university of koblenz
Computer Science

Faculty 4: Computer Science

ist Institut für Softwaretechnik

Institut für Softwaretechnik

Recognition of Layout Patterns in Historical Legal Texts

Proposal for a Master Thesis

submitted by

Mohammad Obaidullah Tusher

January 27, 2026

Supervisor: Prof. Dr. Jan Jürjens
Institute for Software Technology/Institut für Softwaretechnik

2nd Supervisor: Thomas Reiser
Institute for Software Technology/Institut für Softwaretechnik

Abstract

This thesis addresses the challenge of automated layout analysis for historical German VET and CVET regulations, currently digitized through manual TEI (Text Encoding Initiative) XML workflows at the University of Koblenz under the supervision of Thomas Reiser. Existing rule-based methods prove insufficient for these documents, which exhibit complex hierarchical structures and variable typography spanning 1938–2022. We propose adapting LayoutLMv3 [10], a multimodal transformer architecture that jointly processes document images, OCR text, and spatial layout information. Our approach leverages transfer learning from large-scale document corpora, fine-tuning the model on a limited set of manually annotated historical pages for 18-class semantic segmentation. Drawing methodological guidance from recent work on the Heimatkunde dataset [1] and visual representation techniques from BEIT [2], we target systematic recognition of headings, section markers, and enumerations essential for reconstructing document hierarchy [12]. A complementary unsupervised baseline using OCR-based features provides comparative context [5]. Expected contributions include empirical evaluation of multimodal transformers on limited historical annotations and measurable improvements over preliminary convolutional baselines (mean IoU 0.397). While computational resources and modest annotation volumes constrain the scope, initial experiments suggest that transfer learning from modern document understanding models can meaningfully improve layout analysis for historical legal texts within standard academic computing environments.

Contents

List of Acronyms and Key Terms

Acronyms

BIBB	Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung)
BERT	Bidirectional Encoder Representations from Transformers
BEiT	BERT Pre-Training of Image Transformers
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CVET	Continuing Vocational Education and Training
DiT	Document Image Transformer
DPI	Dots Per Inch
DRFN	Document Recognition Fusion Network
dVAE	discrete Variational Autoencoder
FCN	Fully Convolutional Network
IoU	Intersection over Union
IIT-CDIP	Illinois Institute of Technology Complex Document Information Processing (dataset)
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MCP	Model Context Protocol
MFCN	Multimodal Fully Convolutional Network
MIM	Masked Image Modeling
MLM	Masked Language Modeling
OCR	Optical Character Recognition
PCA	Principal Component Analysis
PDF	Portable Document Format
RGB	Red, Green, Blue (color model)
t-SNE	t-Distributed Stochastic Neighbor Embedding
TEI	Text Encoding Initiative
TF-IDF	Term Frequency-Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
VET	Vocational Education and Training
ViT	Vision Transformer
WP	Work Package
WPA	Word-Patch Alignment
XML	Extensible Markup Language

Key Terms

Antiqua Modern serif typeface (as opposed to Fraktur blackletter).

Blackletter/Fraktur Gothic typeface commonly used in German documents until mid-20th century, characterized by dense, angular strokes that complicate OCR.

Dice Coefficient Similarity metric for segmentation evaluation: $\frac{2|A \cap B|}{|A|+|B|}$, where A is predicted and B is ground truth.

F1-Score Harmonic mean of precision and recall: $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.

Focal Loss Modified cross-entropy loss that down-weights easy examples to focus learning on hard, misclassified cases; designed for class imbalance.

Fine-tuning Training a pre-trained model on a specific downstream task using supervised learning.

long s (*ſ*) Historical German typographic character resembling 'f', frequently confused by OCR systems.

Macro-averaging Computing metrics independently for each class, then averaging (weights all classes equally regardless of frequency).

Multimodal Processing multiple input types simultaneously (e.g., text, images, spatial layout).

Self-supervised Learning Training paradigm where models learn representations from unlabeled data using automatically generated pseudo-labels (e.g., masked token prediction).

Semantic Segmentation Pixel-level classification assigning each pixel to a semantic category (e.g., "heading," "paragraph").

Transfer Learning Reusing knowledge learned on one task/dataset (e.g., modern documents) for a different but related task (e.g., historical documents).

Transformer Neural network architecture based on self-attention mechanisms, enabling modeling of long-range dependencies.

1 Introduction

The Federal Institute for Vocational Education and Training (BIBB) maintains a vast archive of historical vocational education (VET) regulations spanning the last century (1938–2022). While these documents preserve essential records of industrial transformation, they currently exist only as scanned images. The institute's ongoing digitization initiative aims to convert these records into structured TEI (Text Encoding Initiative) XML to make them computationally accessible. However, this process is currently bottlenecked by the difficulty of automated layout analysis. As noted in foundational surveys by [12], traditional rule-based methods struggle to generalize when confronted with the irregularities and evolving layouts inherent in historical materials.

1.1 Challenges in Historical Document Processing

The digitization of this corpus faces two distinct challenges: typographic variation and data scarcity. First, the documents exhibit extreme layout drift, transitioning from dense blackletter (Fraktur) typography in the 1930s to typewriter formats in the 1970s and computer typesetting in the 1990s. Fraktur scripts present a particular hurdle for Optical Character Recognition (OCR); even specialized engines suffer from the "vocabulary gap," often confusing characters like 'f' and the long s (*ſ*) [15]. Recent evaluations indicate that while LSTM-based models have improved performance, error rates on degraded historical pages can still exceed 15% [3, 14, 13]. This noise complicates downstream layout analysis, as models must be robust to corrupted text inputs. Second, the project operates under severe data constraints. The corpus consists of 668 pages, of which only 87 (approx. 13%) have been manually annotated with pixel-level masks for the required 18-class taxonomy. Preliminary experiments using a Multimodal Fully Convolutional Network (MFCN) on this subset yielded a mean Intersection over Union (IoU) of only 0.397. These results suggest that purely convolutional architectures or simple rule-based heuristics are insufficient for reconstructing complex document hierarchies [6].

1.2 Motivation and Proposed Solution

To address these limitations, this thesis proposes adapting LayoutLMv3 [10], a multimodal transformer architecture, to the specific domain of historical German legal texts. Unlike previous approaches, LayoutLMv3 unifies textual, visual, and spatial information, a paradigm that has shown remarkable success on modern benchmarks. Recent work on the Heimatkunde dataset [1] demonstrated that such multimodal transformers can effectively handle historical German documents, provided they are adapted to the specific domain. Our approach leverages the model's ability to model long-range dependencies essential for linking section headers to subsequent enumerations and its robustness to OCR noise via visual context [5]. By utilizing transfer learning from large-scale pre-training (drawing conceptually from masked modeling techniques in BEiT [2]), we aim to overcome the scarcity of annotated data. The goal is to move beyond simple text extraction to a semantic reconstruction of the document hierarchy, enabling the automated identification of headings, sections, and enumerations required for the BIBB digitization workflow.

1.3 Research Questions and Objectives

The challenge of transforming historical document images into structured, machine-readable data has occupied researchers for decades, yet certain domains remain underexplored. German

vocational education and training regulations present a particularly compelling case: these documents span nearly a century of typographic evolution, from ornate Fraktur scripts in the 1930s to standardized typewriter formatting in the 1980s. Current digitization workflows at the Federal Institute for Vocational Education and Training (BIBB) rely on manually crafted rules to identify structural elements headings, section markers, enumerated lists, footnotes but such heuristic approaches struggle when confronted with the sheer diversity of layouts across different eras [3]. Machine learning offers an alternative path, one that might learn to recognize patterns from examples rather than requiring explicit programming of every typographic convention. Still, applying modern deep learning to historical documents raises its own questions. How can methods handle degraded scans where ink has faded or pages have yellowed? What happens when training data is scarce when thousands of documents exist but only a small fraction bear the pixel-level annotations that supervised methods typically demand? And perhaps most fundamentally, can computational approaches actually surpass the rule-based systems that domain experts have refined over years of manual transcription work? These questions frame the investigation that follows.

