International Journal on Document Analysis and Recognition (IJDAR) (2024) 27:335–347
https://doi.org/10.1007/s10032-024-00475-w

SPECIAL ISSUE PAPER

Approximate ground truth generation for semantic labeling of historical documents with minimal human effort

Najoua Rahal[1] · **Lars Vögtlin**[1] · **Rolf Ingold**[1]

Received: 14 November 2023 / Revised: 15 March 2024 / Accepted: 8 May 2024 / Published online: 12 June 2024

Abstract
Deep learning approaches have shown high performance for layout analysis of historical documents, provided that enough labeled data is available. This is not an issue for generic tasks such as image binarization, text graphics separation, or text line and text block detection but can become an impediment for more specialized tasks specific to one or a few books only. This paper addresses layout analysis of medieval books with rich and complex layouts, for which no labeled data is initially available. The proposed strategy consists of training an initial model with artificial data created to reflect the rules a deep neural network should learn. Then, the model is iteratively fine-tuned by mixing the artificial data with real data obtained by previous predictions, post-processed, and manually selected by an expert user. Such a strategy needs less human effort than manual ground truthing. The approach is qualitatively and quantitatively assessed and shows that the system converges to an accurate model that finally produces approximate ground truth stable and good enough to train a final model to solve the targeted task with high accuracy.

Keywords Historical documents · Layout analysis · Deep learning · Synthetic data

1 Introduction

Historical documents contain a wealth of valuable information representing the cultural heritage of human history and civilization. Thanks to digitization, these documents are preserved in a digital format. With the emergence of computer vision, it is becoming feasible to analyze them effectively. This momentum is particularly vivid in the research topic of layout analysis, for which there has been a notable increase in research initiatives over the last few decades [1].

Layout analysis is the crux of document image processing and the prerequisite step for text recognition. It enables the splitting of a document into semantic homogeneous units such as background, text blocks, tables, etc [2]. The main challenges of layout analysis of historical documents are the heterogeneity and complexity of layouts and degradations. With the introduction of deep neural networks, recent research approaches have reported significant progress in historical document layout analysis by providing pixel-wise annotations. They generally focus on text line, title, figure, and table detection [3, 4]. With a vast amount of ground truth used for training, these approaches perform reasonably well and achieve near-perfect accuracy. However, they usually cannot cope with more specialized needs when it comes to accurately analyzing specific and more complex layout structures.

In this paper, we address the issues raised by some richly decorated medieval manuscript documents characterized by various types of ornaments and decorations to be distinguished. Also, many challenges arise such as decorative text, ink bleed-though, decorated and colorful objects, etc. Of course, in theory, it should be possible to adapt existing deep neural networks by using transfer learning [5, 6]. The problem we face with such a strategy is the lack of labeled training data. This problem can be solved in two ways: annotating manually real-world data or generating synthetic data. Annotating manually real-world data is costly in terms of time and requires a lot of manpower and expertise. Synthetic data generation can be a way to solve this problem. Many approaches,

✉ Najoua Rahal
najoua.rahal@unifr.ch

Lars Vögtlin
lars.voegtlin@unifr.ch

Rolf Ingold
rolf.ingold@unifr.ch

[1] Document Image and Voice Analysis Group (DIVA), University of Fribourg, 1700 Fribourg, Switzerland

the majority of which are based on generative adversarial networks (GANs) [7], have been proposed for synthetic data generation in multiple domains: semantic segmentation [8, 9], handwritten text recognition, for both contemporary documents [10, 11] and historical documents [12, 13], as well as scene text detection and recognition [14, 15].

In this paper, we propose a novel training strategy that considerably reduces the effort required to produce such labeled data. The idea consists of a bootstrapping approach starting with artificial data and progressively including real data by replacing them with the pages that obtained the best predictions of the previously trained model. With such an approach, the human effort is reduced to selecting the documents to be included in the following training step.

The main contributions of this paper can be summarized as follows:

1. We propose a new iterative strategy for semantic labeling requiring less human effort.
2. Based on that strategy, we provide approximate ground truth for a new dataset.

To the best of our knowledge, we are the first to produce an approximate high-quality ground truth based on an iterative process.

The remainder of this paper is organized as follows: Sect. 2, the most relevant existing related works are reviewed. The data and tasks we address are presented in Sect. 3. In Sect. 4, we describe our proposed strategy. Our experiments and their evaluations are presented in Sect. 5. Finally, our conclusion and future work are presented in Sect. 6.

2 Related works

With the burst of deep neural networks and their improvements in computer vision and natural language processing, researchers have recently explored valuable deep learning-based approaches for the semantic segmentation of historical document images. This section reviews notable existing techniques tackling the closest related works to ours.

Xu et al. [16] proposed a deep, fully Convolutional Neural Network (FCN) for page segmentation of historical handwritten documents of the DIVA-HisDB database [17]. The proposed network, based on the VGG 16-layer network [18], was trained to predict a pixel's class as background, main text body, comment, or decoration. However, several modifications are applied to fit the use case of page segmentation, such as low-level processing in an earlier stage of the network and additional convolutional layers before the last phases. Then, heuristic post-processing was adopted to refine the coarse segmentation results by reducing noises and correcting misclassified pixels by connected component analysis.

Renton et al. [5] introduced an FCN using dilated convolutions for text line segmentation of historical document images. The proposed network was applied on the cBAD dataset [19] to detect only the text class at an x-height area.

For baseline detection in historical documents, Grüning et al. [20] proposed ARU-Net, a variant of the U-Net model. It consists of two stages; the first relies on assigning each pixel to one of three classes: baseline, separator, or other. The second stage focuses on a bottom-up clustering approach to build baselines. ARU-Net has been evaluated on DIVA-HisDB [17] and cBAD [19] datasets to detect only text at baseline level. This model used data augmentation to artificially increase the amount of training data and, thus, obtain better results.

Oliveira et al. [21] proposed the dhSegment, a multi-task FCN followed by a post-processing block for pixel classification. The tasks are mainly page extraction, baseline detection, document layout analysis, ornament detection, and photo-collection extraction. To accomplish these tasks, dhSegment used dilated convolution blocks. The cBAD dataset [19]was used to evaluate the page extraction and baseline detection tasks. The document layout analysis task aimed to assign each pixel to one of the following classes: text region, decoration, comment, or background. This method was evaluated on the DIVA-HisDB dataset [17]. We note that the ornament detection and photo-collection extraction were evaluated using private manually annotated datasets.

Boillet et al. [22] presented the Doc-UFCN model, inspired by the dhSegment model [21], for text line segmentation. The difference between DocUFCN and dhSegment lies in the used encoder. The encoder of dhSegment is the ResNet-50 [23] architecture, pre-trained on natural scene images, while the encoder of Doc-UFCN is fully trained on historical document images. Also, the Doc-UFCN model had less trainable parameters than other state-of-the-art networks. They proved that pre-training an FCN model on multiple datasets and fine-tuning on small datasets improves text line segmentation. The Doc-UFCN model was evaluated on Balsac [24], Horae [25], READ-BAD [26], and DIVA-HisDB [17] datasets.

Rahal et al. [27] addressed two sub-tasks of layout analysis of historical documents: page segmentation and text line detection. The paper proposed L-U-Net, a new variant of the U-Net model, with dilated convolutions and a constant number of 16 filters at each block. They showed that a model with much fewer parameters can perform better while being lighter for training than the most popular models of the U-Net family. The presented model was evaluated on DIVA-HisDB [17] and cBAD [19] datasets.

