A One-Shot Learning Approach to Document Layout Segmentation of Ancient Arabic Manuscripts

Axel De Nardin[*,1,2] , Silvia Zottin[*,1] , Claudio Piciarelli[1] , Emanuela Colombi[1] , Gian Luca Foresti[1]

[1]University of Udine, Udine, Italy
[2]University of Trieste, Trieste, Italy
{denardin.axel, zottin.silvia}@spes.uniud.it
{claudio.piciarelli, emanuela.colombi, gianluca.foresti}@uniud.it

Abstract

Document layout segmentation is a challenging task due to the variability and complexity of document layouts. Ancient manuscripts in particular are often damaged by age, have very irregular layouts, and are characterized by progressive editing from different authors over a large time window. All these factors make the semantic segmentation process of specific areas, such as main text and side text, very difficult. However, the study of these manuscripts turns out to be fundamental for historians and humanists, so much so that in recent years the demand for machine learning approaches aimed at simplifying the extraction of information from these documents has consistently increased, leading document layout analysis to become an increasingly important research area. In order for machine learning techniques to be applied effectively to this task, however, a large amount of correctly and precisely labeled images is required for their training. This is obviously a limitation for this field of research as ground truth must be precisely and manually crafted by expert humanists, making it a very time-consuming process. In this paper, with the aim of overcoming this limitation, we present an efficient document layout segmentation framework, which while being trained on only one labeled page per manuscript still achieves state-of-the-art performance compared to other popular approaches trained on all the available data when tested on a challenging dataset of ancient Arabic manuscripts.

1. Introduction

As the amount of historical documents archived in digital libraries around the world keeps increasing over the years, to preserve their content, the role of systems aiming at the analysis of this content gets more and more important. The main motivation behind the adoption of these systems is that the raw format in which the documents are stored in their digital form makes it very time expensive to analyze their contents manually, therefore a necessary step to simplify this process is represented by the automated transcription of the documents. To perform this task, however, a preliminary step is necessary. This step is represented by the binarization, or even better, the layout analysis of the pages of the documents [5]. Document Layout Analysis (DLA) includes several fundamental activities for document image processing that allow for simplifying other more complex related tasks such as optical character recognition [23], automatic text transcription [16] and writer identification [24]. In particular, a central task of DLA is page segmentation which consists of the subdivision of a given document image into semantically meaningful regions (*e.g.* main text, side text, and background).

Over the years, algorithms and methods dealing with DLA have become increasingly popular, especially when it comes to ancient and handwritten documents, which pose an extra challenge compared to printed ones [30]. The reason behind this added complexity is that the manuscripts have complex and very variable layouts, with different and uneven writing styles throughout the book. Furthermore, in many manuscripts, around the main text, there are notes and paratexts arranged marginally, with different orientations, writing styles, and font sizes, but often similar to the main text, which makes feature extraction and semantic segmentation very difficult. As a further complication, ancient documents suffer from different levels of degradation due to aging and bad conservation over time, such as scratches and holes in the page, ink stains and text fading, noise, and bleed-through.

[*]These authors contributed equally to this work

Typically, to address semantic segmentation problems, Ground Truth maps (GT) are needed, to train the segmentation models and test their performances. When it comes to ancient manuscripts, obtaining these GTs turns out to be very difficult as the segmentation must be done by an expert in the document domain area, who knows how the classification of the different regions of the document pages must be performed. Moreover, the creation of the labeled data must be pixel-precise which is a very time-consuming process [18].

In this paper, we present a one-shot document layout segmentation approach, which addresses the aforementioned issues. In particular, we apply a popular document layout analysis framework to the layout analysis of ancient Arabic manuscripts in a one-shot learning setting, where we rely on only one GT image per manuscript to train the model.

The rest of the paper is structured as follows: Section 2 reviews the state-of-the-art of document layout segmentation, especially those using the same dataset used for our method; Section 3 describes the proposed one-shot learning approach; the experimental results and evaluations are drawn in Section 4, where we provide a thorough comparison between our approach and the current state-of-the-art; finally, in Section 5 the conclusions and the future works are provided.

