Signal Processing: Image Communication
journal homepage: www.elsevier.com

A Dilated MultiRes Visual Attention U-Net for Historical Document Image Binarization

Nikolaos Detsikas[a], Nikolaos Mitianoudis[a], Nikolaos Papamarkos[a]

[a]*Electrical and Computer Engineering Department, Democritus University of Thrace, University Campus Xanthi-Kimmeria, Xanthi 67100, Greece*

ABSTRACT

The task of binarization of historical document images has been in the forefront of image processing research, during the digital transition of libraries. The process of storing and transcribing valuable historical printed or handwritten material can salvage world cultural heritage and make it available online without physical attendance. The task of binarization can be viewed as a pre-processing step that attempts to separate the printed/handwritten characters in the image from possible noise and stains, which will assist in the Optical Character Recognition (OCR) process. Many approaches have been proposed before, including deep learning based approaches. In this article, we propose a U-Net style deep learning architecture that incorporates many other developments of deep learning, including residual connections, multi-resolution connections, visual attention blocks and dilated convolution blocks for upsampling. The novelties in the proposed DMVAnet lie in the use of these elements in combination in a novel U-Net style architecture and the application of DMVAnet in image binarization for the first time. In addition, the proposed DMVAnet is a very computationally lightweight network that performs very close or even better than the state-of-the-art approaches with a fraction of the network size and parameters. Finally, it can be used on platforms with restricted processing power and system resources, such as mobile devices and through scaling can result in inference times that allow for real-time applications.

1. Introduction

Document Image Binarization is the process of separating the text from its environment in a document image. The input image is segmented into two layers of information, one for the background and one for the text. The output image of the process is essentially a binary map, where the value of each pixel represents whether it belongs to the text or the background layer.

The ability to extract text from an image is critical in many applications. Firstly, scanning handwritten or printed documents and extracting the text allows the digitisation of the written cultural heritage. In addition to that, Optical Character Recognition can greatly benefit from improved text binarization algorithms, since extracting the scanned text is a critical task of the process. Moreover, captured images and video feeds often contain text that must be extracted and processed. Written language appears everywhere in our urban environments and text binarization is a method, through which it can be processed by the machines of the present and future. For these reasons, the problem of Document Image Binarization is one of the earliest in the image processing literature and has been addressed with numerous methods, both based on traditional image processing (often heuristic methods) and machine learning or deep learning.

One of the most well-known approaches is Otsu's method [1], which performs automatic global image thresholding in its basic form. Sauvola [2] and Niblack [3] binarization algorithms operate by calculating local thresholds for every pixel, based on statistical information from the pixel neighbourhood.

These early methods suffer from the need of parameter tuning. The parameters have to be tuned differently for each binarization context. Heuristic methods have been introduced that combat this problem. Howe [4] binarizes the image by minimising a global energy function, based on a Markov Random Field model and performing automatic parameter tuning. Su et

e-mail: ndetsika@ee.duth.gr (Nikolaos Detsikas), nmitiano@ee.duth.gr (Nikolaos Mitianoudis), papamark@ee.duth.gr (Nikolaos Papamarkos)

al. [5] construct an adaptive contrast map and based on that and the Canny edge detection map, the method detects the text stroke edges, which are used to estimate local threshold values for binarizing the image. Lelore and Bouchara [6] introduce the FAIR algorithm, which applies a modified Canny edge detection algorithm and clusters the resulting pixels. To tackle the defects of parameter selection, the process is applied twice with different parameters and the final results are merged. Nachi et al. [7] use phase congruency feature maps, based on Kovesi's phase congruency model, as well as a phase-derived denoised image in order to produce a final binarized version of the input. Mitianoudis and Papamarkos [8] address the Document Image Binarization problem by first removing the background with a long-window low-pass filtering process. The resulting image is binarized using Local Co-occurrence Mapping (LCM), that exploits common local character properties, when identifying the character pixels and Mixture of Gaussians (MoG) clustering. Finally, a mathematical morphology step removes misclassified or noisy items. Jie at al. [9] perform a background removal process, compute a gradient map from the output and extract the structural symmetric pixels (SSPs) to calculate local thresholds for the binarization process. Bhowmik et al [10] employ Game Theory concepts, such as two-player non-zero-sum non-cooperative game and the Nash equilibrium, in order to extract image features that are further fed to a K-means clustering step for classifying the pixels into foreground and background groups. In all these methods, there are certain parameters that need to be defined by hand. Global values can be used, but the result may not be optimal for each separate case.

Heuristic methods have considerably lower performance compared to the recently introduced deep learning methods. Moreover, even without the need for parameter tuning, they still cannot handle extensive input image variations as effectively as deep neural networks. In this paper, we focus on recent deep learning methods, we evaluate their performance and suggest an innovative architecture for addressing the problem more efficiently. In [11], Li et al. suggest a deep learning approach to the classic *Sauvola* binarization method for eliminating the hyperparameters dependency of the algorithm and replacing it with trainable parameters. The method implements the conventional *Sauvola* algorithm into a Deep Neural Network, uses attention mechanisms for bypassing the need for manually specifying a window size and finally uses deep networks for calculating the adaptive binarization thresholds. In [12], He and Schomaker suggest an iterative deep learning framework for improving the input images by removing noise and degradations that usually prevent efficient binarization. The framework "learns" the noise and degradations present on the original image and after several iterations produces a uniform variant that can be finally binarized with any existing binarization process. Therefore, the proposed framework acts as a deep learning augmentation preprocessing step for any binarization process. Calvo-Zaragoza and Gallego [13] propose the *Selectional Auto-Encoder (SAE)*, an auto-encoder architecture, implemented only with convolutional machine learning layers, that aims at learning a mapping between the input image and a binary representation of foreground (text) and background pixels. Similarly to the basic auto-encoder architecture, the *SAE* gradually downsamples and upsamples the image in the encoder and decoder parts respectively, until reaching the final prediction layer, where the binary map of the input is produced. Vo et al. [14] use a multiscale hierarchical approach consisting of three Deep Supervised Networks (DSN) in order to separate text from the background noise. By using different feature scales, the model tries to optimize the classification of image pixels over large areas as well as those over the text boundaries. Zhao et al. [15] view the binarization as an image generative task and employ conditional Generative Adversarial Networks (cGANs) in order to synthesize the binarized output images from degraded document images. A two-stage generator is employed for producing binary maps, based on the learned input and ground truth images. Another generative adversarial model solution is proposed by De et al. [16], who use a U-Net generator to binarize images and later discriminate with two distinct discriminator networks, one for higher level and one for lower level features. Peng et al. [17] address the Document Image Binarization task by introducing a convolutional attention block for the block input features that are most likely to produce the desired output targets. In [18], Kang et al. implement a model (CMU-Net), consisting of multiple pretrained U-Nets, connected in a cascaded manner. Each U-Net is trained to perform an image binarization task, such as erosion, dilation, binarization and Canny-edge detection. In conventional image processing, the above methods serve as steps in a binarization algorithm. The goal of the authors is to reproduce such a pipeline of methods with individually trained U-Nets, interconnected with appropriate skip connections. The CT-Net, is an architecture proposed by He and Schomaker [19], who describe a novel T-net architecture that consists of one encoder and two decoders, the first of which performs an image enhancement task and the other a binarization task. In order to achieve better performance, the T-net blocks are cascaded resulting in a CT-net model. The connections between the T-nets are placed along the enhancement outputs, so that each T-net along the pipeline receives a more and more enhanced image as input. Finally, Jemni at al. [20] train a multipart GAN architecture to enhance images before binarization tasks. The architecture consists of a U-Net generator that produces a clean binary version of the degraded input image, followed by a discriminator that decides whether the cleaned images come from the generator or are ground truth images. The final binary image is passed through a Convolutional Recurrent Neural Network (CRNN), which serves as a Handwritten text line recogniser that outputs plain text. Finally, Ju et al. [21, 22] introduce a GAN inspired three-stage method for enhancing and binarizing degraded document images (CCDWT-GAN). Stage-1 enhances the image by applying the Discrete Wavelet Transform (DWT) and retaining the Low-Low (LL) subband images. The proposed pipeline continues with stage-2, where each input image channel is trained with independent adversarial networks. The extracted channel-wise information is trained again with the input image in an adversarial network for final document binarization.

Most of the presented deep learning approaches use large architectures that cannot easily be executed on devices with lim-

ited resources. Memory and inference time can be prohibitive issues for general reception and usage. In addition, some deep learning methods act merely as a pre-processing step, or build multiple-step processing pipelines in order to achieve their binarization goals, which again poses restriction on the device capabilities tolerance.

In this paper, we aim to propose a simple, lightweight yet effective deep learning architecture that addresses the binarization problem without using pre- and post- processing or ensemble of different networks. We start from a baseline U-Net architecture and we introduce a number of modifications that lead to a comparatively low-complexity architecture that manages to outperform most of the state-of-the-art (SOTA) binarization methods. First, we introduce residual connections [23] and replace the convolution blocks on each layer by residual blocks, as described in Section 2.2. Consequently, we weigh the skip connections with visual attention blocks, before concatenating them to the decoder layers (Section 2.3). Next, we replace the residual blocks previously added with multi-resolution blocks and also add residual paths along the skip connections. Both blocks and the resulting architecture are described in Section 2.4. Finally, we further optimize our network with atrous convolutions (Section 2.5).

