

# Dynamic instance generation for few-shot handwritten document layout segmentation (short paper)

Axel De Nardin<sup>1</sup>, Silvia Zottin<sup>1</sup>, Matteo Paier<sup>1</sup>, Gian Luca Foresti<sup>1</sup>,  
Emanuela Colombi<sup>2</sup> and Claudio Piciarelli<sup>1</sup>

<sup>1</sup>*Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy*

<sup>2</sup>*Department of Humanities and Cultural Heritage, University of Udine, Udine, Italy*

## Abstract

Historical handwritten document analysis is an important activity to retrieve information about our past. Given that this type of process is slow and time-consuming, the humanities community is searching for new techniques that could aid them in this activity. Document layout analysis is a branch of machine learning that aims to extract semantic informations from digitised documents. Here we propose a new framework for handwritten document layout analysis that differentiates from the current state-of-the-art by the fact that it features few-shot learning, thus allowing for good results with little manually labelled data and the dynamic instance generation process. Our results were obtained using the DIVA - HisDB dataset.

## Keywords

Few-shot learning, Handwritten document layout analysis, Fully-Convolutional Network, Document image segmentation

## 1. Introduction

In the humanities community the study of historical handwritten documents is a crucial activity [1]. For centuries humanists have focused only on text without considering the elements that accompanied it as comments and decoration, in general called paratext. In the latter years, paratext analysis has gained more and more relevance: this data is fundamental to the cultural-historical understanding of the individual manuscript but is also of great philological relevance, because paratexts can be transcribed from one manuscript to another [2].

Despite this importance there are currently not so many studies about them, mainly because paratext analysis is a difficult and expensive task. So, newly developed tools and methods are sought: automated paratext extraction not only saves large amounts of time, but also enables

---

*1st Italian Workshop on Artificial Intelligence for Cultural Heritage (AI4CH22), co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIXIA 2022). 28 November 2022, Udine, Italy.*

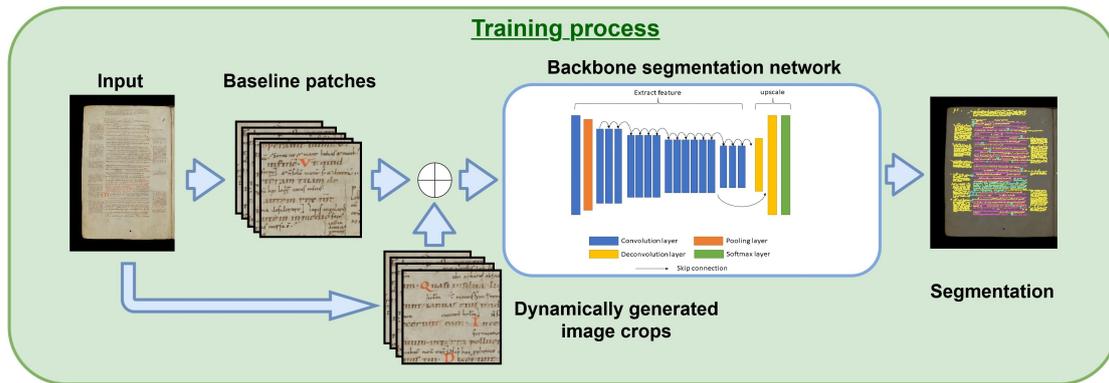
✉ denardin.axel@spes.uniud.it (A. De Nardin); zottin.silvia@spes.uniud.it (S. Zottin); paier.matteo@spes.uniud.it (M. Paier); gianluca.foresti@uniud.it (G. L. Foresti); emanuela.colombi@uniud.it (E. Colombi); claudio.piciarelli@uniud.it (C. Piciarelli)

🆔 0000-0002-0762-708X (A. De Nardin); 0000-0003-0820-7260 (S. Zottin); 0000-0002-8425-6892 (G. L. Foresti); 0000-0002-0384-6664 (E. Colombi); 0000-0001-5305-1520 (C. Piciarelli)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Visual representation of the proposed segmentation framework.

comparisons of the extracted data in rapid time, making it possible to establish connections between seemingly distant manuscripts, connections that escape the human eye and memory.

In order to achieve this we must start from page segmentation of a given document image into semantically meaningful regions (e.g. main text, comments, decorations and background), that is the main focus of this paper.