2. Related Works

To deal with the layout analysis of handwritten historical documents, various approaches have been used, which differ from each other, among other aspects, in terms of the amount of labeled data involved in the training of the respective models.

The supervised approaches solve the layout segmentation task by requiring a large number of labeled data for training. These techniques typically represent the state-of-the-art for the task at hand and address this problem both with classical computer vision methods [3, 9, 17, 19, 22] and by relying on deep learning models [1, 2, 6, 8, 29]. These approaches have very high performance, but as previously discussed they have a high demand for labeled data which are often not available because, as previously mentioned, creating them requires the involvement of a domain expert and a considerable amount of time to segment the images.

Conversely, while unsupervised approaches do not require any labeled data for the training, they are very rare in the context of DLA [7, 15, 28], the main reasons behind this are connected to the complexity and variability of ancient manuscripts' layout and also to the fact that typically, when working with documents of any kind, the different layout classes tend to present very similar characteristics, as they are all typically represented by text in different formats, making it very hard to consistently and correctly identifying them without any external supervision.

To overcome these problems, few-shot solutions have been proposed, which involve the limited use of labeled data for the segmentation task. Classical transfer learning approaches where a feature extraction module pre-trained on a large general-purpose model have been studied to address the lack of annotated training data, as in [27] where a network pre-trained on the popular ImageNet dataset is then fine-tuned on the document images, however, what emerged from this study is that the feature extracted from ImageNet don't seem to be effectively applicable for the downstream task of document layout segmentation. A more recent approach, proposed in [13], combined a robust segmentation network with effective data augmentation and segmentation refinement strategies that allowed to reach state-of-the-art performance for the layout segmentation task on the popular Diva-HisDB [26] dataset while relying on only two labeled instances to train the model.

2.1. Ancient Arabic Manuscripts Layout Analysis

The papers addressing the document layout segmentation task on the Bukhari et al. [6] dataset are very limited. In [6], in addition to presenting the dataset, they propose an approach based on a combination of feature extraction from the connected components and a supervised multilayer perception classifier to define the segmentation class. In [3] instead, a learning-free method was proposed to detect the main text area in ancient manuscripts. First, the main text area is coarsely segmented using a texture-based filter and then, as an energy minimization task, it is refined. In [4] a Fully Convolutional Network is trained with the aim of segmenting into main text and side text regions by dense pixel prediction. Instead in [1], a Siamese neural network is used, trained on different patches in order to be able to calculate their similarity and the distance matrix, in such a way as to segment each page of the manuscripts into the main text and side text classes. A similar approach was proposed in [15], where the Siamese neural network is trained to differentiate between patches using their properties such as number of foreground pixels, and average component height and width. Then a principal component analysis of the feature map is applied to segment the page into main and side text regions.

Differently in our work, we propose a one-shot solution for document layout segmentation, which involves only one labeled image per manuscript to train the model, with better performance than the state-of-the-art approaches for the same dataset.

3. Method

The approach proposed in this paper is inspired by the recent works on few-shot document layout analysis described in [13, 14]. The framework consists of 3 main components:

