation Criteria

To evaluate the proposed method, I use the same criteria as CMU-Nets' criteria. There are four criteria: FM, F_{ps} , PSNR, and DRD. Since the explanation of these four criteria is demonstrated in Section 4.2.1, the detailed explanation is omitted.

7.2 Experimental Results

7.2.1 Quantitative Evaluations

In this section, I compare the quantitative results of ACMU-Nets with the DIBCO 2017 winner's method and CMU-Nets by using the DIBCO 2017 dataset. In Table 7.1, I obtained the three highest values compared with the others. ACMU-Nets achieved the value of FM of 92.19 and F_{ps} of 94.07. In particular, the value of PSNR is up by 8.08%. According to these results, I confirmed that SE-blocks can improve the performance of CMU-Nets.

Also, I conducted a comparative evaluation with the revised training method. The revised training method is obtained a higher value in all criteria in Table 7.2. However, the difference of the values between the proposed method without the revised training

Table 7.1: Comparative Results on DIBCO 2017

Method	FM	F_{ps}	PSNR	DRD
Otsu [20]	77.73	77.89	13.85	15.54
Sauvola and Pietikäinen [22]	77.11	84.1	14.25	8.85
Zhao et al. [142]	90.73	92.58	17.83	3.58
Competition winner [78]	91.04	92.86	18.28	3.40
CMU-Nets [133]	91.57	93.55	15.85	2.92
ACMU-Nets	92.19	94.07	17.13	2.69

Table 7.2: Results of the proposed method with the revised training method on DIBCO 2017

Method	FM	F_{ps}	PSNR	DRD
Conventional training method	92.07	93.86	16.89	2.71
Revised training method	92.19	94.07	17.13	2.69

method and the proposed method is not a large value. According to this result, if CMU-Nets are designed for more complex tasks, the proposed method can be considered to avoid a problem with the order of arranged U-Net modules.

7.2.2 Qualitative Evaluations

In Fig. 7-1, ACMU-Nets show robust document image binarization results compared with the others. I can find out that the proposed method controls global and local features well according to erased noisy parts and fore-ground pixels' thickness. Indeed, since conventional CMU-Nets use dilation and erosion U-Net modules, there is a tiny difference in fore-ground pixels' thickness compared with the ground-truth image. However, since the SE-blocks can calibrate the necessary features, lots of noisy parts can be erased.

7.2.3 Effectiveness of SE-Nets in ACMU-Nets

As shown in Fig. 7-2, I confirmed that the inter-module skip-connections with densely connected SE-Nets (6 SE-Nets) cannot give a big performance boost to CMU-Nets.



Figure 7-1: Comparison results with the DIBCO 2017 winner's, CMU-Nets, and the proposed method. Images in the columns from left to right correspond to the original images, ground truth images, the result images of the DIBCO 2017 winner, the result images of CMU-Nets, and the result images of ACMU-Nets.

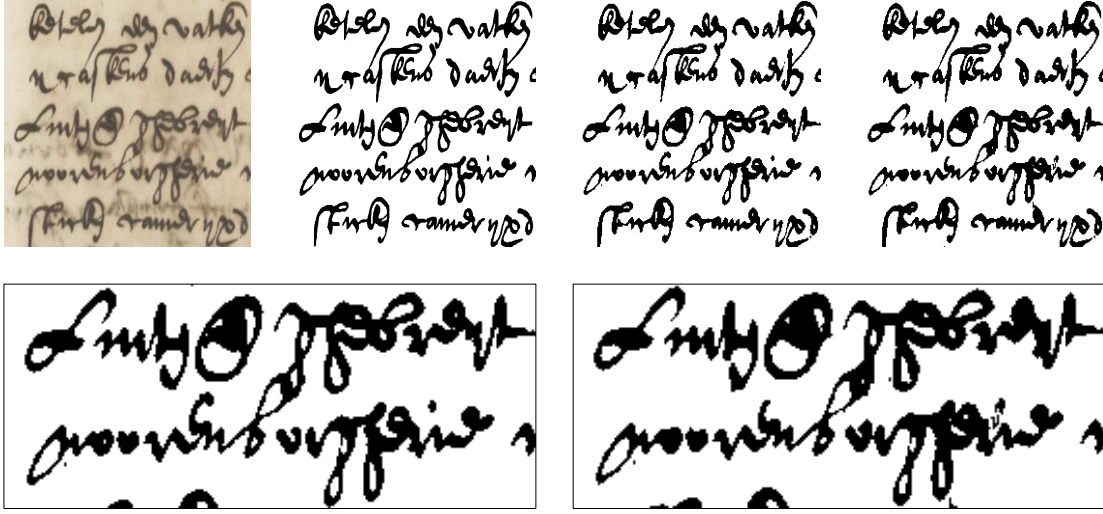


Figure 7-2: The output images of CMU-Nets with densely attached SE-blocks in the inter-module skip-connections. Images from the upper left to upper right correspond to the original image, ground truth image, the output image of CMU-Nets, the output image of CMU-Nets with densely attached SE-blocks in the inter-module skip-connections. The lower images are enlarged images of CMU-Nets and CMU-Nets with densely attached SE-blocks in the inter-module skip-connections. Since the result images of CMU-Nets with densely attached SE-blocks is thicker than the output of CMU-Nets, the utilization of densely attached SE-blocks is not proper in CMU-Nets.

