

PH.D. THESIS IN
COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

Document Segmentation under Limited Data Conditions: Transfer Learning and Few-Shot Approaches

CANDIDATE

Silvia Zottin

SUPERVISOR

Prof. Gian Luca Foresti

Cycle XXVIII — A.Y. 2024-2025

INSTITUTE CONTACTS

Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Università degli Studi di Udine
Via delle Scienze, 206
33100 Udine — Italy
+39 0432 558400
<https://www.dmif.uniud.it/>

AUTHOR'S CONTACTS

Silvia Zottin
Via delle Scienze, 206
33100 Udine — Italy
silvia.zottin@uniud.it

© 2026 Silvia Zottin

This work is shared under the Creative Commons 4.0 License Attribution-NonCommercial-ShareAlike.

To my parents, family, and friends,
for their love and support throughout this journey

Acknowledgements

A PhD journey is indeed an emotional rollercoaster, marked by highs and lows, pressure, tight deadlines, rejections, but also acceptances. All of these experiences have shaped me both professionally and personally. While there were many challenges and moments of doubt, the achievements and successes have been even more rewarding, motivating me to keep going. I am, therefore, deeply grateful for my decision to embark on this doctoral path, even if at times I found myself wondering, “*What have I gotten myself into?*”.

I would like to extend my heartfelt gratitude to my supervisor, Prof. Gian Luca Foresti, for his guidance, support, and encouragement throughout my PhD journey and my development in the academic world. I am also profoundly grateful to Prof. Claudio Piciarelli, who first introduced me to the world of academic research. Their expertise, advice, and unwavering support have been invaluable in helping me navigate both the obstacles and opportunities that this research has presented.

I am grateful to the AVML Lab, which came together somewhat by chance but has become a true source of satisfaction and inspiration. Thanks to the entire team, past and present (including Raffa the giraffe!), who have shared this journey with me laughing, joking, but above all, working, working, and working! Special thanks to my colleague Axel, who shared the world of research with me and made this PhD journey more exciting and, as we like to say, a little “*sparkling*”!

I would like to express my deepest gratitude to my family, whose love, encouragement, and unwavering support have been the foundation of everything I have achieved. Year after year, they have stood by me and guided me through every decision. I am also thankful to my friends, whose companionship, encouragement, and humor have made this journey brighter, more joyful, and truly memorable.

Thank you all from the bottom of my heart!

Abstract

This thesis addresses the problem of document segmentation under limited data conditions, focusing on industrial and historical documents. Document layout analysis is essential for converting unstructured visual information into machine-readable formats, supporting applications ranging from industrial automation to cultural heritage preservation. However, real-world scenarios often involve scarce annotated data and high variability in document layouts, which challenges traditional segmentation approaches that rely on large labeled datasets. To overcome these limitations, this thesis explores transfer learning and few-shot learning strategies, supported by the creation of two novel datasets specifically designed for industrial and historical documents.

A comprehensive analysis of low-resource document layout segmentation highlights how annotation scarcity and structural diversity affect model generalization. Extensive experiments demonstrate that domain-specific pre-training enhances performance more effectively than generic large-scale pre-training, while a novel few-shot learning framework achieves accurate segmentation using only a small number of annotated examples. The proposed methods also exhibit strong cross-domain generalization, addressing the challenges posed by variability and low-resource scenarios. Overall, the thesis shows that effective document segmentation can be achieved with minimal annotation effort, bridging the gap between theoretical advances and real-world applications and contributing to both industrial efficiency and cultural heritage preservation.

List of Publications

- **Zottin, S.**, De Nardin, A., Branca, G., Piciarelli, C., Foresti, G.L. (2026). IC-DAR 2025 Competition on FEw-Shot Text Line Segmentation of Ancient Hand-written Documents (FEST). In Document Analysis and Recognition – ICDAR 2025. Lecture Notes in Computer Science, vol 16027. Springer, Cham.
- Toma, A., Salvati, D., De Nardin, A., **Zottin, S.**, Scagnetto, I., Ferrin, G., Drioli, C., & Foresti, G. L. (2025). Deep acoustic learning on unmanned aerial vehicles for real-time human and drone detection. Integrated Computer-Aided Engineering, 10692509261417106.
- Piciarelli, C., **Zottin, S.**, De Nardin, A., & Foresti, G. L. (2025, September). Segmenting Anything in Architecture: A Tailored Approach to Segmenting Floor Plan Documents. In International Conference on Image Analysis and Processing (pp. 416-428). Cham: Springer Nature Switzerland.
- Madni, H. A., Shujat, H., De Nardin, A., **Zottin, S.**, & Foresti, G. L. (2025). Fs-BAD: Data-efficient feature reconstruction for few-shot brain anomaly detection. Pattern Recognition Letters.
- De Nardin, A., **Zottin, S.**, Toma, A., Piciarelli, C., & Foresti, G. L. (2025). A multibranch and multiattention framework for floor plan segmentation. Computer-Aided Civil and Infrastructure Engineering.
- De Nardin, A., **Zottin, S.**, Piciarelli, C., & Foresti, G. L. (2025). Deep Learning-Based Intrusion Detection Systems for Phishing Email Detection: A Short Survey. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7567-7575).

- Antonutti, D., De Nardin, A., **Zottin, S.**, Piciarelli, C., & Foresti, G. L. (2025). Multi-modal Analysis of Bi-Parametric MRI Slices for Lesion Detection in Prostate Cancer Screening. IEEE Access.
- Madni, H. A., Shujat, H., De Nardin, A., **Zottin, S.**, & Foresti, G. L. (2025). Unsupervised Brain MRI Anomaly Detection via Inter-Realization Channels. International journal of neural systems, 2550047.
- Madni, H. A., Umer, R. M., **Zottin, S.**, Marr, C., & Foresti, G. L. (2025). FL-W3S: Cross-domain federated learning for weakly supervised semantic segmentation of white blood cells. International Journal of Medical Informatics, 195, 105806.
- **Zottin, S.**, De Nardin, A., Branca, G., Colombi, E., Piciarelli, C., Shujat, H., & Foresti, G. L. (2025). Exploring few-shot text line segmentation approaches in challenging ancient manuscripts. In CEUR WORKSHOP PROCEEDINGS (Vol. 3937). CEUR-WS.
- **Zottin, S.**, De Nardin, A., Foresti, G. L., Colombi, E., & Piciarelli, C. (2024, August). ICDAR 2024 Competition on Few-Shot and Many-Shot Layout Segmentation of Ancient Manuscripts (SAM). In International Conference on Document Analysis and Recognition (pp. 315-331). Cham: Springer Nature Switzerland.
- De Nardin, A., **Zottin, S.**, Piciarelli, C., Foresti, G. L., & Colombi, E. (2024). In-domain versus out-of-domain transfer learning for document layout analysis. International Journal on Document Analysis and Recognition (IJDAR), 1-15.
- **Zottin, S.**, De Nardin, A., Colombi, E., Piciarelli, C., Pavan, F., & Foresti, G. L. (2024). U-DIADS-Bib: a full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts. Neural Computing and Applications, 36(20), 11777-11789.
- De Nardin, A., **Zottin, S.**, Piciarelli, C., Colombi, E., & Foresti, G. L. (2024). A one-shot learning approach to document layout segmentation of ancient Arabic manuscripts. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 8127-8136).

- De Nardin, A., **Zottin, S.**, Piciarelli, C., Colombi, E., & Foresti, G. L. (2023). Few-shot pixel-precise document layout segmentation via dynamic instance generation and local thresholding. *International Journal of Neural Systems*, 33(10), 2350052.
- De Nardin, A., **Zottin, S.**, Colombi, E., Piciarelli, C., & Foresti, G. L. (2023, September). Is imagenet always the best option? An overview on transfer learning strategies for document layout analysis. In *International Conference on Image Analysis and Processing* (pp. 489-499). Cham: Springer Nature Switzerland.
- De Nardin, A., **Zottin, S.**, Paier, M., Foresti, G. L., Colombi, E., & Piciarelli, C. (2023). Efficient few-shot learning for pixel-precise handwritten document layout analysis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3680-3688).
- De Nardin, A., **Zottin, S.**, Paier, M., Foresti, G. L., Colombi, E., & Piciarelli, C. (2022). Dynamic instance generation for few-shot handwritten document layout segmentation (short paper). In *CEUR Workshop Proceedings* (Vol. 3286, pp. 26-34). CEUR-WS.

Contents

1	Introduction	1
1.1	Image Segmentation in Computer Vision	1
1.2	Document Layout Analysis	5
1.3	Challenges of Limited Data Conditions	8
1.4	Contributions and Thesis Organization	9
2	Related Works	13
2.1	General Purpose Semantic Segmentation	13
2.2	Document Layout Segmentation	15
3	Datasets and Challenges	19
3.1	Industrial Dataset	20
3.1.1	TobaccoSeg dataset	22
3.2	Historical Datasets	27
3.2.1	U-DIADS-Bib	28
3.2.2	DIVA-HisDB	39
3.2.3	Bukhari dataset	41
3.3	Common Challenges	43
4	Effective Transfer Learning for Document Layout Analysis	47
4.1	Methods	48
4.1.1	Segmentation architecture	48
4.1.2	Transfer learning strategies	50
4.1.3	Pre-training datasets	51
4.1.4	Evaluation datasets	54
4.2	Experiments	54
4.2.1	Training setup	54
4.2.2	Metrics	55
4.3	Results	56
4.3.1	Transferability of the learned features	56
4.3.2	Impact on convergence time	57
4.3.3	Performance in low-data regimes	58

4.3.4	Qualitative results	60
4.3.5	Discussion	61
5	Few-Shot Learning for Document Layout Segmentation	65
5.1	Method	66
5.1.1	Segmentation backbone	66
5.1.2	Dynamic instance generation	68
5.1.3	Segmentation refinement	69
5.2	Experiments	70
5.2.1	Datasets	70
5.2.2	Training and inference setup	71
5.2.3	Evaluation metrics	71
5.3	Results	72
5.3.1	Quantitative results	72
5.3.2	Qualitative results	75
5.3.3	Ablation study	78
5.3.4	Discussion	81
6	Conclusions and Future Work	83

List of Tables

3.1	Classes distribution (%) for TobaccoSeg dataset.	24
3.2	Performance of the different semantic segmentation models on TobaccoSeg dataset.	25
3.3	Pixel-level class distribution (%) for each manuscript in U-DIADS-Bib and its few-shot variant U-DIADS-BibFS.	33
3.4	Comparison of the performance of the different semantic segmentation models chosen tested on U-DIADS-Bib dataset.	36
3.5	Comparison of the performance of selected segmentation models versus the state-of-the-art on the U-DIADS-BibFS dataset.	38
3.6	Classes distribution (%) for each manuscripts of Diva-HisDB [64] (CB55, CSG18 and CSG863) dataset.	41
3.7	Classes distribution (%) for Bukhari et al. [10] dataset.	43
4.1	Compact overview of key information regarding the datasets used for our analysis.	52
4.2	Results for the different pre-training strategies.	55
5.1	Quantitative results for the document layout segmentation task in DIVA-HisDB dataset.	73
5.2	Quantitative results for the document layout segmentation task in U-DIADS-BibFS, Bukhari and TobaccoSeg datasets.	74
5.3	Comparison between the use of different neural network architectures as the segmentation backbone for our model.	78
5.4	Comparison between the adoption of different patch sizes during the instance generation process of our framework.	79
5.5	Ablation study for the document layout segmentation framework. . . .	79

List of Figures

1.1	Example of the three main image segmentation tasks.	2
1.2	Examples of image segmentation in different application fields: (a) Medical imaging [54], (b) Industrial anomaly detection [7], and (c) Floor plan segmentation [28].	4
1.3	Visual examples of the three main DLA tasks across different types of documents.	6
3.1	Samples of TobaccoSeg images with their corresponding ground truth masks.	21
3.2	Confusion matrices obtained by applying DeepLabV3+ [16] to the TobaccoSeg dataset.	26
3.3	Images showing a page of each manuscript of U-DIADS-Bib dataset as well as its corresponding ground truth mask.	28
3.4	Samples of six segmentation classes in U-DIADS-Bib dataset: Main Text (3.4a), Decoration (3.4b), Title (3.4c), Chapter Headings (3.4d), Paratext (3.4e) and Background (3.4f).	30
3.5	Confusion matrices obtained by applying DeepLabV3+ [16] to the four ancient manuscripts of the U-DIADS-Bib dataset.	37
3.6	Confusion matrices obtained by applying the model proposed by De Nardin et al. [23] to the four ancient manuscripts of the U-DIADS-BibFS dataset.	39
3.7	Examples from the DIVA-HisDB [64] dataset with corresponding ground truth masks.	40
3.8	Sample pages and ground truth masks from the Bukhari et al. [10] dataset.	42
3.9	Examples of difficulties and challenges present in document images from the U-DIADS-Bib, TobaccoSeg dataset, and DIVA-HisDB.	44
4.1	Visual representation of the training strategies explored.	49
4.2	Examples from the three datasets used as pre-training data: ImageNet-1k with the class “cup”, COCO, and U-DIADS-Bib.	53
4.3	Learning curves representing the evolution of the IoU scores for the 5 target document classes using different initialization strategies and keeping the encoder frozen at the time of fine-tuning on the target data.	58

4.4	Learning curves representing the evolution of the IoU scores for the 5 target document classes using different initialization strategies and fine-tuning the entire model on the target data.	59
4.5	Performance (IoU) of the segmentation model on the test set of the 5 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.	60
4.6	Performance (IoU) of the segmentation model on the test set of the 5 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.	61
4.7	Qualitative comparison of the segmentation maps of CSG18 manuscript obtained by employing the different training strategies considered in the analysis.	62
4.8	Qualitative comparison of the segmentation maps of TobaccoSeg dataset obtained by employing the different training strategies considered in the analysis.	63
5.1	Visual representation of the segmentation pipeline for the document layout analysis framework.	67
5.2	Illustration of the instance generation process.	69
5.3	Qualitative results for the document layout segmentation task.	76
5.4	Overview of the main instances of misclassification for the proposed approach.	77
5.5	Qualitative results showing the effects of the segmentation refinement process.	80

1

Introduction

1.1 Image Segmentation in Computer Vision

Image segmentation is one of the core problems in computer vision, as it seeks to partition an image into coherent and meaningful regions that correspond to objects or structural elements. Unlike image classification, where the task is to assign a single label to the entire image, or object detection, which localizes entities by placing bounding boxes around them, segmentation works at a much finer level of detail. In fact, each pixel in the image is assigned a label, producing a dense prediction that precisely outlines the boundaries of objects and separates them from their surroundings. This pixel-level representation provides a richer and more informative description of the scene, making segmentation an indispensable tool in applications that demand accurate localization, precise boundary information, or structural interpretation.

From a formal perspective, image segmentation can be defined as the problem of classifying each pixel in an image according to one or more criteria, which can be semantic, structural, or contextual. Depending on the type of information required, segmentation tasks are commonly grouped into three major categories: semantic, instance, and panoptic segmentation [43] (Fig. 1.1).

Semantic segmentation aims at associating every pixel with a semantic category, grouping together all pixels that belong to the same type of object or region. The outcome is a partition of the image into homogeneous semantic areas, without distinction between different occurrences of the same class. As an example, in Fig. 1.1(b), all surfboards are represented as belonging to the same category, while other pixels are

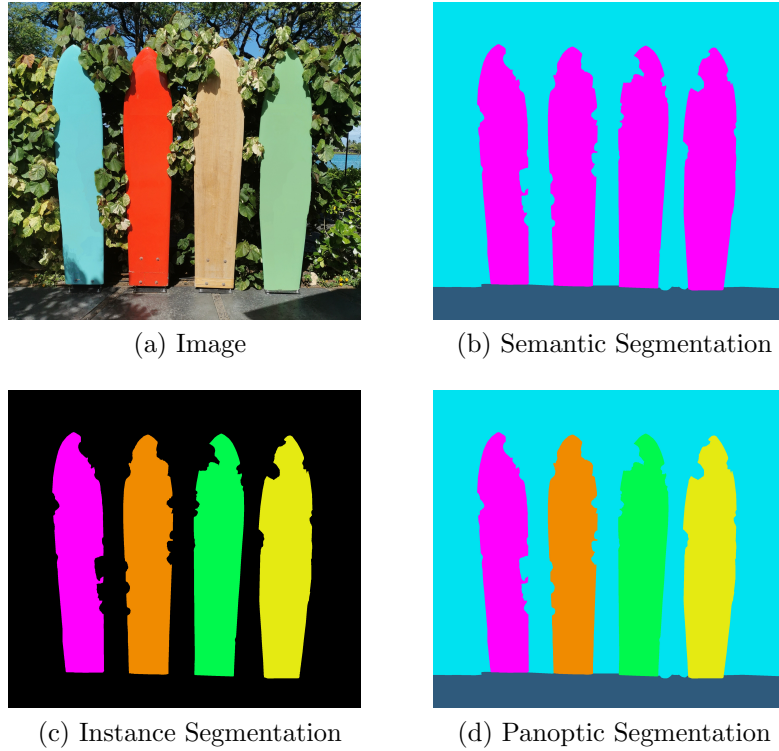


Figure 1.1: Example of the three main image segmentation tasks: (a) input image, (b) semantic segmentation, (c) instance segmentation, and (d) panoptic segmentation.

labeled as background or floor.

Instance segmentation builds upon this idea by adding the capability to distinguish individual objects of the same class. Instead of merging them into a single semantic region, each occurrence is segmented separately, thus producing object-level masks enriched with class labels. In Fig. 1.1(c), the four surfboards are identified not only as “surfboard” but also as four distinct entities, each isolated with its own mask.

Panoptic segmentation represents a more recent paradigm that unifies the previous two approaches. It provides per-pixel labels that encode both semantic categories and instance-level information, delivering a holistic and coherent representation of the scene. As shown in Fig. 1.1(d), panoptic segmentation does not simply classify pixels as surfboards, floor, or background, but also differentiates between each surfboard individually. This type of segmentation is particularly valuable in domains where a high degree of perceptual detail and scene understanding is required, such as robotics, autonomous driving, and architectural analysis, where both the nature of the objects

and their individuality are equally important.

The study of segmentation has a long history in image analysis. Early approaches relied on low-level cues such as color, texture, or intensity, and employed techniques like thresholding, region growing, clustering, or watershed algorithms [72]. While these approaches were computationally efficient and interpretable, they struggled in complex scenes with cluttered backgrounds, varying illumination, or occlusion, and they lacked the ability to capture semantic meaning.

A major breakthrough came with learning-based approaches. Initially, machine learning models such as random forests or support vector machines were applied to handcrafted features, offering improved robustness in specific domains. However, these methods still relied heavily on human-designed descriptors, which limited their ability to generalize across diverse visual scenarios.

With the advent of deep learning, segmentation reached an entirely new level of accuracy and flexibility [52]. Convolutional Neural Networks introduced the ability to learn hierarchical feature representations directly from raw data, eliminating the need for manual feature engineering. Architectures such as Fully Convolutional Networks demonstrated that dense per-pixel predictions could be generated efficiently by adapting classification networks into fully convolutional forms. Later, encoder-decoder designs such as U-Net [61] and SegNet [5] further improved localization by combining low-level spatial details with high-level semantic understanding.

More recently, advances in attention mechanisms and transformer-based architectures have pushed the boundaries even further, enabling models to capture long-range dependencies and global context, which are essential in complex scenes [42]. Combined with large annotated datasets and powerful GPUs, deep segmentation models now achieve unprecedented levels of performance. Nevertheless, this progress comes at a cost: the reliance on massive amounts of labeled data, high computational demands, and the difficulty of transferring models to new domains remain significant challenges.

The applications of segmentation are vast and span multiple domains (Fig. 1.2). In medical imaging, it is employed for tasks such as tumor detection [3] and organ delineation [54], providing clinicians with precise structural information. In autonomous driving, semantic segmentation enables vehicles to build a detailed understanding of road environments by identifying lanes, pedestrians, and traffic signs at the pixel level [69], thereby supporting safe navigation. In industrial settings, segmentation is often used for anomaly detection [62, 7], where defects such as cracks, scratches, or

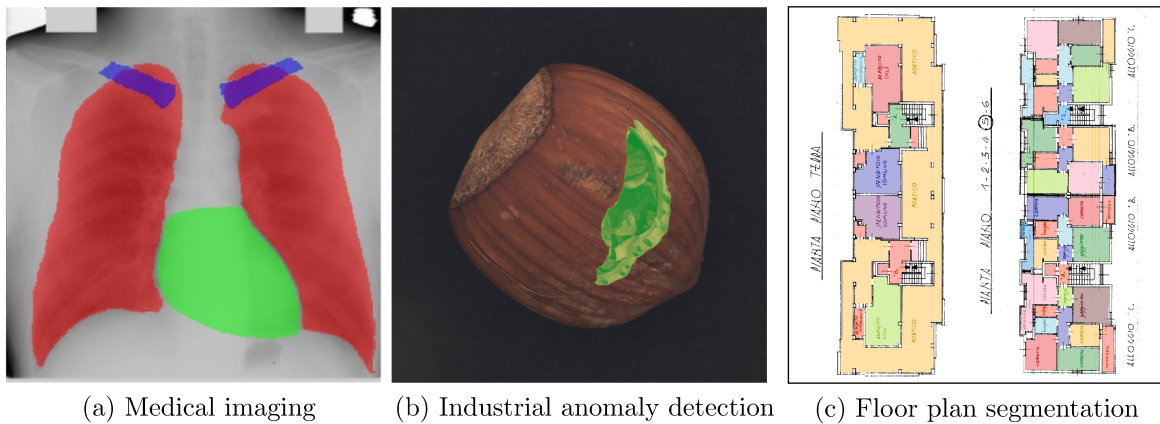


Figure 1.2: Examples of image segmentation in different application fields: (a) Medical imaging [54], (b) Industrial anomaly detection [7], and (c) Floor plan segmentation [28].

surface irregularities are identified at fine granularity, ensuring quality control in manufacturing pipelines. In the field of architecture, segmentation methods are increasingly applied to building floor plans [28]. Automatic analysis of architectural drawings makes it possible to detect and segment individual rooms, distinguish structural elements such as walls, doors, and windows, and ultimately compute room areas and the overall surface of an apartment or building. This type of approach reduces manual effort and enables accurate and scalable analysis for renovation, real estate assessment, or energy efficiency planning. Additional domains, such as biometrics [67], robotics [37], security [26] and, most importantly for the focus of this thesis, document analysis, also rely heavily on segmentation to extract structured knowledge from images.

Despite remarkable advances, image segmentation remains a challenging problem. The variability of object appearance, the presence of small or ambiguous classes, complex boundaries, and frequent occlusions introduce significant difficulties. Furthermore, deep segmentation models typically require large annotated datasets to achieve competitive results, and their performance often degrades when applied to new domains or in scenarios with limited training data. These challenges highlight the dual nature of segmentation research: while it has already matured into a reliable tool for many real-world applications, it continues to face fundamental obstacles that motivate ongoing investigation, particularly in specialized contexts such as document layout analysis.

