International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-024-00497-4

ORIGINAL PAPER

In-domain versus out-of-domain transfer learning for document layout analysis

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

Received: 13 May 2024 / Revised: 18 July 2024 / Accepted: 2 August 2024

Abstract
Data availability is a big concern in the field of document analysis, especially when working on tasks that require a high degree of precision when it comes to the definition of the ground truths on which to train deep learning models. A notable example is represented by the task of document layout analysis in handwritten documents, which requires pixel-precise segmentation maps to highlight the different layout components of each document page. These segmentation maps are typically very time-consuming and require a high degree of domain knowledge to be defined, as they are intrinsically characterized by the content of the text. For this reason in the present work, we explore the effects of different initialization strategies for deep learning models employed for this type of task by relying on both in-domain and cross-domain datasets for their pre-training. To test the employed models we use two publicly available datasets with heterogeneous characteristics both regarding their structure as well as the languages of the contained documents. We show how a combination of cross-domain and in-domain transfer learning approaches leads to the best overall performance of the models, as well as speeding up their convergence process.

Keywords Document analysis · Layout segmentation · Semantic segmentation · Transfer learning

1 Introduction

Document layout analysis refers to the task of identifying the different semantically meaningful regions of a document page and grouping them based on a set of pre-defined classes such as text, decorations, and titles. Being able to understand the layout of a document's page is of paramount importance for both humanities scholars and computer scientists, as it represents the first step towards the extraction and further analysis of their contents, making it easier to perform other tasks such as optical character recognition [1], automatic text transcription [2], writer identification [3] and text-line segmentation [4–6]. In recent years, there has been an increasing effort to try to automate this process; however, unlike printed, well-structured, documents for which very promising results have been obtained, performing this task on heavily edited, handwritten documents has proven to be particularly challenging, especially when it comes to ancient manuscripts. These are characterized by various degrees of degradation, inconsistent conditions in the capture of the instances of datasets, and large amounts of additions and corrections that are heavily intertwined with the main text. These characteristics make it nearly impossible to rely on the most popular techniques adopted for printed documents, which typically rely on bounding boxes to group together the different elements of the page layout [7].

The alternative is represented by the adoption of pixel-level segmentation maps that allow for the higher degree of precision needed to overcome this problem. However, compared to typical segmentation tasks in a natural environment, the areas representing the different regions of ancient manuscript layouts are typically characterized by very small

✉ Axel De Nardin
axel.denardin@uniud.it

Silvia Zottin
zottin.silvia@spes.uniud.it

Claudio Piciarelli
claudio.piciarelli@uniud.it

Gian Luca Foresti
gianluca.foresti@uniud.it

Emanuela Colombi
emanuela.colombi@uniud.it

[1] Department of Mathematics, Informatics and Physics, University of Udine, Via delle Scienze, 206, 33100 Udine, UD, Italy

[2] Department of Humanities and Cultural Heritage, University of Udine, Vicolo Florio, 2, 33100 Udine, UD, Italy

Published online: 19 August 2024

regions with jagged edges that make it really time-consuming to produce the corresponding segmentation maps [8]. Furthermore, an understanding of the content of the pages is typically needed to make sure that the different regions are classified correctly, meaning that the segmentation process must be supervised by an individual with appropriate domain knowledge. Both of these requirements lead to a scarcity in the availability of data for this type of task, with the majority of available datasets providing at most a couple dozen images as the corresponding training set. Some examples of such datasets are represented by Diva-HisDB [9], Bukhari [10] and the very recent U-DIADS-BiB [11].

In the past few years, this problem has been tackled by various authors [12–14], who developed a set of few-shot-learning-oriented frameworks specifically aiming at leveraging the small amount of data available to generate more and more accurate predictions for the task at hand, producing results that are on par or even surpass previously available state-of-the-art approaches that relied on much more data.

In the present paper, we tackle the problem from another point of view by exploring different transfer learning approaches as a way to make good use of alternative data sources to pre-train our models. In particular, we analyze the effectiveness of different initialization approaches for the encoder component of a selected semantic segmentation model, such as training from scratch, in-domain transfer learning, cross-domain transfer learning, and a combination of these last two. Specifically:

- we provide a thorough analysis of the effects of different initialization and transfer learning strategies on the performance of the segmentation network, particularly when working in a low data setting,
- we show how the features learned from pre-training on large-sized general-purpose datasets are generally not effectively transferable to the specific domain of document image analysis.
- we show, on the contrary, how relying on domain-specific data for pre-training, even in a small amount, leads to a substantial improvement in the performance on the downstream task, especially when working with few training instances on the target dataset.

The rest of the paper is organized as follows: in Sect. 2 a review of the related works in this field of research is provided, an overview of the methodology adopted is provided in Sect. 3 and a discussion of the experimental setup and the corresponding results is reported in Sect. 4. Finally, in Sect. 5 we summarize our findings and propose a direction for future works.

2 Related works

The scarcity of extensively labeled data in the field of ancient manuscript analysis can be attributed to the specialized knowledge and substantial time and financial resources required for its creation, particularly when dealing with documents featuring intricate layouts. Consequently, a logical progression involves the development of systems capable of delivering commendable performance with limited annotated data. Nevertheless, the literature currently provides only a limited number of works showcasing such systems.

In [15], a few-shot learning technique named deep and syntax, designed for segmenting historical handwritten registers, is presented. Few-shot learning, a paradigm enabling models to generalize from a limited number of examples, proves beneficial in scenarios with restricted annotated data. The suggested method utilizes a hybrid system that relies on recurrent patterns to delineate individual records. This hybrid system integrates U-shaped neural networks, typically employed in image segmentation tasks, with logical rules such as filtering and text alignment.

Another example of a few-shot learning strategy for document layout segmentation has been introduced in [16]. In this approach, only two ground truth images per manuscript are employed to train the segmentation model, yielding results comparable to supervised models that currently represent the state-of-the-art for this task. The proposed framework combines a novel data augmentation method, with a segmentation refinement module employing a traditional computer vision approach for local thresholding. This integration fully exploits the limited dataset available while still achieving competitive performance compared to other supervised methods, as highlighted in their in-depth and analytical work [12].

A further, more recent approach is the one presented in [14], where a one-shot learning approach is introduced for the layout segmentation of ancient Arabic documents. In this paper, the authors introduce an efficient framework that, despite being trained on only one labeled page per manuscript, achieves state-of-the-art performance compared to other approaches tested on a challenging dataset of ancient Arabic manuscripts. This method consists of three main components, a semantic segmentation backbone, a dynamic instance generation module, and a segmentation refinement module, and aims to overcome the limitation of requiring extensive manual labeling for training machine learning models in this field.

Finally, in [13] the authors tackle the challenge of limited ground truth availability by proposing an unsupervised deep learning approach for page segmentation. Their method involves the use of a Siamese neural network to differentiate between patches based on quantifiable properties, with a specific emphasis on the count of foreground pixels. The

goal is to ensure that spatially adjacent patches demonstrate similarities in their measurable characteristics. Following the training of the network, the acquired features are then utilized for the task of page segmentation.

Transfer learning approach Transfer learning pre-trained deep networks is an approach to derive advantages from the representations learned on a large and general-purpose database while having relatively few examples to train a model [17, 18]. In the literature, numerous works employ transfer learning techniques as they have wide application in many domains, such as in the medical [19, 20], biometric [21], agricultural [22], industrial [23] and robotic [24] fields. Conversely, in the field of ancient document layout analysis, the effectiveness of transfer learning techniques has not been extensively explored, as there are only a few works in the literature addressing this topic.

In an investigation conducted in [25], it was determined that the outcomes of the semantic segmentation problem, whether employing training from scratch or cross-domain learning from a pre-existing model, are contingent upon the specific characteristics of the test dataset. The accuracy of segmentation varies significantly across datasets, irrespective of the model architecture or training methodology employed. To arrive at these findings, the researchers initiated their model's encoder with pre-trained weights from ImageNet and systematically compared its performance with models trained from scratch using an ancient document dataset.