1.3.1 Research Questions

The central aim of this thesis is to adapt multimodal transformer architectures for the semantic segmentation of historical German VET regulations. To guide this investigation, we define one main research question supported by two specific sub-questions:
Main Question: How can transfer learning with multimodal transformers improve semantic layout analysis of historical German Vocational Education and Training (VET) documents compared to existing rule-based and convolutional approaches?
RQ 1: How can multimodal models remain robust against historical challenges, specifically blackletter scripts (Fraktur), inconsistent typography across decades, and degraded scan quality?
RQ 2: To what extent can transfer learning from modern document corpora compensate for the scarcity of pixel-level annotations in the target historical dataset?

1.3.2 Elaboration on Research Questions

Main Question: Improving Semantic Analysis This inquiry goes beyond simply testing whether machine learning works on historical documents prior work suggests it can [6, 1]. The core challenge lies in designing an adaptation strategy for the specific constraints of the BIBB archive. The term "improve" is defined here by two metrics: quantitative segmentation accuracy (measured via IoU and Dice coefficients) and practical utility for the digitization pipeline. A model that achieves high accuracy on clean 1990s documents but fails on degraded 1940s pages offers limited value. Therefore, improvement implies developing a system that generalizes across the corpus's temporal range (1938–2022) better than the existing baseline, despite the noise inherent in the source material.
RQ 1: Robustness Against Historical Variation This question addresses the reality that historical materials rarely resemble the clean, standardized PDFs encountered in standard pretraining datasets. Fraktur typography presents a severe bottleneck for OCR systems designed for Antiqua fonts, with character error rates often exceeding 20% on poorly preserved documents [15]. Since multimodal models like LayoutLMv3 rely on OCR-derived bounding boxes and text embeddings, high error rates can introduce substantial noise.

We must investigate whether the model learns to be robust to these misrecognitions using visual cues to correct for textual errors or if corrupted embeddings fundamentally undermine

performance. While domain-adaptive training can mitigate some OCR difficulties [16] and specialized engines like Calamari show promise [13], the downstream effect on layout analysis remains an empirical question. Furthermore, the corpus exhibits extreme typographic drift: 1938 regulations feature dense paragraphs and ornate headers, while 1970s documents use monospaced typewriter fonts. We aim to determine if a single model can learn era-invariant patterns or if the temporal variance requires specific architectural adaptations.

RQ 2: Overcoming Data Scarcity This question tackles the project's primary constraint: the corpus contains 668 pages, but currently only 87 possess pixel-level segmentation masks for the required eighteen semantic classes. Traditional supervised learning typically fails in such low-resource settings. We address this through transfer learning from the LayoutLMv3 model pre-trained on large-scale modern document corpora like IIT-CDIP [10]. The working hypothesis is that general document understanding capabilities such as recognizing that bold, centered text typically indicates headings or that indented lists represent enumerations can transfer effectively from modern English documents to historical German regulations when fine-tuned on domain-specific examples.

However, we acknowledge that a minimum threshold of high-quality ground truth is required for stable fine-tuning across all 18 semantic classes. Therefore, this research incorporates an explicit contingency for annotation expansion: should preliminary experiments indicate that the initial 87 pages are insufficient for convergence, we will manually annotate additional documents to reach a target of approximately 120–150 pages. This expanded dataset will provide sufficient examples for the model to adapt its pre-trained representations to the specific visual and textual patterns of historical VET regulations.

1.4 Research Objectives

From the research questions emerge concrete objectives that define the scope of this thesis. These objectives translate conceptual inquiries into actionable experimental goals, balancing the ambition of using state-of-the-art transformers with the pragmatic constraints of a master's thesis. The primary objective is to develop and evaluate a semantic segmentation pipeline based on the LayoutLMv3 architecture, employing a transfer learning strategy to address the scarcity of annotated historical data. Rather than conducting computationally expensive pre-training from scratch, this research will leverage model weights pre-trained on large-scale modern document corpora (such as IIT-CDIP) and fine-tune them specifically for the BIBB historical archive. This involves adapting the model to a fine-grained 18-class taxonomy and implementing strategies to mitigate severe class imbalance, such as specialized loss functions or data augmentation, given that dominant classes like "Paragraph" vastly outnumber rare elements like "Centered Text." The core goal is to empirically validate whether general document understanding capabilities acquired from modern data can transfer effectively to the domain of historical German regulations despite the noise introduced by Fraktur typography and OCR errors.

To quantify the improvements offered by this transformer-based approach, the second objective focuses on rigorous benchmarking against established baselines. We will implement a heuristic baseline that extracts hand-crafted features from OCR output such as bounding box dimensions, spatial coordinates, and TF-IDF text vectors and applies unsupervised clustering (e.g., K-means) to establish a lower bound for performance. This comparison allows us to determine which semantic categories require the sophisticated contextual reasoning of a transformer and which can be separated by simple geometry alone. Additionally, the proposed model will be compared against the preliminary Multimodal Fully Convolutional Network (MFCN) results to assess whether the global attention mechanisms of LayoutLMv3 successfully resolve

the issue of spatial fragmentation observed in earlier convolutional experiments. Methodological insights regarding these comparisons will be drawn from recent work on the comparable Heimatkunde dataset [1].
Finally, should the primary objectives be met ahead of schedule, a third optional objective involves an exploratory analysis of the model's internal representations. This includes using dimensionality reduction techniques, such as t-SNE, to visualize whether the model's embeddings form coherent clusters corresponding to semantic classes before the final classification layer. Such analysis would illuminate what the model actually learns about document structure specifically, whether it distinguishes "Section Headers" from "Main Headers" based on semantic context or merely visual font size. Furthermore, this phase may explore integrating Conditional Random Fields (CRFs) on top of the transformer output to explicitly model the sequential logic of legal documents, thereby addressing any potential lack of structural coherence in the pixel-level predictions. These objectives collectively aim to provide a comprehensive evaluation of transfer learning for historical document analysis.

1.5 Expected Contributions

This research aims to bridge the gap between modern multimodal deep learning and the specific constraints of historical document processing. By adapting transformer-based architectures to the digitization of German Vocational Education and Training (VET) regulations, this thesis offers contributions across three dimensions: dataset creation, methodological evaluation, and practical institutional utility.

1.5.1 Dataset Contribution: A Complex Historical Benchmark

This work introduces a specialized corpus that fills a notable gap in resources for historical document layout analysis. While datasets like PubLayNet [19] cover modern scientific documents and the Heimatkunde collection [1] addresses historical books, neither captures the specific complexity of legal regulations spanning the 20th century. Our dataset distinguishes itself through:

- **Temporal Diversity:** Spanning 1938–2022, the corpus captures extreme typographic drift, from Fraktur scripts to typewriter fonts and modern computer typesetting.
- **Fine-Grained Taxonomy:** Unlike the seven-class scheme used in comparable historical works, our 18-class taxonomy captures dense structural hierarchies (nested enumerations, specific legal section markers, date stamps) required for TEI XML transcription.
- **Realistic Degradation:** The material reflects real-world archival conditions, including inconsistent preservation and varying scan qualities.

We anticipate that the annotated subset and the accompanying unlabeled corpus will serve as a valuable benchmark for testing model robustness against "vocabulary gaps" and layout drift in historical German documents.