The paper proposed by Rahal et al. [6] addressed the text line detection and classification when only a few annotated training data are available. It presented two novel training strategies: pre-training the networks with controlled data and

morphological operators. The first strategy consisted of using artificial data applied to the real task. The second strategy used real data to pre-train the network on a pretext task with morphological data automatically generated. These strategies proved that pre-training with either artificial data or a pretext task can improve the final training. It was also shown that an architecture with fewer parameters can perform better while being faster for training. The experiments were carried out on CB55 and CSG18, two subsets of DIVA-HisDB [17].

The authors, in [28], presented SwinDocSegmenter, transformer-based [29] approach for instance-level semantic segmentation of complex document images including historical documents. Historical Japanese benchmark dataset [30] was used for the evaluation (the other datasets are not mentioned because they do not contain historical documents). It contained seven class labels: body, row, title, bio, name, position, and other. The proposed approach used a model with one billion parameters, making it impossible to be trained with limited data.

Let us summarize the relevant state-of-the-art approaches done previously. Deep learning-based approaches have demonstrated significant success in substantially solving many challenges of historical document image segmentation tasks at a pixel level. They achieved near-perfect accuracy in different mono- and multi-task problems. However, many of these models tend to overfit on small training datasets due to their millions of parameters. To avoid this problem, previous approaches proposed data augmentation [20], self-supervised learning [6], and transfer learning strategies [6, 22] to address the semantic segmentation of historical documents when only a few labeled data for training is available. Other experiences [27] showed that a smaller network with fewer parameters is well-suited for the semantic segmentation of historical document images. It prevents overfitting and achieves competitive results using fewer computing resources.

One of the most pressing issues of deep learning-based approaches is that they usually cannot cope with more challenging data when no annotations are available. In the last few years, this issue has become an active research area that has attracted the research community's interest. It is always an issue that has not been entirely solved, and state-of-the-art approaches still have a vast space to enhance. In this paper, we propose a novel training strategy to address the above-mentioned challenge. Our strategy is detailed in Sect. 4.

3 Tasks and datasets

3.1 Tasks description

Layout analysis is a fundamental process that remains the prerequisite of many historical document analysis tasks, such as text recognition. It aims to segment a document image into regions of interest. There are several methods to locate the regions of interest. Our work considers two main tasks: complex layout labeling and text line detection combined with classification. For complex layout labeling, we consider seven classes: *background*, *decoration*, *body* (main part of content), *text line*, *large drop caps*, *small drop caps*, and *filler* (filling the white spaces). For text line detection and classification, we consider four classes: *background*, *highlights*, *main text*, and *glosses*.

Table 1 Post-processing parameters for UTP-110 prediction enhancement

Layer	Sequence of operation	Kernel
Body+*	Opening	(51,51)
	Closing	(91,91)
Text line	Opening	(5,5)
	Closing	(3,75)
Filler	Opening	(5,5)
	Closing	(5,101)
Large drop caps	Opening	(5,5)
	Closing	(5,51)
Small drop caps	Opening	(5,5)
	Closing	(5,15)
Decor	Closing	(91,91)
	Opening	(55,55)

(*) Body+ includes Body, Text line, Large drop caps, Small drop caps, and Filler

3.2 Real data

Two real datasets were considered for our experiments: UTP-110 and CB55.

UTP-110 is part of the medieval manuscript collection Utopia, armarium codicum bibliophilorum, Cod. 110. The book, which the Master of Charles VIII wrote in the early 16th century in French, is entitled Book of hours.[1] It has 300 pages mostly in Latin, partly in French, publicly accessible from the digital library e-codices[2] The document images have been resized to 640×960 pixels, keeping the original image ratio. This medium resolution is convenient for capturing the layout structure of entire images with adequate precision and minimizing computing time without losing too much information. The documents of this manuscript raise real challenges for layout analysis, such as different types of ornaments, decorative text, faded writing, and ink bleed-through.

[1] https://museumsandcollections.unimelb.edu.au/__data/assets/pdf_file/0020/2031545/05_Maddocks_Book-of-hours16.pdf.

[2] https://www.e-codices.unifr.ch/en/searchresult/list/one/utp/0110.

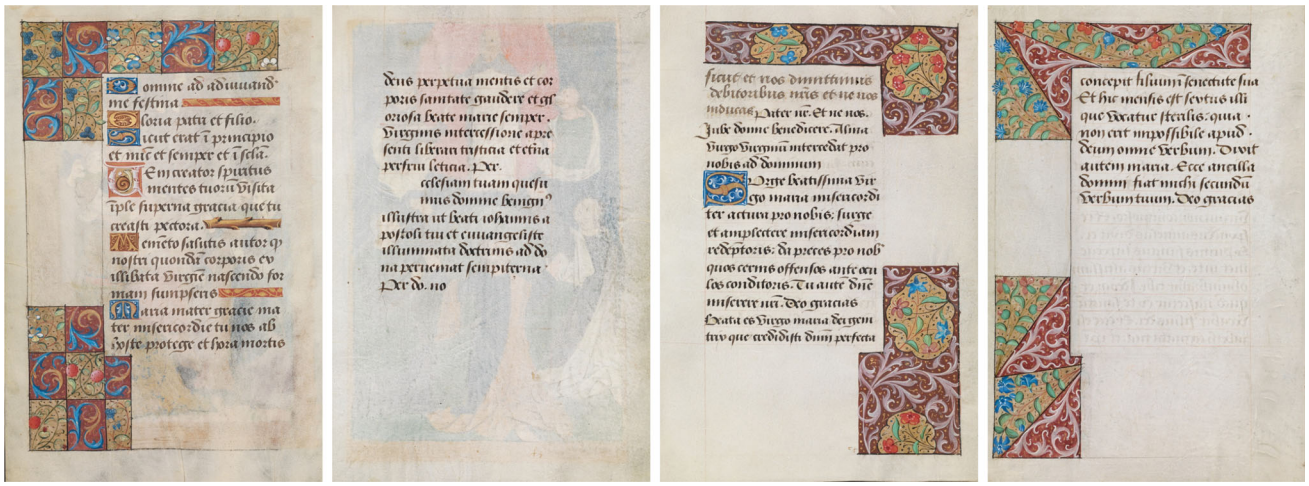


Fig. 1 Four example pages of the UTP-110 dataset

CB55, Codex Guarneri, which was written in the first half of the 14th century in Italy [31], is a sub-set of DIVA-HisDB [17]. It consists of 120 pages and is composed of the Inferno and Purgatorio from Dante's Divina Commedia. The images have been cropped and resized to 960 × 1344 pixels. This corresponds to a medium resolution considered appropriate to capture the layout structure of entire pages with sufficient precision. The ground truth was created semi-automatically by adding an average x-height for text lines to the existing DIVA-HisDB baseline annotation. This dataset has just been used to assess the proposed strategy on a task we already had results of previous experiments.

3.3 Artificial data generation

The artificial dataset was designed for an initial training step to provide an initial model that captures the general layout rules of the complex layout structures. Based on previous observations [6], our aim was not to generate documents with realistic appearance nor degradation. We focused on significant variability in geometric characteristics, writing styles, and ornaments.

