Tramba: A Hybrid Architecture for Table Understanding

Md. Sayeed Abid

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master's (Computer Science) at
Concordia University
Montréal, Québec, Canada

20 July 2025

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Md. Sayeed Abid**

Entitled: **Tramba: A Hybrid Architecture for Table Understanding**

and submitted in partial fulfillment of the requirements for the degree of

Master's (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair
Dr. Tse-Hsun (Peter) Chen

_____ Examiner
Dr. Xinxin Zuo

_____ Examiner
Dr. Tse-Hsun (Peter) Chen

_____ Supervisor
Dr. Yang Wang

_____ Co-supervisor
Dr. Ching Yee Suen, PhD

Approved by _____
Dr. Denis Pankratov, Graduate Program Director

06 August 2025 _____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Tramba: A Hybrid Architecture for Table Understanding

Md. Sayeed Abid

The increasing complexity and density of document images—particularly in scientific and industrial contexts—have posed significant challenges for traditional transformer-based models, due to their quadratic attention complexity and reliance on extensive computational resources. In response, this thesis proposes a novel hybrid vision architecture that integrates the Vision Mamba encoder with the Detection Transformer (DETR) framework to address the tasks of table detection and structure recognition. Leveraging Mamba's state space modeling, which reduces computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, the proposed architecture retains competitive representational power while improving scalability and training efficiency. Vision Mamba is a state space sequence model designed for vision tasks, offering linear-time computation and efficient long-range dependency modeling through a bidirectional convolutional structure. DETR, in contrast, is an end-to-end object detection framework that formulates detection as a direct set prediction problem using a transformer-based encoder-decoder and learnable object queries. In our hybrid model, we replace DETR's standard transformer encoder with a Mamba-based encoder stack, preserving the core object query mechanism while enabling lightweight and efficient sequential processing. Through extensive experiments on the PubTables-1M dataset, which is one of the largest datasets for table extraction tasks, we demonstrate that our model outperforms Faster R-CNN on both detection and structure recognition tasks, and approaches the performance of full DETR models—despite using only one-third of the encoder-decoder layers and fewer training epochs. These results highlight the architecture's efficiency and adaptability, offering strong performance under constrained training budgets. Beyond empirical gains, the modular design of the model facilitates extensibility, including integration with large language models (LLMs) for advanced multimodal tasks such as document question answering, layout-based information retrieval, and regulatory content parsing. Finally, the lightweight nature of the Mamba encoder makes the model well-suited for deployment in enterprise-scale document processing systems, where throughput and latency are critical. This thesis thus introduces a promising direction for rethinking vision transformers through hardware-efficient sequence modeling, contributing meaningfully to the advancement of document AI and structured visual understanding. The source code is available at: github.com/SayeedAbid/Tramba

Statement of Originality

I hereby declare that I am the sole author of this thesis. All ideas and inventions attributed to others have been properly referenced. I understand that my thesis may be made electronically available to the public.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Yang Wang and my co-supervisor, Dr. Ching Yee Suen, PhD, for their exceptional guidance, patience, insights, and encouragement. Without their supervision and invaluable support, this thesis would not have been possible.

Special thanks to my thesis examiners, Dr. Xinxin Zuo and Dr. Tse-Hsun (Peter) Chen for their extremely valuable and constructive suggestions.

Lastly, I want to thank my family for their constant support and understanding during this period. Your love and encouragement have been my driving force.

Contents

List of Figures

List of Tables

List of Acronyms

CNN Convolutional Neural Network

ViT Vision Transformer

DEIT Data Efficient Image Transformer

SSMs State Space Models

DETR Detection Transformer

TD Table Detection

TSR Table Structure Recognition

TE Table Extraction

S4 Structured State Space Sequence model

SISR Single-Image Super-Resolution

BST Block-State Transformer

RPN Region Proposal Networks

NMS Non-Maximum Suppression

SSD Single Shot MultiBox Detector

YOLO You Only Look Once

ConvNets Convolutional Neural Networks

ResNet Residual Network

NLP Natural Language Processing

BERT Bidirectional Encoder Representations from Transformers

MoE Mixture-of-Experts

OCR Optical Character Recognition

LSSL Linear State-Space Layer

DSS Diagonal State Space

GIoU Generalized Intersection over Union

Chapter 1

Introduction

In this chapter, we present the motivation behind this research, formally define the problem addressed, review relevant literature, and summarize the key contributions of the thesis. Specifically, we focus on the task of extracting structured information from visually complex tables in documents, a problem that involves accurately detecting tables and recovering their underlying structure from raw visual data. The literature review section provides an overview of existing approaches to table detection and explores recent developments in hybrid modeling strategies for document understanding, particularly those that integrate state space models and transformer-based architectures.