Since the role of inter-module skip-connections in CMU-Net is to transmit localization information between modularized U-Nets, densely attached SE-Nets in the inter-module skip-connections can degrade the performance of ACMU-Nets. As shown in Fig. 7-2, the result of ACMU-Nets with densely attached SE-Nets in the inter-module skip-connections is thicker than the result of CMU-Nets. It means that excessive interference occurred in binarization process of ACMU-Nets.

According to Fig. 7-3, I found better performance improvement results than randomly initialized CMU-Nets compared with CMU-Nets with densely attached SE-Nets in the inter-module skip-connections. Since CMU-Nets use pre-trained U-Net modules, it is judged that the effect of the SE-block is much smaller, and thus the performance change is less. Therefore, the method of densely attached SE-Nets in

the inter-module skip-connections can improve the performance of randomly initialized CMU-Nets. In contrast, CMU-Nets, which are organized with pre-trained U-Net modules, cannot be improved by adopting the method of densely attached SE-Nets in the inter-module skip-connections. According to Fig. 7-2 and Fig. 7-3, excessive utilization of SE-Nets cannot improve the performance of ACMU-Nets. Moreover, in ACMU-Nets, I confirmed that the role of the inter-module skip-connections is to transmit global localization information between the connected U-Net modules.

Furthermore, since CMU-Nets utilize pre-trained U-Nets and inter-module skip-connections, parameter increase due to the SE-Net utilization should be considered when ACMU-Nets are designed.

7.3 Summary

In this chapter, ACMU-Nets are analyzed by quantitative and qualitative results. Since ACMU-Nets have two different strategies: modified training methods of individual U-Net modules and SE-Net utilization, the experimental analyses prove that the proposed methods can improve the ACMU-Nets' performance. For evaluating ACMU-Nets, the DIBCO 2017 dataset is used for the model test.

According to the evaluation of ACMU-Nets, I confirmed that the proposed model shows the improvement of document image binarization. ACMU-Nets achieve FM of 92.19% and F_{ps} of 94.07% for the DIBCO 2017 dataset. To evaluate the proposed SE-Net utilization method in ACMU-Nets, various comparative results were demonstrated. According to these comparative results, excessive SE-Net utilization in CMU-Nets cannot improve the performance of ACMU-Nets. Furthermore, excessive SE-Net utilization can occur excessive parameter increase in ACMU-Nets since ACMU-Nets are organized with four U-Nets and various skip-connections. Therefore, the proposed SE-Net utilization method is proper to improve the cascading architecture of U-Nets included pre-trained U-Net modules.