In the most recent study [26], the authors present an overview of domain-specific transfer learning for document layout segmentation. They demonstrate that utilizing document-related images for pre-training yields consistently enhanced performance and faster convergence compared to training from scratch or relying on a large, general-purpose dataset like ImageNet.

One limitation common to both of these works, however, is that they explore only the use of ImageNet as a potential out-of-domain dataset for pre-training, which focuses on the classification of instances at an image level instead of a pixel one.

In this paper, we extend the examination of the efficacy and potential advantages of employing different transfer learning strategies for the layout analysis of ancient manuscripts in contrast to training the model from the ground up. Specifically, we provide a more in-depth analysis of what kind of data represents the best for the pre-training process in the context of document layout segmentation. For this reason, we introduce the use of an additional cross-domain dataset, namely MS-COCO, specifically tailored towards semantic segmentation in images, and we also expand the analysis to hybrid strategies involving the combination of both cross-domain and in-domain data for pre-training our model.

3 Methods

3.1 Segmentation architecture

As the semantic segmentation model for our experiments, we opted for the popular DeepLabv3 [27] architecture. DeepLabv3 is a ResNet-based architecture that employs atrous (dilated) convolutions in cascade or in parallel with different dilation levels. This approach allows for retaining a larger spatial resolution for the feature maps throughout the network architecture compared to models relying heavily on striding and pooling layers. The key aspect that makes atrous convolutions effective in the context of semantic segmentation tasks is that they allow for the creation of deeper neural network architectures, while at the same time providing output feature maps that are larger than those of a traditional deep CNN architecture and without any increase in the amount of computation needed. While we are aware there are more recent and sophisticated models for semantic segmentation, the focus of this work is not on obtaining the best possible results or on setting a new state-of-the-art but on providing a better understanding of which transfer learning strategies lead to the best results when all the other conditions are kept as consistent as possible. For this reason, we chose DeepLabv3 as it represents a well-tested, reliable, and easily accessible architecture, which makes it easier to reproduce our results.

3.2 Transfer learning strategies

Transfer learning is referred to as the process through which the knowledge learned from a task (upstream) is re-used to boost performance on a related task (downstream). In this work, we explore three different pre-training strategies characterized by the different types of upstream data employed to initialize the segmentation model. Furthermore, we studied the effects of two transfer learning pipelines to adapt the trained models to the target data. An overview of the whole setup is provided in Fig. 1.

3.2.1 Upstream data sources

Cross-domain transfer learning As previously mentioned, in our analysis, we explore three different pre-training strategies. The first one is represented by cross-domain pre-training, meaning that the dataset characterizing the upstream task is substantially different from the one characterizing the downstream task, either by content, objective, or both. This has been one of the most popular approaches for transfer learning since the release of large-scale general-purpose datasets such as ImageNet1K, and it has proven to be very effective in a wide variety of application fields. The main advantage of this strategy is that this type of dataset is

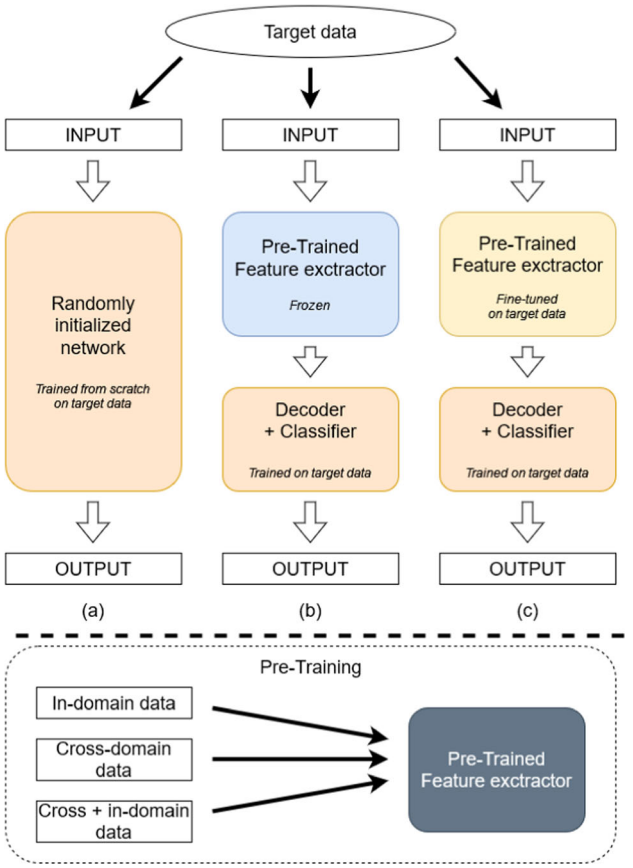


Fig. 1 Visual representation of the different training strategies we explored in this study. **a** represents the traditional training from scratch approach which we used as our baseline, **b** shows a transfer-learning approach in which the feature extractor component of the network is frozen during the training process on the target downstream task and finally **c** shows a further transfer-learning strategy in which the entire model is fine-tuned on the downstream task dataset. For both (**b**) and (**c**) we explored the use of in-domain and cross-domain data to pre-train the selected model as well as a combination of the two

very easily available in an already structured format nowadays. Furthermore, deep learning models pre-trained on this kind of dataset are provided by many open-source libraries, making them easily accessible. However, when working on downstream tasks characterized by data not belonging to the domain of natural images, it becomes harder to learn effectively transferable features through this approach. A common example is represented by the medical imaging field [28].

In-domain transfer learning The second pre-training strategy we analyze is the in-domain one, meaning that, for the pre-training of our model, we employ a dataset that shares the same domain as the one we will use for the downstream task we are interested in, even though the specific instances are different. While the features learned through this approach are typically more applicable to the downstream task, making it more reliable for domain-specific applications, the data needed is not always easily available, as their limited scope is a deterrent to the very time-consuming task of building a specific dataset for this purpose.

Combining cross-domain and in-domain transfer learning The final strategy we explore is represented by the combination of cross-domain and in-domain pre-training. The way we combine the two previously described approaches is by performing an initial training of our model on large-scale cross-domain datasets to learn a general set of features. As a second step, we perform a fine-tuning process on the in-domain dataset, in our case U-DIADS-Bib, to learn a set of domain-specific features. During this second phase, we don't freeze any of the model's weights allowing the feature-extractor to be tailored to the specific nature of the features characterizing the in-domain dataset. Other than the potential improvement in performance that this approach could lead to, we believe it could also represent a way to reduce the amount of domain-specific data needed compared to relying exclusively on an in-domain transfer learning strategy.

3.2.2 Transfer learning pipelines

The two transfer learning pipelines we studied are shown in Fig. 1b, c and differ by the way the feature extraction module of the pre-trained model is employed. In (b), this module is frozen when performing the training process on the downstream task, while in (c), it's fine-tuned on the target data, allowing it to learn its peculiarities and therefore improve its effectiveness on the target task at the cost of a higher computational effort during the training process. An important thing to notice is that in both pipelines, the weights of the decoder and classifier modules of the model are initialized randomly before the training step on the downstream task. When working with either cross-domain or in-domain data individually as the input to pre-train our model, we explore the effects of both pipelines to better understand their respective effects on the final performance. When combining both types of data, we perform a first pre-training step on the cross-domain data following pipeline (a), then fine-tune the entire model on the in-domain data of the upstream task following pipeline (c) and, finally, we perform the transfer learning step on the target downstream data by once again exploring the effects of either freezing the feature extractor or fine-tuning the whole model.

Table 1 provides a schematic overview of all the dataset presented in this section.