We wrote a dedicated Python program to create this artificial dataset using the image processing package Pillow. This program comprises several modules that individually produce the basic building blocks of such documents.

- The background generator simulates the color and texture of the parchment with some smooth local brightness and hue changes.
- The ink generator produces a layer that simulates the color of the written text; similarly to the background, the ink color is undergoing some local color changes.
- The text generator produces a grayscale image to be used as a transparency mask to combine ink and background layers. Six different fonts with different sizes have been used to get enough variability. Additionally, line spacing is also randomized.
- The decoration generator produces large decorations surrounding the central text blocks. It has been designed to generate a large set of different colors with a texture that combines painted surfaces and strokes modeled by splines with different stroke widths.
- A drop-caps generator generates drop caps composed of a colored background on top of which a large capital is drawn; two different sizes, stretching respectively over one and two text lines, can be produced.
- An additional graphic generator simulates additional decorations such as fillers that fill the gaps of text lines at the end of paragraphs.
- Finally, the layout manager uses elastic rules to construct a layout structure with randomized positions and content for left- and right-side pages.

All these modules have been designed to simultaneously produce the synthetic document images as PNG files and their associated ground truth presented as indexed color images in GIF format.

Figure 2 shows some samples of artificial image elements. Figure 3 shows some samples generated by the specific modules. Figure 4 illustrates the global layout structure of an entire page with its associated ground truth.

4 Proposed strategy

4.1 Contribution

Let $D_L = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ denotes the labeled training dataset. x_i denotes the data sample and y_i denotes the

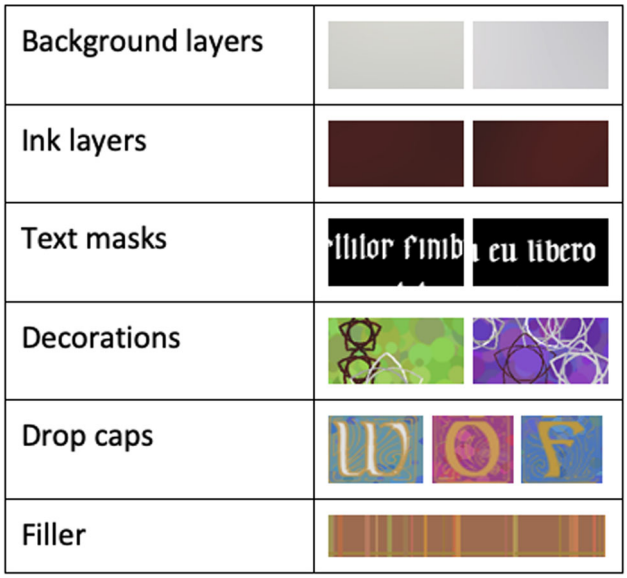


Fig. 2 Samples of artificial image elements

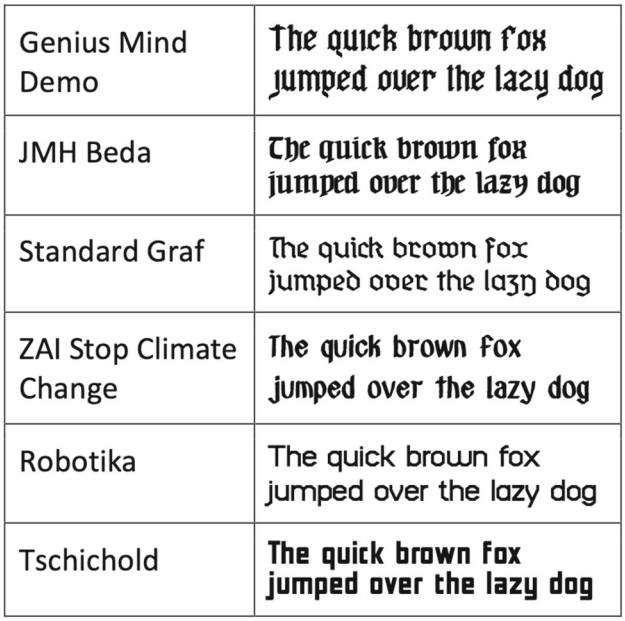


Fig. 3 List of fonts used for generating artificial data

corresponding mask label. Let $D_U = \{\hat{x_1}, \ldots, \hat{x_n}\}$ denotes the real unlabeled data. Let $P_A = \{\hat{y_1}, \ldots, \hat{y_n}\}$ denotes the prediction of the real data D_U after post-processing. As shown in Fig. 5, our strategy aims to iteratively select best-predicted documents P_A, which are then aggregated back into the labeled data D_L. Thus, the model's classification performance M trained on the updated labeled data D_L is maximized. In the first iteration of the process, an initial model is trained from scratch on D_L containing only artificial data. We ask a human expert to select the visually most reliable predictions. Afterwards, we pass the selected predictions through a post-processing algorithm to reduce noise. Then, combined with artificial data, we use these cleaned predictions P_A as provisional ground truth to fine-tune a new model. This process is repeated until the performance of prediction does not further improve. Ultimately, we can obtain an approximate and high-quality ground truth estimation.

4.2 Post-processing

To enhance prediction quality, we implement a simple and effective post-processing algorithm. The goal is to remove noise, fill gaps, and smooth the borders. To process the result of a prediction, we first separate the labeled image into six binary layers corresponding to *Body* (B), *Decoration* (D), *Text line* (T), *Large drop caps* (L), *Small drop caps* (S), and *Filler* (F) (see Fig. 6b). Every binary layer is processed with a combination of opening and closing morphological operations (see Fig. 6c). Table 1 indicates the kernels used for each layer. When the kernel size of a binary layer exceeds the margin, some preliminary padding is required to avoid errors on the borders.

Table 2 Details of the experimental protocol for training on the UTP-110 dataset

	Training		Validation	
	Real	Artificial	Real	Artificial
M_1	–	360	–	72
M_2 fine-tuned from M_1	30	60	10	20
M_3 fine-tuned from M_2	60	60	10	20
M_4 fine-tuned from M_3	75	60	25	20

Finally, the binary layers are combined to generate the prediction after post-processing (see Fig. 6d).

5 Experiments and evaluation

To thoroughly evaluate the proposed strategy's performance, we have conducted a series of experiments on the two datasets: UTP-110 and CB55. We investigate the effectiveness of the successive fine-tuning iterations. The quantitative and qualitative analysis indicates that our strategy can produce high-quality ground truth for text line detection and classification, as well as for complex layout labeling tasks. We describe the network architecture we used in our experiments in Sect. 5.1. The experimental setup is shown in Sects. 5.2 and 5.3. Our qualitative and quantitative analyses are detailed in Sects. 5.4 and 5.5. Finally, we assess the stability by evaluating the model using 5-fold cross-validation.

