

Recognition of Layout Patterns in Historical Legal Texts

A Thesis on Digitizing German VET Regulations Using Multimodal Transformers

Student: Mohammad Obaidullah Tusher | **Supervisor:** Prof. Dr. Jan Jürjens | **Co-Supervisor:** Thomas Reiser

Institution: University of Koblenz, Institute for Software Technology

Introduction & Context

The Federal Institute for Vocational Education and Training (BIBB) maintains an archive of German VET (Vocational Education and Training) regulations. This archive contains thousands of documents spanning nearly a century (1938–2022). The goal is to transform scanned images into structured TEI XML for computational analysis.

The Problem: Current digitization relies on manual entry or rigid rule-based heuristics which fail on historical documents.

The Context

BIBB maintains an archive of
German VET regulations

The Scope

Thousands of documents spanning
nearly a century (1938–2022)

The Goal

Transform scanned images into
structured TEI XML for
computational analysis

The Core Challenge

This project faces two interconnected layers of complexity that make traditional approaches insufficient.

Visual Complexity

Typography: Evolution from Fraktur (Blackletter) to modern Antiqua/Sans-serif

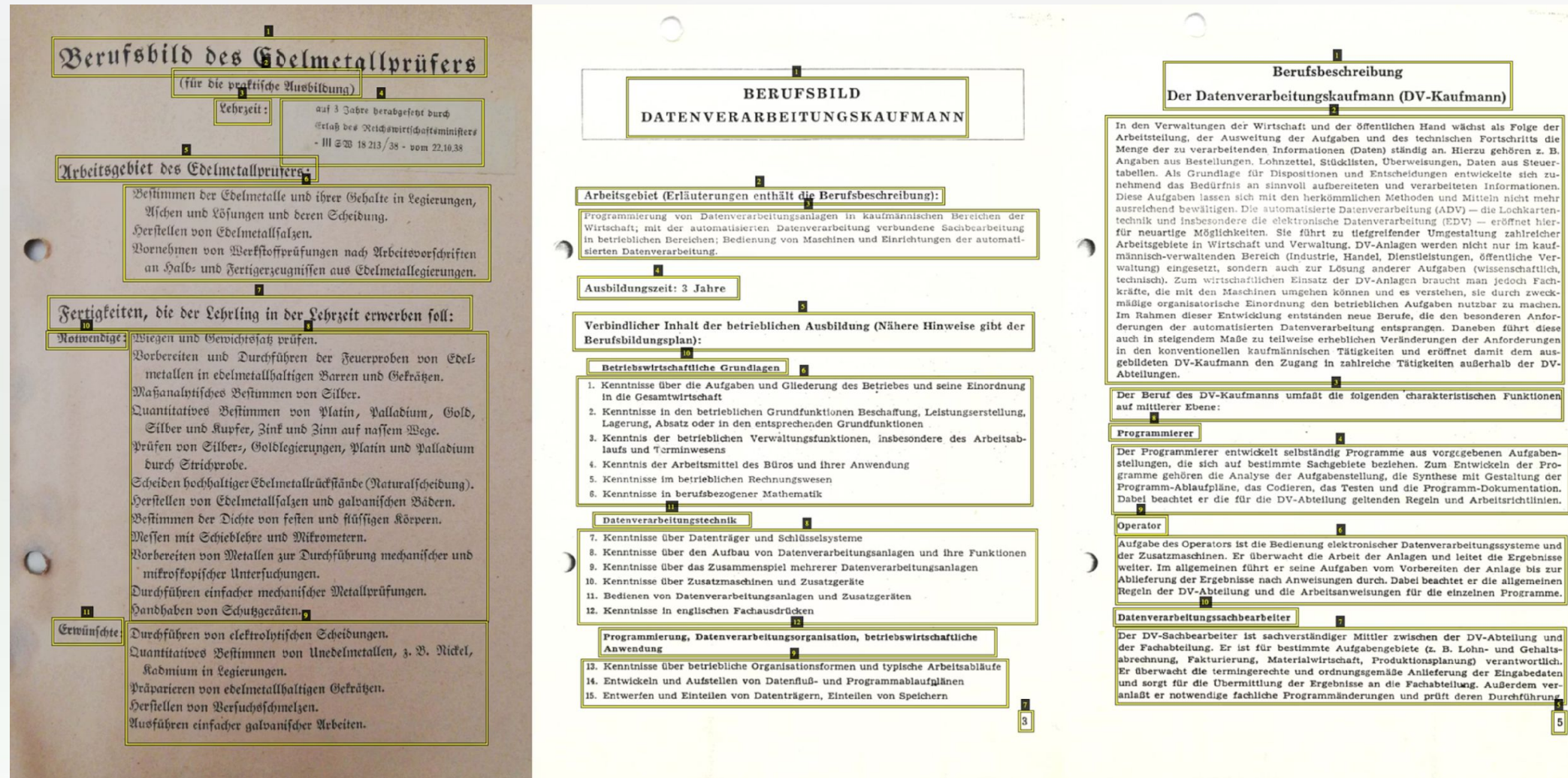
Layout Drift: 1930s dense text blocks vs. 1970s typewriter formatting vs. 2020s digital layouts

