DP-LinkNet: A convolutional network for historical document image binarization

Wei Xiong[1,2,*], Xiuhong Jia[1], Dichun Yang[1], Meihui Ai[1], Lirong Li[1], and Song Wang[2,*]
[1] School of Electrical and Electronic Engineering, Hubei University of Technology
Wuhan, Hubei 430068 China
[e-mail: xw@mail.hbut.edu.cn]
[2] Department of Computer Science and Engineering, University of South Carolina
Columbia, SC 29201 USA
[e-mail: songwang@cec.sc.edu]
[*]Corresponding authors: Wei Xiong and Song Wang

Received October 17, 2020; revised January 19, 2021; revised March 1, 2021; accepted March 30, 2021; published May 31, 2021

Abstract

Document image binarization is an important pre-processing step in document analysis and archiving. The state-of-the-art models for document image binarization are variants of encoder-decoder architectures, such as FCN (*fully convolutional network*) and U-Net. Despite their success, they still suffer from three limitations: (1) reduced feature map resolution due to consecutive strided pooling or convolutions, (2) multiple scales of target objects, and (3) reduced localization accuracy due to the built-in invariance of *deep convolutional neural networks* (DCNNs). To overcome these three challenges, we propose an improved semantic segmentation model, referred to as DP-LinkNet, which adopts the D-LinkNet architecture as its backbone, with the proposed *hybrid dilated convolution* (HDC) and *spatial pyramid pooling* (SPP) modules between the encoder and the decoder. Extensive experiments are conducted on recent *document image binarization competition* (DIBCO) and *handwritten document image binarization competition* (H-DIBCO) benchmark datasets. Results show that our proposed DP-LinkNet outperforms other state-of-the-art techniques by a large margin. Our implementation and the pre-trained models are available at https://github.com/beargolden/DP-LinkNet.

Keywords: Degraded document image binarization, semantic segmentation, DP-LinkNet, encoder-decoder architecture, hybrid dilated convolution (HDC), spatial pyramid pooling (SPP)

This work was supported in part by National Natural Science Foundation of China (61571182, 61601177), Natural Science Foundation of Hubei Province of China (2019CFB530, 2019ZYYD020), China Scholarship Council (201808420418), and Hubei Provincial Department of Education (B2019042).

http://doi.org/10.3837/tiis.2021.05.011 **ISSN : 1976-7277**

1. Introduction

The purpose of document image binarization (also referred to as segmentation) is to convert gray-scale or color images into binary images by labeling each pixel as foreground (black) or background (white). Binarization plays a key role in *document analysis and recognition* (DAR) systems, and the performance of this stage is crucial for subsequent tasks, such as *optical character recognition* (OCR) [1]. However, binarization is extremely challenging due to the long storage time and poor storage environment of historical documents, resulting in severe degradation, to name a few, paper aging, text stroke fading, ink bleed through, page stains, and library seal. **Fig. 1** shows several degraded historical document image samples selected from *document image binarization competition* (DIBCO) datasets and the Bickley diary dataset.

(a) comprehensive degradation (b) paper creasing and text stroke fading (c) page stains (d) library seal and paper creasing (e) paper aging and text stroke fading (f) ink bleed through (g) papyrus paper fragments

Fig. 1. Degraded historical document image samples, (a) selected from the Bickley diary dataset and (b)-(g) selected from the DIBCO series dataset

DIBCO 2009 [2], 2011 [3], 2013 [4], 2017 [5], 2019 [6] and H-DIBCO 2010 [7], 2012 [8], 2014 [9], 2016 [10], 2018 [11] show recent advances in the binarization of degraded historical document images. We have participated in such academic competitions since 2017. Our Laplacian energy-based segmentation method won the first place in ICFHR 2018 competition on *handwritten document image binarization* (H-DIBCO 2018) [11]. Later, our improved binarization method based on D-LinkNet won the first place in ICDAR 2019 time-quality binarization competition on photographed document images taken by Motorola Z1 and Galaxy Note4 with flash off, and second and third places on binarization of photographed document images taken by the same mobile devices with flash on, respectively [12].

This paper presents our winning algorithm in ICDAR 2019 time-quality binarization competition on photographed document images. The proposed method is based on the D-LinkNet architecture [13]. To the best of our knowledge, manual extraction of text stroke features using traditional feature engineering is inadequate and subject to bias, especially when dealing with extremely degraded or severely damaged pages of historical antiquities.

Therefore, deep learning-based approaches are not only a good alternative, but also the current trend. The state-of-the-art models for semantic segmentation are variants of encoder-decoder architectures, such as FCN [14] and U-Net [15]. Despite their success, they still suffer from three limitations: (1) reduced feature map resolution due to consecutive strided pooling or convolutions, (2) multiple scales of target objects, and (3) reduced localization accuracy due to the built-in invariance of deep CNNs.

To overcome the first obstacle and efficiently generate dense feature maps, we adopt a **hybrid dilated convolution** layer to make intermediate feature maps denser. Compared to standard convolutional layers that use larger convolution kernels, dilated convolutions can increase the receptive field size without decreasing the spatial resolution of intermediate feature maps. To solve the second problem, the images can be cropped at different scales and then the feature maps can be fused. Although this approach is effective, it introduces too much computational cost. Inspired by the **spatial pyramid pooling**, we subsample the input feature maps at different rates to further encode global contextual information, and the target information can be obtained at different scales. The built-in invariance of deep convolutional neural networks is desirable for classification tasks, but may hinder dense prediction tasks, e.g., semantic segmentation, where abstraction of spatial information is undesirable. One way to address the third problem is to add **skip connections** to extract features at different levels, and fuse them at the decoder to obtain segmentation results.

Our contributions are three folds. First, we propose a combination of hybrid dilated convolution and spatial pyramid pooling to further aggregate multi-scale contextual features and encode multi-scale global contextual information. Second, we integrate the proposed hybrid dilated convolution and spatial pyramid pooling modules into the encoder-decoder architecture for degraded historical document image binarization. Last but not least, we apply a test time augmentation strategy to further improve the robustness of the proposed document image binarization method, and the experimental results show that our method outperforms other state-of-the-art techniques by a large margin.

The remainder of this paper is organized as follows. Section 2 reviews the related work on degraded document image binarization. Section 3 and 4 present the proposed architecture and the detailed implementations, respectively. Section 5 evaluates the binarization performance and time complexity of the proposed model. Section 6 concludes the paper.

2. Related Work

Document image binarization or segmentation can be roughly divided into global and local thresholding methods [16, 17].

Global thresholding, e.g., Otsu's method [18], computes an optimal threshold for the entire image to maximize the interclass variance or equivalently minimize the intraclass variance of foreground and background pixels. Global thresholding can give satisfactory results when the image histogram follows a bimodal distribution, but will generally fail when dealing with low-quality images.

Locally adaptive thresholding estimates a different threshold for each pixel in the image based on the analysis of neighborhood statistics, e.g., Niblack's [19], Sauvola's [20], and Wolf's [21] methods use local mean and standard deviation, while Bernsen's [22] and Herk's [23] methods employ local contrast. Local thresholding approaches generally have better performance than global counterparts. But the main drawbacks of these local methods are that the thresholding performance depends heavily on the sliding window size and thus on the text stroke width.

Degraded historical document image binarization has been a hot topic in the past decade since the first DIBCO was held in 2009. Previous studies on document image binarization are often based on edge detection. For instance, Su et al. [24, 25] propose to use local maximum and minimum to detect high-contrast pixels, usually located near text stroke edges. Lu et al. [26] present a binarization technique based on document background estimation and stroke edge detection. Jia et al. [27] present an effective method for document image binarization using *structural symmetric pixels* (SSPs), which are located at the edges of text strokes, and can be extracted from those with large gradient magnitude and opposite gradient direction.

Markov random fields (MRFs) [28] and *conditional random fields* (CRFs) [29] are used for degraded document image binarization. Howe [30, 31] presents an energy-based segmentation that uses graph cut optimization to solve the energy minimization problem of the objective function, which combines the Laplacian operator and the Canny edge detector. Since text edge detection is embedded in energy-based segmentation, such methods are less effective for dealing with low-contrast or severely degraded document images.

Active contour models, also referred to as snakes, are proposed for degraded historical document image binarization. Rivest-Hénault et al. [32] introduce a local linear level set framework, and Hadjadj et al. [33] also present a technique based on active contour evolution. Since the level set method is a local optimization method, it has a high time complexity and a tendency to fall into the nearest local minimum.

