Few-Shot Pixel-Precise Document Layout Segmentation via Dynamic Instance Generation and Local Thresholding

Few-shot pixel-precise document layout segmentation via dynamic instance generation and local thresholding

Axel De Nardin
Department of Mathematics, Computer Science and Physics, Università degli Studi di Udine
Via delle Scienze 206, 33100 Udine, Italy
Email: denardin.axel@spes.uniud.it

Silvia Zottin
Department of Mathematics, Computer Science and Physics, Università degli Studi di Udine
Via delle Scienze 206, 33100 Udine, Italy
Email: zottin.silvia@spes.uniud.it

Claudio Piciarelli
Department of Mathematics, Computer Science and Physics, Università degli Studi di Udine
Via delle Scienze 206, 33100 Udine, Italy
Email: claudio.piciarelli@uniud.it

Emanuela Colombi
Department of Humanities and Cultural Heritage, Università degli Studi di Udine
Vicolo Florio 2/b, 33100 Udine, Italy
Email: colombi.emanuela@uniud.it

Gian Luca Foresti
Department of Mathematics, Computer Science and Physics, Università degli Studi di Udine
Via delle Scienze 206, 33100 Udine, Italy
Email: gianluca.foresti@uniud.it

Over the years, the humanities community has increasingly requested the creation of artificial intelligence frameworks to help the study of cultural heritage. Document Layout segmentation, which aims at identifying the different structural components of a document page, is a particularly interesting task connected to this trend, specifically when it comes to handwritten texts. While there are many effective approaches to this problem, they all rely on large amounts of data for the training of the underlying models, which is rarely possible in a real-world scenario, as the process of producing the ground truth segmentation task with the required precision to the pixel level is a very time-consuming task and often requires a certain degree of domain knowledge regarding the documents at hand. For this reason, in the present paper, we propose an effective few-shot learning framework for document layout segmentation relying on two novel components, namely a dynamic instance generation and a segmentation refinement module. This approach is able of achieving performances comparable to the current state of the art on the popular Diva-HisDB dataset, while relying on just a fraction of the available data.

Keywords: Document layout segmentation ; Few-shot learning ; Image segmentation ; Handwritten documents analysis ; Layout analysis

1. Introduction

During the last few decades, many libraries and archives have focused their activities on ensuring worldwide access to their cultural heritage documents in digital form. Consequently, many challenges and open issues have been raised by researchers, his-

torians, humanists, and computer scientists working on cultural heritage documents to optimize the management, navigation, and analysis of digitized document images [1].

In DIA, two main types of analysis can be highlighted. The first one is Optical Character Recognition (OCR), used to derive the meaning of the document characters and words, and the other is layout analysis to determine the formatting of the document page and text. These tasks may be performed separately, or the results from one analysis may be used to aid or correct the other [2].

OCR lies at the core of the discipline of pattern recognition where the objective is to understand and recognize language characters from different idioms, either handwritten or printed [3; 4]. The goal of OCR is to detect the characters contained in a document image and to transfer these into digital text.

Layout analysis is the process of identifying and recognizing the physical and logical organization and structure of document images [5; 6]. Document layout analysis includes three main tasks each with a specific purpose and which leads to the study of different characteristics of the document. These three key subprocesses of layout analysis are page segmentation, text line segmentation and baseline detection.

Page segmentation is a prerequisite step of DIA. Page segmentation is the process that segments the document images into different semantically meaningful regions like main text, paratexts, decorations and background. In particular, the page segmentation of historical manuscripts allows humanists to study documents more quickly and easily because it allows the paratexts (i.e all the semantic elements which are part of the foreground but don't belong to the main text) to be analyzed separately.However, performing this task in historical manuscripts is much more difficult than in printed documents [7], due to many variations, such as layout structure, decoration, different writing styles, texture, and degradation.

Text line segmentation aims at identifying the areas in the document corresponding to each text line. Text line extraction is one of the previous stages of the OCR and document transcription process [8].

Finally, baseline detection is the task that has the objective of identifying in the document image the virtual line where characters of the text rest upon and descenders extend below [9].

So, the tasks of document image layout analysis refer to the segmentation of a given document image into semantically meaningful regions, such as main text, paratexts, decorations, and background or even the detection of individual baseline or text lines [10].

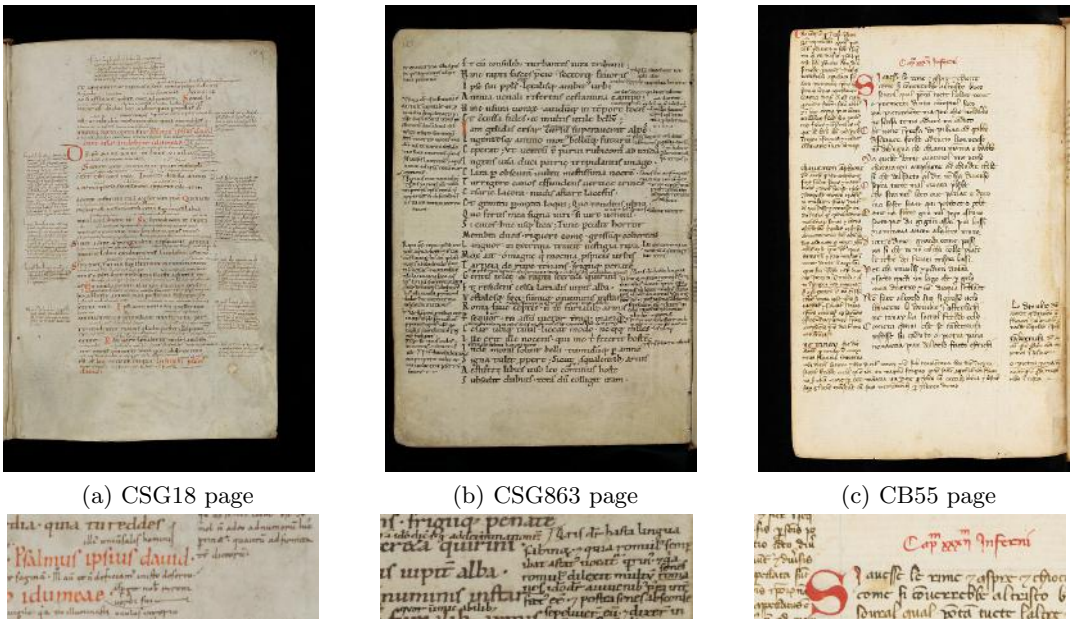
In recent years many machine learning and deep learning algorithms have been used to perform this task and there are many fields in which it is applied to the real context. For example, it is widely used for the problem of anomaly detection in visual inspection in industrial production, biomedical image analysis or infrastructure defects [11–14]. Further deep learning models for semantic segmentation are used for autonomous driving, identifying obstacles such as pedestrians, sidewalks, poles, and other cars [15].

Machine learning and deep learning models for semantic segmentation have also been used to perform the task of layout analysis in historical documents [5] and especially in our work to perform page segmentation.

In order to develop machine learning and deep learning-based approaches and compare the performance of different segmentation methods, ground truth is needed. For ground truth to be suitable for training accurate deep learning models, the annotation of the segmentation masks must be as precise as possible down to the pixel level [16]. The disadvantage is that the pixel-precise annotation of the entire historical document page dataset is a very time-consuming process and requires domain-specific knowledge, which only an expert humanist can satisfy, especially when working with historical manuscripts [17], making this type of information rarely available in a real-world scenario. Nonetheless, few-shot learning approaches in the context of document layout segmentation are still under-explored in the literature.

For this reason, in the present work, which extends earlier work [18], we propose a novel few-shot learning framework for efficient pixel-precise page segmentation of historical manuscripts, which is able to accurately segment the different component of a document page (e.g. text, paratext, images) achieving results comparable to the current state-of-the-art approaches on the popular Diva-HisDB dataset (Fig. 1) while using only a fraction of the available data for the training process.