3.3 Pre-training datasets

In this section, we will give a brief description of the three datasets employed to perform the pre-training of the selected model architecture. We decided to analyze the effectiveness of the features learned through training on datasets with different characteristics. In particular, we selected two

Table 1 Compact overview of key information regarding the datasets used for our analysis

Dataset	# Images	# Classes	#Images type	Task	Role
ImageNet1K	1.281.167	1000	Natural	Classification	Out-domain source
COCO	∼200.000	80	Natural	Segmentation	Out-domain source
U-DIADS-Bib	200	6	Documents	Segmentation	In-domain source
DIVA-HisDB	150	4	Documents	Segmentation	Target
Bukhari	32	3	Documents	Segmentation	Target

large cross-domain datasets, namely ImageNet-1K [29] and COCO [30] which are employed, respectively, for classification and semantic segmentation tasks, and represent our cross-domain data sources, as well as a small, recently published in-domain dataset that focuses on semantic segmentation specifically applied to the layout analysis of ancient manuscripts, called U-DIADS-Bib [11].

3.3.1 ImageNet-1k

ImageNet-1k is a hierarchically structured dataset consisting of 1.281.167 natural images focused on the classification task, with its instances being organized into 1000 different categories. This dataset represents probably the most popular resource for pre-training computer vision models, and it has been successfully employed in a wide variety of application fields since its release in 2012.

3.3.2 COCO

The COCO (Common Objects in Context) dataset, introduced in 2014, contains over 200k labeled images covering 80 different object categories, which appear in around 1.5 million individual instances in the dataset images. While this dataset is much smaller than the previously presented ImageNet-1k, its main advantage is that it focuses mainly on the tasks of object detection and semantic segmentation, making it more relatable to the downstream task, we are analyzing in this work.

3.3.3 U-DIADS-Bib

U-DIADS-Bib[1] [11] is a recently published dataset focusing specifically on the layout segmentation of ancient manuscripts. It consists of a total of 200 images, representing the pages of 4 different manuscripts written either in Latin or Syriac. Each of the dataset instances can contain up to 6 different segmentation classes, namely: the main text of the manuscript, decorations, titles, chapter headings, additional paratexts, and finally the background of the pages. The key characteristics of this dataset are the improved precision in the definition of the ground truths compared to previously available ones as well as its heterogeneity, which made it an ideal candidate for our analysis of in-domain transfer learning. In fact, the instances of this dataset are characterized by high variability in the layout structure, and significant interclass similarity, which forces the segmentation model to learn subtle differences between the different layout components defining the 6 layout classes and a combination of textual information and pictures, allowing the model to generalize to learn a set of features that is not exclusively focused on textual structures. Fig. 2a shows a sample page of 2 of the documents characterizing the dataset, together with the corresponding segmentation maps, highlighting its key features.

3.4 Evaluation datasets

In the following section, we present the two datasets selected to evaluate the different initialization strategies for our segmentation model. We relied on the two most popular datasets for layout analysis of handwritten documents, which are the DIVA-HisDB dataset [9] and the Bukhari dataset [10].

3.4.1 DIVA-HisDB

The first dataset selected to evaluate our model on is the DIVA-HisDB dataset [9], a historical document dataset consisting of a total of 150, high-resolution, pixel-annotated pages coming from three different medieval manuscripts, identified as CSG18, CSG863, and CB55 and characterized by complex and heterogeneous layouts as well as different levels of degradation. Each of the pages can contain up to four different segmentation classes, categorized as main text, comments, decorations, and background. For each of the documents, 20 images are typically reserved for training, 10 for validation, and 20 more for the testing process. Like U-DIADS-Bib, the Diva-HisDB dataset also consists of both textual and graphical layout components, allowing us to understand if the model correctly learns to distinguish between the two as well as discriminating between the three different textual classes. Furthermore, compared with the U-DIADS-bib dataset, it is characterized by very intertwined layout components, with paratexts that heavily overlap with the main text sections of the document (Fig. 2b). This spe-

[1] https://ai4ch.uniud.it/udiadsbib.

Fig. 2 Samples of a set of selected instanced coming from both the source in-domain dataset U-Diads-Bib (**a**) as well as the two target evaluation datasets Diva-HisDB (**b**) and Bukhari (**c**) together with the respective ground truth segmentation maps

cific aspect makes it necessary for the model to provide precise segmentation to avoid misclassification between the two classes..

3.4.2 Bukhari dataset

The second dataset we selected for the evaluation process in this paper is the one presented by Bukhari et al. [10], which represents the most popular one for the task of document layout segmentation on historical Arabic manuscripts. It consists of 32 images, each representing a page from one of three different Arabic historical manuscripts. Out of all the samples, 24 are typically used for the training process, while the remaining 8 are used for testing. A peculiar characteristic of this dataset is that the text presents different orientations within the page on which it is written (Fig. 2c), thus allowing us to test our model robustness in this type of edge scenario, which could represent an important feature for tasks where this is a common occurrence, such as text de-wrapping [31, 32].

4 Experiments

4.1 Training setup

As previously stated for all our experiments we relied on the DeepLabv3 architecture for semantic segmentation which was trained following one of the three pipelines reported in Fig. 1. The training process on the target datasets was performed on a total of 200 epochs, using the Adam optimizer with a learning rate of $1e^{-}3$ with a batch size of 20. Furthermore, an early stop condition was introduced after the first 50 epochs in case the model loss on the validation set didn't decrease over the last 20 epochs. All the instances of the employed datasets were resized to 672×1008, keeping the aspect ratio of the original images intact to avoid artifacts.

Losses The loss we adopted to train our models is a combination of the dice loss and the weighted cross-entropy loss, where the weight of each class is represented by the square root of the inverse frequency of that class in the instances belonging to the corresponding document, as proposed in [12] to account for the substantial class imbalance characterizing all the document datasets, detailed in Table 3.

Table 2 Tabular overview of the performances obtained by fine-tuning the selected model on the target datasets document classes when initialized with different strategies

Source dataset	None (Trained from scratch)	CB55				CSG18				CSG863				Bukhari			
		Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
		0.552	0.652	0.439	0.576	0.571	0.833	0.512	0.653	0.614	0.719	0.504	0.647	0.534	0.607	0.431	0.562
Frozen encoder	ImageNet1K	***0.502***	***0.622***	***0.399***	***0.533***	***0.515***	***0.626***	***0.425***	***0.556***	***0.520***	***0.631***	***0.428***	***0.562***	***0.487***	***0.534***	***0.385***	***0.505***
	COCO	***0.526***	***0.625***	***0.410***	***0.542***	***0.540***	***0.639***	***0.441***	***0.574***	***0.560***	***0.667***	***0.464***	***0.602***	***0.501***	***0.586***	***0.401***	***0.529***
	U-DIADS-Bib	0.581	0.717	0.484	0.625	0.582	***0***.825	0.524	0.663	0.619	0.778	0.539	0.680	0.546	0.648	0.447	0.582
	ImageNet1K+U-DIADS-Bib	0.573	0.723	0.481	0.624	0.583	**0.836**	0.527	0.667	0.616	0.791	0.542	0.682	0.556	0.675	0.458	0.597
	COCO+U-DIADS-Bib	**0.596**	**0.746**	**0.508**	**0.651**	**0.606**	***0***.830	**0.545**	**0.683**	**0.652**	**0.798**	**0.567**	**0.705**	**0.569**	**0.705**	**0.475**	**0.614**
Finetuned encoder	ImageNet1K	0.575	0.730	0.483	0.627	0.578	***0***.831	0.520	0.660	0.619	0.773	0.533	0.675	0.565	0.674	0.467	0.604
	COCO	0.619	0.783	0.536	0.680	0.608	**0.886**	0.560	0.698	0.654	0.851	0.588	0.726	**0.599**	0.720	**0.504**	**0.644**
	U-DIADS-Bib	0.592	0.754	0.500	0.645	0.590	0.853	0.537	0.676	0.627	0.817	0.557	0.697	0.578	0.700	0.481	0.621
	ImageNet1K+U-DIADS-Bib	0.580	0.765	0.495	0.640	0.588	0.849	0.533	0.673	0.626	0.810	0.551	0.692	0.571	0.692	0.474	0.613
	COCO+U-DIADS-Bib	**0.626**	**0.806**	**0.546**	**0.690**	**0.613**	0.882	**0.564**	**0.702**	**0.657**	**0.862**	**0.597**	**0.733**	**0.599**	**0.728**	**0.504**	**0.644**