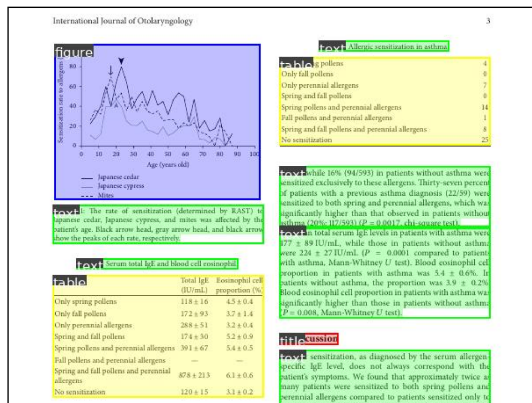
1.2 Document Layout Analysis

Document Image Analysis (DIA) is a broad research field within computer vision and pattern recognition that focuses on the automatic interpretation of documents in their various forms [48]. Its ultimate goal is to convert unstructured or semi-structured visual representations, such as scanned pages, handwritten notes, historical manuscripts, forms, or technical drawings, into structured and machine-readable information. By doing so, DIA enables a wide range of applications, from digital preservation and information retrieval to workflow automation and semantic content analysis.

The scope of DIA is intentionally wide, encompassing tasks at different levels of abstraction. At the lowest level, DIA includes image enhancement and preprocessing operations, such as binarization [63] and noise removal [38], which prepare document images for subsequent processing stages. At an intermediate level, it involves segmentation and structural analysis, in which the physical and logical organization of a document is extracted. This includes tasks such as page segmentation [25], text line segmentation [75], table recognition [34], and layout understanding [48]. Finally, at the highest level, DIA addresses semantic interpretation, involving Optical Character Recognition (OCR) [55], automatic text transcription [32], and writer identification [56].

Historically, document analysis has been closely tied to the development of OCR systems. Early research in the field was driven by the need to digitize printed text efficiently and reliably. However, over the years, its scope has expanded to include complex documents characterized by heterogeneous layouts, handwritten annotations, diagrams, mathematical formulas, and graphical elements. Modern DIA methods must therefore integrate computer vision, machine learning, and natural language processing to address this variety.

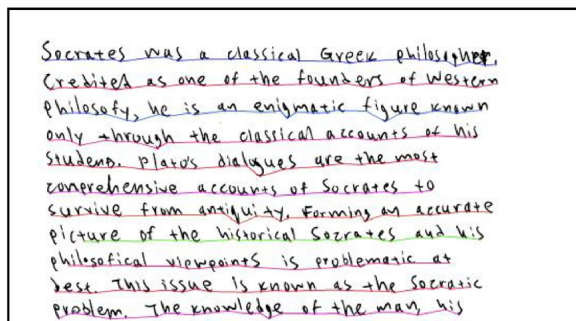
The importance of DIA extends across several domains. In industry, it supports the automation of document-intensive processes, such as invoice management, form processing, and quality-control reporting. In cultural heritage, it enables the systematic digitization, transcription, and study of historical manuscripts and archival records, providing new opportunities for scholars in the digital humanities. In legal, administrative, and educational contexts, DIA ensures faster access to information, reduces manual workload, and increases compliance and traceability. Despite the diversity of these use cases, a common challenge persists: documents encountered in the real world



(a) Page segmentation



(b) Text line segmentation



(c) Baseline detection

Figure 1.3: Visual examples of the three main DLA tasks across different types of documents: (a) page segmentation in a printed document [74], (b) text line segmentation in a historical manuscript [75], and (c) baseline detection in a handwritten letter [59].

are often highly variable, noisy, and poorly structured.

Within this landscape, Document Layout Analysis (DLA) emerges as a central component, providing the structural foundation upon which higher-level interpretation tasks rely. DLA focuses on uncovering the physical and logical organization of a document by partitioning it into semantically meaningful regions [8]. These regions can include text blocks, titles, tables, figures, footnotes, marginal notes, or decorative elements, each carrying a specific role within the document's structure. Without this step, tasks such as OCR, transcription, and information extraction cannot be carried out effectively. DLA is commonly structured into three fundamental sub-processes, each addressing a specific level of granularity in the structural interpretation of documents. These tasks are not sequential but operate independently, each providing complementary information that enriches the representation of the document image and enables

advanced analysis. Figure 1.3 illustrates a visual example for each task that comprises the DLA.

Page segmentation (Fig. 1.3(a)) partitions a document page into distinct layout components, such as main body text, titles, figures, tables, marginalia, decorations, or background. The goal is to separate semantically meaningful regions that fulfill different communicative roles within the document. Accurate page segmentation is essential, as it provides structural context for other analytical processes. In this thesis, we specifically address the task of page segmentation.

Text line segmentation (Fig. 1.3(b)) identifies and segments individual lines of text. This task is crucial for OCR and automatic transcription. In printed documents, where the structure is regular and well-aligned, text line extraction is largely solved. Handwritten sources, however, pose significant challenges: lines may be curved or slanted, characters may overlap across lines, and additional material such as marginal glosses, annotations, or decorative ornaments can obscure clear separation. Sophisticated models are therefore required to handle both local variations in handwriting and global layout irregularities [2, 76].

Finally, baseline detection (Fig. 1.3(c)) identifies the virtual line on which the characters of a text rest, with ascenders extending above and descenders below. This information provides a structural reference that supports precise character alignment, word segmentation, and transcription, which are critical for accurate text recognition [30].

The importance of document segmentation spans both modern and historical contexts. The modern category includes a broad spectrum of contemporary materials, ranging from industrial records and administrative forms to business reports, invoices, technical schematics, and digitally generated layouts, whose structure and content vary widely across domains. Within industrial environments, in particular, organizations produce large volumes of semi-structured documents such as inspection reports, production logs, forms, and technical drawings, whose manual processing is both costly and error-prone. Automating layout segmentation accelerates digitization workflows, enhances compliance and traceability, and reduces reliance on human intervention. However, industrial documents often feature highly heterogeneous layouts, mixing handwritten and printed text, stamps, signatures, and annotations, which pose unique segmentation challenges. Conversely, historical manuscripts represent a radically different but equally demanding scenario. These artifacts, often centuries old, exhibit non-standardized layouts: annotations in the margins, interwoven text and il-

illustrations, decorative borders, and a wide range of handwriting styles. In addition, degradation caused by fading ink, bleed-through, or page warping makes segmentation particularly complex. Yet, accurate layout analysis is fundamental for the digital humanities, as it enables systematic study, transcription, and preservation of cultural heritage.

Although modern industrial and historical documents differ substantially in form and purpose, they share important challenges: variability, noise, and the lack of large annotated datasets. In both contexts, segmentation is not an end in itself but rather a gateway to broader goals: workflow automation in industry and enhanced accessibility of knowledge in the humanities. This dual perspective highlights the need for robust and adaptable DLA methods, capable of generalizing across diverse document types and use cases.

1.3 Challenges of Limited Data Conditions

One of the most critical challenges in DLA, and in particular in page segmentation, is the scarcity of annotated training data. Unlike large-scale natural image datasets, which can contain millions of labeled examples across diverse categories, document datasets, especially in industrial or historical contexts, rarely reach such scale. Producing pixel-level ground truth is an expensive and labor-intensive task, often requiring expert knowledge to ensure precision in delineating regions of interest.

The lack of sufficient annotated data has several consequences. Models trained on small datasets are prone to overfitting, achieving good performance on the training set but generalizing poorly to unseen documents. Domain variability exacerbates the problem: industrial forms may differ drastically across organizations and even across years within the same company, while historical manuscripts exhibit wide variation in script styles, page layouts, and degradation patterns. Domain shifts are therefore the norm rather than the exception, meaning that a model optimized for one dataset may fail when applied to another.

These challenges underscore the need for methods that can learn effectively in low-data regimes. Transfer learning, by leveraging knowledge from large-scale vision datasets, provides a practical way to bootstrap model performance with minimal supervision. Few-shot learning pushes this paradigm further, aiming to generalize from only a handful of annotated examples. Complementary approaches such as domain adap-

tation, semi-supervised learning, and data augmentation provide additional means to mitigate domain variability and compensate for annotation scarcity. Addressing these issues is not only a theoretical necessity but also a practical requirement for deploying DLA systems in real-world scenarios. For this reason, the development of robust methodologies for Page Segmentation under limited data conditions constitutes the central research direction of this thesis.

1.4 Contributions and Thesis Organization

This thesis addresses the problem of document image segmentation under limited data conditions, with an emphasis on both industrial and historical documents. The main contributions can be summarized as follows:

- A systematic analysis of the challenges encountered in low-data regimes, highlighting the specific issues of variability, and annotation scarcity in document layout segmentation.
- The creation of two novel datasets specifically designed for document layout analysis: one composed of industrial documents and one of historical handwritten manuscripts, both publicly released to support further research in the field.
- A comprehensive evaluation of transfer learning strategies, investigating how models pre-trained on large-scale natural image datasets can be adapted to the document domain.
- The design of a few-shot learning approach for document layout segmentation, demonstrating how accurate results can be achieved with only a few annotated samples.
- Experimental validation on a diverse set of datasets, including both industrial documentation and historical manuscripts, ensuring that the proposed methods generalize across contexts and are applicable in real world scenarios.

By pursuing these objectives, the thesis aims to bridge the gap between the data requirements of modern deep learning methods and the practical constraints of document collections. The results are expected to contribute to both industrial automation

and cultural heritage preservation, offering solutions that are technically effective and broadly relevant.

The remainder of the thesis is organized as follows. Chapter 2 surveys prior work on image segmentation and document layout analysis. Chapter 3 introduces the industrial and historical datasets used throughout the study, and discusses the common challenges they pose. Chapter 4 investigates the impact of transfer learning on document segmentation, while Chapter 5 presents a novel few-shot framework that integrates deep semantic models with traditional image processing. Finally, Chapter 6 summarizes the main findings and outlines directions for future research.

2

Related Works

Ongoing efforts to enhance the accuracy of image segmentation systems have produced a variety of powerful deep learning models and frameworks, suitable for a broad range of applications. This chapter provides a brief overview of the most notable of these systems, focusing in particular on those developed for document layout segmentation.

2.1 General Purpose Semantic Segmentation

Over the last decade, deep learning has profoundly transformed the field of semantic segmentation, enabling models to learn rich hierarchical and contextual representations and produce high-resolution dense predictions. General-purpose semantic segmentation architectures are designed to operate effectively across diverse datasets and scenarios, balancing accuracy, efficiency, and scalability. Several review studies are available in the literature that analyze semantic segmentation models, highlighting its applications across multiple domains [53, 71]. Among the numerous segmentation models proposed, some of the most notable architectures are described below, and they are also employed in experiments presented in this thesis.

Fully Convolutional Networks (FCNs) [49] represent the foundational deep learning approach to semantic segmentation. FCNs replace fully connected layers with fully convolutional layers, allowing dense predictions on images of arbitrary size. They follow an encoder-decoder paradigm: the encoder extracts hierarchical feature maps, while the decoder restores spatial resolution through upsampling layers and learned deconvolution filters. Skip connections propagate fine spatial details from shallow lay-

ers to deeper semantic representations, improving localization accuracy and boundary delineation. FCNs established the basic principles of end-to-end learning for pixel-wise prediction, inspiring subsequent architectures focused on multi-scale context and refinement.

Building upon the need for global contextual reasoning, the Pyramid Scene Parsing Network (PSPNet) [73] introduced a Pyramid Pooling Module that captures information from multiple spatial scales. Using a ResNet backbone, PSPNet first extracts hierarchical features and then enriches them by pooling from progressively coarser spatial regions, integrating both local and global context. The decoder refines these feature maps through convolutional layers to produce high-resolution segmentation outputs, demonstrating the importance of multi-scale contextual aggregation for complex scenes.

The DeepLab family of models represents another influential line of research in semantic segmentation. DeepLabV3 [15] improves upon earlier DeepLab versions by introducing Atrous Spatial Pyramid Pooling (ASPP), which combines multiple parallel atrous (dilated) convolutions with different dilation rates to capture multi-scale contextual information efficiently. By using atrous convolutions, DeepLabV3 preserves high-resolution feature maps while enlarging the receptive field, allowing the network to incorporate both local and global context without losing spatial detail. Unlike previous versions, DeepLabV3 does not initially rely on an explicit decoder, focusing on enriching encoder features with multi-scale context before producing the segmentation map, which improves accuracy especially in large objects and complex scenes.

DeepLabV3+ [16] represents the latest iteration, introduced by Google in 2018, of the DeepLab framework series of segmentation models. The success of this model is attributed to a set of architectural choices, some inherited from its predecessors, such as atrous convolution and ASPP, and others representing a departure, notably the addition of a lightweight decoder module. This decoder upsamples encoder outputs and integrates lower-level feature maps to refine object boundaries and spatial details. By combining multi-scale context with precise boundary reconstruction, DeepLabV3+ achieves state-of-the-art performance in general-purpose semantic segmentation.

For applications demanding computational efficiency, the Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) [36] offers a lightweight alternative to ASPP-based architectures. By simplifying the multi-scale pooling operations and reducing the number of parameters, LRASPP maintains competitive accuracy while enabling real-time inference, making it particularly suitable for mobile or embedded systems.

2.2 Document Layout Segmentation

Semantic segmentation in the context of document layout analysis, commonly referred to as document layout segmentation, presents unique challenges compared to general-purpose semantic segmentation, particularly when the documents of interest are handwritten. Several surveys have been conducted on DLA tasks for both historical and contemporary documents, including printed and handwritten materials [8, 50]. However, solutions specifically targeting handwritten documents remain relatively scarce.

Page segmentation, which is the focus of this thesis, remains an open problem within the machine learning community. The techniques developed for this task are generally classified into three categories: top-down, bottom-up, and hybrid approaches [8].

Top-down approaches assume that pages have a well-defined structure and layout. The page segmentation process starts from the whole page and cuts it into homogeneous zones. In these methods, various characteristics of the document page structure are considered, such as white space between text regions, size of text blocks and the measures between main texts and paratexts [18, 4, 19]. In general, the top-down methods are easily applicable but not suitable for complex layouts such as handwritten historical documents. In addition, these methods depend on the layout structure of the document, so they have a low generalization capability.

The bottom-up strategy derives document analysis dynamically from smaller granularity data levels, such as pixels, groups of pixels and connected components, to generate larger and different regions with uniform elements [51, 50, 41]. The main advantage of bottom-up methods is that these techniques do not require any prior knowledge of the document layout structure, and, for this reason, this strategy is preferred when working with manuscripts. However, usually, these techniques demand many labeled data that is often not available, especially in the domain of historical documents where highly specialized expertise is needed. That is why the request for methods with few-shot and unsupervised learning is increasingly necessary and required.

Although research for this task is well established, there are still many challenging issues that neither bottom-up nor top-down strategies can adequately address. For this reason, a strategy has been identified that derives from the integration of the other two main ones, called the hybrid strategy [6, 68, 44].

Over the years, many techniques have been used to address the page segmentation task. Recently, BinMakhashen and Mahmoud [9] proposed a learning-free and hybrid

document layout analysis for manuscripts. First, the proposed method locates the main content initially using top-down white space analysis. Then, it extracts template features representing the manuscripts author’s writing behavior. After that, moving windows are used to scan the manuscript page and define main-content boundaries more precisely. Chen et al. [12] used a convolutional autoencoder to learn the features directly from the pixel intensity values and train a Support Vector Machine (SVM) to segment the page without any assumption of specific shapes of document layouts. A similar approach was proposed by Mehri et al. [51] with the method based on learning texture features. This method for extracting texture features from images uses the linear iterative clustering of super-pixels, Gabor descriptors and the co-occurrence matrix of the gray level. Then an SVM is used to classify pixels into foreground and background. The page segmentation problem can approach as a pixel labeling problem such as the work by Chen et al. [11], where the features are learned directly from randomly selected image patches by using stacked convolutional autoencoders. With an SVM trained with the extracted features of the central pixels of the super-pixels, an image is segmented into semantic regions. Finally, the segmentation results are refined by a connected components-based smoothing procedure. Following the same idea, in Chen et al. [13] local features are learned with stacked convolutional autoencoders in an unsupervised manner for the purpose of initial labeling. Then a conditional random field model is applied for modeling the contextual information to improve the segmentation results. More recent, Jo et al. [40] present a method that separates handwritten and machine-printed components even when they overlap in documents, using a U-Net trained in an end-to-end manner. Their approach avoids conventional preprocessing and connected-component extraction, employs a cross-entropy-based loss to handle class imbalance, and relies on a novel data synthesis method to generate realistic overlapping training samples. Another interesting approach was proposed by Xu et al. [70] with a Multi-task Layout Analysis framework (MLA) to solve all layout analysis tasks simultaneously for the layout analysis of historical handwritten documents, which established the current state of the art in this research field. The model trains a multi-task fully convolutional network to predict pixel-wise classes and as the final step a heuristic-based post-processing is adopted to reduce noise and correct misclassification. The prediction of the four branches was combined to produce the result of layout analysis tasks. Tran et al. [68] proposed a hybrid method for page segmentation problems. In the first stage, the text and non-text elements are classified by using a

minimum homogeneity algorithm which is the combination of connected component analysis and multilevel homogeneity structure. Then, in the second stage, a new homogeneity structure is combined with an adaptive mathematical morphology in the text document to get a set of text regions. Davoudi et al. [20] proposed a novel method for document layout analysis that reduces the need for labeled data. This method is a dictionary-based feature learning model where a sparse autoencoder is first trained in an unsupervised manner on a document's image patch. Then, the latent representation of image patches is then used to classify pixels into various region categories of the document using a feed-forward neural network. Also, Alaasam et al. [1] used the patching of the document image to train a siamese network model that takes an input a pair of patches and gives as an output a distance that corresponds to their similarity. The trained model is also used to calculate a distance matrix which in turn is used to cluster the patches of a page as either main text, side text or a background patch. Studer et al. [65] tackle the problem of the limited presence of annotated data by introducing the use of pre-trained segmentation models on images from a different domain and then fine-tuning them on handwritten documents. The results demonstrated that on some manuscripts pre-training on ImageNet increases the performance, but on others, the pre-trained network performs much worse. Also Droby et al. [31] try to tackle the problem of the limited presence of ground truth by presenting an unsupervised deep learning method for page segmentation. In this work a Siamese neural network is trained to differentiate between patches using their measurable properties such as the number of foreground pixels so that spatially nearby patches are similar. The network's learned features are used for page segmentation. Finally, Tarride et al. [66] propose the few-shot learning approach Deep&Syntax to segment handwritten registers. Their work uses a hybrid system that exploits recurring patterns to delimit each record, combining U-shaped networks and logical rules such as filter and text alignment. While the presented approaches have different degrees of effectiveness when trying to solve the document layout segmentation task, they all rely on large amounts of data for their training.

3

Datasets and Challenges

This chapter provides a comprehensive overview of the datasets employed in this thesis, covering both industrial and historical document collections. Each dataset presents unique characteristics and challenges, reflecting the diversity of real-world document types and the complexity of their layouts.

For the experiments conducted in this thesis, two datasets were specifically created: TobaccoSeg and U-DIADS-Bib. In particular the U-DIADS-Bib dataset [77] was developed through a collaborative effort between humanities scholars and computer scientists, combining domain expertise with methodological innovation. Alongside these newly developed datasets, the experimental evaluation also makes use of several well-established resources from the literature, which complement our datasets and provide additional variability and benchmarking opportunities.

In addition to presenting these datasets, this chapter also highlights the common challenges encountered in document layout analysis. Understanding these challenges is essential for contextualizing the experiments and motivating the choices made in the design of models and evaluation strategies presented in the following chapters.

The topics addressed in this chapter, including the creation of the datasets and their methodological foundations, have been the subject of the following publications: [77, 78].

3.1 Industrial Dataset

Modern industries generate a vast amount of documentation on a daily basis, including quality control reports, production records, safety checklists, and maintenance logs. This constant flow of documents is crucial for ensuring regulatory compliance, supporting decision-making processes, and maintaining overall operational efficiency. As a result, even medium-sized companies can accumulate extensive archives of heterogeneous, mixed-format documents that must be processed, stored, and analyzed efficiently.

Industrial documents such as forms, invoices, logs, and production notes represent a significant challenge for document image segmentation. A common characteristic of these materials is the coexistence of printed templates with handwritten entries, often filled in by operators directly on the shop floor. Consequently, the documents are not only noisy, affected by handwriting variability, smudges, stains, and scanning artifacts, but also structurally complex, frequently combining tables, free-text notes, stamps, and annotations in irregular layouts. In many cases, operators are required to transcribe information manually into paper forms, which are later digitized and integrated into other digital systems. This multi-step process increases the risk of transcription errors and inefficiencies, further underlining the need for automated segmentation and information extraction. Automating document segmentation and information extraction can drastically reduce time and costs while improving data reliability. More importantly, structured and accurately segmented data unlocks strategic opportunities: from production traceability and predictive maintenance to process optimization. Thus, automation not only addresses operational inefficiencies but also enables advanced data-driven decision-making and long-term competitiveness.

Despite the relevance of industrial documentation, there is a striking lack of publicly available datasets for layout analysis in this domain. Most companies manage large proprietary repositories, but confidentiality and intellectual property concerns prevent their release. The few existing datasets of real industrial documents are limited and are usually labeled for other tasks, such as classification.

Even when available internally, industrial datasets are difficult to exploit: they often suffer from inconsistent annotations, incomplete labeling, or heterogeneous formats. Moreover, unlike general-purpose document collections, they typically contain sensitive data, such as proprietary processes, supplier contracts, or safety incidents, making public release highly restricted. The scarcity of standardized benchmarks forces researchers

dation for the creation of the TobaccoSeg dataset, which is introduced in the following section. Unlike the original dataset, which is mainly employed for classification tasks at the document level, TobaccoSeg was specifically designed with a focus on layout segmentation in real industrial documents.

3.1.1 TobaccoSeg dataset

The TobaccoSeg dataset consists of 50 manually segmented images derived from the Tobacco-3482 dataset [45]. The original Tobacco-3482 dataset contains 3,482 single-page document images, each labeled as one of ten categories: Advertisement, Email, Form, Letter, Memo, News, Note, Report, Resume, and Scientific. These images were sampled from the IIT Complex Document Information Processing Test Collection, which in turn sources its documents from the Truth Tobacco Industry Documents (TTID)¹. The TTID is a large collection of internal documents from major US tobacco companies, released as part of legal settlements [46]. Documents in the TTID have one or multiple manual category annotations that include the ten categories used in Tobacco-3482. The primary task associated with Tobacco-3482 dataset is document classification.

Segmentation analysis

For the creation of TobaccoSeg, we selected 50 images containing handwritten annotations and prepared them for the task of document layout segmentation. In many industrial workflows, operators are required to fill in printed forms and tables by hand, and these handwritten inputs must later be digitized in order to be analyzed and integrated with other digital data sources. The TobaccoSeg dataset was designed to reflect this real-world challenge, focusing on the coexistence of printed and handwritten elements within the same page layout. Common use cases include pre-printed forms and tables annotated manually by operators, as well as financial statements or reports enriched with handwritten notes in the margins. The segmentation of these handwritten parts represents the first step toward automatic transcription into digital text, thus reducing the time required for information acquisition.

Based on these considerations, we defined three segmentation classes:

¹<https://www.industrydocuments.ucsf.edu/tobacco/>

- **Handwritten:** includes all manually written elements, such as text, signatures, numbers, underlines, and checkmarks.
- **Printed:** includes any machine-printed content, such as text, forms, tables, logos, and other printed elements.
- **Background:** covers the background of the page and any visible outline in the scanned image.