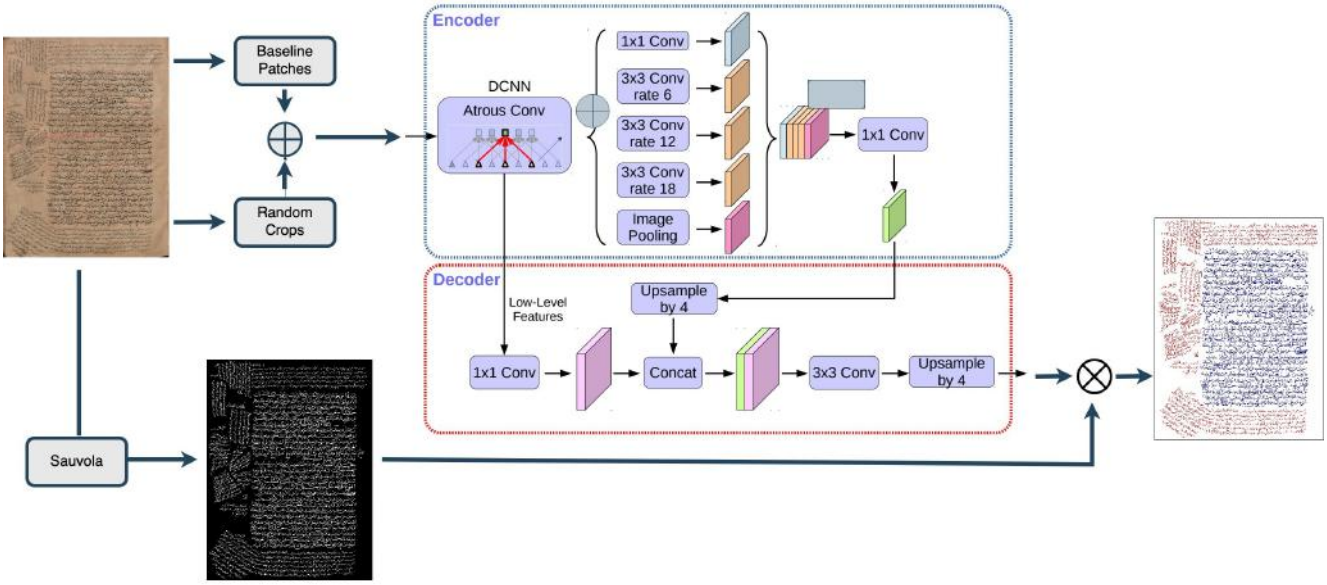


Figure 1. Visual representation of the segmentation pipeline characterizing the adopted framework. From the input image are extracted two sets of patches, the baseline patches are extracted by drawing a grid over the input image while the random crops, as the name suggests, are patches that are extracted randomly from various locations inside the image. At training time both sets are fed into the backbone model which tries to predict the corresponding segmentation maps. At inference time only the baseline patches are used, furthermore, the obtained segmentation maps are refined via the application of the Sauvola thresholding algorithm.

a semantic segmentation backbone, a dynamic instance generation module, and a segmentation refinement module.

A visual representation of the framework is provided in Fig. 1.

3.1. Backbone

Choosing an effective and efficient segmentation backbone is a key step in providing accurate segmentation predictions for the task at hand, especially when working in a few-shot or one-shot setting. Differently from [13], which relies on the DeepLabv3 [10] network, we opted for the more recent DeepLabv3+ [11] architecture as the backbone of choice for the present work. There are two main differences between the two architectures. The first one is represented by the model used for the encoder model. While the former adopts a traditional ResNet [20], the latter relies on a more sophisticated Aligned XCeption [12] network in which all the max pooling layers have been replaced by depth-wise convolutional blocks. The second difference is represented by the introduction of a simple decoder module in the DeepLabv3+ pipeline, which proved to be an effective way to increase the segmentation performance of the network by adding only a small computational overhead. It takes the high-level features from the encoder and refines them through upsampling and convolutional operations to capture fine-grained details, leading to better boundary delineation and more accurate segmentation masks.

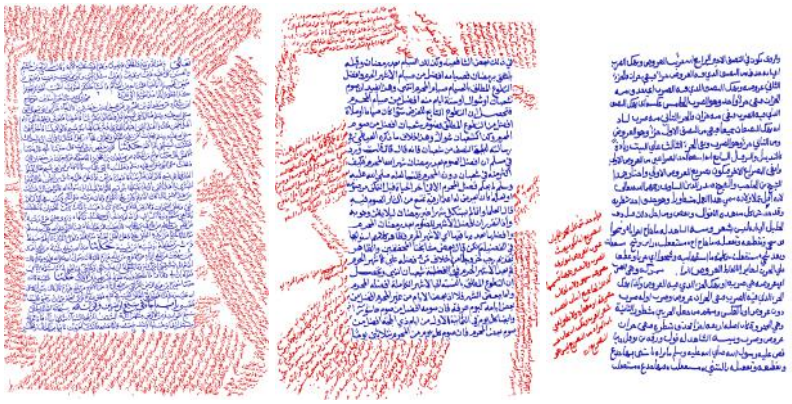
3.2. Dynamic Instance Generation

When working in a low data setting, and especially in a one-shot learning one, the adoption of effective data augmentation strategies is of uttermost importance to be able to fully leverage the small amount of information available. The Dynamic Instance Generation module serves exactly this purpose by generating a set of sub-instances from the input instances at each training epoch. This process is carried out in two steps. The first one involves splitting the input image into a set of n non-overlapping patches that cover its full surface. These patches were kept consistent across all training epochs and also between the training and testing phases. In the second step, introduced to further enhance the model's generalization capabilities, a small set of k crops is extracted at random locations around the image. The key idea behind this approach is that, in the context of document layout analysis, given patches that are large enough, the structure of their content is essentially representative of the overall structure of the whole page. This allows for the effective capture of features describing its different layout components.