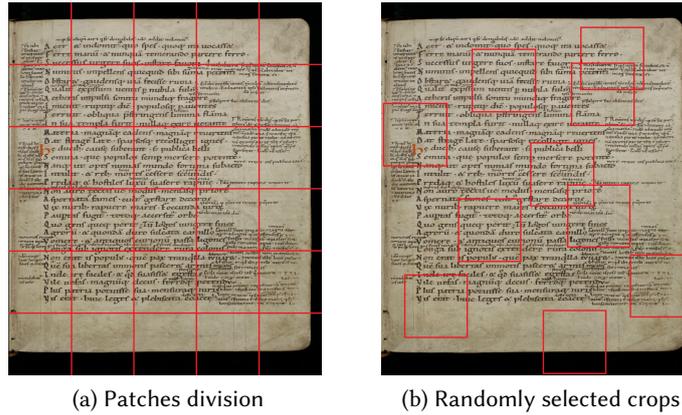
Page segmentation is a well-known open problem in the machine learning community. Due to the non-uniformity and integrity of the images, many of the approaches adopted to solve this problem rely on a fully supervised learning paradigm [3, 4, 5]. An exhaustive survey on document layout analysis is in Binmakhshen and Mahmoud [6].

In contrast, our paradigm tries to achieve few-shot learning, given that in real world applications usually the ground truth is limited in size [7]. To overcome this limit we introduce a dynamic instance generation process that allow to improve the limited data available in this scenario. By combining this with a fully-convolutional network we are able to achieve greater performances adding image patching. We used DIVA-HisDB to test the proposed framework [8].

The rest of this paper is organized as follows. Section 2 describes the components defining the proposed framework. Section 3 reports the details of our experimental setup. Finally, in Section 4, are drawn the conclusions and discuss the future work.

## 2. Proposed Method

In this section, we present the details of the proposed framework with a brief description of the key components. First, we introduce a Fully-Convolutional Network model with a ResNet-50 backbone. This is used for a segmentation tasks in our framework. Moreover we present our training process characterized by the dynamic instance generation. In Fig. 1 it is presented a visual representation of our framework.



**Figure 2:** An example of the dynamically enhanced training data. In (2a) a page of the CSG863 manuscript class is divided into non-overlapping, fixed-size, patches that cover the entire input image. In (2b) 10 crops with the same size of the patches are randomly selected in the same page to improve the training set.

## 2.1. Backbone network

The main component of our framework is a Fully-Convolutional Network (FCN) model [9], a ResNet-based deep model that combines layers of the feature hierarchy and refines the spatial precision of the output. This architecture is largely adopted in the context of image semantic segmentation.

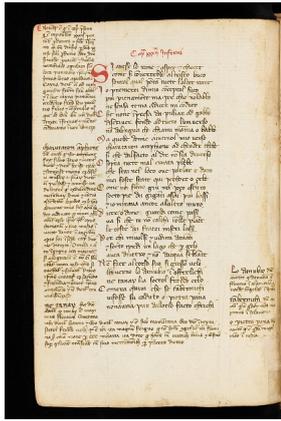
The network is composed of a downsampling and upsampling path. The first is used to extract and interpret the context, instead second one allows the localization. The network is able to combine coarse, high layer information with fine, low layer information. Furthermore, the multilayer outputs are followed by deconvolutional layers for bilinear upsampling to pixel-dense outputs. This allows for pixel-level identification of class labels and predict segmentation masks.

The FCN-based methods learn a mapping from pixels to pixels, without extracting the region proposals. This architecture employs solely locally connected layers (such as convolution, pooling and upsampling) and avoids the use of dense layers. This means that requires less parameters and then making the network faster to train and enables to make predictions on inputs of variable sizes.

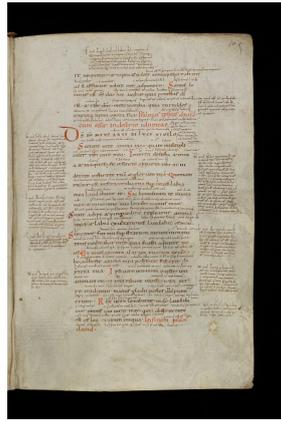
## 2.2. Dynamically enhanced training data

In this paper we present a few-shot learning system and the key element of this process is the maximization of the exploitation of the available data.

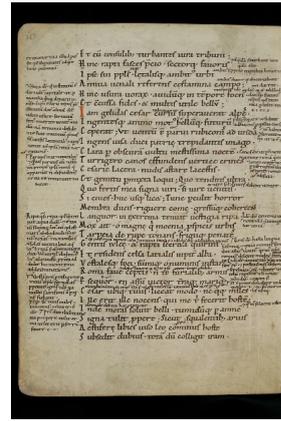
Usually, to capture global contextual information about images the model is trained on whole images, however we believe that a great extent of the same information can also be retrieved from smaller sections of the document pages. For this reason, to improve the efficiency of our training setup reducing the number of annotated images, we decided to split each page of the manuscript in a set  $P$  of non-overlapping, fixed-size, patches that cover the entire input image.