Ground truth construction

The annotation and creation of the Ground Truth (GT) masks for TobaccoSeg were performed manually. As previously noted, manual annotation of images is a highly time-consuming task. Consequently, since our goal was to produce a dataset with a large number of annotated pages, pixel-precise segmentation, and minimal, if any, noise, we followed the segmentation pipeline originally proposed for the creation of the U-DIADS-Bib dataset [77], as described in Section 3.2.1, with some modifications.

First, a subset of 10 images per manuscript was selected and binarized using the Sauvola thresholding technique [63] combined with morphological operators, providing an initial basis for annotation. These images were then manually segmented at the pixel level, using different colors to represent only the text and background classes. In this initial step, the two segmentation classes were merged into a single class called “Text”, in order to obtain a less detailed but well-defined segmentation across the entire dataset with minimal noise. The next step involved training a machine learning model on these subsets and their corresponding ground truths to generate a coarse segmentation of the entire dataset. For this purpose, we followed the framework proposed in [24], which implements a few-shot, pixel-precise document layout segmentation method with high performance. Finally, the remaining semantic class (Handwritten) was introduced, and all ground truth masks were carefully refined and corrected by comparing them to the original images. It is important to note that, although the process is computer-aided, the final result is always verified and defined manually, ensuring accuracy and avoiding potential biases or errors in the dataset.

As a result, all document pages in the TobaccoSeg dataset have GT masks annotated with three semantic classes: Handwritten, Printed, and Background. Each pixel within a document image is labeled according to its corresponding content type, ensuring that the dataset provides a precise and unambiguous representation of the layout structure.

Classes	Background	Printed	Handwritten
%	91.33	6.75	1.92

Table 3.1: Classes distribution (%) for TobaccoSeg dataset.

The annotations are encoded using specific RGB values to clearly distinguish between the three classes, allowing both human interpreters and automated algorithms to reliably identify the different regions. The RGB encoding scheme for the three classes is as follows:

- Background: RGB(255, 255, 255), white.
- Printed: RGB(255, 0, 255), magenta.
- Handwritten: RGB(0, 255, 0), lime.

Examples of pages from the dataset, along with their corresponding GT masks, are illustrated in Fig. 3.1. The dataset is divided into 10 images for training, 10 for validation, and 30 for testing. As shown in Table 3.1, TobaccoSeg is characterized by a significant imbalance among the different layout classes, which makes the segmentation task particularly challenging.

Benchmarking and training setup

To analyze our data and provide a benchmark for future studies, we selected five popular deep-learning-based semantic segmentation approaches: FCN [49], Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) [36], DeepLabV3 [15], its enhanced version DeepLabV3+ [16], and Pyramid Scene Parsing Network (PSPNet) [73].

All models were trained using the same configuration: the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5} . Training proceeded for a maximum of 200 epochs with an early stopping mechanism, which monitors the validation loss and halts training if no improvement occurs over 20 consecutive epochs. To ensure that the models undergo a minimum number of iterations and reduce the risk of early convergence to local minima, a buffer of 50 epochs was introduced. All models employ a ResNet34 backbone, except for LRASPP, which uses the MobileNetV3-Large architecture.

Dataset	Metric (weight/macro)	Backbone				
		FCN [49]	LRSAPP [36]	DeepLabV3 [15]	DeepLabV3+ [16]	PSPNET [73]
TobaccoSeg	Precision	0.955/ 0.678	0.941/0.615	0.945/0.612	0.956 /0.647	0.942/0.615
	Recall	0.933/0.854	0.908/0.652	0.909/0.855	0.955/0.869	0.899/0.835
	IoU	0.897/0.623	0.868/0.481	0.868/0.558	0.905/0.626	0.859/0.527
	F1-Score	0.940/ 0.751	0.917/0.585	0.922/0.684	0.945 /0.747	0.915/0.651

Table 3.2: Weighted average and macro average performance of the different semantic segmentation models chosen tested on TobaccoSeg dataset. The best-performing results are shown in bold.

Given the high imbalance between semantic classes in the dataset, we adopted a weighted cross-entropy loss function (Eq. 3.1), where the weight for each class is inversely proportional to its frequency (F_i) within the corresponding manuscript. This weighting scheme encourages the models to focus on under-represented classes, mitigating the effects of class imbalance:

$$W_i = \sqrt{\frac{1}{F_i}} \quad (3.1)$$

Evaluation metrics

The performance of the semantic segmentation tasks is evaluated using Precision, Recall, Intersection over Union (IoU), and F1-Score. The definitions of these metrics are reported in Eqs. 3.2–3.5, where TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.3)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (3.4)$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.5)$$

For each metric, a weighted average is computed based on the class distribution in TobaccoSeg. In addition, the macro average is also reported, which treats all segmentation classes as equally important.

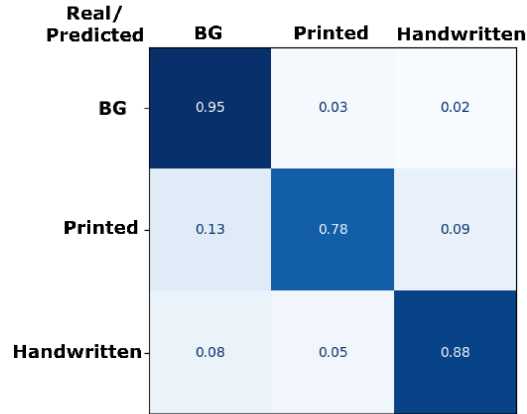


Figure 3.2: Confusion matrices obtained by applying DeepLabV3+ [16] to the TobaccoSeg dataset.

Results

Table 3.2 reports the performance of the selected semantic segmentation models when evaluated on the proposed TobaccoSeg dataset.

Overall, the results confirm that DeepLabV3+ [16] is the most effective architecture among those considered, achieving the best scores across both weighted and macro-averaged metrics. This indicates not only its ability to adapt to the imbalanced class distribution in the dataset but also its robustness when all classes are treated as equally important. The only exceptions are observed for the macro-averaged Precision and F1-Score, where FCN [49] slightly outperforms the other approaches. This outcome suggests that, despite being one of the earliest deep learning architectures for semantic segmentation, FCN still demonstrates competitive performance in terms of minimizing false positives and balancing precision with recall across classes.

To further contextualize these results, we provide in Fig. 3.2 a visualization of the confusion matrices generated by the best-performing model, namely DeepLabV3+. This representation offers a more granular perspective on how the model handles each segmentation class. As shown, DeepLabV3+ achieves high accuracy for the Handwritten and Background classes, while performance decreases considerably for the Printed class.

3.2 Historical Datasets

In recent years, interest in cultural heritage has also been growing within the IT community. Collaboration between computer scientists and humanities scholars is becoming increasingly prominent, fostering interdisciplinary approaches that enrich both fields. For computer scientists, these collaborations provide access to complex cultural heritage data, which can be analyzed using advanced computational methods to uncover patterns, structures, and relationships that may not be immediately apparent. For humanities scholars, the involvement of IT experts significantly streamlines research processes, automating labor-intensive tasks and enabling new kinds of analyses that would be impractical to perform manually. This symbiotic relationship is contributing to a more nuanced and efficient exploration of cultural heritage, bridging the gap between traditional scholarship and modern technological tools.

One of the activities that humanities scholars are particularly interested in is the identification of both the logical and physical structure of documents, as well as their organization. This process is essential for understanding the contents of a manuscript in depth, as it allows researchers to detect structural patterns, segment meaningful units, and trace connections between different works of the same genre or period. The task often involves highlighting and extracting specific sets of elements, such as headings, paragraphs, tables, illustrations, marginalia, or annotations. In the field of IT, this activity is commonly referred to as document layout segmentation, and it encompasses techniques for automatically detecting and classifying various components of a document, often using image processing, pattern recognition, and machine learning methods. By bridging the gap between visual representation and semantic understanding, document layout segmentation plays a crucial role in enabling both automated analysis and enhanced human interpretation of cultural heritage materials.

In the experiments presented in the following chapters, we relied on several historical datasets, namely the U-DIADS-Bib dataset [77], DIVA-HisDB [64], and the Arabic dataset by Bukhari et al. [10]. While the last two datasets are well-established resources in the literature, U-DIADS-Bib was created as part of this research and developed through close collaboration with humanities scholars, ensuring both historical accuracy and methodological rigor. All three datasets are employed in our experimental evaluation, each contributing different challenges and characteristics relevant to the tasks addressed in this thesis.

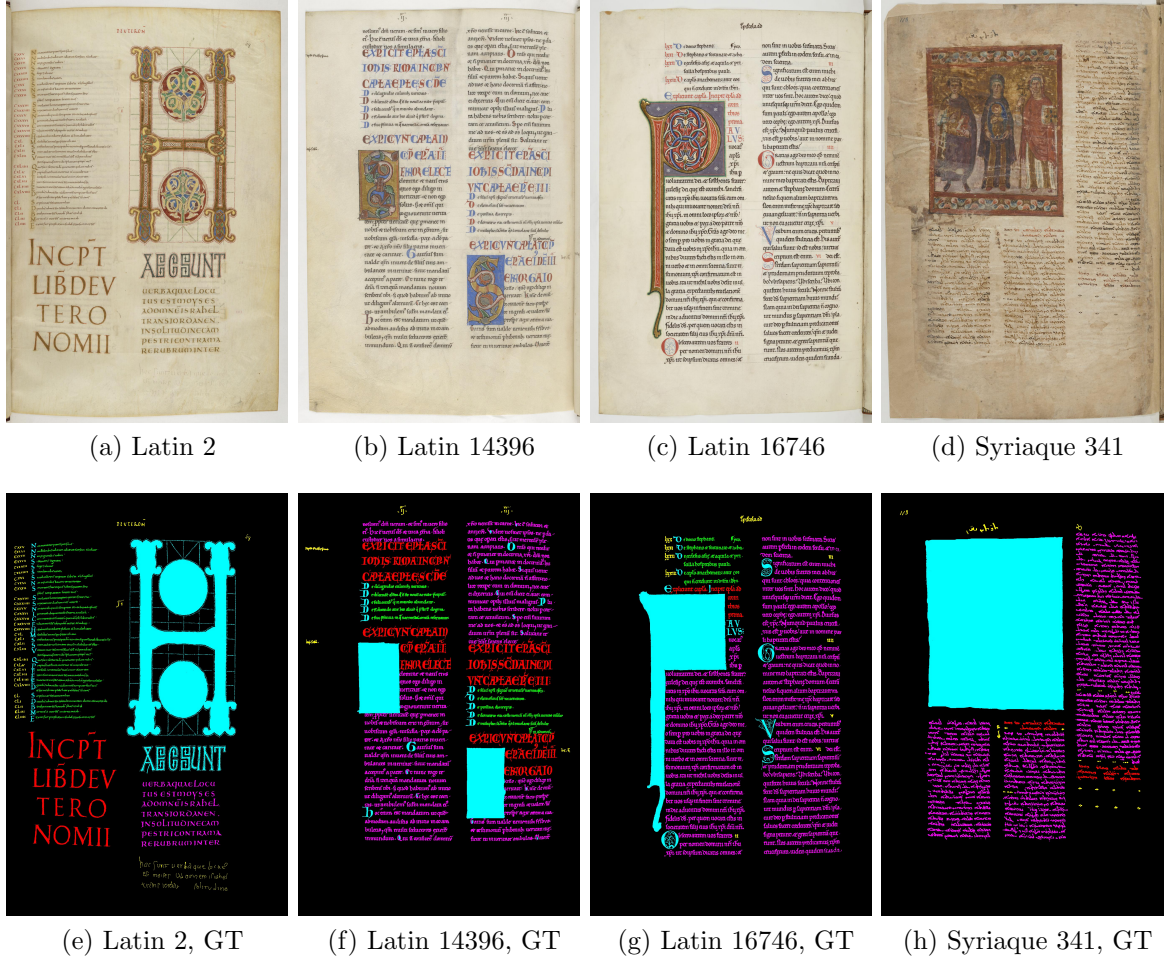


Figure 3.3: Images showing a page of each manuscript (3.3a– 3.3d) of U-DIADS-Bib dataset as well as its corresponding ground truth mask (3.3e– 3.3h), in which each color represents a different semantic class of the document layout.

3.2.1 U-DIADS-Bib

Uniud - Document Image Analysis DataSet - Bible version (U-DIADS-Bib) [77] is composed of 200 images, 50 for each of the 4 different manuscripts that characterize it. These handwritten books were selected in collaboration with humanities scholars partners considering both the complexity of their layout and the presence of significant and semantically distinguishable elements. In particular, the images of the four manuscripts were collected from the French digital library Gallica². All manuscripts are Latin and Syriac Bibles published between the 6th and 12th centuries A.D. which

²<https://gallica.bnf.fr>

will be briefly described hereafter:

- Paris, Bibliothèque nationale de France, Latin 2³. The manuscript, known as the Second Bible of Charles the Bald, was produced between A.D. 871 and 877 at the Abbey of Saint-Amand (Haute-France); it was kept in the Abbey of Saint-Denis between A.D. 877 and 1595 and later transferred to the Royal Library of France. It is composed of 444 parchment pages and the layout is structured in two columns.
- Paris, Bibliothèque nationale de France, Latin 14396⁴. The manuscript was produced between A.D. 1145 and 1150 in the Abbey of Saint-Victor (Paris). It contains the biblical text from the Book of Ezra to the Book of Revelation and it is probably the final part of a three-volume Bible; this codex, known as Genesis-Kings, is the first volume. The manuscript is composed of 170 parchment pages and the layout is structured in two columns.
- Paris, Bibliothèque nationale de France, Latin 16746⁵. The manuscript was produced between A.D. 1170 and 1190 at the Abbey of Saint Bertin (Pas-de-Calais). It contains the New Testament and it is the final part of a four-volume Bible (Paris, Bibliothèque nationale de France Latin 16743-16746). The Bible had been kept for a long time in the Capuchin convent of Saint-Honoré, in Paris. The manuscript is composed of 176 parchment pages and the layout is structured in two columns.
- Paris, Bibliothèque nationale de France, Syriac 341⁶. The manuscript was produced between the 6th and 7th centuries A.D. probably in the Monastery of Baquqa in Iraq. It contains the Old Testament in the Syriac Peshitta version and portions of the New Testament. The manuscript is composed of 256 parchment and paper pages, and the layout is structured in three columns.

For the dataset, we selected a representative set of 50 pages from each manuscript. Fig. 3.3 illustrates some examples of the defined GT and corresponding original image for each manuscript of the U-DIADS-Bib. Each selected pixel is marked by a color that symbolizes the corresponding content type.

³<https://gallica.bnf.fr/ark:/12148/btv1b8452767n>

⁴<https://gallica.bnf.fr/ark:/12148/btv1b84429190>

⁵<https://gallica.bnf.fr/ark:/12148/btv1b85144288>

⁶<https://gallica.bnf.fr/ark:/12148/btv1b10527102b>

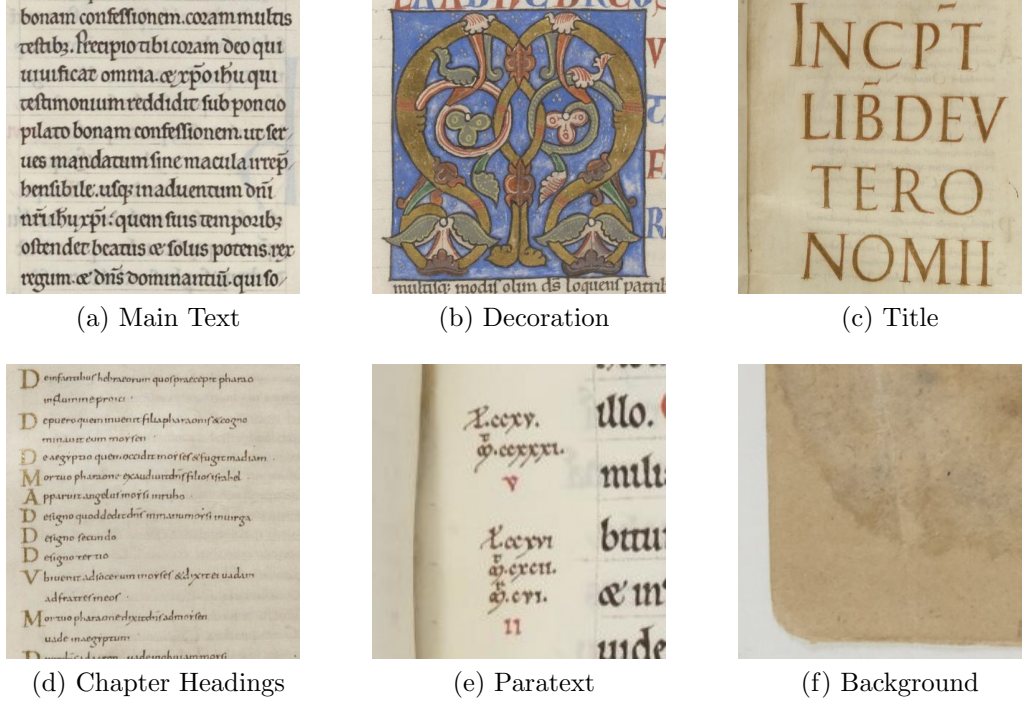


Figure 3.4: Samples of six segmentation classes in U-DIADS-Bib dataset: Main Text (3.4a), Decoration (3.4b), Title (3.4c), Chapter Headings (3.4d), Paratext (3.4e) and Background (3.4f).

Segmentation analysis

The choice of biblical texts with the homogeneity of the contents, and the abundance of material from different centuries, scriptures and layouts that characterize them, allows for the challenge of better highlighting the differences in the graphic and textual elements that accompany the text but are not part of it. It is important to remark that these decisions were supported by scholars in the humanities, which are experts in the respective area. Indeed, from a historical perspective, the Bible holds immense cultural significance as one of the most widespread and influential texts in the world. It has profoundly impacted human societies across different spatial and temporal contexts. In particular, the biblical manuscripts represent a rich diversity of historical layouts and textual variations, due to their wide diffusion over the centuries and in the world.

The six segmentation classes highlighted by humanities scholars for the page segmentation task are visible in Fig.3.4 and are:

- **Main Text.** This class includes the writing area and represents the core and

central content of the book. This class includes punctuation and pause marks. It can be structured in different layouts such as one or two columns.

- **Decoration.** This class includes both figurative elements in the proper sense (miniatures and decorated initials), and elements of minimal decoration, such as initials with a graphic element or color distinguishing them from the rest of the text.
- **Title.** More properly *'incipit and explicit formulae'*, identifiable by the use of a different ink color and/or by the adoption of display scripts: monumental (square) or rustic capital, uncials or mixed capital/uncials script.
- **Chapter Headings.** The chapter headings function in ancient manuscripts was to facilitate the retrieval of a particular chapter or passage, but according to different scansion and interpretation from one set to another. From a graphical point of view, they are often recognizable by a script similar to that of the main text but smaller in size.
- **Paratext.** This class consists of several elements such as glosses, marginal and interlinear notes, corrections, paragraph numbering, possession notes from different periods, page or fascicle numbering. In general, includes all hand annotations that do not fit into the other classes.
- **Background.** This class includes the background of the page and any outline visible in the scanned image.

However, there is also a twofold motivation behind this choice from a humanities scholars point of view: firstly, the classification of paratexts is in many cases neither unambiguous nor immediate (e.g. it is not easy to distinguish at a glance a textual variant from the integration of a passage omitted by mistake in the original text or the correction of an erroneous text), and the use of broader classes avoids stiffening the interpretation of paratexts into dead ends. Furthermore, the broader aim is to encourage a wider use of document layout analysis from an interdisciplinary point of view: the extraction of paratextual elements of different natures can in fact provide material for scholars of different historical, textual and librarian disciplines, from philology to paleography to art history to the history of ancient and modern libraries.

Ground truth construction

The annotation of the U-DIADS-Bib is the result of the collaboration between humanities scholars and computer scientists. As previously addressed, manual annotation of images is a very time-consuming task, especially in the context of document segmentation where the different layout components can be very small and detailed, whilst annotations provided by algorithms tend to present many inaccuracies and are prone to the introduction of noise. Consequently, since our goal was to produce a dataset with a large number of annotated pages, pixel-precise segmentation and very limited, if any, noise, we have defined a segmentation pipeline that involves the alternation of humanities scholars and algorithmic work, in order to optimize the expected results.

First of all, after having chosen the manuscripts, a subset of 50 images was selected in such a way as to represent all the chosen segmentation classes for each manuscript. A subset of 10 images per manuscript was selected and binarized using the Sauvola threshold technique [63] and morphological operators in order to give the human experts a starting point to work on. Then, experts in manuscript texts have manually segmented at pixel level with different colors these few images per manuscript that contain examples of all expected classes.

The next step was to train a machine learning model with these subsets of images and the related GTs segmented by humanities scholars to obtain a coarse segmentation of the entire dataset. To achieve this, the framework proposed in [24] was followed, where a few-shot pixel-precise document layout segmentation method with high performance is presented. The segmentation was done for 4 classes, so as to obtain a less detailed but well-defined segmentation on the whole dataset and with the presence of almost zero noise. Finally, the humanities scholars introduced the other missing semantic classes and meticulously refined and corrected all the color masks of the GTs by comparing them to the original images. It is worth noting that, despite the task being computer-aided, the final result is always defined by a human expert, thus avoiding possible biases or errors in the dataset.

U-DIADS-Bib is therefore composed of 50 original color page images for each manuscript, stored in JPEG image format with resolution 1344×2016 px. Each page is associated with the corresponding GT data, stored in a PNG image with the same size as the original one. GTs contain six different and non-overlapping annotated classes (background, comment, decoration, text, title and chapter headings) encoded by RGB

Classes	Lat 2/FS	Lat 14396/FS	Lat 16746/FS	Syr 341/FS
Background	92.8/92.9	89.2	88	85.1
Paratext	0.1	0.1	0.3	0.2
Decoration	1.5/1.3	2	3	2.8/2.6
Main Text	4.7/4.8	7.6/7.5	7.8	11.9/12
Title	0.4	0.5	0.1	0.1
Chapter Headings	0.5	0.6/0.7	0.8	0

Table 3.3: Pixel-level class distribution (%) for each manuscript in U-DIADS-Bib and its few-shot variant U-DIADS-BibFS (FS). Values in FS are omitted when identical to those in U-DIADS-Bib.

value as follow:

- Background (BG): RGB(0,0,0), Black.
- Paratext: RGB(255,255,0), Yellow.
- Decoration: RGB(0,255,255), Cyan.
- Main Text: RGB(255,0,255), Magenta.
- Title: RGB(255,0,0), Red.
- Chapter Headings (CH): RGB(0,255,0), Lime.

The Syriac 341 manuscript contains five semantic classes and the missing one is the Chapter Headings.

For each of the manuscripts included in U-DIADS-Bib, 10 images have been selected for the training set, 10 for the validation set, and the remaining 30 for the test. As shown in Table 3.3 the proposed dataset is characterized by a noticeable imbalance between the different layout classes, making the layout segmentation task even more challenging, due to the fact that the classes that appear less consistently in the layout of the selected pages are much harder to identify for the models employed for the segmentation task.