The proposed Dilated MultiRes Visual Attention U-Net (DMVAnet) architecture is a composite U-Net network with the following key advantages, contributions and innovations:

- DMVAnet exhibits state-of-the-art performance in a single-step approach with comparatively low complexity
- DMVAnet contains carefully selected features from previous networks to form a novel architecture that has not been applied or tested in the context of image binarization before, to the best of our knowledge
- DMVAnet is a single network, trained only once, while related SOTA methods, either consist of multiple separate deep neural networks, different pre- or post- processing steps, or have to be applied iteratively. Due to the above, they feature complex implementation steps or serve as an enhancement step that should be followed by other binarization methods
- Most SOTA networks feature far more complex architectures (with up to 32 times more parameters than DMVAnet), yielding no or minimal performance gain compared to their added complexity

The paper is organised, as follows. In Section 2, the procedure of constructing the proposed architecture is described in an illustrative step-by-step manner. Section 3 contains an ablation study that justifies the appropriate changes and choices in the proposed deep learning architecture. Section 4 compares the computational complexity of the proposed approach with other SOTA approaches. Section 5 compares the proposed approach with other state-of-the-art methods on commonly-used datasets by the community. Finally, Section 6 concludes the article and proposes steps for future work.

2. The Proposed DMVAnet

In this section, we start from a baseline U-Net architecture and gradually build a semantic segmentation model that incorporates elements from modern deep learning structures, improving the performance of the final proposed model. The proposed modifications to the basic U-Net network are gradually introduced, either by adding new blocks or by altering key architectural components. This procedure has been validated through experiments, the most indicative of which will appear in the ablation study. In this study, we will be presenting only structural changes that led to performance improvement. However, many other architectures (U-Net with dense connections [24], UNet++ [25] and DeeplabV3+ [26]) were also tested without improving performance.

2.1. Basic U-Net architecture

The baseline architecture consists of a simple U-Net [27], a simple encoder-decoder network with skip connections, which is the basis of most deep learning architectures used in semantic segmentation problems.

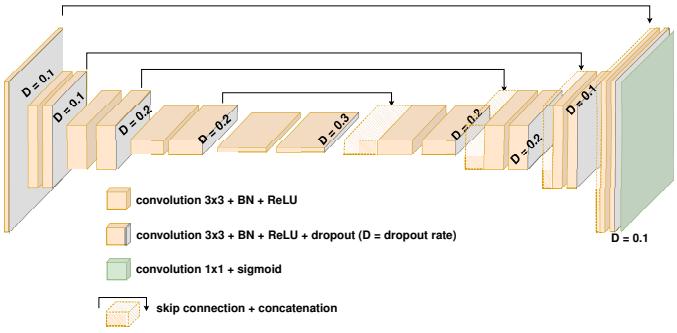

Fig. 1: Modified U-Net architecture.

U-Net architectures primarily consist of an encoder and a decoder block. The encoder gradually transforms the input image into feature maps of encoded information of lower resolution. The encoded information is then gradually decoded by the decoder into the desired form of information. For semantic segmentation tasks this is a list of binary output maps, one for each semantic class, where 1's mark the pixels that belong to the class. For document image binarization, a single binary map is produced, with 0's being the background and 1's the text pixels. The output map resolution is the same as or similar to the input resolution.

Starting with a baseline U-net, to tackle overfitting, we have included a *Dropout* layer [28], on each encoder and decoder level, with gradually increasing and decreasing rate respectively. In addition, our experiments have shown that upscaling with transpose convolution layers [29] performed better than upsampling with any other interpolation method. The parameters chosen were 1×1 kernel, strides of size 2 and appropriate input padding ('same'), such that the output feature maps dimensions are doubled with the stride effect. Another benefit of using transpose convolution for upscaling the image is that it can reduce the number of output filters at the same time, while

mathematically upscaling would require a subsequent convolution layer for this purpose. Fig. 1 shows the modified U-Net, which will be the starting point for further modifications that will help us improve its performance on the Document Image Binarization task.

2.2. Residual U-Net (Res-U-Net)

A residual U-Net is a U-Net network with residual connections along the paths of the encoder and decoder. The residual connections were introduced in the seminal paper on ResNets [23]. The residual connections are implemented in a residual connection block, as shown in Fig. 2.

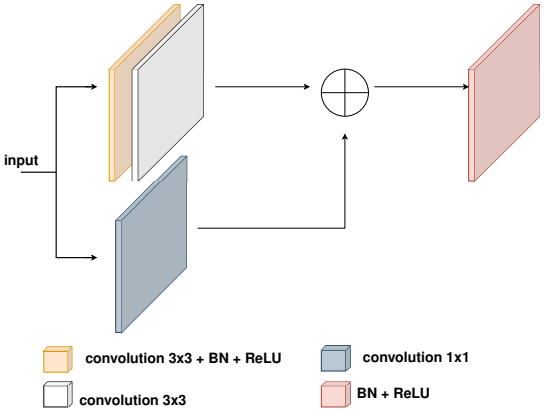

Fig. 2: A Residual block used in the Res-U-Net architecture.

Training a deep neural network can exhibit the vanishing gradient problem, where the error back propagation may have insignificant effect on early network layers. This happens because of the continuous application of the chain rule for computing the gradients. The deeper the network, the stronger the effect might become. The residual block identity connection has shown that it can protect the network from this shortcoming [23]. Essentially, the residual U-Net architecture is a basic U-Net, where we have replaced the convolutions on each level of the encoder and decoder with the residual block. The resulting network is shown in Fig. 3.

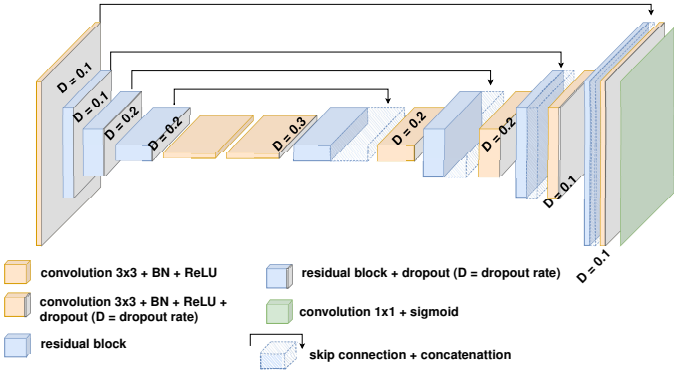

Fig. 3: Residual U-Net architecture.

2.3. Visual attention U-Net (VAnet)

We further extended the residual U-Net architecture, by adding a visual attention block. In general, the visual attention block is used to scale its input data in order to focus on those image parts that contain the most relevant to the task information [30]. In U-Net architectures, the visual attention block is used for scaling the skip connection encoder output with the use of the respective decoder level output, before concatenating the two sets of feature maps [31]. The visual attention block used here is shown in Fig. 4.

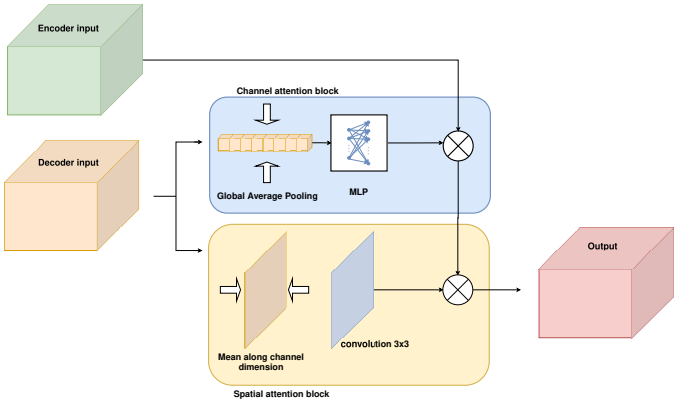

Fig. 4: The proposed Visual attention block.

The *visual attention block* used in this work, consists of two modules, the *Channel attention module* and the *Spatial attention module*, similarly to [30]. Within the context of a U-Net architecture, the decoder input is modulated within each block in order to provide a weight to scale the input coming from the encoder skip connection.

The *Channel attention module* first performs a global average pooling operation on the decoder input, reducing its size to $1 \times 1 \times C$, essentially reducing the input features maps to a single vector of size $C \times 1$, where C is the number of input feature maps. Subsequently, a 2-layer fully-connected neural network maps the vector to the channel attention output, which is multiplied element-wise with the encoder skip connection input.

The *Channel attention module* focuses on the information captured across the channels. By collapsing the 2D decoder feature maps, through the 2D averaging operation, the spatial information is ignored. The remaining channel vector acts as a vector of "what" was represented on the collapsed feature maps, ignoring the "where". The neural network, that follows, acts as a regressor that predicts the scale that should be applied to each encoder skip connection input features map. This way, the decoder input acts as a guide, as to how important each individual feature map of the encoder is.

On the other hand, the *Spatial attention module* focuses on the spatial characteristics of the feature maps. The channel dimension collapses by averaging and the result is the 2D information of the feature maps. The convolution operation that follows, along with the sigmoid activation function, creates a binary representation of "where" important information exists on the feature maps.

The selected form of the visual attention mechanism was the result of experimentation, based on the architectures described in [30] and [31]. The selected architecture was the one that performed best in the context of the Document Image Binarization problem in our experiments.