Statistical learning-based approaches are proposed to binarize historical document images as well. Chen et al. [34] propose a parallel non-parametric binarization method for degraded document images. Xiong et al. [35] present a *support vector machine* (SVM) based technique for binarization of degraded document images. Bhowmik et al. [36] introduced a *game theory inspired document image binarization* (GiB) technique. In general, the main disadvantage of these learning-based methods is that only low-level or hand-crafted features are used to obtain segmentation results. As a result, it is difficult to design representative features for different applications, and the designed features work well for one type of images, but often fail on another. Therefore, a general method for extracting features is still lacking.

With the development of *convolutional neural networks* (CNNs) in the field of document analysis and recognition, multi-scale and adaptive feature extraction and learning using deep network models has become a feasible approach for document image binarization. Tensmeyer and Martinez [37] present a *fully convolutional network* (FCN), which combines F-measure and pseudo F-measure losses for document image binarization tasks. Vo et al. [38] propose a hierarchical *deep supervised network* (DSN) for binarization of degraded document images. Calvo-Zaragoza and Gallego [39] choose the *residual encoder- decoder network* (RED-Net) [40] as the backbone of their *selectional auto-encoder* (SAE) architecture for document image binarization. Bezmaternykh et al. [41] present a historical document image binarization based on U-Net [15], originally designed for biomedical image segmentation. Zhao et al. [42] consider binarization as an image-to-image generation task and propose a binarization method for document images with *conditional generative adversarial networks* (cGANs). Peng et al. [43] propose a deep learning framework to infer the probabilities of text regions through a multi-resolution attentional model, which is then fed into a *convolutional conditional random field* (ConvCRF) to obtain the resulting binary image.

Although the application of convolutional neural networks can significantly improve the binarization performance, a common challenge for all these deep network models is how to find a more generalized scale selection or combination approach to better extract contextual information, which is also the motivation for the proposed architecture in this paper.

3. Proposed Architecture: DP-LinkNet

The proposed DP-LinkNet uses D-LinkNet [13] and LinkNet [44], with a pre-trained encoder as the backbone. As depicted in **Fig. 2**, it consists of four main parts: the encoder (part A), the *hybrid dilated convolution* (HDC) module (part B), the *spatial pyramid pooling* (SPP) module (part C), and the decoder (part D). The encoder extracts text stroke features with deep semantic information. The HDC module expands the receptive field size and aggregates multi-scale contextual features, while the SPP module encodes the output of the HDC with multi-kernel pooling. The combination of the HDC and SPP modules will produce enriched higher-level abstract feature maps. The decoder then maps the low-resolution feature map output from the central part back to the size of the input image for pixel-by-pixel classification. Although there are several subtle and important differences, what distinguishes our proposed DP-LinkNet from the two models mentioned above is that the LinkNet [44] contains only part A and D, while D-LinkNet [13] additionally contains part B.

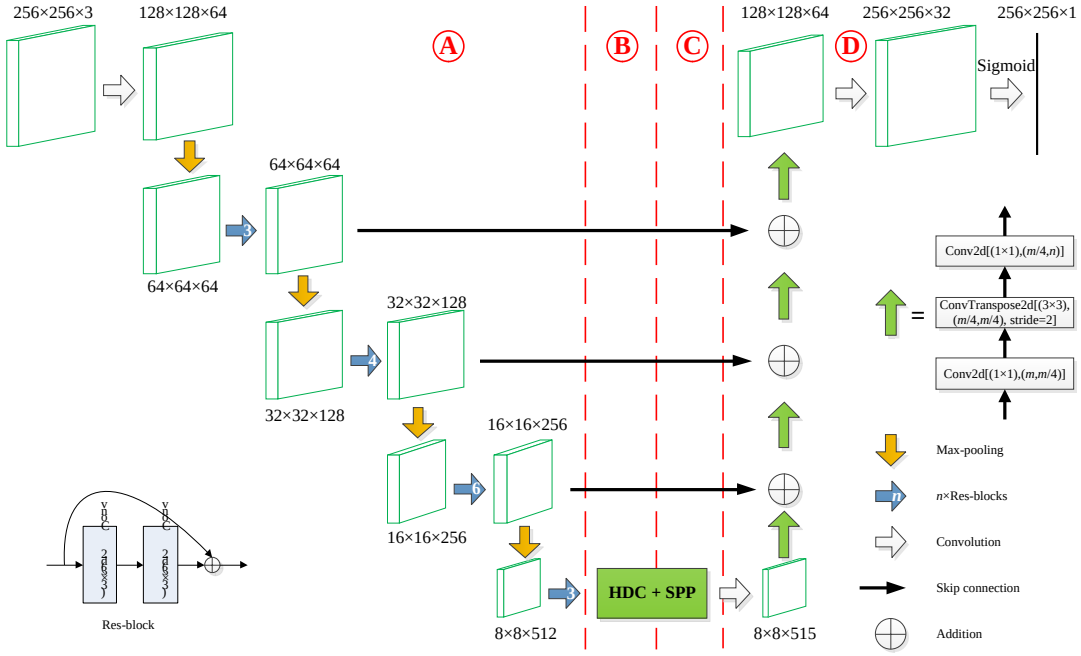


Fig. 2. The proposed DP-LinkNet architecture

3.1 Part A: Encoder

The original LinkNet [44] adopts the 18-layer ResNet as its encoder, while the D-LinkNet [13] employs the 34-layer ResNet as its encoder. Considering that the D-LinkNet achieved the first place in CVPR 2018 DeepGlobe Road Extraction Challenge, we decide to use the D-LinkNet pre-trained on the ImageNet [45] dataset as the encoder of our proposed architecture.

The first layer of our model is a 7×7 convolution layer with 64 output channels and a stride of 2, followed by a *batch normalization* (BN) layer, a *rectified linear unit* (ReLU) activation layer, and a 3×3 maximum pooling layer with a stride of 2. The rest of the encoder follows the four ResNet-34 encoder modules, consisting of 3, 4, 6, and 3 *residual blocks* (Res-blocks, as shown in the lower left panel of **Fig. 2**), respectively. The number of channels in the first module is the same as the number of input channels of the module. Since a 3×3 convolution

layer with a stride of 2 has already been used in the first residual block of each subsequent module, the number of channels is doubled and the spatial resolution of feature maps is reduced by half, compared to the previous module.

3.2 Part B: Hybrid Dilated Convolution (HDC)

At the encoder, a series of convolutions and downsampling are performed. This helps extract high-level features, but reduces the resolution of the feature map and may lead to the loss of spatial information. Dilated convolution is an alternative to downsampling operations and has been widely used for semantic or instance segmentation [46]. It is generally available in two types of modes, namely parallel mode [47] and cascade mode [48], both of which have demonstrated significant improvement in segmentation accuracy.

In dilated convolutions, the dilation rate r indicates inserting $r-1$ zeros between the kernel weights or subsampling the feature map by a factor of $r-1$. For a 1-dilated convolutional kernel of size $k \times k$, the size of the resulting r-dilated convolution kernel is:

$$\hat{k} = k + (k-1) \times (r-1) = r \times (k-1) + 1 \tag{1}$$

Considering the stride s in the l^{th} layer ($l = 1, 2, \ldots, L$), the *receptive field* (RF) size of each r-dilated convolution layer is:

$$\begin{cases} RF_l = s_{l-1} \times RF_{l-1} + \left(\hat{k}_{l-1} - s_{l-1}\right) \\ RF_0 = 1 \end{cases} \tag{2}$$

We solve the recursive equation and obtain:

$$RF_L = 1 + \sum_{l=1}^{L} \left(\left(\hat{k}_l - 1\right) \prod_{i=1}^{l-1} s_i \right) \tag{3}$$

Inspired by the fact that dilated convolutions exponentially increase the receptive field size without decreasing the spatial resolution of intermediate feature maps, we exploit the advantages of both modes and combine them using a shortcut connection. The proposed hybrid dilated convolution module is shown in **Fig. 3**. It contains dilated convolutions in both parallel and cascade modes. Each branch consists of 1 to 3 cascaded dilated convolutions with a kernel size of 3×3 and dilation rates of 1, 2, and 4, respectively. Therefore, the receptive field size of each branch will be 3, 7, and 15, respectively. As mentioned in Subsection 3.1, the encoder of ResNet-34 contains 5 strided convolution layers, equivalently downsampling layers, so if an image patch of 256×256 goes through the encoder part, the size of the output feature map will be 8×8.