Moreover, as we believe that real-world use of layout analysis techniques will be broadly adopted only if the amount of required manually-labeled training data is reduced, we also provided a standardized few-shot version of our dataset (U-DIADS-BibFS), which we hope will encourage further efforts towards this goal. U-DIADS-BibFS offers 43 images per manuscript, where only 3 images are provided for training,

10 for validation and 30 for testing. For each manuscript, the 3 training images were appropriately chosen so that they included all the segmentation classes. The detail of the semantic component distributions of this few-shot dataset are shown in Table 3.3.

The dataset is freely available for download on the U-DIADS-Bib repository⁷. This dataset was also employed in the International Competition on Few-Shot and Many-Shot Layout Segmentation of Ancient Manuscripts (SAM), a challenge we organized at ICDAR 2024 [78].

Benchmarking and training setup

In order to analyze our data and to provide a benchmark for future studies, we selected a set of 5 popular deep-learning-based semantic segmentation approaches, namely FCN [49], Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) [36], DeepLabV3 [15], its improvement represented by DeepLabV3+ [16], and Pyramid Scene Parsing Network (PSPNet) [73]. Also, we include the results obtained by state-of-the-art for the task of document layout segmentation in a few-shot learning setting, presented by De Nardin et al. [23].

All the models were trained using the same setup: the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5} . The models are trained for a maximum of 200 epochs with an early stop mechanism that monitors the validation loss at every epoch and stops the training process if it did not improve over the last 20 consecutive epochs. However, to ensure that the model is trained for a minimum number of iterations, thus reducing the risk of convergence to a local minimum at the beginning of the training process, a buffer of 50 epochs is added.

Due to the high imbalance between the semantic classes in each document category of the dataset, we have decided to adopt a weighted cross-entropy loss function (Eq. 3.6), where the weight for each class is inversely proportional to the frequency of that element in the corresponding manuscript (F_i). The idea behind this choice is to allow the selected segmentation models to focus more on the classes that are less represented in the dataset, therefore mitigating the challenge introduced by its imbalanced nature.

$$W_i = \sqrt{\frac{1}{F_i}} \quad (3.6)$$

All models rely on a residual network as their backbone, in particular the ResNet34

⁷<https://sites.google.com/view/avml-lab-udiadsbib/home/>

version, except LRASPP which is based on MobileNetV3-Large architecture.

Evaluation Metrics

The performances for these semantic segmentation tasks are measured using Precision, Recall, Intersection over Union (IoU) and F1-Score. Metric definitions are reported in Eq. 3.7– 3.10, where TP, FP and FN stand respectively for True Positives, False Positives and False Negatives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.7)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.8)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (3.9)$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.10)$$

For each metric, a weighted average is performed, based on each class distribution in each manuscript. Furthermore, the macro average for each metric is also presented to consider the weight of all segmentation classes equally important. These evaluation metrics are calculated individually for each manuscript, and a final evaluation of a model is obtained by averaging the metrics of the four handwritten documents.

Results

Table 3.4 shows the different performances of the aforementioned semantic segmentation models on the proposed U-DIADS-Bib dataset. The results are provided both for a single manuscript and for the full dataset by averaging the scores obtained for each manuscript class. In particular, both the values of the weighted average and the macro average of the metrics are presented. The reason behind this decision is that a consensus on which of the two averaging approaches is better suited for the layout segmentation task has not been reached yet. In fact, while it would be tempting to consider all segmentation classes as equally important this could lead to penalizing excessively the score of a model that made some minor mistake on one of the classes that appear more sparsely in the dataset, which is not always the desired effect in a real-world application. As expected the macro averaged scores tend to be much lower

Manuscript	Metric (weight/macro)	Backbone				
		FCN [49]	LRSAPP [36]	DeepLabV3 [15]	DeepLabV3+ [16]	PSPNET [73]
Latin 2	Precision	0.955/0.542	0.953/0.489	0.954/0.491	0.962/0.593	0.932/0.217
	Recall	0.926/0.708	0.898/0.499	0.907/0.811	0.936/0.828	0.890/0.297
	IoU	0.892/0.425	0.862/0.373	0.870/0.451	0.906/0.544	0.851/0.200
	F1-Score	0.935/0.535	0.916/0.460	0.923/0.585	0.945/0.670	0.904/0.235
Latin 14396	Precision	0.953/0.519	0.939/0.616	0.951/0.590	0.961/0.659	0.951/0.560
	Recall	0.936/0.752	0.906/0.589	0.918/ 0.902	0.944/0.842	0.917/0.779
	IoU	0.898/0.500	0.867/0.417	0.875/0.556	0.910/0.609	0.873/0.531
	F1-Score	0.940/0.586	0.919/0.531	0.928/0.683	0.949/0.730	0.927/0.633
Latin 16746	Precision	0.948/0.564	0.945/0.614	0.947/0.586	0.958/0.664	0.926/0.389
	Recall	0.937/0.787	0.893/0.892	0.905/ 0.953	0.939/0.932	0.899/0.460
	IoU	0.897/0.541	0.842/0.555	0.856/0.560	0.900/0.633	0.845/0.362
	F1-Score	0.939/0.638	0.908/0.694	0.917/0.693	0.944/0.757	0.906/0.412
Syriaque 341	Precision	0.942/0.669	0.912/0.575	0.927/0.625	0.943/0.745	0.916/0.491
	Recall	0.928/ 0.851	0.878/0.436	0.905/0.739	0.931/0.686	0.899/0.528
	IoU	0.879/ 0.600	0.808/0.353	0.846/0.551	0.883/0.530	0.837/0.448
	F1-Score	0.932/ 0.719	0.885/0.433	0.912/0.665	0.934/0.617	0.904/0.509
Mean	Precision	0.950/0.574	0.937/0.574	0.945/0.573	0.956/0.665	0.931/0.414
	Recall	0.932/0.775	0.894/0.604	0.909/ 0.851	0.938/0.822	0.901/0.516
	IoU	0.892/0.517	0.845/0.425	0.862/0.530	0.900/0.579	0.852/0.385
	F1-Score	0.937/0.620	0.907/0.530	0.920/0.657	0.943/0.694	0.910/0.447

Table 3.4: Comparison of the weighted average and macro average performance of the different semantic segmentation models chosen tested on U-DIADS-Bib dataset. The best-performing results for each manuscript and for full dataset are shown in bold.

than their weighted average counterparts as typically the least represented classes are also the hardest ones to classify correctly.

As we can see, the model that presents the overall best performance for both weighted and macro averages on the full dataset is DeepLabV3+ [16]. The only exception is represented by the macro averaged recall, for which DeepLabV3 [15] achieves the best results. A possible reason behind this is that both systems introduce, as part of their architecture, a way of effectively leveraging the global context of the image instead of just relying on local information. In contrast, the FCN, which doesn't implement any kind of global information, shows a much lower performance on the task at hand. While LRSAPP and PSPNet also rely on global information to perform the segmentation, they likely need a much larger amount of data to be properly trained and reach convergence.

In order to provide a more in-depth insight into the results obtained on the presented dataset we provide a visualization of the confusion matrices obtained by applying the best performing model, namely DeepLabV3+, on each of the manuscript classes that make up U-DIADS-Bib (Fig. 3.5). We can notice how each of the manuscripts presents different criticalities, in terms of segmentation classes that are particularly difficult to

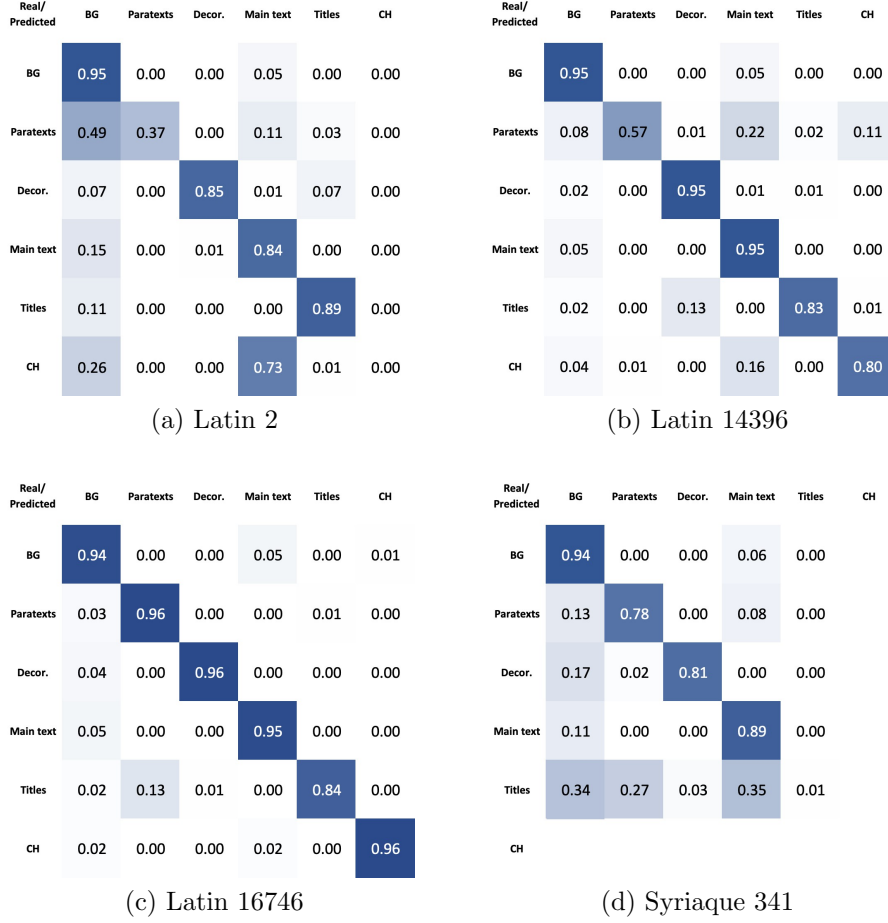


Figure 3.5: Confusion matrices obtained by applying DeepLabV3+ [16] to the four ancient manuscripts of the U-DIADS-Bib dataset.

identify correctly. In particular, DeepLabV3+ obtained particularly low accuracies for the Paratext and Title classes on the Latin 14396 and Syriac 341 manuscripts respectively, while both classes represented a challenge when working with the Latin 2 manuscript. The performance on the Latin 16746 manuscript, on the other hand, appears to be consistent for all the segmentation classes, with only a small drop in accuracy for the Title.

Few-shot setting

As previously stated, as part of the present work we wanted to emphasize the importance of being able to work with a small amount of data while retaining good

Manuscript Few-Shot	Metric (weight/macro)	Backbone					
		FCN [49]	LRSAPP [36]	DeepLabV3 [15]	DeepLabV3+ [16]	PSPNET [73]	De Nardin et al. [23]
Latin 2	Precision	0.873/0.225	0.907/0.559	0.939/0.442	0.958/0.537	0.928/0.355	0.973/0.636
	Recall	0.930/0.194	0.927/0.203	0.897/0.570	0.932/0.756	0.868/0.296	0.972/0.787
	IoU	0.867/0.177	0.809/0.186	0.863/0.369	0.902/0.485	0.887/0.190	0.950/0.547
	F1-Score	0.899/0.200	0.871/0.216	0.914/0.467	0.941/0.602	0.822/0.227	0.970/0.685
Latin 14396	Precision	0.845/0.426	0.852/0.398	0.945/0.549	0.958/0.705	0.925/0.377	0.971/0.719
	Recall	0.908/0.612	0.888/0.234	0.908/ 0.794	0.941/0.789	0.912/0.447	0.973/0.718
	IoU	0.833/0.408	0.866/0.206	0.863/0.496	0.905/0.612	0.915/0.350	0.949/0.629
	F1-Score	0.874/0.485	0.808/0.257	0.920/0.618	0.946/ 0.736	0.860/0.405	0.970/0.707
Latin 16746	Precision	0.839/0.312	0.897/0.472	0.934/0.525	0.941/0.560	0.875/0.246	0.952/0.714
	Recall	0.880/0.553	0.869/0.547	0.887/ 0.895	0.911/0.731	0.858/0.467	0.959/0.743
	IoU	0.810/0.289	0.867/0.288	0.839/0.483	0.863/0.464	0.857/0.228	0.921/0.636
	F1-Score	0.858/0.341	0.796/0.378	0.904/0.610	0.920/0.598	0.789/0.282	0.951/0.761
Syriaque 341	Precision	0.755/0.320	0.861/0.308	0.890/0.487	0.937/ 0.654	0.895/0.476	0.946/0.602
	Recall	0.862/0.329	0.883/0.321	0.864/0.502	0.926/0.520	0.862/0.436	0.949/0.549
	IoU	0.745/0.277	0.802/0.269	0.789/0.365	0.873/0.456	0.787/0.354	0.907/0.473
	F1-Score	0.804/0.323	0.872/0.314	0.872/0.463	0.928/0.546	0.871/0.436	0.946/0.564
Mean	Precision	0.828/0.321	0.879/0.434	0.927/0.501	0.949/0.614	0.906/0.364	0.961/0.668
	Recall	0.895/0.422	0.917/0.326	0.889/0.690	0.928/ 0.699	0.875/0.412	0.963/0.699
	IoU	0.829/0.288	0.879/0.237	0.859/0.428	0.900/0.504	0.883/0.281	0.942/0.571
	F1-Score	0.844/0.337	0.819/0.291	0.882/0.540	0.920/0.621	0.815/0.338	0.950/0.679

Table 3.5: Comparison of the weighted average and macro average performance of the different semantic segmentation models chosen and the current state-of-the-art tested on U-DIADS-BibFS dataset. The best-performing results for each manuscripts and for full dataset are shown in bold.

performance for the task of document layout analysis. For this reason, in Table 3.5, we also present the results obtained by the different segmentation approaches on U-DIADS-BibFS, the few-shot version of the proposed dataset. As previously described, only three images were adopted for the training process. Consistent with the full version of the dataset, the results are provided both for a single manuscript and for the full dataset and for weighted and macro averages.

In this setting, the De Nardin et al. [23], which represents the current state-of-the-art for few-shot document layout segmentation, achieved the best results for all the metrics, both when using the weighted and macro averaging strategies on the full dataset, while being outperformed by either DeepLabV3 or DeepLabV3+ on some of the metrics when looking at the individual manuscripts scores.

As we did for the full dataset, in Figure 3.6 we provide the results obtained by the best-performing model on U-DIADS-BibFS. As expected, for most of the manuscripts we can observe a deterioration of the scores obtained for the different classes. In particular, we can notice a substantial drop in accuracy for the Paratext and Decoration classes on the Latin 14396 and Syriaque 341 manuscripts. Similarly, the Decoration and CH classes of the Latin 16746 manuscript appear to be much harder to identify in the few-shot learning setting compared to the full dataset one. Surprisingly, the

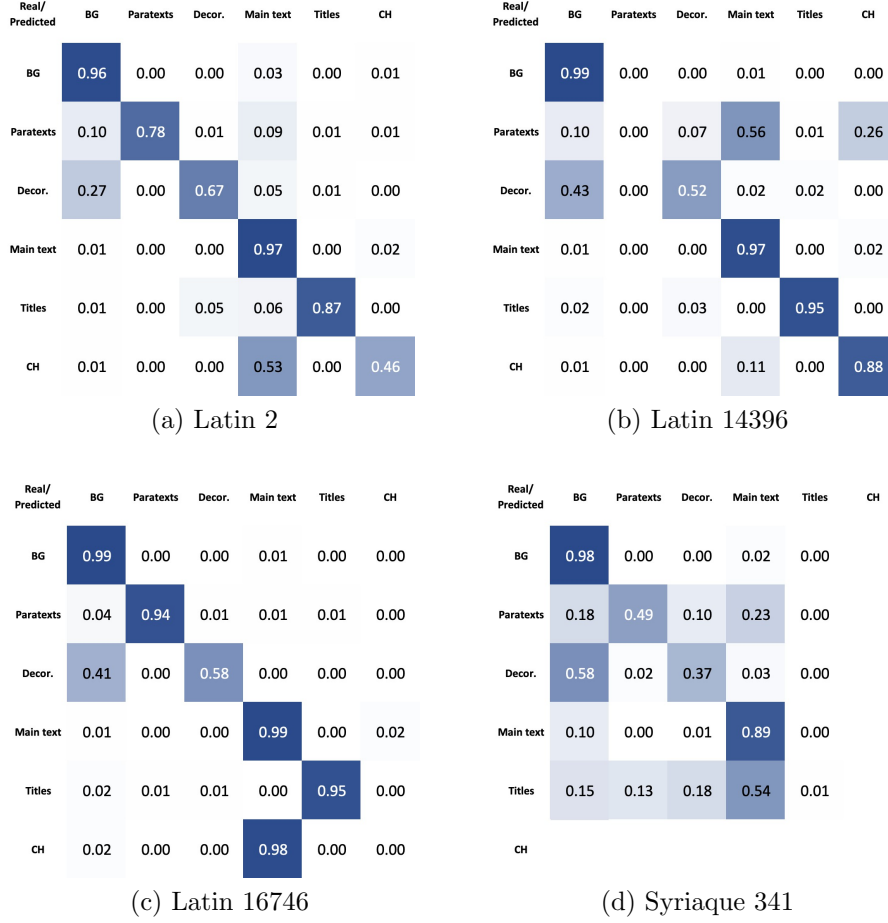
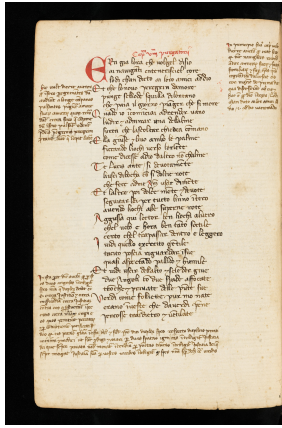


Figure 3.6: Confusion matrices obtained by applying the model proposed by De Nardin et al. [23] to the four ancient manuscripts of the U-DIADS-BibFS dataset.

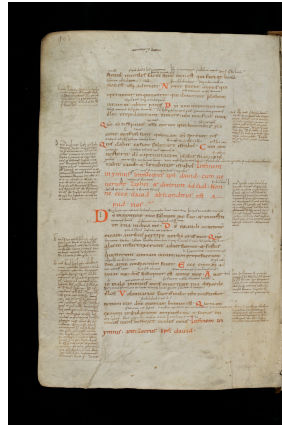
performance on the Latin 2 manuscript is overall better in this scenario, where we can notice a substantial improvement in accuracy for the Paratext, Main Text and Chapter Headings classes respectively, counterbalanced only by a noticeable drop in accuracy for the Decoration class.

3.2.2 DIVA-HisDB

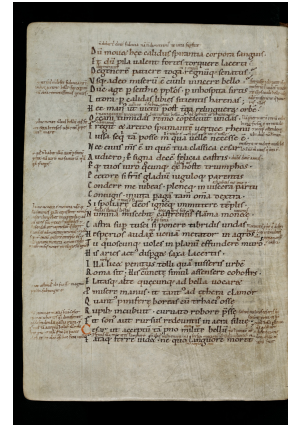
The DIVA-HisDB dataset [64] is one of the most widely used benchmarks for document layout analysis. It consists of a total of 150 images, equally distributed across three different manuscripts (50 per manuscript). Out of these, 60 images are commonly used for training, 30 for validation, and the remaining 60 for testing. The manuscripts were



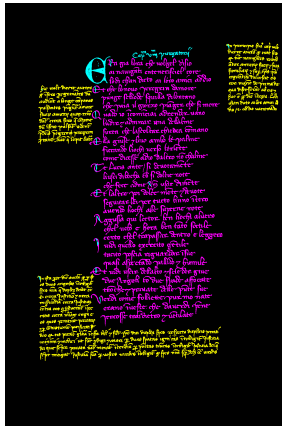
(a) CB55



(b) CSG18



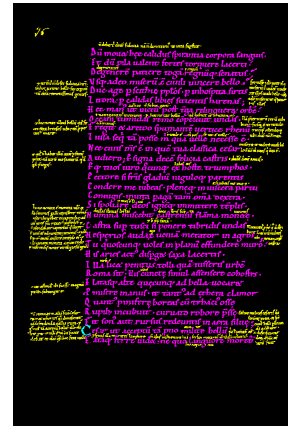
(c) CSG863



(d) CB55, GT



(e) CSG18, GT



(f) CSG863, GT

Figure 3.7: Examples from the DIVA-HisDB [64] dataset (3.7a– 3.7c) with corresponding ground truth masks (3.7d– 3.7f). Magenta highlights the main text, yellow the comments, cyan the decorations, and black the background.

carefully selected for their heterogeneity and the complexity of their layouts, making the dataset a particularly challenging benchmark. The three manuscripts included in DIVA-HisDB are:

- Cod. Bodmer 55 (CB55), Cologny-Genève, Fondation Martin Bodmer. Known as the Codex Guarneri, this manuscript was written in Italy during the first half of the 14th century. It contains Dante’s Divina Commedia (Inferno and Purgatorio), written in chancery script and enriched with marginal glosses in Latin.
- Cod. Sang. 18 (CSG18), St. Gallen, Stiftsbibliothek. The portion included in the dataset corresponds to codicological unit 4, which contains a Psalterium

Manuscript	Background	Comment	Decoration	Text
CB55	82.41	8.36	0.55	8.68
CSG18	85.16	6.78	1.47	6.59
CSG863	77.82	6.35	1.83	14.00

Table 3.6: Classes distribution (%) for each manuscripts of Diva-HisDB [64] (CB55, CSG18 and CSG863) dataset.

glossatum with extensive commentary. It was most likely written in St. Gall in the second half of the 11th century.

- Cod. Sang. 863 (CSG863), St. Gallen, Stiftsbibliothek. Dating to the second half of the 11th century, this manuscript contains Lucan’s *Pharsalia libri decem*. Each book begins with a decorated red initial and two lines in display scripts (uncial and capitalis rustica). Red incipits and explicits frame the text, while marginal commentary, interlinear glosses, reference marks, and corrections further increase the layout complexity.

All documents in the dataset are annotated with four semantic classes: main text, comments, decorations, and background. Examples of the manuscript images alongside their pixel-level ground truth masks are shown in Fig. 3.7. A detailed breakdown of the semantic component distributions across the three manuscripts is reported in Tab. 3.3.

3.2.3 Bukhari dataset

For some of the experiments presented in this thesis, we made use of the Arabic dataset introduced by Bukhari et al. [10]. This dataset represents the most widely adopted benchmark for the task of document layout segmentation on historical Arabic manuscripts and has been frequently employed in the literature as a reference point for evaluating the effectiveness of segmentation frameworks.

It consists of 32 images, each corresponding to a page from one of three different Arabic historical manuscripts. The pages were scanned from a private library located in the old city of Jerusalem, as well as from the Islamic Manuscripts Digitization Project at the Leipzig University Library⁸.