1.1 Motivation and Background

1.1.1 Motivation

The advent of transformer-based models has led to significant breakthroughs in visual understanding tasks, including image classification, object detection, and table structure recognition. One such instance is the PubTables-1M framework, which leverages a transformer-based detection pipeline to extract and parse complex tabular structures from scientific documents accurately [80]. While transformer architectures such as ViT and DEIT have demonstrated exceptional performance, their computational inefficiency due to the quadratic complexity of self-attention severely limits their scalability, especially in tasks that require processing large image sequences or dense visual layouts.

Recently, State Space Models (SSMs) have emerged as a compelling alternative due to their ability to model long-range dependencies with linear computational complexity. Among them, Mamba, a hardware-aware, input-dependent variant of SSMs, has achieved performance comparable to transformers on sequential modeling tasks while significantly reducing memory and compute requirements [31]. However, the application of Mamba in computer vision is still nascent and faces unique

challenges due to the nature of visual data, which is spatial rather than purely sequential.

Motivated by the limitations of transformers and the potential of Mamba, this thesis investigates the replacement of transformer components in the PubTables-1M pipeline with Mamba-based modules. The goal is to evaluate whether Mamba can match or exceed transformer performance in table structure recognition while significantly improving computational efficiency. The inspiration to pursue this hybridization stems from recent work that integrates Mamba with vision backbones in innovative ways[31, 91, 12], which have shown promising results in bridging the performance-efficiency gap.

This work thus aims to push the frontier by answering a simple yet powerful question: *Can a purely transformer-based vision pipeline be re-architected with Mamba while preserving task performance and improving computational throughput?*

1.1.2 Background

End-to-End Object Detection with Transformers (DETR)

Traditional object detection systems rely heavily on complex pipelines, incorporating hand-engineered components like anchor generation, region proposal networks, and non-maximum suppression. DETR, introduced by Facebook AI, redefines this process by framing object detection as a direct set prediction problem. This novel approach eliminates the need for duplicate removal or anchor-based heuristics, offering a cleaner, fully end-to-end formulation [7]. At its core, DETR employs a transformer encoder-decoder architecture that globally models object interactions and contextual cues across an image. It processes CNN-extracted image features alongside a fixed number of learned object queries, producing all bounding box predictions simultaneously.

Unlike conventional methods that treat object detection as a classification-regression problem over a dense set of proposals, DETR leverages a bipartite matching loss to ensure one-to-one correspondence between predictions and ground truth objects. This enables the model to avoid redundant detections and post-processing steps. While DETR demonstrates strong performance on large objects, its reliance on global attention introduces inefficiencies in modeling fine-grained details of smaller objects. Furthermore, training DETR requires longer schedules and higher data augmentation due to its lack of inductive biases.

Despite these challenges, DETR achieves competitive results with established detectors like Faster R-CNN and shows extensibility to tasks such as panoptic segmentation—underscoring its potential as a general-purpose detection framework.

Deep Learning for Table Detection and Structure Recognition

Table detection (TD) and table structure recognition (TSR) are foundational components of table extraction, aiming to localize tabular regions and interpret their internal grid-like structure, comprising rows, columns, and cells. Early approaches relied heavily on handcrafted rules and visual heuristics, which often failed to generalize across documents with diverse layouts and complex formatting.

The emergence of deep learning, particularly transformer-based object detection frameworks, revolutionized this field by enabling more robust and scalable modeling. The PubTables-1M dataset exemplifies this progress, offering a large-scale, richly annotated corpus of one million tables from scientific literature, supporting all three subtasks of TD, TSR, and functional analysis. The use of DETR-style transformer architectures in this work unified these subtasks under a shared encoder-decoder framework, producing state-of-the-art results without the need for handcrafted features. However, the deployment of such transformer-based models remains computationally intensive due to their quadratic attention complexity, which becomes a critical bottleneck in real-time or large-scale document processing scenarios [80].

