Unconstrained Handwritten Arabic Text-lines Segmentation based on AR2U-Net

Takwa Ben Aïcha Gader
University of Tunis
ENSIT-LaTICE, Tunis, Tunisia
takwa.ben.aichaa@gmail.com

Afef Kacem Echi
University of Tunis
ENSIT-LaTICE, Tunis, Tunisia
afef.kacem@ensit.rnu.tn

***Abstract*—Text-lines are hard to segment in the context of Arabic manuscripts, because of the narrowly spaced text-lines with touching or overlapping components, the varying spaces between words, the ascendant or descendant letters, special marks, and dots, calligraphy, etc. In this work, we proposed a system to automatically extract text-lines from images of unconstrained handwritten Arabic texts. Each text-line is detected by its baseline. The proposed system is based on text-line masks which are predicted by a deep neural network called AR2U-Net: a Recurrent Residual convolutional neural network based on the U-Net model with an Attention mechanism. We adjusted the AR2U-Net model to allow a pixel-wise classification and therefore to separate text-lines pixels from the background one. We tested it on BADAM: a Public Dataset for Baseline Detection in Arabic script Manuscripts that involves complex layouts as well as curved and arbitrarily oriented text-lines and overlaps between adjacent text-lines, words, or sub-words. Our model achieves the best performances with a Precision of 0.932% which competes with current state-of-the-art approaches.**

***Keywords*-Text-line segmentation; Baseline detection; Deep learning; U-Net; AU-Net; ResU-Net; RCNN; RU-Net; AR2U-Net;**

I. Introduction

Text-line segmentation is recognized as being an important step for off-line handwritten text recognition because inaccurate segmentation will cause errors in the recognition step. But this task remains a challenge. This can be attributed to many reasons: 1) the huge variability of handwritings among writers, 2) the cluttered writing style which may introduce ligatures between parts of words and makes overlap the adjacent text-lines and 3) the fluctuation of the baseline which results in different variations in the inclination within the same text-line and makes the problem much complicated. The language characteristics also can hold back achieving significant results in the segmentation problem solution, for example, the diacritic which lie below and above the words, the letters extensions (with ascenders or/and descenders) which easily introduce connections between successive text-lines, the words which are often divided into letters and sub-words and the spaces between them are variable, etc. Figure 1 shows an example of used documents.

The existing handwritten text-line segmentation methods can be categorized as top-down, bottom-up, hybrid methods or machine learning methods. Top-down methods, generally

Figure 1: Example of used documents.

based on projection profile, Hough transform, Gaussian filters, process the whole image and recursively subdivide it into smaller blocks to isolate the desired part. They assume that gap between adjacent text-lines is significant and the text-lines are reasonably straight, which may not be faithful in handwritten texts. Bottom-up methods use simple heuristic rules, analyzing the geometric relationships between neighboring blocks such as the distance or the overlap. As mentioned by [12], these methods have the merit to deal with noise problems and writing variation. But, as most of top-down methods, bottom-up methods require prior knowledge about the documents, such as text-line inter-spaces, its orientation, etc. to guide the segmentation. Hybrid methods combine bottom-up grouping and top-down partitioning in different ways. They must therefore mix different image processing methods to consider all possible features which make them complicated in computation and the design of a robust combination scheme is non-trivial.

Recently, an increasing interest in using segmentation-free

978-1-7281-9966-5/20/$31.00 ©2020 IEEE
DOI 10.1109/ICFHR2020.2020.00070

and statistical feature analysis methods in various sub-fields and tasks related to document image analysis. These methods treat the image as a whole without any prior knowledge. The literature review shows the success of deep learning-based methods for text-line segmentation of regular handwritten document images. In this work, we investigate the use of these methods for unconstrained handwritten Arabic text segmentation as well. More precisely, we used a U-Net based model, named AR2U-Net, for a text-line segmentation system which can handle variations in handwriting styles and overlaps between adjacent text-lines, words or sub-words, as it will be proven later. This model was first used in the medical image segmentation field [14]. We adjusted it to fight against the pixel location information loss that may be caused by the pooling and convolution operations by modifying the model inputs. This modification will be well explained later.

This paper is divided into four sections. The first section gives a brief overview of some related works. The second section outlines the used model. The third section discusses the obtained results. The final section draws our conclusions and gives some prospects.