Of the 32 samples, 24 are typically used for training, while the remaining 8 are reserved for testing. The dataset defines three semantic segmentation classes: main

⁸<http://www.islamic-manuscripts.net>

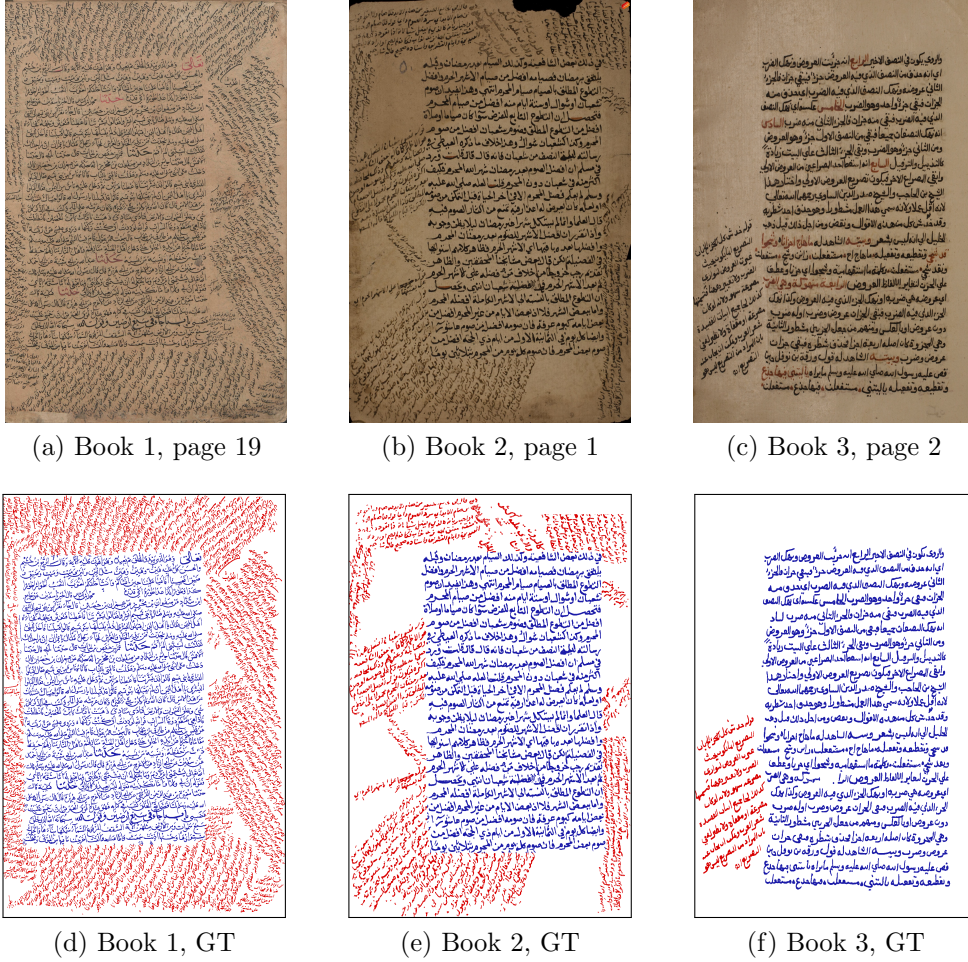


Figure 3.8: Sample pages (3.8a– 3.8c) and GT masks (3.8d– 3.8f) from the Bukhari et al. [10] dataset. Blue marks the main text, red the side text, and white the background.

text, side text, and background. A distinctive feature of this dataset is its irregular layout format and various writing styles, which poses a significant challenge for segmentation methods. The manuscripts often include marginal notes (side text) serving as commentary on the main text contained within the page frame. While the main text is mostly arranged in horizontal lines, side notes exhibit highly variable orientations and positions across the page margins with diverse writing styles. This variability contributes to the complexity and irregularity of the page layout.

Both main text and side text are annotated at the pixel level. Examples of document images together with their ground truth masks are illustrated in Fig. 3.8. A detailed distribution of the semantic components is reported in Tab. 3.7.

Classes	Background	Side Text	Main Text
%	86.07	4.71	9.22

Table 3.7: Classes distribution (%) for Bukhari et al. [10] dataset.

3.3 Common Challenges

Document layout segmentation presents several challenges across both industrial and historical domains. Although the types, purposes, and time periods of documents may differ, several recurring obstacles complicate dataset construction, model training, and evaluation. Figure 3.9 illustrates some typical difficulties encountered in handwritten document segmentation. These examples highlight how the task of layout segmentation requires not only sophisticated algorithms capable of handling diverse structural variations but also carefully designed datasets that faithfully capture the complexity of real-world documents.

Unlike printed documents such as scientific articles or newspapers, which generally follow standardized templates with well-aligned text blocks, tables, and headings, handwritten materials present a radically different scenario. Manuscripts, notes, and annotated forms often feature irregular layouts, including variable line spacing, changing text orientations, marginalia, interwoven illustrations, and inconsistent alignment. This stark contrast between structured printed material and free-form handwritten content makes generalization across document types difficult and demands methods capable of handling both highly regular and highly irregular layouts.

A central challenge is the high variability of document structures. Industrial forms may differ not only between organizations but also across departments or over time, combining tables, handwritten notes, stamps, and annotations in inconsistent ways. Historical manuscripts exhibit even greater heterogeneity, with diverse column structures, ornamental borders, marginalia, illustrations, and multiple handwriting styles. Such variability requires models that can generalize without overfitting to specific templates or scripts.

Noise and degradation further complicate layout segmentation. Historical documents often show ink fading, bleed-through, stains, tears, or warped pages, while industrial records may suffer from smudges, scanning artifacts, uneven lighting, or low-resolution captures. These issues obscure layout boundaries, making pixel-level segmentation particularly challenging. The problem is exacerbated by the scarcity of



Figure 3.9: Examples of difficulties and challenges present in document images from the U-DIADS-Bib, TobaccoSeg dataset, and DIVA-HisDB.

reliable GT annotations. Producing accurate GT requires significant time and domain expertise, since distinguishing elements such as main text, paratext, decorations, or handwritten notes is far from trivial. Consequently, most DLA datasets remain small, limiting the ability of models to generalize effectively.

Two additional obstacles are class imbalance and inter-class similarity. In many collections, body text dominates, while titles, marginalia, decorations, or annotations are far less frequent, biasing models toward majority classes. At the same time, visually similar regions, for example, annotations blending with marginal notes or headings resembling body text, make consistent classification difficult, even for human annotators. Together, these issues complicate both training and evaluation, especially in low-data regimes.

These challenges have direct implications for machine learning. Standard bounding-box methods, effective for printed material, are inadequate for handwritten or degraded

documents. Pixel-level segmentation is often required to capture small and complex elements. Ground truth maps are central to this process, serving both as training targets and as the basis for quantitative evaluation. However, producing pixel-accurate annotations is notoriously difficult. On the one hand, the task demands expert knowledge of historical scripts and layouts; on the other, document heterogeneity, stemming from varying degradation, irregular structures, ink stains, scratches, or miscalibrated scans, makes automation unreliable. As a result, GT creation remains a bottleneck, introducing noise when automated and consuming considerable resources when performed manually.

While recent years have seen efforts to automate segmentation, success has been limited largely to well-structured printed documents. Handwritten and heavily edited manuscripts remain particularly challenging due to inconsistent acquisition conditions, numerous corrections and additions, and widespread degradation. These characteristics render bounding-box methods ineffective and necessitate pixel-level segmentation maps. Compared to natural image segmentation, however, manuscript regions are smaller, irregular, and harder to annotate, requiring careful expert supervision. The resulting scarcity of data, most datasets contain only a few dozen training images, further limits model performance.

The consequences are significant. Models trained on small datasets risk overfitting, achieving high accuracy on training data but poor generalization to unseen documents. Domain variability compounds the problem: industrial forms evolve across organizations and time, while historical manuscripts differ drastically in script styles, layouts, and degradation patterns. Robust, generalizable segmentation models must therefore contend with domain shifts as the norm rather than the exception.

Finally, the scarcity of open datasets contrasts sharply with the vast amount of documents produced in industrial contexts. Although data exist in abundance, restricted access prevents large-scale training. This reality underscores the need for methods capable of working under limited data conditions. Addressing this challenge is the primary objective of this thesis, which focuses on data-efficient strategies such as transfer learning and few-shot learning to develop robust and generalizable document layout segmentation models.

4

Effective Transfer Learning for Document Layout Analysis

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 truth. A notable example is represented by the task of Document Layout Analysis of handwritten documents, which requires pixel-precise segmentation maps to highlight the different layout components of each document page. The creation of segmentation maps is typically very time-consuming and requires a high degree of domain knowledge to be defined, as they are intrinsically characterized by the content of the text to be segmented.

For this reason, in the present chapter, we explore the effects of different transfer learning strategies for the task of document layout segmentation, to provide a better understanding of which of them allows for a maximal reduction in the amount of data needed for the target task while retaining the performance of the employed model as much as possible. In particular, we study the effects of both in-domain and cross-domain datasets for the pre-training of a selected segmentation model, which is then fine-tuned on a set of popular, publicly available, document layout segmentation 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 pre-training data leads to the best overall performance of the models, as well as speeding up their convergence process, regardless if the whole model is fine-tuned on the downstream task dataset, or just the decoder module.

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 small amounts, leads to a substantial improvement in performance on the downstream task, especially when working with few training instances on the target dataset.

The topics addressed in this chapter have been the subject of the following publications: [21, 27].

4.1 Methods

This section presents a detailed overview of the selected architecture and training configuration, along with a description of the datasets employed in the experiments and the evaluation metrics used.

4.1.1 Segmentation architecture

As the semantic segmentation model for our experiments, we opted for the popular DeepLabV3 [14] 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

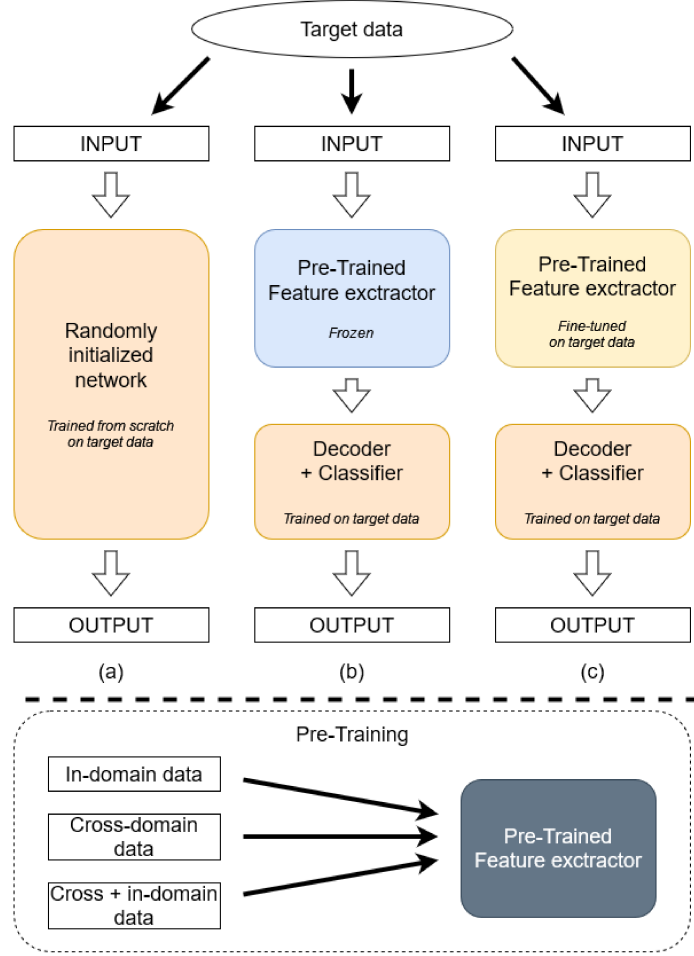


Figure 4.1: Visual representation of the training strategies explored: (a) training from scratch (baseline), (b) transfer learning with a frozen feature extractor, and (c) full model fine-tuning. For (b) and (c), pre-training was performed using in-domain, cross-domain, or combined data.

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.

4.1.2 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.4.1.

Upstream data sources

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 ImageNet-1K, 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 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 [60].

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.

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.

Transfer learning pipelines

The two transfer learning pipelines we studied are shown in Fig. 4.1(b) and (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.

Tab. 4.1 provides a schematic overview of all the datasets presented in this section.

4.1.3 Pre-training datasets

In this section, we will give a brief description of the 3 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 2 large cross-domain datasets, namely ImageNet-1K [29] and

Dataset	# Images	# Classes	#Images type	Task	Role
ImageNet-1K	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
TobaccoSeg	50	3	Documents	Segmentation	Target

Table 4.1: Compact overview of key information regarding the datasets used for our analysis.

COCO [47] 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 documents, called U-DIADS-Bib [77]. An example image from each pre-training dataset is shown in Fig. 4.2.

ImageNet-1k

ImageNet-1k [29] is a hierarchically structured dataset consisting of 1,281,167 natural images, primarily designed for the image classification task. Its instances are organized into 1,000 distinct categories, covering a wide range of everyday objects and scenes. 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. Its large scale, diversity, and standardized structure have made it an essential benchmark for evaluating and comparing the performance of deep learning architectures.

COCO

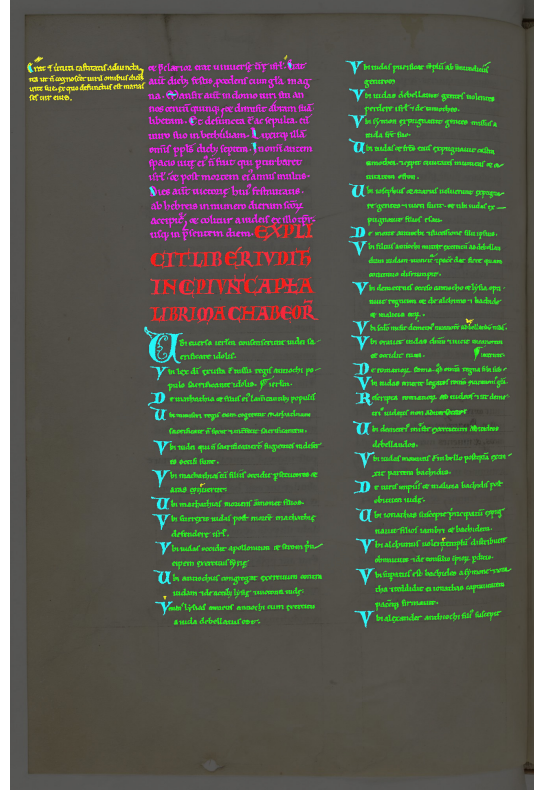
The COCO (Common Objects in Context) dataset [47], introduced in 2014, contains over 200k labeled images covering 80 different object categories. These categories are represented by approximately 1.5 million individual instances distributed across the dataset images, capturing a wide variety of real-world contexts and object interactions. Although COCO is considerably smaller than the previously discussed ImageNet-1k dataset, it offers distinct advantages for tasks beyond image classification. In particular, COCO is explicitly designed for object detection, segmentation, and captioning, providing detailed instance-level annotations that are crucial for these tasks. This



(a) ImageNet-1k



(b) COCO



(c) U-DIADS-Bib

Figure 4.2: Examples from the three datasets used as pre-training data: ImageNet-1k with the class “cup”, COCO, and U-DIADS-Bib.

makes the dataset highly relevant for downstream tasks that involve understanding the structure and layout of images, such as semantic segmentation and document analysis, which are central to the work presented in this study.

U-DIADS-Bib

U-DIADS-Bib [77] is a recently published dataset focusing specifically on the layout segmentation of manuscripts. A detailed description of this dataset is provided in Chapter 3. 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 inter-class 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.

4.1.4 Evaluation datasets

We selected three datasets to evaluate the different initialization strategies for our segmentation model. Two of them are the most widely used datasets for handwritten document layout analysis, DIVA-HisDB [64] and the Bukhari dataset [10], while the third is TobaccoSeg industrial dataset. These datasets have been previously introduced and are described in detail in Chapter 3.

4.2 Experiments

In this section, we present the experimental evaluation of our proposed approach, detailing the training procedures, dataset preparation, and evaluation metrics.

4.2.1 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. 4.1. 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.

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 [24] to account for the substantial class imbalance characterizing all the document datasets, detailed in Chapter 3.

Source dataset	Metric	Trained from scratch	Frozen Encoder					Finetuned Encoder				
			ImageNet1K	COCO	U-DIADS-Bib	ImageNet1K + U-DIADS-Bib	COCO + U-DIADS-Bib	ImageNet1K	COCO	U-DIADS-Bib	ImageNet1K + U-DIADS-Bib	COCO + U-DIADS-Bib
CB55	Prec	0.552	<u>0.502</u>	<u>0.526</u>	<u>0.581</u>	0.573	0.596	0.575	<u>0.619</u>	0.592	0.580	0.626
	Rec	0.652	0.622	0.625	0.717	<u>0.723</u>	0.746	0.730	<u>0.783</u>	0.754	0.765	0.806
	IoU	0.439	0.399	0.410	<u>0.484</u>	0.481	0.508	0.483	<u>0.536</u>	0.500	0.495	0.546
	F1	0.576	0.533	0.542	<u>0.625</u>	0.624	0.651	0.627	<u>0.680</u>	0.645	0.640	0.690
CSG18	Prec	0.571	0.515	0.540	0.582	<u>0.583</u>	0.606	0.578	<u>0.608</u>	0.590	0.588	0.613
	Rec	0.733	0.626	0.639	0.825	0.836	<u>0.830</u>	0.831	0.886	0.853	0.849	<u>0.882</u>
	IoU	0.512	0.425	0.441	0.524	<u>0.527</u>	0.545	0.520	<u>0.560</u>	0.537	0.533	0.564
	F1	0.653	0.556	0.574	0.663	<u>0.667</u>	0.683	0.660	<u>0.698</u>	0.676	0.673	0.702
CSG863	Prec	0.614	0.520	0.560	0.619	0.616	0.652	0.619	<u>0.654</u>	0.627	0.626	0.657
	Rec	0.719	0.631	0.667	0.778	<u>0.791</u>	0.798	0.773	<u>0.851</u>	0.817	0.810	0.862
	IoU	0.504	0.428	0.464	0.539	<u>0.542</u>	0.567	0.533	<u>0.588</u>	0.557	0.551	0.597
	F1	0.647	0.562	0.602	0.680	<u>0.682</u>	0.705	0.675	<u>0.726</u>	0.697	0.692	0.733
Bukhari	Prec	0.534	0.487	0.501	0.546	<u>0.556</u>	0.569	0.565	0.599	<u>0.578</u>	0.571	0.599
	Rec	0.607	0.534	0.586	0.648	<u>0.675</u>	0.705	0.674	<u>0.720</u>	0.700	0.692	0.728
	IoU	0.431	0.385	0.401	0.447	<u>0.458</u>	0.475	0.467	0.504	<u>0.481</u>	0.474	0.504
	F1	0.562	0.505	0.529	0.582	<u>0.597</u>	0.614	0.604	0.644	<u>0.621</u>	0.613	0.644
TobaccoSeg	Prec	0.509	0.477	0.506	<u>0.520</u>	0.509	0.525	0.522	0.543	<u>0.547</u>	0.543	0.587
	Rec	0.573	0.546	0.569	0.614	<u>0.657</u>	0.676	0.749	0.830	0.725	0.699	<u>0.828</u>
	IoU	0.359	0.346	0.348	<u>0.441</u>	0.428	0.446	0.444	0.514	0.479	0.471	0.531
	F1	0.412	0.401	0.406	<u>0.551</u>	0.542	0.552	0.559	<u>0.609</u>	0.603	0.592	0.660

Table 4.2: Results for the different pre-training strategies. 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.

4.2.2 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. Metric definitions are reported in Eq. 4.1– 4.4, where TP, FP and FN stand respectively for True Positives, False positives and False Negatives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.2)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (4.3)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

Each metric was calculated individually for each document class of the 3 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

In this section, we also analyze the impact of different pre-training strategies on the convergence time, stability, and data efficiency of our models. Furthermore, we provide a qualitative assessment of the segmentation results and discuss the implications of the observed performance trends, highlighting how the choice of source datasets influences both quantitative metrics and training dynamics.

4.3.1 Transferability of the learned features

Tab. 4.2 shows the performances, in terms of the 4 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 bold italic the scores achieved by pre-trained models that do not improve over the random initialization baseline, as these cases provide an important indication of the limitations of certain transfer strategies. From this analysis, we can observe that, pre-training on out-of-domain datasets with no further fine-tuning of the encoder module, 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 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. 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 U-DIADS-Bib 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 ImageNet-1K and U-DIADS-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 Fig. 4.3 and Fig. 4.4 the learning curves representing the evolution of the Intersection over Union (IoU) on the validation set throughout the 200 training epochs for pipeline Fig. 4.1(b) and (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

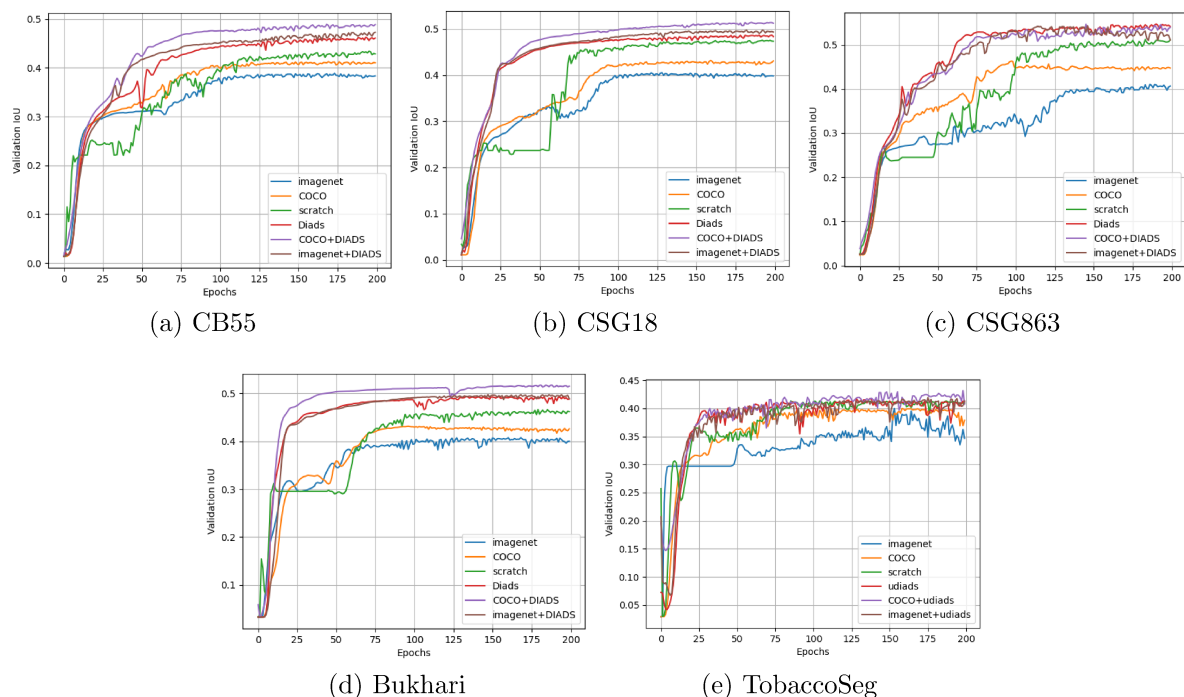


Figure 4.3: Learning curves representing the evolution of the IoU scores for the 5 target document classes using different initialization strategies and keeping the encoder frozen at the time of fine-tuning on the target data.

instability of the training process during the first 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. 4.5 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

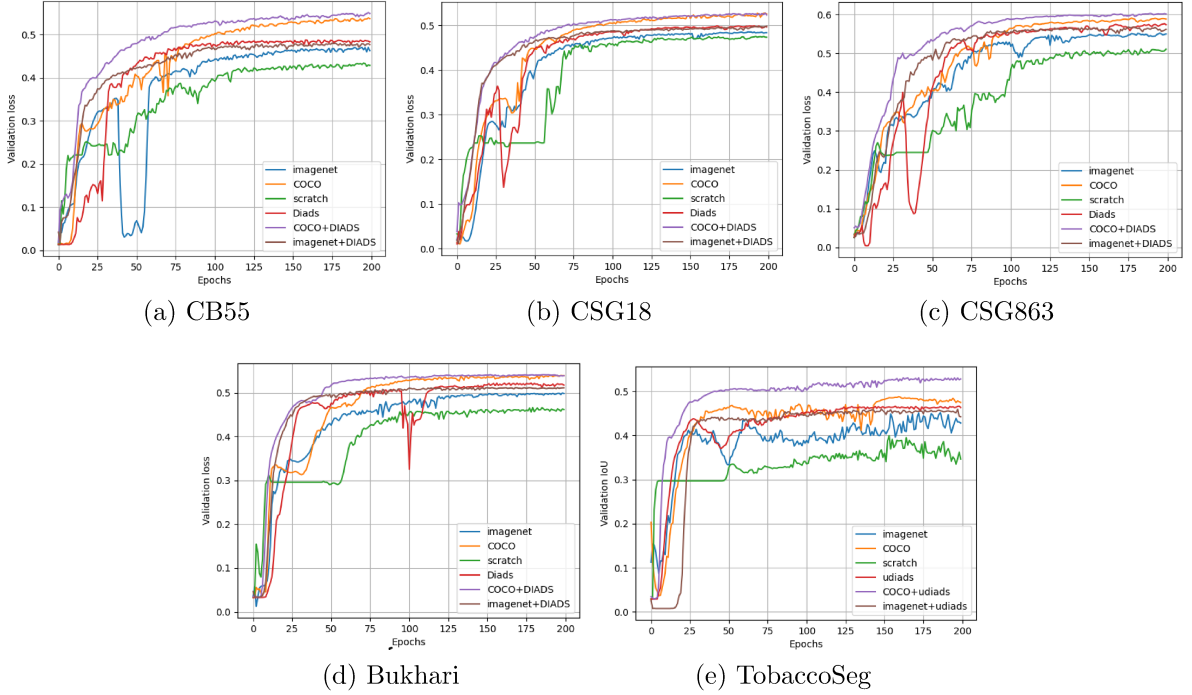


Figure 4.4: Learning curves representing the evolution of the IoU scores for the 5 target document classes using different initialization strategies and fine-tuning the entire model on the target data.