Mamba and the Emergence of Linear-Time Sequence Models

Recent advances in State Space Models (SSMs) have introduced new directions for efficient sequence modeling. One such innovation is Mamba, which builds on the structured state space sequence model (S4) and employs a selective scan mechanism to process sequences in linear time. Unlike transformers that scale poorly with input length due to their attention mechanism, Mamba is designed to retain long-range contextual understanding while significantly reducing the computational footprint [12].

Originally tailored for natural language processing, Mamba's autoregressive and input-dependent design proves well-suited for sequential tasks. However, this same sequential bias poses challenges when applied to vision tasks, where data is inherently spatial and two-dimensional. Early efforts such as VMamba, Vision-Mamba, and EfficientVMamba attempted to adapt Mamba for vision applications, but struggled with global context modeling and spatial feature extraction limitations. These issues highlight the need for further architectural innovation to fully harness Mamba's efficiency in vision domains [31] [54].

Toward Hybrid Architectures: Combining Mamba with Transformers

To overcome the respective shortcomings of pure transformer and pure Mamba architectures, several recent studies have proposed hybrid models that strategically integrate both paradigms.

These models aim to combine the high-throughput, linear scalability of Mamba with the powerful global modeling capabilities of self-attention.

MambaVision is a pioneering hybrid vision backbone that incorporates transformer blocks in the final stages of a Mamba-based network. This selective integration enables the model to recover long-range spatial dependencies lost in purely sequential Mamba processing, while preserving superior throughput and accuracy trade-offs. The model achieves strong performance across multiple vision benchmarks, demonstrating that Mamba and transformers can be jointly optimized for visual perception tasks [31].

Similarly, **TranMamba** introduces a lightweight architecture for single-image super-resolution (SISR) by alternating between Mamba and transformer blocks. This architectural interleaving balances local feature extraction with global context modeling, leading to state-of-the-art results with fewer parameters and reduced computational costs [91].

Building on this momentum, **TransMamba** tackles the challenge of architectural transition by enabling cross-architecture knowledge distillation—transferring representational knowledge from pretrained transformer models to Mamba-based networks. Using feature alignment and adaptive bidirectional distillation, TransMamba accelerates Mamba training while enhancing performance across vision and multimodal tasks such as image classification, VQA, and text-video retrieval [12].

Synthesis and Research Opportunity

Collectively, these developments signal a convergence in architectural design: transformer models offer expressive power but are computationally expensive, while Mamba introduces a pathway toward efficient yet capable alternatives. Hybrid designs, as explored in MambaVision, TranMamba, and TransMamba, illustrate that these paradigms are not mutually exclusive but complementary.

Yet, despite these promising directions, no work has yet explored replacing the full transformer encoder-decoder stack in a DETR-based structured vision model such as PubTables-1M with a Mamba-driven counterpart.

This gap presents a unique research opportunity. In this thesis, we aim to develop and evaluate a novel hybrid architecture that substitutes the transformer blocks in DETR with a Mamba-based encoder-decoder framework, addressing the limitations of attention-based models in terms of efficiency, while assessing the viability of Mamba in complex structured vision tasks.

1.2 Research Gap

Despite the theoretical and empirical promise of Mamba, very few have attempted a full-scale replacement of transformers in a complex vision pipeline like table structure recognition. While hybrid

models exist, they often retain transformer layers due to concerns over Mamba's spatial modeling limitations. This thesis takes a bolder step by modifying the DETR pipeline and integrating Mamba as the primary sequential processor, replacing attention-based components.

In doing so, this research addresses the following gaps:

- **Scalability of Mamba in Document Understanding:** Can Mamba handle spatially complex layouts like tables?
- **Architectural Efficiency:** What are the trade-offs in computational performance versus accuracy?
- **Transferability of Learning:** Can pre-trained transformer models guide Mamba-based models through knowledge distillation?