The best and second-best results for each transfer learning pipeline are reported in bold and underlined respectively, while the bolditalic values represent the instances in which a pre-trained model performed worse than the baseline model trained from scratch

Table 3 Classes distribution (%) at pixel level for each manuscript class of the three datasets employed for the analysis

Dataset	Document class	BG	Paratext	Decor.	Main text	Title	Chapter headings	Total
U-DIADS-Bib	Lat 2	92.80	0.10	1.50	4.70	0.40	0.50	100
	Lat 14,396	89.20	0.10	2.00	7.60	0.50	0.60	100
	Lat 16,746	88.00	0.30	3.00	7.80	0.10	0.80	100
	Syr 341	85.10	0.20	2.80	11.90	0.10	0	100
Diva-HisDB	CB55	82.41	8.36	0.55	8.68	–	–	100
	CS18	85.16	6.78	1.47	6.59	–	–	100
	CS863	77.82	6.35	1.83	14.00	–	–	100
Bukhari	Bukhari	86.07	4.71	–	9.22	–	–	100

4.2 Metrics

To evaluate the different initialization strategies we relied on 4 popular metrics in the field of document layout segmentation, namely precision, recall, Intersection over Union (IoU) and F1-Score, which are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

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

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

each metric was calculated individually for each document class of the two target datasets. Furthermore, a macro average of the scores obtained for the different segmentation classes was performed, to ensure that each of them contributes equally to the final score.

4.3 Results

4.3.1 Transferability of the learned features

Table 2 shows the performances, in terms of the four selected evaluation metrics, obtained by our model when initialized employing different strategies and trained on the target dataset following each of the three aforementioned pipelines. In this table, the best and second-best performing systems for each transfer learning pipeline are highlighted in bold and underlined, respectively. Furthermore, we mark in red the scores achieved by pre-trained models that don't improve over the random initialization baseline. From this analysis, we can observe that, pre-training on out-of-domain datasets with no further fine-tuning of the encoder module, consistently leads to a drop in performance compared to random initialization, meaning that the features learned during the training on the upstream task have no real overlap with the ones needed to perform the downstream task. In contrast, pre-training on U-DIADS-Bib, representing the in-domain source dataset, consistently leads to an increase in performance compared to the baseline, with improvements in the scores for the selected metrics going from 1 to 6% for all the classes of the target datasets. The only exception is represented by the score obtained on the recall metric for the CSG18 document class of the DIVA-HisDB dataset. On the other hand, when fine-tuning the whole model on the downstream task data, every pre-training strategy consistently leads to an improvement over the baseline approach. This implies that even if the source and target datasets belong to very different domains, pre-training on the former still leads to a better starting point for the training of the latter compared to random initialization.

Furthermore, we can clearly observe how combining the cross-domain and in-domain initialization strategies consistently leads to the best overall results, with the setup represented by pre-training on COCO and fine-tuning on the DIVA-HisDB dataset outperforming all the other initialization strategies on all the selected metrics, regardless of the way the encoder weights are treated during the final training step, with the only exception of the recall for the CSG18 class. It is interesting to observe how, even though pre-training on natural image datasets doesn't provide any real benefit over random initialization when used individually while freezing the feature extractor module of the segmentation model, it actually represents a valid strategy when combined with a fine-tuning step on an in-domain dataset such as U-DIADS-Bib, especially when using COCO as the source dataset. In fact, we can observe how the hybrid strategy involving the combination of the COCO and U-DIADS-Bib datasets consistently achieves better results compared to relying exclusively on the latter. This means that, even in this scenario, the pre-training on COCO still leads to a more robust initialization than a random one. On the other hand, the results obtained by combining ImageNet1K and U-DIADS-

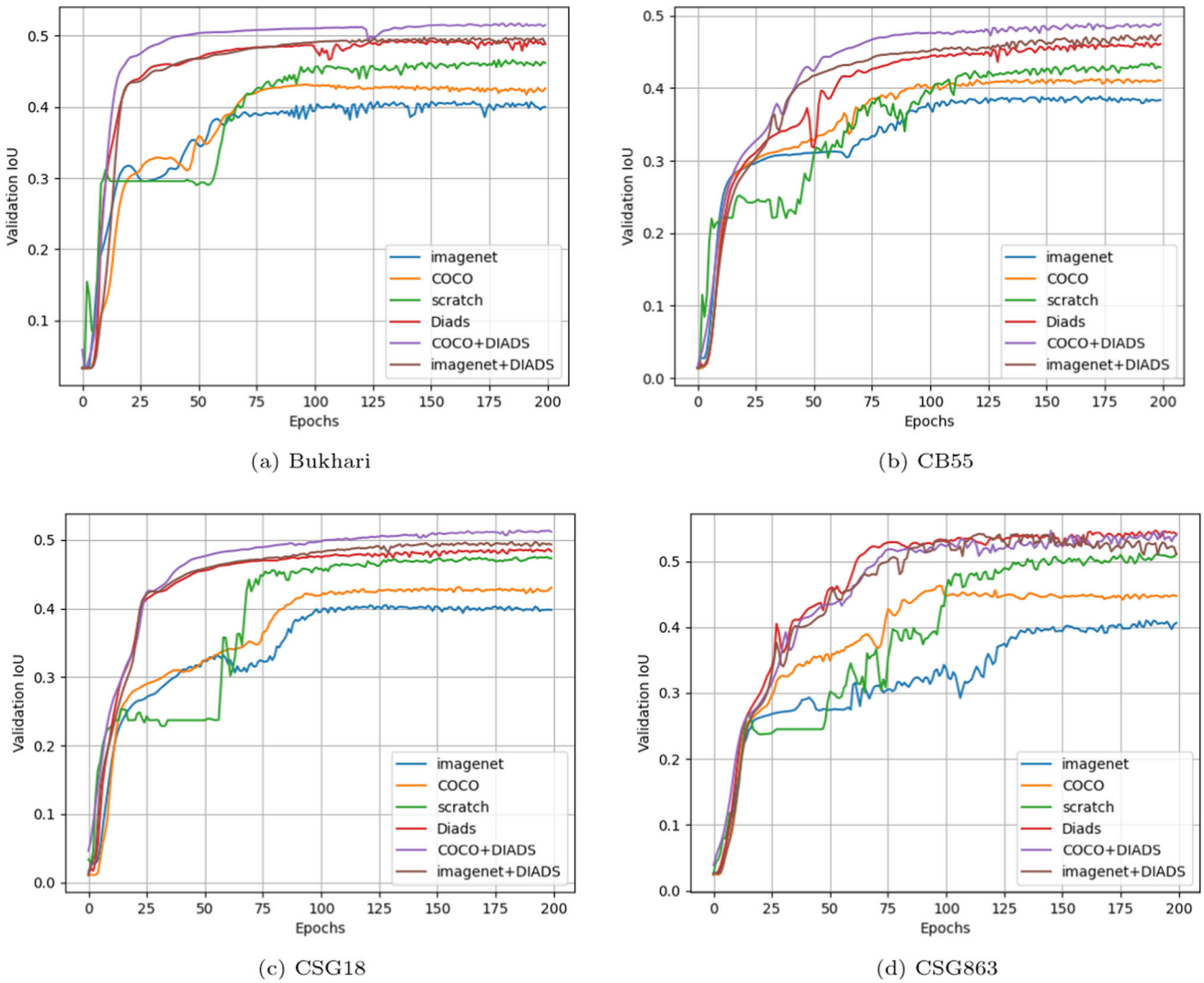


Fig. 3 Learning curves representing the evolution of the IoU scores for the four target document classes using different initialization strategies and keeping the encoder frozen at the time of fine-tuning on the target data

Bib are overall comparable to the simple in-domain transfer learning strategy, with only slight improvements on some of the metrics.