According to Zhou [13] and most other researchers, the HDC can only contain two stacked cascades of dilated convolution branches at the most, so that the receptive field size of the cascaded dilated convolutions can be comparable to the output feature map size of the last stage of the encoder, or the former is slightly smaller than the latter. However, we found a temporarily unexplained phenomenon in our experiments that the segmentation performance is slightly higher, when the HDC contains a branch with a maximum receptive field size of 15, than the case when the HDC does not contain this branch. Therefore, in our proposed model, the three cascaded dilated convolution branches are summed with the feature map itself before feeding into the subsequent spatial pyramid pooling module.

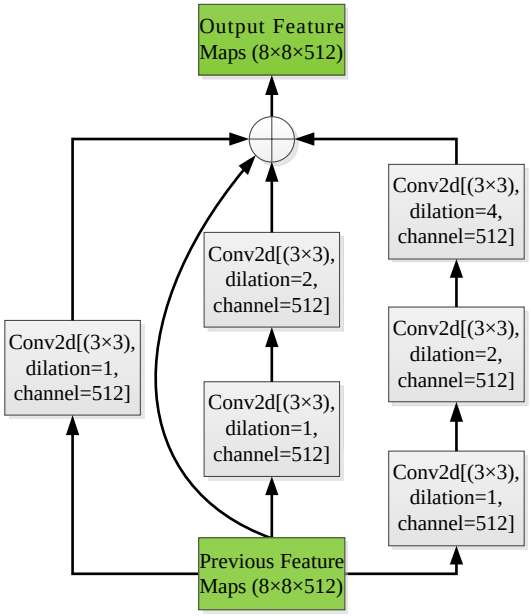


Fig. 3. Hybrid Dilated Convolution (HDC) module

3.3 Part C: Spatial Pyramid Pooling (SPP)

One of the major challenges in the binarization of degraded historical document images is the diversity of text strokes and the complexity of document degradations, and it is difficult to find a general algorithm to handle a variety of low-quality documents. Although the previous HDC module is able to obtain different sizes of receptive fields and aggregate multi-scale features through parallel and cascaded dilated convolutions, we believe that it is still difficult to detect or identify objects of different sizes with a fixed-sized field-of-view.

Inspired by the SPP-Net [49] for image classification and object detection, we adopt the *spatial pyramid pooling* (SPP) strategy to solve the above problem, but with several subtle and important differences. As presented in **Fig. 4**, our proposed SPP module encodes the global contextual information with three different sizes of receptive fields. We apply three maximum pooling operations to the feature maps output from the HDC, with convolution kernel sizes of 2×2, 3×3, and 5×5, respectively. The output shape of the multi-size pooling can be precisely described as:

$$h_{out} = \left\lfloor \frac{h_{in} + 2 \times p - d \times (k-1) - 1}{s} + 1 \right\rfloor$$
$$w_{out} = \left\lfloor \frac{w_{in} + 2 \times p - d \times (k-1) - 1}{s} + 1 \right\rfloor \quad (4)$$

where h_{in} and w_{in} are the input height and width of the feature map, h_{out} and w_{out} are the output height and width, k and s denote the kernel size and the stride (whose default value is the same as the kernel size) of the window, p denotes the number of implicitly zero-padded points on both sides, and d is the dilation rate. In our implementation, we set $p = 0$ and $d = 1$ in the SPP module.

With our spatial pyramid pooling settings, the input feature map can be represented at different scales, like a multi-level image pyramid representation. When the input feature map is at different scales, the deep network extracts features at different scales. Interestingly, the coarsest pyramid level has a single bin that covers the entire feature map. This is actually a

"global pooling" operation that has been studied in several works simultaneously, for instance, global average pooling and global maximum pooling.

In classification applications, when the network input is an image of arbitrary size, we can perform convolutions and pooling until the network is about to connect to the *fully-connected* (FC) layer, and convert the arbitrary-sized feature maps into fixed-sized feature vectors by the spatial pyramid pooling, i.e., extracting fixed-sized feature vectors using multi-scale features. However, for image segmentation, which can be viewed as pixel-level classification, we then up-sample the three low-resolution feature maps to the same size as the input feature maps, and finally, concatenate the input feature maps with the three up-sampled feature maps.

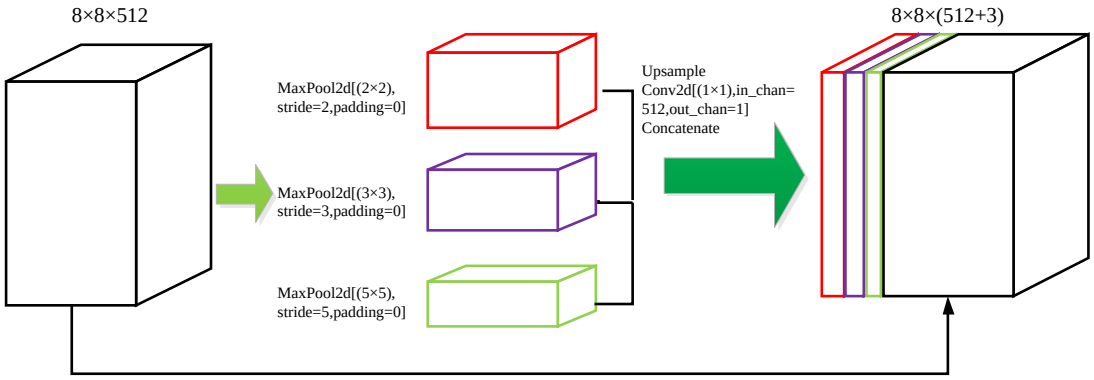


Fig. 4. Spatial Pyramid Pooling (SPP) module

3.4 Part D: Decoder

The decoder of our proposed architecture remains consistent with the original D-LinkNet [13] and is computationally efficient. It has 4 decoder blocks, each containing a 1×1 convolution, a 3×3 transposed convolution with a stride of 2, and a 1×1 convolution, as shown in the legend of **Fig. 2**. Also, the skip connections combine the coarse-grained, deep and high-level semantic features from the decoder with the fine-grained, shallow and low-level visual features from the encoder, which can compensate for the spatial information loss caused by consecutive strided convolution or pooling operations.

4. Implementation Details

4.1 Training Dataset and Data Augmentation

We have collected 50+ degraded document images from the *recognition and enrichment of archival documents* (READ) project[1] and 20+ Greek and Latin papyri documents from Google as training data, 20% of which is used as validation data.

Given a color document image, it is first cropped into image patches of size 128×128, then fed into our proposed DP-LinkNet model for training or prediction, and the output binary image patches are seamlessly stitched together to generate the resulting binary image.

Bezmaternykh et al. [41] have shown that data augmentation is crucial to provide network robustness against different types of degradation or deformation. Therefore, we have done

[1] http://read.transkribus.eu/

ambitious data augmentation, including horizontal flipping, vertical flipping, diagonal flipping, color jittering, image shifting, and scaling.

4.2 Loss Function

In order to train the network and obtain an optimal model, this paper uses the sum of *binary cross entropy* (BCE) and dice coefficient loss as the loss function, which is defined as:

$$Loss\left(y_{true}, \hat{y}_{pred}\right) = \underbrace{1 - \frac{2y_{true}\hat{y}_{pred}}{y_{true} + \hat{y}_{pred}}}_{\text{Dice Coefficient Loss}} \underbrace{-\left(y_{true}\log\left(\hat{y}_{pred}\right) + \left(1 - y_{true}\right)\log\left(1 - \hat{y}_{pred}\right)\right)}_{\text{Binary Cross Entropy Loss}} \tag{5}$$

where y_{true} is the *ground truth* (GT) label, and $\hat{y}_{pred}$ is the predicted probability of the model. The Adam optimizer is selected for parameter optimization. The initial learning rate is set to 2×10^{-4}, and reduced by 5 for 5 times while observing the training loss gradually decreasing. The batch size is fixed to 32, and the number of epochs is set to 500, with early stop strategy to avoid overfitting.

4.3 Prediction Phase

To improve the robustness of the proposed document image binarization method, this paper employs a *test time augmentation* (TTA) strategy [50], which is a means of data augmentation on the test set, including horizontal flipping, vertical flipping, and diagonal flipping (equivalent to generating $2^3 = 8$ augmented patches for each test patch). The eight predictions are then averaged to produce the final prediction map.