The rest of the paper is organized as follows: in



(a) CSG18 page (b) CSG863 page (c) CB55 page

(d) CSG18 detail (e) CSG863 detail (f) CB55 detail

Fig. 1: Samples from the three representative manuscripts (CSG18, CSG863 and CB55) present in DIVA-HisDB. Fig. 1a– 1c show a full page for each manuscripts, while Fig. 1d– 1f show a detail extracted from each of them.

Section 2 we provide a review of the recent works in the context of document layout analysis. Sections 3 and 4 contain a detailed description of the proposed framework and the experimental setup used to train and test it. Section 5 provides an in-depth description of the results achieved by our model, both from a quantitative and a qualitative perspective. Finally, in Section 6, the conclusions are drawn.

2. Related Works

Different approaches have been proposed to tackle the DIA tasks, especially for historical manuscripts. In general, several surveys have been created concerning the DIA tasks for both ancient and recent documents, both printed and handwritten [19; 20]. But the proposed solutions for DIA tasks in ancient handwritten documents are few.

Page segmentation, which is the task that we address in this work, is an open problem in the machine-learning community. The techniques employed for this task are usually divided into three categories: top-down, bottom-up and hybrid [19].

Top-down approaches assume that pages have a well-defined structure and layout. The page segmentation process starts from the whole page and cuts it into homogeneous zones. In these methods, various characteristics of the document page structure are considered, such as white space between text regions, size of text blocks and the measures between main texts and paratexts [21; 22]. In general, the top-down methods are easily applicable but not suitable for complex layouts such as handwritten historical documents. In addition, these methods depend on the layout structure of the document, so they have a low generalization capability.

The bottom-up strategy derives document analysis dynamically from smaller granularity data levels, such as pixels, groups of pixels and connected components, to generate larger and different regions with uniform elements [23; 20; 24]. The main advantage of bottom-up methods is that these techniques do not require any prior knowledge of the document layout structure, and, for this reason, this strategy is preferred when working with ancient manuscripts. However, usually, these techniques demand many labeled data that is often not available, especially in the domain of historical documents where highly specialized expertise is needed. That is why the request for methods with few-shot and unsupervised learning is increasingly necessary and required.

Although research for this task is well established, there are still many challenging issues that

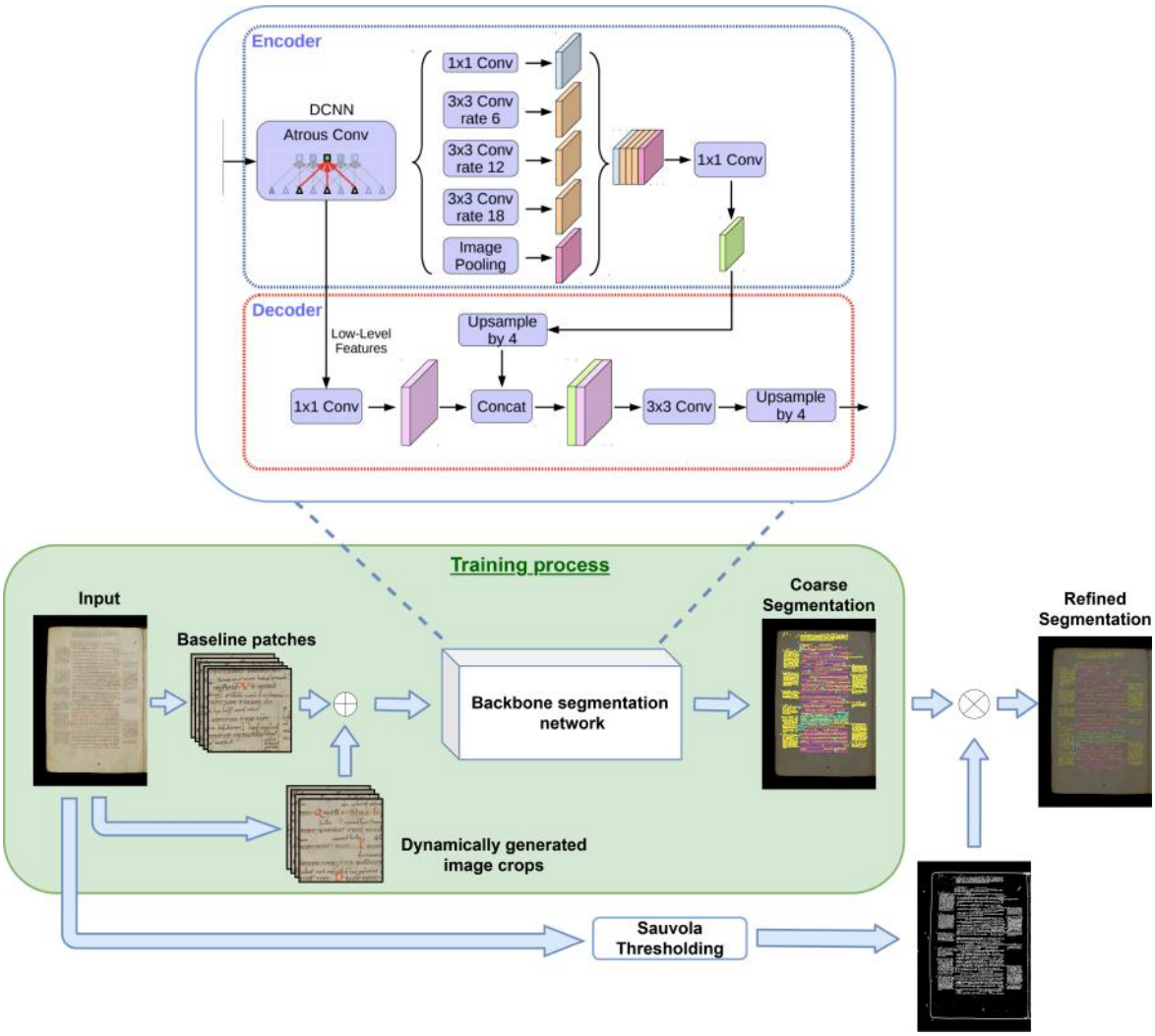


Fig. 2: Visual representation of the segmentation pipeline for the proposed framework. The green area represents the processes carried out during the training phase, where each input image is split into 2 sets of patches: the baseline patches, which are non-overlapping patches of size $k \times k$ providing a complete representation of the original image and a set of C random crops which are extracted from random locations of the image at each training epoch. These 2 sets of patches are then combined and given in input to the backbone segmentation model which provides a predicted coarse segmentation map for each of them. These maps are compared with the ground truth ones through the application of a weighted cross-entropy loss. At inference time the dynamic instance generation step is removed while a segmentation refinement process is applied to the outputs of the backbone architecture to obtain more precise segmentation maps

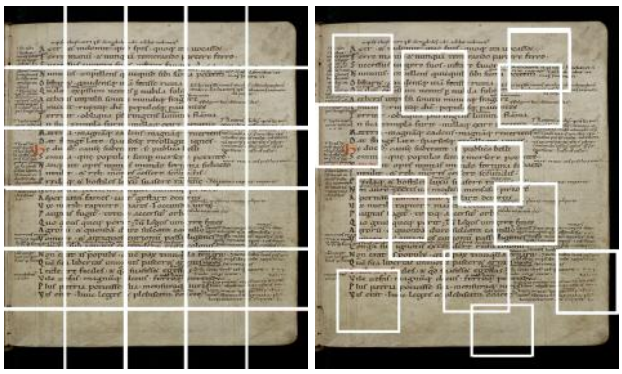
neither bottom-up nor top-down strategies can adequately address. For this reason, a strategy has been identified that derives from the integration of the other two main ones, called the hybrid strategy [25; 26].