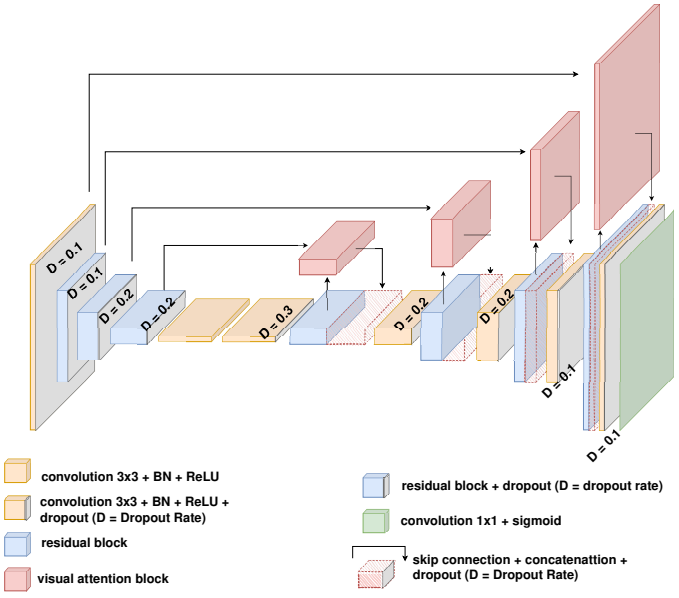

Fig. 5: A U-Net with Visual Attention modules.

In our implementation, we have used the *visual attention block* on the decoder levels of the U-Net, where we connect the encoder skip connection and the decoder input from the previous level (after the transpose convolution). The block output is propagated to the rest of the decoder level blocks. The resulting architecture is shown in Fig. 5.

2.4. MultiRes Visual Attention U-Net (MVAnet)

In [32], Ibtehaz and Rahman propose the replacement of *Inception blocks* with *MultiRes blocks* as a means to enhance the performance of a U-Net network, while minimising the additional memory overhead from the increased complexity.

A simple *Inception block* consists of the concatenation of the output of convolutions at different scales, such as 3×3, 5×5 and 7×7. Introduced in [33], convolutions at different sizes are used in order to capture information at different image scales. The larger convolutional layers of the *Inception block* perform better at various contexts but cause significant overhead to the memory requirements of the network.

The *MultiRes block* replaces the larger convolutions with a sequence of 3×3 convolutional layers. The 5×5 is replaced with two 3×3 layers and the 7×7 with three 3×3 layers. In a sequence of only three 3×3 layers, the intermediate outputs represent all the above convolutions, as depicted in Fig. 6. The block is finalized by concatenating the intermediate outputs, while a residual connection completes the *MultiRes block*. As also described in [32], the number of filters in the three consecutive convolutional layers is gradually increased, for keeping the memory requirements low. In the opposite case, the convolution layer cascade would keep the memory requirements large.

In addition to the *MultiRes block*, the authors also suggest replacing the simple skip connections with *Residual paths (Res paths)* as depicted in Fig. 7. The *Res paths* pass the encoder output through a series of convolutional blocks with skip connections, before the concatenation with the respective decoder

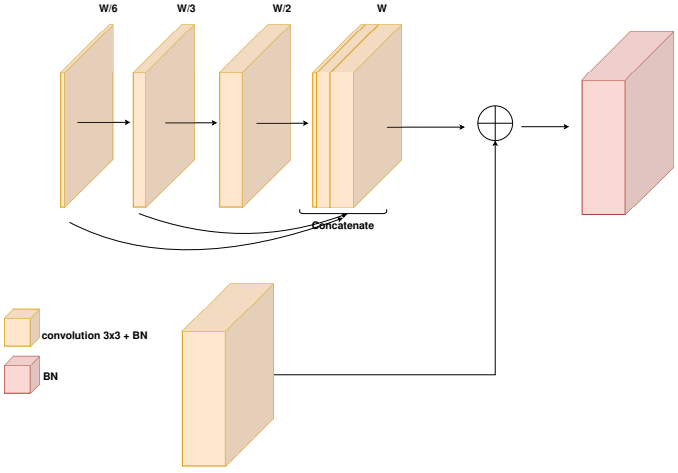

Fig. 6: The MultiRes block used in the U-Net concept.

features. The rationale behind this architectural trait is that the decoder features go through much more processing before their concatenation with the respective encoder features. This processing imbalance arguably causes incompatibilities in the semantic information carried by the concatenated encoder and decoder features maps. The additional convolutions along the *Res paths* aim at bridging this gap.

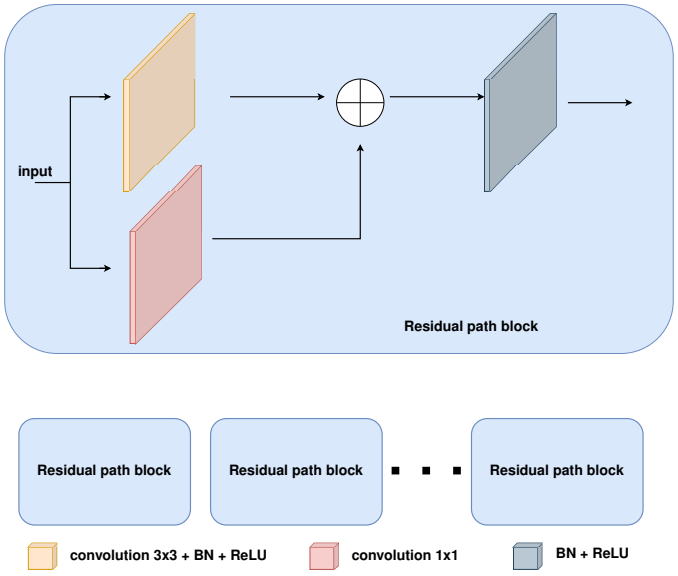

Fig. 7: The Residual path used in the MultiRes Visual Attention U-Net.

The number of processing blocks on the *Res paths* along the skip connections is not constant. The deeper the skip connection, the shorter the *Res paths*, since less additional processing is needed for the encoder feature maps.

In our implementation, we have replaced the residual blocks on all encoder and decoder levels of the *Visual attention U-Net* of the previous section, with the *MultiRes block* and also replaced the skip connections with the *Res path* scheme, described above and shown in Fig. 8.

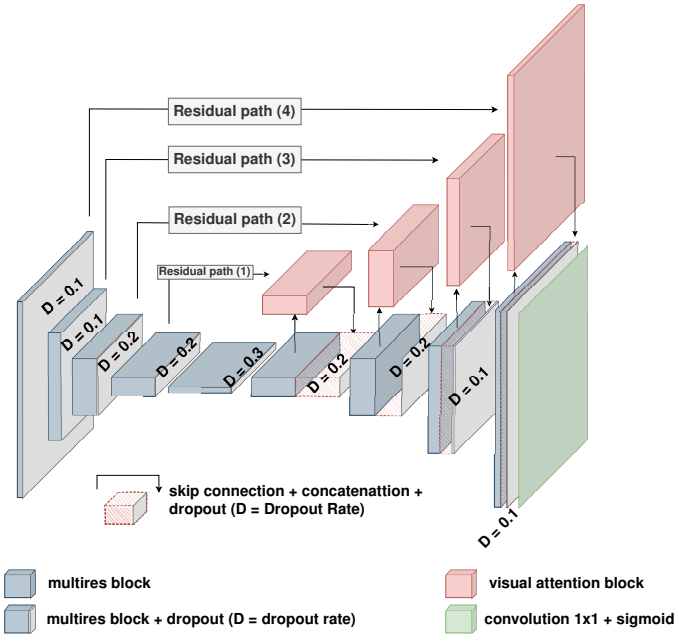

Fig. 8: MultiRes Visual Attention U-Net.

2.5. Dilated MultiRes Visual Attention U-Net (DMVAnet)

Inspired by the DeepLab network series [34, 26, 35] we incorporate a-trous (dilated) convolutions in our network layers. Dilated convolutions remedy the spatial resolution loss after a series of consecutive pooling layers. Dilated convolutions enlarge the receptive field of filters and also allow for dense feature extraction without excessive memory requirements.

If all convolutions are replaced by their dilated counterparts, network feature maps can ideally retain the original high resolution. However, such an architecture would introduce excessive resource requirements. For this reason, we only convert some of our network levels. If we only replace the last two encoder layers before the bottleneck, the network size, in terms of trainable parameters, already becomes too large. If we replace the last layer only, we do not benefit enough. In order to balance between the two concepts, we introduce an extra encoder (and decoder) layer right before the bottleneck and apply dilated convolution on that as well.

Fig. 9 shows the introduction of dilated convolutions in our MultiRes U-Net structure. We have added an extra encoder level with dilated convolutions and also used dilated convolutions in the bottleneck layer. We have also added an extra decoder level without increasing the feature maps dimensions. The feature maps dimensions of the first two decoder layers remain the same, while only the filter number increases. The residual paths and visual attention blocks follow the size of the encoder levels they are connected to. The network in Fig. 9 constitutes the proposed Dilated MultiRes Visual Attention U-Net (DMVAnet).

3. Ablation study

In this section, we present an extensive ablation study, in order to quantify the benefits of introducing several features to the proposed architecture, which served as a guide regarding the final network architecture and training process. In order to make the study more transparent unbiased against any specific network adaptation, we have performed the ablation study on the basic U-Net model (Section 2.1).

Several metrics were estimated for the evaluation purposes, but the final study ranking was based on the *F-measure*, *Distance Reciprocal Distortion (DRD)* and *PSNR*. Results were ranked according to each of these metrics, with the best receiving the highest value. When n different items were evaluated, the available ranks were $0, \ldots, n-1$. The overall rank was the sum of the individual metric ranks.