1.3 Overview and Thesis Contribution

This section presents an overview of each chapter and the entire thesis:

1.3.1 Background and Literature Review

This chapter provides foundational context and a comprehensive background necessary for understanding the advancements discussed throughout this report. The primary aim of this chapter is to offer an in-depth overview of existing literature and key methodologies relevant to the subsequent development and analysis presented in later chapters. It systematically explores crucial aspects such as object detection, vision backbone architectures, state space models (SSMs) tailored specifically for visual applications, and the innovative approaches adopted towards hybrid architectures. In the Literature Review section, the chapter begins with an extensive analysis of object detection methodologies. It provides insights into classical and modern approaches, outlining significant contributions and advancements, such as the role of convolutional neural networks (CNNs) and Transformer-based models, and highlighting state-of-the-art techniques and their applications across various visual tasks. Subsequently, the Architecture for Vision Backbone section delves deeply into the frameworks that have shaped contemporary vision models. This includes a discussion on the evolution from traditional CNN architectures, through the advent and dominance of Vision Transformers, to the recent innovations in SSM-based vision backbones. Each architectural paradigm is assessed for its strengths, limitations, and specific application scenarios. In addressing the State Space Model for Visual Application, the chapter highlights recent breakthroughs in adapting traditional SSM techniques to complex visual processing tasks. Key studies such as the Structured State Space (S4) model and its extensions to multi-dimensional visual data are thoroughly reviewed.

This section emphasizes how these models effectively capture temporal and spatial dependencies, addressing previously challenging computational constraints and improving performance across diverse visual tasks. Further, the Modeling Approach section provides a detailed examination of methodologies employed to integrate these diverse architectures into cohesive, high-performing systems. It highlights critical developments in model fusion techniques, attention mechanisms, and strategies for efficiently combining features from different sources. The focus is placed on identifying and explaining methodological innovations crucial for advancing model performance and scalability. Finally, the chapter explores Approaches towards Hybrid Architecture, underscoring recent and innovative strategies that leverage the complementary advantages of different model families, such as Transformers and SSMs. Various hybrid models, including the Jamba and Block-State Transformer (BST), are discussed, demonstrating their ability to enhance model performance by integrating short-range contextual modeling with long-range dependencies. Overall, this chapter sets the stage for the detailed methodologies and experimental evaluations presented in subsequent chapters, ensuring that readers are well-equipped with the theoretical and practical context needed for understanding the contributions and implications of this report.

1.3.2 Tramba: A Hybrid Architecture for Table Understanding

This chapter introduces Tramba, our proposed hybrid architecture, and thoroughly discusses its methodology, detailed algorithmic structure, and technical intricacies. This chapter provides a clear exposition of the design principles underlying Tramba, highlighting how it combines Transformer and Mamba-based architectures into a cohesive model. The methodology section outlines the specific innovations introduced in Tramba, such as the incorporation of Mamba-based layers for capturing extensive temporal dependencies and Transformer layers for precise local feature extraction. The algorithmic framework section meticulously describes Tramba's model architecture, elucidating each layer's functionalities, interactions, and integration within the broader network. Technical details include the architectural composition of hybrid layers, gating mechanisms, optimization strategies, and parameter configurations. Additionally, this chapter provides comprehensive equations and pseudocode, clearly defining input-output transformations at each computational step, facilitating deeper technical understanding. Moreover, the chapter addresses implementation specifics, including training procedures, loss functions, hyperparameter tuning, and computational efficiency considerations. The detailed technical exposition ensures clarity regarding Tramba's innovative approach to hybrid modeling, serving as a cornerstone for evaluating its performance in subsequent experimental analyses and comparative studies outlined in later chapters.

1.3.3 Experiments

This chapter presents an extensive experimental evaluation of the proposed Tramba architecture. It outlines a detailed description of experimental setups, datasets employed, and evaluation metrics utilized to assess the performance of the hybrid model comprehensively. This chapter systematically reports experimental results, providing quantitative and qualitative analyses and visualizations to elucidate the performance gains achieved by Tramba. Comparative studies are thoroughly presented, benchmarking Tramba against established models and hybrid architectures, clearly illustrating its strengths and areas of improvement. These evaluations provide robust evidence for the effectiveness of Tramba, offering insights into its scalability, accuracy, and computational efficiency.

1.3.4 Conclusion and Future Work

This chapter concludes the report by summarizing the primary findings and contributions made throughout the study. It highlights the significance of the proposed Tramba architecture and its potential implications in advancing visual and multimodal modeling. Additionally, the chapter identifies and discusses limitations encountered during the research and outlines concrete avenues for future exploration. Prospects for subsequent research, focusing on further optimization, broader applications, and innovative extensions of the Tramba model, are clearly delineated, providing direction for continued investigation and advancement in hybrid modeling frameworks.