3.3. Segmentation Refinement

The final component of the adopted framework is a segmentation refinement approach based on Sauvola Thresholding [25], a traditional image binarization approach. The Sauvola Algorithm performs image binarization by calcu-

(a) Book 1, page 19 (b) Book 2, page 1 (c) Book 3, page 2



(d) Book 1 ground truth (e) Book 2 ground truth (f) Book 3 ground truth

Figure 2. The three pages (2a– 2c), one for each manuscript, and the corresponding ground truth masks (2d– 2f) selected from the Bukhari et al. [6] dataset, used for our training set. The areas highlighted in blue and red represent the main and side text components respectively, while the white areas correspond to the background.

lating local thresholding on a set of small windows that make up the original image. It is an adaptive thresholding technique, meaning that the threshold value is computed locally for each pixel in the image based on its local neighborhood.

The equation used to calculate the threshold value is the following:

$$T = \mu(N) \times \left(1 + k \times (\frac{\sigma(N)}{R} - 1)\right) \qquad (1)$$

where N is the local window of size $n \times n$, $\mu(N)$ and $\sigma(N)$ are, respectively, the corresponding mean and standard deviation, and R is the dynamic range of the standard deviation. Finally, k is a manually selected parameter that regulates the value of the local threshold. More specifically, a larger value of k leads to a lower threshold for the local window which in turn results in missed part of text from the document image, on the contrary, a small k usually leads to blurry and noisy segmentations.

4. Experimental Results

In this section we provide a detailed description of the training and testing phase of our approach, describing the dataset used and the evaluation metrics adopted. We compared our performance with the state-of-the-art results for the same dataset.

4.1. Dataset

The dataset we selected to test our approach in this paper is the one presented by Bukhari et al. [6], which represents the most popular one for the task of document layout segmentation on historical Arabic manuscripts. It consists of 32 images each representing a page from one of three different Arabic historical manuscripts. Out of all the samples 24 are typically used for the training process while the remaining 8 are used for the testing. In the present work, however, we relied on just one image from each manuscript to train our model, bringing the total size of the training set to just 3 images. Furthermore, 4 images were used as our validation set. Finally, the test set was kept consistent with the other works involving the use of the dataset to provide

a fair comparison with their approaches. In this dataset, the semantic segmentation classes are: main text, side text, and background. The instances selected for the training set are shown in Fig. 2.

4.2. Metrics

The performance of our method is evaluated by calculating the F-score. This metric, also called the F1 score, is the harmonic mean of Precision and Recall. In particular, Precision (Eq. 2) measures the proportion of True Positive (TP) predictions among all positive predictions made by the model, then TP and False Positive (FP). Recall (Eq. 3), on the other hand, measures the proportion of TP predictions among all actual positive instances in the data, then TP and False Negative (FN).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F-score combines Precision and Recall into a single score using the following equation:

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F-score ranges between 0 and 1, where a score of 1 indicates perfect precision and recall, and a score of 0 indicates poor performance. The F-score of both the main text and the side text is presented separately, consistently with the previous works performed on this dataset, by keeping only the pixels of the class of interest and setting all the pixels of the other class as background. Furthermore, the overall average of the model's performance is also provided.

4.3. Training setup

The backbone is trained with an early stopping mechanism to prevent over-fitting. In this approach, the model is trained for a maximum of 200 epochs and, at each epoch, the validation loss is monitored. If the validation loss does not improve over the last 20 consecutive epochs, the training process is stopped, even if the maximum number of epochs has not been reached. To prevent the model from converging too quickly to a sub-optimal solution, a buffer of 50 epochs has been added. The ADAM optimizer [21] was adopted for the network training process with a learning rate of 10^{-3} and a weight decay of 10^{-5}.