We can also observe how, regardless of the pre-training strategy employed, fine-tuning the whole model on the target dataset consistently leads to better performance compared to training only the decoder and segmentation modules while keeping the encoder module frozen. This behavior is expected as the model can more effectively learn a set of bespoke features on the target dataset following this approach.

4.3.2 Impact on convergence time

As a further result, we show in Figs. 3 and 4 the learning curves representing the evolution of the Intersection over Union (IoU) on the validation set throughout the 200 training epochs for pipeline Fig. 1b, c respectively. As we can observe, in both scenarios, all the strategies involving the use of an in-domain dataset, both on its own or combined with a pre-training step on a natural image dataset, lead to faster convergence of the model on all the document classes characterizing the target datasets, compared to both random initialization and pre-training exclusively on out-of-domain datasets. Additionally, in-domain and hybrid pre-training strategies consistently allow for a much more stable learning process, drastically reducing the performance spike characterizing the other strategies. On the other hand, when pre-training exclusively on the cross-domain datasets and freezing the encoder module, the convergence time is comparable to that of the model trained from scratch, with the downside of the final IoU being higher. While fine-tuning the whole model on the target data, we can observe a marked instability of the training process during the first

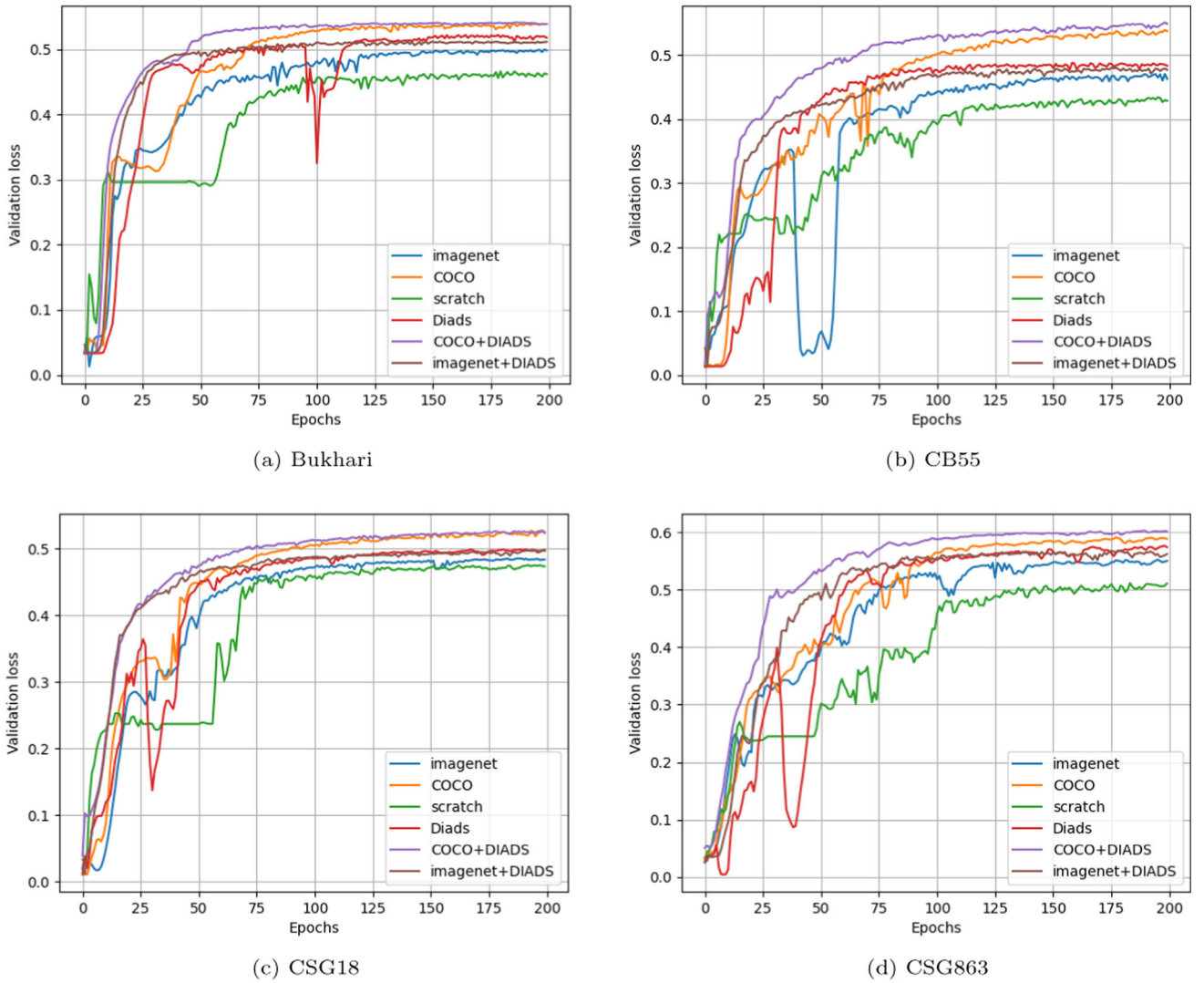
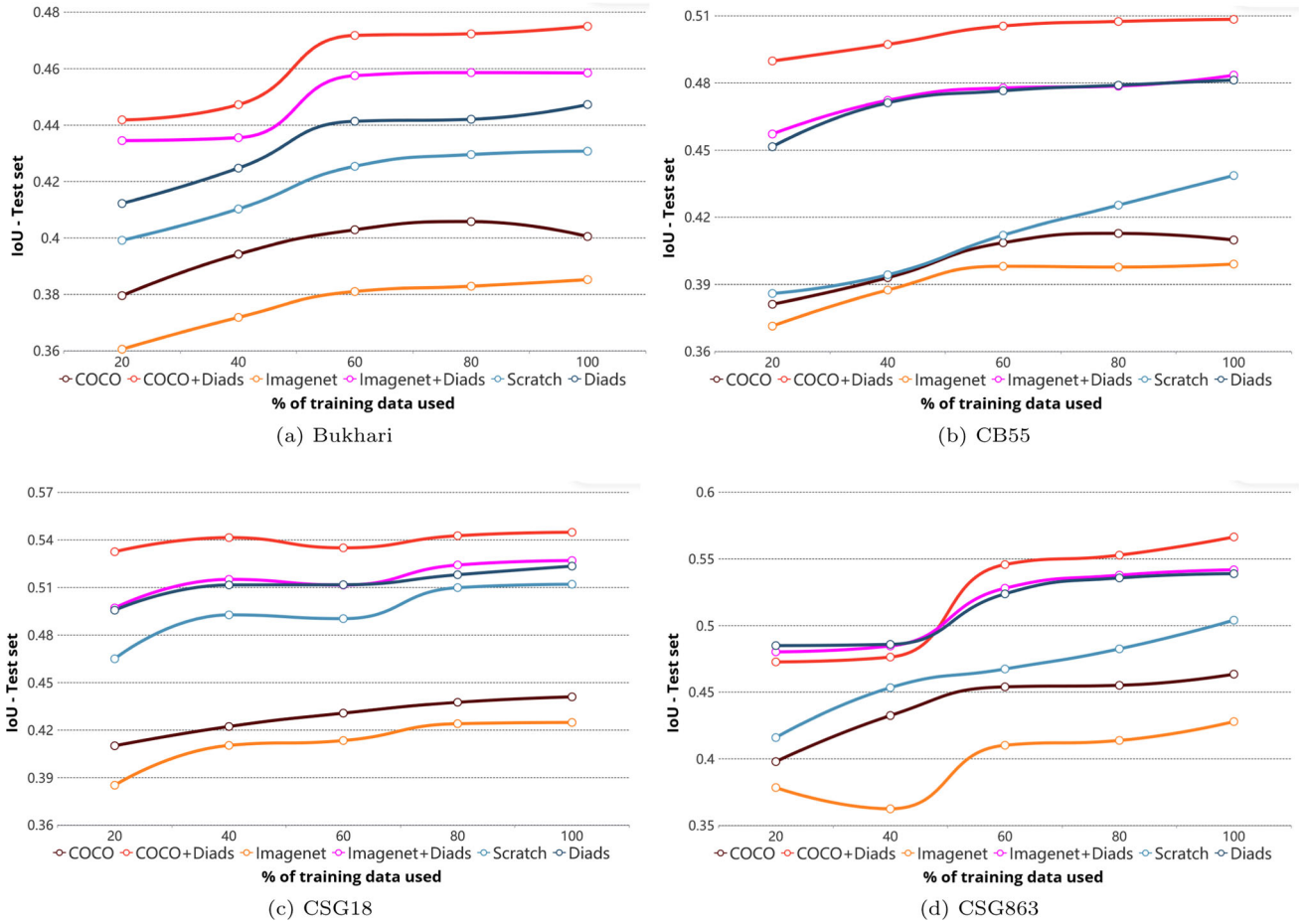