1.4 Thesis Organization

The rest of this thesis is organized as follows: Chapter 2 discusses the literature review and background studies of this thesis. Chapter 3 presents our hybrid architecture and it's overall methodology Chapter 4 describes the results of our architecture. Chapter 5 concludes the thesis and discusses the potential research doors for future work.

Chapter 2

Background and Literature Review

In this chapter, we begin by introducing the fundamental concepts underlying object detection, including vision backbone architectures and the application of state space models in visual tasks. We then outline the modeling strategies employed in hybrid architectures that integrate these components. Furthermore, a comprehensive review of existing literature is provided to contextualize the proposed work, identify prevailing methodologies, and highlight the limitations and research gaps that this thesis seeks to address.

2.1 Literature Review

2.1.1 Object Detection

Object detection has undergone significant evolution over the past decade, transitioning from handcrafted feature methods to deep learning architectures that achieve remarkable accuracy. Modern approaches typically predict bounding boxes and class labels for objects within images, with two-stage detectors like Faster R-CNN [74] pioneering region proposal networks (RPNs) that generate candidate regions before refinement [74, 6]. These methods rely heavily on handcrafted components, including anchor generation and non-maximum suppression (NMS) to manage duplicate predictions and encode spatial priors.

In contrast, single-stage detectors like SSD (Single Shot MultiBox Detector) [53] and the YOLO (You Only Look Once) series [72] predict object locations and classes directly from dense grids or anchor points. YOLOv3 [71] introduced multi-scale predictions using anchor boxes, while YOLOv4 further optimized backbone architectures and training strategies. These approaches streamline detection but still depend on heuristics for matching predictions to ground truth and post-processing. CenterNet [96] later reimagined detection by treating objects as points in a heatmap, enabling

efficient keypoint-based localization.

Recent research has simplified pipelines by reframing detection as direct set prediction, eliminating many hand-designed components. The DETR (DEtection TRansformer) model [7] exemplifies this shift with its transformer-based encoder-decoder architecture that predicts a fixed set of detections in parallel. Using bipartite matching loss, DETR enforces unique prediction-to-ground-truth assignments. This approach removes the necessity for NMS and anchor generation, as the model learns to output non-redundant, globally consistent predictions. Deformable DETR [97] enhanced this approach with deformable attention mechanisms, improving convergence speed and small-object detection.

Multi-stage refinement architectures like Cascade R-CNN [6] address quality limitations through progressive detection refinement across sequential stages. This approach demonstrates how complex pipelines can improve high-precision detection despite increased computational demands.

The evolution highlights three key trends:

- **Reduced handcrafting through end-to-end learning** [7, 97]
- **Efficiency optimizations in single-stage models** [72, 53]
- **Specialized architectures for quality or speed tradeoffs** [6, 95]

In summary, the evolution of object detection has moved from complex, hand-engineered pipelines toward more streamlined, end-to-end models that leverage advances in deep learning architectures. While transformer-based models capture global context effectively through self-attention, challenges remain in training efficiency and small-object detection. Nevertheless, the trajectory continues toward unified architectures that minimize inductive biases while maintaining competitive performance across benchmarks like COCO.

2.1.2 Architecture for Vision Backbone

The development of generic vision backbones has undergone significant paradigm shifts, transitioning from convolutional networks to transformer-based models and exploring new state-space approaches. Early computer vision systems relied on convolutional neural networks (ConvNets) pioneered by [42], which became the de facto standard through architectures like AlexNet [41], VGGNet [79], and ResNet [32]. These models leveraged spatial inductive biases for efficient feature extraction, with later innovations like EfficientNet [83] optimizing accuracy-efficiency trade-offs and RegNet [69] systematically scaling model dimensions. The landscape transformed with Vision Transformers (ViT) [22], which treated images as sequences of patch tokens and applied pure transformer architectures. ViT demonstrated that global self-attention could outperform ConvNets at