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 4 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 data compared to the randomly initialized model trained on the entire training set for all the document classes except for CSG863.

In Fig. 4.6, 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

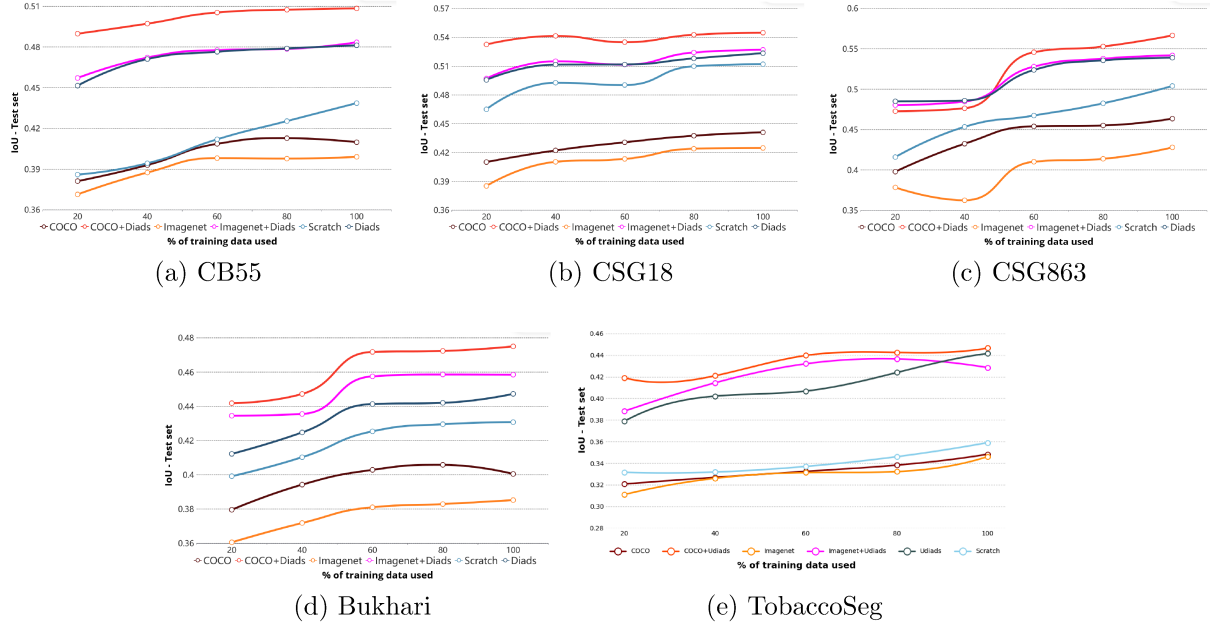


Figure 4.5: Performance (IoU) of the segmentation model on the test set of the 5 document classes when initialized with different strategies while relying on increasing percentages of the available data for the training. Only the decoder and classifier modules of the model were trained on the target data, while the feature-extractor was kept frozen.

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 across all four of the considered document classes.

4.3.4 Qualitative results

For completeness in Fig. 4.7 and Fig. 4.8 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

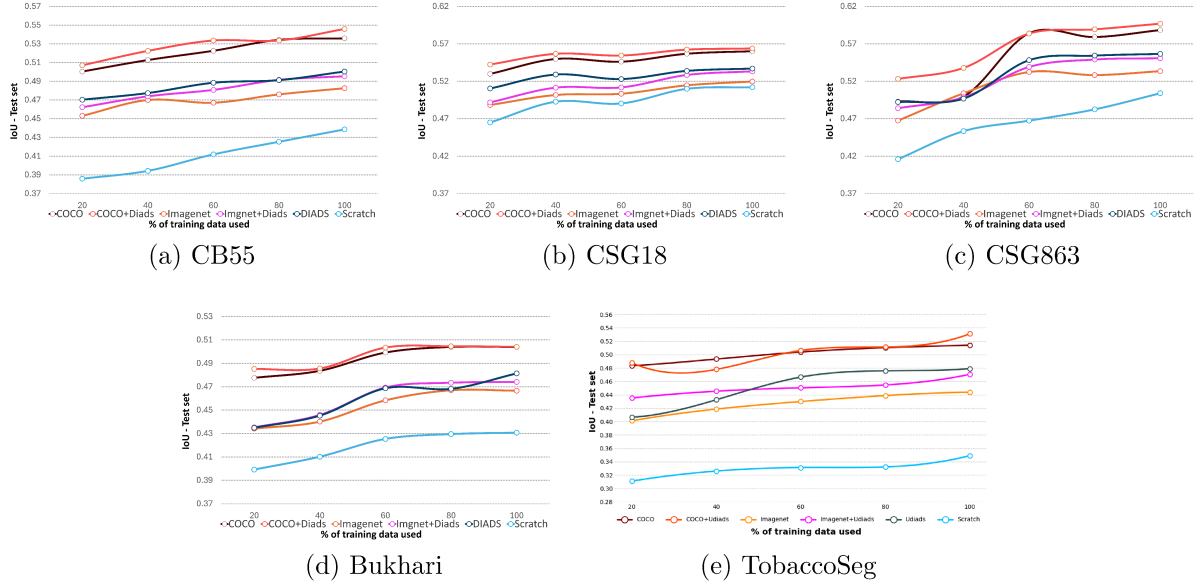


Figure 4.6: Performance (IoU) of the segmentation model on the test set of the 5 document classes when initialized with different strategies while relying on increasing percentages of the available data for the training. The model was fine-tuned entirely on the target data, with no frozen modules.

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.

4.3.5 Discussion

The previously presented results are very eloquent regarding the significant impact on performance that the adoption of different transfer learning strategies, as well as different source datasets for pre-training, has 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 behaviour, 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,

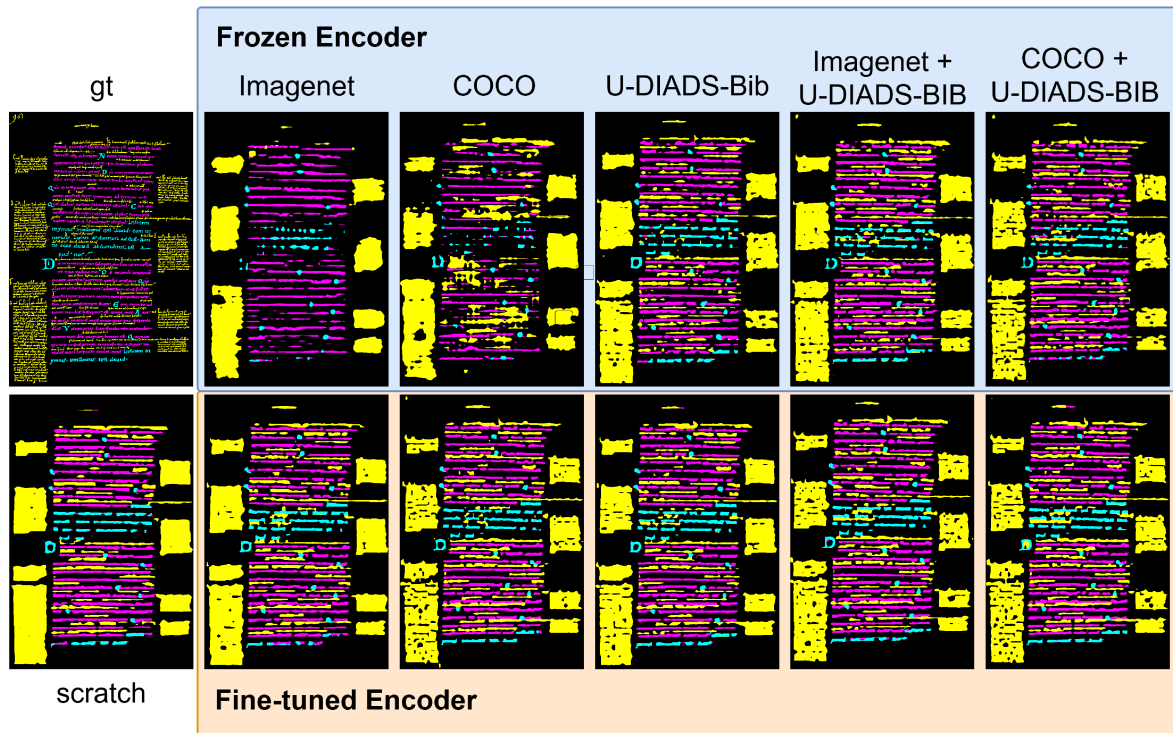


Figure 4.7: Qualitative comparison of the segmentation maps of CSG18 manuscript obtained by employing the different training strategies considered in the analysis.

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 has led to the second-best overall performance across all the considered initialization strategies. The reason why ImageNet is not as effective 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 with document images.

We have further shown how introducing a pre-training step focusing on a domain-

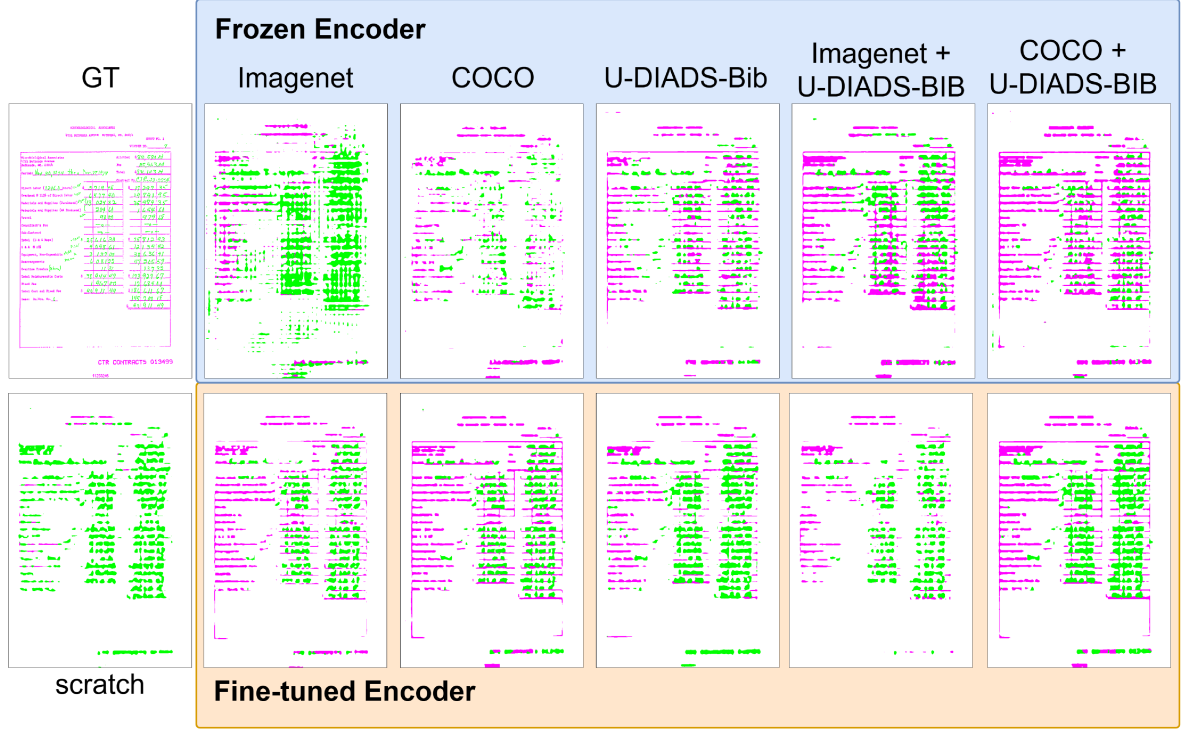


Figure 4.8: Qualitative comparison of the segmentation maps of TobaccoSeg dataset obtained by employing the different training strategies considered in the analysis.

specific dataset, even one as small as the U-DIADS-Bib 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 convergence time and stability of the training process. Moreover, 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

Few-Shot Learning for Document Layout Segmentation

In recent years, machine learning and deep learning-based methods have increasingly been adopted for document layout analysis, as they offer a more effective alternative to traditional approaches. However, these methods rely on large amounts of carefully defined ground-truth maps in order to be properly trained and meaningfully compared with other techniques.

To be suitable for training accurate deep learning models, ground-truth annotations must provide pixel-level precision in the segmentation masks [33]. This requirement represents a significant drawback, since the pixel-accurate annotation of entire historical document datasets is both extremely time-consuming and dependent on domain-specific expertise, particularly when dealing with historical manuscripts, where only specialized humanities scholars can ensure the necessary level of precision [58]. As a consequence, such high-quality ground-truth data are rarely available in real-world scenarios.

Despite this challenge, few-shot learning approaches for document layout segmentation remain relatively underexplored in the current literature. In this chapter, we therefore introduce a novel few-shot learning framework for efficient, pixel-precise page segmentation of documents [23, 25, 24]. The proposed approach is capable of accurately segmenting the different components of a document page (e.g., text, paratext, images), achieving performance comparable to state-of-the-art methods while requiring only a fraction of the training data typically needed. Moreover, a key feature of our system is

its ability to generalize across domains: it has been designed to work robustly not only on complex historical manuscripts but also on contemporary industrial documents, which often present very different structural and visual characteristics. Specifically:

- we design a dynamic instance generation strategy that maximizes the variability of the training data while keeping the annotation requirements to a minimum;
- we adopt a robust segmentation backbone tailored to work effectively in a few-shot regime, ensuring high-quality pixel-level predictions even with limited supervision;
- we introduce a refinement module that leverages adaptive thresholding techniques to further improve the sharpness and accuracy of the segmentation maps;
- we demonstrate that our framework generalizes across domains, providing reliable results both on complex historical manuscripts and on contemporary industrial documents.

The topics addressed in this chapter have been the subject of the following publications: [22, 23, 24, 25].

5.1 Method

The proposed approach is built on three core components, namely a robust segmentation backbone used to retrieve the semantic components of each document page, a dynamic instance generation module that allows us to fully leverage the limited amount of data available at training time, and finally a segmentation refinement module that makes it possible to further improve the quality of the segmentation maps produced by our model. A visual representation of the proposed framework pipeline is reported in Fig. 5.1.

5.1.1 Segmentation backbone

Adoption of a robust backbone is a crucial step in each Deep Learning framework. When working in a few-shot setting in particular we need a network that is able to capture a sufficient level of detail while being given in input just a handful of samples. For this reason, we selected DeepLabV3+ [16] as the backbone of our framework.

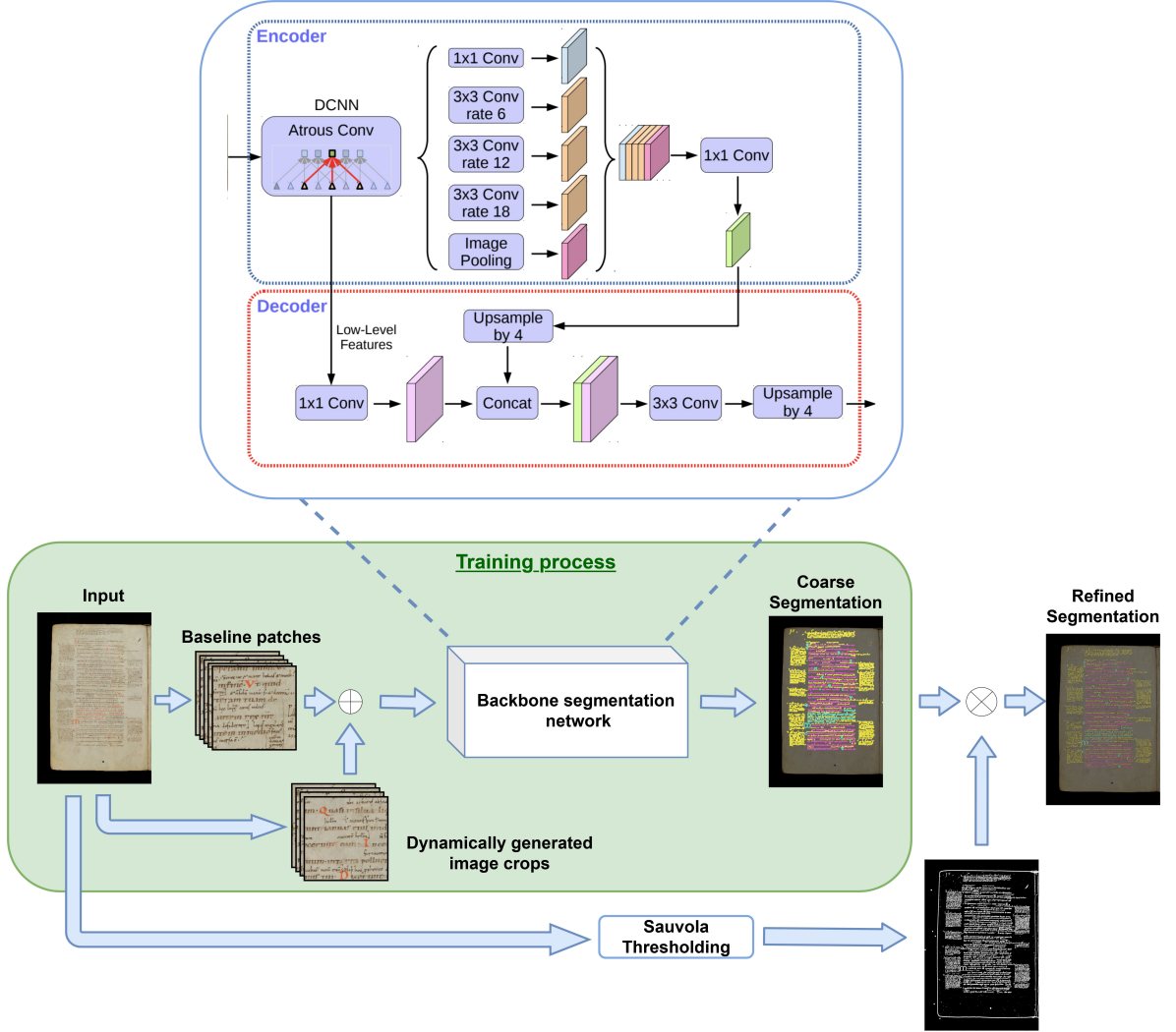


Figure 5.1: Visual representation of the segmentation pipeline for the proposed framework. The green area represents the processes carried out during the training phase, where each input image is split into 2 sets of patches: the baseline patches, which are non-overlapping patches of size $k \times k$ providing a complete representation of the original image and a set of C random crops which are extracted from random locations of the image at each training epoch. These 2 sets of patches are then combined and given in input to the backbone segmentation model which provides a predicted coarse segmentation map for each of them. These maps are compared with the ground truth ones through the application of a weighted cross-entropy loss. At inference time the dynamic instance generation step is removed while a segmentation refinement process is applied to the outputs of the backbone architecture to obtain more precise segmentation maps.

DeepLabV3+ is a popular pixel-wise semantic segmentation model built on its predecessor DeepLabV3 [15]. The latter is a ResNet [35] based architecture heavily relying on atrous convolutions which are employed both in parallel and in a cascade in order to enlarge the receptive fields of the filters and consequently retain a higher spatial resolution throughout the network. The key advantage of this approach is that it allows for deeper neural networks that provide larger feature maps at no additional computational cost. Finally, the Atrous Spatial Pyramid Pooling (ASPP) is introduced in DeepLabV3 as a way of capturing features at different scales in the original image by relying on a heterogeneous set of dilation rates in the network. DeepLabV3+ introduces two substantial changes compared to the aforementioned architecture. The first one regards the substitution of the ResNet encoder with a custom version of the Aligned Xception [17] model in which all max pooling operations are replaced by depth-wise separable convolution. Furthermore, it adds a simple yet effective decoder which refines the segmentation results. The decoder module employs depth-wise separable convolutions to enhance the spatial resolution of the feature maps, resulting in sharper and more detailed output segmentation maps.