Degradation: Scan noise, aging paper, ink bleed

Structural Complexity

Need to distinguish: Main Headings vs. Section Headings vs. Enumerations vs. Paragraph Bodies.

The annotated pages



Problem Statement & Motivation

Limitations of Current Methods

Rule-Based Systems

Rigid. Fails when a document deviates from the template.

CNNs (Convolutional Neural Networks)

Good at local features, but struggle with long-range dependencies (e.g., linking a header to a list at the bottom of the page).

The Data Constraint

- Total Corpus: ~668 pages

Annotated Data: Only **87 pages** (approx. 13%) have pixel-level ground truth

The Research Gap: How do we train a sophisticated model on such limited data?

Research Questions

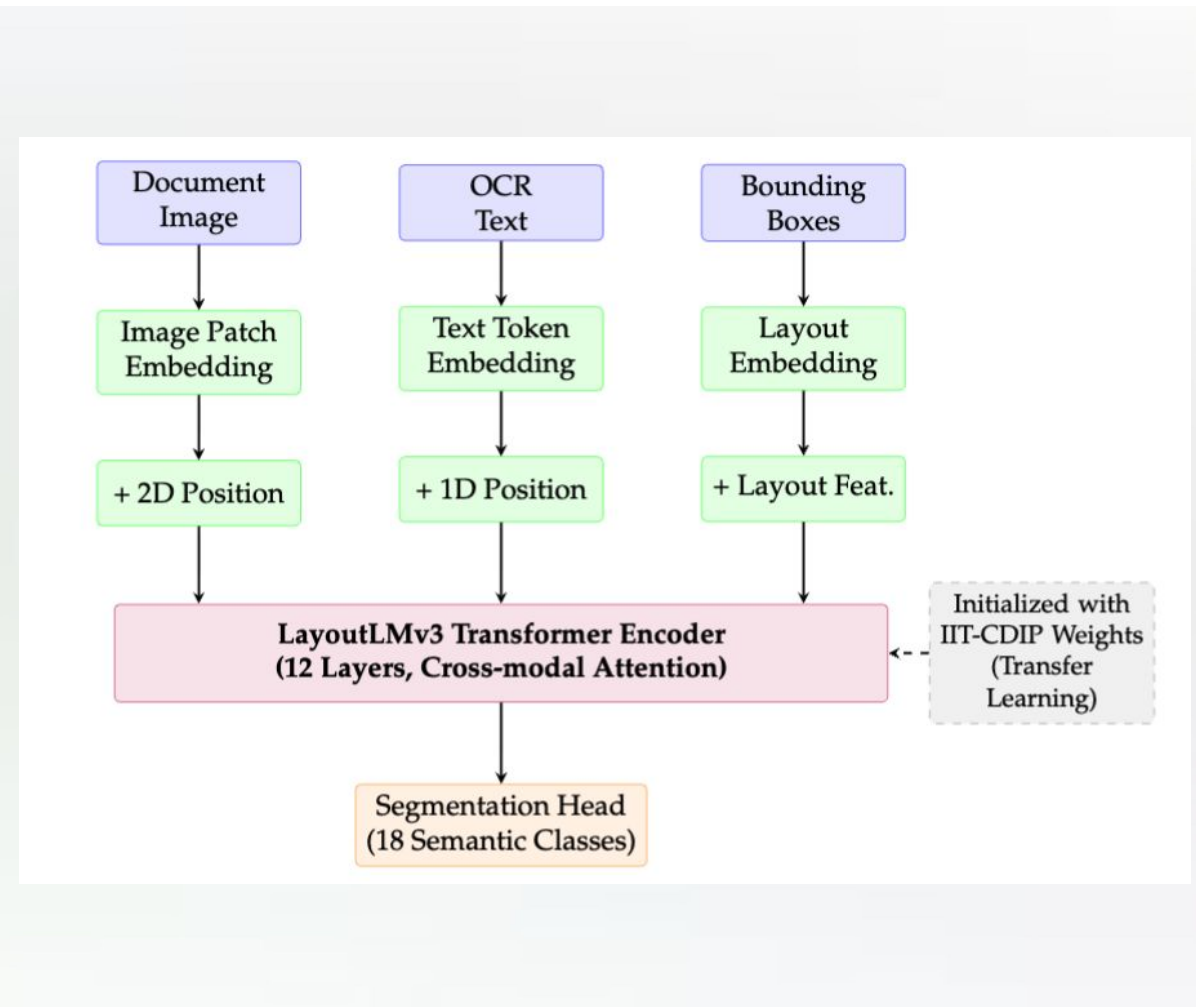
This thesis will answer three critical questions about handling historical document digitization:

- 1 Robustness**
How can models handle historical fonts (Fraktur) and diverse layouts with high OCR noise?
- 2 Low-Resource Learning**
How can we cope with limited annotated training data (87 pages) while learning useful patterns?
- 3 Comparative Performance**
Can machine learning approaches actually outperform the current rule-based systems used at BIBB?

Proposed Methodology:

LayoutLMv3

We fine-tune **LayoutLMv3** (pre-trained on IIT-CDIP) for **18-class semantic segmentation** of historical VET regulations.



Text

OCR tokens (noisy but informative)



Image

Typography + separators (font size, boldness, lines)