II. State-of-the-Art

In the literature, there are many text-line segmentation methods for handwritten documents. Smearing method [3] which fills the space between consecutive foreground pixels can be used on skewed documents [4] as well. Grouping method aggregates pixels or connected components in a bottom-up strategy and is superior in case of skewed and curved text-lines [5]. Machine learning algorithms, a type of grouping method, handle text-line segmentation as a pixel classification problem. Pixel classification can be done in a sliding window manner [6] which is not desirable due to redundant and expensive computation of overlapping areas in the sliding windows. On the other hand, the dense prediction does not suffer from redundant computation and has been successfully used for text-line segmentation of handwritten documents [7]. However, the text-line extraction of challenging documents has not been extensively studied.

Renton and al. [15], used an FCN variant with dilated convolutions for x-heights extraction, knowing that the text-line was presented by their x-heights. In [17], Kurar Barakat and al. used an FCN for text-lines masks prediction. In [16] Oliveira and al. proposed a CNN-based model for a pixel-wise prediction. A multi-usage model for pages extraction, layout analysis, paragraphs detection, and baselines extraction. The model returns the binary mask of the polygonal lines representing baselines. Mechi and al. [18] used a U-net model: an FCN variant for x-heights detection. Likewise, Neche and al. [19] used RU-net, an FCN variant for a pixel-wise classification for x-heights extraction.

Grüning et al. [20] proposed to segment text-lines, using an ARU-Net which is a U-Net extended with an Attention mechanism (A) and a Residual structure (R). This network returns two maps: the detected baseline map and the text-line beginning and end map. A super-pixel extraction is then performed on the output of the ARU-Net and the extracted ones are further clustered to build baselines. They reached 95% as correct text-line detection rate, on the cBad dataset. But a lot of processing and computations are needed, which makes it time-consuming. In [22] authors proposed a new Recurrent Residual Convolutional Neural Network (RRCNN) based on the U-Net model, which is called R2U-Net. The principal advantages of such architecture are: 1) a residual unit helps when training deep architecture, 2) feature accumulation with recurrent residual convolutional layers ensure better feature representation for segmentation tasks and 3) it allows designing better U-Net architecture using the same number of network parameters with better performance for segmentation tasks. The proposed model was tested on three benchmark datasets: blood vessel retina images, skin cancer lung lesion segmentation. The experimental results show superior performance on segmentation tasks compared to equivalent models including U-Net and residual U-Net (ResU-Net) [22].

Inspired by the work in [22], we propose a deep neural network, called AR2U-Net, which is a U-Net extended with an attention mechanism (A), a residual structure (R) and recurrent convolutional layers (R) to be described below. This model aims to detect then extract baselines in unconstrained handwritten Arabic texts from the BADAM data-set which seems to be significantly more difficult than comparable datasets (cBAD, KHATT, etc.).

III. Proposed Approach

To detect baselines which are virtual lines on which most characters rest, we perform a pixel-wise classification to separate text-lines pixels from the background one, using an AR2U-Net. Note that the first studies on the pixel labeling were carried out within the last few years. Most of the proposed architectures rely on Convolutional Neural Networks (CNNs) [8] whom their direct application for semantic segmentation is done be [9]. The authors presented Fully Convolutional Network (FCN) which combines local features to produce more meaningful high-level features using pooling layers. Pooling reduces the spatial dimension. Thus, the result suffers from a coarse resolution. This problem was remedied by [10] using a de-convolutional network on the sub-sampled output of the FCN. In [11] authors proposed to furthermore introduce shortcuts between layers of the same spatial dimension in the U-Net to make easy the combination of local low-level features and global higher-level features and to reduce the vanishing gradient problems [2].

In summary, U-Net is built upon the FCN and modified in a way that yields better segmentation in medical imaging.

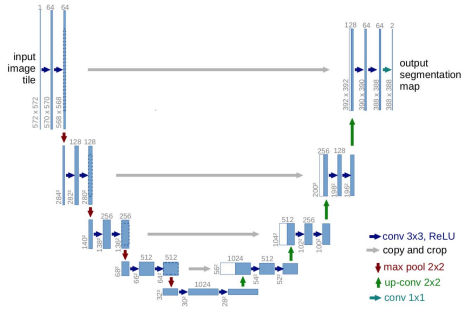

Figure 2: U-Net's architecture [13].