5. Experimental Results and Analysis

5.1 Ablation Study

Datasets: We have conducted a comprehensive ablation study to evaluate the performance of LinkNet, D-LinkNet, and our proposed DP-LinkNet. For this purpose, this study uses a relatively small dataset of historical document images, namely the diary of John R. Bickley, who was a lecturer of modern languages at the University of Pittsburgh in the mid-twentieth century. The travel diary records his trip to Europe in August 1949. The diary briefly describes the sights, restaurants, and hotels that Bickley visited in France, Spain, Switzerland, and Italy. Most of the diary is written in English, with a small portion written in French. **Fig. 1**(a) shows a sample of the original diary images.

Metrics: We adopt evaluation measures used in recent documenet image binarization competitions, including FM (*F-measure*), pFM (*pseudo F-measure*), PSNR (*peak signal-to-noise ratio*), NRM (*negative rate metric*), DRD (*distance reciprocal distortion*), and MPM (*misclassification penalty metric*). The first two metrics, namely FM and pFM, reach their best values at 1 and the worst at 0. The PSNR measures how close a binary image to the GT image, so the higher the PSNR value, the better. In contrast to the former three metrics, the binarization performance is better for lower NRM, DRD, and MPM values. Due to space limitations, we omit definitions of those evaluation measures, but readers can refer to [6, 7] for more information.

Results: **Table 1** summarizes the results of this ablation study. As can be seen from the data in the table, the LinkNet, which contains only part A and D of the proposed architecture, has the worst performance among all evaluation metrics of the three models. With an additional hybrid dilated convolution module, the evaluation performance of D-LinkNet,

which contains part A, B, and D of the proposed architecture, is better than that of the LinkNet. Our proposed DP-LinkNet, with the help of the additional spatial pyramid pooling module, outperforms the other two counterparts, and there is no significant increase in the number of parameters, compared to D-LinkNet. Our experimental results suggest that the combination of HDC and SPP modules can effectively further improve the segmentation performance of deep networks.

Table 1. Ablation study on LinkNet, D-LinkNet, and the proposed DP-LinkNet

Architecture	Params	FM(%)	pFM(%)	PSNR(dB)	DRD	MPM(‰)
LinkNet34	21,642,401	91.65	92.70	16.83	1.84	0.52
D-LinkNet34	28,736,321	91.94	93.34	17.05	1.75	0.44
DP-LinkNet34	28,738,244	**92.81**	**94.30**	**17.56**	**1.53**	**0.34**

5.2 More Segmentation Experiments

Test Datasets: We have conducted extensive experiments to evaluate the performance of our proposed DP-LinkNet architecture. For this purpose, this study uses ten recent document image binarization competition datasets from 2009 to 2019, such as DIBCO 2009 [2], 2011 [3], 2013 [4], 2017 [5], 2019 [6] and H-DIBCO 2010 [7], 2012 [8], 2014 [9], 2016 [10], 2018 [11] benchmark datasets, covering 36 machine-printed and 90 handwritten document images as well as their corresponding ground truth images. The historical documents in these datasets are originated from the *recognition and enrichment of archival documents* (READ) project, which contains a variety of collections from the fifteenth to the nineteenth century. An additional 10 historical document images are from the DIBCO 2019 Iliad papyrus dataset, which reflects the diversity of literary papyri from the third century BCE to the sixth century CE, using different types of papyri quality, inks, and handwriting styles. A total of 136 test images contain representative historical document degradation, such as severely damaged pages, ink bleed through, page stains, text stroke fading, background texture, and artifacts. **Fig. 1**(b)-(g) present some of the DIBCO and H-DIBCO image samples.

Performance Evaluation: In the first experiment, we have compared the proposed method with those that achieved TOP 3 performance in the annual document image binarization competition during 2009-2019. The evaluation results are listed in **Table 2**, and those for the TOP 3 winners of the year are copied from the corresponding competition reports, respectively. It can be seen from the table that our proposed method achieves the best performance in all the evaluation metrics by a large margin. In general terms, this indicates that our proposed method can better segment text pixels and preserve text strokes.

In the second experiment, we have further compared our proposed method with the Otsu's global thresholding [18], Niblack's [19], Sauvola's [20], and Wolf's [21] local thresholding, Jia's SSPs [27], Bhowmik's GiB [36], Vo's DSN [38], Gallego's SAE [39], Bezmaternykh's U-Net [41], Zhao's cGAN [42], and Peng's attention-based [43] methods for all the 10 DIBCO and H-DIBCO testing datasets. The algorithms involved in the comparison are implemented according to the open source or executable code provided by the original authors, and the performance evaluation results are presented in **Table 3**.

The ranking score, used in recent DIBCO and H-DIBCO competitions, is also adopted to evaluate the overall performance of each binarization method. It is to sort the accumulated ranking values of each binarization method over all the evaluation metrics and all the testing images [9]. Let $R^i(n, m)$ be the ranking value of the i^{th} method concerning the n^{th} image when using the m^{th} metric, the final ranking Score_i for each binarization method i is given by:

$$\text{Score}_i = \sum_{n=1}^{N}\sum_{m=1}^{M} R^i(n,m) \tag{6}$$

Table 2. Performance evaluation results of our proposed method against the TOP 3 winners in the DIBCO or H-DIBCO annual competition (best results highlighted in bold)

Dataset	Method	FM(%)	pFM(%)	PSNR(dB)	NRM(%)	DRD	MPM(‰)
DIBCO 2009	Rank 1	91.24		18.66	4.31		0.55
	Rank 2	90.06		18.23	4.75		0.89
	Rank 3	89.34		17.79	5.32		1.90
	Proposed	**96.39**		**22.16**	**1.30**		**0.10**
H-DIBCO 2010	Rank 1	91.50	93.58	19.78	5.98		0.49
		89.70	95.15	19.15	8.18		0.29
	Rank 2	91.78	94.43	19.67	4.77		1.33
	Rank 3	89.73	90.11	18.90	5.78		0.41
	Proposed	**96.19**	**97.06**	**22.95**	**1.29**		**0.10**
DIBCO 2011	Rank 1	80.86		16.13		104.48	64.43
	Rank 2	85.20		17.16		15.66	9.07
	Rank 3	88.74		17.84		5.36	8.68
	Proposed	**96.27**		**22.23**		**1.01**	**0.11**
H-DIBCO 2012	Rank 1	89.47	90.18	21.80		3.44	
	Rank 2	92.85	93.34	20.57		2.66	
	Rank 3	91.54	93.30	20.14		3.05	
	Proposed	**96.90**	**97.62**	**23.99**		**0.84**	
DIBCO 2013	Rank 1	92.12	94.19	20.68		3.10	
	Rank 2	92.70	93.19	21.29		3.18	
	Rank 3	91.81	92.67	20.68		4.02	
	Proposed	**97.15**	**97.77**	**24.09**		**0.78**	
H-DIBCO 2014	Rank 1	96.88	97.65	22.66		0.90	
	Rank 2	96.63	97.46	22.40		1.00	
	Rank 3	93.35	96.05	19.45		2.19	
	Proposed	**97.47**	**98.05**	**23.46**		**0.66**	
H-DIBCO 2016	Rank 1	87.61	91.28	18.11		5.21	
	Rank 2	88.72	91.84	18.45		3.86	
	Rank 3	88.47	91.71	18.29		3.93	
	Proposed	**96.29**	**97.03**	**23.04**		**1.05**	
DIBCO 2017	Rank 1	91.04	92.86	18.28		3.40	
	Rank 2	89.67	91.03	17.58		4.35	
	Rank 3	89.42	91.52	17.61		3.56	
	Proposed	**95.52**	**96.46**	**20.83**		**1.31**	
H-DIBCO 2018	Rank 1	88.34	90.24	19.11		4.92	
	Rank 2	73.45	75.94	14.62		26.24	
	Rank 3	70.01	74.68	13.58		17.45	
	Proposed	**95.99**	**96.85**	**22.71**		**1.09**	
DIBCO 2019	Rank 1	72.88	72.15	14.48		16.24	
	Rank 2	71.63	70.78	14.15		16.71	
	Rank 3	70.43	69.84	15.31		8.05	
	Proposed	**87.67**	**87.56**	**18.63**		**2.38**	

where N and M denote the total number of testing images and evaluation metrics, respectively. Therefore, a lower ranking score indicates that the algorithm has better overall performance.