Fig. 4 Learning curves representing the evolution of the IoU scores for the four target document classes using different initialization strategies and fine-tuning the entire model on the target data

50 to 75 epochs for the models pre-trained exclusively on cross-domain or in-domain data, but after that point they substantially stabilize, leading to a higher final IoU compared to the model trained from scratch. As previously mentioned, this phenomenon does not occur in cases where a mixed pre-training strategy is employed. The validation loss curves characterizing them are very stable throughout the training process and consistently lead to the best overall performance compared to all the other approaches.

4.3.3 Performance in low-data regimes

Finally, we provide an analysis of the performance achieved through the different initialization strategies when artificially limiting the amount of data available from the target datasets to train the model. In particular, in Fig. 5 we show the results obtained by our models when trained only on 20%, 40%, 60%, and 80% of the training sets for the downstream tasks, with a frozen encoder module. From this analysis, it becomes even more evident how, in this scenario, pre-training exclusively on out-of-domain data sources doesn't lead to any real benefit compared to training our model from scratch, even when the amount of data available for training on the target task is very small (in the 20% setting, we have around four images available per document class). On the other hand, in-domain and cross-domain pre-training strategies allow for substantially improved performance both when working in high-data and especially in low-data regimes. In particular, we can see how the model resulting from the most effective initialization strategy, namely the hybrid strategy relying on COCO and U-DIADS-Bib as the source datasets, achieves better performance when trained on the 20% of the training



(a) Bukhari (b) CB55

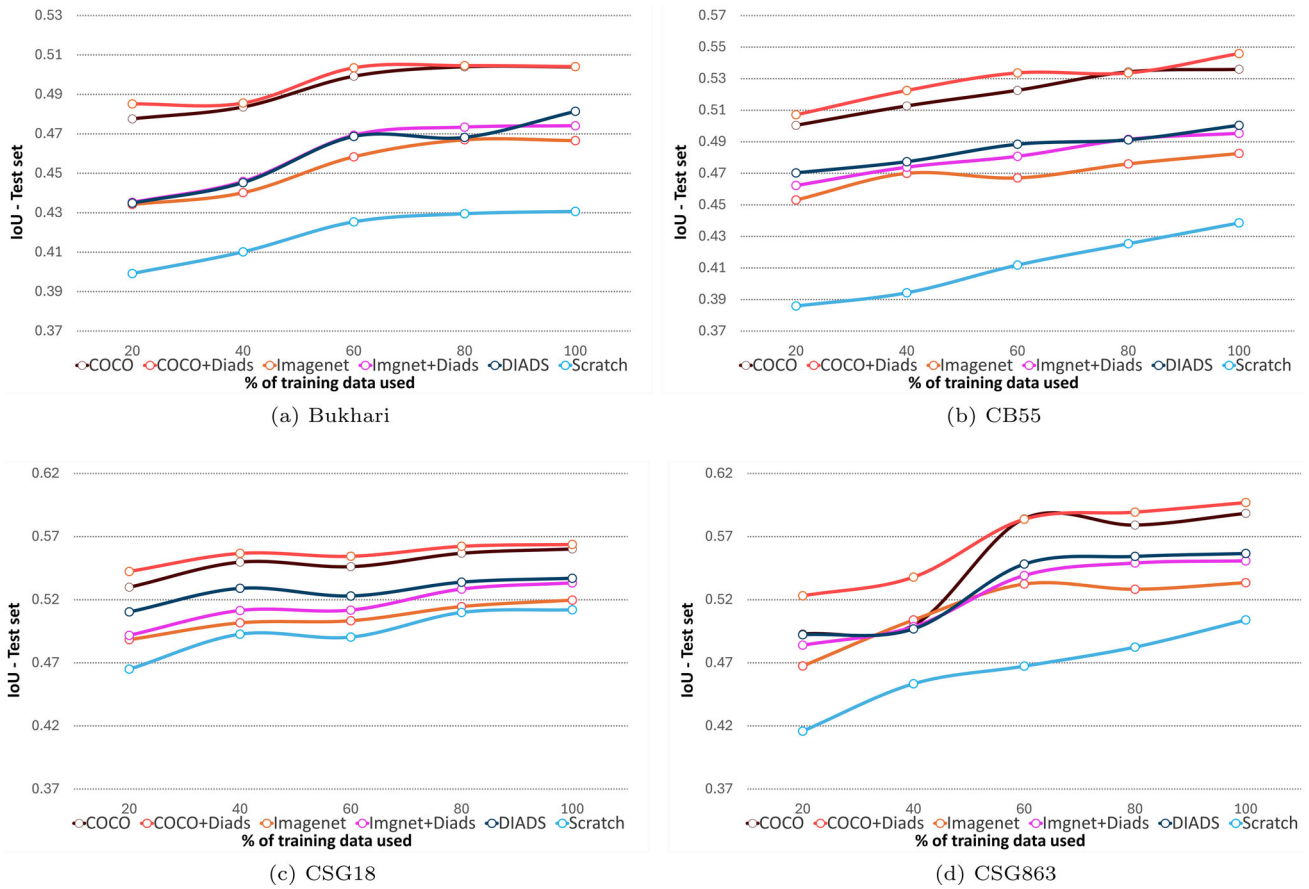
(c) CSG18 (d) CSG863

Fig. 5 Performance (IoU) of the segmentation model on the test set of the 4 document classes when initialized with different strategies while relying on increasing percentages of the available data for the training. Only the decoder and classifier modules of the model where trained on the target data, while the feature-extractor was kept frozen

data compared to the randomly initialized model trained on the entire training set for all the document classes except for CSG863. In Fig. 6, we show the same curves for the scenario in which the whole model is fine-tuned on the downstream task dataset. In this case, we can observe how all the pre-training strategies lead to an improvement in performance over training from scratch, regardless of the amount of data available for the fine-tuning process. In particular, we can see how, in this case, both the in-domain and mixed pre-training strategies, even when fine-tuned on a small portion of the target dataset, consistently outperform the model trained from scratch on the whole dataset on all four of the considered document classes.

4.3.4 Qualitative results

For completeness in Fig. 7 we provide a sample of the segmentation maps resulting from each training strategy considered in this analysis. As we can observe, consistently with the previously presented quantitative results, when freezing the encoder module of the segmentation model during the fine-tuning step on the target dataset, the segmentation maps obtained by relying on Imagenet and coco as pre-training datasets are much worse than the ones obtained with the other strategies, including training from scratch. In particular, the corresponding segmentation maps are completely missing large portions of the textual information while at the same time misclassifying other regions. When it comes to the remaining strategies, the visual difference between the generated segmentation maps are less obvious, but still characterized by sparse misclassifications especially involving the less frequent layout classes. Furthermore, the best-performing approaches, which are the ones involving the use of U-DIADS-Bib in the pre-training step, present segmentation maps characterized by a higher precision in the identification of the different layout components.