Compared to FCN, the two main differences are: 1) the U-Net is symmetric and 2) the skip connections between the down-sampling path and the up-sampling path apply a concatenation operator instead of a sum. These skip connections intend to provide local information to the global information while up-sampling. Because of the U-Net symmetry, the network has a large number of feature maps in the up-sampling path, which allows transferring information. By comparison, the basic FCN architecture only had the number of classes feature maps in its up-sampling path. The spatial dimensions in each scale space of the U-Net are all the same (see Figure 2) for a schematic representation of a U-Net. Thus, the output of the U-Net is a feature map (Z features in Figure 2) of the same spatial dimension as the input.

The used AR2U-Net model is in line with a variation of that reported in [22] with some changes. In [22], the authors proposed a Recurrent Residual Convolutional Neural Network (RRCN) based on the U-Net model (R2U-Net) for medical image segmentation. The objective is to ensure better performance during the training and the testing phase. For that, they used a Residual unit (R) which helps when training very deep architecture and Recurrent Residual (R2) convolutional layers for feature accumulation which allows better feature representation for segmentation tasks. In this work, we propose an extension of the U-Net architecture using an RRCN with an attention mechanism (A) that we call AR2U-Net. The attention mechanism is used to search in specific zones in an image, just like a human does, where he can orient itself in its research to specific zones of an image when searching for a precise pattern. These basic model concepts can be defined as follows:

1) **Recurrent Convolutional Unit**: This unit is used on RCNN which is a neural network with recurrent convolutional layers. These layers allow features accumulation and therefore a better representation of the features (see Figure 4 of [22]). Note that the recurrent convolutional operation includes one single convolution layer in the proposed AR2U-Net's architecture, as shown in Figure 3. RU-Net is a U-Net with Recurrent blocs.
2) **Residual Convolutional Unit**: In very deep neural networks, problems such as accuracy saturation or training performance degradation could be encountered. Deep Residual Networks (Res-Net) solve these problems using residual blocks (see Figure 4 of [22]). The residual connections allow such networks to be trained efficiently by decomposing them into smaller and simpler networks interconnected by shortcuts. The shortcut enables easy identity propagation and error back-propagation. R2U-Net is an RU-Net with residual blocs.

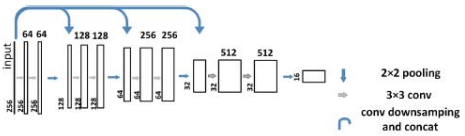

Figure 4: In the encoder, the input is convolved and down-sampled to form feature maps of different scales, and then concatenated with their respective layers [21].

3) **Attention-guided dense-up sampling block (AU block)**: To explicitly incorporate the potential to handle various writing styles, we use a pixel-wise spatial attention mechanism. This mechanism allows the selective use of visual information, focusing on image content at different positions and scales. Note that the basic U-Net uses de-convolution to up-sample the feature maps. In AR2U-Net, we use the bilinear up-sampling method as done by [23]. This method serves to extract the most important information from high-level and low-level features. To do so, high-level features are simply up-sampled and concatenated with the low-level features after the convolution layer (see Figure 3). Then, the important information is selected from the concatenation result. Such a mechanism can easily improve both the sensitivity and accuracy of the model for labeling tasks by ignoring feature activations in irrelevant regions. AR2U-Net is an R2U-Net with an attention mechanism.

In Figure 3, we display the used AR2U-Net's architecture with convolutional encoding and decoding units using attention mechanism, recurrent convolutional layers (RCL) and residual layers based on the U-Net.

As it can be seen, the AR2U-NET takes as input multi-scale images to avoid the loss of pixel location information, possibly due to convolution and pooling operations. More precisely, in the encoder part of the AR2U-Net model, the multi-scales inputs are convolved and down-sampled to return features maps of different sizes. These features maps are then concatenated with those of the corresponding scales generated by the pooling layer in the encoder (see Figure 4).