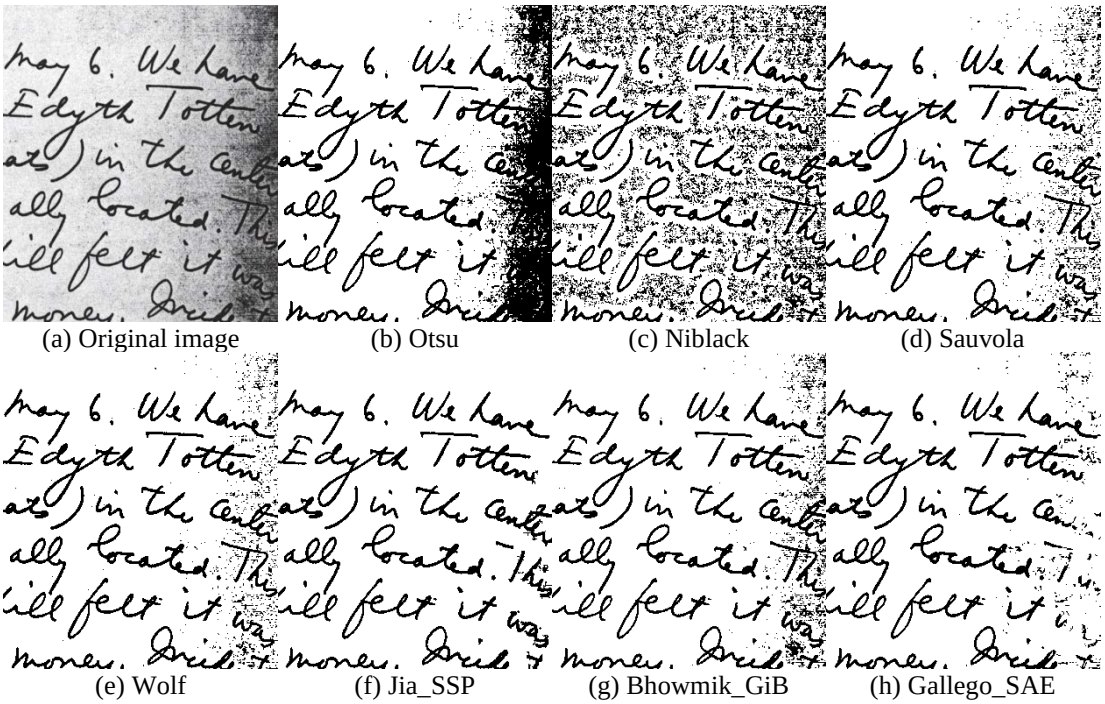
As we can see from **Table 3**, all the evaluation metrics (FM, pFM, PSNR, and DRD) of our method achieve the best performance among all the comparative methods, and the overall performance of the proposed method is also the best based on the ranking Score scheme. This

implies that our method is robust to various types and levels of document degradation, and can preserve text strokes better.

Table 3. Performance evaluation results of our proposed method against the state-of-the-art techniques on the 10 DIBCO and H-DIBCO testing datasets (best results highlighted in bold)

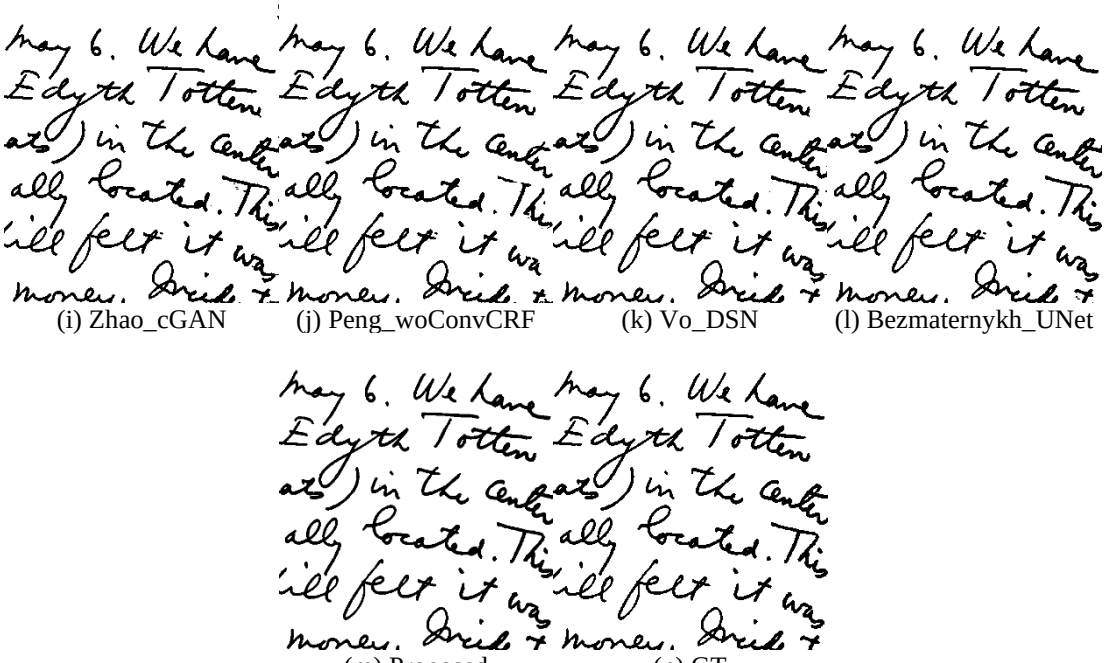
Rank	Method	FM(%)	pFM(%)	PSNR(dB)	DRD	Score
1	Proposed DP-LinkNet	**95.13**	**95.80**	**22.13**	**1.19**	**1109**
2	Bezmaternykh's UNet	89.29	90.53	21.32	3.29	2341
3	Vo's DSN	88.04	90.81	18.94	4.47	2946
4	Peng's woConvCRF	86.09	87.40	18.99	4.83	3216
5	Zhao's cGAN	87.45	88.87	18.81	5.56	3531
6	Wolf's	78.67	82.89	16.28	7.80	4851
7	Sauvola's	79.12	82.95	16.07	8.61	5281
8	Bhowmik's GiB	83.16	87.72	16.72	8.82	5316
9	Gallego's SAE	79.22	81.12	16.09	9.75	5910
10	Jia's SSP	85.05	87.24	17.91	9.74	6219
11	Otsu's	74.22	76.99	14.54	30.36	17116
12	Niblack's	41.12	41.57	6.67	91.23	50335

Fig. 5, **Fig. 6**, and **Fig. 7** present three sample images (HW1.png in DIBCO 2011, PR05.bmp in DIBCO 2013, and 12.bmp in CATEGORY2 of DIBCO 2019) and the resulting binary images generated by the selected methods involved in the comparison. It can be seen from the figures that Otsu's and Niblack's methods generally fail to produce reasonable results. Sauvola's and Wolf's methods tend to remove too many text strokes. Jia's SSP fails to extract low-contrast text pixels, especially on recent DIBCO 2019 Iliad papyri images. Gallego's SAE tends to produce ghost text pixels in the real background region. Compared to Bhowmik's GiB, Zhao's cGAN, Vo's DSN, Peng's attention-based and Bezmaternykh's UNet approaches, our proposed method can better preserve text strokes and produce better visual quality.