1.5.2 Practical Utility and Institutional Impact

For the Federal Institute for Vocational Education and Training (BIBB), this research offers a pathway to automate the transcription of thousands of regulations currently locked in image formats. Current workflows rely on manual annotation or rigid rule-based heuristics that struggle with the formatting variations inherent in a century-long archive [12]. By developing a model capable of pre-segmenting documents with high semantic accuracy, we aim to shift the human role from "annotator" to "validator," potentially reducing processing time significantly. Furthermore, this work explicitly tackles the challenge of OCR noise in historical German texts [15, 13, 3]. By analyzing how multimodal transformers cope with corrupted text embeddings versus visual cues, we hope to identify robust digitization strategies that are applicable to other German archives facing similar blackletter recognition challenges. Finally, this thesis prioritizes methodological transparency over simple metric chasing. By conducting ablation studies and qualitative error analysis, we aim to document not just that the model works, but why it fails providing insights into the limits of transformer-based layout analysis when applied to the noisy reality of historical archives [4].

1.6 Proposed Approach

To address these limitations, this thesis proposes adapting LayoutLMv3 [10], a multimodal transformer architecture, to the specific domain of historical German legal texts through supervised fine-tuning. Unlike previous approaches, LayoutLMv3 unifies textual, visual, and spatial information within a single transformer backbone.

Our methodology relies on a transfer learning strategy that leverages the model's existing capabilities acquired through large-scale pre-training on modern document corpora such as IIT-CDIP (11 million documents) [10]. Rather than training from scratch or conducting additional domain-specific pre-training both computationally prohibitive within a master's thesis timeline we fine-tune these pre-trained weights directly on our annotated historical dataset. The pipeline proceeds as follows:

1. **Multimodal Input Processing:** We process the 87+ annotated pages to extract three synchronized modalities: image patches (capturing visual features like font weight and separators), OCR text tokens (capturing semantic content), and bounding box coordinates (capturing spatial layout).
2. **Supervised Fine-tuning:** The pre-trained transformer is fine-tuned on our 18-class taxonomy. To handle the "vocabulary gap" between modern English pre-training data and historical German Fraktur, we employ extensive data augmentation and explore domain-specific text normalization.
3. **Segmentation Head:** A pixel-level classification head projects the transformer's contextualized outputs onto the 18 semantic classes (e.g., Section Headers, Enumerations, Footnotes).

This approach draws methodological guidance from recent work on the Heimatkunde dataset [1], which demonstrated that multimodal transformers initialized on modern data can generalize effectively to historical blackletter documents through fine-tuning alone. By focusing our resources on supervised adaptation rather than expensive pre-training, we can rigorously evaluate whether the model's existing structural knowledge is sufficient to overcome the noise introduced by historical OCR and layout drift [15].

Thesis Statement: This thesis evaluates the transferability of the LayoutLMv3 multimodal transformer to historical German Vocational Education and Training (VET) documents. We aim to demonstrate that fine-tuning a model pre-trained on modern documents offers a robust and data-efficient alternative to rule-based systems, effectively handling blackletter typography and complex layout hierarchies despite limited annotated training data.

2 Literature Review

The field of document analysis has undergone a fundamental transformation from rule-based heuristics to data-driven, multimodal learning. This section examines the trajectory of these methodological advances, identifies gaps regarding historical legal documents, and positions our transfer learning approach within this landscape. A comprehensive comparison of these approaches is provided in Table 1.

2.1 From Rule-Based Heuristics to Vision-Only Models

Early work in document structure analysis relied on algorithmic methods that encoded explicit geometric rules. As surveyed by Mao et al. [12], foundational techniques like the X-Y Cut algorithm or Docstrum utilized recursive whitespace partitioning and nearest-neighbor clustering to identify text regions. While effective for standardized modern layouts, these methods falter when applied to historical archives. Our VET corpus (1938–2022) exhibits irregularities that violate the rigid assumptions of these algorithms: Fraktur typography blurs character boundaries, mechanical typesetting creates inconsistent spacing, and aging artifacts like ink bleed confuse connected-component analysis. The limitations of hand-crafted rules drove a shift toward deep learning, initially dominated by vision-only approaches. Chen et al. [6] and Wick and Puppe [17] demonstrated that Convolutional Neural Networks (CNNs) could learn to segment historical manuscripts by extracting visual features directly from pixel data. However, these models treated documents purely as images, ignoring the rich semantic content encoded in the text. This is a critical deficiency for legal regulations, where the structural role of a text block distinguishing a "Section Header" from a "Paragraph" is often defined by specific terminology (e.g., "§ 3") rather than visual appearance alone.

2.2 Multimodal Transformers and LayoutLMv3

To bridge the gap between visual layout and textual semantics, recent research has coalesced around multimodal transformer architectures. LayoutLMv3, introduced by Huang et al. [10], represents the current state-of-the-art in this domain. Unlike its predecessors, LayoutLMv3 employs a unified architecture that jointly processes text tokens, layout coordinates, and image patches without relying on a separate CNN backbone. A key innovation relevant to our work is the model's visual processing strategy, which draws on the concept of discrete visual tokens introduced in BEiT [2]. By discretizing image patches into "visual words," the model can align linguistic semantics with visual structures during its initial pre-training on the massive IIT-CDIP dataset (11 million documents). This pre-training equips the model with a general understanding of document logic for example, that bold, centered text at the top of a page usually signifies a title. Our research investigates the transferability of this general knowledge to the specific domain of historical German VET regulations. While LayoutLMv3 achieves remarkable results on modern benchmarks like PubLayNet, it has not been systematically evaluated

on documents featuring blackletter typography or the extreme layout drift observed over the 20th century. We hypothesize that fine-tuning these pre-trained weights will allow the model to adapt to historical idiosyncrasies more effectively than training a vision-only model from scratch, essentially bypassing the need for massive annotated historical datasets.

2.3 Heimatkunde: Benchmarking Historical Layout Analysis

The most relevant precedent for our work is the recent introduction of the Heimatkunde dataset by Baloun et al. [1]. This corpus consists of late 19th-century Czech district histories printed in Fraktur, providing a rare benchmark for multimodal analysis on historical German-language materials. Their experiments demonstrated that a fusion of BERT text embeddings and Vision Transformer features significantly outperformed unimodal baselines, validating the necessity of a multimodal approach for historical documents. However, significant differences distinguish our VET corpus from Heimatkunde. First, while Heimatkunde represents a static snapshot of 19th-century printing, our dataset captures nearly a century of temporal evolution, requiring a model that remains robust as typography shifts from Fraktur to Antiqua and layouts transition from manual to digital typesetting. Second, our 18-class taxonomy is significantly more granular than their 7-class scheme, designed to capture the dense legal hierarchies (nested enumerations, specific section markers) required for TEI XML transcription. Consequently, our work extends their findings by evaluating whether the transfer learning paradigm remains effective under stricter semantic requirements and greater temporal variation.

Table 1: Comprehensive Literature Review Comparison. This table contextualizes our proposed transfer learning approach against classical rule-based methods and recent multimodal benchmarks.