Over the years, many techniques have been used to address the page segmentation task. Recently, [27] proposed a learning-free and hybrid document layout analysis for historical manuscripts. First, the proposed method locates the main content initially using top-down white space analysis. Then, it extracts template features representing the manuscripts author's writing behavior. After that, moving windows are used to scan the manuscript page and define main-content boundaries more precisely. [28] used a convolutional autoencoder to learn the features directly from the pixel intensity values and train

a Support Vector Machine (SVM) to segment the page without any assumption of specific shapes of document layouts. A similar approach was proposed by [23] with the method based on learning texture features. This method for extracting texture features from images uses the linear iterative clustering of super-pixels, Gabor descriptors and the co-occurrence matrix of the gray level. Then an SVM is used to classify pixels into foreground and background. The page segmentation problem can approach as a pixel labeling problem such as the work by [29], where the features are learned directly from randomly selected image patches by using stacked convolutional autoencoders. With an SVM trained with the extracted features of the central pixels of the super-pixels, an image is segmented into semantic regions. Finally, the segmentation results are refined by a connected components-based smoothing procedure. Following the same idea, in [30] local features are learned with stacked convolutional autoencoders in an unsupervised manner for the purpose of initial labeling. Then a conditional random field model is applied for modeling the contextual information to improve the segmentation results. Another interesting approach was proposed by [31] with a multi-task framework to solve all layout analysis tasks simultaneously. The model trains a multi-task fully convolutional network to predict pixel-wise classes and as the final step a heuristic-based post-processing is adopted to reduce noise and correct misclassification. The prediction of the four branches was combined to produce the result of layout analysis tasks. [26] proposed a hybrid method for page segmentation problems. In the first stage, the text and non-text elements are classified by using a minimum homogeneity algorithm which is the combination of connected component analysis and multilevel homogeneity structure. Then, in the second stage, a new homogeneity structure is combined with an adaptive mathematical morphology in the text document to get a set of text regions. [32] proposed a novel method for document layout analysis that reduces the need for labeled data. This method is a dictionary-based feature learning model where a sparse autoencoder is first trained in an unsupervised manner on a document's image patch. Then, the latent representation of image patches is then used to classify pixels into various region categories of the document using a feed-forward neural network. Also, [33] used the patching of the document image to train a siamese network model that takes an input a pair of patches and gives as an output a distance that corresponds to their similarity. The trained model is also used to calculate a distance matrix which in turn is used to cluster the patches of a page as either main text, side text or a background patch. [34] tackle the problem of the limited presence of annotated data by introducing the use of pre-trained segmentation models on images from a different domain and then fine-tuning them on historical handwritten documents. The results demonstrated that on some manuscripts pre-training on ImageNet increases the performance, but on others, the pre-trained network performs much worse. Also [35] try to tackle the problem of the limited presence of ground truth by presenting an unsupervised deep learning method for page segmentation. In this work a Siamese neural network is trained to differentiate between patches using their measurable properties such as the number of foreground pixels so that spatially nearby patches are similar. The network's learned features are used for page segmentation. Finally, [36] propose the few-shot learning approach Deep&Syntax to segment historical handwritten registers. Their work uses a hybrid system that exploits recurring patterns to delimit each record, combining U-shaped networks and logical rules such as filter and text alignment. While the presented approaches have different degrees of effectiveness when trying to solve the document layout segmentation task, they all rely on large amounts of data for their training. The main contribution we bring with the present work is the ability to achieve similar, or even better performance while relying on just a fraction of the available data.

3. Proposed approach

The proposed approach is built on three core components, namely a robust segmentation backbone used to retrieve the semantic components of each document page, a dynamic instance generation module that allows us to fully leverage the limited amount of data available at training time and finally a segmentation refinement module that makes it possible to further improve the quality of the segmentation maps produced by our model. A visual representation of the proposed framework pipeline is reported in Fig. 2.



(a) Baseline patches (b) Randomly selected crops

Fig. 3: Representation of the instance generation process of the 2 sets of patches used to train our model: in 3a is shown the generation process for baseline non-overlapping patches, while 3b provides a visual depiction of our dynamic crop generation process

3.1. *Segmentation backbone*

Adoption of a robust backbone is a crucial step in each Deep Learning framework. When working in a few-shot setting in particular we need a network that is able to capture a sufficient level of detail while being given in input just a handful of samples. For this reason, we selected DeepLabV3+ [37] as the backbone of our framework. DeepLabV3+ is a popular pixel-wise semantic segmentation model built on its predecessor DeepLabV3 [38]. The latter is a ResNet [39] based architecture heavily relying on atrous convolutions which are employed both in parallel and in a cascade in order to enlarge the receptive fields of the filters and consequently retain a higher spatial resolution throughout the network. The key advantage of this approach is that it allows for deeper neural networks that provide larger feature maps at no additional computational cost. Finally, the Atrous Spatial Pyramid Pooling (ASPP) is introduced in DeepLabV3 as a way of capturing features at different scales in the original image by relying on a heterogeneous set of dilation rates in the network. DeepLabV3+ introduces two substantial changes compared to the aforementioned architecture. The first one regards the substitution of the ResNet encoder with a custom version of the Aligned XCeption [40] model in which all max pooling operations are replaced by depth-wise separable convolution. Furthermore, it adds a simple yet effective decoder which refines the segmentation results. The decoder module employs depth-wise separable convolutions to enhance the spatial resolution of the feature maps, resulting in sharper and more detailed output segmentation maps.

3.2. *Dynamic instance generation*

The dynamic instance generation module is a key component of the training pipeline of our framework. The key idea behind it is that it efficiently exploits the small amount of data available at training time. To do so, instead of relying on the full document pages as the instances of our dataset, we split them into two sets of smaller patches. The first ones, which we will refer to as baseline patches, consist of a set of non-overlapping sub-regions of size $m \times n$ extracted from the original input image in order to cover its entire surface and are kept consistent between the training and inference time (Fig. 3a). In addition to the baseline patches, as a way to further improve the generalization capabilities of our model, we also generate a small set of k potentially overlapping crops of the same size as the baseline patches which are extracted from random locations of the original image (Fig. 3b). This process is carried out at each epoch during training time, while at inference time no additional crops are generated as they are not needed to obtain the final segmentation mask. While relying on sub-patches of the original images is a common approach in computer vision-related tasks, in most cases, these patches are either limited to the ones corresponding to our baseline ones, which leads to losing potentially useful information contained in the data. As an alternative approach, they may generate a large number of patches in advance, without considering the varying complexity of different datasets [31] . As a consequence, excessive amounts of potentially unnecessary data is produced. Our dynamic instance generation approach addresses both limitations effectively at the cost of a very small computational overhead at training time.

3.3. *Segmentation refinement*

Our segmentation refinement module is based on the Sauvola thresholding algorithm for document binarization [41]. The Sauvola thresholding algorithm is an evolution of Niblack's method [42], which introduced the idea of a dynamic threshold that is calculated based on the mean and standard deviation of the gray levels of a local window inside an image. The

main drawback of Niblack's approach is that it didn't perform well for images with a light-textured background as it would result in very noisy binarization masks. Sauvola solved this problem by introducing the dynamic range of the standard deviation as an additional term in the equation used to calculate the local threshold, which has the effect of amplifying the contribution of the standard deviation in an adaptive manner throughout the image. The resulting equation adopted by the Sauvola algorithm is shown in Eq. 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.

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

The refined segmentation masks are then obtained by performing the Hadamard product between the layout segmentation predictions provided by our backbone and the mask resulting from running the Sauvola algorithm on the corresponding images of the dataset.