(a) Original image (b) Otsu (c) Niblack (d) Sauvola

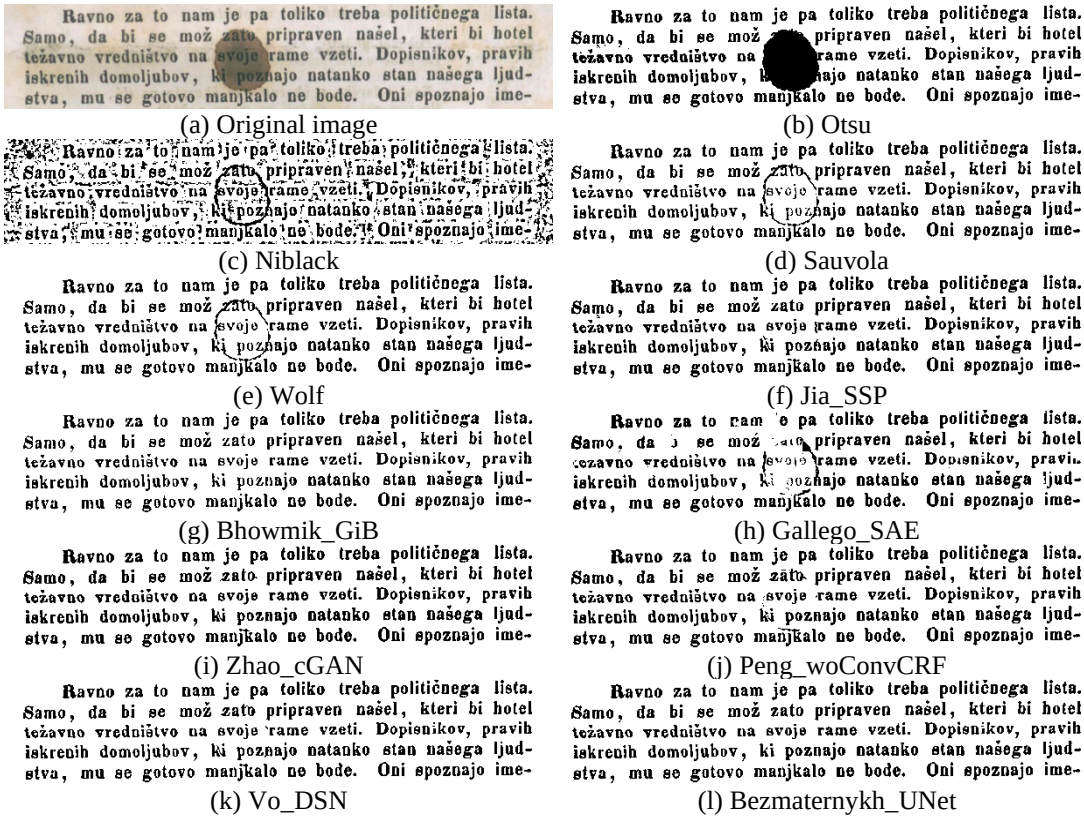
(e) Wolf (f) Jia_SSP (g) Bhowmik_GiB (h) Gallego_SAE



(i) Zhao_cGAN (j) Peng_woConvCRF (k) Vo_DSN (l) Bezmaternykh_UNet

(m) Proposed (n) GT

Fig. 5. Binarization results of selected methods for HW1 in DIBCO 2011



(a) Original image (b) Otsu

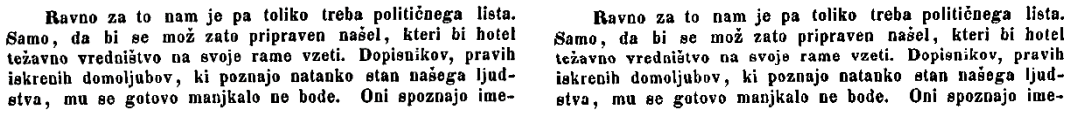
(c) Niblack (d) Sauvola

(e) Wolf (f) Jia_SSP

(g) Bhowmik_GiB (h) Gallego_SAE

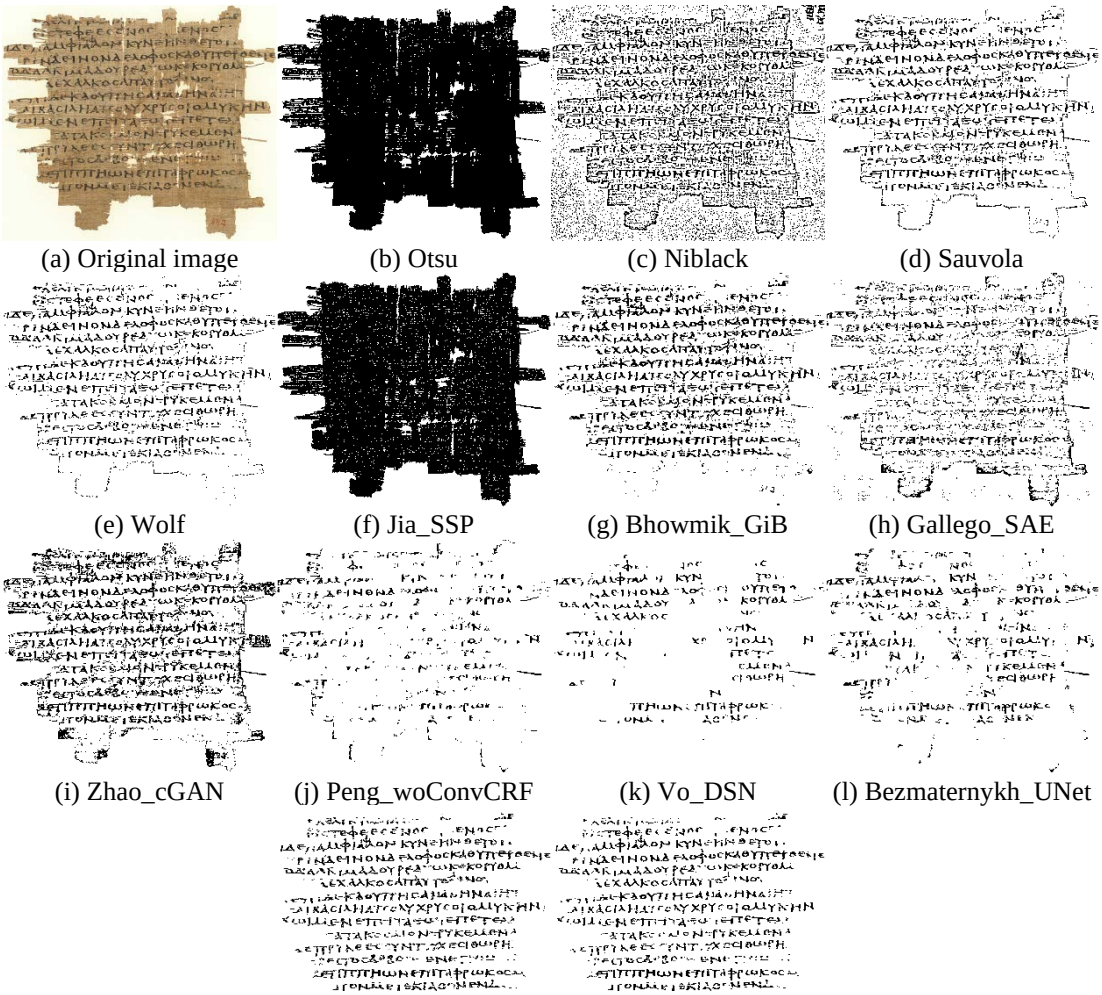
(i) Zhao_cGAN (j) Peng_woConvCRF

(k) Vo_DSN (l) Bezmaternykh_UNet



(m) Proposed (n) GT

Fig. 6. Binarization results of selected methods for PR05 in DIBCO 2013



(a) Original image (b) Otsu (c) Niblack (d) Sauvola

(e) Wolf (f) Jia_SSP (g) Bhowmik_GiB (h) Gallego_SAE

(i) Zhao_cGAN (j) Peng_woConvCRF (k) Vo_DSN (l) Bezmaternykh_UNet

(m) Proposed (n) GT

Fig. 7. Binarization results of selected methods for CATEGORY2_12 in DIBCO 2019

5.3 Time Complexity

Since the execution efficiency is related to the size of the input image, the running time of each binarization algorithm is evaluated in *seconds per megapixel* (sec/MP). The experiments are conducted on the following hardware and software platforms. The operating system is 64-bit Ubuntu 16.04 LTS (Xenial Xerus), the CPU is AMD Ryzen5 2600 processor, the system memory is 8GB; the graphics card model is NVIDIA GeForce GTX 1060 (with 6GB of video memory), the GPU acceleration library is CUDA 9.0/CUDNN 7.3.1, and the deep learning framework is based on PyTorch.

In terms of the programming languages used in other approaches, the Otsu's, Niblack's, Sauvola's, and Wolf's methods are reproduced in MatLab scripts. Bhowmik's GiB also uses MatLab, but is compiled into an executable format. Jia's SSP is written in C++, and the deep learning models are all Python-based. However, the deep learning framework used by Vo's DSN and Bezmaternykh's U-Net is Caffe. Zhao's cGAN uses PyTorch, while Gallego's SAE and Peng's attention-based techniques adopt TensorFlow. Therefore, we can only roughly evaluate the average running time of each binarization method, as illustrated in **Fig. 8**.