Approach	**Modality**	**Doc Type**	**Lang**	**Key Innovation**	**Training Strategy**	**Strengths**	**Limitations for VET**
Mao et al. (2003) [12]	N/A	General docs	Agnostic	Taxonomy of rule-based structure analysis	N/A (Heuristic)	Low computational cost; interpretable rules	Fails with historical variance and Fraktur typography
Chen et al. (2015) [6]	Vision	Hist. manuscripts	Various	Convolutional Autoencoders for feature learning	Unsupervised Reconstruction	Handles visual noise better than rules	Ignores textual semantics; misses legal context
Wick & Puppe (2018) [17]	Vision	Hist. docs	Various	Fully Convolutional Networks (FCN)	Supervised (Pixel-level)	Effective for complex page layouts	No text integration; requires large labeled sets
Cao et al. (2022) [5]	Vision + Text	Hist. docs	Various	Hybrid rule-based features + supervised classifiers	Supervised	Bridges classical and modern methods	Requires manual feature engineering; not end-to-end
Huang et al. (LayoutLMv3, 2022) [10]	Vis + Text + Layout	Modern business docs	Multilingual	Unified transformer for text, image, and layout	Self-supervised on IIT-CDIP (11M docs)	SOTA on modern docs; unified architecture	Pre-trained on modern English; untested on Fraktur
Baloun et al. (Heimatkunde, 2024) [1]	Vis + Text	Hist. books	German (Fraktur)	First multimodal benchmark for historical Fraktur	Supervised (Transfer from PubLayNet)	Validates multimodal approach for history	Limited taxonomy (7 classes); single time period
Our Approach	**Vis + Text + Layout**	**VET Regs (1938–2022)**	**German (Mixed)**	**Adapting LayoutLMv3 to fine-grained legal taxonomy**	**Supervised Fine-tuning (Transfer Learning)**	**Models temporal drift; 18-class legal taxonomy**	**Relies on quality of pre-trained weights; OCR noise sensitivity**

3 Methodology

The proposed methodology addresses the semantic layout analysis of historical German Vocational Education and Training (VET) documents through a transfer learning approach centered on LayoutLMv3 [10]. This strategy diverges from earlier rule-based heuristics [12] and vision-only convolutional methods [6, 17] by explicitly modeling the interdependence between document appearance, linguistic content, and geometric layout. We recognize that the semantic understanding of legal regulations emerges from the interaction of these modalities rather than from any single one in isolation. The methodological design draws inspiration from recent advances in multimodal historical document analysis, particularly the Heimatkunde dataset [1], which demonstrated effective fusion strategies on German blackletter materials, and BEiT's discrete token reconstruction approach [2]. Given the severe annotation constraints where only 87 of 668 corpus pages bear pixel-level labels, this approach focuses on extracting maximum representational value from pre-trained weights adapted to our specific domain.

3.1 Data Preparation

The empirical foundation of this research rests upon a corpus of approximately 668 historical document pages drawn from the Federal Institute for Vocational Education and Training (BIBB) occupational archive. These materials comprise German VET regulations spanning nearly a century from 1938 through 2022, representing a substantial temporal range that encompasses fundamental shifts in typographic conventions. The collection exhibits considerable heterogeneity, ranging from early documents featuring dense Fraktur typography with ornate section headings to mid-century typewriter formatting and recent computer-typeset layouts. This diversity offers a realistic test bed for evaluating whether machine learning approaches can generalize across historical typographic variation rather than merely optimizing for a narrow temporal window. The preparation pipeline follows the protocol established by Baloun et al. [1], subjecting each document page to systematic preprocessing to extract visual, textual, and spatial information streams. Visual representations are generated by converting scanned images to normalized grayscale at a consistent resolution, applying adaptive noise reduction to mitigate degradation artifacts common in older materials such as yellowing and ink bleed-through. For text extraction, we evaluate two OCR strategies on a representative 50-page subset: embedded PDF text layers versus Tesseract 5.0 equipped with specialized German Fraktur language models [13]. Since prior literature indicates that character error rates on degraded historical texts can exceed 30% [15], the method exhibiting superior recall for text region detection will be selected for corpus-wide preprocessing to ensure the multimodal model receives robust textual inputs.

3.1.1 Annotation Examples and Class Taxonomy

Of the full corpus, 87 pages have been manually annotated with pixel-level segmentation masks identifying 18 semantic document classes. These annotations distinguish structural elements critical for TEI (Text Encoding Initiative) XML transcription, including main headings, section headings, section markers prefixed with paragraph symbols (§), enumerated lists at multiple nesting levels, paragraph bodies, and footnotes. The taxonomy reflects the hierarchical structure characteristic of German legal regulations, where numbered paragraphs organize substantive requirements and nested enumerations specify training criteria.

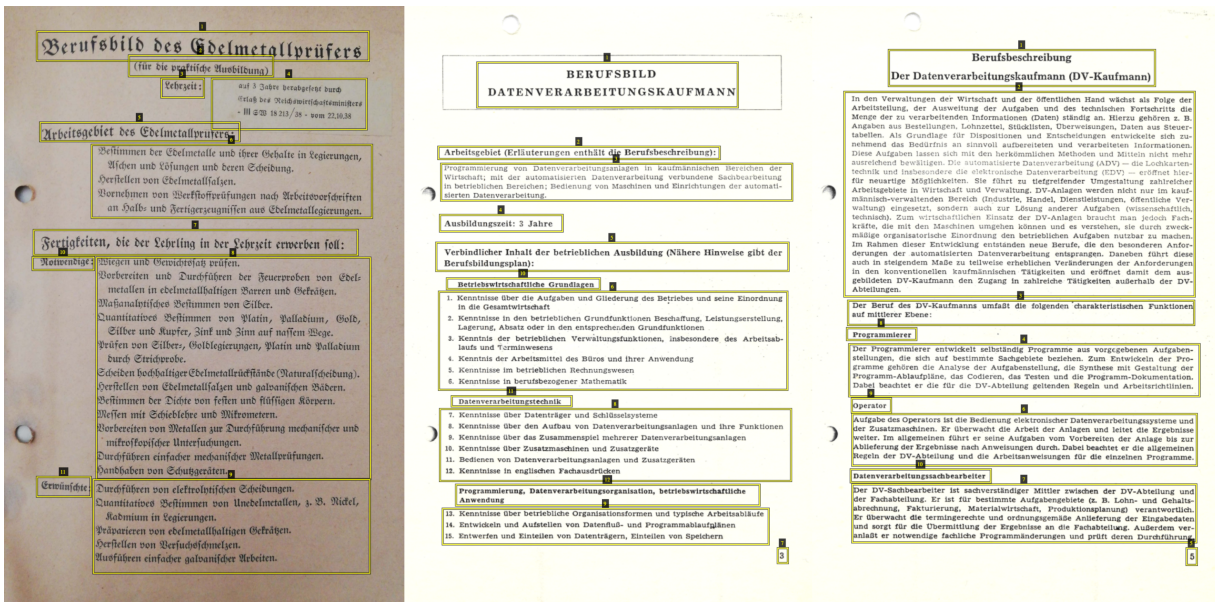

Figure 1: Example annotation from the 87-page labeled dataset, showing pixel-level segmentation masks for 18 semantic classes on a 1950s VET regulation. The document exhibits mixed Fraktur-Antiqua typography typical of the transitional period. Note the fine-grained distinctions between section markers (§ symbols), section headings (following text), and nested enumeration levels categories that rule-based methods frequently confuse.

As illustrated in Figure 1, this fine-grained categorization allows for the reconstruction of complex document hierarchies that rule-based methods frequently confuse. The dataset exhibits substantial class imbalance, as paragraph bodies comprise roughly 45% of annotated regions while specialized elements like centered text appear in fewer than 1% of annotations. This imbalance poses methodological challenges that will be addressed through the use of Focal Loss during fine-tuning. Furthermore, optional augmentation strategies may be explored if initial experiments reveal severe data scarcity constraints. Synthetic generation of additional training examples through typographic transformations could artificially expand the labeled set, though such augmentation risks introducing unrealistic artifacts that degrade generalization [7]. Finally, it is important to note that document structures spanning multiple pages are acknowledged as beyond the current scope. The LayoutLMv3 architecture processes pages independently, lacking mechanisms for cross-page reasoning [10], and thus this thesis focuses on establishing effective single-page semantic segmentation as a necessary foundation for future multi-page reconstruction efforts.