Due to the imbalanced segmentation classes in the dataset, as a loss function, we used a combination of Jaccard loss and Dice loss. Jaccard loss measures the dissimilarity between the predicted segmentation mask and the GT mask, based on the intersection and union of the two masks. The equation of Jaccard loss is the following:

$$\mathcal{L}_{\text{Jaccard}} = 1 - \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

	Main Text	Side Text	Average
Bukhari et al. [6]	0.950	0.950	0.950
Barakat and El-Sana [4]	0.950	0.800	0.875
Alasaam et al. [1]	0.986	0.969	0.978
Droby et al. [15]	0.986	0.970	0.978
Asi et al. [3]	**0.992**	0.985	0.988
Ours	0.989	**0.991**	**0.990**

Table 1. Comparison between the F-score of our model and the state-of-the-art for Bukhari et al. [6] dataset. The best-performing model is reported in bold.

where TP, FP, and FN stand respectively for True Positives, False positives and False Negatives. Also, Dice loss measures the dissimilarity between the predicted segmentation mask and the GT mask but is based on their overlap. The equation of Dice loss is the following:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

In our segmentation problem, Dice loss was generalized to handle multi-class segmentation problems by calculating the Dice loss separately for each class and then taking the average over all classes.

In general, Dice loss is more sensitive to small differences between the predicted and GT masks, while Jaccard loss is more sensitive to larger differences. By combining the two loss functions, the model can learn to produce more accurate and robust segmentation masks. The loss adopted to train our network is therefore defined as:

$$\mathcal{L} = \mathcal{L}_{\text{Jaccard}} + \mathcal{L}_{\text{Dice}} \tag{7}$$

Before the network training, the original high-resolution input image is resized to the fixed shape of 1344×2016 px. During each epoch the images of the training set are split into patches of size 672×672 px and the dynamic instance generation module increases the number of patches with 12 random crops of the same size.

Finally, for the segmentation refinement module and, in particular, for the Sauvola Thresholding a window size of 31×31 px and a k value equal to 0.2 was chosen.

4.4. Results

We compared the performance of our approach with the current state-of-the-art models for document layout segmentation on the Bukhari et al. [6] dataset. In Table 1 we provide all the quantitative results, which were taken from the respective papers, except for some overall averages of F-scores that we calculated considering the performance for the main text and side text. Our approach, even though it was trained on just one image for each book in the dataset, outperformed all the competition on the side text, achieving

(a) Book 1 page 1, Original (b) Ground Truth (c) Asi et al. [3] (d) Ours

(e) Book 1 page 18, Original (f) Ground Truth (g) Barakat and El-Sana [4] (h) Ours

(i) Book 3 page 1, Original (j) Ground Truth (k) Barakat and El-Sana [4] (l) Ours

Figure 3. A qualitative comparison between our framework and the competition one. Each row represents a different instance of the Bukhari et al. [6] dataset. In the first column, the original images are shown, while in the remaining columns we provide the ground truth segmentation maps and the results produced by Asi et al. [3], Barakat and El-Sana [4] and Ours.

a 0.991 F-score, which represents a 0.6% improvement over the previous state-of-the-art, while for the main text class, it achieved the second best result with a difference of only 0.3% compared to the best one being represented by Asi et al. [3].

In general, the proposed framework presents the overall best performance compared to the others, achieving an improvement of over 1% F1 score for the average performance compared to all the competition approaches, with the exception of the previous state-of-the-art over which it improved the results by 0.2%. While the reported improvement over the previous state-of-the-art approach may seem very marginal the key contribution of our work is represented by the one-shot setting in which our model was trained. It is important to keep in mind that all the competition models were trained on approximately 8 times the amount of data used for our experiments and our approach still managed to outperform all of them. Furthermore, in Table 2 we show the F-scores for each individual sample contained in the test set, compared to Barakat and El-Sana [4] results (except the ones for the 5th sample which were not provided in the original paper). As we can see, The pro-