It can be seen from the bar chart that the binarization algorithms based on simple statistical features (e.g., Otsu's, Niblack's, Sauvola's, and Wolf's) are relatively less computationally intensive and have faster processing speed, but the segmentation performance is poor. Our proposed binarization method, using the TTA strategy and performing on-the-fly data augmentation in the testing phase, is much faster than most CNN-based techniques and those based on complex text stroke features. Experimental results indicate that the proposed DP-LinkNet is able to extract text features better, with fewer parameters and without deepening the network hierarchy.

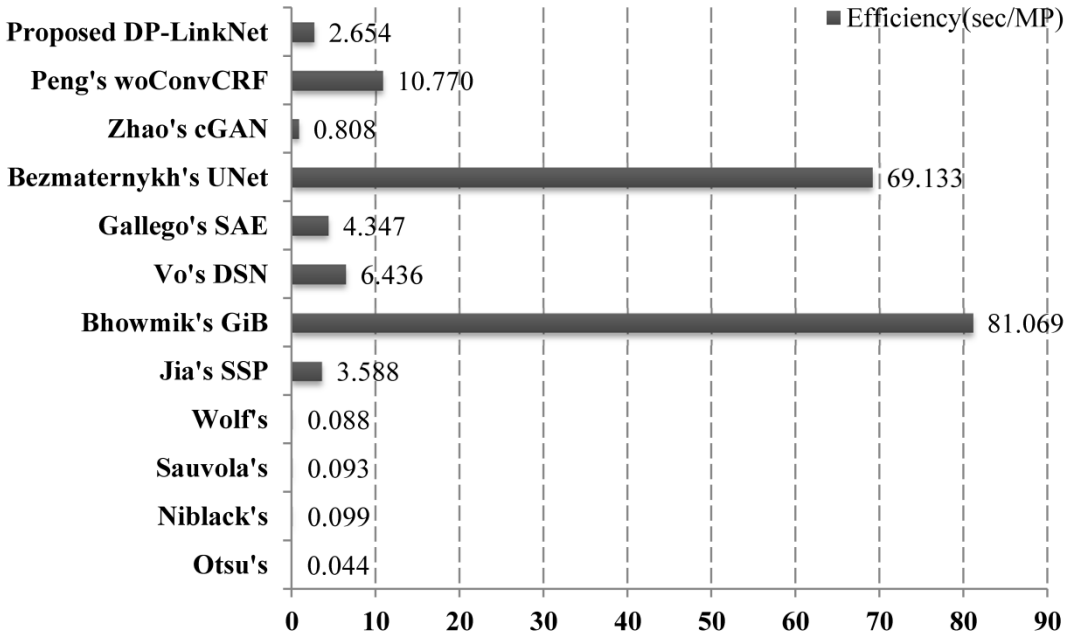


Fig. 8. Average running time comparison in *seconds per megapixel* (sec/MP)

6. Conclusion

This paper presents a semantic segmentation network, named DP-LinkNet, for more accurate binarization of degraded historical document images. The improved performance is mainly attributed to its hybrid dilated convolution and spatial pyramid pooling blocks located between the encoder and the decoder. The hybrid dilated convolution expands the receptive field size, while the spatial pyramid pooling encodes the aggregated multi-scale features. Detailed spatial information is still maintained by skip connections, which combine the coarse-grained, deep, and high-level semantic features from the decoder with the fine-grained, shallow, and low-level visual features from the encoder. We have conducted extensive experiments on recent DIBCO and H-DIBCO benchmark datasets. The results show that our proposed DP-LinkNet

outperforms other state-of-the-art techniques by a large margin.

Appendix

Computing Receptive Fields of Convolutional Neural Networks

Consider a *fully convolutional network* (FCN) with L layers, $l = 1, 2,\ldots, L$. Define k_l and s_l to denote the kernel size and stride of the l^{th} layer. The dilation rate r_l indicates inserting $r_l - 1$ zeros (holes) between the kernel weights or subsampling the feature map $\boldsymbol{f}_l \in \mathbb{R}^{h_l \times w_l \times d_l}$ by a factor of $r_l - 1$, where h_l, w_l, and d_l denote the height, width, and depth of the corresponding feature map, respectively. Therefore, the virtual kernel size $\hat{k}_l$ of r_l-dilated convolutions in the l^{th} layer is:

$$\hat{k}_l = k_l + (k_l - 1) \times (r_l - 1) = r_l \times (k_l - 1) + 1 .$$

The receptive field size RF_l corresponds to the number of features in the feature map $\boldsymbol{f}_l$, which contributes to generate one feature in the final output feature map $\boldsymbol{f}_L$. Note that $RF_0 = 1$.

Since each feature from $\boldsymbol{f}_{l-1}$ is directly connected to $\hat{k}_{l-1}$ features from $\boldsymbol{f}_l$, we obtain RF_1 for the first layer:

$$RF_1 = s_0 \times RF_0 - \left(s_0 - \hat{k}_0 \right).$$

The first term $s_0 \times RF_0$ covers the entire region where the features come from, but it covers $s_0 - \hat{k}_0$ more features, which should be subtracted. Like this, we obtain RF_2 for the second layer:

$$RF_2 = s_1 \times RF_1 - \left(s_1 - \hat{k}_1 \right).$$

Therefore, we get the general recursive equation:

$$\begin{cases} RF_l = s_{l-1} \times RF_{l-1} - \left(s_{l-1} - \hat{k}_{l-1} \right) \\ RF_0 = 1 \end{cases}$$

Readers may refer to the web page[2] for more information on solving such first-order non-homogeneous recurrence relations with variable coefficients, and we finally obtain:

$$RF_L = 1 + \sum_{l=1}^{L} \left(\left(\hat{k}_l - 1 \right) \prod_{i=1}^{l-1} s_i \right)$$

References

[1] S. Eskenazi, P. Gomez-Krämer, J.-M. Ogier, "A comprehensive survey of mostly textual document segmentation algorithms since 2008," *Pattern Recognition*, vol. 64, pp. 1-14, 2017. Article (CrossRef Link)

[2] B. Gatos, K. Ntirogiannis, I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*, Barcelona, SPAIN, pp. 1375-1382, 2009. Article (CrossRef Link)

[2] https://en.wikipedia.org/wiki/Recurrence_relation#Solving_first-order_non-homogeneous_recurrence_relations_with_variable_coefficients

[3] I. Pratikakis, B. Gatos, K. Ntirogiannis, "Icdar 2011 document image binarization contest (dibco 2011)," in *Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, Beijing, CHINA, pp. 1506-1510, 2011. Article (CrossRef Link)

[4] I. Pratikakis, B. Gatos, K. Ntirogiannis, "Icdar 2013 document image binarization contest (dibco 2013)," in *Proc. of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*, Washington, DC, USA, pp. 1471-1476, 2013. Article (CrossRef Link)

[5] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, "Icdar 2017 competition on document image binarization (dibco 2017)," in *Proc. of the 14th International Conference on Document Analysis and Recognition (ICDAR 2017)*, Kyoto, JAPAN, pp. 1395-1403, 2017. Article (CrossRef Link)

[6] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, I. Marthot-Santaniello, "Icdar 2019 competition on document image binarization (dibco 2019)," in *Proc. of the 15th International Conference on Document Analysis and Recognition (ICDAR 2019)*, Sydney, AUSTRALIA, 2019. Article (CrossRef Link)

[7] I. Pratikakis, B. Gatos, K. Ntirogiannis, "H-dibco 2010 - handwritten document image binarization competition," in *Proc. of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010)*, Kolkata, INDIA, pp. 727-732, 2010. Article (CrossRef Link)

[8] I. Pratikakis, B. Gatos, K. Ntirogiannis, "Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012)," in *Proc. of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, Bari, ITALY, pp. 817-822, 2012. Article (CrossRef Link)

[9] K. Ntirogiannis, B. Gatos, I. Pratikakis, "Icfhr 2014 competition on handwritten document image binarization (h-dibco 2014)," in *Proc. of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014)*, Hersonissos, GREECE, pp. 809-813, 2014. Article (CrossRef Link)

[10] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, "Icfhr 2016 handwritten document image binarization contest (h-dibco 2016)," in *Proc. of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016)*, Shenzhen, CHINA, pp. 619-623, 2016.

[11] I. Pratikakis, K. Zagoris, P. Kaddas, B. Gatos, "Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018)," in *Proc. of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2018)*, Niagara Falls, USA, pp. 489-493, 2018. Article (CrossRef Link)

[12] R. D. Lins, E. Kavallieratou, E. B. Smith, R. B. Bernardino, D. M. d. Jesus, "Icdar 2019 time-quality binarization competition," in *Proc. of the 15th International Conference on Document Analysis and Recognition (ICDAR 2019)*, Sydney, AUSTRALIA, 2019.