3.2 Model Architecture: LayoutLMv3 Adaptation

The architectural foundation for this research is LayoutLMv3 [10], a unified multimodal transformer that processes documents through the simultaneous encoding of textual content, visual representations, and spatial layout. This approach represents a methodological departure from our preliminary Multimodal Fully Convolutional Network (MFCN) experiments, which achieved only modest performance on the annotated subset. Unlike convolutional architectures that process images through hierarchical feature maps and typically require late fusion of independently trained encoders, LayoutLMv3 embeds all modalities within a shared transformer backbone. This design enables cross-modal attention mechanisms to learn interdependencies during the forward pass rather than imposing them through rigid architectural constraints.

Our implementation leverages a transfer learning strategy, initializing the model with weights pre-trained on the massive IIT-CDIP dataset (11 million modern documents). This pre-training phase, conducted by the original LayoutLMv3 authors [10], employed self-supervised objectives Masked Language Modeling (MLM), Masked Image Modeling (MIM) inspired by BEiT [2], and Word-Patch Alignment. We do not repeat this pre-training process. Instead, we directly fine-tune these publicly available pre-trained weights on our annotated historical dataset using supervised learning. Understanding these pre-training objectives is valuable for interpreting the model's capabilities: the MLM objective enables semantic reasoning about text, while the MIM objective trains the model to reconstruct discrete visual tokens, allowing it to distinguish typographic features such as font weight and alignment patterns. By initializing with these capabilities rather than training from scratch, we benefit from learned document structures such as the hierarchical relationship between headings and body text without requiring massive computational resources or millions of annotated pages.

The unified design supports multi-modal feature fusion at every transformer layer. When processing a complex element like a section heading, the model simultaneously attends to textual signals (e.g., the section marker "§"), visual cues (bold weight, larger font size), and spatial context (centering or top-of-page positioning). This cross-modal attention allows evidence from one modality to compensate for ambiguity in another. For instance, if OCR recognizances the word "Abschnitt" due to Fraktur character confusion, the visual embedding of the distinctive heading typography may still enable correct classification. Such robustness is particularly valuable for historical documents where text recognition errors remain inevitable despite recent advances in OCR technology [15, 13].

Adapting this pre-trained architecture to historical German legal regulations requires domain-specific preprocessing to bridge the gap between the modern pre-training data and our archival corpus. Textual preprocessing incorporates normalization strategies to handle noisy OCR output characteristic of blackletter typography; historical orthographic conventions such as the long s (*ſ*) and various ligatures are mapped to modern Unicode equivalents to reduce vocabulary fragmentation. Furthermore, the text encoder's vocabulary is augmented with domain-specific legal terms extracted from the corpus to minimize out-of-vocabulary tokens. Visual preprocessing applies adaptive histogram equalization to compensate for contrast variations introduced by aging, such as faded ink or yellowed paper, while image resolution is standardized to match the model's configuration. Finally, layout preservation is ensured by normalizing bounding box coordinates to relative positions, achieving invariance across different scan dimensions.

The workflow illustrated in Figure 2 depicts the complete pipeline from document input to prediction. Page images are divided into patches, OCR text is tokenized, and bounding box coordinates are embedded as continuous features. These three streams merge within the transformer encoder, where multi-head self-attention mechanisms compute dependencies both within and across modalities. By initializing this encoder with pre-trained weights, we enable the model to recognize high-level document structures such as the hierarchical relationship between a section heading and a subsequent list without requiring the massive dataset usually needed to learn such patterns from scratch. This capability addresses the limitations of our earlier convolutional experiments, where the lack of global context led to spatially fragmented predictions for long document elements [1].

3.3 Training Procedure

This research employs a single-phase supervised fine-tuning approach rather than multi-stage pre-training. We initialize the LayoutLMv3 model with publicly available weights pre-trained

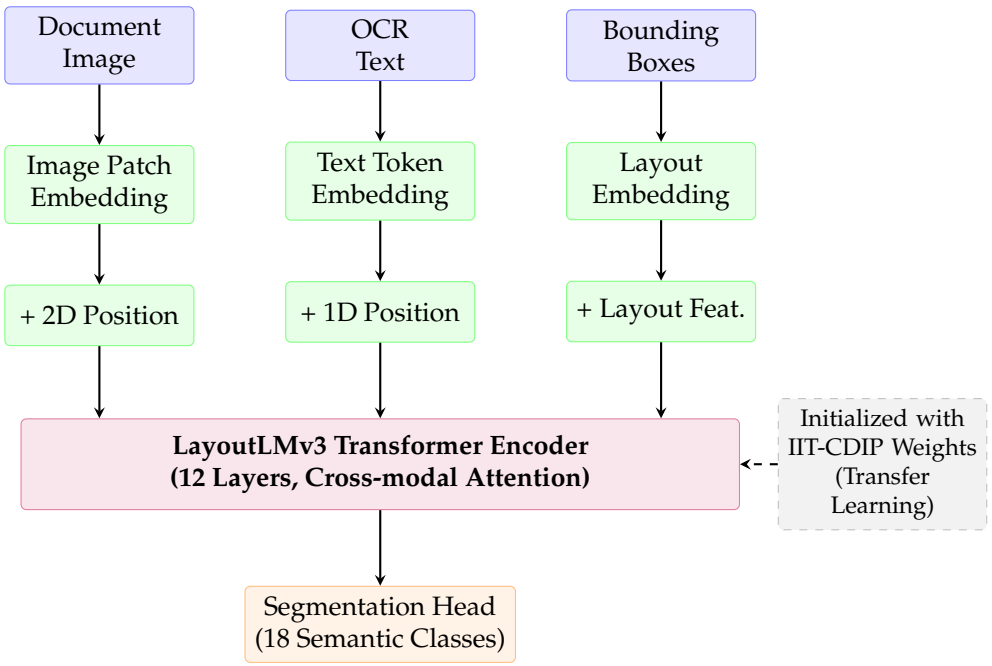

Figure 2: Adapted LayoutLMv3 architecture for fine-tuning. The model processes three synchronized modalities images, text, and layout through a transformer backbone initialized with weights from large-scale pre-training. A task-specific segmentation head projects the contextualized embeddings onto our 18-class historical taxonomy.

on the IIT-CDIP dataset [10] and directly fine-tune on our annotated historical pages. We do not conduct additional self-supervised pre-training on the unlabeled portion of our 668-page corpus, as this would require substantial computational resources beyond the scope of a master's thesis. Our focus is on evaluating how effectively modern document understanding capabilities transfer to the historical domain through supervised adaptation alone.

The training procedure employs a supervised fine-tuning strategy that adapts the pre-trained LayoutLMv3 model to our specific historical domain. This approach addresses the project's primary constraint annotation scarcity by transferring general document understanding capabilities from large-scale public pre-training to the specific task of semantic segmentation on our 87 annotated pages [10].

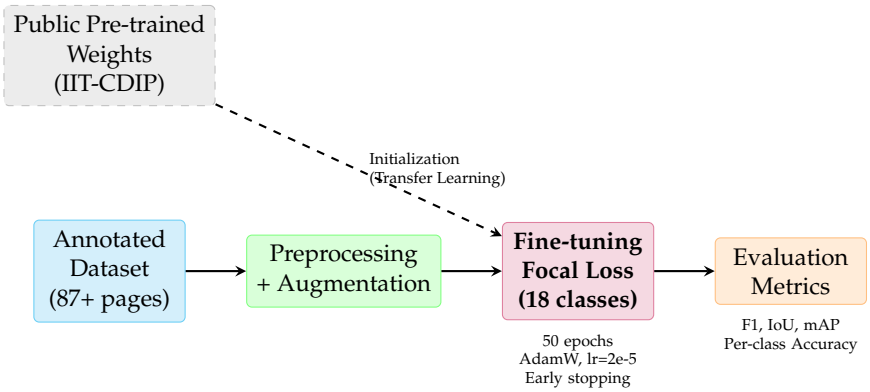

Figure 3: Single-phase transfer learning workflow. The model is initialized with weights pre-trained on modern public corpora (IIT-CDIP) and fine-tuned on the historical annotated dataset. This bypasses the need for large-scale domain-specific pre-training while leveraging learned visual-textual representations.