It is apparent from Figure 3 that there are many differ-

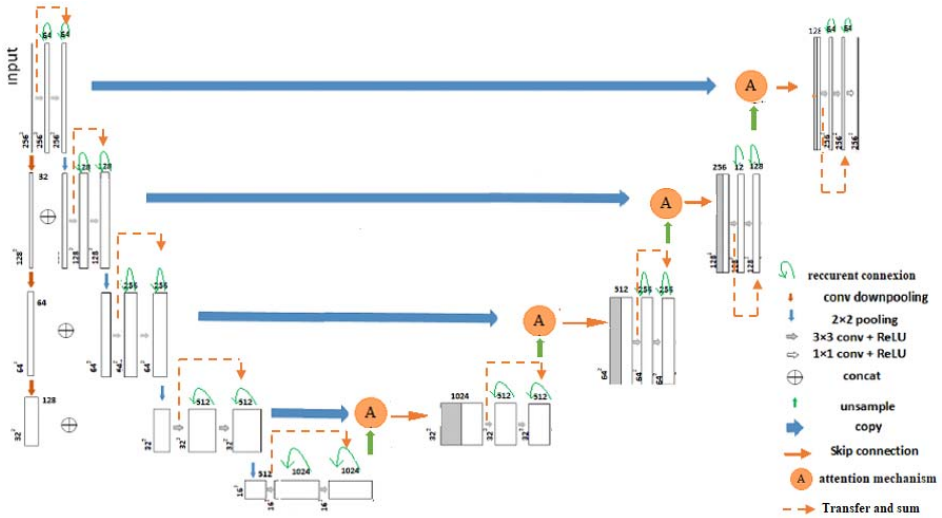

Figure 3: AR2U-Net's architecture.

ences between the used AR2U-Net and the basic U-Net model. Like U-Net, the AR2U-Net's architecture consists of convolutional encoding and decoding units, but the RCLs with residual units are used instead of regular forward convolutional layers in both the encoding and decoding units. The residual unit with RCLs has the merit to develop an efficient deep model. Also, the U-Net only shows the benefit during the training process in the form of better convergence whereas the AR2U-Net model is beneficial for both training and testing phases thanks to the useful feature accumulation method included in the RCL units of the model. Thanks to the feature accumulation, we have been able to extract very low-level features that are essential for text-line segmentation. As shown in Figure 3, the AR2U-Net has an asymmetrical architecture, unlike the U-NET model. More precisely in the decoder part, the U-NET model uses a de-convolution to up-sample the features map while the AR2U-Net model uses a bilinear up-sampling. Also, we have replaced the cropping and copying unit in the U-Net model by only concatenation operations which results in better performance. Finally, we multi-scale the AR2U-Net's input to have multi-scale features maps. Theses features maps are concatenated with the encoder layers inputs of the same size to avoid the loss of pixel location information.

IV. Experimental Results

To assess the accuracy of the proposed AR2U-Net model and to compare it to those of some related works, we used BADAM [**?**]: a publicly available and freely licensed dataset that has 400 annotated document images from from four digital collections of Arabic language manuscripts of different domains and periods. We first transformed the XML baselines information to ground truth images which contain two classes: 1) white for the ground truth and 2) black for the baseline. We then used 260 for the training set and 140 for the testing set. For all our experiments, the Adam Optimizer [24] and a learning rate of 10^{-5} are used. The network has been trained for 100 epochs. The metrics used to evaluate learning is classification loss. Table I summarizes the parameter settings. The model was trained on a Hp Z-440 workstation. During the training, the current state of the model at each step was evaluated on a hold-out validation dataset that is not part of the training dataset (we took 20% of the training set for the validation) and the best model was saved. Figure 5 displays the training and validation curves. Figure 6 shows an example of model prediction on an image of the datatest.

To evaluate the baseline extraction, we computed the Precision, Recall and F-measure. The obtained results are represented in Table II and can be compared to that of the only machine learning method used for baseline extraction and tested on the BADAM database as far as we know. Note that the obtained Precision, Recall, and F-measure values are slightly higher than those of the model reported in [**?**].

Table I: Parameter settings

Image pre-processing	260 paragraph images: 140 for training and 150 for testing. Images of different sizes. No further processing
Training setting	Initial weights: 0.1 for the 2 classes. Initial learning rate: 10^5. Optimizer : Adam. Initial number of epochs: 100. Batch-size per epoch: 128. Evaluation metrics: Accuracy, Loss and F-measure

Table II: Results for the BADAM dataset.