[13] L. Zhou, C. Zhang, M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR 2018)*, Salt Lake City, UT, USA, pp. 192-196, 2018. Article (CrossRef Link)

[14] E. Shelhamer, J. Long, T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2017. Article (CrossRef Link)

[15] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Munich, GERMANY, pp. 234-241, 2015. Article (CrossRef Link)

[16] M. Sezgin, B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146-168, 2004. Article (CrossRef Link)

[17] N. P. Challa, R. V. K. Mehta, "Applications of image processing techniques on palm leaf manuscripts - a survey," *Helix*, vol. 7, no. 5, pp. 2013-2017, 2017.

[18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979. Article (CrossRef Link)

[19] W. Niblack, *An introduction to digital image processing*. Englewood Cliffs, New Jersey: Prentice-Hall International Inc., 1986.
[20] J. Sauvola, M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225-236, 2000. Article (CrossRef Link)
[21] C. Wolf, J.-M. Jolion, "Extraction and recognition of artificial text in multimedia documents," *Pattern Analysis and Applications*, vol. 6, no. 4, pp. 309-326, 2004. Article (CrossRef Link)
[22] J. Bernsen, "Dynamic thresholding for gray-level images," in *Proc. of the 8th International Conference on Pattern Recognition (ICPR 1986)*, Paris, pp. 1251-1255, 1986.
[23] M. van Herk, "A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels," *Pattern Recognition Letters*, vol. 13, no. 7, pp. 517-521, 1992. Article (CrossRef Link)
[24] B. Su, S. Lu, C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *Proc. of the* the *9th IAPR International Workshop on Document Analysis Systems (DAS 2010)*, Boston, Massachusetts, USA, pp. 159-166, 2010. Article (CrossRef Link)
[25] B. Su, S. Lu, C. L. Tan, "Robust document image binarization technique for degraded document images," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1408-1417, 2013. Article (CrossRef Link)
[26] S. Lu, B. Su, C. L. Tan, "Document image binarization using background estimation and stroke edges," *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 303-314, 2010. Article (CrossRef Link)
[27] F. Jia, C. Shi, K. He, C. Wang, B. Xiao, "Degraded document image binarization using structural symmetry of strokes," *Pattern Recognition*, vol. 74, pp. 225-240, 2018. Article (CrossRef Link)
[28] Q. N. Vo, S. H. Kim, H. J. Yang, G. Lee, "An mrf model for binarization of music scores with complex background," *Pattern Recognition Letters*, vol. 69, pp. 88-95, 2016. Article (CrossRef Link)
[29] E. Ahmadi, Z. Azimifar, M. Shams, M. Famouri, M. J. Shafiee, "Document image binarization using a discriminative structural classifier," *Pattern Recognition Letters*, vol. 63, pp. 36-42, 2015. Article (CrossRef Link)
[30] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition*, vol. 16, no. 3, pp. 247-258, 2013. Article (CrossRef Link)
[31] N. R. Howe, "A laplacian energy for document binarization," in *Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, Beijing, CHINA, pp. 6-10, 2011. Article (CrossRef Link)
[32] D. Rivest-Hénault, R. F. Moghaddam, M. Cheriet, "A local linear level set method for the binarization of degraded historical document images," *International Journal on Document Analysis and Recognition*, vol. 15, no. 2, pp. 101-124, 2012. Article (CrossRef Link)
[33] Z. Hadjadj, M. Cheriet, A. Meziane, Y. Cherfa, "A new efficient binarization method: Application to degraded historical document images," *Signal, Image and Video Processing*, vol. 11, pp. 1155-1162, 2017. Article (CrossRef Link)
[34] X. Chen, L. Lin, Y. Gao, "Parallel nonparametric binarization for degraded document images," *Neurocomputing*, vol. 189, pp. 43-52, 2016. Article (CrossRef Link)
[35] W. Xiong, J. Xu, Z. Xiong, J. Wang, M. Liu, "Degraded historical document image binarization using local features and support vector machine (svm)," *Optik*, vol. 164, pp. 218-223, 2018. Article (CrossRef Link)
[36] S. Bhowmik, R. Sarkar, B. Das, D. Doermann, "Gib: A game theory inspired binarization technique for degraded document images," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1443-1455, 2019. Article (CrossRef Link)
[37] C. Tensmeyer, T. Martinez, "Document image binarization with fully convolutional neural networks," in *Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*, Kyoto, Japan, pp. 99-104, 2017. Article (CrossRef Link)
[38] Q. N. Vo, S. H. Kim, H. J. Yang, G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, pp. 568-586, 2018. Article (CrossRef Link)

[39] J. Calvo-Zaragoza, A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86,pp. 37-47, 2019. Article (CrossRef Link)
[40] X.-J. Mao, C. Shen, Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. of the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, pp. 2810-2818, 2016.
[41] P. V. Bezmaternykh, D. A. Ilin, D. P. Nikolaev, "U-net-bin: Hacking the document image binarization contest," *Computer Optics*, vol. 43, no. 5, pp. 825-832, 2019. Article (CrossRef Link)
[42] J. Zhao, C. Shi, F. Jia, Y. Wang, B. Xiao, "Document image binarization with cascaded generators of conditional generative adversarial networks," *Pattern Recognition*, vol. 96, 2019. Article (CrossRef Link)
[43] X. Peng, C. Wang, H. Cao, "Document binarization via multi-resolutional attention model with drd loss," in *Proc. of the 15th IAPR International Conference on Document Analysis and Recognition (ICDAR 2019)*, Sydney, NSW, Australia, pp. 45-50, 2019. Article (CrossRef Link)
[44] A. Chaurasia, E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. of the 2017 IEEE Visual Communications and Image Processing (VCIP 2017)*, St. Petersburg, FL, USA, pp. 1-4, 2017. Article (CrossRef Link)
[45] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, FL, pp. 248-255, 2009. Article (CrossRef Link)
[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2018. Article (CrossRef Link)
[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the 15th European Conference on Computer Vision (ECCV 2018)*, Munich, GERMANY, pp. 833-851, 2018. Article (CrossRef Link)
[48] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. of the 4th International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, 2016.
[49] K. He, X. Zhang, S. Ren, J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015. Article (CrossRef Link)
[50] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34-45, 2019. Article (CrossRef Link)

Wei Xiong received the B.S. degree in electronic engineering and the Ph.D. degree in signal and information processing, both from Wuhan University, Hubei, China, in 2003 and 2010, respectively. He is an Associate Professor with the School of Electrical and Electronic Engineering, Hubei University of Technology, Hubei, China. He was a visiting professor with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA. His research interests include computer vision, pattern recognition, deep learning, and artificial intelligence.

Xiuhong Jia received the B.S. degree in electronic information engineering from Shanxi Normal University, Shanxi, China, in 2017, and the M.S. degree in electrical engineering from Hubei University of Technology, Hubei, China, in 2020, respectively. Her research interests include computer vision, deep learning, document recognition, and artificial intelligence.

Dichun Yang received the B.S. degree in electronic engineering and automation from Hubei University of Technology, Hubei, China, in 2018. He is a M.S. candidate in navigation guidance and control with the School of Electrical and Electronic Engineering, Hubei University of Technology, Hubei, China. His research interests include computer vision, deep learning, and artificial intelligence.

Meihui Ai received the B.S. degree in communication engineering from Hubei University of Technology, Hubei, China, in 2018. She is a M.S. candidate in control engineering with the School of Electrical and Electronic Engineering, Hubei University of Technology, Hubei, China. Her research interests include computer vision, deep learning, scene text detection and recognition.

Lirong Li received the M.S. degree and Ph.D. degree, both in pattern recognition and artificial intelligence, from Huazhong University of Science and Technology, Wuhan, Hubei, China, in 2004 and 2017, respectively. She has been working as a teacher in the School of Electrical and Electronic Engineering, Hubei University of Technology since July 2004. Her research interests include computer vision and artificial intelligence, multispectral image processing.

Song Wang (*Senior Member, IEEE*) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana Champaign (UIUC) in 2002. From 1998 to 2002, he also worked as a Research Assistant with Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, where he is currently a Professor. His research interests include computer vision, medical image processing, and machine learning. He is a Senior Member of the IEEE Computer Society. He is currently serving as an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition Letters, and Electronics Letters.