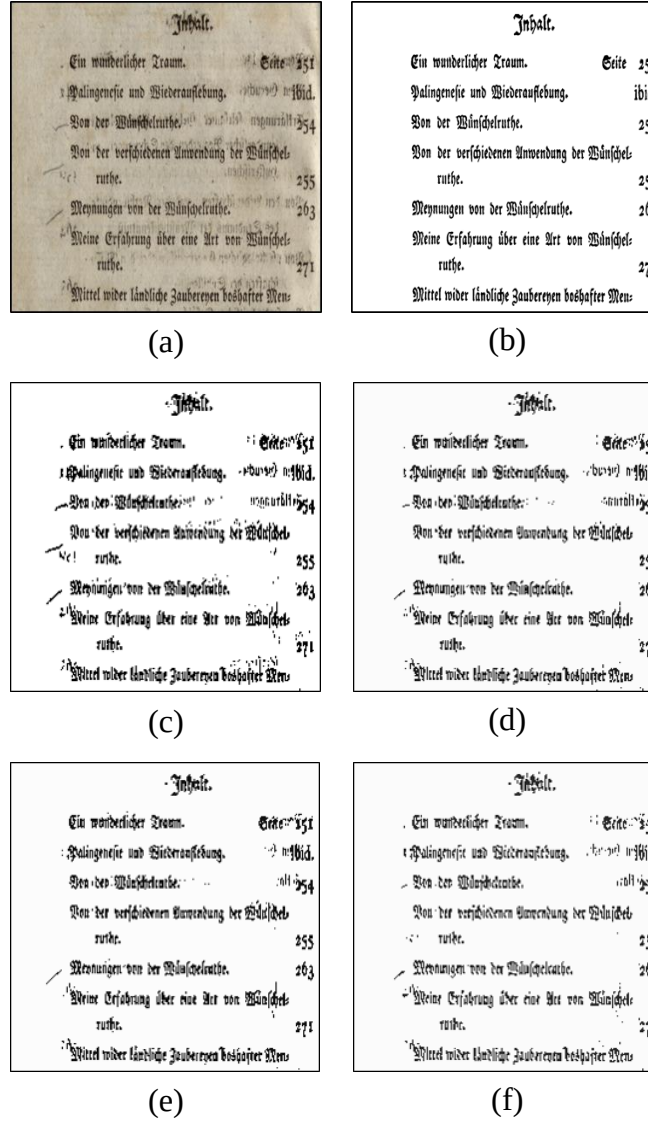


Figure 7-3: Effectiveness of SE-blocks between CMU-Nets and randomly initialized CMU-Nets. Images from upper left to lower right correspond to the original image, the ground truth image, the output image of randomly initialized CMU-Nets, the output image of randomly initialized CMU-Nets with densely attached SE-blocks in the inter-module skip-connections, the output image of CMU-Nets, and the output image of CMU-Nets with densely attached SE-blocks in the inter-module skip-connections. By comparing these output images, the method of densely attached SE-Nets in the inter-module skip-connections is not an effective strategy to improve CMU-Nets' performance compared with randomly initialized CMU-Nets.

Moreover, I adopted the revised training method for making pre-trained U-Net modules. By the quantitative evaluation of the revised training method, the proposed training method can improve the performance of ACMU-Nets. Moreover, unlike a document image binarization task, setting the optimal order of arranging pre-processing techniques is difficult in various complex image-to-image transformation tasks like biomedical image analysis. Therefore, the revised training method of individual U-Nets can improve ACMU-Nets' performance even though the arrangement of individual U-Net modules is revised.

Chapter 8

Conclusion

8.1 Conclusion

This thesis presented two novel models: CMU-Nets and ACMU-Nets. Since the purposes of this thesis are as follows: optimizing model training of U-Net-based model with insufficient training data and improving the performance of CMU-Nets at the test phase by using SE-Nets. These two proposed models demonstrated a significant improvement in document image binarization.

Firstly, I suggested CMU-Nets that were trained for historical document image binarization to prove the performance improvement with insufficient training data. CMU-Nets are organized with pre-trained U-Net modules. These U-Net modules were trained by using a sufficient amount of training data such as the MSCOCO dataset. Since the DIBCO data as the target training data is insufficient to train U-Net modules, I adopted a transfer learning method by training U-Net modules with the MSCOCO dataset. In the case of DIBCO 2017 winner [104], 70,000 newly generated training samples were utilized for the network training. However, since CMU-Nets can be optimized by only 800 images at the fine-tuning, I confirmed that the strategy of U-Net modularization enables the network to train with insufficient data.

After then, I cascaded these pre-trained U-Net modules by the optimal order of cascading U-Net modules. The cascading U-Net modules were connected by the proposed inter-module skip-connection method. The inter-module skip-connections were utilized for effective pixel-wise localization information between the cascaded U-Nets. In the end, I conduct a fine-tuning process of the cascading modular U-Nets by using the DIBCO datasets.

In the CMU-Nets research, I demonstrated the ability of CMU-Nets by quantitative and qualitative results. The proposed method proved that the utilization of pre-trained modular U-Nets and the proposed scheme of inter-module skip-connections in cascading U-Nets achieved robust binarization performance. The experiments were evaluated on all currently existing DIBCO and H-DIBCO datasets. It was able to maintain an average FM of $94.50 \pm 4.79\%$ and an average F_{ps} of $95.12 \pm 3.50\%$ for all DIBCO and H-DIBCO datasets. Furthermore, I justified some of the design choices, such as the proposed inter-module skip-connection scheme and module order.

Secondly, I suggested ACMU-Nets that were trained for historical document image binarization to prove the performance improvement at the test phase by adopting SE-Nets. The proposed model is based on CMU-Nets, utilized SE-Nets for improving the performance of the modified CMU-Nets. Also, I adopted the revised training method for making pre-trained U-Net modules.

According to the analysis of the quantitative and qualitative results, ACMU-Nets can be considered if someone wants to make cascading modularized U-Nets with SE-blocks. I also find out that simple usage of SE-blocks in cascading U-Nets can not make dramatic performance improvement. ACMU-Nets achieve FM of 92.19% and F_{ps} of 94.07% for the DIBCO 2017 dataset. Compared with the existing CMU-Nets, ACMU-Nets were improved (CMU-Nets: FM = 91.57% and F_{ps} = 93.55%).

In this thesis, by diverse quantitative and qualitative results, I confirmed that CMU-Nets and ACMU-Nets can conduct an effective complex image transformation task such as historical document image binarization. Since CMU-Nets are organized

with pre-trained modularized U-Nets as pre-processing, various complex image transformations can be solved by cascading them like rule-based systems.