The evaluation metrics were measured separately on three evaluation datasets, the DIBCO 2011 [36], the H-DIBCO 2014 [37] and H-DIBCO 2016 [38], in order to investigate the consistency of the binarization process in different scenarios. The training datasets were DIBCO 2009 [39], H-DIBCO 2010 [40], H-DIBCO 2012 [41], Bickley-diary [42] and Synchromedia Multispectral datasets [43]

Individual ablation studies were performed in a cascaded manner, meaning that whenever a study resulted in a conclusion, the selected parameter option was finalized for the subsequent studies. For example, when the best upscaling method was identified, it was used in the subsequent experimentation.

3.1. Dropout usage

We compared our model with and without dropout layers and the results have shown that dropout gives marginally better results in two out of the three evaluation datasets. This led to the inclusion of dropout layers. The results are shown in Table 1.

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
without dropout	91.19	18.90	3.45	2
with dropout	**91.6**	**19.12**	**3.18**	**1**

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
without dropout	92.61	21.71	2.58	2
with dropout	**93.93**	**21.90**	**2.10**	**1**

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
without dropout	**89.90**	**18.87**	**3.52**	**1**
with dropout	89.75	18.81	3.59	2

Table 1: Ablation study on the use of dropout layers

3.2. Upscaling method

Since upscaling is a crucial part in U-Net architectures, we have examined whether it is better to upscale with the use of

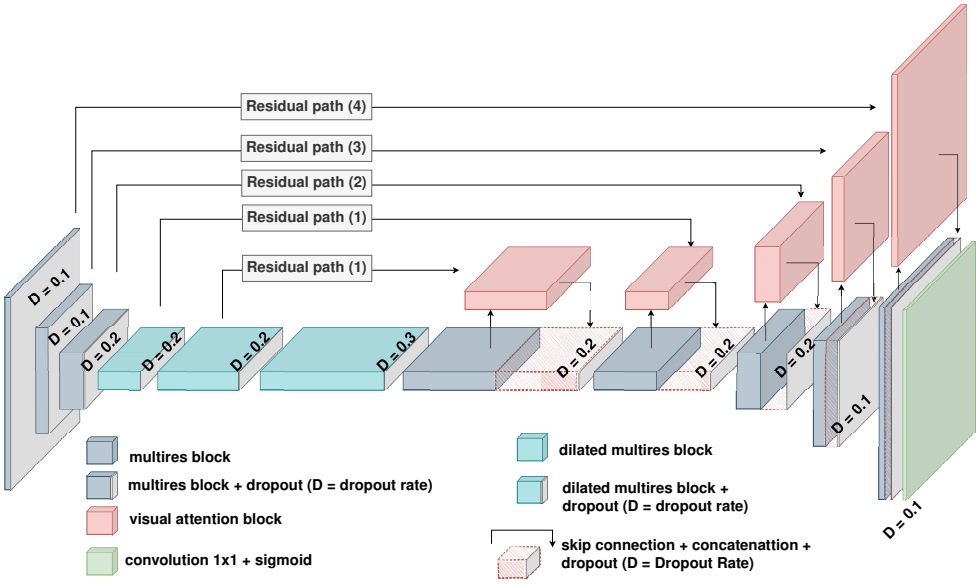

Fig. 9: The proposed Dilated MultiRes Visual Attention U-Net (DMVAnet)

a transpose convolution or with an upscaling interpolation algorithm. Transpose convolution also offers choices on the kernel size (among other parameters) that can result in the desired resolution. In our study, we have examined 1×1 and 3×3 kernel sizes, to keep complexity low. In total, the methods examined were transpose convolution with 1×1 kernel (k1), transpose convolution with 3×3 kernel (k3), Nearest-Neighbor (NN) interpolation and Bilinear (Bil) interpolation. The results are shown in Table 2.

The results show that upscaling with a transpose convolution and a 1×1 kernel is better in two out of three cases. In DIBCO2011, where it is not consistently best, it still yields the best PSNR score, therefore it was chosen as the preferred upscaling method. Transpose convolution provides the additional advantage of adjusting the number of filters at the same time, while an interpolation-based upscaling function would require a subsequent convolutional layer.

3.3. Loss function choice

The choice of a loss function is a vital part of the training of deep learning systems. We have examined the following loss functions:

- Binary Cross-Entropy (BCE)
- Mean Square Error (MSE)
- Dice loss (Dice)
- Inverse Peak Signal-to-Noise Ratio (InvPSNR)
- Differentiable F-Measure (DFM)

The Binary Cross-Entropy function is defined, as follows:

$$\mathcal{L} = -\frac{1}{N} \cdot \sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{1}$$

where N is the number of samples, y_i is the label for sample i and $\hat{y}_i$ is the predicted value for that sample. The Binary Cross-Entropy function is commonly used in binary classification tasks.

The Mean Square Error function definition is given by:

$$\mathcal{L} = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

where again N is the number of samples, y_i is the label for sample i and $\hat{y}_i$ is the predicted value for that sample. It is a commonly used measure of the performance of an estimator.

Document image datasets, such as the ones we use in this study, are inherently imbalanced, since the text pixels are considerably fewer than the background pixels. *Dice Loss* is a dominant loss function that aims to remedy such imbalances. Based on the Dice coefficient and originally proposed as a loss function in [44] and [45], it is mathematically formulated as:

$$\mathcal{L} = 1 - \frac{2 \sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i + \hat{y}_i} \tag{3}$$

In semantic segmentation, the Peak Signal-to-Noise Ratio is a measure of how the representation error (noise) compares to the maximum ground truth signal power. In the case of binary discrete signals, the metric is defined as

$$PSNR = 10 \cdot \log_{10} \frac{1}{MSE} \tag{4}$$

where $MSE = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$.

Since the loss function is a measure of cost or error, we use the inverse PSNR (InvPSNR):

$$\mathcal{L} = \frac{1}{PSNR} \tag{5}$$

In [46], Pastor-Pellicer et al. proposed the use of F-Measure as the loss function for problems with imbalanced datasets.

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
Transpose conv2d k1	88.17	**17.80**	5.38	3
Transpose conv2d k3	88.50	17.72	**4.71**	2
Nearest-Neighbor	**88.62**	17.72	4.96	**1**
Bilinear	88.30	17.70	4.99	4

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
Transpose conv2d k1	**89.17**	**20.68**	**3.42**	**1**
Transpose conv2d k3	88.54	20.23	3.66	2
Nearest-neighbor	87.79	20.13	3.86	4
Bilinear	89.14	20.39	3.54	3

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
Transpose conv2d k1	**89.16**	**18.57**	**3.90**	**1**
Transpose conv2d k3	89.04	18.47	3.98	2
Nearest-neighbor	86.91	18.30	4.53	4
Bilinear	88.44	18.44	4.21	3

Table 2: Ablation study on various upscaling methods used for the baseline U-Net training

More precisely, they suggested a differentiable version of the F-Measure function that can be incorporated into the back propagation process. It is given by the following formula:

$$\mathcal{L} = -\frac{(1+\beta^2)\sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} \hat{y}_i + \beta^2 \cdot y_i} \tag{6}$$

where y_i is the label for sample i, $\hat{y}_i$ is the predicted value for that sample and β is the weighting factor between precision and recall of the regular F-Measure. In our experiments, we used $\beta = 1$. We use the negative sign to change the monotonicity of the function, so that it can be used as a cost function.

Table 3 shows that the Dice loss provided the best results. For this reason, it was selected and at the same time linear combinations with some of the remaining loss functions were also examined. The selected combinations are shown in the following list:

- Dice + DFM
- Dice + MSE
- Dice + 10*InvPSNR

The weight in the Dice/InvPSNR combination was applied to scale the loss functions to comparable levels. Adjusting weights was attempted to other combinations as well, but did not provide any improvement against the simple addition.

The performance of all loss functions is summarised in Table 3. The Dice loss consistently performs better than the other

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
BCE	88.17	17.80	5.38	7
MSE	88.21	17.72	5.27	6
Dice	**91.13**	**18.94**	**3.57**	**1**
InvPSNR	89.21	18.19	4.58	5
DFM	87.39	17.52	5.31	8
Dice+DFM	90.17	18.73	3.73	4
Dice+MSE	90.85	18.77	3.58	2
Dice+(10*1.0/PSNR)	90.89	18.76	3.61	3

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
BCE	**89.17**	20.68	**3.42**	3
MSE	87.47	20.22	3.88	7
Dice	88.71	**21.33**	3.46	**1**
InvPSNR	87.67	20.77	3.79	6
DFM	85.23	19.63	4.44	8
Dice+DFM	88.05	21.03	3.61	5
Dice+MSE	88.27	21.17	3.57	4
Dice+(10*1.0/PSNR)	88.25	21.29	3.56	2

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
BCE	89.16	18.57	3.90	5
MSE	88.04	18.36	4.25	8
Dice	89.61	18.72	3.69	2
InvPSNR	88.26	18.51	4.07	6
DFM	87.93	18.43	4.15	7
Dice+DFM	**89.67**	**18.76**	**3.62**	**1**
Dice+MSE	89.16	18.66	3.80	4
Dice+(10*1.0/PSNR)	89.32	18.69	3.73	3

Table 3: Ablation study on various loss functions used for the baseline U-Net training

loss functions. Other loss functions that may occasionally appear to provide better scores, tend to behave much worse in the remaining cases. On the other hand, Dice loss, even when it is not ranked to the top on some metric of some evaluation dataset, it is still among the best. Since generalisation is a desired element for machine learning solutions, we selected Dice loss as the loss function of our study.