4. Experimental setup

In this section, we outline a detailed description of the dataset adopted for the experiments and the training setup. Furthermore, the metrics used to evaluate and compare the performance of the proposed approach are presented, together with the results of the ablation study.

Table 1: Classes distribution (%) for each manuscripts of Diva-HisDB [43] (CB55, CSG18 and CSG863), and for Bukhari et al. [44] dataset

Manuscript	BG	Comment	Decoration	Text
CB55	82.41	8.36	0.55	8.68
CSG18	85.16	6.78	1.47	6.59
CSG863	77.82	6.35	1.83	14.00
Bukhari et al.	86.07	4.71	—	9.22

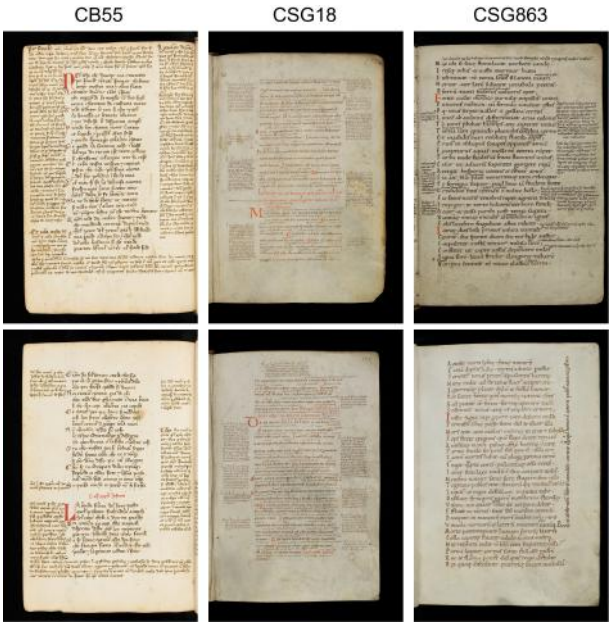


Fig. 4: Instances selected from each manuscript in DIVA-HisDB as the training set for the proposed approach. Each of them was chosen to effectively represent the characteristics for the corresponding class

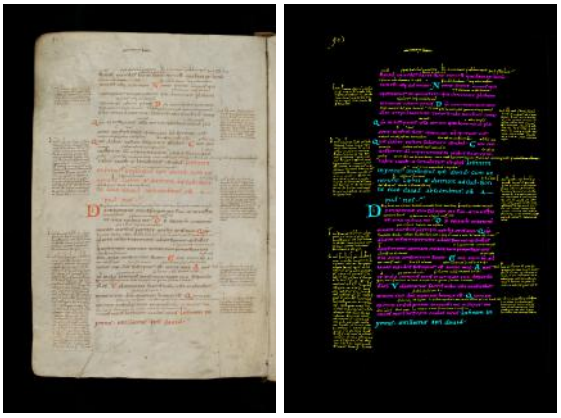
4.1. *Dataset*

To train and test our model we selected the popular Diva-HisDB dataset [43]. Diva-HisDB is a collection of 3 medieval manuscripts (CB55, CSG18 and CSG863) selected for their heterogeneity and layout complexity. All the documents contained in the dataset are characterized by 4 classes of semantic components, namely main text, comments, decorations and background (BG), with very unbalanced distributions making the dataset particularly challenging for a few shot settings as the less common classes are present in a very small amount or not at all in some of the instances. A detail of the semantic component distributions for each manuscript is provided in Tab. 1.

Furthermore, the manuscripts provide a high degree of heterogeneity concerning the level of degradation of the pages, the epoch in which they were written, both inter and intra class differences in the pages layout and in writing styles, as both the CSG18 and CSG863 were written by an unspecified number of authors. The dataset consists of a total of 150 instances, 50 for each document class, of these 60 are typically used for training, 30 for validation and another 60 for testing the models. In the present work,

we relied on just 6 images, 2 for each class, for training our model (Fig. 4). For each of the document pages, the dataset provides a corresponding ground truth segmentation mask as shown in Fig. 5.

Finally, to further validate the robustness of our approach we also tested it on the dataset proposed by Bukhari et al.[44] which 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, while for the purpose of this paper we only relied on 3 images, one for each manuscript, to train our model. A detail of the semantic component distributions is provided in Tab. 1.



(a) Original page (b) Page ground truth

Fig. 5: Images showing a page of the CSG18 manuscript (5a) as well as its corresponding ground truth mask (5b), in which the magenta areas represent the main text, while the yellow and cyan areas represent the comments and decorations respectively. Finally, the black area represents the background of the image

4.2. *Training and inference setup*

Our model was trained using the popular Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5}. The maximum number of epochs for which it was allowed to run has been set to 200 with an early stop in case the validation loss didn't improve in the last 20 epochs and a buffer of 50 epochs which guarantees that the model will be trained at least for the specified amount of iterations. During each epoch, a set of 10 dynamic crops of size 672×672 px has been generated in addition to the baseline patches of the same sizes extracted from the original image. This process led to a maximum of 4012 instances being generated for each document class during training, in case the model needed all the 200 epochs in order to converge. In order to be able to fit them in the GPU memory the images of the dataset have been resized from their original high resolution (up to $4.8k \times 6.8k$ px), down to a size of 1344×2016 px. The loss function selected to train the model is a weighted Cross Entropy Loss [45] in which the weight for each semantic element class is inversely proportional to the frequency of that element in that dataset and, more precisely is calculated as the square root of 1 over the square root of the occurrence frequency of the corresponding element in the dataset (Eq. 2).

$$W_i = \sqrt{\frac{1}{F_i}} \tag{2}$$

This specific choice was made to take into account the high imbalance between the semantic classes distribution in each document category of the datasets (Tab. 1). Our model was trained separetly and from scratch on each document class. Regarding the inference setup, the main choice involved in it is represented by the hyper parameters of the segmentation refinement algorithm, namely the window size which was kept consistent at 15×15 px for all document classes and the control value k, which regulates the value of the threshold in the local window (the higher the k value, the lower the threshold) and was set at the value of 0.01 for all classes.

4.3. *Evaluation metrics*

In order to evaluate the performance of our proposed approach we use different metrics such as Precision, Recall, Intersection over Union (IoU) and F1-Score. These evaluation metrics are calculated individually for each one of the manuscripts that compose DIVA-HisDB dataset. Metric definitions are reported in Eq. 3– 6, where TP, FP and FN stand respectively for True Positives, False positives and False Negatives. For each metric a weighted average is performed, based on each class frequency in each manuscript. The final evaluation of a model is then obtained by averaging the metrics of all pages of the

Table 2: Comparison between the performance of our model and the competition on the 4 selected metrics. The best and second-best performing models are reported in a bold and underlined fashion respectively while FS indicates the models trained in a few-shot setting by using the same set of images selected for our framework