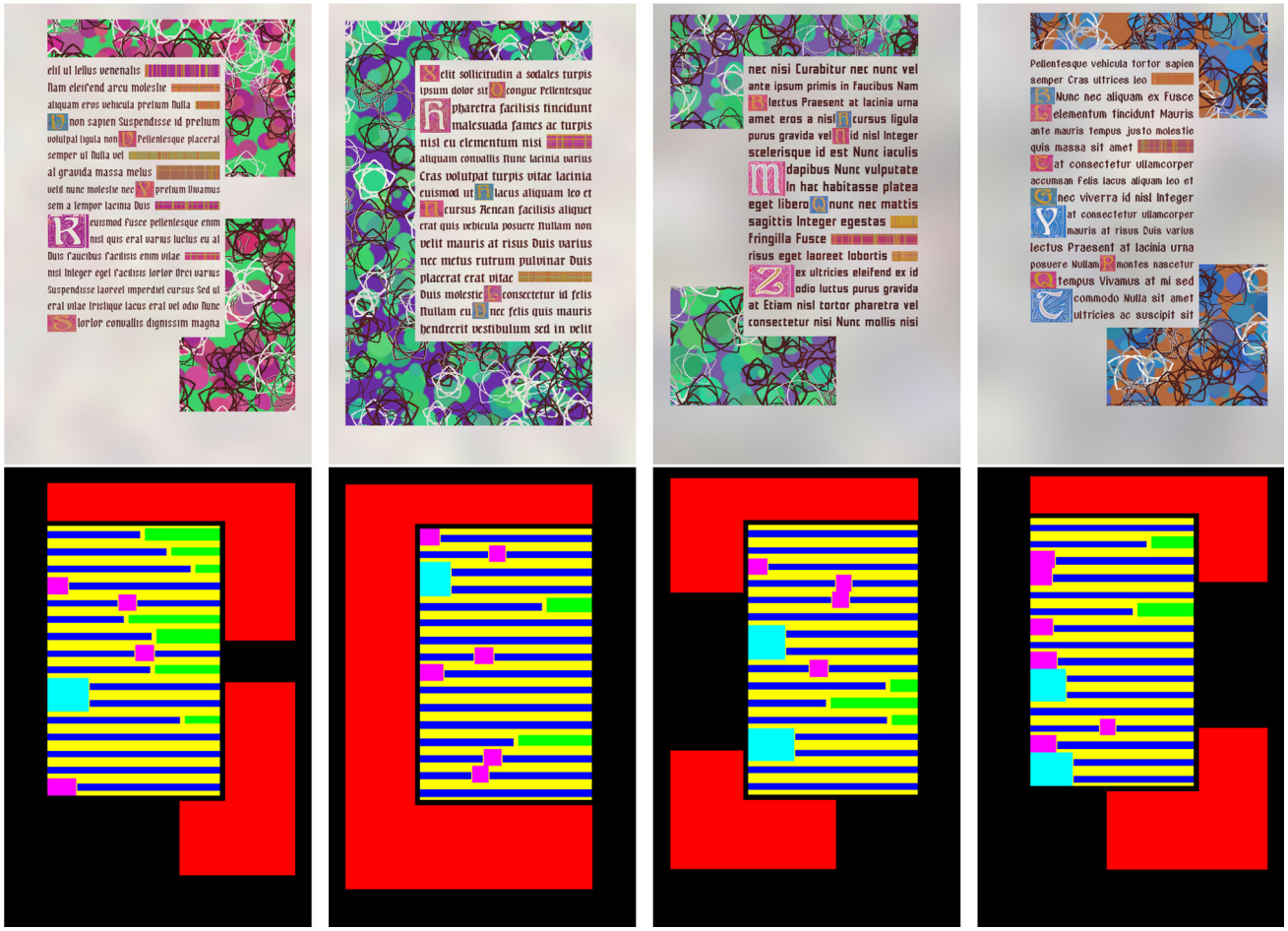


Fig. 4 Samples of artificial images with their corresponding ground truth used in our experiments. In the ground truth image, black represents the *background*, red *decoration*, yellow *body*, blue *text line*, cyan *large drop caps*, magenta *small drop caps*, and green *filler*

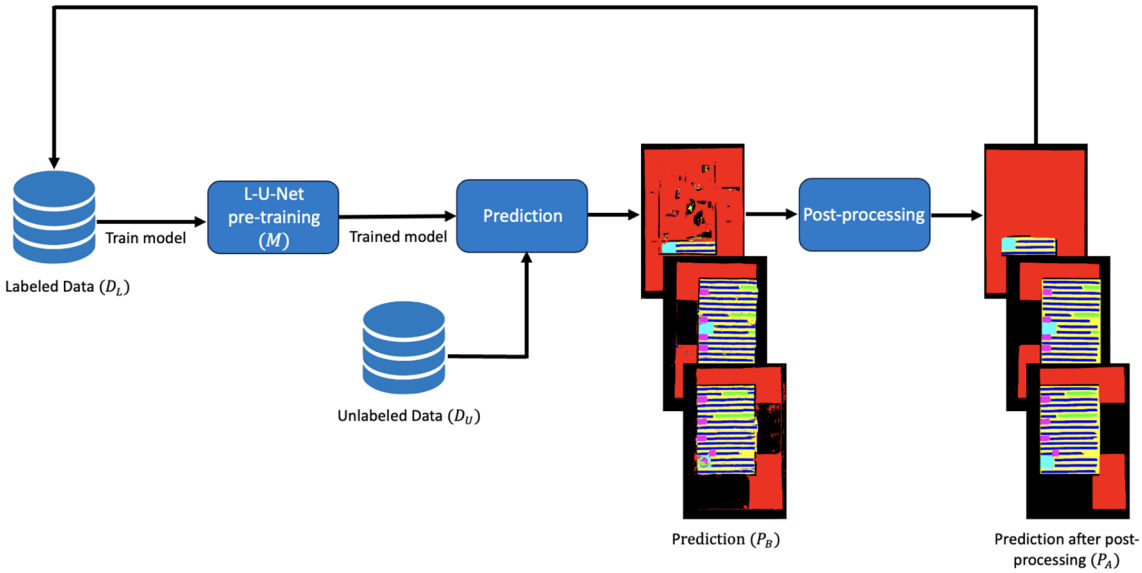


Fig. 5 Iterative process of ground truth generation. The L-U-Net model M is first trained on the labeled dataset D_L, which is then used to predict the unlabeled real data D_U. The predicted data are post-processed, and the best documents are qualitatively selected and aggregated back into the labeled data. The process is repeated