3.4. Boundary conditions

Images in the training and evaluation datasets are split into 256×256 patches. However, the original image dimensions are not always multiples of 256. This implies that some strategy should be applied for the patches that lie around the image boundaries. We choose not to crop the image to dimensions multiple of 256, so as not to lose information. Instead, we expand the image appropriately and symmetrically on all sides. This enables a range of options among which the simpler ones

are to fill the new space with a solid colour. This option is rejected, since it modifies the image in an unnatural way.

We examine two options, reflecting the image with respect to the boundary line (*reflect 101*) and replicating the image edge. In both cases, the ground truth images are modified in the same way. Results are shown in Table 4, where *Reflect 101* is clearly better. It is an anticipated outcome, since it modifies the boundary information in a more realistic way, resulting in a network that generalises better. In the opposite case, the network learns patterns that are unlikely to appear in evaluation scenarios.

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
Reflect 101	**91.61**	**19.12**	**3.18**	**1**
Replicate	91.13	18.94	3.57	2

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
Reflect 101	**93.93**	**21.90**	**2.10**	**1**
Replicate	88.71	21.33	3.46	2

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
Reflect 101	**89.76**	**18.82**	**3.59**	**1**
Replicate	89.61	18.72	3.69	2

Table 4: Ablation study on boundary conditions when adjusting the dimensions of the training and evaluation images to multiples of the patch dimension (256)

3.5. Architectures Comparison

All previous ablation study steps were performed on the basic U-Net architecture of Fig. 1. In this final step, we had to evaluate the architecture enhancements we have presented, determine the one that performs best and verify that combining the proposed architectural blocks does improve the performance. In order not to underestimate the learning capacity of the networks, they were trained for 150 epochs with all data augmentations, as presented in the following section. The results are shown in Table 5 and the overall ranking in Table 6.

The Dilated MultiRes Visual Attention U-Net (DMVAnet) appears to be the best of the architectures we have evaluated. The overall ranking does show that combining all the proposed modifications improves the performance.

As it can be seen in Table 5, moving from VAnet to MVAnet seems to produce a worse network. In two out of the three evaluation datasets, DIBCO 2011 and H-DIBCO 2016, the MVAnet is ranked lower than its parent, the VAnet. In DIBCO 2011 it is ranked 3rd while in H-DIBCO 2016 4th. An observation such as this could support the argument that perhaps adding the MVAnet extensions produces an under-performing network. This made the authors apply the subsequent step of modifications, the dilated convolutions, to both the VAnet and MVAnet and examine which performs best. The results are shown in Table 7. This extra ablation step ensures that the dilated convolutions should be applied to the MVAnet and thus the MultiRes step should not be skipped in the chain of modifications that we proposed.

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
Baseline	92.24	19.7	2.88	4
Res-U-Net	91.75	19.51	3.04	5
VAnet	93.11	20.26	2.42	2
MVAnet	92.45	20.03	2.7	3
DMVAnet	**94.3**	**20.93**	**2.01**	**1**

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
Baseline	95.66	22.49	1.6	5
Res-U-Net	95.74	22.23	1.56	4
VAnet	95.9	22.62	1.5	3
MVAnet	95.98	22.96	1.54	2
DMVAnet	**97.73**	**24.21**	**0.75**	**1**

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
Baseline	90.45	18.93	3.45	5
Res-U-Net	90.61	19.1	3.36	3
VAnet	**90.63**	**19.18**	**3.24**	**1**
MVAnet	90.43	18.99	3.42	4
DMVAnet	90.62	19.08	3.34	2

Table 5: Architecture evaluation results

Method	Rank
Baseline	5
Resnet	4
VAnet	2
MVAnet	3
DMVAnet	**1**

Table 6: Architecture evaluation overall ranking

4. Complexity Comparison

Table 8 lists the complexity of the proposed DMVAnet method in terms of training parameters compared against the complexities of the SOTA methods that are listed in the experiments. It is clear that the proposed DMVAnet is the smallest network in this list, after SauvolaNet and SAE, which are the most lightweight binarization networks. The remaining SOTA methods feature 1.3 to 32 times more network parameters, which does not necessarily translate to equivalent performance, as it will be shown during the experiments in the forthcoming section.

DIBCO 2011

Method	F-measure	PSNR	DRD	Rank
DVAnet	93.89	20.40	2.17	2
DMVANet	**94.3**	**20.93**	**2.01**	**1**

H-DIBCO 2014

Method	F-measure	PSNR	DRD	Rank
DVAnet	96.23	23.04	1.05	2
DMVANet	**97.73**	**24.21**	**0.75**	**1**

H-DIBCO 2016

Method	F-measure	PSNR	DRD	Rank
DVAnet	**90.89**	**19.22**	**3.02**	**1**
DMVANet	90.62	19.08	3.34	2

Table 7: Dilated convolutions comparisons between DVAnet and DMVAnet.

Method	Parameters
DMVAnet	6.5M
MRAtt [17]	8.5M
CMU-Net [18]	13M
DeepOtsu [12]	17M
HTR [20]	19M
DSN [14]	24M
CT-Net [19]	45M
cGANs [15]	57M
CCDWT-GAN [21] [22]	179M
DD-GAN [16]	210M
PDNet [47]	-

Table 8: Comparison based on network size in terms of number of trainable parameters.

5. Performance Comparison

In this section, we evaluate the proposed DMVAnet model against state-of-the-art binarization approaches. In order to compare against reported results in modern literature, we replicate two experiments, as described in two state-of-the-art approaches [11], [19]. The architecture of the DMVAnet model remains the same in each experiment, however, each time the training and evaluation sets are changed in order to match those described in each experiment. Specific details are provided in the respective experiment section. The architectures that were used in our comparison contained both traditional image processing based methods and deep learning based methods. More specifically, the following methods were examined: Otsu [1], Sauvola [2], Su et al [5], Howe [4], Lelore et al. [6], Nafchi et al. [7], Mitianoudis et al. [8], Jia et al. [9], GiB [10], SauvolaNet [11], HTR [20], CMU-Net [18], CT-Net [19], cGANs [15], DeepOtsu [12], DSN [14], MRAtt [17], SAE [13], DD-GAN [16], PDNet [47], CT-Net-3 [19], CTada -Net-3 [19] In addition to the above, we also compare against CCDWT-GAN [21, 22], since it is a very recent method and has been evaluated under the same conditions (training and evaluation datasets).

In all experiments, the training images were split in 256×256 patches. Since the original image dimensions are not multiples of 256, the border interpolation method used was the *"reflect 101"*, which preserves the last column or row and reflects an appropriate number of the preceding or succeeding rows or columns.

The following augmentations were added to the training samples in a cascaded manner:

- Scale augmentation: by 1.7 on each dimension, on all training samples.
- Contrast augmentation: contrast was reduced with the transformation $0.3 \times I + 90$, where I is the pixel intensity (on each RGB channels), on all training samples.
- Noise augmentation: Gaussian and Salt & Pepper noise was added to the training samples. The Gaussian noise variance was 0.02 and the Salt & Pepper noise probability was 0.05. Each of these two types of noise was added with a probability of 0.5 on all training samples.
- Geometrical augmentation: training samples were rotated and flipped. Rotation was done in multiples of 90 degrees and flipped was done along both axes. The rotation angle and the flip direction were randomly chosen. Both rotation and flip were added with probability of 0.7 each.

Training of the proposed DMVAnet was performed for 150 epochs using the Adam optimizer and a learning rate of $\eta = 0.0001$ (the default Adam learning rate of $\eta = 0.001$ did converge faster but produced slightly inferior results). Training and evaluation was done on an Ubuntu 20.04 PC with 64GB RAM, an Intel i9 2.5 GHz 16-Core CPU and an NVIDIA GeForce RTX 3090 GPU. The architecture was developed in Python v3.8.10 and Tensorflow v2.10.0. The developed code is available via the following url[1].

5.1. Comparisons & results

In both following experiments, for each evaluation dataset, predictions were made by directly applying our model to each evaluation dataset. Result metrics were collected and combined to produce the results in Tables 9, 10, 11, 12.

In addition to the metrics, predicted images were produced and can be seen in Fig. 10, 11, 12, 13. Images are quite homogeneous in background, since this is the usual form in text binarization datasets, but contain certain common noise elements, such as stains, poor lighting conditions, textures and ink bleeding.

[1] `https://github.com/detsikas/DMVAnet`

Method	DIBCO 2011	H-DIBCO 2014	H-DIBCO 2016	Overall rank
Otsu [1]	9	10	9	9
Howe [4]	8	5	10	8
MRAtt [17]	6	9	2	6
DeepOtsu [12]	5	7	3	5
SAE [13]	7	8	6	7
DSN [14]	4	3	7	4
DD-GAN [16]	-	4	8	-
cGANs [15]	3	6	**1**	3
Sauvola [2]	10	11	11	10
Sauvola MS[48]	11	12	12	11
SauvolaNet [11]	**1**	**1**	5	**1**
DMVAnet	2	2	4	2

Table 9: Overall ranking for Experiment 1 (Dashes show that the method is not ranked because metrics are missing).