As illustrated in Figure 3, the workflow begins by initializing the LayoutLMv3 transformer with weights pre-trained on the IIT-CDIP dataset. The model receives a trimodal input consisting of image patches, OCR text tokens, and layout coordinates, where each input token is associated with a ground truth label from our 18-class taxonomy. A task-specific classification head maps the transformer's contextualized embeddings to class probabilities. The dataset is partitioned into training (70 pages), validation (10 pages), and test (7 pages) splits using stratified sampling to ensure proportional representation of temporal periods and rare classes.
Training minimizes Focal Loss [11], a choice necessitated by the severe class imbalance documented in Section 3.1, where paragraph bodies dominate 45% of the data. Unlike standard Cross-Entropy, Focal Loss down-weights easy examples and focuses the model's attention on hard, misclassified examples. This formulation is critical for learning minority classes like section markers and footnotes, which are essential for reconstructing the document hierarchy [12].
To prevent overfitting on the small dataset, we apply rigorous data augmentation during training. Geometric perturbations such as random cropping and rotation simulate scanning inconsistencies, while photometric augmentations including Gaussian noise and contrast adjustment replicate historical degradation artifacts like yellowing and faded ink [6, 18].

3.3.1 Fine-tuning Protocol

The model receives a trimodal input consisting of image patches, OCR text tokens, and layout coordinates. Each input token is associated with a ground truth label from our 18-class taxonomy. A task-specific classification head maps the transformer's contextualized embeddings to class probabilities. The dataset is partitioned into training (70 pages), validation (10 pages), and test (7 pages) splits using stratified sampling to ensure proportional representation of temporal periods and rare classes. Training minimizes Focal Loss [11], a choice necessitated by the severe class imbalance documented in Section 3.1 (where paragraph bodies dominate 45% of the data). Unlike standard Cross-Entropy, Focal Loss down-weights easy examples and focuses the model's attention on hard, misclassified examples. This is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

where $\gamma = 2.0$ reduces the loss contribution of well-classified tokens, preventing the model from achieving high accuracy by simply predicting the majority class. This formulation is critical for learning minority classes like section markers and footnotes, which are essential for reconstructing the document hierarchy [12].

To prevent overfitting on the small dataset, we apply rigorous data augmentation during training. Geometric perturbations (random cropping, rotation $\pm 2^\circ$) simulate scanning inconsistencies, while photometric augmentations (Gaussian noise, contrast adjustment) replicate historical degradation artifacts like yellowing and faded ink [6, 18].

3.4 Evaluation Framework

The evaluation strategy balances quantitative rigor with qualitative insight to assess whether the transformer-based approach offers a tangible improvement over rule-based systems for BIBB's digitization workflow.

3.4.1 Quantitative Metrics

Performance is measured on the held-out test set using standard segmentation metrics established in the document analysis literature [19, 7]. The primary indicator of segmentation quality is the Mean Intersection over Union (mIoU), a strict metric that penalizes both false positives and false negatives to provide a robust measure of boundary accuracy. To account for the dataset's significant class imbalance, we also compute the Macro-Averaged F1-Score. Unlike micro-averaging, which can be skewed by dominant classes, macro-averaging treats all classes equally, ensuring that performance on rare but critical categories such as "Centered Text" or "Document IDs" contributes as much to the final score as the abundant "Paragraph" class. Furthermore, we conduct a granular per-class breakdown to specifically analyze structural elements such as headings, enumerations, and section markers, which are essential for the downstream TEI XML transcription process. These results will be benchmarked against two distinct baselines: the preliminary Multimodal Fully Convolutional Network (MFCN) experiments, which achieved a mean IoU of 0.397 [18], and a rule-based heuristic baseline implemented following standard layout analysis logic [12].

3.4.2 Success Criteria

Success is defined through a tiered framework that establishes clear performance thresholds ranging from viability to state-of-the-art performance. Minimum viable success is marked by a mean IoU exceeding 0.45, a threshold that represents a measurable improvement over the MFCN baseline and demonstrates the model's fundamental ability to handle historical Fraktur inputs despite OCR noise [15]. Beyond this baseline, the target success criteria aim for a mean IoU greater than 0.52 and a Macro-F1 score exceeding 0.60, with the crucial requirement of consistent performance across both older Fraktur-based documents (1930s) and modern layouts (1990s). Achieving this level would validate the model's practical utility for reducing the manual annotation effort in the BIBB workflow [1]. Finally, aspirational or "stretch" success is defined by a mean IoU exceeding 0.65, a performance level that would approach state-of-the-art results on comparable historical datasets such as Heimatkunde.

3.4.3 Qualitative Analysis

Beyond numbers, we conduct a systematic error analysis on a sample of test predictions. We categorize failures into boundary imprecision, class confusion (e.g., Section vs. Subsection), and OCR-induced errors. This visual inspection is vital for identifying "plausible but wrong" predictions such as identifying a single word as a heading due to bold font that metrics alone might miss [7].

3.4.4 Feasibility and Limitations

The core scope preprocessing, fine-tuning, and evaluation is achievable within the standard master's thesis timeline using available GPU resources (e.g., NVIDIA A100). We explicitly exclude cross-page structure reconstruction from the current project, focusing solely on establishing a robust single-page segmentation baseline. The primary risk remains OCR quality; if character error rates on Fraktur documents exceed 20%, the multimodal advantage may diminish, a limitation we will quantify through our ablation studies [13].

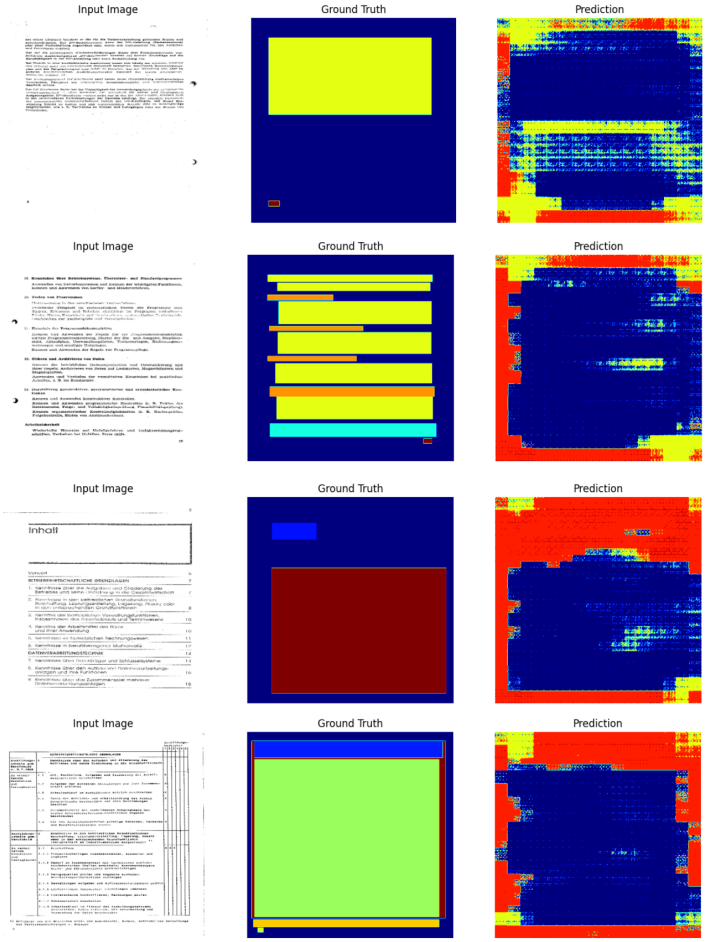

Figure 4: Preliminary baseline predictions demonstrating fragmentation. The proposed transformer model aims to resolve this lack of spatial coherence by modeling global dependencies.