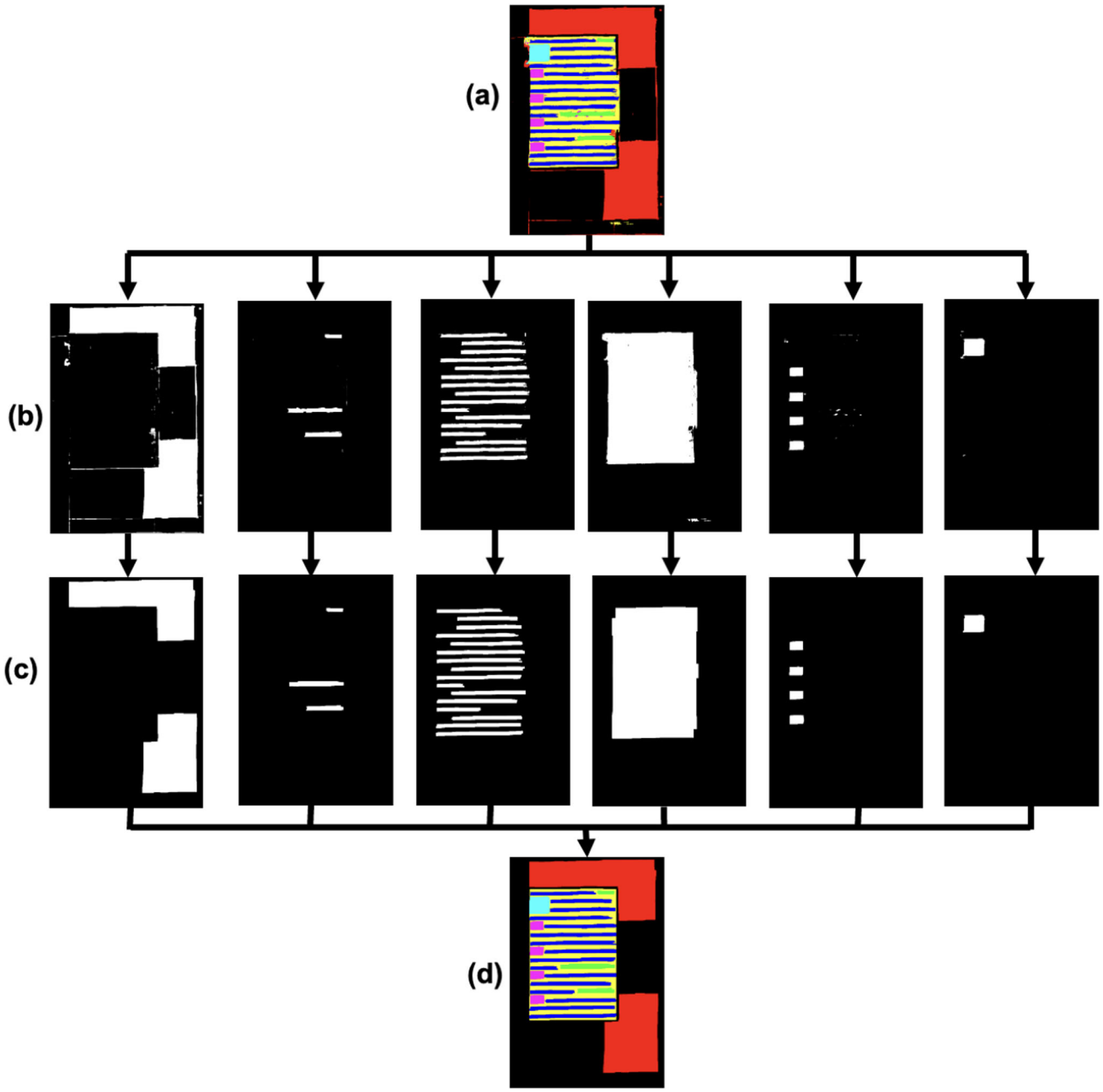


Fig. 6 Post-processing steps for prediction enhancement

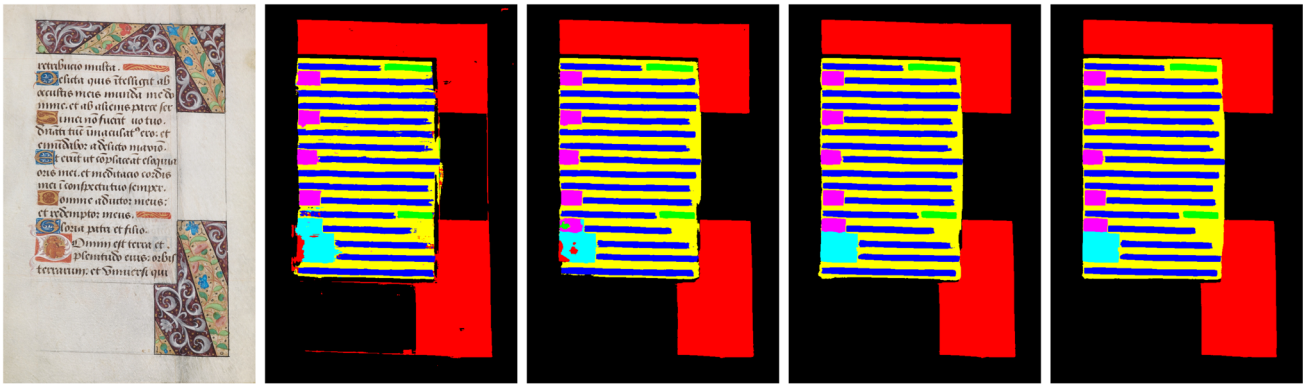
5.1 Network architecture

Our study is based on the L-U-Net neural network, described in [27]. It is a state-of-the-art approach for page segmentation and text line detection tasks. This small U-Net like [32] FCN consists of 65'634 trainable parameters. It comprises an encoder, a decoder, and a last convolution layer for classification. The encoder consists of four dilated blocks. Each block has 16 filters and consists of four dilated convolutions with dilation 1, 1, 2, and 2, respectively. A max-pooling layer follows each block. The decoder comprises four convolutional blocks, each consisting of a straightforward convolution followed by one transposed convolution. The code of the L-U-Net model is publicly available .[3]

5.2 Experimental protocol and metrics

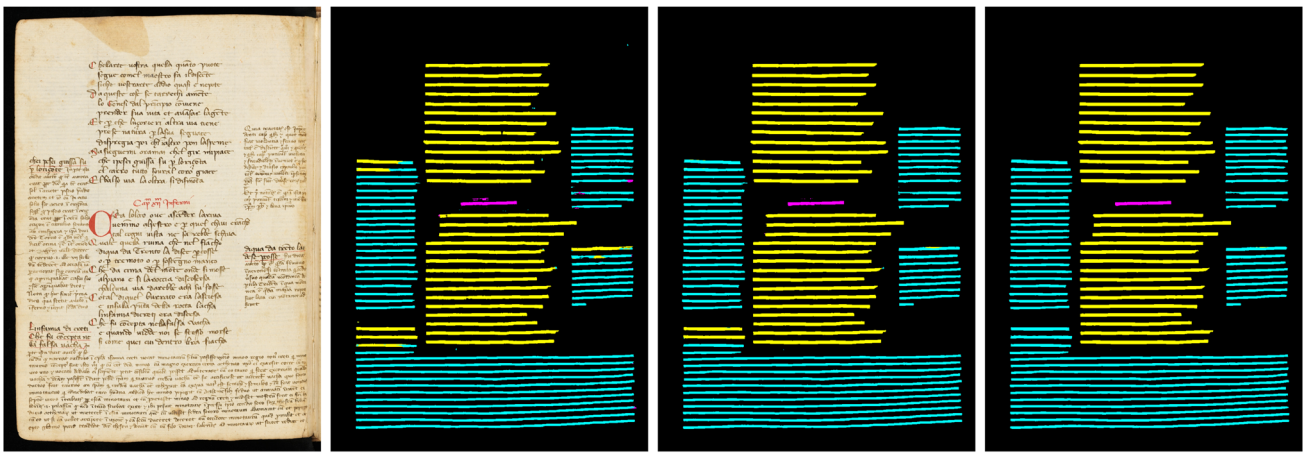
To thoroughly evaluate and validate the performance of the iterative process for approximate ground truth generation,

[3] https://github.com/DIVA-DIA/Layout-Analysis-using-a-Light-CNN



(a) Original image (b) Iteration 1 (c) Iteration 2 (d) Iteration 3 (e) Iteration 4

Fig. 7 Iterative ground truth generation on a UTP-110 image



(a) Original image (b) Iteration 1 (c) Iteration 2 (d) Iteration 3

Fig. 8 Iterative ground truth generation on a CB55 image

we pre-train the neural network model using an increasing number of real data with their labels for training. This allows us to control the quality of the prediction at each iteration. The basic idea relies on the importance of better controlling the training process of neural networks by paying great attention to the training strategy. From our past experiment, we know that the initialization of networks strongly impacts their final performance [33, 34]. This experiment aims to find the best pre-training dataset for the following iteration. In Table 2, we denote the data split of the different iterations. For the first iteration, we used 360 artificial documents for training and 72 artificial documents for validation. The dataset consists of a balanced set of the different fonts shown in Fig. 3. We analyze the impact of the training size at each iteration. The test is carried out on 118 real documents. In addition, as described in Table 2, during the fine-tuning, we drastically reduced the number of artificial data to balance real and artificial data used for training. We notice that the number of real images used for training increases from one iteration to another since the iterative fine-tuning leads to increasingly useful results. One interesting observation from the experiments is that the successive fine-tuning of the model performs better than the training from scratch.