Sample	Main Text		Side Text		#Book	#Train Samples	
	[4]	Our	[4]	Our		[4]	Our
1	**0.990**	0.989	0.980	**0.993**	1	20	1
2	0.990	**0.991**	0.980	**0.993**	1	20	1
3	**1.00**	0.987	**1.00**	0.996	1	20	1
4	**1.00**	0.991	**1.00**	0.988	1	20	1
5	-	**0.989**	-	**0.992**	1	-	1
6	**0.990**	**0.990**	0.950	**0.993**	1	20	1
7	0.850	**0.989**	0.100	**0.978**	2	3	1
8	0.830	**0.985**	0.510	**0.993**	3	1	1

Table 2. F-scores of our approach and that presented in Barakat and El-Sana [4], on each test sample and number of train samples from the corresponding book of the test sample (in bold the best performing approach).

Book	Main Text			Side Text		
	Baseline	Base+Crops	Full Framework	Baseline	Base+Crops	Full Framework
1	0.827	0.950	**0.990**	0.932	0.956	**0.993**
2	0.787	0.966	**0.989**	0.916	0.945	**0.978**
3	0.897	0.965	**0.985**	0.932	0.981	**0.993**
Full Dataset	0.873	0.953	**0.989**	0.933	0.958	**0.991**

Table 3. F-score reached by the three versions of our framework on the Bukhari dataset. Namely, the baseline without either the dynamic crop generation or the segmentation refinement module. The one relying only on the former and finally the full framework. The results are presented subdivided both for the three books and for the entire dataset.

posed method achieves results comparable to Barakat and El-Sana on the first 6 test samples and substantially outperforms it in the 7th and 8th test samples, especially for the side text class where the competition approach achieves very poor performance, likely due to the reduced amount of training samples available (respectively 2nd and 3rd book to which these samples belong). In general, we can observe how our approach can fully leverage the small amount of data available (just one sample for each document class), achieving high and consistent performance across all the images present in the test set, clearly showing that it's well-defined and robust.

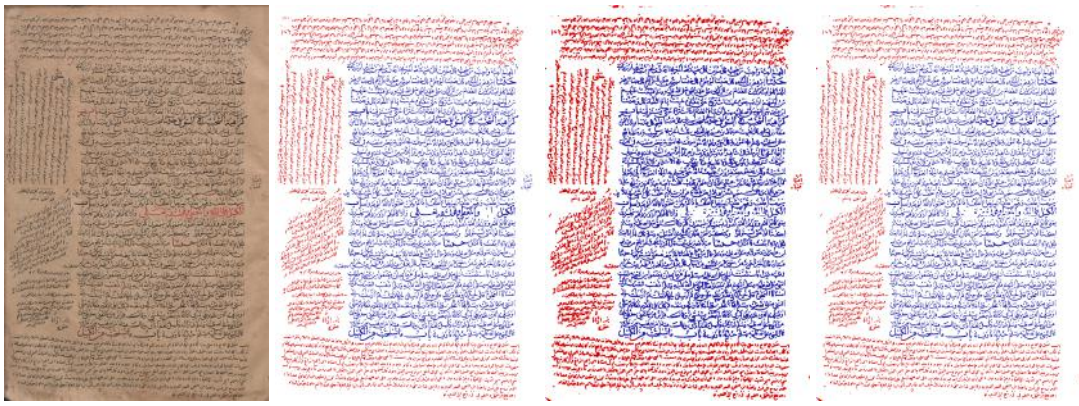
Finally, in Figure 3 we provide the qualitative results achieved by the proposed framework on three document pages from the Bukhari et al. [6] dataset. The respective segmentation maps produced by our model are compared against the Ground Truth masks and the results obtained in Asi et al. [3] and Barakat and El-Sana [4]. The main text components are highlighted in blue, the side text components are highlighted in red and the background is white. As we can see our approach compares favourably to both the competition models. Compared to the Barakat and El-Sana it manages to achieve very good performance even on the most challenging instance represented by Fig.3(i), due to the low amount of training data available for Book 3 of the dataset. Furthermore, it still manages to provide a more accurate segmentation than both competition methods even when the latter are trained on a much larger amount of data, such as the one available for Book 1. In fact, we can observe how the proposed framework manages to correct most of the mistakes introduced by both the Asi et al. [3] and Barakat and El-Sana [4] approaches in the side-text regions.