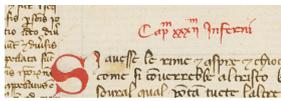
(a) CB55 page



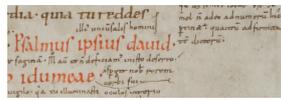
(b) CSG18 page



(c) CSG863 page



(d) CB55 detail



(e) CSG18 detail



(f) CSG863 detail

**Figure 3:** Samples from the 3 manuscripts (CB55, CSG18 and CSG863) presents in DIVA-HisDB [8]. Fig. 3a– 3c show a full page for each manuscripts, while Fig. 3d– 3f show a detail extracted from each of them.

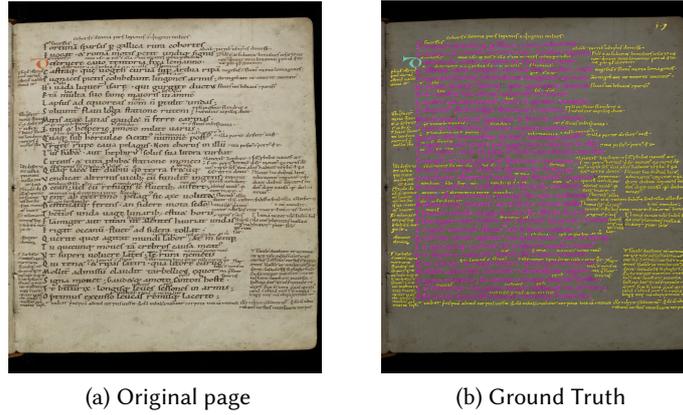
All of these patches forms the base training set.

The cardinality of  $P$  cannot raise indefinitely: the size of the single patches must be large enough to allow capturing contextual information from the corresponding represented area of the original image. To overcome this limitation, we introduce a dynamic instance generation process. At each epoch we retrieve a set  $C$  of randomly selected crops of the same size as the patches. We then train the segmentation network using as additional instances these patches, together with the corresponding segmentation maps. So, the patches selected to cover the entire image are always the same, while the crops will change at each epoch and will be taken at random in the entire image. These crops, being randomly selected, can also overlap each other. An example of this process can be seen in Fig. 2.

This enables us to further improve the efficiency of our training process while trying to enhance the generalization capacities of our model.

### 3. Experiments

In this section, we offer a description of the used dataset by highlighting its characteristics. We also describe the detailed training setup adopted for the experiments. Then, we outline the metrics used for the evaluation of our framework as well as provide an ablation study, aimed at supporting the effectiveness of the choices defining the proposed system.



**Figure 4:** A page of the CSG863 manuscript class ( 4a) and the corresponding ground truth segmentation map ( 4b): the magenta areas represent the main text, the yellow the comments and the cyan the decorations.

### 3.1. Dataset

To train and test our system we selected the DIVA-HisDB dataset [8]. It is a historical handwritten document dataset consisting of a high-resolution and RGB color annotated pages coming from three different medieval manuscripts, identified as CB55, CSG18 and CSG863. These documents have complex and heterogeneous layouts. Moreover, as an additional challenge, they have different levels of degradation. A sample page for each manuscripts are reported in Fig. 3.

The dataset consisting of a total of 150 images where, for each manuscripts, 20 are typically used for training, 10 for validation and another 20 for testing. For the present work, we only relied on 2 images for each manuscript to train the model.

DIVA-HisDB supplies pixel-level ground truth segmentation (Fig. 4) for the layout of each image, which distinguishes between 4 classes of elements: background, main text, comments and decoration.

### 3.2. Hyperparameters setup

For the training of the proposed model, the loss function selected is a weighted cross-entropy loss. This is due to the fact that DIVA-HisDB, and in general all of the historical manuscripts, is a very imbalance between classes. Detail of DIVA-HisDB distribution is provided in Tab. 1.

	Background	Comments	Decoration	Text
<b>CB55</b>	82.41	8.36	0.55	8.68
<b>CSG18</b>	85.16	6.78	1.47	6.59
<b>CSG863</b>	77.82	6.35	1.83	14.00

**Table 1**  
Classes distribution (%).

Then, the weights for each class are calculated by taking the square root of 1 over the class frequency in the dataset (Eq. 1, where  $F_i$  represents the frequency (%) of class in the corresponding class dataset).

$$W_i = \sqrt{\frac{1}{F_i}} \quad (1)$$

The ADAM optimizer with a learning rate of  $1e^{-3}$  and a weight decay of  $1e^{-5}$  were used. The maximum number of epochs during training model was 200 with an early stop if the network didn't improve over the last 20 iterations after 50 epochs.

The original images in the dataset have high spatial resolution (up to  $4.8k \times 6.8k$  px), thus, to reduce the model's computational complexity, they have been resized to  $1120 \times 1344$  px.