	CB55				CSG18				CSG863				Mean			
Backbone	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
FCN (FS)	0.894	0.883	0.783	0.863	0.874	0.885	0.797	0.863	0.915	0.907	0.826	0.895	0.894	0.892	0.802	0.874
FCN	0.902	0.900	0.815	0.884	0.930	0.930	0.869	0.919	0.923	0.919	0.847	0.909	0.918	0.916	0.844	0.904
LRSAPP (FS)	0.847	0.837	0.718	0.808	0.919	0.913	0.871	0.911	0.869	0.864	0.757	0.842	0.878	0.871	0.782	0.854
LRSAPP	0.880	0.883	0.789	0.864	0.921	0.927	0.868	0.918	0.911	0.910	0.833	0.899	0.904	0.907	0.830	0.894
PSPNET (FS)	0.876	0.868	0.761	0.846	0.906	0.905	0.829	0.890	0.913	0.896	0.817	0.888	0.898	0.890	0.802	0.875
PSPNET	0.887	0.894	0.811	0.880	0.912	0.920	0.857	0.910	0.913	0.915	0.845	0.906	0.904	0.910	0.838	0.899
DeepLabV3 (FS)	0.893	0.883	0.784	0.863	0.901	0.895	0.806	0.873	0.864	0.853	0.737	0.828	0.886	0.877	0.776	0.855
DeepLabV3	0.905	0.901	0.817	0.886	0.930	0.931	0.871	0.920	0.920	0.914	0.839	0.903	0.918	0.915	0.842	0.903
DeepLabV3+ (FS)	0.908	0.903	0.821	0.888	0.931	0.929	0.867	0.918	0.936	0.933	0.875	0.927	0.925	0.922	0.854	0.911
DeepLabV3+	0.943	0.945	0.896	0.939	0.961	0.962	0.929	0.959	0.965	0.965	0.935	0.964	0.956	0.957	0.920	0.954
MLA	-	-	-	-	-	-	-	-	-	-	-	-	0.965	**0.995**	**0.989**	**0.995**
Ours	**0.989**	**0.987**	**0.977**	**0.988**	**0.983**	**0.982**	**0.967**	**0.982**	**0.986**	**0.983**	**0.971**	**0.984**	**0.986**	0.984	0.972	0.985

three manuscripts.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

5. Results

In the following section, we provide a thorough comparison between the results achieved by the proposed framework and a set of popular semantic segmentation approaches, namely DeepLabV3 [38], its improvement represented by DeepLabV3+ [37], FCN [46], Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) [47] and Pyramid Scene Parsing Network (PSPNet) [48], furthermore we also include the results obtained by current state of the art for the task of document layout segmentation, which we will refer to as MLA [31]. The comparison focuses both on a quantitative and a qualitative perspective in order to provide a complete overview of the quality of the model's predicted segmentations. To this end, we also provide a discussion about the critical cases in which our approach fails to provide the correct segmentation for the corresponding instances. All the models, excluding MLA for which we gathered the results from the respective paper, have been personally tested by us keeping the training and evaluation settings as consistent as possible.

5.1. *Quantitative results*

In Tab. 2 the quantitative results achieved by our proposed framework for all the selected metrics across all the document classes contained in the DivaHisDB dataset, are shown and compared with the competitor models. In particular, for all the models, excluding MLA, we provide both the results obtained by training them on the entire available dataset and the ones obtained by training the model only on the subset of 2 pages selected for our approach (FS = Few-Shot setting). Unfortunately, MLA authors provided only the mean scores for the selected metrics and some implementation details were missing, leading our attempt at reimplementing their work to achieve sub-optimal results. As we can see our model is consistently capable of outperforming the other semantic segmentation networks on all the metrics, re-

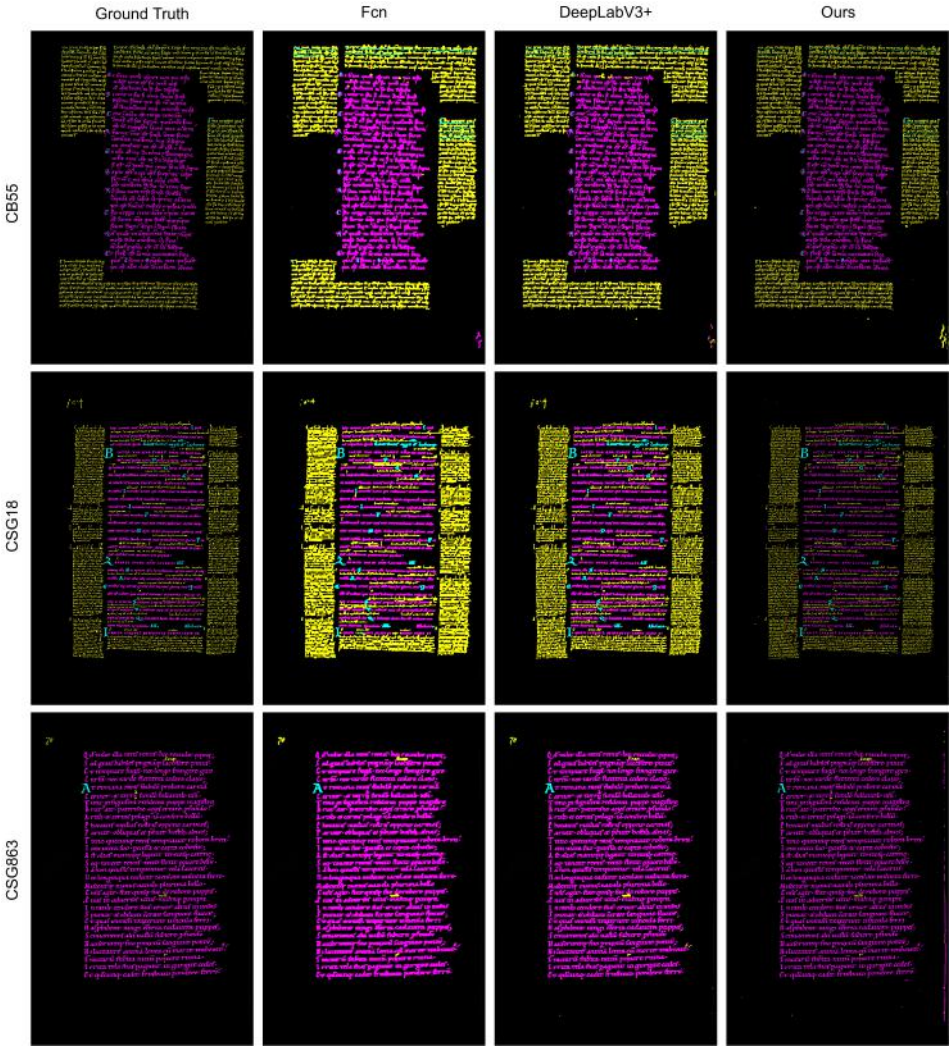


Fig. 6: Image showing a qualitative comparison between our framework and the competition ones. Each row represents a zoomed area belonging to a different instance of the dataset, representing the three classes of manuscript contained in it. In the first column, the ground truth segmentation maps for the 3 images are shown, while on the remaining columns we provide the results produced by the three systems, FCN, DeepLabV3+ and Ours respectively

gardless of the setup in which they have been trained. In particular, compared to the second best performing approach, being represented by DeepLabV3+, our model achieves an mean improvement of 7.7% when the former is trained in a few-shot setting with a peak improvement of 11.8% for the IoU metric. While when DeepLabv3+ is trained using the full training set, our approach outperforms it by a still substantial mean of 3.5% (5.2% for the IoU metric) while using only a fraction of the available data.

Furthermore, our framework achieves very close performance even when compared with the current state-of-the-art MLA, even surpassing it by 2.1% on the mean precision metric. As for the remaining metrics our model performance is still comparable to that of MLA with a difference going from 1.7% for the IoU metric, to as little as 1% for the F1-score. It is important to notice, however, that MLA is trained on around 180000 instances extracted from all the images of the training set, while our framework, as previously mentioned, extracts at most 4012 unique instances from just 2 of the available images in the training set, resulting in a reduction of the needed data by a factor approximately 45.

Finally in Tab. 3 we show the comparison between our model and the competition on the Bukhari dataset for Arabic manuscript layout segmentation. As we can see our framework achieves the best performance compared to all the other approaches even when they are trained using the full training set. In particular, compared to the single best performing model, being represented by DeepLabV3+ our achieves a 2-4% improvement across all metrics against its fully trained configuration and around a 4-9% performance improvement against the few shot version of the model.