Model	Precision [25]	Recall [25]	F-measure [25]
[**?**]	0.941	0.901	0.924
AR2U-Net	**0.932**	**0.943**	**0.937**

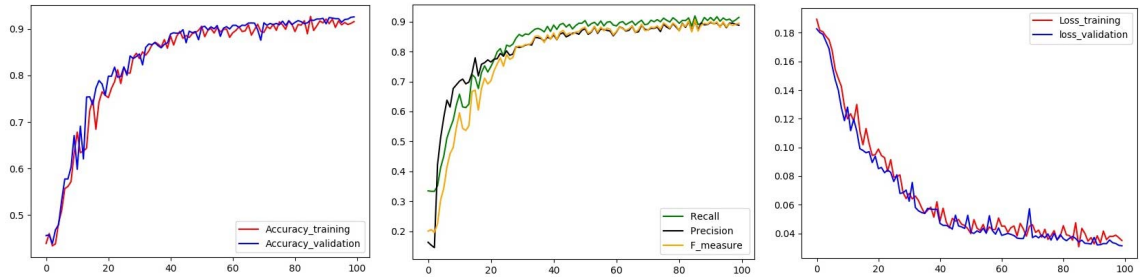

Figure 5: Training curves: a) Training and Validation Accuracy, b) Recall, Precision, F-measure Metrics, and c) Training and Validation Loss.

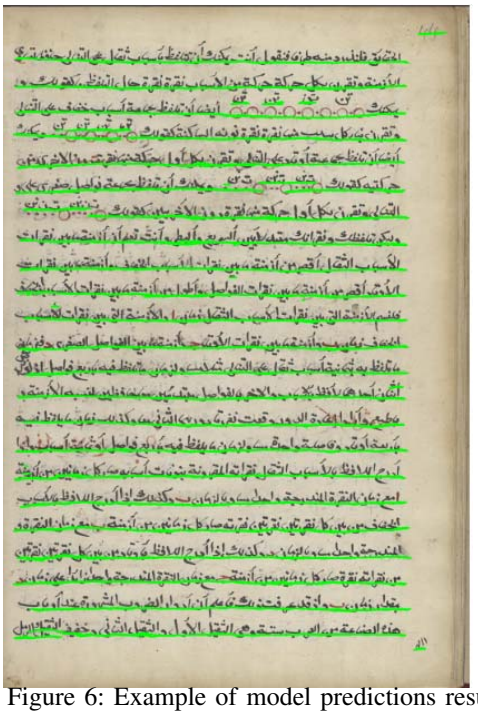

Figure 6: Example of model predictions result.

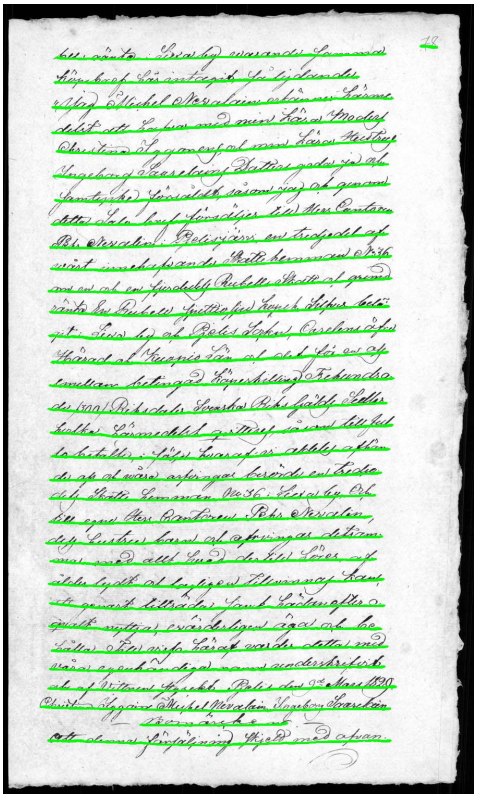

Figure 7: AR2U-Net tested on Latin handwritten text.

Maybe the results seem to be a bit lower compared to some Arabic manuscripts text-line segmentation systems. But it is noteworthy that experiments are carried on the BADAM database: a too complex compared to other databases such as KHATT, INF-ENIT, AHDB. As expected and shown in Figure 7, our system is also proven to be script-independent.

V. Conclusion and Future Work

The contributions of this work can be summarized as follows.

- the feature accumulation method which is provided by recurrent layers for a better features representation.
- the residual layers to improve training performance for very deep neural networks. Also, they used to fight against problems such as accuracy saturation and training performance degradation.
- the attention mechanism to handle with different writing styles by the selective use of visual information and the focus on image content at different positions and scales.
- the use of multi-scale inputs that generate multi-scale features maps to be concatenated with the inputs of the encoder's layers. This avoids the loss of pixel location information possibly due to convolution and pooling operations.
- the efficiency in terms of the number of network

parameters: The used model is designed to have the same number of network parameters when compared to U-Net, ResU-Net, RU-Net, and R2U-Net. Note that the recurrent and residual operations do not increase the number of network parameters. However, they have a significant impact on training and testing performance, as shown through a set of experiments on the public dataset BADAM.