4 Preliminary Results and Discussion

The evolution of this research has followed a trajectory characteristic of exploratory work in document understanding: initial experiments established feasibility constraints, revealing both the promise and limitations of baseline approaches before motivating a methodological pivot toward sophisticated multimodal transformers. Early investigations centered on a Document Recognition Fusion Network (DRFN), a convolutional architecture combining visual feature extraction with embedded textual representations [18, 6]. These preliminary experiments served primarily as proof-of-concept validation, demonstrating that machine learning could be applied to historical legal documents while simultaneously exposing fundamental architectural constraints that necessitated rethinking the approach.

4.1 Context and Transition from Early Experiments

The initial DRFN experiments operated on a limited annotated dataset comprising approximately 30-40 document pages with pixel-level semantic annotations across 18 structural classes. The network architecture employed a fully convolutional design, processing document images through successive downsampling and upsampling layers to produce dense segmentation masks [17]. Textual information, extracted via OCR using Tesseract with German language models [13], was embedded through pre-trained German FastText vectors and integrated with visual features through concatenation at intermediate layers [18]. Quantitative evaluation revealed substantial challenges. IoU metrics remained modest across all semantic classes, with the mean IoU approaching only 0.40 and the Dice coefficient achieving approximately 0.49 [6]. These figures, while demonstrating that the network learned to distinguish document regions beyond random chance, fell considerably short of the performance levels required for practical deployment. Qualitative inspection confirmed these limitations, as the model produced fragmented, inconsistent segmentations where ground truth annotations delineated coherent semantic regions [7]. Visual examination revealed that the network struggled to maintain spatial coherence, often fragmenting continuous sections into disconnected components and exhibiting class confusion between semantically similar elements [18]. This failure was attributed to the purely convolutional architecture's inability to model long-range dependencies [9] and the inefficiency of simple concatenation for multimodal fusion [10].

4.2 Methodological Transition to LayoutLMv3

The decision to adopt LayoutLMv3 as the primary methodological framework emerged from convergent evidence in recent document understanding literature [10, 1]. LayoutLMv3 represents a fundamental architectural departure from the convolutional paradigm. Where DRFN processed visual and textual modalities through separate pathways before mechanically fusing them, LayoutLMv3 employs a unified transformer encoder that processes text tokens, image patches, and spatial layout coordinates within a shared representational space [10, 2]. The architectural unification enables crossmodal attention mechanisms to learn interdependencies between modalities, allowing the network to discover that the word "§ 3" in a particular visual context should be classified as a section marker, while identical text in a different context represents something else entirely. Such contextual reasoning proves difficult for convolutional architectures but emerges naturally from transformer self-attention [8]. Furthermore, by initializing with pre-trained weights rather than training from scratch, we aim to bypass the limitations of our small annotated dataset. The LayoutLMv3 model we employ has already learned general document understanding patterns from 11 million modern documents [10]. Our task

is to adapt these learned representations to historical German VET regulations through supervised fine-tuning on our domain-specific annotations, addressing the annotation scarcity constraint that hindered the purely supervised DRFN approach.

4.3 Current Research Status and Data Preparation

The present stage centers on implementing the fine-tuning pipeline and finalizing the dataset. Data preparation has advanced substantially, with the corpus of approximately 668 historical document pages collected from the BIBB archive and converted to standardized image formats. Two OCR strategies have been evaluated on a representative 50-page subset: embedded PDF text layers and Tesseract 5.0 with specialized German Fraktur language models [13]. Early assessment suggests that for documents from the 1970s onward, embedded PDF OCR achieves acceptable accuracy. However, for earlier materials featuring Fraktur typography, Tesseract proves necessary despite increased computational cost, as error rates remain in the 10-15% range for degraded scans [15, 16]. The annotation expansion effort represents a critical parallel workstream. The initial 87 manually annotated pages constitute barely 13% of the full corpus and exhibit imbalanced class distribution [11]. To address this, additional annotation has commenced using Label Studio to expand the dataset to approximately 120-150 pages before commencing the final fine-tuning runs. This expansion aims to balance annotation quality, temporal diversity, and class coverage, ensuring the model is not biased toward specific time periods [19].

4.4 Discussion and Interpretation

The transition from early convolutional experiments to the current LayoutLMv3 implementation reflects accumulated empirical evidence about what document understanding requires. The preliminary results contributed valuable knowledge: document layout analysis demands long-range dependency modeling [9] and sophisticated multimodal reasoning [10]. The decision to utilize transfer learning rests on the hypothesis that general-purpose document representations learned from modern documents can transfer effectively to historical tasks when fine-tuned [1]. However, several aspects of this hypothesis merit scrutiny. First, the extent to which modern pre-training weights (from IIT-CDIP) can bridge the domain gap to historical VET documents (1938-2022) remains an empirical question. The VET documents exhibit considerable typographic diversity, and the "vocabulary gap" caused by Fraktur OCR errors may undermine the model's textual understanding [15]. Second, the 18-class taxonomy, while designed for TEI XML encoding [12, 4], introduces label noise due to ambiguous cases, such as distinguishing "main" from "section" headings across decades of changing conventions [7]. The ongoing expansion of the annotated dataset with strict quality control mechanisms aims to mitigate these concerns, positioning the research to rigorously evaluate whether multimodal transformers can succeed where convolutional baselines failed.

5 Work Packages and Project Timeline

The successful completion of this master's thesis requires systematic progression through interconnected research activities, organized into discrete work packages that build upon one another. The timeline spans approximately six months of full-time equivalent work, accounting for the staged nature of machine learning projects where infrastructure development and data preparation must precede model training and evaluation.

5.1 Work Package Structure

The project decomposes into six primary work packages (WPs), reflecting both logical dependencies and practical resource allocation.

WP1: Data Collection and Quality Assessment constitutes the foundational activity. This package encompasses the acquisition of document scans from the BIBB archive and their conversion to standardized 300 DPI grayscale images. A critical subtask involves evaluating OCR quality by comparing embedded PDF text layers against Tesseract-based extraction [13]. This evaluation determines which strategy best balances accuracy and computational cost across the corpus's typographic diversity [15].

WP2: Annotation Expansion addresses the constraint that only 87 of 668 corpus pages currently bear pixel-level annotations. This package centers on expanding the dataset to approximately 120–150 pages using Label Studio. The process involves training annotators to apply the 18-class taxonomy consistently, with strict guidelines to disambiguate borderline cases (e.g., Main Heading vs. Section Heading). Inter-annotator agreement analysis on a 10% overlap subset will quantify consistency [7], ensuring high-quality ground truth for the supervised learning phase.

WP3: Model Implementation and Transfer Learning operationalizes the LayoutLMv3 architecture. Instead of training from scratch, we implement the transfer learning pipeline by initializing the model with weights pre-trained on the IIT-CDIP dataset [10]. Key tasks include adapting the tokenizer to handle German historical orthography (e.g., ligatures, long s (*ſ*)) and developing the data loading pipeline to generate the trimodal inputs (image patches, text tokens, layout coordinates) required by the transformer.

WP4: Supervised Fine-tuning adapts the pre-trained model to our specific semantic segmentation task. Training employs the expanded annotated dataset partitioned into training, validation, and test splits. We minimize Focal Loss [11] to address the severe class imbalance where paragraph bodies dominate specialized classes like footnotes. Hyperparameter optimization will explore learning rates, batch sizes, and augmentation strategies (geometric and photometric perturbations) to maximize generalization. This package also includes the implementation of the rule-based and MFCN baselines for comparative assessment [18].