8.2 Future Work

For future work, I hope that CMU-Nets and ACMU-Nets can contribute to the significant improvement of CNN-based models. I explain several potential research directions to further explore CMU-Nets and ACMU-Nets as follows.

- In this thesis, CMU-Nets and ACMU-Nets are suggested for historical document image binarization by cascading their pre-trained U-Net modules. When the proposed model was cascaded, pre-trained U-Net modules for pre-processing are connected by the optimal order. However, since the optimal order was decided by the general order of rule-based binarization systems, the optimal order for other tasks should be newly decided. Therefore, in future work, I want to design an automatic system to find the optimal cascading order.
- CNN-based models of cascading architecture can utilize the strategies of CMU-Nets for improving the performance of these models. Since CMU-Nets demonstrated a significant improvement by using pre-trained U-Net modules for solving an insufficient training data issue, the cascading architecture of CNN-based models can be trained by using insufficient data. Moreover, the proposed inter-module skip-connections can be considered when researchers design the cascading architecture of image transformation networks.
- The research of ACMU-Nets inspires researchers to utilize the attention mechanism in the cascading architecture of image transformation networks. The utilization strategy of ACMU-Nets can be applied to downsize the network size of cascading image-to-image transformation models as well as improve the performance of these models simultaneously.

- The downsizing of CMU-Nets should be considered. Even though CMU-Nets and ACMU-Nets have shown significant performance in complex image transformation tasks, these models have lots of layers since these models are organized with four U-Net models. Therefore, for downsizing CMU-Nets, we can adopt the distillation process.
- For controlling the output of the proposed model, I want to design the proposed model that included tunable parameters. The proposed model with tunable parameters can generate the output for the purpose by using insufficient training data. In addition, since the proposed model with tunable parameters can generate the optimal output by only learning the tunable parameters, an additional fine-tuning process does not need.

Bibliography

- [1] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” in *Graphics Gems IV*. Elsevier, 1994, pp. 474–485.
- [2] R. Garcia, T. Nicosevici, and X. Cufi, “On the way to solve lighting problems in underwater imaging,” in *OCEANS’02 MTS/IEEE*, vol. 2. IEEE, 2002, pp. 1018–1024.
- [3] O. Chum, T. Pajdla, and P. Sturm, “The geometric error for homographies,” *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 86–102, 2005.
- [4] E. Dubrofsky and R. J. Woodham, “Combining line and point correspondences for homography estimation,” in *International Symposium on Visual Computing*. Springer, 2008, pp. 202–213.
- [5] H.-C. Lee, *Introduction to color imaging science*. Cambridge University Press, 2005.
- [6] B. Funt, V. Cardei, and K. Barnard, “Learning color constancy,” in *Proceedings of Color and Imaging Conference*, vol. 1996, no. 1. Society for Imaging Science and Technology, 1996, pp. 58–60.
- [7] P. J. Green, “Bayesian reconstructions from emission tomography data using a modified em algorithm,” *IEEE Transactions on Medical Imaging*, vol. 9, no. 1, pp. 84–93, 1990.
- [8] D. L. Snyder, M. I. Miller, L. J. Thomas, and D. G. Politte, “Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography,” *IEEE Transactions on Medical Imaging*, vol. 6, no. 3, pp. 228–238, 1987.
- [9] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.
- [10] M. Rabbani, “Jpeg2000: Image compression fundamentals, standards and practice,” *Journal of Electronic Imaging*, vol. 11, no. 2, p. 286, 2002.

- [11] S. Di Zenzo, L. Cinque, and S. Levialdi, "Run-based algorithms for binary image analysis and processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 83–89, 1996.
- [12] S. Uchida, "Image processing and recognition for biological images," *Development, growth & differentiation*, vol. 55, no. 4, pp. 523–549, 2013.
- [13] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
- [14] P. J. Besl and R. C. Jain, "Three-dimensional object recognition," *ACM Computing Surveys*, vol. 17, no. 1, pp. 75–145, 1985.
- [15] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proceedings of the International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.
- [16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 886–893.
- [17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proceedings of the European Conference on Computer Vision*. Springer, 2006, pp. 404–417.
- [18] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive kalman filter," *Journal of Visual Communication and Image Representation*, vol. 17, no. 6, pp. 1190–1208, 2006.
- [19] M. S. Allili and D. Ziou, "Object tracking in videos using adaptive mixture models and active contours," *Neurocomputing*, vol. 71, no. 10-12, pp. 2001–2011, 2008.
- [20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, jan 1979.
- [21] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.
- [22] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, feb 2000.
- [23] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2005, pp. 60–65.