Note that comparison against the recently proposed state-of-the-art method tested on the BADAM dataset shows superior performance, achieved by the AR2U-Net model, against the equivalent model with the same number of network parameters. To further our research we plan to increase the training set by annotating more images of the BADAM dataset and to use a special post-treatment to decrease the error rate and to segment document images into images of text-lines bounding boxes that correspond to each baseline detected. We also intend to evaluate the proposed model on other databases to be able to compare it to other machine-learning-based approaches. To increase the performance of our system, we plan this concatenation allows the avoidance of location information loss.

REFERENCES

[1] N. Ouwayed and A. Belaïd, "A general approach for multi-oriented text line extraction of handwritten documents", IJDAR, pp. 1-18, 2012.

[2] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", PMLR, vol. 9, pp. 249-256, 2010.

[3] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system", IBM journal of research and development, vol. 26, no. 6, pp. 647-656, 1982.

[4] A. Alaei, U. Pal, and P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation", Pattern Recognition, vol. 44, no. 4, pp. 917-928, 2011.

[5] R. Cohen, I. Dinstein, J. El-Sana, and K. Kedem, "Using scale-space anisotropic smoothing for text line extraction in historical documents", ICIAR, pp. 349-358, 2014.

[6] J. Pastor-Pellicer, M. Z. Afzal, M. Liwicki, and M. J. Castro-Bleda, "Complete system for text line extraction using convolutional neural networks and watershed transform", DAS, pp. 30-35, 2016.

[7] G. Renton, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet, "Handwritten text line segmentation using fully convolutional network", ICDAR, pp. 5-9, 2017.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", IEEE, vol. 86(11), pp. 2278-2323, 1998.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", CVPR, pp. 3431-3440, 2015.

[10] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation", CVPR, pp. 1520-1528, 2015.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation", MICCAI, pp. 234-241, 2015.

[12] A. Belaïd and N. Ouwayed, "Segmentation of Ancient Arabic Documents", In Guide to OCR for Arabic Scripts, Märgner, Volker and El Abed, Haikal editor, Springer London publisher, pp. 103-122, 2012.

[13] $https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5$

[14] $https://github.com/LeeJunHyun/Image\_Segmentation$

[15] G. Renton, Y. Soullard, C. Chatelain, S. Adam, C. Kermorvant, and T. Paquet, "Fully convolutional network with dilated convolutions for handwritten text line segmentation", IJDAR, vol. 21, no. 3, pp. 177-186, 2018.

[16] S. A. Oliveira, B. Seguin, and F. Kaplan, "dhSegment: A generic deeplearning approach for document segmentation", ICFHR, pp. 7-12, 2018.

[17] B. Kurar Barakat, A. Droby, M. Kassis, and J. El-Sana, "Text line segmentation for challenging handwritten document images using fully convolutional network", ICFHR, pp. 374-379, 2018.

[18] O. Mechi, M. Mehri, R. Ingold, and N. E. B. Amara, "Text line segmentation in historical document images using an adaptive U-Net architecture", ICDAR, pp. 369-374, 2019.

[19] C. Neche, A. Belaïd, and A. Kacem Echi, "Arabic handwritten documents segmentation into text-lines and words using deep learning", ICDAR, pp. 19-24, 2019.

[20] T. Grüning, G. Leifert, T. Strauss and R. Labahn, "A Two-Stage Method for Text Line Detection in Historical Documents", 2018.

[21] J. Zhang, J. Du, H. Liu, X. Hou, Y. Zhao, and M. Ding, "LU-NET: An Improved U-Net for Ventricular Segmentation", IEEE Access, vol. 7, pp. 92539-92546, 2019.

[22] M. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation", CVPR, 2018.

[23] H. Sun, Ch. Li, B.Liu, Z. Liu, J. Luo, H. Zheng, D. D. Feng, S. Wang,"AU-Net: Attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms", CVPR, 2019.

[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", ICLR, pp. 1-15, 2015.

[25] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, "READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents", DAS, pp. 351-356, 2018.