	Inference Time (per instance)
Coarse segmentation	0.41s
Segmentation refinement	0.01s
Total inference time	0.42s

Table 4. Processing times of the two main components of the adopted framework. Measurements are indicated in seconds.

4.5. Ablation study

To conclude, in this section, we provide the results of the ablation study we conducted to validate the improvements introduced by the dynamic instance generation module and the segmentation refinement module in the context of layout segmentation for ancient Arabic manuscripts. In Table 3 we report the results obtained by the base version of the framework, relying only on the baseline patches, with those obtained by the version including the dynamic crops but not the segmentation refinement process, and with the full framework. Specifically, we show the performance achieved by the three versions of the framework, in terms of F1-score, on the individual books composing the dataset as well as on the dataset in its entirety.



(a) Original (b) Ground Truth (c) Coarse segmentation (d) Refined Segmentation

Figure 4. Qualitative comparison between the segmentation maps produced by our framework with and without the introduction of the segmentation refinement process. Fig. 4a and Fig. 4b show the original picture and the corresponding Ground Truth respectively. Fig. 4c shows the coarse segmentation mask obtained from our model without using the refinement module. Fig. 4d shows the result of segmentation prediction obtained with our full framework.

As we can see for all three subsets as well as for the full dataset there is a consistent and substantial improvement in performance with the introduction of each module of our framework. This is particularly evident for the main text class where we can observe an improvement ranging from around 7% for book number 3, to an impressive 18% for book 2 with just the introduction of the dynamic crop generation module. With a total improvement of 11.7% across the dataset when going from the baseline version to the full framework. On the side text class the improvement is not as large as for the main text class but the full framework still consistently achieves a 6% increase in F1-score across all the books, as well as on the full dataset, over the baseline approach.

Finally, in Figure 4 we report a sample qualitative result for one of the instances present in the test portion of the adopted dataset. In particular, we compare the coarse and refined segmentation predictions provided by our approach with the ground truth mask provided for the corresponding instance of the test set, where the red and blue portions of the image represent the side and main text of the page respectively. The coarse segmentation is achieved by our approach without using the refinement module described above. As we can observe while the coarse output of the network correctly identifies all the layout components of the selected instance, it partially lacks precision, providing masks that typically extend into the background portion of the image. On the other hand, when the refinement process is applied, the final segmentation prediction closely resembles the one provided by the ground truth masks.

For completeness, as in [3], we calculated the inference times of our framework when executed on a consumer-grade GPU, namely the NVidia GeForce RTX 3090, and using Python's built-in profiler. In particular, we report in Table 4, the execution times for both the coarse segmentation and the segmentation refinement modules of our framework individually, as well as the total time needed to perform the end-to-end inference process leading to the final segmentation masks. As we can observe, the segmentation refinement step introduces very little overhead in the inference process.

5. Conclusion and Future Works

In this paper, we presented a low-data semantic segmentation framework capable of achieving state-of-the-art results on a popular dataset for historical Arabic document layout segmentation. We have shown that, even while being trained in a one-shot setting, our framework achieves consistently better results on the task at hand compared to previous methods trained on the full available dataset.

A downside of the proposed approach is represented by the need to manually set the hyper-parameters for the Sauvola Thresholding Algorithm used in our segmentation refinement module, which has a noticeable impact on the performance of the segmentation task. In future works, we plan to address this problem by automatizing the parameter selection process.

Aknowledgments

Partial financial support was received from Piano Nazionale di Ripresa e Resilienza DD 3277 del 30 dicembre 2021 (PNRR Missione 4, Componente 2, Investimento 1.5) - Interconnected Nord-Est Innovation Ecosystem. Partial financial support was received from Strategic Departmental Plan on Artificial Intelligence, Department of Mathematics, Computer Science and Physics, University of Udine.