- [24] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [25] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [26] R. Girshick, “Fast R-CNN,” in *Proceedings of the 2015 IEEE International Conference of Computer Vision*, 2015, pp. 1440–1448.
- [27] S. He and L. Schomaker, “DeepOtsu: Document enhancement and binarization using iterative deep learning,” *Pattern Recognition*, vol. 91, pp. 379–390, 2019.
- [28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [29] R. Qian, Y. Yue, F. Coenen, and B. Zhang, “Traffic Sign Recognition with Convolutional Neural Network Based on Max Pooling Positions,” in *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2016, pp. 578–582.
- [30] B. K. Iwana, V. Frinken, and S. Uchida, “A Robust Dissimilarity-based Neural Network for Temporal Pattern Recognition,” in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2016, pp. 265–270.
- [31] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “CNN-RNN: A Unified Framework for Multi-label Image Classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285–2294.
- [32] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, “A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network,” in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, 2007, pp. 11–16.
- [33] J. Lee, B. K. Iwana, S. Ide, H. Hayashi, and S. Uchida, “Globally optimal object tracking with complementary use of single shot multibox detector and fully convolutional network,” in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2017, pp. 110–122.
- [34] H. Song, D. Suehiro, and S. Uchida, “Adaptive aggregation of arbitrary online trackers with a regret bound,” in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 681–689.

- [35] G. Atarsaikhan, B. K. Iwana, A. Narusawa, K. Yanai, and S. Uchida, “Neural font style transfer,” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, nov 2017.
- [36] H. Hayashi, K. Abe, and S. Uchida, “GlyphGAN: Style-consistent font generation based on generative adversarial networks,” *Knowledge-Based Systems*, vol. 186, p. 104927, dec 2019.
- [37] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, dec 2017.
- [38] J. Lee, H. Hayashi, W. Ohyama, and S. Uchida, “Page segmentation using a convolutional neural network with trainable co-occurrence features,” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2019, pp. 1023–1028.
- [39] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2015, pp. 3431–3440.
- [40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.
- [41] J. Kittler and J. Illingworth, “Minimum error thresholding,” *Pattern Recognition*, vol. 19, no. 1, pp. 41 – 47, 1986.
- [42] B. Javidi and C.-J. Kuo, “Joint transform image correlation using a binary spatial light modulator at the fourier plane,” *Applied Optics*, vol. 27, no. 4, pp. 663–665, 1988.
- [43] C. Tensmeyer and T. Martinez, “Document image binarization with fully convolutional neural networks,” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, nov 2017, pp. 99–104.
- [44] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, “Up-snet: A unified panoptic segmentation network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8818–8826.
- [45] K. He, J. Sun, and X. Tang, “Guided image filtering,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 1–14.

- [46] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, “O-cnn: Octree-based convolutional neural networks for 3d shape analysis,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.
- [47] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-scale Image Recognition,” in *Proceedings of the International Conference on Learning Representations*, 2015.
- [48] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 818–833.
- [49] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [50] H. Jiang, T. Shi, Z. Bai, and L. Huang, “Ahnnet: An application of attention mechanism and hybrid connection for liver tumor segmentation in ct volumes,” *IEEE Access*, vol. 7, pp. 24 898–24 909, 2019.
- [51] K. Nishimura, R. Bise *et al.*, “Weakly supervised cell instance segmentation by propagating from detection response,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 649–657.
- [52] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, “Binarization of degraded document images based on hierarchical deep supervised network,” *Pattern Recognition*, vol. 74, pp. 568–586, 2018.
- [53] S. Kang, B. K. Iwana, and S. Uchida, “Complex image processing with less data—document image binarization by integrating multiple pre-trained u-net modules,” *Pattern Recognition*, vol. 109, p. 107577, 2021.
- [54] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.
- [55] P. Biasutti, V. Lepetit, J.-F. Aujol, M. Brédif, and A. Bugeau, “Lu-net: An efficient network for 3d lidar point cloud semantic segmentation based on end-to-end-learned 3d features and u-net,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [56] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

- [57] Z. Yang and R. Nevatia, “A multi-scale cascade fully convolutional network face detector,” in *Proceedings of the International Conference on Pattern Recognition*. IEEE, dec 2016, pp. 633–638.
- [58] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.
- [59] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, “Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks,” in *Proceedings of the International Conference on Computer Vision*, 2015, pp. 2956–2964.
- [60] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in *Proceedings of the British Machine Vision Conference*, 2018.
- [61] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, “BAM: Bottleneck attention module,” *arXiv preprint arXiv:1807.06514*, 2018.
- [62] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.
- [63] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
- [64] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [65] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 843–852.
- [66] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.
- [67] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

- [68] J. Munoz-Bulnes, C. Fernandez, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, “Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection,” in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2017, pp. 366–371.
- [69] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2. IEEE, 2001, pp. 416–423.
- [70] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” *arXiv preprint arXiv:1902.09063*, 2019.
- [71] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.
- [72] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *Proceedings of the AAAI conference on Artificial Intelligence*, 2020.
- [73] L. Y. Pratt, “Discriminability-based transfer between neural networks,” in *Advances in Neural Information Processing Systems*, 1993, pp. 204–211.
- [74] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual Attention Network for Image Classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.
- [75] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi *et al.*, “USE-Net: incorporating Squeeze-and-Excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets,” *arXiv preprint arXiv:1904.08254*, 2019.
- [76] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “H-DIBCO 2010 - handwritten document image binarization competition,” in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, nov 2010, pp. 727–732.
- [77] —, “ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012),” in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, sep 2012, pp. 817–822.