To evaluate the proposed strategy for the text line detection and classification task, we use the CB55 dataset. The training from scratch uses 120 artificial documents for training and 30 artificial documents for validation. For more details about the artificial data generation for the CB55 dataset, please refer to our earlier work [6]. The same split as the UTP-110 dataset is used to fine-tune the other iterations. The test used 120 authentic documents.

To assess the prediction quality across the different iterations and provide an exhaustive quantitative evaluation of the proposed strategy, we computed the standard per-pixel accuracy metrics: IoU, Precision, Recall, and F1-measure. We note that the background and body classes are not considered in the quantitative evaluations.

5.3 Implementation details

All experiments were implemented in PyTorch with Apple M2 GPU, 12 core, and 32GB RAM. During each model's training, we set the batch size to five over a maximum of 150 epochs. ADAM [35] was used as the optimizer, in which epsilon was set at $1e^{-5}$, and the learning rate was set to $1e^{-3}$. Cross-entropy loss served as the loss function. We saved the best model based on the highest Intersection-over-Union (IoU) value achieved on the validation set and used it to make predictions on the test set.

5.4 Qualitative analysis

Fig. 7 illustrates the qualitative differences between the iterations of our proposed strategy on a UTP-110 image without post-processing. The first iteration has proven to be very efficient in recognizing *decoration* and *filler* classes for most images, using only the artificial data. On the other hand, the difficulties encountered were recognizing the *small drop cap* and the *large drop cap* classes and the distinction between these two very similar classes. Also, sometimes, there is misclassification of *line* as *filler* due to the faded ink of the text lines. Successive iterations progressively resolved these difficulties for the majority of the images. As we can see, the performance without post-processing is convincing.

As shown in Fig. 8, the iterative process can accurately detect and classify text lines. The errors caused by the confusion of *main text* and *glosses* are gradually solved for most samples across the iterative process. Through the qualitative comparison presented in Fig. 9, we observe that our strategy can generate approximate ground truth comparable to the manually annotated one.

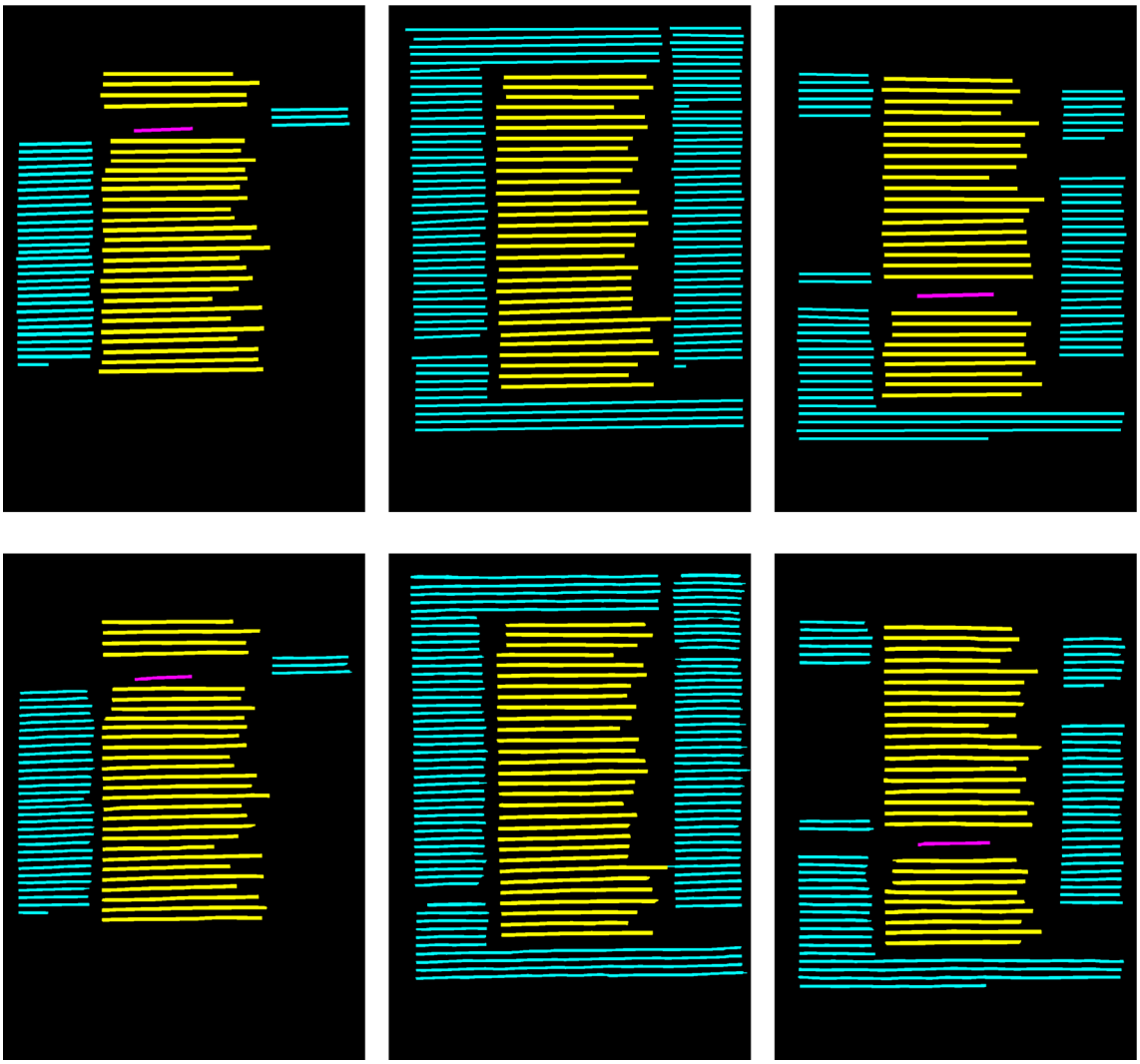


Fig. 9 Comparison of the generated ground truth with our strategy with the original ground truth, manually annotated, for text line detection and classification task on CB55 dataset. Top: original ground truth. Bottom: approximate ground truth

Table 3 Details of the iterative process (IoU) for training UTP-110 dataset

	Initialization	Epochs	Decor	Text	BDC	SDC	Filler
Iter1: M_1	Scratch	137	99.86	98.22	95.37	85.99	88.45
Iter2: M_2	M_1	138	99.98	99.31	97.34	96.88	92.75
Iter3: M_3	M_2	60	99.98	99.45	97.83	**98.44**	94.15
Iter4: M_4	M_3	85	**99.99**	**99.46**	**98.39**	98.35	**94.59**
M_5	Scratch	595	99.98	99.34	96.36	97.03	93.94