Table 3: Comparison between the performance of our model and the competition on the Bukhari dataset. The best and second-best performing models are reported in a bold and underlined fashion respectively while FS indicates the models trained in a few-shot setting by using the same set of images selected for our framework

Backbone	Prec	Rec	IoU	F1
FCN (FS)	0.836	0.875	0.788	0.853
FCN	0.865	0.899	0.824	0.879
LRSAPP (FS)	0.806	0.858	0.742	0.805
LRSAPP	0.899	0.876	0.806	0.884
PSPNET (FS)	0.843	0.859	0.770	0.846
PSPNET	0.911	0.861	0.790	0.875
DeepLabV3 (FS)	0.879	0.815	0.735	0.836
DeepLabV3	0.908	0.871	0.802	0.883
DeepLabV3+ (FS)	0.929	0.907	0.850	0.914
DeepLabV3+	<u>0.956</u>	<u>0.943</u>	<u>0.902</u>	<u>0.946</u>
Ours	**0.970**	**0.966**	**0.940**	**0.967**

5.2. *Qualitative results*

Fig. 6 shows the segmentation maps produced by our model for three document pages belonging, respectively, to the three document class present in the Diva-HisDB dataset and compared with the ones predicted by the FCN and DeepLabV3+ models, both trained on the whole available training set. Furthermore, the corresponding ground truth segmentation is provided as a reference.

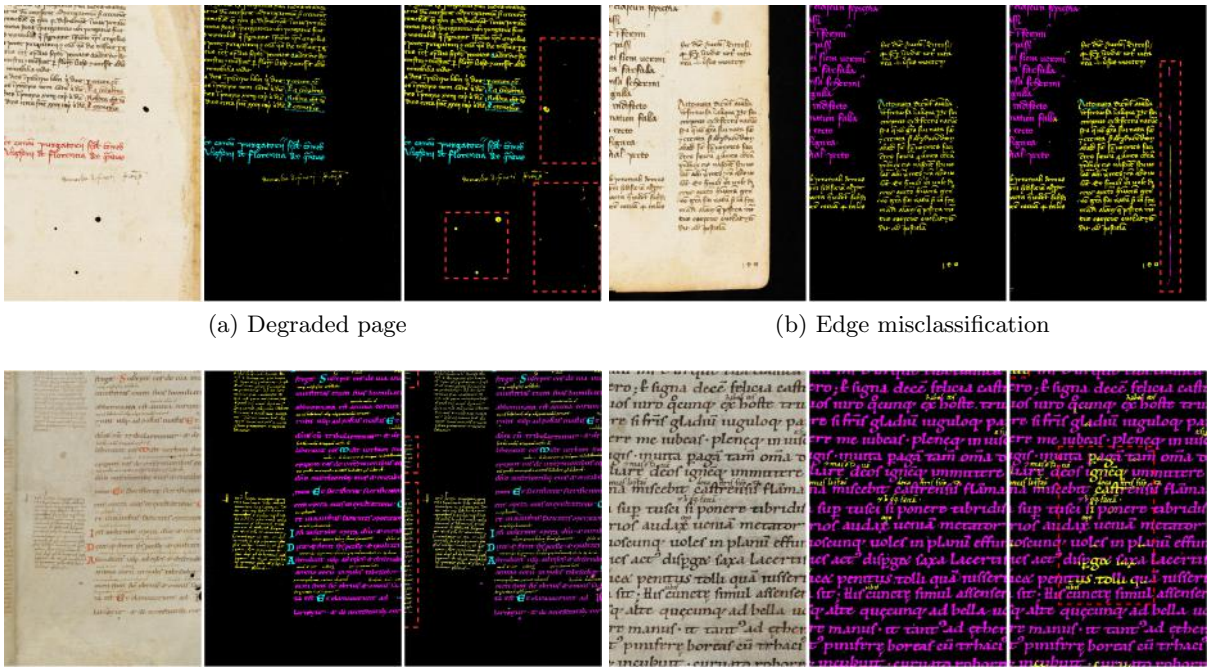
While the maps produced by FCN are typically correct and with very limited amounts of noise, they tend to be very coarse, especially when observed in the areas of the pages where the text is smaller and the different components more intertwined. DeeplabV3+ provides a higher level of detail, in particular when looking at the main text component (magenta segmentation). Finally, our model provides visibly more precise segmentation maps than the competition when compared to the ground truth ones.

5.2.1. *Fail cases*

As already mentioned in the previous section the main drawback of the presented approach is that compared to the competition it introduces more noise in the provided segmentations. For completeness, in Fig. 7 we provide some more criticalities of the proposed framework together with the original image and the corresponding ground truth. In particular, other than the typical misclassification of foreground elements (Fig. 7d) we can notice three main instances of recurrent mistakes. The first one is represented by the edge of the pages of the documents which, being lighter than the black background introduces an area of high contrast which is identified both by the model and by the thresholding algorithm as part of the text (Fig. 7b). A similar occurrence can be observed for degraded areas in the page's background, these areas are, in fact, typically darker than the rest of the background and are once again misclassified as foreground elements (Fig. 7a). Finally, we have the misclassification caused by the text belonging to the page adjacent to the currently analyzed instance, which while correctly identified as part of the text by our model, is not included in the ground truth segmentations (Fig. 7c). This last case, however, is highly dependent on the coarse cropping process of the instances of the Diva-HisDB dataset which doesn't precisely include only the elements of the current page and, as such, is easily solvable by refining the crops.

5.3. *Ablation study*

In this section, we provide the details regarding the ablation study we conducted in order to obtain the final version of the proposed framework. In particular,



(a) Degraded page (b) Edge misclassification

(c) Adjacent page text (d) Foreground misclassification

Fig. 7: Overview of the main instances of misclassification for the proposed approach. From the top left corner we have: 7a degraded spots in the page background being misclassified as foreground, 7b the same type of misclassification involving page edges, 7c Text belonging to the adjacent page being recognized as part of the current one, 7d a simple case of misclassified foreground elements in particular involving the main text being mistaken as part of the comments

we show the effects that different segmentation backbones and patch sizes for the generated instances have on the performance of our approach for the task at hand. Furthermore, we provide a comparison between the performance of the baseline model and the models enhanced with the additional modules introduced in this paper in order to provide proof of their effectiveness.

5.3.1. *Backbones*

Tab. 4 shows a comparison of the performance of our framework when using different backbones for the segmentation module. For this comparison, we selected a set of recent and popular semantic segmentation networks (DeepLabV3 [38], DeepLabV3+ [37], FCN [46], LRASPP [47] and PSPNet [48]). To allow for a fair comparison all the models have been trained and tested with the exact same setup, with 2 images for each document class as the training set and a consistent patch size of 672×672 px. As we can see all the models provide reasonably good performance on the task at hand achieving an IoU higher than 70% and a performance of over 80% for all the remaining metrics. From this analysis emerges that DeepLabV3+ consistently outperforms all other models on each of the selected metrics and on all the document classes present in the dataset, achieving an mean improvement of 6.23% over the second-best model, being represented by its predecessor DeepLabV3. A particularly interesting boost in performance is achieved for the IoU metrics where an increase of almost 9% is obtained by the former over the latter.