- [78] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, “ICDAR 2017 competition on document image binarization (DIBCO 2017),” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, nov 2017, pp. 1395–1403.
- [79] B. Gatos, K. Ntirogiannis, and I. Pratikakis, “ICDAR 2009 document image binarization contest (DIBCO 2009),” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 1375–1382.
- [80] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [81] S. Behnke, *Hierarchical neural networks for image interpretation*. Springer, 2003, vol. 2766.
- [82] P. Simard, D. Steinkraus, and J. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 3, no. 2003. IEEE, 2003.
- [83] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2015.
- [84] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [85] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, “Convolutional Neural Network Committees for Handwritten Character Classification,” in *Proceedings of the International Conference on Document Analysis and Recognition*, 2011, pp. 1135–1139.
- [86] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification,” in *Processing of the International Joint Conference on Artificial Intelligence*, 2011.
- [87] “A survey of multi-source domain adaptation,” *Information Fusion*, vol. 24, pp. 84 – 92, 2015.
- [88] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2015, pp. 1180–1189.

- [89] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.
- [90] R. Kasturi, L. O’gorman, and V. Govindaraju, “Document Image Analysis: A primer,” *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002.
- [91] M. S. Hitam, E. A. Awalludin, W. N. J. H. W. Yussof, and Z. Bachok, “Mixture contrast limited adaptive histogram equalization for underwater image enhancement,” in *Proceedings of the International Conference on Computer Applications Technology (ICCAT)*. IEEE, 2013, pp. 1–5.
- [92] A. Doucet, G. Kazai, and J.-L. Meunier, “Icdar 2011 book structure extraction competition,” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 1501–1505.
- [93] K. Y. Wong, R. G. Casey, and F. M. Wahl, “Document analysis system,” *IBM Journal of Research and Development*, vol. 26, no. 6, pp. 647–656, nov 1982.
- [94] F. Farooq, V. Govindaraju, and M. Perrone, “Pre-processing methods for handwritten arabic documents,” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2005, pp. 267–271.
- [95] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, “Ocr binarization and image pre-processing for searching historical documents,” *Pattern Recognition*, vol. 40, no. 2, pp. 389–397, 2007.
- [96] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the International Conference on Machine Learning*, 2010.
- [97] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for document image classification,” in *Proceedings of the International Conference on Pattern Recognition*. IEEE, 2014, pp. 3168–3172.
- [98] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, “Convolutional neural networks for page segmentation of historical document images,” in *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 2017, pp. 965–970.
- [99] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, “AON: Towards arbitrarily-oriented text recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5571–5579.

- [100] X.-H. Li, F. Yin, and C.-L. Liu, "Page object detection from pdf document images by deep structured prediction and supervised clustering," in *Proceedings of the International Conference on Pattern Recognition*. IEEE, 2018, pp. 3627–3632.
- [101] C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," in *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 2017, pp. 388–393.
- [102] I.-J. Kim and X. Xie, "Handwritten Hangul Recognition Using Deep Convolutional Neural Networks," *International Journal on Document Analysis and Recognition*, vol. 18, pp. 1–13, 2015.
- [103] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning Spatial-semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1903–1917, 2018.
- [104] P. Bezmaternykh, D. Ilin, and D. Nikolaev, "U-Net-bin: hacking the document image binarization contest," *Computer Optics*, vol. 43, pp. 825–832, 10 2019.
- [105] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018)," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, aug 2018, pp. 489–493.
- [106] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of Neural Networks Using Dropconnect," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1058–1066.
- [107] S. Uchida, S. Ide, B. K. Iwana, and A. Zhu, "A further step to perfect accuracy by training CNN with larger data," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, oct 2016, pp. 405–410.
- [108] C. Wick and F. Puppe, "Fully convolutional neural networks for page segmentation of historical document images," in *Proceedings of the 2018 International Workshop on Document Analysis Systems*. IEEE, apr 2018, pp. 287–292.
- [109] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the International Conference on 3D Vision*. IEEE, oct 2016.
- [110] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