DIBCO 2011

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	82.1	84.80	15.7	9.0
Howe [4]	91.7	92.00	19.3	3.4
MRAtt [17]	93.16	95.23	19.78	2.2
DeepOtsu [12]	93.4	95.80	19.9	1.9
SAE [13]	92.77	95.68	19.55	2.52
DSN [14]	93.3	96.40	20.1	2.0
DD-GAN [16]	-	-	-	-
cGANs [15]	93.81	95.26	20.3	**1.82**
Sauvola [2]	82.1	87.70	15.6	8.5
Sauvola MS[48]	79.7	81.78	14.91	11.67
SauvolaNet [11]	**94.32**	**96.40**	20.55	1.97
DMVAnet	94.3	96.32	**20.93**	2
Rank	2	3	1	4

H-DIBCO 2014

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	91.7	95.70	18.7	2.7
Howe [4]	96.5	97.40	22.2	1.1
MRAtt [17]	94.9	95.98	21.09	1.85
DeepOtsu [12]	95.9	97.20	22.1	0.9
SAE [13]	95.81	96.78	21.26	1.0
DSN [14]	96.7	97.60	23.2	0.7
DD-GAN [16]	96.27	97.66	22.60	1.27
cGANs [15]	96.41	97.55	22.12	1.07
Sauvola [2]	84.7	87.80	17.8	2.6
Sauvola MS[48]	85.83	86.83	17.81	4.88
SauvolaNet [11]	**97.83**	**98.74**	24.13	**0.65**
DMVAnet	97.74	98.52	**24.21**	0.75
Rank	2	2	1	3

H-DIBCO 2016

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	86.6	89.90	17.8	5.6
Howe [4]	87.5	82.30	18.1	5.4
MRAtt [17]	**91.68**	94.71	19.59	2.93
DeepOtsu [12]	91.4	94.30	19.6	2.9
SAE [13]	90.72	92.62	18.79	3.28
DSN* [14]	90.1	83.60	19.0	3.5
DD-GAN [16]	89.98	85.23	18.83	3.61
cGANs [15]	91.66	94.58	**19.64**	**2.82**
Sauvola [2]	84.6	88.40	17.1	6.3
Sauvola MS[48]	79.84	81.61	14.76	11.50
SauvolaNet [11]	90.25	95.26	18.97	3.51
DMVAnet	90.63	**95.35**	19.08	3.34
Rank	5	1	4	5

Table 10: Comparison on DIBCO 2011, H-DIBCO 2014, H-DIBCO 2016 datasets following the guidelines of Experiment 1 (source [11]).

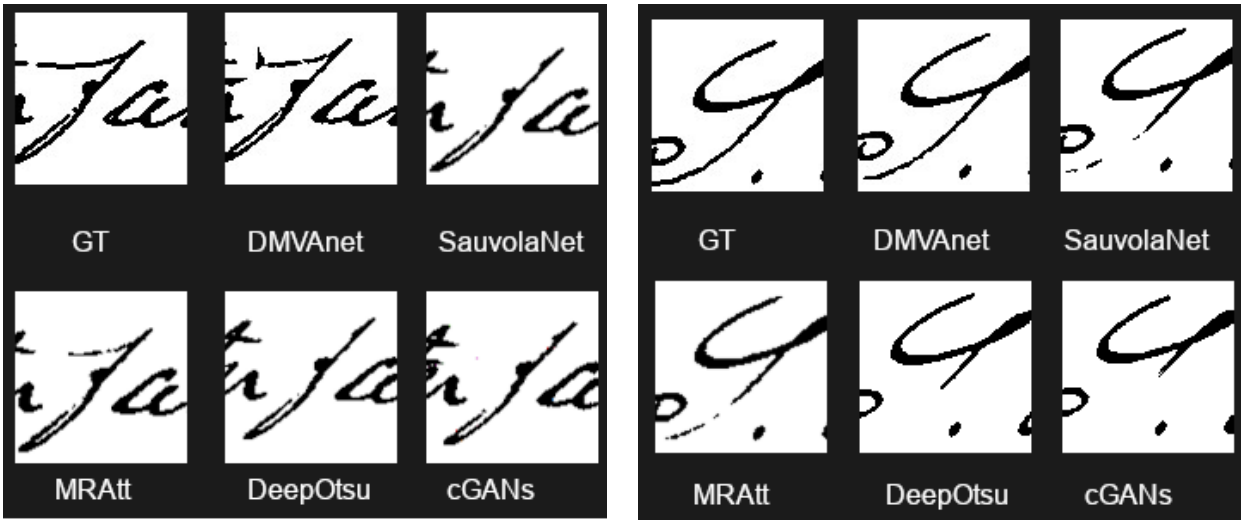

Fig. 10: Prediction details of the DMVAnet against the competitor methods of experiment 1.

Fig. 11: Sample Binarization results from the methods in Experiment 1 [11]. GT denotes the ground truth binarization and DMVAnet the proposed method.

Method	2009	2011	2014	2016	2017	2018	Overall rank
Otsu [1]	11	14	12	10	9	9	7
Sauvola [2]	10	13	13	13	8	8	8
Su et al [5]	7	12	11	12	-	-	-
Howe [4]	4	10	6	9	6	6	5
Lelore et al. [6]	5	8	9	11	-	-	-
Nafchi et al. [7]	-	-	-	-	-	-	-
Mitianoudis et al. [8]	9	11	-	-	-	-	-
Jia et al. [9]	6	9	10	6	7	7	6
GiB [10]	-	-	-	-	-	-	-
DeepOtsu [12]	-	6	8	3	-	-	-
DSN [14]	-	7	3	7	-	-	-
PDNet [47]	-	-	-	-	-	-	-
cGANs [15]	2	5	7	**1**	5	4	3
CT-Net-3 [19]	8	**1**	**1**	8	**1**	3	4
CTada -Net-3 [19]	3	3	4	5	2	**1**	2
CCDWT-GAN [21, 22]	-	4	5	2	4	2	-
DMVAnet	**1**	2	2	4	3	5	**1**

Table 11: Overall ranking for Experiment 2 (Dashes show that the method is not ranked because metrics are missing).

Fig. 12: Prediction details of the DMVAnet against some of the competitor methods of experiment 2. The red annotations come from the authors of [19] and show missing text pixels.

DIBCO 2009

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	78.6	80.53	15.31	22.57
Sauvola [2]	85.37	89.08	16.37	7.08
Su et al [5]	93.02	94.61	19.41	2.64
Howe [4]	94.04	95.06	20.43	2.10
Lelore et al. [6]	93.93	95.10	20.21	2.17
Nafchi et al. [7]	93.36	-	19.55	-
Mitianoudis et al. [8]	90.27	92.69	18.08	3.71
Jia et al. [9]	93.05	94.60	19.29	2.40
GiB [10]	92.5	-	19.26	2.41
DeepOtsu [12]	-	-	-	-
DSN [14]	-	-	-	-
PDNet [47]	91.5	-	19.25	3.06
cGANs [15]	94.1	95.26	20.30	1.82
CT-Net-3 [19]	92.08	94.31	19.77	3.58
CTada -Net-3 [19]	94.18	95.80	20.50	2.56
CCDWT-GAN [21, 22]	-	-	-	-
DMVAnet	**95.7**	**96.84**	**21.42**	**1.35**
Rank	1	1	1	1

DIBCO 2011

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	82.10	85.96	15.72	8.95
Sauvola [2]	82.14	87.70	15.65	8.50
Su et al [5]	87.83	90.24	17.71	4.66
Howe [4]	90.79	92.28	19.01	4.46
Lelore et al. [6]	92.48	94.11	19.37	2.97
Nafchi et al. [7]	92.57	-	19.29	2.28
Mitianoudis et al. [8]	89.13	93.79	17.90	3.47
Jia et al. [9]	91.92	95.09	18.98	2.64
GiB [10]	90.33	-	18.29	2.99
DeepOtsu [12]	93.4	95.8	19.9	1.9
DSN [14]	93.3	96.4	20.1	2.0
PDNet [47]	91.87	-	19.07	2.57
cGANs [15]	93.81	95.70	20.26	1.81
CT-Net-3 [19]	**95.27**	**97.24**	**21.50**	**1.37**
CTada -Net-3 [19]	94.17	96.92	20.76	1.69
CCDWT-GAN [21, 22]	94.08	97.08	20.51	1.75
DMVAnet	94.68	96.71	21	1.62
Rank	2	4	2	2

H-DIBCO 2014

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	91.62	95.69	18.72	2.65
Sauvola [2]	84.70	87.88	17.81	4.77
Su et al [5]	94.38	95.94	20.31	1.95
Howe [4]	96.49	97.38	22.24	1.08
Lelore et al. [6]	96.14	96.73	21.88	1.25
Jia et al. [9]	94.98	97.18	20.56	1.50
Mitianoudis et al. [8]	87.57	-	18.43	-
GiB [10]	94.00	-	19.93	1.79
DeepOtsu [12]	95.9	97.2	22.1	0.9
DSN [14]	96.66	97.59	23.23	0.79
PDNet [47]	89.99	-	20.52	7.42
cGANs [15]	96.41	97.55	22.12	1.07
CT-Net-3 [19]	**97.70**	**98.74**	**23.92**	**0.65**
CTada -Net-3 [19]	96.91	97.93	22.62	0.88
CCDWT-GAN [21, 22]	96.65	98.19	22.27	0.96
DMVAnet	97.55	98.58	23.62	0.71
Rank	2	2	2	2