5.3.2. *Patch sizes*

A further comparison has been performed by exploring the adoption of different sizes for the crops of the instances being provided to the backbone networks. In particular, we selected 3 different sizes, going from the standard 224×224 which is the size used by all the pre-trained models available in PyTorch, to a much larger 672×672. The results of this comparison are shown in Tab. 5. In this case, the dif-

Table 4: Comparison between the use of different neural network architectures as the segmentation backbone for our model, in bold is reported the best performing model

	CB55				CSG18				CSG863				Mean			
Backbone	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
FCN	0,871	0,850	0,728	0,820	0,887	0,876	0,773	0,847	0,884	0,874	0,770	0,851	0,881	0,867	0,757	0,839
LRASPP	0,801	0,770	0,614	0,718	0,815	0,835	0,718	0,797	0,919	0,908	0,858	0,912	0,845	0,838	0,730	0,809
PSPNET	0,849	0,828	0,694	0,792	0,877	0,869	0,761	0,838	0,901	0,887	0,801	0,876	0,876	0,861	0,752	0,835
DeeplabV3	0,873	0,853	0,734	0,824	0,891	0,881	0,781	0,854	0,882	0,869	0,762	0,845	0,882	0,868	0,759	0,841
DeeplabV3+	**0.918**	**0.908**	**0.827**	**0.894**	**0.926**	**0.923**	**0.855**	**0.910**	**0.931**	**0.927**	**0.863**	**0.917**	**0.925**	**0.919**	**0.848**	**0.907**

Table 5: Comparison between the adoption of different patch sizes during the instance generation process of our framework, in bold is reported the best performing model

	CB55				CSG18				CSG863				Mean			
Patch size	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
224	0.911	0.900	0.813	0.884	0.920	0.916	0.843	0.900	0.919	0.917	0.846	0.904	0.917	0.911	0.834	0.896
336	0.916	0.906	0.823	0.891	0.925	0.920	0.849	0.905	0.928	0.926	0.860	0.916	0.923	0.917	0.844	0.904
672	**0.918**	**0.908**	**0.827**	**0.894**	**0.926**	**0.923**	**0.855**	**0.910**	**0.931**	**0.927**	**0.863**	**0.917**	**0.925**	**0.919**	**0.848**	**0.907**

ference in performance wasn't as substantial as the one resulting from the adoption of different types of segmentation backbones. In particular, we can notice that the difference between the best and the worst performing models, which are the ones adopting the largest and smallest patch sizes respectively, is on mean around 1%. A potential explanation behind the improved performance corresponding to the adoption of larger patch sizes is that the model to which they are given in input is able to capture a higher amount of contextual information regarding the layout of the original image from which they are extracted, allowing for more accurate segmentation.

5.3.3. *Framework modules*

Finally, we provide a comparison between different versions of our framework in which we systematically introduce the original modules presented in this paper, namely the dynamic instance generation and the segmentation refinement ones. In particular, in Tab. 6 we show the performance obtained by our baseline model, in which the images have been split into patches but without the addition of either the dynamically generated crops or the segmentation refinement process, as well as the one achieved by introducing these 2 techniques singularly and in a combined fashion, which represents our full framework pipeline. As we can see each of the additional modules leads to a substantial improvement in performance over the baseline approach with the best performance being achieved with the use of both modules. More specifically the final framework achieves an improvement in performance going from 6.8% for the precision metric to a very substantial 13.3% for the Intersection over Union one, with an mean improvement of 9% across all metrics when compared to the baseline approach.

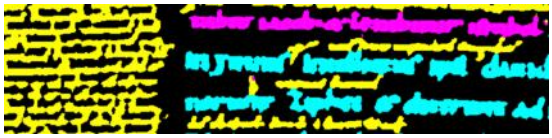
As additional proof of the effectiveness of the proposed approach. In Fig. 8 we provide a qualitative comparison between the segmentation masks

Table 6: Results of the ablation study. Each row shows the performance of the different versions of our system across all the selected metrics for the 4 classes of manuscripts composing the DIVA-HisDB dataset. The last four columns show the mean scores achieved by the models across the different classes

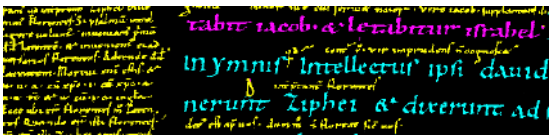
	CB55				CSG18				CSG863				Mean			
	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
Ours (baseline)	0.907	0.900	0.815	0.884	0.926	0.923	0.860	0.912	0.917	0.914	0.840	0.900	0.917	0.912	0.838	0.899
Ours (w/ dynamic crop gen.)	0.918	0.908	0.827	0.894	0.926	0.923	0.855	0.912	0.931	0.927	0.863	0.917	0.925	0.919	0.848	0.907
Ours (w/ seg. refinement)	0.979	0.978	0.967	0.976	0.981	0.978	0.963	0.979	0.982	0.980	0.965	0.980	0.981	0.979	0.965	0.978
Ours (w/ both)	**0.989**	**0.987**	**0.977**	**0.988**	**0.983**	**0.982**	**0.967**	**0.982**	**0.986**	**0.983**	**0.971**	**0.984**	**0.986**	**0.984**	**0.972**	**0.985**

provided by the baseline and the final framework, while also showing the corresponding ground truth as a reference.

(a) Ground Truth



(b) Coarse Prediction



(c) Refined Prediction

Fig. 8: Qualitative results showing the effects of the segmentation refinement process. Fig. 8a shows the original ground truth for a zoomed area of the original image. Fig. 8b shows the coarse segmentation mask obtained by the model. Finally Fig. 8c shows the segmentation prediction resulting from the refinement process

6. Conclusions

In this paper, we proposed an effective framework that tackles the underexplored problem of few-shot document layout analysis by introducing two original modules, namely the dynamic instance generation and segmentation refinement ones which help the core image segmentation backbone to fully leverage the small amount of training data available in order to achieve pixel-precise segmentations of the document pages. When compared to other popular image segmentation algorithms, our model consistently outperforms them, while relying only on a fraction of the training data and with a computational load that is comparable to the one of the original backbone segmentation network adopted, being represented by DeepLabV3+. Furthermore, when compared to the current State of the Art framework, our approach achieves comparable performance on all the selected metrics. While the reported results are very promising, there are still some criticalities we plan to address in the future, specifically by investigating more effective segmentation refinement strategies.

Acknowledgements

Partial financial support was received from Piano Nazionale di Ripresa e Resilienza (PNRR) DD 3277 del 30 dicembre 2021 (PNRR Missione 4, Componente 2, Investimento 1.5) - iNEST.

References

1. O. Mechi, M. Mehri, R. Ingold and N. Essoukri Ben Amara, Text line segmentation in historical document images using an adaptive u-net architecture, *2019 International Conference on Document Analysis and Recognition (ICDAR)*, (Sidney, Australia, 2019), pp. 369–374.
2. R. Kasturi, L. O'Gorman and V. Govindaraju, Document image analysis: A primer, *Sadhana* **27** (Feb 2002) 3–22.
3. D. Berchmans and S. S. Kumar, Optical character recognition: An overview and an insight, *2014*

International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), (Kanyakumari District, India, 2014), pp. 1361–1365.