- [111] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: Learning dense volumetric segmentation from sparse annotation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.
- [112] P. Esser and E. Sutter, “A variational u-net for conditional appearance and shape generation,” in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018, pp. 8857–8866.
- [113] A. Sevastopolsky, S. Drapak, K. Kiselev, B. M. Snyder, J. D. Keenan, and A. Georgievskaya, “Stack-U-Net: refinement network for improved optic disc and cup image segmentation,” in *Proceedings of the Medical Imaging 2019: Image Processing*, vol. 10949. SPIE, mar 2019.
- [114] 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.
- [115] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Chellappa, “Stacked u-nets for ground material segmentation in remote sensing imagery,” in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2018.
- [116] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and H. K. Maier-Hein, “nnU-Net: Self-adapting Framework for U-Net-based Medical Image Segmentation,” *arXiv preprint arXiv:1809.10486*, 2018.
- [117] Z. Tang, X. Peng, S. Geng, Y. Zhu, and D. N. Metaxas, “CU-Net: Coupled U-Nets,” in *Proceedings of the British Machine Vision Conference*, 2018.
- [118] P. W. Palumbo, P. Swaminathan, and S. N. Srihari, “Document Image Binarization: Evaluation Of Algorithms,” in *Proceedings of the Applications of Digital Image Processing IX*, A. G. Tescher, Ed. SPIE, dec 1986.
- [119] F. Shafait, D. Keysers, and T. M. Breuel, “Efficient implementation of local adaptive thresholding techniques using integral images,” in *Document recognition and retrieval XV*, vol. 6815. International Society for Optics and Photonics, 2008, p. 681510.
- [120] R. Estrada and C. Tomasi, “Manuscript bleed-through removal via hysteresis thresholding,” in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 753–757.

- [121] S. Lu and C. Tan, "Binarization of badly illuminated document images through shading estimation and compensation," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2007, pp. 312–316.
- [122] Y. Akbari, A. S. Britto, S. Al-maadeed, and L. S. Oliveira, "Binarization of degraded document images using convolutional neural networks based on predicted two-channel images," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2019.
- [123] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, "Degraded document image binarization using structural symmetry of strokes," *Pattern Recognition*, vol. 74, pp. 225–240, feb 2018.
- [124] X. Peng, C. Wang, and H. Cao, "Document binarization via multi-resolutional attention model with DRD loss," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2019.
- [125] R. Mondal, D. Chakraborty, and B. Chanda, "Learning 2d morphological network for old document image binarization," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2019.
- [126] F. Hollaus, S. Brenner, and R. Sablatnig, "CNN based binarization of Multi-Spectral document images," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2019.
- [127] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2016.
- [128] T. Mondal, M. Coustaty, P. Gomez-Kramer, and J.-M. Ogier, "Learning free document image binarization based on fast fuzzy c-means clustering," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2019.
- [129] W. Niblack, "An introduction to digital image processing," *Prentice Hall*, 01 1986.
- [130] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2010, pp. 2963–2970.
- [131] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition*, vol. 16, no. 3, pp. 247–258, sep 2012.

- [132] D. Ghosh, T. Dube, and A. Shivaprasad, “Script Recognition: A Review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010.
- [133] S. Kang, S. Uchida, and B. K. Iwana, “Cascading modular u-nets for document image binarization,” in *Proceedings of the International Conference on Document Analysis and Recognition*, 2019, pp. 675–680.
- [134] 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 *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
- [135] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, “COCO-Text: Dataset and benchmark for text detection and recognition in natural images,” *arXiv preprint arXiv:1601.07140*, 2016.
- [136] K. Wongsritong, K. Kittayaruasiriwat, F. Cheevasuvit, K. Dejhan, and A. Somboonkaew, “Contrast enhancement using multipeak histogram equalization with brightness preserving,” in *Proceedings of the IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems*. IEEE, 1998, pp. 455–458.
- [137] Y. C. Hum, K. W. Lai, and M. I. M. Salim, “Multiobjectives bihistogram equalization for image contrast enhancement,” *Complexity*, vol. 20, no. 2, pp. 22–36, jan 2014.
- [138] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 12 2015.
- [139] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “Performance evaluation methodology for historical document image binarization,” *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 595–609, feb 2013.
- [140] H. Lu, A. Kot, and Y. Shi, “Distance-reciprocal distortion measure for binary document images,” *IEEE Signal Processing Letter*, vol. 11, no. 2, pp. 228–231, feb 2004.
- [141] J. Guo, C. He, and X. Zhang, “Nonlinear edge-preserving diffusion with adaptive source for document images binarization,” *Applied Mathematics and Computation*, vol. 351, pp. 8–22, 2019.
- [142] J. Zhao, C. Shi, F. Jia, Y. Wang, and B. Xiao, “Document image binarization with cascaded generators of conditional generative adversarial networks,” *Pattern Recognition*, vol. 96, p. 106968, 2019.