H-DIBCO 2016

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	86.59	89.92	17.79	5.58
Sauvola [2]	84.64	88.39	17.09	6.27
Su et al [5]	84.75	88.94	17.64	5.64
Howe [4]	87.47	92.28	18.05	5.35
Lelore et al. [6]	87.21	88.48	17.36	5.27
Jia et al. [9]	90.48	93.27	19.30	3.97
Mitianoudis et al. [8]	86.89	-	17.60	-
GiB [10]	91.15	-	19.18	3.20
DeepOtsu [12]	91.4	94.3	19.6	2.9
DSN [14]	90.10	93.57	19.01	3.58
PDNet [47]	90.18	-	18.99	3.61
cGANs [15]	**91.66**	94.58	19.64	**2.82**
CT-Net-3 [19]	89.62	91.60	18.63	4.70
CTada -Net-3 [19]	91.07	94.34	19.22	3.29
CCDWT-GAN [21, 22]	91.46	**96.32**	**19.66**	2.94
DMVAnet	90.83	95.23	19.23	3.2
Rank	6	2	5	4

DIBCO 2017

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	77.73	77.89	13.85	15.54
Sauvola [2]	77.11	84.10	14.25	8.85
Su et al [5]	-	-	-	-
Howe [4]	90.10	90.95	18.52	5.12
Lelore et al. [6]	-	-	-	-
Jia et al. [9]	85.59	86.38	16.39	7.99
Mitianoudis et al. [8]	-	-	-	-
GiB [10]	-	-	-	-
DeepOtsu [12]	-	-	-	-
DSN [14]	-	-	-	-
PDNet [47]	-	-	-	-
cGANs [15]	90.73	92.58	17.83	3.58
CT-Net-3 [19]	**92.72**	94.31	**19.17**	2.79
CTada -Net-3 [19]	92.65	94.73	**19.17**	2.65
CCDWT-GAN [21, 22]	90.95	93.79	18.57	2.94
DMVAnet	92.2	**95.14**	18.73	**2.6**
Rank	3	1	3	1

H-DIBCO 2018

Method	FM	F_{ps}	PSNR	DRD
Otsu [1]	51.45	53.05	9.74	59.07
Sauvola [2]	67.81	74.08	13.78	17.69
Su et al [5]	-	-	-	-
Howe [4]	80.84	82.85	16.67	11.96
Lelore et al. [6]	-	-	-	-
Jia et al. [9]	76.52	79.90	17.00	8.11
Mitianoudis et al. [8]	-	-	-	-
GiB [10]	-	-	-	-
DeepOtsu [12]	-	-	-	-
DSN [14]	-	-	-	-
PDNet [47]	-	-	-	-
cGANs [15]	87.73	90.60	18.37	4.58
CT-Net-3 [19]	88.90	91.45	18.84	5.58
CTada -Net-3 [19]	**92.23**	94.97	**20.13**	**2.70**
CCDWT-GAN [21, 22]	91.66	**95.53**	20.02	2.81
DMVAnet	85.9	89.45	18.16	6.99
Rank	5	5	5	5

Table 12: Comparison on DIBCO 2009, DIBCO 2011, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 and H-DIBCO 2018 datasets following the guidelines of Experiment 2. Dashes indicate missing metrics. (source [19])

Fig. 13: Sample Binarization results from the methods in Experiment 2 [19]. GT denotes the ground truth binarization and DMVAnet the proposed method.

5.1.1. Experiment 1 [11]

First, we compare the DMVAnet model to the scenario described in [11]. In this experiment, the training datasets are DIBCO 2009 [39], H-DIBCO 2010 [40] and H-DIBCO 2012 [41] datasets, as well as the Bickley-diary [42] and the Synchromedia Multispectral dataset [43]. The evaluation datasets are the DIBCO 2011 [36], the H-DIBCO 2014 [37] and H-DIBCO 2016 [38] datasets.

We first show the overall ranking against the competitor methods over all the evaluation datasets in Table 9. Our method ranks second only to the SauvolaNet [11] method. To calculate the ranking, we rank all methods for each metric and each evaluation dataset. We add the metric ranks and calculate the total rank for each evaluation dataset by sorting the sums from lowest to highest. The individual dataset ranks are shown in the corresponding columns of Table 9. Finally, we add the individual dataset ranks and calculate the overall rank, which is the last column. Methods that do not have values for all metrics and all evaluation datasets are not considered in the ranking. In the overall ranking, it can be observed that the proposed DMVAnet is the second network in performance overall in Experiment 1.

Table 10 shows more analytically the performance of the DMVAnet model in comparison with other SOTA Binarization methods. The last row in each table indicates the relative rank of the proposed method among the comparison results. Our method ranks among the top for the DIBCO 2011 and H-DIBCO 2014 datasets, scoring the best PSNR in both datasets and the second best FM and F_{ps}. The H-DIBCO 2016 dataset is one of the most challenging to binarize because the document images exhibit very strong bleed-through of the back page ink. Our model performance is lower on that dataset in terms of FM, PSNR and DRD, however, it gets the first place in terms of F_{ps}.

Table 8 shows, with the exception of SauvolaNet and SAE (which is outperformed by our model), that our model has significantly fewer parameters compared to all other methods. As described before, the simplicity of our model also lies in the fact that, unlike its competitors, it is a one-shot prediction method, with a single DNN without any pre- or post- processing steps. It is clear that the proposed DMVAnet method ranks second in performance in this Experiment.

Finally, Fig. 10 shows some image details of the binarized images between the main competitor methods of Experiment 1. The overall good performance of the proposed method is evident, which shows that the proposed method is attentive to details. In addition, Fig. 11 shows predictions of complete images by most competing methods that take part in the experiment.

5.1.2. Experiment 2 [19]

In each of the experiments described in [19], a DIBCO dataset is used as the evaluation dataset and the remaining are treated as training datasets. In total, the DIBCO datasets used both for training and evaluation purposes are DIBCO 2009, H-DIBCO 2010, DIBCO 2011, H-DIBCO 2012, DIBCO 2013 [49], H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 [50] and H-DIBCO 2018 [51].

The training sets also include the Bickley-diary dataset, the Persian Heritage Image Binarization dataset (PHIDB) [52] and the Synchromedia Multispectral dataset [43].

We performed the DIBCO 2009, DIBCO 2011, H-DIBCO 2014, H-DIBCO 2016, DIBCO 2017 and H-DIBCO 2018 experiments. Table 11 shows the overall ranking against the competitor methods over all evaluation datasets. The ranking mechanism is the same as the one described in Experiment 1. Despite, the fluctuation in individual datasets, our method ranks first against all other methods in the experiments, which implies that in general the proposed method outperforms more complex networks and offers a reliable lightweight architecture for the problem of document image binarization.

Extensive results are listed in Table 12. Our method is the top ranking method for DIBCO 2009. For DIBCO 2011 and H-DIBCO 2014, it is outperformed only by CT-Net-3, which is of much higher complexity (seven times more parameters, see Table 8). With the exception of F_{ps}, where it is first, our method is ranked lower in the H-DIBCO 2016 dataset experiment, but still, all the outperforming deep learning methods have significantly higher complexity. As mentioned before, the H-DIBCO 2016 is a particularly challenging dataset, due to the strong bleed-through exhibited. The proposed DMVAnet exhibits a similar performance to the one observed in Experiment 1. For DIBCO 2017, our method exhibits the best pseudo f-measure and DRD values, while it is outperformed only by CT-Net-3 method variations. Finally, H-DIBCO 2018 dataset experiment renders our method fifth among the examined methods, all of which though have much higher complexity. Again, H-DIBCO 2018 presents special challenges such as the strong bleed-through, strong paper stains and page margins not seen in other datasets. CCDWT-GAN [21, 22] lacks overall ranking, since no results were reported for DIBCO 2009. However, even in the theoretical case that the CCDWT-GAN method is ranked first in DIBCO 2009, overall ranking calculations still render our method top.

Prediction details on several images of the dataset used in Experiment 2 are depicted in Fig. 12 against the main competitors of the proposed DMVAnet. It can be observed that the proposed network is very attentive to details. Finally, Fig. 13 shows predictions of complete document images by the main competitor methods of Experiment 2.

6. Conclusions and future work

We have presented a single-step one-shot light-weight deep learning network for Document Image Binarization that requires no pre- or post- processing steps. The proposed DMVAnet combines a basic U-Net architecture with elements from modern deep learning architectures, including visual attention blocks, multi resolution blocks, residual connections and dilated convolutions that enhances its performance without inhibiting computational efficiency. The DMVAnet's performance was benchmarked with State of the Art methods on the popular (H-)DIBCO datasets and demonstrated that it exhibits better or comparable performance, but with much smaller complexity in terms of number of training parameters.

The rationale behind the architecture elements added to our architecture, was that each one individually is a well-established deep learning architectural trait that addresses prob-

lems, encountered in deep networks, and ensures specific optimization gains. All have benefits and drawbacks and had to be studied thoroughly within our model context. For instance, adding dilated convolutions stems from the need for wider model field-of-view. At the same time, balancing the number of dilated convolutions, prevents excessive model memory footprint growth. Wider convolutional kernels were implicitly introduced with the MultiRes blocks, which, at the same time, extend the network length and increase the need and dependence on residual connections. Careful combined application and study of all the described modifications to the initial basic U-net model, enabled the design of a highly effective, yet very low-cost binarization deep learning model.

For future research, continuing on the path of visual attention research, we will investigate more complex deep learning attention architectures, such as transformer networks. Even though transformer networks had been primarily introduced for sequential data problems, such as Natural Language Processing (NLP), they process the entire input at once and take advantage of contextual information through their innate attention mechanism. Due to these properties, the transformer network adaptation on image semantic segmentation task is a very promising and challenging task that should be investigated and extended with other successful and well-established contemporary deep learning architectural blocks

References

[1] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66. doi:10.1109/TSMC.1979.4310076.