References

[1] Reem Alaasam, Berat Kurar, and Jihad El-Sana. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 738–742, 2019. 2, 5

[2] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12, 2018. 2

[3] Abedelkadir Asi, Rafi Cohen, Klara Kedem, Jihad El-Sana, and Itshak Dinstein. A coarse-to-fine approach for layout analysis of ancient manuscripts. In *2014 14th International Cofnference on Frontiers in Handwriting Recognition*, pages 140–145, 2014. 2, 5, 6, 7, 8

[4] Berat Kurar Barakat and Jihad El-Sana. Binarization free layout analysis for arabic historical documents using fully convolutional networks. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 151–155, 2018. 2, 5, 6, 7

[5] Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), oct 2019. 1

[6] Syed Saqib Bukhari, Thomas M. Breuel, Abedelkadir Asi, and Jihad El-Sana. Layout analysis for arabic historical document images using machine learning. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 639–644, 2012. 2, 4, 5, 6, 7

[7] Kai Chen, Cheng-Lin Liu, Mathias Seuret, Marcus Liwicki, Jean Hennebert, and Rolf Ingold. Page segmentation for historical document images based on superpixel classification with unsupervised feature learning. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 299–304, 2016. 2

[8] Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 965–970, 2017. 2

[9] Kai Chen, Hao Wei, Jean Hennebert, Rolf Ingold, and Marcus Liwicki. Page segmentation for historical handwritten document images using color and texture features. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 488–493, 2014. 2

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 3

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing. 3

[12] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 3

[13] Axel De Nardin, Silvia Zottin, Matteo Paier, Gian Luca Foresti, Emanuela Colombi, and Claudio Piciarelli. Efficient few-shot learning for pixel-precise handwritten document layout analysis. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3669–3677, 2023. 2, 3

[14] Axel De Nardin, Silvia Zottin, Claudio Piciarelli, Emanuela Colombi, and Gian Luca Foresti. Few-shot pixel-precise document layout segmentation via dynamic instance generation and local thresholding. *International Journal of Neural Systems*, 33(10):2350052, 2023. PMID: 37567858. 2

[15] Ahmad Droby, Berat Kurar Barakat, Borak Madi, Reem Alaasam, and Jihad El-Sana. Unsupervised deep learning for handwritten page segmentation. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 240–245, 2020. 2, 5

[16] Andreas Fischer, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. Automatic transcription of handwritten medieval documents. In *2009 15th International Conference on Virtual Systems and Multimedia*, pages 137–142, 2009. 1

[17] Angelika Garz, Robert Sablatnig, and Markus Diem. Layout analysis for historical manuscripts using sift features. In *2011 International Conference on Document Analysis and Recognition*, pages 508–512, 2011. 2

[18] Angelika Garz, Mathias Seuret, Fotini Simistira, Andreas Fischer, and Rolf Ingold. Creating ground truth for historical manuscripts with document graphs and scribbling interaction. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 126–131, 2016. 2

[19] Costantino Grana, Daniele Borghesani, and Rita Cucchiara. Automatic segmentation of digitalized historical manuscripts. *Multimedia Tools and Applications*, 55(3):483–506, Dec 2011. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[22] Maroua Mehri, Nibal Nayef, Pierre Héroux, Petra Gomez-Krämer, and Rémy Mullot. Learning texture features for enhancement and segmentation of historical document images. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pages 47–54, 2015. 2

[23] Sonika Rani Narang, M. K. Jindal, and Munish Kumar. Ancient text recognition: a review. *Artificial Intelligence Review*, 53(8):5517–5558, Dec 2020. 1

[24] Karl Ni, Patrick Callier, and Bradley Hatch. Writer identification in noisy handwritten documents. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1177–1186, 2017. 1

[25] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000. 3

[26] Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476, 2016. 2

[27] Linda Studer, Michele Alberti, Vinaychandran Pondenkandath, Pinar Goktepe, Thomas Kolonko, Andreas Fischer, Marcus Liwicki, and Rolf Ingold. A comprehensive study of imagenet pre-training for historical document image analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 720–725, 2019. 2

[28] Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1220–1224, 2013. 2

[29] Yue Xu, Fei Yin, Zhaoxiang Zhang, and Cheng-Lin Liu. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *IJCAI*, pages 1057–1063, 2018. 2

[30] Chunhu Zhang, Mayire Ibrayim, and Askar Hamdulla. A methodological study of document layout analysis. In *2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI)*, pages 12–17, 2022. 1