WP5: Evaluation and Analysis assesses model performance through quantitative metrics and qualitative error analysis. We compute standard segmentation metrics (mIoU, Macro-F1) on the held-out test set [19], with a specific focus on minority classes critical for document structure reconstruction. Qualitative analysis involves inspecting failure modes such as boundary imprecision or class confusion to understand the model's limitations [7].

WP6: Thesis Documentation runs concurrently with the later phases, transforming experimental results into the final manuscript. This includes drafting chapters, creating high-quality visualizations of the architecture and results, and ensuring rigorous academic formatting.

5.2 Timeline and Gantt Chart

The project timeline allocates these packages across 24 weeks. The schedule acknowledges that data annotation (WP2) is a time-intensive human process, while model training (WP4) is computationally intensive but faster to execute given the transfer learning approach. As illustrated in Figure 5, the schedule emphasizes parallel execution. Weeks 5–12 see simultaneous progress on annotation (WP2) and model implementation (WP3). This efficiency ensures that once the dataset is ready, the model infrastructure is immediately available for fine-tuning. The timeline includes slack periods to accommodate risks such as poor OCR quality or slower-than-expected convergence [16].

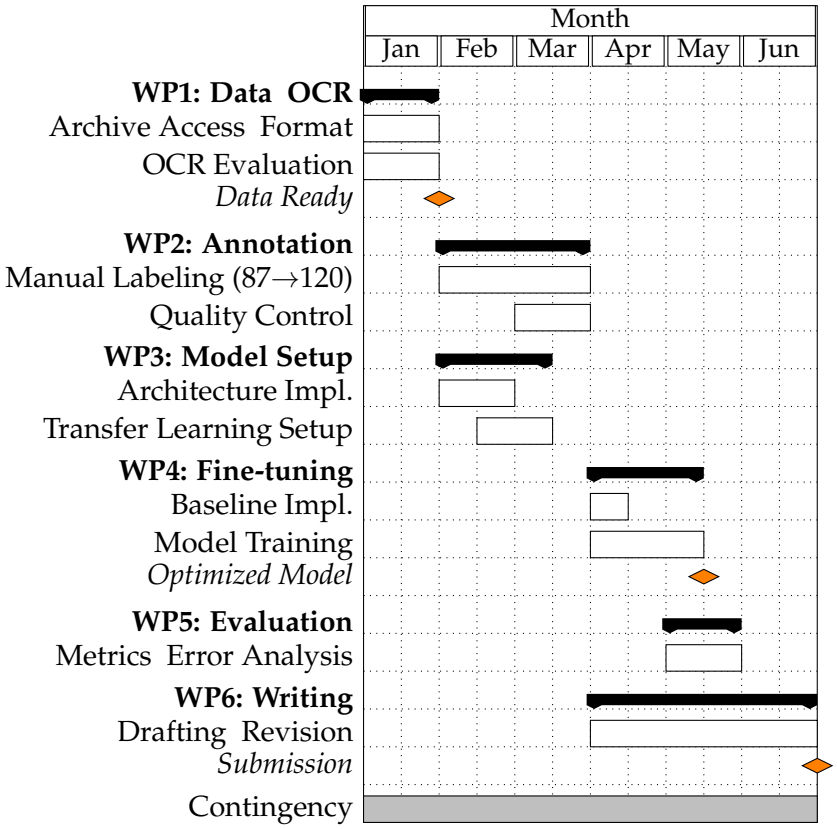

Figure 5: Project Timeline. WP2 (Annotation) and WP3 (Model Setup) run in parallel to maximize efficiency. The fine-tuning phase (WP4) commences once sufficient data is labeled. A continuous buffer (red) accounts for potential delays.

5.3 Deliverables and Success Criteria

The primary deliverable is the thesis manuscript, supported by technical artifacts: the curated corpus with documented OCR quality, the expanded annotated dataset, and the fine-tuned LayoutLMv3 model weights. Success is defined not merely by achieving a specific metric threshold, but by generating rigorous empirical evidence regarding the viability of multimodal transformers for historical documents. Achieving a mean IoU improvement of 10–15 percentage points over the MFCN baseline would constitute a significant contribution [19], validating the transfer learning approach for the BIBB digitization initiative. Conversely, documenting failure modes such as the model's inability to handle Fraktur OCR noise would provide equally valuable insights for future research [17].

References

[1] Josef Baloun et al. „Heimatkunde: Dataset for Multi-Modal Historical Document Analysis“. In: *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*. INSTICC. SciTePress, 2024, pp. 995–1001. ISBN: 978-989-758-680-4. DOI: 10.5220/0012428500003636.

[2] Hangbo Bao et al. „BEiT: BERT Pre-Training of Image Transformers“. In: *International Conference on Learning Representations (ICLR)*. 2022. URL: https://arxiv.org/abs/2106.08254.

[3] Thomas M. Breuel. „High Performance Text Recognition Using a Hybrid Convolutional-LSTM Implementation“. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 11–16. DOI: 10.1109/ICDAR.2017.12.

[4] Ruiyang Cao et al. „Extracting Variable-Depth Logical Document Hierarchy from Long Documents: Method, Evaluation, and Application“. In: *Journal of Computer Science and Technology* 37.3 (2022), pp. 699–718. DOI: 10.1007/s11390-021-1076-7. URL: http://dx.doi.org/10.1007/s11390-021-1076-7.

[5] Shuyang Cao and Lu Wang. „HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization“. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland, 2022, pp. 786–807. DOI: 10.18653/v1/2022.acl-long.58.

[6] Kai Chen et al. „Page Segmentation of Historical Document Images with Convolutional Autoencoders“. In: *13th International Conference on Document Analysis and Recognition (ICDAR)*. Tunis, Tunisia, 2015, pp. 1011–1015. DOI: 10.1109/ICDAR.2015.7333914.

[7] Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. „Scenario Driven In-depth Performance Evaluation of Document Layout Analysis Methods“. In: *2011 International Conference on Document Analysis and Recognition (ICDAR)*. Framework for evaluating layout analysis with limited ground truth. Beijing, China, 2011, pp. 1404–1408. DOI: 10.1109/ICDAR.2011.282.

[8] Jacob Devlin et al. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[9] Alexey Dosovitskiy et al. „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[10] Yupan Huang et al. „LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking“. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4083–4091. DOI: 10.1145/3503161.3548112.

[11] Tsung-Yi Lin et al. „Focal Loss for Dense Object Detection“. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, Oct. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.324.

[12] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. „Document structure analysis algorithms: A literature survey“. In: *Document Recognition and Retrieval X*. Vol. 5010. SPIE. 2003, pp. 197–207. DOI: 10.1117/12.476326.

[13] Christian Reul et al. „OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings“. In: *Applied Sciences* 9.22 (2019), p. 4853. DOI: 10.3390/app9224853. URL: https://www.mdpi.com/2076-3417/9/22/4853.

[14] Christian Reul et al. „State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines“. In: 2018. arXiv: 1810.03436 [cs.CV]. URL: https://arxiv.org/abs/1810.03436.

[15] Uwe Springmann and Anke Lüdeling. „OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus“. In: *Digital Humanities Quarterly* 11.2 (2017). URL: http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.

[16] Uwe Springmann et al. „Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin“. In: (2018). arXiv: 1809.05501 [cs.CL]. URL: https://arxiv.org/abs/1809.05501.

[17] Christoph Wick and Frank Puppe. „Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images“. In: *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 287–292. DOI: 10.1109/DAS.2018.39.

[18] Xiao Yang et al. „Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 2017, pp. 5315–5324. DOI: 10.1109/CVPR.2017.462.

[19] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. „PubLayNet: Largest Dataset Ever for Document Layout Analysis“. In: *15th International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, Australia, 2019, pp. 1015–1022. DOI: 10.1109/ICDAR.2019.00166.