Decor: Decoration. SDC: Small drop caps. BDC: Big drop caps
The bold values indicate the best results

Table 4 Details of the iterative process (IoU) for training CB55 dataset

	Initialization	Epochs	Highlight	Main text	Gloss
Iter1: M_1	Scratch	150	87.23	99.05	98.13
Iter2: M_2	M_1	107	99.29	99.88	99.77
Iter3: M_3	M_2	23	**99.75**	**99.91**	**99.82**

The bold values indicate the best results

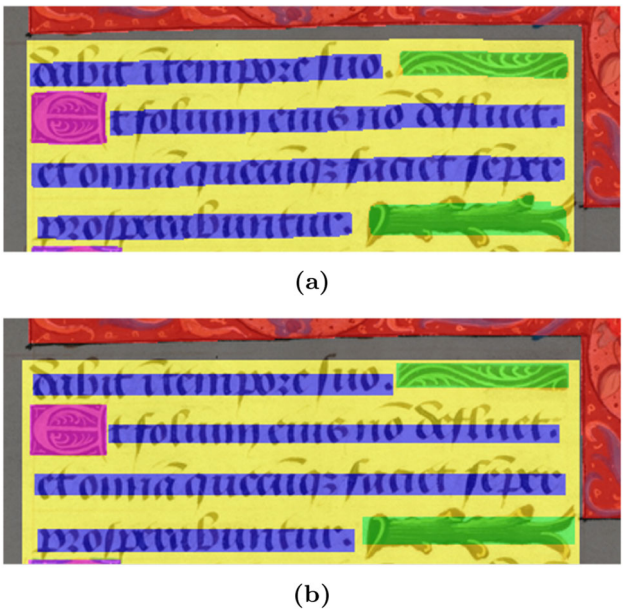


Fig. 10 Comparison between automatically generated approximate ground truth (a) and manually produced ground truth showing inaccurate line delimitation (b)

5.5 Quantitative analysis

To assess the proposed strategy, we believe a quantitative evaluation is necessary. To do this, we faced the problem that no manually labeled ground truth was available. Instead, we selected a representative set of ten top-quality results obtained with a preliminary experiment, on which we applied the post-processing method in a supervised manner, that is, by tuning the parameters individually on each page to obtain optimal results. We considered these results close enough to manually labeled ground truth to use them as the test set for the quantitative evaluation. To guarantee a scientifically sound approach, the test set images were finally discarded for the iterative steps of the final experiment described in this paper.

Consistent with Fig. 7 and Fig. 8, Table 3 and Table 4 show that the iterative process leads to improved results for UTP-110 and CB55 datasets, respectively. Four iterations for UTP-110 and three for CB55 were sufficient to obtain our best predictions.

The quantitative results are obtained using the IoU metric for each class. The last iteration achieves the best result and surpasses its predecessors, especially the first iteration, by a large margin of 4.58% and 5.03% IoU for UTP-110 and CB55 datasets, respectively. Even though the difference between the third and fourth iteration is small, the IoU performance can be further improved by 0.19%. Also, in Table 3, we can observe that the results obtained by $M4$ are higher than those of $M5$, which is trained from scratch. This statement is correct even when the model $M5$ is trained for several training steps equivalent to the entire iterative process. This demonstrates that fine-tuning has significant effects and typically results in better performance.

The quantitative and qualitative analysis of samples provided in Fig. 7, Fig. 8, Table 3 and Table 4 show that our strategy achieves highest performance for complex layout labeling and text detection and classification tasks. The outputs are very promising to prove the robustness of our strategy.

5.6 Analysis of ground truth quality

For our final evaluation, as explained above, we used a test set that was obtained with a previous model. We chose this to align with our goal to study the convergence of the general strategy. However, it is usually considered bad practice because of inherent biases. To complete the study, we compared the quantitative results with manually generated annotations. To do so, we selected the test set of fold 3 of the

Table 5 Comparatif evaluation (IoU) for different type of ground truth

	Decor	Text	BDC	SDC	Filler
Baseline	99.96	99.00	88.93	92.55	93.39
Manually edited GT	99.96	98.32	87.94	91.18	87.41
Improvement	0.00	0.69	1.11	1.48	6.39
Manually produced GT	99.95	97.67	86.28	89.35	84.52
Improvement	0.01	1.34	2.98	3.46	9.50

Table 6 Evaluation of complex layout labeling task on UTP-110 dataset using the generated ground truth

	IoU (%)	Precision (%)	Recall (%)	F1 (%)
Fold1	89.15	91.23	97.08	93.42
Fold2	91.52	94.04	97.08	95.30
Fold3	94.76	97.89	96.72	97.27
Fold4	93.15	95.62	97.05	96.31
Fold5	94.22	97.52	96.45	96.95
Average	**92.56**	**95.26**	**96.88**	**95.85**

The bold values indicate the average results

final experiment, and two different procedures were considered:

- **Manually edited ground truth**: in that case, a user was asked to manually correct the ground truth as used in the baseline experiment using an image editing software with a transparent layer to display the original document.
- **Manually produced ground truth**: in that case, the user just obtained the original document and was asked to produce the ground truth from scratch.

During the procedure, we registered the time it took to complete the work. To edit one page of ground truth, it took 4min 30 s on average, whereas for producing it from scratch, 18min 40 s were needed.

More importantly, we evaluated our final model with the two sets of manual ground truth. The results are reported in Table 5. At first glance, these results confirm the bias since the evaluation with this new data is degraded by 1.8% with the manually corrected ground truth and by 3.4% with the manually produced ground truth. However, having a closer look at the annotations, it appears that the manually produced ground truth also has flows, which are inherent to the limitations of the tools used and the time that can be reasonably invested by the annotator. Figure 10 shows an example comparing approximate and manual ground truth on the same image crop. We can observe the inaccurate line delimitation of the manual ground truth. Therefore, based on visual inspection, we come to the conclusion that the evaluation with the approximate ground truth is currently the most relevant.

5.7 Final discussion

In this section, we estimate the ground truth quality generated by our strategy on L-U-Net neural network [27]. As no distinctive separation in training, validation, and test partition exists, we have undertaken 5-fold cross-validation on the UTP-110 dataset and reported the mean metrics scores. A set of 60, 20, and 20 documents are taken as the training, validation, and test set, respectively. As shown in Table 6, we achieved an IoU, precision, recall, and F1-score of 92.56%, 95.26%, 96.88%, and 95.85%, respectively, which proves the training stability of the model.

The training strategy was also applied and compared with two state-of-the-art models: Adaptive U-Net [36] and DocUFCN [22]. All models have similar results for most classes, but L-U-Net performed significantly better in detecting the big (BDC) and small (SDC) initial classes.

Finally, our evaluation demonstrates its acceptable effectiveness for semantic segmentation tasks where high-quality performance is obtained. Still, it shows that there is room for improvement.