5.1.2 Dynamic instance generation

The dynamic instance generation module is a key component of the training pipeline of our framework. The key idea behind it is that it efficiently exploits the small amount of data available at training time [22]. To do so, instead of relying on the full document pages as the instances of our dataset, we split them into two sets of smaller patches. The first ones, which we will refer to as baseline patches, consist of a set of non-overlapping sub-regions of size $m \times n$ extracted from the original input image in order to cover its entire surface and are kept consistent between the training and inference time (Fig. 5.2a). In addition to the baseline patches, as a way to further improve the generalization capabilities of our model, we also generate a small set of k potentially overlapping crops of the same size as the baseline patches which are extracted from random locations of the original image (Fig. 5.2b). This process is carried out at each epoch during training time, while at inference time no additional crops are generated as they are not needed to obtain the final segmentation mask. While relying on sub-patches of the original images is a common approach in computer vision-related tasks, in most cases, these patches are either limited to the ones corresponding to our

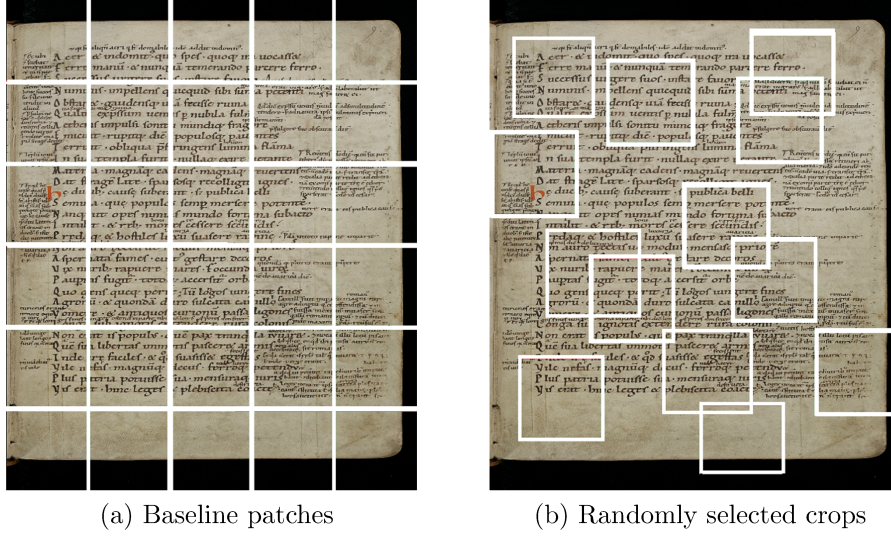


Figure 5.2: Representation of the instance generation process of the 2 sets of patches used to train our model: in 5.2a is shown the generation process for baseline non-overlapping patches, while 5.2b provides a visual depiction of our dynamic crop generation process.

baseline ones, which leads to losing potentially useful information contained in the data. As an alternative approach, they may generate a large number of patches in advance, without considering the varying complexity of different datasets [70]. As a consequence, excessive amounts of potentially unnecessary data is produced.

Our dynamic instance generation approach addresses both limitations effectively at the cost of a very small computational overhead at training time.

5.1.3 Segmentation refinement

Our segmentation refinement module is based on the Sauvola thresholding algorithm for document binarization [63]. The Sauvola thresholding algorithm is an evolution of Niblack’s method [57], which introduced the idea of a dynamic threshold that is calculated based on the mean and standard deviation of the gray levels of a local window inside an image. The main drawback of Niblack’s approach is that it didn’t perform well for images with a light-textured background as it would result in very noisy binarization masks. Sauvola solved this problem by introducing the dynamic range of the standard deviation as an additional term in the equation used to calculate the local threshold, which has the effect of amplifying the contribution of the standard

deviation in an adaptive manner throughout the image. The resulting equation adopted by the Sauvola algorithm is shown in Eq. 5.1, where N is the local window of size $n \times n$, $\mu(N)$ and $\sigma(N)$ are, respectively, the corresponding mean and standard deviation, and R is the dynamic range of the standard deviation. Finally, k is a manually selected parameter that regulates the value of the local threshold.

$$T = \mu(N) \times \left(1 + k \times \left(\frac{\sigma(N)}{R} - 1 \right) \right) \quad (5.1)$$

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

5.2 Experiments

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

5.2.1 Datasets

To train and evaluate our model, we selected several datasets covering a wide range of document types: the popular Diva-HisDB dataset [64], the Bukhari dataset [10], the U-DIADS-Bib FS version, and, to further assess the robustness of our approach, the industrial dataset we constructed, TobaccoSeg. A complete description of all these datasets is provided in Chapter 3. In this work, we conduct experiments in a few-shot setting. For Diva-HisDB, we used only 6 images in total (2 for each class) to train our model. For the Bukhari dataset, we relied on just 3 images, one per manuscript. For U-DIADS-Bib, we employed its FS version, using 3 images per manuscript for training. Finally, for TobaccoSeg, we used 3 images out of the 10 available for training.

By selecting datasets that span from historical manuscripts to industrial documents, with varying layouts and segmentation classes, we ensure that the proposed methods generalize across different contexts and remain applicable in real-world scenarios.

5.2.2 Training and inference setup

Our model was trained using the popular Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5} . The maximum number of epochs for which it was allowed to run has been set to 200 with an early stop in case the validation loss didn't improve in the last 20 epochs and a buffer of 50 epochs which guarantees that the model will be trained at least for the specified amount of iterations. During each epoch, a set of 10 dynamic crops of size 672×672 px has been generated in addition to the baseline patches of the same sizes extracted from the original image. This process led to a maximum of 4012 instances being generated for each document class during training, in case the model needed all the 200 epochs in order to converge. In order to be able to fit them in the GPU memory the images of the dataset have been resized from their original high resolution (up to $4.8k \times 6.8k$ px), down to a size of 1344×2016 px. The loss function selected to train the model is a weighted Cross Entropy Loss [39] in which the weight for each semantic element class is inversely proportional to the frequency of that element in that dataset and, more precisely is calculated as the square root of 1 over the square root of the occurrence frequency of the corresponding element in the dataset (Eq. 5.2).

$$W_i = \sqrt{\frac{1}{F_i}} \quad (5.2)$$

This specific choice was made to take into account the high imbalance between the semantic class distribution in each document category of the datasets descritte nel Capitolo 3. Our model was trained separately and from scratch on each document class.

Regarding the inference setup, the main choice involved in it is represented by the hyperparameters of the segmentation refinement algorithm, namely the window size which was kept consistent at 15×15 px for all document classes and the control value k , which regulates the value of the threshold in the local window (the higher the k value, the lower the threshold) and was set at the value of 0.01 for all classes.

5.2.3 Evaluation metrics

In order to evaluate the performance of our proposed approach we use different metrics such as Precision, Recall, IoU and F1-Score. These evaluation metrics are calculated individually for each one of the manuscripts that compose DIVA-HisDB and U-DIADS-

Bib dataset. Metric definitions are reported in Eq. 5.3–5.6, where TP, FP and FN stand respectively for True Positives, False positives and False Negatives. For each metric a weighted average is performed, based on each class frequency in each manuscript. The final evaluation of a model is then obtained by averaging the metrics of all pages of the three manuscripts.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5.3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5.4)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (5.5)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.6)$$

5.3 Results

In the following section, we provide a thorough comparison between the results achieved by the proposed framework and a set of popular semantic segmentation approaches, namely DeepLabV3 [15], its improved version DeepLabV3+ [16], FCN [49], Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) [36], and Pyramid Scene Parsing Network (PSPNet) [73]. Furthermore, we also include the results obtained by the current state-of-the-art method for the task of document layout segmentation, referred to as MLA [70], although these results are only available for the DIVA-HisDB dataset. The comparison focuses both on a quantitative and a qualitative perspective in order to provide a complete overview of the quality of the model’s predicted segmentations. To this end, we also provide a discussion about the critical cases in which our approach fails to provide the correct segmentation for the corresponding instances. All the models, excluding MLA for which we gathered the results from the respective paper, have been personally tested by us keeping the training and evaluation settings as consistent as possible.

5.3.1 Quantitative results

In Tab. 5.1 the quantitative results achieved by our proposed framework for all the selected metrics across all the document classes contained in the Diva-HisDB dataset

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

Table 5.1: Comparison between the performance of our model and the competition on the 4 selected metrics in DIVA-HisDB dataset. The best and second-best performing models are reported in a bold and underlined fashion respectively while * indicates the models trained in a few-shot setting by using the same set of images selected for our framework.

are shown and compared with the competitor models. In particular, for all the models, excluding MLA, we provide both the results obtained by training them on the entire available dataset and the ones obtained by training the model only on the subset of 2 pages selected for our approach. Unfortunately, MLA authors provided only the mean scores for the selected metrics and some implementation details were missing, leading our attempt at reimplementing their work to achieve sub-optimal results. As we can see our model is consistently capable of outperforming the other semantic segmentation networks on all the metrics, regardless of the setup in which they have been trained. In particular, compared to the second-best performing approach, being represented by DeepLabV3+, our model achieves a mean improvement of 7.7% when the former is trained in a few-shot setting with a peak improvement of 11.8% for the IoU metric. While, when DeepLabV3+ is trained using the full training set, our approach outperforms it by a still substantial mean of 3.5% (5.2% for the IoU metric) while using only a fraction of the available data.

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

Source dataset	Metric	Backbone										
		FCN*	FCN	LRSAPP*	LRSAPP	PSPNET*	PSPNET	DLV3*	DLV3	DLV3+*	DLV3+	Ours
Latin 2	Prec	0.873	0.955	0.907	0.953	0.928	0.932	0.939	0.954	0.958	<u>0.962</u>	0.973
	Rec	0.930	0.926	0.927	0.898	0.868	0.890	0.897	0.907	0.932	<u>0.936</u>	0.972
	IoU	0.867	0.892	0.809	0.862	0.887	0.851	0.863	0.870	0.902	<u>0.906</u>	0.950
	F1	0.899	0.935	0.871	0.916	0.822	0.904	0.914	0.923	0.941	<u>0.945</u>	0.970
Latin 14396	Prec	0.845	0.953	0.852	0.939	0.925	0.951	0.945	0.951	0.958	<u>0.961</u>	0.971
	Rec	0.908	0.936	0.888	0.906	0.912	0.917	0.908	0.918	0.941	<u>0.944</u>	0.973
	IoU	0.833	0.898	0.866	0.867	0.915	0.873	0.863	0.875	0.905	<u>0.910</u>	0.949
	F1	0.874	0.940	0.808	0.919	0.860	0.927	0.920	0.928	0.946	<u>0.949</u>	0.970
Latin 16746	Prec	0.839	0.948	0.897	0.945	0.875	0.926	0.934	0.947	0.941	0.958	<u>0.952</u>
	Rec	0.880	0.937	0.869	0.893	0.858	0.899	0.887	0.905	0.911	<u>0.939</u>	0.959
	IoU	0.810	0.897	0.867	0.842	0.857	0.845	0.839	0.856	0.863	<u>0.900</u>	0.921
	F1	0.858	0.939	0.796	0.908	0.789	0.906	0.904	0.917	0.920	<u>0.944</u>	0.951
Syriaque 341	Prec	0.755	0.942	0.861	0.912	0.895	0.916	0.890	0.927	0.937	<u>0.943</u>	0.946
	Rec	0.862	0.928	0.883	0.878	0.862	0.899	0.864	0.905	0.926	<u>0.931</u>	0.949
	IoU	0.745	0.879	0.802	0.808	0.787	0.837	0.789	0.846	0.873	<u>0.883</u>	0.907
	F1	0.804	0.932	0.872	0.885	0.871	0.904	0.872	0.912	0.928	<u>0.934</u>	0.946
Mean U-DIADS-Bib	Prec	0.828	0.950	0.879	0.937	0.906	0.931	0.927	0.945	0.949	<u>0.956</u>	0.961
	Rec	0.895	0.932	0.917	0.894	0.875	0.901	0.889	0.909	0.928	<u>0.938</u>	0.963
	IoU	0.829	0.892	0.879	0.845	0.883	0.852	0.859	0.862	<u>0.900</u>	<u>0.900</u>	0.942
	F1	0.844	0.937	0.819	0.907	0.815	0.910	0.882	0.920	0.920	<u>0.943</u>	0.950
Bukhari	Prec	0.836	0.865	0.806	0.899	0.843	0.911	0.879	0.908	0.929	<u>0.956</u>	0.970
	Rec	0.875	0.899	0.858	0.876	0.859	0.861	0.815	0.871	0.907	<u>0.943</u>	0.966
	IoU	0.788	0.824	0.742	0.806	0.770	0.790	0.735	0.802	0.850	<u>0.902</u>	0.940
	F1	0.853	0.879	0.805	0.884	0.846	0.875	0.836	0.883	0.914	<u>0.946</u>	0.967
TobaccoSeg	Prec	0.841	0.955	0.880	0.941	0.933	0.942	0.940	0.945	0.951	<u>0.956</u>	0.963
	Rec	0.917	0.933	0.915	0.908	0.901	0.899	0.897	0.909	0.935	<u>0.955</u>	0.964
	IoU	0.841	0.897	0.848	0.868	0.857	0.859	0.858	0.868	0.903	<u>0.905</u>	0.936
	F1	0.877	0.940	0.889	0.917	0.913	0.915	0.914	0.922	0.941	<u>0.945</u>	0.963

Table 5.2: Comparison between the performance of our model and the competition on the U-DIADS-BibFS, Bukhari and TobaccoSeg datasets. The best and second-best performing models are reported in a bold and underlined fashion respectively while * indicates the models trained in a few-shot setting by using the same set of images selected for our framework.

available images in the training set, resulting in a reduction of the needed data by a factor approximately 45.

In Tab. 5.2, we present a comprehensive comparison between our model and competing approaches on the U-DIADS-BibFS, Bukhari, and TobaccoSeg datasets. As shown, our framework consistently achieves the best performance across all settings, even when compared to models trained on the full training set. The only exception is observed in the precision metric for manuscript Latin 16746 in U-DIADS-BibFS, where our model ranks second, although the difference with the best-performing method remains marginal.

A closer analysis reveals that, when compared to the strongest baseline represented

by DeepLabV3+, our approach achieves a consistent improvement of 2-4% across all metrics in its fully trained configuration. The advantage becomes even more evident in the few-shot scenario, where our framework surpasses DeepLabV3+ by approximately 4-9%, clearly demonstrating its capacity to learn effectively even in the presence of limited annotated data. These results confirm that our framework is not only competitive, but in fact superior to widely adopted semantic segmentation methods, regardless of the training regime considered.

It is also important to emphasize that such improvements are not confined to a specific type of dataset. On the contrary, they are observed across collections of very different nature and complexity: from historical manuscripts with highly variable layouts, degraded conditions, and irregular calligraphy (U-DIADS-BibFS, Bukhari), to industrial documents characterized by fine-grained structural elements and standardized layouts (TobaccoSeg). This breadth of evaluation demonstrates the robustness and adaptability of our method, highlighting its ability to generalize effectively beyond the distribution of the training data and to operate reliably across heterogeneous real-world scenarios.

5.3.2 Qualitative results

Fig. 5.3 shows the segmentation maps produced by our model for four document pages belonging, respectively, to the Diva-HisDB, U-DIADS-Bib, Bukhari, and TobaccoSeg datasets. Our approach is compared with the predictions of the FCN and DeepLabV3+ models, both trained on the entire available training set. For reference, the corresponding ground-truth segmentation maps are also reported.

The maps produced by FCN often display segmentation errors, with frequent class confusions in the U-DIADS-Bib and Bukhari datasets in particular. Moreover, across all datasets, FCN outputs are affected by a significant amount of noise and tend to be coarse, especially in regions containing small text or where different components are densely intertwined. DeepLabV3+ achieves a higher level of detail, particularly in the representation of text regions. Finally, our model yields segmentation maps that are consistently sharper and more precise than those of the competing approaches, closely matching the ground truth across all the analyzed datasets.

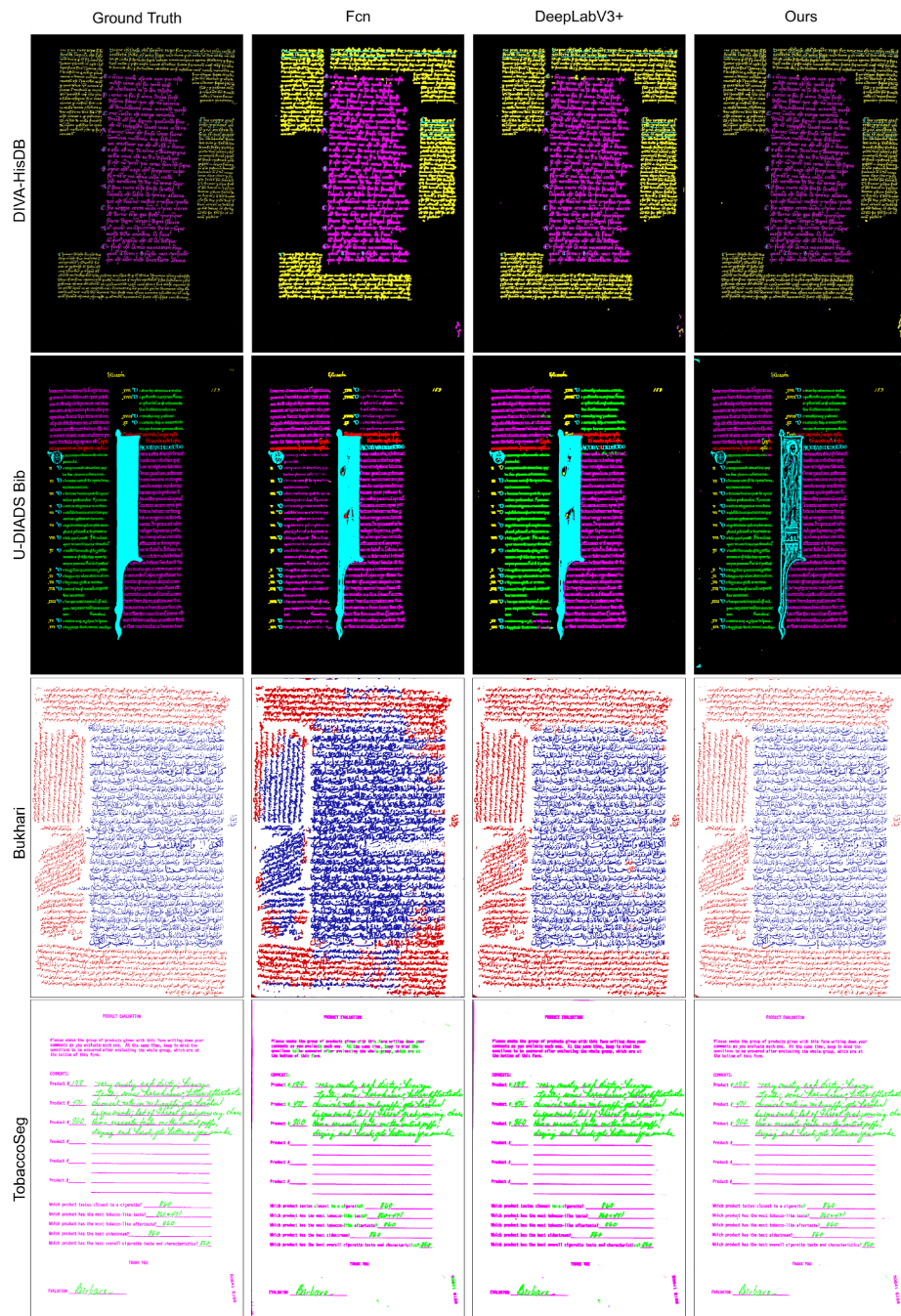
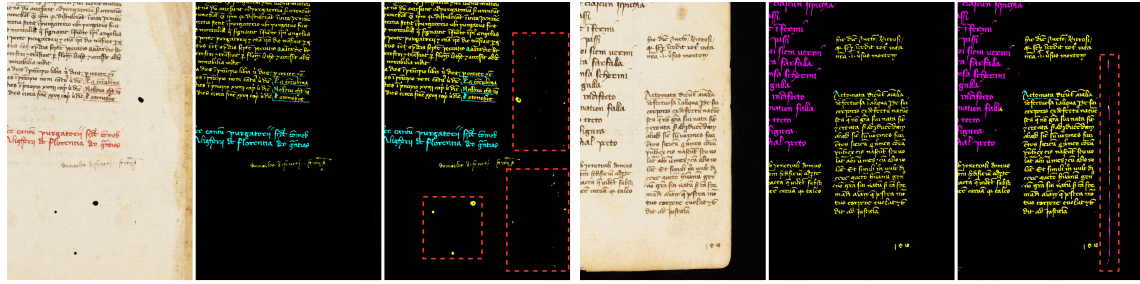
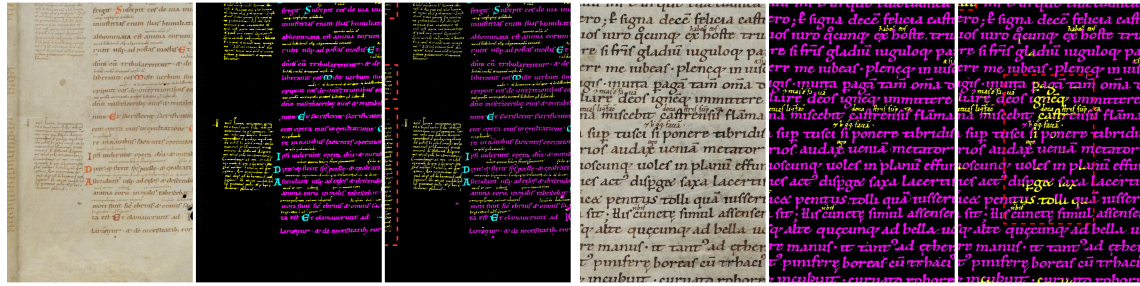


Figure 5.3: Image showing a qualitative comparison between our framework and competing approaches. Each row corresponds to a different sample from the various datasets. The first column shows the ground truth segmentation maps for the four selected images, while the remaining columns display the results produced by FCN, DeepLabV3+, and our method, respectively.



(a) Degraded page

(b) Edge misclassification



(c) Adjacent page text

(d) Foreground misclassification

Figure 5.4: Overview of the main instances of misclassification for the proposed approach. From the top left corner we have: 5.4a degraded spots in the page background being misclassified as foreground, 5.4b the same type of misclassification involving page edges, 5.4c Text belonging to the adjacent page being recognized as part of the current one, 5.4d a simple case of misclassified foreground elements in particular involving the main text being mistaken as part of the comments.

Fail cases

As already mentioned in the previous section the main drawback of the presented approach is that compared to the competition it introduces more noise in the provided segmentations. For completeness, in Fig. 5.4 we provide some more criticalities of the proposed framework together with the original image and the corresponding ground truth. In particular, other than the typical misclassification of foreground elements (Fig. 5.4d) we can notice three main instances of recurrent mistakes. The first one is represented by the edge of the pages of the documents which, being lighter than the black background introduces an area of high contrast that is identified both by the model and by the thresholding algorithm as part of the text (Fig. 5.4b). A similar occurrence can be observed for degraded areas in the page's background, these areas are, in fact, typically darker than the rest of the background and are once again

Backbone	CB55				CSG18				CSG863				Mean			
	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
FCN	0,871	0,850	0,728	0,820	0,887	0,876	0,773	0,847	0,884	0,874	0,770	0,851	0,881	0,867	0,757	0,839
LRASPP	0,801	0,770	0,614	0,718	0,815	0,835	0,718	0,797	0,919	0,908	0,858	0,912	0,845	0,838	0,730	0,809
PSPNET	0,849	0,828	0,694	0,792	0,877	0,869	0,761	0,838	0,901	0,887	0,801	0,876	0,876	0,861	0,752	0,835
DeepLabV3	0,873	0,853	0,734	0,824	0,891	0,881	0,781	0,854	0,882	0,869	0,762	0,845	0,882	0,868	0,759	0,841
DeepLabV3+	0.918	0.908	0.827	0.894	0.926	0.923	0.855	0.910	0.931	0.927	0.863	0.917	0.925	0.919	0.848	0.907

Table 5.3: Comparison between the use of different neural network architectures as the segmentation backbone for our model, in bold is reported the best-performing model.

misclassified as foreground elements (Fig. 5.4a). Finally, we have the misclassification caused by the text belonging to the page adjacent to the currently analyzed instance, which while correctly identified as part of the text by our model, is not included in the ground truth segmentations (Fig. 5.4c). This last case, however, is highly dependent on the coarse cropping process of the instances of the different datasets which don't precisely include only the elements of the current page and, as such, is easily solvable by refining the crops.