(a) Bukhari (b) CB55

(c) CSG18 (d) CSG863

Fig. 6 Performance (Iou) of the segmentation model on the test set of the four document classes when initialized with different strategies while relying on increasing percentages of the available data for the training. The model was fine-tuned entirely on the target data, with no frozen modules

4.3.5 Discussion

The previously presented results are very eloquent regarding the big impact on performance that the adoption of different transfer learning strategies, as well as different source datasets for pre-training, have when it comes to the task of document layout analysis. Differently from other application fields, the features learned exclusively from general-purpose datasets, such as ImageNet and COCO, don't seem to be directly transferable to the domain-specific task at hand. This behavior, which has already been observed in other application fields where the downstream task focuses on data that is substantially different from natural images, such as medical imaging, is likely to be attributed to the fact that natural images are typically characterized by larger structural components and a higher degree of inter-class similarity compared to document images, leading the pre-trained model to learn features, especially the high-level ones extracted by the deeper layers of the network, that don't capture the fine detail needed to work on document analysis tasks.

On the other hand, pre-training on natural images still provides a better starting point compared to random initialization when it comes to fine-tuning the model on the target data. This is particularly true when COCO is used as the pre-training dataset, which led to the second-best overall performance across all the considered initialization strategies. The reason why ImageNet is not as effective as a choice for pre-training is that, by being structured around a classification task, it leads the encoder module of the model to focus more on the general context of the images instead of on the fine details they contain, which is less suitable for segmentation tasks, particularly when working on document images.

We have further shown, how introducing a pre-training step focusing on a domain-specific dataset, even of a very small size such as the U-DIADS-Bib one employed in this study, greatly improves the transferability of the learned features to the downstream task. In particular, pre-training a segmentation model on document images, either exclusively, or in combination with a preliminary training step on natural images, not only consistently leads to improved performance on the downstream task but also positively affects the con-

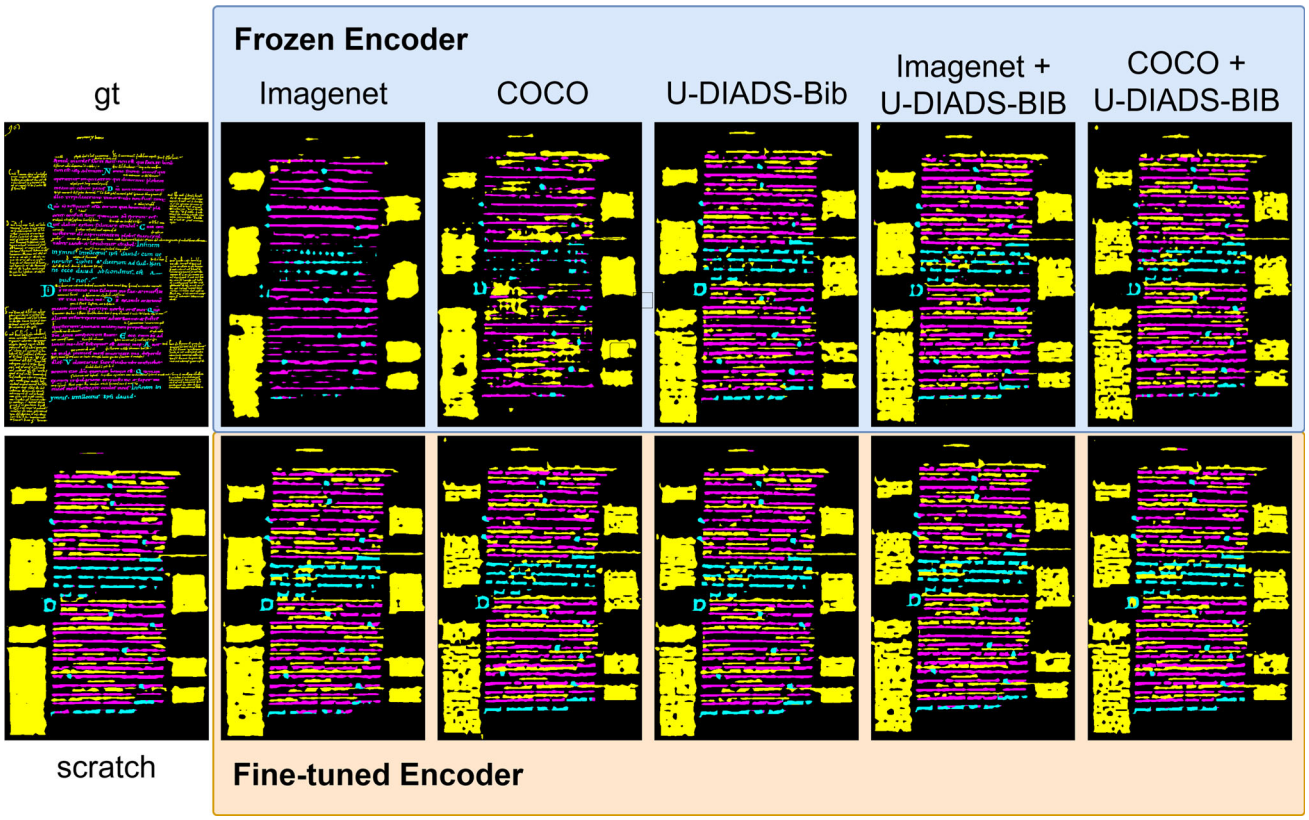


Fig. 7 Qualitative comparison of the segmentation maps obtained by employing the different training strategies considered in the anlaysis

vergence time and stability of the training process. More so, the combination of COCO and U-DIADS-Bib as pre-training data sources allows for a substantial reduction in the amount of data needed to effectively tackle the downstream task, with the pre-trained model achieving a better IoU score than the model trained from scratch while relying on as little as five times less data for the fine-tuning process.

5 Conclusions

In this work, we compared four different initialization approaches: random initialization, cross-domain initialization through the popular ImageNet1K and COCO datasets, in-domain initialization through the U-DIADS-Bib dataset, and hybrid initialization combining the pre-training of both the ImageNet1K and COCO datasets with a fine-tuning step on the in-domain one. Furthermore, we explored two different fine-tuning strategies involving, respectively, the training of the whole model and the training of exclusively the decoder and segmentation modules on the downstream task datasets. We tested the different approaches on two publicly available target datasets for document layout analysis, the DIVA-HisDB and the Bukhari one. We found out that, differently from other application areas, pre-training on large-scale, general-purpose datasets consisting of natural images doesn't bring any real benefit and is actually detrimental when working with downstream tasks revolving around document layout segmentation, both in terms of convergence speed as well as in terms of the overall performance of the model on the target dataset, unless the entire model is fine-tuned on the target data, leading to the intuition that the features learned from cross-domain data are not transferable directly to the domain of manuscript analysis. However, these learned features still represent a better starting point compared to the random initialization of the model weights. On the other hand, transfer learning strategies revolving around the use of in-domain source data, as well as hybrid strategies that make use of both in-domain and out-of-domain data, consistently lead to increased effectiveness and efficiency, in terms of the amount of data needed, on the downstream segmentation task, regardless of the training strategy employed on the target dataset.

In particular, we have shown how pre-training on the COCO dataset, followed by a fine-tuning process on U-DIADS-Bib, led to the best overall performance on the target task while at the same time substantially speeding up the convergence time of the model and, leading to a much more stable training process. Furthermore, we have shown how this approach allows for much more efficient use of the available

data, achieving better performance than the randomly initialized model even when trained on only 20% of the data, compared to the latter when trained on the entire dataset available for the downstream task.