To train the proposed model have been selected 2 images for each manuscript and divided into patches of size  $224 \times 224$  px, for a total of 60 patches for each manuscript. This set is then enhanced by generating 10 random crops of the same size for each image as part of our dynamic training routine generating a maximum of 4000 instances if the model needs all the epochs to converge.

### 3.3. Metrics

The performance at pixel level of proposed approach is evaluated by Precision, Recall, Intersection over Union (IoU) and F1-Score. These metrics were calculated individually for each manuscripts following the definition reported in Eq. 2– 5, where TP, FP and FN stand respectively for True Positives, False positives and False Negatives, and then a weighted average, based on each class frequency has been performed.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (4)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

### 3.4. Results

We have conducted an ablation study on the different versions of the proposed framework. The baseline approach consists of using whole images as backbone network input. Our improved model is obtain by running the same network on a patch level with dynamic crop generation. A comparison of the two versions is presented in Tab. 2. We report both the scores for the individual manuscripts as well as the final averaged ones.

In general, as we can observe, the image split into patches and the dynamic crop generation determine an improvement in the framework performance with an average improvement across all the metrics. Our framework raises all of the selected metrics in all of the classes, in particular in CB55.

	CB55				CSG18				CSG863				Mean			
	Prec	Rec	IoU	F1												
<i>Ours (baseline)</i>	0.814	0.829	0.705	0.795	0.853	0.862	0.756	0.831	0.893	0.881	0.782	0.862	0.853	0.857	0.748	0.829
<i>Ours (w/ dynamic crop gen.)</i>	<b>0.867</b>	<b>0.853</b>	<b>0.735</b>	<b>0.824</b>	<b>0.869</b>	<b>0.873</b>	<b>0.775</b>	<b>0.845</b>	<b>0.894</b>	<b>0.885</b>	<b>0.789</b>	<b>0.867</b>	<b>0.877</b>	<b>0.870</b>	<b>0.766</b>	<b>0.845</b>

**Table 2**

Results of the ablation study. Each rows shows the performance of the different versions of our system across all the selected metrics for the 4 manuscripts of DIVA-HisDB dataset. The last four columns show the average scores.



**Figure 5:** Image shows the comparison between the segmentation results obtained by our framework and the ground truth provided by DIVA-HisDB. Each column represents a different instance of the three classes of manuscripts.

In Fig. 5 we provide some examples of the results. In particular, we show the comparison between our results and the provided ground truth. As can be seen, our model provides a different level of precision (a coarser segmentation map) compared with the fidelity of the ground truth segmentation map.

## 4. Conclusions

In this paper, we presented a few-shot learning framework for the layout segmentation for the historical handwritten document. In particular, we introduced a dynamic instance generation module that allowed us to increase the model performance, while maintaining the requirement of few segmented images for training. This was an important aspect of our research because manually annotating the ground truth of manuscripts is a heavy task, thus reducing the number of required segmented pages greatly helps humanists.

For future works, we would like to refine our results in order to get pixel-level segmentation, as the ground truth. This would provide similar results to the current state-of-the-art for historical document layout segmentation, but maintaining few-shot learning, that in our opinion is a fundamental feature.

## References

- [1] F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold, M. Liwicki, Icdar2017 competition on layout analysis for challenging medieval manuscripts, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 1, IEEE, 2017, pp. 1361–1370.
- [2] P. Andrist, Toward a definition of paratexts and paratextuality: the case of ancient greek manuscripts, *Bible as Notepad. Tracing Annotations and Annotation Practices* (2018) 130–149.
- [3] M. Mehri, N. Nayef, P. Héroux, P. Gomez-Krämer, R. Mullet, Learning texture features for enhancement and segmentation of historical document images, in: *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, 2015, pp. 47–54.
- [4] Y. Xu, F. Yin, Z. Zhang, C.-L. Liu, et al., Multi-task layout analysis for historical handwritten documents using fully convolutional networks., in: *IJCAI*, 2018, pp. 1057–1063.
- [5] S. A. Oliveira, B. Seguin, F. Kaplan, dhsegment: A generic deep-learning approach for document segmentation, in: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2018, pp. 7–12.
- [6] G. M. Binmakhashen, S. A. Mahmoud, Document layout analysis: a comprehensive survey, *ACM Computing Surveys (CSUR)* 52 (2019) 1–36.
- [7] A. Garz, M. Seuret, F. Simistira, A. Fischer, R. Ingold, Creating ground truth for historical manuscripts with document graphs and scribbling interaction, in: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, IEEE, 2016, pp. 126–131.
- [8] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, R. Ingold, Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts, in: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2016, pp. 471–476.
- [9] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.