5.3.3 Ablation study

In this section, we provide the details regarding the ablation study we conducted in order to obtain the final version of the proposed framework. In particular, we show the effects that different segmentation backbones and patch sizes for the generated instances have on the performance of our approach for the task at hand. Furthermore, we provide a comparison between the performance of the baseline model and the models enhanced with the additional modules introduced in this work in order to provide proof of their effectiveness.

Backbones

Tab. 5.3 shows a comparison of the performance of our framework when using different backbones for the segmentation module. For this comparison, we selected a set of recent and popular semantic segmentation networks (DeepLabV3 [15], DeepLabV3+ [16], FCN [49], LRASPP [36] and PSPNet [73]). To allow for a fair comparison, all the models have been trained and tested with the exact same setup, with 2 images for each document class as the training set and a consistent patch size of 672×672 px. As we can see, all the models provide reasonably good performance on the task at hand, achieving an IoU higher than 70% and a performance of over 80% for all the remaining

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

Table 5.4: Comparison between the adoption of different patch sizes during the instance generation process of our framework, in bold is reported the best-performing model.

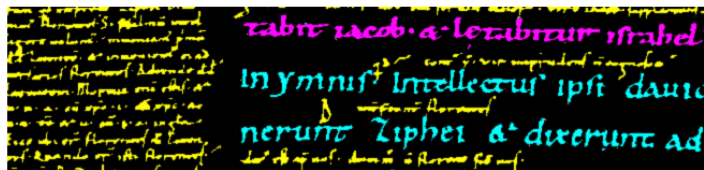
Our Modules	CB55				CSG18				CSG863				Mean			
	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1	Prec	Rec	IoU	F1
<i>Ours (baseline)</i>	0.907	0.900	0.815	0.884	0.926	0.923	0.860	0.912	0.917	0.914	0.840	0.900	0.917	0.912	0.838	0.899
<i>Ours (w/ dynamic crop gen.)</i>	0.918	0.908	0.827	0.894	0.926	0.923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

Table 5.5: Results of the ablation study. Each row shows the performance of the different versions of our system across all the selected metrics for the 4 classes of manuscripts composing the DIVA-HisDB dataset. The last four columns show the mean scores achieved by the models across the different classe.

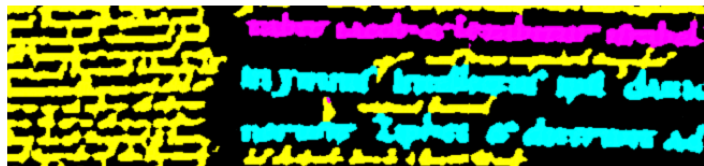
metrics. From this analysis, it emerges that DeepLabV3+ consistently outperforms all other models on each of the selected metrics and on all the document classes present in the dataset, achieving a mean improvement of 6.23% over the second-best model, represented by its predecessor DeepLabV3. A particularly interesting boost in performance is achieved for the IoU metrics, where an increase of almost 9% is obtained by the former over the latter.

Patch sizes

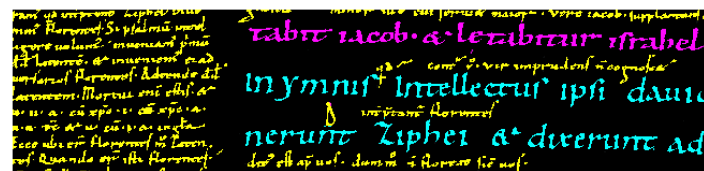
A further comparison has been performed by exploring the adoption of different sizes for the crops of the instances being provided to the backbone networks. In particular, we selected 3 different sizes, going from the standard 224×224 which is the size used by all the pre-trained models available in PyTorch, to a much larger 672×672 . The results of this comparison are shown in Tab. 5.4. In this case, the difference in performance wasn't as substantial as the one resulting from the adoption of different types of segmentation backbones. In particular, we can notice that the difference between the best and the worst performing models, which are the ones adopting the largest and smallest patch sizes respectively, is on mean around 1%. A potential explanation behind the improved performance corresponding to the adoption of larger patch sizes is that the model to which they are given in input is able to capture a higher amount of contextual information regarding the layout of the original image from which they are extracted,



(a) Ground Truth



(b) Coarse Prediction



(c) Refined Prediction

Figure 5.5: Qualitative results showing the effects of the segmentation refinement process. Fig. 5.5a shows the original ground truth for a zoomed area of the original image. Fig. 5.5b shows the coarse segmentation mask obtained by the model. Finally, Fig. 5.5c shows the segmentation prediction resulting from the refinement process.

allowing for more accurate segmentation.

Framework modules

Finally, we provide a comparison between different versions of our framework in which we systematically introduce the original modules presented in this work, namely the dynamic instance generation and the segmentation refinement ones. In particular, in Tab. 5.5 we show the performance obtained by our baseline model, in which the images have been split into patches but without the addition of either the dynamically generated crops or the segmentation refinement process, as well as the one achieved by introducing these 2 techniques singularly and in a combined fashion, which represents our full framework pipeline. As we can see, each of the additional modules leads to a substantial improvement in performance over the baseline approach with the best performance being achieved with the use of both modules. More specifically, the final framework achieves an improvement in performance going from 6.8% for the precision

metric to a very substantial 13.3% for the Intersection over Union one, with a mean improvement of 9% across all metrics when compared to the baseline approach.

As additional proof of the effectiveness of the proposed approach. In Fig. 5.5 we provide a qualitative comparison between the segmentation masks provided by the baseline and the final framework, while also showing the corresponding ground truth as a reference.

5.3.4 Discussion

In this chapter, we proposed an effective framework to address the underexplored problem of few-shot document layout analysis. Our approach introduces two original modules, the dynamic instance generation and the segmentation refinement, which enhance the capabilities of a core image segmentation backbone, enabling it to fully exploit the limited training data available and to achieve pixel-precise page segmentations.

When compared to widely adopted image segmentation algorithms, our model consistently outperforms them while requiring only a fraction of the training data and maintaining a computational cost comparable to that of the backbone network (DeepLabV3+). Moreover, in comparison with the current state-of-the-art framework, our method achieves comparable performance across all standard evaluation metrics on the DIVA-HisDB dataset. Extending the evaluation to three additional datasets (U-DIADS-Bib, Bukhari, and TobaccoSeg) further demonstrates the robustness and versatility of our approach, showing that it generalizes effectively across domains and delivers reliable results both on complex historical manuscripts and on contemporary industrial documents.

Although the results are very promising, some limitations remain. In future work, we plan to investigate more advanced refinement strategies to further enhance the accuracy and sharpness of the segmentation maps.

6

Conclusions and Future Work

This thesis has addressed the problem of document image segmentation under limited data conditions, with a particular focus on industrial and historical documents. The central motivation behind this work has been the dual challenge posed by the scarcity of annotated data and the high degree of domain variability that characterizes real-world scenarios. In many practical contexts, producing pixel-precise ground truth is not only costly but often infeasible, especially for historical documents, where the complexity of layouts, degradations, and diverse scripts further complicates annotation efforts. Similarly, in industrial settings, document formats can vary significantly between companies, departments, or even production batches, making large-scale data collection and standardization impractical. To tackle these challenges, this research proposed novel methodological contributions that go beyond the direct application of existing approaches. Alongside the use of transfer learning and few-shot principles, the thesis introduced new architectural solutions and adaptation mechanisms specifically designed for low-resource document segmentation. These include the design of a few-shot learning framework tailored to document structures, customized pre-training and fine-tuning pipelines, and architectural refinements allowing models to generalize across heterogeneous document domains. These contributions were further supported by the creation of new datasets and extensive experimental validation.

The research unfolded through several key contributions. First, a systematic study of the challenges associated with low-resource document layout analysis was carried out, highlighting how annotation scarcity and structural variability directly affect model generalization. Second, two novel datasets were introduced, one for industrial docu-

ments and one for historical manuscripts, both designed to support reproducible research in this field. Third, the thesis provided a comparative analysis of transfer learning strategies together with an original domain-specific pre-training pipeline, demonstrating that even limited targeted pre-training may outperform large-scale generic alternatives. Finally, a novel few-shot segmentation framework was proposed, combining architectural modifications and prototype-based adaptation strategies. This framework showed that accurate segmentation can be achieved with only a handful of annotated samples and demonstrated strong cross-domain generalization between industrial and historical collections.

Taken together, these results indicate that the combination of knowledge transfer and few-shot generalization constitutes a promising direction for developing document layout analysis systems that are both effective and practical. The work confirms that carefully designed training strategies can compensate for the absence of large annotated corpora and that models trained in this way can adapt to diverse document domains, bridging the gap between theoretical research and real-world applicability.

Despite these advances, several open challenges remain and define potential avenues for future research. One important direction concerns domain adaptation and continuous generalization: although the proposed approaches exhibit transferability across contexts, extreme domain shifts remain difficult to handle. Methods based on meta-learning, unsupervised adaptation, or continual learning could provide more robust solutions when encountering entirely new types of documents. Another promising perspective lies in the integration of multimodal approaches that combine visual and textual information. Vision-language models, in particular, could enhance segmentation with semantic understanding, which is especially relevant for degraded or complex historical materials.

The creation of training data remains a fundamental bottleneck. Advances in synthetic data generation, annotation-efficient pipelines, and human-in-the-loop systems could substantially reduce the costs associated with building ground truth and accelerate the release of new benchmarks. Furthermore, the practical deployment of segmentation systems will require more efficient models, optimized for resource-constrained environments typical of both industrial applications and large-scale digital humanities projects. Model compression and lightweight architectures therefore represent an essential line of future investigation.

Finally, closer interdisciplinary collaborations will be crucial, especially for histor-

ical documents, where the involvement of domain experts not only ensures accurate annotations but also helps to define segmentation criteria that are meaningful from a scholarly perspective.

In conclusion, this thesis has shown that effective document segmentation in low-data regimes is achievable through transfer learning and few-shot strategies. At the same time, it has highlighted the importance of continuing research aimed at improving robustness, scalability, and semantic awareness. By advancing along these directions, future systems will be able to support both industrial efficiency and cultural heritage preservation, ultimately making document layout analysis a mature and versatile technology for real-world deployment.

Bibliography

- [1] Reem Alaasam, Berat Kurar, and Jihad El-Sana. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 738–742, Sydney, Australia, 2019.
- [2] Michele Alberti, Lars Vöglin, Vinaychandran Pondenkandath, Mathias Seuret, Rolf Ingold, and Marcus Liwicki. Labeling, cutting, grouping: An efficient text line segmentation method for medieval manuscripts. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1200–1206, 2019.
- [3] Davide Antonutti, Axel de Nardin, Silvia Zottin, Claudio Piciarelli, and Gian Luca Foresti. Multi-modal analysis of bi-parametric mri slices for lesion detection in prostate cancer screening. *IEEE Access*, 13:115735–115747, 2025.
- [4] Abedelkadir Asi, Rafi Cohen, Klara Kedem, Jihad El-Sana, and Itshak Dinstein. A coarse-to-fine approach for layout analysis of ancient manuscripts. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 140–145, Crete, Greece, 2014.
- [5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.
- [6] P. Barlas, S. Adam, C. Chatelain, and T. Paquet. A typed and handwritten text block segmentation system for heterogeneous and complex documents. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 46–50, Tours, France, 2014.
- [7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [8] Galal M. Binmakhashen and Sabri A. Mahmoud. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6), oct 2019.
 - [9] Galal M. BinMakhashen and Sabri A. Mahmoud. Historical document layout analysis using anisotropic diffusion and geometric features. *International Journal on Digital Libraries*, 21(3):329–342, Sep 2020.
 - [10] Syed Saqib Bukhari, Thomas M. Breuel, Abedelkadir Asi, and Jihad El-Sana. Layout analysis for arabic historical document images using machine learning. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 639–644, 2012.
 - [11] Kai Chen, Cheng-Lin Liu, Mathias Seuret, Marcus Liwicki, Jean Hennebert, and Rolf Ingold. Page segmentation for historical document images based on superpixel classification with unsupervised feature learning. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 299–304, Santorini, Greece, 2016.
 - [12] Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert, and Rolf Ingold. Page segmentation of historical document images with convolutional autoencoders. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1011–1015, Nancy, France, 2015.
 - [13] Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert, Cheng-Lin Liu, and Rolf Ingold. Page segmentation for historical handwritten document images using conditional random fields. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 90–95, Shenzhen, China, 2016.
 - [14] L. Chen, G. Papapandreou, F. Schroff, and A. Hartwig. Rethinking atrous convolution for semantic image segmentation. In *Conference on computer vision and pattern recognition (CVPR)*. *IEEE/CVF*, volume 6, 2017.
 - [15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

- [16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.
- [17] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.
- [18] Rafi Cohen, Abedelkadir Asi, Klara Kedem, Jihad El-Sana, and Itshak Dinstein. Robust text and drawing segmentation algorithm for historical documents. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, HIP '13*, page 110–117, New York, NY, USA, 2013. Association for Computing Machinery.
- [19] Rafi Cohen, Itshak Dinstein, Jihad El-Sana, and Klara Kedem. Using scale-space anisotropic smoothing for text line extraction in historical documents. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, pages 349–358, Cham, 2014. Springer International Publishing.
- [20] Homa Davoudi, Marco Fiorucci, and Arianna Traviglia. Ancient document layout analysis: Autoencoders meet sparse coding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5936–5942, Milan, Italy, 2021.
- [21] Axel De Nardin, Silvia Zottin, Emanuela Colombi, Claudio Piciarelli, and Gian Luca Foresti. Is imagenet always the best option? an overview on transfer learning strategies for document layout analysis. In Gian Luca Foresti, Andrea Fusiello, and Edwin Hancock, editors, *Image Analysis and Processing - ICIAP 2023 Workshops*, pages 489–499, Cham, 2024. Springer Nature Switzerland.
- [22] Axel De Nardin, Silvia Zottin, Matteo Paier, Gian Luca Foresti, Emanuela Colombi, and Claudio Piciarelli. Dynamic instance generation for few-shot handwritten document layout segmentation (short paper). In *CEUR Workshop Proceedings*, volume 3286, page 26 – 34, 2022.
- [23] Axel De Nardin, Silvia Zottin, Matteo Paier, Gian Luca Foresti, Emanuela Colombi, and Claudio Piciarelli. Efficient few-shot learning for pixel-precise hand-

- written document layout analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3680–3688, Waikoloa, Hawaii, January 2023.
- [24] Axel De Nardin, Silvia Zottin, Claudio Piciarelli, Emanuela Colombi, and Gian Luca Foresti. Few-shot pixel-precise document layout segmentation via dynamic instance generation and local thresholding. *International Journal of Neural Systems*, 33(10):2350052, 2023.
- [25] Axel De Nardin, Silvia Zottin, Claudio Piciarelli, Emanuela Colombi, and Gian Luca Foresti. 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)*, pages 8112–8121, 2024.
- [26] Axel De Nardin, Silvia Zottin, Claudio Piciarelli, and Gian Luca Foresti. Deep learning-based intrusion detection systems for phishing email detection: A short survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 7567–7575, October 2025.
- [27] Axel De Nardin, Silvia Zottin, Claudio Piciarelli, Gian Luca Foresti, and Emanuela Colombi. In-domain versus out-of-domain transfer learning for document layout analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, 28(2):161–175, Jun 2025.
- [28] Axel De Nardin, Silvia Zottin, Andrea Toma, Claudio Piciarelli, and Gian Luca Foresti. A multibranch and multiattention framework for floor plan segmentation. *Computer-Aided Civil and Infrastructure Engineering*, 2025.
- [29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [30] Markus Diem, Florian Kleber, Stefan Fiel, Tobias Grüning, and Basilis Gatos. cbad: Icdar2017 competition on baseline detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1355–1360, Kyoto, Japan, 2017.

- [31] Ahmad Droby, Berat Kurar Barakat, Borak Madi, Reem Alaasam, and Jihad El-Sana. Unsupervised deep learning for handwritten page segmentation. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 240–245, Dortmund, Germany, 2020.
- [32] Andreas Fischer, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. Automatic transcription of handwritten medieval documents. In *2009 15th International Conference on Virtual Systems and Multimedia*, pages 137–142, 2009.
- [33] Angelika Garz, Mathias Seuret, Fotini Simistira, Andreas Fischer, and Rolf Ingold. Creating ground truth for historical manuscripts with document graphs and scribbling interaction. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 126–131, Santorini, Greece, 2016.
- [34] Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, Muhammad Adnan Afzal, Muhammad Ahtsham Afzal, and Muhammad Zeshan Afzal. Current status and performance analysis of table recognition in document images with deep neural networks. *IEEE Access*, 9:87663–87685, 2021.
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, Nevada, 2016.
- [36] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, Seoul, South Korea, 2019.
- [37] Juana Valeria Hurtado and Abhinav Valada. Chapter 12 - semantic scene segmentation for robotics. In Alexandros Iosifidis and Anastasios Tefas, editors, *Deep Learning for Robot Perception and Cognition*, pages 279–311. Academic Press, 2022.
- [38] Pranjal Jadhav, Mayuree Sawal, Anushka Zagade, Prerna Kamble, and Prajakta Deshpande. Pix2pix generative adversarial network with resnet for document

- image denoising. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1489–1494, 2022.
- [39] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020.
- [40] Junho Jo, Hyung Il Koo, Jae Woong Soh, and Nam Ik Cho. Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications*, 79(43):32137–32150, Nov 2020.
- [41] Nicholas Journet, Jean-Yves Ramel, Rémy Mullot, and Véronique Eglin. Document image characterization using a multiresolution analysis of the texture: application to old documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 11(1):9–18, Oct 2008.
- [42] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), September 2022.
- [43] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [44] Florian Kleber, Robert Sablatnig, Melanie Gau, and Heinz Miklas. Ancient document analysis based on text line extraction. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.
- [45] Jayant Kumar and David Doermann. Unsupervised classification of structurally similar document images. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1225–1229, 2013.
- [46] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 665–666, New York, NY, USA, 2006. Association for Computing Machinery.

- [47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [48] Francesco Lombardi and Simone Marinai. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10):110, Oct 2020.
- [49] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
- [50] Maroua Mehri, Pierre Héroux, Petra Gomez-Krämer, and Rémy Mullot. Texture feature benchmarking and evaluation for historical document image analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, 20(1):1–35, Mar 2017.
- [51] Maroua Mehri, Nibal Nayef, Pierre Héroux, Petra Gomez-Krämer, and Rémy Mullot. Learning texture features for enhancement and segmentation of historical document images. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, HIP '15*, page 47–54, New York, NY, USA, 2015. Association for Computing Machinery.
- [52] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.
- [53] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.
- [54] Lesia Mochurad. Approach for enhancing the accuracy of semantic segmentation of chest x-ray images by edge detection and deep learning integration. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025.

- [55] Sonika Rani Narang, M. K. Jindal, and Munish Kumar. Ancient text recognition: a review. *Artificial Intelligence Review*, 53(8):5517–5558, Dec 2020.
- [56] Karl Ni, Patrick Callier, and Bradley Hatch. Writer identification in noisy handwritten documents. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1177–1186, 2017.
- [57] W Niblack. *An Introduction to Digital Image Processing*. Prentice hall, Englewood Cliffs, 1986.
- [58] Konstantina Nikolaidou, Mathias Seuret, Hamam Mokayed, and Marcus Liwicki. A survey of historical document image datasets. *Int. J. Doc. Anal. Recognit.*, 25(4):305–338, dec 2022.
- [59] Moisés Pastor. Text baseline detection, a single page trained system. *Pattern Recognition*, 94:149–161, 2019.
- [60] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
- [62] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14298–14308, 2022.
- [63] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.
- [64] Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. Diva-hisdb: A precisely annotated large dataset of chal-

- lenging medieval manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476, Shenzhen, China, 2016.
- [65] Linda Studer, Michele Alberti, Vinaychandran Pondekandath, Pinar Goktepe, Thomas Kolonko, Andreas Fischer, Marcus Liwicki, and Rolf Ingold. A comprehensive study of imagenet pre-training for historical document image analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 720–725, Sydney, Australia, 2019.
- [66] Solène Tarride, Aurélie Lemaitre, Bertrand Coüasnon, and Sophie Tardivel. Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(1):77–96, Jun 2021.
- [67] K Thamarai Selvan, N. Shanmugapriya, K Senthil Kumar, and N Somasundaram. Automated segmentation of palmprint images for biometrics using computer vision. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, pages 1–6, 2025.
- [68] Tuan Anh Tran, In Seop Na, and Soo Hyung Kim. Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology. *International Journal on Document Analysis and Recognition (IJDAR)*, 19(3):191–209, Sep 2016.
- [69] Hai Wang, Yanyan Chen, Yingfeng Cai, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. Sfnet-n: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21405–21417, 2022.
- [70] Yue Xu, Fei Yin, Zhaoxiang Zhang, and Cheng-Lin Liu. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI’18*, page 1057–1063, Stockholm, Sweden, 2018. AAAI Press.
- [71] Hongshan Yu, Zhengeng Yang, Lei Tan, Yaonan Wang, Wei Sun, Mingui Sun, and Yandong Tang. Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304:82–103, 2018.

- [72] Ying Yu, Chunping Wang, Qiang Fu, Renke Kou, Fuyu Huang, Boxiong Yang, Tingting Yang, and Mingliang Gao. Techniques and challenges of image segmentation: A review. *Electronics*, 12(5), 2023.
- [73] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, Honolulu, hawaii, 2017.
- [74] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, 2019.
- [75] Silvia Zottin, Axel De Nardin, Giuseppe Branca, Emanuela Colombi, Claudio Piciarelli, Hafsa Shujat, and Gian Luca Foresti. Exploring few-shot text line segmentation approaches in challenging ancient manuscripts. In *CEUR Workshop Proceedings*, volume 3937, 2025.
- [76] Silvia Zottin, Axel De Nardin, Giuseppe Branca, Claudio Piciarelli, and Gian Luca Foresti. Icdar 2025 competition on few-shot text line segmentation of ancient handwritten documents (fest). In Xu-Cheng Yin, Dimosthenis Karatzas, and Daniel Lopresti, editors, *Document Analysis and Recognition – ICDAR 2025*, pages 586–602, Cham, 2026. Springer Nature Switzerland.
- [77] Silvia Zottin, Axel De Nardin, Emanuela Colombi, Claudio Piciarelli, Filippo Pavan, and Gian Luca Foresti. U-diads-bib: a full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts. *Neural Computing and Applications*, pages 1–13, 2024.
- [78] Silvia Zottin, Axel De Nardin, Gian Luca Foresti, Emanuela Colombi, and Claudio Piciarelli. Icdar 2024 competition on few-shot and many-shot layout segmentation of ancient manuscripts (sam). In Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng, editors, *Document Analysis and Recognition - ICDAR 2024*, pages 315–331, Cham, 2024. Springer Nature Switzerland.