6 Conclusion

In this paper, we have described a strategy that can be used to generate approximate ground truth for document analysis tasks in an effective way. The processing strategy consists of training an initial model with only synthetic data and then fine-tuning it with iterative training steps, during which more and more reliable real data is introduced. With this approach, no initial ground truth is needed and the human effort is significantly reduced compared to manual ground truth annotation. In practice, the required human effort covers three different aspects:

- The first human effort is needed to generate appropriate artificial data. If the term appropriate is interpreted as realistic, the effort could become tremendous and would not be competitive with the manual labeling of real documents. However, for our experiment, we showed that artificial document images reflecting the layout rules are sufficient; there is no need to simulate degradation artifacts. In that case, the programming effort is not high. For this first contribution, we developed the algorithm from scratch, but if an appropriate library provides the appropriate tools, we claim that the programming effort can be reduced to a maximum of one day.
- The second effort that was needed was to develop a post-processing tool to clean and smooth prediction results. The chosen approach is based on logical and morphological operations, where only the parameters have to be tuned correctly. This needs at most a few hours.

- Finally, the third human effort concerns the visual inspection to select the appropriate images providing good quality post-processed predictions that can be used for the next training round. This last task can be achieved in a few seconds per page and does not require high technical expertise; it can be assigned to the end user.

To assess the method, we provided both a qualitative and a quantitative evaluation. The last experiment proves that, in the end, the system converges to a model that is able to generate approximate ground truth, which is not perfect but of acceptable quality. We also estimated the quality of this ground truth with a cross-validation procedure achieving a consistent IoU value of 92.56%. Finally, an additional experiment has shown that automatically generated ground truth is even more precise and reliable than manual annotations produced too quickly and with an inappropriate tool.

In the future, we plan to generalize the proposed strategy and apply it to other documents in the UTP collection with similar layout complexity. The aim is to provide a large dataset of more than a thousand pages with approximate ground truth. Furthermore, we are currently developing an interactive tool to manually correct ground truth in an effective way. With this tool, it will be possible to generate nearly perfect ground truth, allowing us to assess the quality of the approximate ground truth on a substantial dataset.

Author Contributions We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Funding Open access funding provided by University of Fribourg.

Data Availibility 'Not applicable' for that section

Declarations

Conflict of interest The authors declare that they have no Conflict of interest. The authors declare they have no financial interests.

Ethics approval 'Not applicable' for that section

Consent to participate 'Not applicable' for that section

References

1. Lombardi, F., Marinai, S.: Deep learning for historical document analysis and recognition-a survey. J. Imaging **6**(10), 110 (2020)
2. Ma, W., Zhang, H., Jin, L., Wu, S., Wang, J., Wang, Y.: Joint layout analysis, character detection and recognition for historical document digitization. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 31–36. IEEE (2020)
3. Chen, K., Seuret, M., Hennebert, J., Ingold, R.: Convolutional neural networks for page segmentation of historical document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 965–970 IEEE (2017)
4. Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Lee Giles, C.: Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5315–5324 (2017)
5. Renton, G., Soullard, Y., Chatelain, C., Adam, S., Kermorvant, C., Paquet, T.: Fully convolutional network with dilated convolutions for handwritten text line segmentation. Int. J. Doc. Anal. Recogn. **21**, 177–186 (2018)
6. Rahal, N., Vögtlin, L., Ingold, R.: Historical document image analysis using controlled data for pre-training. In: International Journal on Document Analysis and Recognition (IJDAR), pp. 1–14 (2023)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27**, 62 (2014)
8. Pondenkandath, V., Alberti, M., Diatta, M., Ingold, R., Liwicki, M.: Historical document synthesis with generative adversarial networks. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 5, pp. 146–151. IEEE (2019)
9. Bartz, C., Raetz, H., Otholt, J., Meinel, C., Yang, H.: Synthesis in style: Semantic segmentation of historical documents using synthetic data. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3878–3884. IEEE (2022)
10. Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., Ingold, R.: Generating synthetic handwritten historical documents with ocr constrained gans. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16, pp. 610–625. Springer (2021)
11. Vidal-Gorène, C., Camps, J.-B., Clérice, T.: Synthetic lines from historical manuscripts: an experiment using gan and style transfer. In: ViDiScript-Visual Processing of Digital Manuscripts: Workflows, Pipelines, Best Practices at ICIAP 2023 (2023)
12. Shen, H., Li, J., Lin, J., Wu, W.: A multi-level synthesis strategy for online handwritten chemical equation recognition. In: International Conference on Document Analysis and Recognition, pp. 202–217 . Springer(2023)
13. Poddar, A., Dey, S., Jawanpuria, P., Mukhopadhyay, J., Kumar Biswas, P.: Tbm-gan: Synthetic document generation with degraded background. In: International Conference on Document Analysis and Recognition, pp. 366–383. Springer (2023)
14. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)
15. Yim, M., Kim, Y., Cho, H.-C., Park, S.: Synthtiger: Synthetic text image generator towards better text recognition models. In: Inter-

national Conference on Document Analysis and Recognition, pp. 109–124. Springer (2021)
16. Xu, Y., He, W., Yin, F., Liu, C.-L.: Page segmentation for historical handwritten documents using fully convolutional networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 541–546. IEEE (2017)
17. Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., Ingold, R.: Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 471–476. IEEE (2016)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Diem, M., Kleber, F., Fiel, S., Grüning, T., Gatos, B.: cbad: Icdar2017 competition on baseline detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1355–1360. IEEE (2017)
20. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. Int. J. Doc. Anal. Recogn. **22**(3), 285–302 (2019)
21. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 7–12. IEEE (2018)
22. Boillet, M., Kermorvant, C., Paquet, T.: Multiple document datasets pre-training improves text line detection with deep neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2134–2141. IEEE (2021)
23. Jian, S., Kaiming, H., Shaoqing, R., Xiangyu, Z.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 770–778 (2016)
24. Vézina, H., Bournival, J.-S.: An overview of the balsac population database. SOWING, 183 (2020)
25. Boillet, M., Bonhomme, M.-L., Stutzmann, D., Kermorvant, C.: Horae: an annotated dataset of books of hours. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, pp. 7–12 (2019)
26. Grüning, T., Labahn, R., Diem, M., Kleber, F., Fiel, S.: Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 351–356. IEEE (2018)
27. Rahal, N., Vögtlin, L., Ingold, R.: Layout analysis of historical document images using a light fully convolutional network. In: International Conference on Document Analysis and Recognition, pp. 325–341 . Springer(2023)
28. Banerjee, A., Biswas, S., Lladós, J., Pal, U.: Swindocsegmenter: An end-to-end unified domain adaptive transformer for document instance segmentation. arXiv preprint arXiv:2305.04609 (2023)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
30. Shen, Z., Zhang, K., Dell, M.: A large dataset of historical japanese documents with complex layouts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 548–549 (2020)
31. Allegretti, P., Chen, S., Hu, X., Yang, J.: Catalogo dei codici italiani, cod. bodmer 55. Corona Nova, 44–47 (2003)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241. Springer (2015)
33. Seuret, M., Alberti, M., Liwicki, M., Ingold, R.: Pca-initialized deep neural networks applied to document image analysis. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 877–882. IEEE (2017)
34. Alberti, M., Seuret, M., Pondenkandath, V., Ingold, R., Liwicki, M.: Historical document image segmentation with lda-initialized deep neural networks. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, pp. 95–100 (2017)
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
36. Mechi, O., Mehri, M., Ingold, R., Amara, N.E.B.: Text line segmentation in historical document images using an adaptive u-net architecture. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 369–374. IEEE (2019)