- [143] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR 2011 document image binarization contest (DIBCO 2011),” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, sep 2011, pp. 1506–1510.
- [144] —, “ICDAR 2013 document image binarization contest (DIBCO 2013),” in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, aug 2013, pp. 1471–1476.
- [145] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “ICFHR 2014 competition on handwritten document image binarization (H-DIBCO 2014),” in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, sep 2014, pp. 809–813.
- [146] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, “ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016),” in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, oct 2016, pp. 619–623.
- [147] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
- [148] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.
- [149] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [150] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [151] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-XL: Attentive language models beyond a fixed-length context,” *arXiv preprint arXiv:1901.02860*, 2019.
- [152] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2048–2057.
- [153] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.

- [154] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, “Multimodal residual learning for visual QA,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2016, pp. 361–369.
- [155] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [156] J. Johnson, A. Karpathy, and L. Fei-Fei, “DenseCap: Fully convolutional localization networks for dense captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.
- [157] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.
- [158] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.
- [159] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.
- [160] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, “Human attention in visual question answering: Do humans and deep networks look at the same regions?” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 932–937.
- [161] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.
- [162] V.-Q. Nguyen, M. Suganuma, and T. Okatani, “Efficient attention mechanism for handling all the interactions between many inputs with application to visual dialog,” *arXiv preprint arXiv:1911.11390*, 2019.
- [163] M. R. Farazi, S. H. Khan, and N. Barnes, “Question-Agnostic Attention for Visual Question Answering,” *arXiv preprint arXiv:1908.03289*, 2019.
- [164] V. Mnih, N. Heess, A. Graves *et al.*, “Recurrent models of visual attention,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.

- [165] N. Sarafianos, X. Xu, and I. A. Kakadiaris, “Deep imbalanced attribute classification using visual attention aggregation,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 680–697.
- [166] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, “Learning attentions: residual attentional siamese network for high performance online visual tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4854–4863.
- [167] M. Sun, Y. Yuan, F. Zhou, and E. Ding, “Multi-attention multi-class constraint for fine-grained image recognition,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 805–821.
- [168] X. Cheng, X. Li, J. Yang, and Y. Tai, “Sesr: Single image super resolution with recursive squeeze and excitation networks,” in *Proceedings of the 2018 International Conference on Pattern Recognition*. IEEE, 2018, pp. 147–152.
- [169] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “CBAM: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.
- [170] A. K. Bhunia, A. Konwer, A. K. Bhunia, A. Bhowmick, P. P. Roy, and U. Pal, “Script Identification in Natural Scene Image and Video Frames Using an Attention Based Convolutional-LSTM Network,” *Pattern Recognition*, vol. 85, pp. 172–184, 2019.
- [171] H. Xie, S. Fang, Z.-J. Zha, Y. Yang, Y. Li, and Y. Zhang, “Convolutional Attention Networks for Scene Text Recognition,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–17, 2019.
- [172] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [173] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention u-net for lesion segmentation,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.
- [174] Q. Jin, Z.-P. Meng, C. Sun, L. Wei, and R. Su, “Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans,” *ArXiv*, vol. abs/1811.01328, 2018.
- [175] T. Fan, G. Wang, Y. Li, and H. Wang, “MA-Net: A multi-scale attention network for liver and tumor segmentation,” *IEEE Access*, vol. 8, pp. 179 656–179 665, 2020.

- [176] R. Li, M. Li, J. Li, and Y. Zhou, “Connection sensitive attention u-net for accurate retinal vessel segmentation,” *arXiv preprint arXiv:1903.05558*, 2019.
- [177] “Hierarchical multi-modal fusion fcn with attention model for rgb-d tracking,” *Information Fusion*, vol. 50, pp. 1 – 8, 2019.
- [178] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, “PSAnet: Point-wise spatial attention network for scene parsing,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 267–283.
- [179] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, “Scene segmentation with dual relation-aware attention network,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2020.
- [180] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, “Abdominal multi-organ segmentation with organ-attention networks and statistical fusion,” *Medical Image Analysis*, vol. 55, pp. 88–102, 2019.
- [181] A. G. Roy, N. Navab, and C. Wachinger, “Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks,” *IEEE Transactions on Medical Imaging*, vol. 38, pp. 540–549, 2019.
- [182] W. Zhu, Y. Huang, H. Tang, Z. Qian, N. Du, W. Fan, and X. Xie, “Anatomynet: Deep 3d squeeze-and-excitation U-nets for fast and fully automated whole-volume anatomical segmentation,” *bioRxiv*, p. 392969, 2018.
- [183] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [184] “Fully automated segmentation on brain ischemic and white matter hyperintensities lesions using semantic segmentation networks with squeeze-and-excitation blocks in mri,” *Informatics in Medicine Unlocked*, vol. 21, p. 100440, 2020.