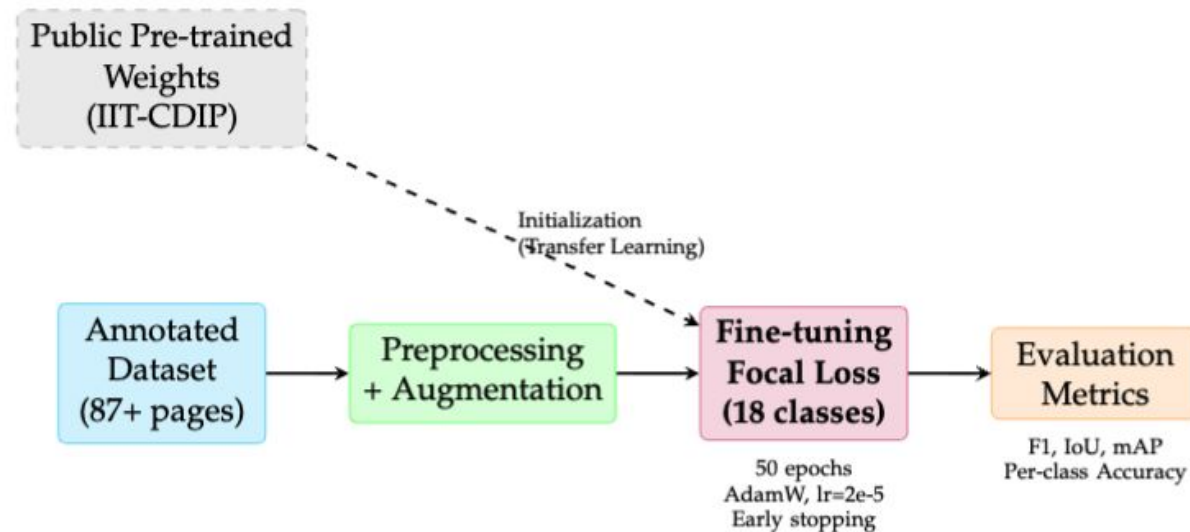
Layout

Spatial coordinates and bounding boxes

Key Advantage: Multimodal attention links text, visuals, and geometry, improving robustness to OCR errors and layout drift.

Single-Phase Transfer Learning Strategy

We fine-tune a publicly pre-trained LayoutLMv3 model on a small labeled subset, no additional self-supervised pre-training is performed.



Block 1: Initialization

Initialization (Pre-trained Weights)

Start from LayoutLMv3 pre-trained on IIT-CDIP

Motivation: reuse general document-structure knowledge (titles, headers, lists)

Block 2 (middle): Supervised Fine-tuning

Supervised Fine-tuning (87 → 120–150 pages if needed)

Task: 18-class semantic segmentation

Inputs: document image + OCR tokens + bounding boxes

Loss: Focal Loss (class imbalance)

Augmentation: mild rotation/crop, noise/contrast (scan variability)

Preliminary Results & Baseline Comparison

Initial Baseline (MFCN/DRFN)