To summarize our findings:

1. we provided a detailed overview of different transfer learning strategies in the context of document layout segmentation.
2. we identified the best initialization approach for transfer learning when working on document images, namely a hybrid initialization relying on a pre-training step on the COCO dataset followed by a fine-tuning step on the U-DIADS-Bib dataset.
3. we have also shown how by following this approach we are able to obtain increased stability during the training process on the target dataset, while also reducing the convergence time.
4. Finally, we showed how this approach allows for a much more efficient fine-tuning of the selected segmentation model on the downstream task data, allowing it to rely on as little to one-fifth of the data to achieve the same performance as the model trained from scratch.

To conclude, while we focused on the specific task of document layout segmentation, we believe our findings are likely applicable to other tasks involving the analysis of documents, both in printed and handwritten form, making it easier to tackle those problems where the scarcity of data has a big impact on the performance of the employed models. As a future effort, we plan to expand our analysis in this direction, to gain a deeper insight into the effect of in-domain transfer learning strategies across different tasks.

Author Contributions A.D.N. Performed the bulk of the experimentation and writing of the paper S.Z. Helped with data collection and writing of the paper C.P. Wrote and reviewed part of the paper G.L.F. reviewed the paper E.C. helped with data collection

Funding Open access funding provided by Università degli Studi di Udine within the CRUI-CARE Agreement. 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) - Interconnected Nord-Est Innovation Ecosystem (iNEST). Partial financial support was received from Strategic Departmental Plan on Artificial Intelligence, Department of Mathematics, Computer Science and Physics, University of Udine.

Data availability Data is provided within the manuscript or supplementary information files

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Narang, S.R., Jindal, M.K., Kumar, M.: Ancient text recognition: a review. Artif. Intell. Rev. **53**(8), 5517–5558 (2020). https://doi.org/10.1007/s10462-020-09827-4
2. Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M.: Automatic transcription of handwritten medieval documents. In: 2009 15th International Conference on Virtual Systems and Multimedia, pp. 137–142 (2009). https://doi.org/10.1109/VSMM.2009.26
3. Ni, K., Callier, P., Hatch, B.: Writer identification in noisy handwritten documents. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1177–1186 (2017). https://doi.org/10.1109/WACV.2017.136
4. Kiessling, B.: A modular region and text line layout analysis system. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 313–318 (2020). https://doi.org/10.1109/ICFHR2020.2020.00064
5. Minj, A., Garai, A., Mandal, S.: Text line segmentation: a FCN based approach. In: Singh, S.K., Roy, P., Raman, B., Nagabhushan, P. (eds.) Computer Vision and Image Processing, pp. 305–316. Springer, Singapore (2021)
6. Dutta, A., Garai, A., Biswas, S., Das, A.K.: Segmentation of text lines using multi-scale CNN from warped printed and handwritten document images. Int. J. Doc. Anal. Recognit. **24**(4), 299–313 (2021). https://doi.org/10.1007/s10032-021-00370-8
7. Zhang, C., Ibrayim, M., Hamdulla, A.: A methodological study of document layout analysis. In: 2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI), pp. 12–17 (2022). https://doi.org/10.1109/VRHCIAI57205.2022.00009
8. Garz, A., Seuret, M., Simistira, F., Fischer, A., Ingold, R.: Creating ground truth for historical manuscripts with document graphs and scribbling interaction. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 126–131 (2016). https://doi.org/10.1109/DAS.2016.29
9. 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 (2016). https://doi.org/10.1109/ICFHR.2016.0093
10. Bukhari, S.S., Breuel, T.M., Asi, A., El-Sana, J.: Layout analysis for Arabic historical document images using machine learning. In: 2012 International Conference on Frontiers in Handwriting Recognition, pp. 639–644 (2012). https://doi.org/10.1109/ICFHR.2012.227
11. Zottin, S., De Nardin, A., Colombi, E., Piciarelli, C., Pavan, F., Foresti, G.L.: U-DIADS-Bib: a full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts. Neu-

ral Comput. Appl. (2024). https://doi.org/10.1007/s00521-023-09356-5
12. De Nardin, A., Zottin, S., Piciarelli, C., Colombi, E., Foresti, G.L.: Few-shot pixel-precise document layout segmentation via dynamic instance generation and local thresholding. Int. J. Neural Syst. **33**(10), 2350052 (2023). https://doi.org/10.1142/S0129065723500521
13. Droby, A., Barakat, B.K., Madi, B., Alaasam, R., El-Sana, J.: Unsupervised deep learning for handwritten page segmentation. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, pp. 240–245 (2020). https://doi.org/10.1109/ICFHR2020.2020.00052
14. De Nardin, A., Zottin, S., Piciarelli, C., Colombi, E., Foresti, G.L.: A one-shot learning approach to document layout segmentation of ancient Arabic manuscripts. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 8112–8121 (2024). https://doi.org/10.1109/WACV57701.2024.00794
15. Tarride, S., Lemaitre, A., Coüasnon, B., Tardivel, S.: Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples. Int J Doc. Anal. Recognit. **24**(1), 77–96 (2021). https://doi.org/10.1007/s10032-021-00362-8
16. De Nardin, A., Zottin, S., Paier, M., Foresti, G.L., Colombi, E., Piciarelli, C.: Efficient few-shot learning for pixel-precise handwritten document layout analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, Hawaii, pp. 3680–3688 (2023). https://doi.org/10.1109/WACV56688.2023.00367
17. Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., Belongie, S., Tsai, Y.-H., Yang, M.-H.: Transfer learning in computer vision tasks: remember where you come from. Image Vis. Comput. **93**, 103853 (2020). https://doi.org/10.1016/j.imavis.2019.103853
18. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proc. IEEE **109**(1), 43–76 (2021). https://doi.org/10.1109/JPROC.2020.3004555
19. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016). https://doi.org/10.1109/TMI.2016.2535302
20. Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P., Rajendra Acharya, U.: Transfer learning techniques for medical image analysis: a review. Biocybern. Biomed. Eng. **42**(1), 79–107 (2022). https://doi.org/10.1016/j.bbe.2021.11.004
21. Boyd, A., Czajka, A., Bowyer, K.: Deep learning-based feature extraction in iris recognition: use existing models, fine-tune or train from scratch? In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9 (2019). https://doi.org/10.1109/BTAS46853.2019.9185978
22. Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W., He, Y.: Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. Comput. Electron. Agric. **167**, 105091 (2019). https://doi.org/10.1016/j.compag.2019.105091
23. Tercan, H., Guajardo, A., Meisen, T.: Industrial transfer learning: boosting machine learning in production. In: 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), vol. 1, pp. 274–279 (2019). https://doi.org/10.1109/INDIN41052.2019.8972099
24. Hua, J., Zeng, L., Li, G., Ju, Z.: Learning for a robot: deep reinforcement learning, imitation learning, transfer learning. Sensors **21**(4), 1278 (2021). https://doi.org/10.3390/s21041278
25. Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M., Ingold, R.: A comprehensive study of imagenet pre-training for historical document image analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 720–725 (2019). https://doi.org/10.1109/ICDAR.2019.00120
26. De Nardin, A., Zottin, S., Colombi, E., Piciarelli, C., Foresti, G.L.: Is imagenet always the best option? An overview on transfer learning strategies for document layout analysis. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds.) Image Analysis and Processing—ICIAP 2023 Workshops, pp. 489–499. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-51026-7_41
27. Chen, L., Papapandreou, G., Schroff, F., Hartwig, A.: Rethinking Atrous convolution for semantic image segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, vol. 6 (2017)
28. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf
29. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009).https://doi.org/10.1109/CVPR.2009.5206848
30. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
31. Garai, A., Biswas, S., Mandal, S., Chaudhuri, B.B.: Automatic rectification of warped Bangla document images. IET Image Processing **14**(1), 74–83 (2020). https://doi.org/10.1049/iet-ipr.2019.0831https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-ipr.2019.0831
32. Garai, A., Biswas, S., Mandal, S., Chaudhuri, B.B.: Automatic dewarping of camera captured born-digital Bangla document images. In: 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–6 (2017). https://doi.org/10.1109/ICAPR.2017.8593157