[2] J. Sauvola, M. Pietikäinen, Adaptive document image binarization, Pattern Recognition 33 (2) (2000) 225–236. doi:https://doi.org/10.1016/S0031-3203(99)00055-2.
URL https://www.sciencedirect.com/science/article/pii/S0031320399000552

[3] W. Niblack, An Introduction to Digital Image Processing, Strandberg Publishing Company, DNK, 1985.

[4] N. R. Howe, Document binarization with automatic parameter tuning, International Journal on Document Analysis and Recognition (IJDAR) 16 (2012) 247–258.

[5] B. Su, S. Lu, C. L. Tan, Robust document image binarization technique for degraded document images, IEEE Transactions on Image Processing 22 (4) (2013) 1408–1417. doi:10.1109/TIP.2012.2231089.

[6] T. Lelore, F. Bouchara, FAIR: A fast algorithm for document image restoration, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 2039–2048. doi:10.1109/TPAMI.2013.63.

[7] H. Nafchi, R. Farrahi Moghaddam, M. Cheriet, Phase-based binarization of ancient document images: Model and applications, IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 23. doi:10.1109/TIP.2014.2322451.

[8] N. Mitianoudis, N. Papamarkos, Document image binarization using local features and gaussian mixture modelling, Image and Vision Computing 38 (2015) 33–51. doi:https://doi.org/10.1016/j.imavis.2015.04.003.
URL https://www.sciencedirect.com/science/article/pii/S0262885615000360

[9] F. Jia, C. Shi, K. He, C. Wang, B. Xiao, Degraded document image binarization using structural symmetry of strokes, Pattern Recognition 74 (2018) 225–240. doi:https://doi.org/10.1016/j.patcog.2017.09.032.
URL https://www.sciencedirect.com/science/article/pii/S0031320317303849

[10] S. Bhowmik, R. Sarkar, B. Das, D. Doermann, GiB: a game theory inspired binarization technique for degraded document images, IEEE Transactions on Image Processing PP (2018) 1–1. doi:10.1109/TIP.2018.2878959.

[11] D. Li, Y. Wu, Y. Zhou, SauvolaNet: Learning adaptive sauvola network for degraded document binarization (2021). arXiv:2105.05521.

[12] S. He, L. Schomaker, DeepOtsu: Document enhancement and binarization using iterative deep learning, Pattern Recognition 91 (2019) 379–390. doi:https://doi.org/10.1016/j.patcog.2019.01.025.
URL https://www.sciencedirect.com/science/article/pii/S0031320319300330

[13] J. Calvo-Zaragoza, A.-J. Gallego, A selectional auto-encoder approach for document image binarization, Pattern Recognition 86 (2019) 37–47. doi:https://doi.org/10.1016/j.patcog.2018.08.011.
URL https://www.sciencedirect.com/science/article/pii/S0031320318303091

[14] Q. N. Vo, S. H. Kim, H. J. Yang, G. Lee, Binarization of degraded document images based on hierarchical deep supervised network, Pattern Recognition 74 (2018) 568–586. doi:https://doi.org/10.1016/j.patcog.2017.08.025.
URL https://www.sciencedirect.com/science/article/pii/S0031320317303394

[15] J. Zhao, C. Shi, F. Jia, Y. Wang, B. Xiao, Document image binarization with cascaded generators of conditional generative adversarial networks, Pattern Recognition 96 (2019) 106968. doi:https://doi.org/10.1016/j.patcog.2019.106968.
URL https://www.sciencedirect.com/science/article/pii/S0031320319302717

[16] R. De, A. Chakraborty, R. Sarkar, Document image binarization using dual discriminator generative adversarial networks, IEEE Signal Processing Letters 27 (2020) 1090–1094. doi:10.1109/LSP.2020.3003828.

[17] X. Peng, C. Wang, H. Cao, Document binarization via multi-resolutional attention model with DRD loss, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 45–50. doi:10.1109/ICDAR.2019.00017.

[18] S. Kang, B. K. Iwana, S. Uchida, Complex image processing with less data—document image binarization by integrating multiple pretrained U-Net modules, Pattern Recognition 109 (2021) 107577. doi:https://doi.org/10.1016/j.patcog.2020.107577.
URL https://www.sciencedirect.com/science/article/pii/S0031320320303800

[19] H. Sheng, L. Schomaker, CT-Net: Cascade T-Shape deep fusion networks for document binarization, Pattern Recognition 118. doi:10.1016/j.patcog.2021.108010.

[20] S. Khamekhem Jemni, M. A. Souibgui, Y. Kessentini, A. Fornés, Enhance to read better: A multi-task adversarial network for handwritten document image enhancement, Pattern Recognition 123 (2022) 108370. doi:https://doi.org/10.1016/j.patcog.2021.108370.
URL https://www.sciencedirect.com/science/article/pii/S0031320321005501

[21] R.-Y. Ju, Y.-S. Lin, J.-S. Chiang, C.-C. Chen, W.-H. Chen, C.-T. Chien, Ccdwt-gan: Generative adversarial networks based on color channel using discrete wavelet transform for document image binarization (2023). arXiv:2305.17420.

[22] R.-Y. Ju, Y.-S. Lin, C.-C. Chen, C.-T. Chien, J.-S. Chiang, Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks (2023). arXiv:2211.16098.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2015). arXiv:1512.03385.

[24] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2261–2269.

[25] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S... 11045 (2018) 3–11.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision, 2018.

[27] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation (2015). arXiv:1505.04597.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[29] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, ArXiv abs/1603.07285.

[30] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module (2018). arXiv:1807.06521.

[31] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning where to look for the pancreas (2018). arXiv:1804.03999.

[32] N. Ibtehaz, M. S. Rahman, MultiResUNet : Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Networks 121 (2020) 74–87. doi:https://doi.org/10.1016/j.neunet.2019.08.025.
URL https://www.sciencedirect.com/science/article/pii/S0893608019302503

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions (2014). arXiv:1409.4842.

[34] 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 (2016). doi:10.48550/ARXIV.1606.00915.
URL https://arxiv.org/abs/1606.00915

[35] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation (2017). doi:10.48550/ARXIV.1706.05587.
URL https://arxiv.org/abs/1706.05587

[36] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICDAR 2011 document image binarization contest (DIBCO 2011), in: 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1506–1510. doi:10.1109/ICDAR.2011.299.

[37] K. Ntirogiannis, B. Gatos, I. Pratikakis, ICFHR2014 competition on handwritten document image binarization (h-dibco 2014), in: 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014, pp. 809–813. doi:10.1109/ICFHR.2014.141.

[38] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, ICFHR2016 handwrit-

ten document image binarization contest (H-DIBCO 2016), in: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 619–623. doi:10.1109/ICFHR.2016.0118.

[39] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), in: 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 1375–1382. doi:10.1109/ICDAR.2009.246.

[40] I. Pratikakis, B. Gatos, K. Ntirogiannis, H-DIBCO 2010 - handwritten document image binarization competition, in: 2010 12th International Conference on Frontiers in Handwriting Recognition, 2010, pp. 727–732. doi:10.1109/ICFHR.2010.118.

[41] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012), in: 2012 International Conference on Frontiers in Handwriting Recognition, 2012, pp. 817–822. doi:10.1109/ICFHR.2012.216.

[42] F. Deng, Z. Wu, Z. Lu, M. S. Brown, Binarizationshop: a user-assisted software suite for converting old documents to black-and-white, in: ACM/IEEE Joint Conference on Digital Libraries, 2010.

[43] R. Hedjam, H. Z. Nafchi, R. F. Moghaddam, M. Kalacska, M. Cheriet, ICDAR 2015 contest on multispectral text extraction (MS-TEx 2015), in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1181–1185. doi:10.1109/ICDAR.2015.7333947.

[44] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. J. Cardoso, , in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, 2017, pp. 240–248. doi:10.1007/978-3-319-67558-9˙28.
URL https://doi.org/10.1007%2F978-3-319-67558-9_28

[45] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation (2016). doi:10.48550/ARXIV.1606.04797.
URL https://arxiv.org/abs/1606.04797

[46] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, M. J. Castro-Bleda, F-measure as the error function to train neural networks, in: I. Rojas, G. Joya, J. Gabestany (Eds.), Advances in Computational Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 376–384.

[47] K. R. Ayyalasomayajula, F. Malmberg, A. Brun, , Pattern Recognition Letters 121 (2019) 52–60. doi:10.1016/j.patrec.2018.05.011.
URL https://doi.org/10.1016%2Fj.patrec.2018.05.011

[48] G. Lazzara, T. Géraud, Efficient multiscale sauvola's binarization, International Journal on Document Analysis and Recognition (IJDAR) 17 (2013) 105–123. doi:10.1007/s10032-013-0209-0.

[49] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICDAR 2013 document image binarization contest (DIBCO 2013), 2011, pp. 1506–1510. doi:10.1109/ICDAR.2011.299.

[50] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, ICDAR2017 competition on document image binarization (DIBCO 2017), 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 01 (2017) 1395–1403.

[51] I. Pratikakis, K. Zagoris, P. Kaddas, B. Gatos, ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018), 2018, pp. 489–493. doi:10.1109/ICFHR-2018.2018.00091.

[52] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, M. Cheriet, An efficient ground truthing tool for binarization of historical manuscripts, in: 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 807–811. doi:10.1109/ICDAR.2013.165.