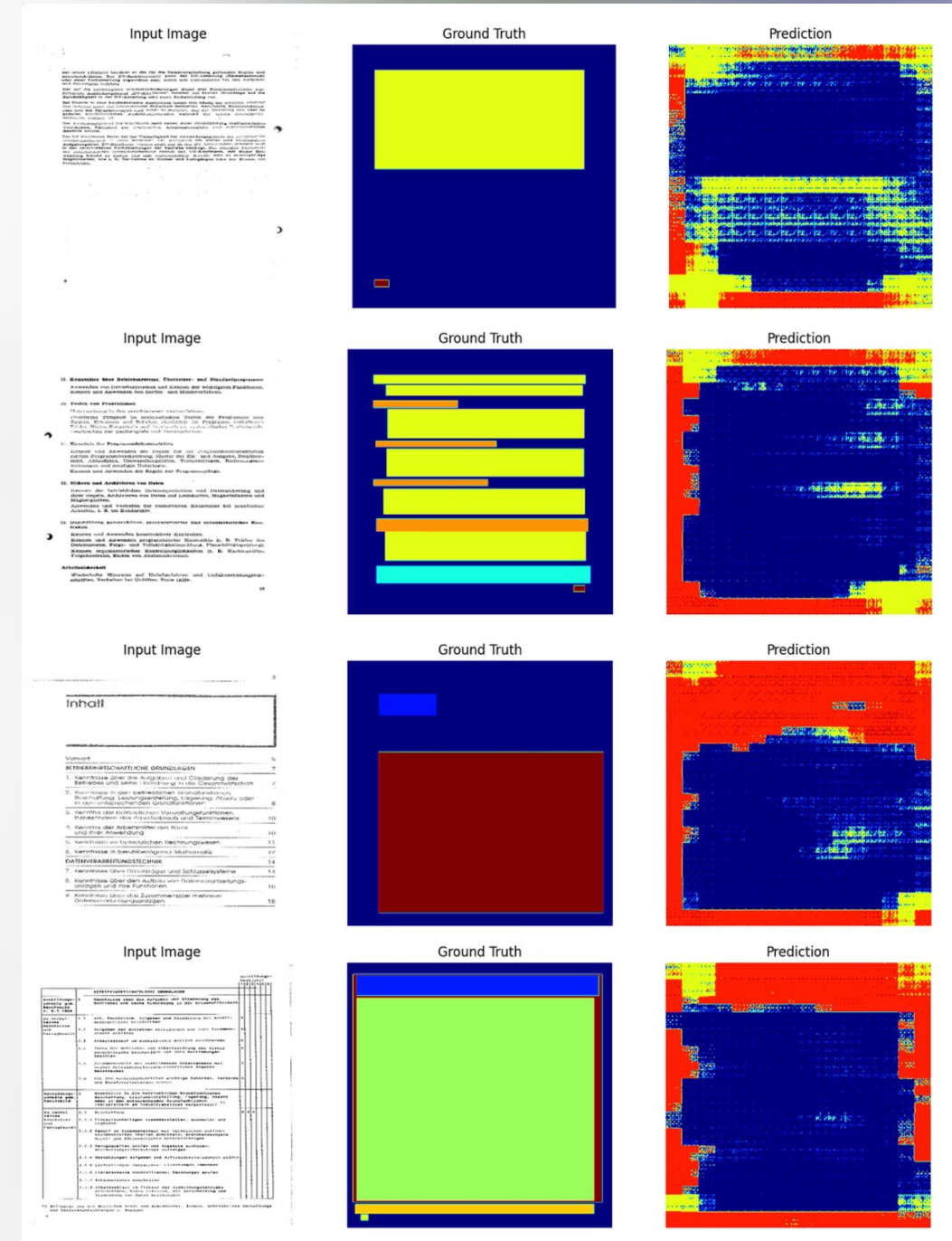
- Mean IoU: ~0.40

Failure Mode: Spatial

Fragmentation. The model broke single paragraphs into disconnected chunks.

Current Status

- Data preparation and OCR evaluation (Tesseract vs. PDF embedding) complete



Evaluation Plan

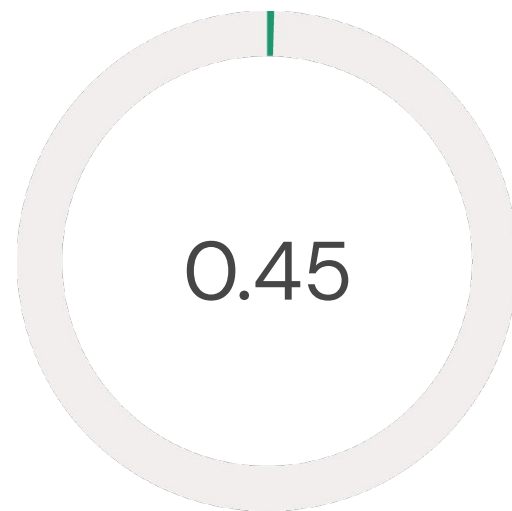
Success will be measured through rigorous quantitative and qualitative evaluation:

Metrics

Intersection over Union (IoU): Primary metric for segmentation accuracy

F1-Score: Critical for minority classes (e.g., Section Markers §)

Tiers of Success



Minimum Target
Beating the baseline



Target Goal
Practical utility for BIBB

Qualitative Analysis: Manual inspection of boundary precision and class confusion.

Project Timeline

The project is structured across six work packages over a six-month timeline:

01

WP 1: Data Collection

Completed

02

WP 2: Annotation Expansion

In Progress

03

WP 3: Model Setup (LayoutLMv3 Fine-tuning Pipeline)

In Progress

04

WP 4: Fine-tuning & Optimization

Upcoming

05

WP 5: Evaluation

Upcoming

06

WP 6: Thesis Writing

Upcoming

Project Timeline

The Gantt Chart showing the 6-month timeline

Expected Contributions

Methodological: Applying LayoutLMv3 to *historical* German legal texts (a low-resource, high-complexity domain).

Practical: A working pipeline for BIBB to assist in digitizing their 100-year archive.

Data: An expanded, annotated dataset for historical layout analysis.