4. S. Drobac and K. Lindén, Optical character recognition with neural networks and post-correction with finite state methods, *Int. J. Doc. Anal. Recognit.* **23** (dec 2020) p. 279–295.
5. F. Lombardi and S. Marinai, Deep learning for historical document analysis and recognition—a survey, *Journal of Imaging* **6** (Oct 2020) p. 110.
6. S. Biswas, P. Riba, J. Lladós and U. Pal, Beyond document object detection: Instance-level segmentation of complex layouts, *Int. J. Doc. Anal. Recognit.* **24** (sep 2021) p. 269–281.
7. J. Y. Ramel, S. Leriche, M. L. Demonet and S. Busson, User-driven page layout analysis of historical printed books, *International Journal of Document Analysis and Recognition (IJDAR)* **9** (Apr 2007) 243–261.
8. O. Mechi, M. Mehri, R. Ingold and N. Essoukri Ben Amara, A two-step framework for text line segmentation in historical arabic and latin document images, *Int. J. Doc. Anal. Recognit.* **24** (sep 2021) p. 197–218.
9. M. Diem, F. Kleber, S. Fiel, T. Grüning and B. Gatos, cbad: Icdar2017 competition on baseline detection, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, **01**, (Kyoto, Japan, 2017), pp. 1355–1360.
10. F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold and M. Liwicki, Icdar2017 competition on layout analysis for challenging medieval manuscripts, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, **01**, (Kyoto, Japan, 2017), pp. 1361–1370.
11. A. De Nardin, P. Mishra, G. L. Foresti and C. Piciarelli, Masked transformer for image anomaly localization, *International Journal of Neural Systems* **32**(07) (2022) p. 2250030, PMID: 35730477.
12. J. Lin, L. Ma and Y. Yao, A spectrum-domain instance segmentation model for casting defects, *Integrated Computer-Aided Engineering* **29** (2022) 63–82, 1.
13. G. Mirzaei and H. Adeli, Segmentation and clustering in brain mri imaging, **30** (2023 2019) 31–44.
14. Z. Wang, Y. Zhang, K. M. Mosalam, Y. Gao and S.-L. Huang, Deep semantic segmentation for visual understanding on construction sites, *Computer-Aided Civil and Infrastructure Engineering* **37** (Feb 2022) 145–162.
15. Ç. Kaymak and A. Uçar, *A Brief Survey and an Application of Semantic Image Segmentation for Autonomous Driving*, *Handbook of Deep Learning Applications* (Springer International Publishing, Cham, 2019), Cham, ch. 9, pp. 161–200.
16. A. Garz, M. Seuret, F. Simistira, A. Fischer and R. Ingold, Creating ground truth for historical manuscripts with document graphs and scribbling interaction, *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, (Santorini, Greece, 2016), pp. 126–131.
17. K. Nikolaidou, M. Seuret, H. Mokayed and M. Liwicki, A survey of historical document image datasets, *Int. J. Doc. Anal. Recognit.* **25** (dec 2022) p. 305–338.
18. A. De Nardin, S. Zottin, M. Paier, G. L. Foresti, E. Colombi and C. Piciarelli, Efficient few-shot learning for pixel-precise handwritten document layout analysis, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Waikoloa, Hawaii, 2023), pp. 3680–3688.
19. G. M. Binmakhashen and S. A. Mahmoud, Document layout analysis: A comprehensive survey, *ACM Comput. Surv.* **52** (oct 2019).
20. M. Mehri, P. Héroux, P. Gomez-Krämer and R. Mullot, Texture feature benchmarking and evaluation for historical document image analysis, *International Journal on Document Analysis and Recognition (IJDAR)* **20** (Mar 2017) 1–35.
21. R. Cohen, A. Asi, K. Kedem, J. El-Sana and I. Dinstein, Robust text and drawing segmentation algorithm for historical documents, *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, *HIP '13*, (Association for Computing Machinery, New York, NY, USA, 2013), p. 110–117.
22. A. Asi, R. Cohen, K. Kedem, J. El-Sana and I. Dinstein, A coarse-to-fine approach for layout analysis of ancient manuscripts, *2014 14th International Conference on Frontiers in Handwriting Recognition*, (Crete, Greece, 2014), pp. 140–145.
23. M. Mehri, N. Nayef, P. Héroux, P. Gomez-Krämer and R. Mullot, Learning texture features for enhancement and segmentation of historical document images, *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, *HIP '15*, (Association for Computing Machinery, New York, NY, USA, 2015), p. 47–54.
24. N. Journet, J.-Y. Ramel, R. Mullot and V. Eglin, Document image characterization using a multiresolution analysis of the texture: application to old documents, *International Journal of Document Analysis and Recognition (IJDAR)* **11** (Oct 2008) 9–18.
25. P. Barlas, S. Adam, C. Chatelain and T. Paquet, A typed and handwritten text block segmentation system for heterogeneous and complex documents, *2014 11th IAPR International Workshop on Document Analysis Systems*, (Tours, France, 2014), pp. 46–50.
26. T. A. Tran, I. S. Na and S. H. Kim, Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology, *International Journal on Document Analysis and Recognition (IJDAR)* **19** (Sep 2016) 191–209.
27. G. M. BinMakhashen and S. A. Mahmoud, Histor-

ical document layout analysis using anisotropic diffusion and geometric features, *International Journal on Digital Libraries* **21** (Sep 2020) 329–342.
28. K. Chen, M. Seuret, M. Liwicki, J. Hennebert and R. Ingold, Page segmentation of historical document images with convolutional autoencoders, *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, (Nancy, France, 2015), pp. 1011–1015.
29. K. Chen, C.-L. Liu, M. Seuret, M. Liwicki, J. Hennebert and R. Ingold, Page segmentation for historical document images based on superpixel classification with unsupervised feature learning, *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, (Santorini,Greece, 2016), pp. 299–304.
30. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, C.-L. Liu and R. Ingold, Page segmentation fosar historical handwritten document images using conditional random fields, *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, (Shenzen, China, 2016), pp. 90–95.
31. Y. Xu, F. Yin, Z. Zhang and C.-L. Liu, Multi-task layout analysis for historical handwritten documents using fully convolutional networks, *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, *IJCAI'18*, (AAAI Press, Stockholm, Sweden, 2018), p. 1057–1063.
32. H. Davoudi, M. Fiorucci and A. Traviglia, Ancient document layout analysis: Autoencoders meet sparse coding, *2020 25th International Conference on Pattern Recognition (ICPR)*, (Milan, Italy, 2021), pp. 5936–5942.
33. R. Alaasam, B. Kurar and J. El-Sana, Layout analysis on challenging historical arabic manuscripts using siamese network, *2019 International Conference on Document Analysis and Recognition (ICDAR)*, (Sydney, Australia, 2019), pp. 738–742.
34. L. Studer, M. Alberti, V. Pondenkandath, P. Goktepe, T. Kolonko, A. Fischer, M. Liwicki and R. Ingold, A comprehensive study of imagenet pretraining for historical document image analysis, *2019 International Conference on Document Analysis and Recognition (ICDAR)*, (Sydney, Australia, 2019), pp. 720–725.
35. A. Droby, B. K. Barakat, B. Madi, R. Alaasam and J. El-Sana, Unsupervised deep learning for handwritten page segmentation, *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, (Dortmund, Germany, 2020), pp. 240–245.
36. S. Tarride, A. Lemaitre, B. Coüasnon and S. Tardivel, Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples, *International Journal on Document Analysis and Recognition (IJDAR)* **24** (Jun 2021) 77–96.
37. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Computer Vision – ECCV 2018*, eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Springer International Publishing, Cham, 2018), pp. 833–851.
38. L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, Rethinking atrous convolution for semantic image segmentation, *arXiv preprint arXiv:1706.05587* (2017).
39. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, Nevada, 2016), pp. 770–778.
40. F. Chollet, Xception: Deep learning with depthwise separable convolutions, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE Computer Society, Los Alamitos, CA, USA, jul 2017), pp. 1800–1807.
41. J. Sauvola and M. Pietikäinen, Adaptive document image binarization, *Pattern Recognition* **33**(2) (2000) 225–236.
42. W. Niblack, *An Introduction to Digital Image Processing* (Prentice hall, Englewood Cliffs, 1986).
43. F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki and R. Ingold, Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts, *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, (Shenzen, China, 2016), pp. 471–476.
44. S. S. Bukhari, T. M. Breuel, A. Asi and J. El-Sana, Layout analysis for arabic historical document images using machine learning, *2012 International Conference on Frontiers in Handwriting Recognition*, (Bari, Italy, 2012), pp. 639–644.
45. S. Jadon, A survey of loss functions for semantic segmentation, *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2020, pp. 1–7.
46. J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, MA, 2015), pp. 3431–3440.
47. A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam and Q. Le, Searching for mobilenetv3, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea, 2019), pp. 1314–1324.
48. H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, hawaii, 2017), pp. 6230–6239.