scale, particularly with large datasets. Some approaches enhance ViT architectures by embedding 2D convolutional priors to inject local inductive biases [87, 21]. This sparked hybrid approaches like CvT [88] and ConViT [24], which integrated convolutional priors into transformers. Concurrently, pyramid structures like PVT [86] enabled multi-scale feature extraction, while Swin Transformer [55] introduced shifted-window attention for hierarchical representation. In response, modern ConvNet revitalization emerged through works like ConvNeXt [56], which reinterpreted ResNet with transformer-inspired training techniques, achieving ViT-level performance. RepLKNet [20] further demonstrated that scaling kernel sizes to 31×31 could enhance long-range modeling in pure ConvNets. Although numerous subsequent works have achieved strong performance and improved efficiency on benchmarks like ImageNet [17] and various downstream vision tasks [50, 94] by incorporating 2D inductive priors into vision transformers, the resurgence of vanilla transformer architectures has become increasingly evident. This shift is largely driven by the rapid growth of large-scale visual pretraining techniques [64, 25, 8] and the widespread adoption of multi-modal learning frameworks [68, 47, 52, 37]. These standard transformer models are once again gaining prominence due to their expansive capacity, compatibility with unified multi-modal representations, and strong alignment with self-supervised learning objectives. However, a significant challenge remains: the quadratic complexity of the self-attention mechanism imposes practical limits on the number of visual tokens, thereby restricting model scalability. While numerous studies have proposed solutions to mitigate this bottleneck [15, 40, 14, 19, 66], the majority focus on general-purpose or language-specific settings. Only a limited subset of this research directly targets vision-specific adaptations for efficient attention computation.

However, the rise of large-scale pretraining [64, 26] and multimodal applications [67, 37] renewed focus on transformer-based models due to their unified representation capabilities. A critical limitation persisted: the quadratic complexity of self-attention constrained sequence length. Efficient transformers addressed this via:

- Low-rank approximations
- Sparse attention
- Dilated mechanisms

Recently, state-space models (SSMs) like Mamba [28] offered linear-time sequence modeling, inspiring pure-SSM vision backbones. These architectures promise to retain ViT's modality-agnostic benefits while enabling extreme-sequence processing—critical for high-resolution medical imaging [85] and video understanding. Current research explores balancing architectural biases: ConvNets offer spatial efficiency, transformers excel at global context, and SSMs enable long-sequence modeling. The field increasingly favors task-adaptive backbones, with quantization (P2-ViT) and pruning

(XGB-based Pruner) enabling deployment across edge devices and data types (RGB, depth, multi-modal). This evolution highlights that no single architecture dominates universally; instead, optimal backbones emerge from synergistic integration of complementary paradigms.

2.1.3 State Space Model for Visual Application

State space models (SSMs) have recently gained significant advancement in visual tasks due to their capability to efficiently model long-range dependencies, an area traditionally dominated by Transformers and convolutional neural networks (CNNs). Initially, Islam et al. [35] utilized a 1D Structured State Space (S4) model to manage temporal dependencies in video classification tasks effectively, highlighting the suitability of SSMs in sequence modeling. Building upon this foundation, [60] expanded the capabilities of 1D S4 to encompass multi-dimensional data structures, including both 2D images and 3D videos, showcasing a broader applicability and enhanced modeling capability in more complex visual scenarios. Further innovations emerged from the hybridization of SSMs with other established methodologies. Islam et al proposed TranS4mer [36], a sophisticated hybrid model that integrates the structural advantages of S4 with self-attention mechanisms, setting new benchmarks in tasks such as movie scene detection. This hybrid approach combines the attention-driven representation of Transformers with the long-range context modeling of SSMs, yielding superior performance. Concurrently, Li et al developed a novel selectivity mechanism within the S4 framework designed specifically for long-form video understanding tasks [47]. This enhancement significantly reduced the computational footprint of traditional S4, addressing the challenge of modeling extensive video data with lower memory usage. Another important step was introduced by Yan et al, who replaced traditional attention-based backbones with an SSM-centric architecture [89]. Their model demonstrated strong capabilities in generating high-resolution imagery and extracting fine-grained features, particularly beneficial for resource-limited environments requiring scalable computation. Expanding these into biomedical domains, Ma et al introduced U-Mamba [58], an innovative hybrid architecture combining CNNs with SSM components to handle intricate dependencies in biomedical image segmentation tasks. This hybrid model notably improved segmentation accuracy by effectively capturing long-range spatial contexts. These diverse yet interrelated works illustrate either specialized applications of SSM or hybrid models integrating SSM with convolutional or attention mechanisms. Liu et al introduced VMamba [54], leveraging multi-directional scanning and hierarchical structures within Mamba for visual recognition tasks. And Zhang et al [90], distinctively emphasizes visual sequence modeling, providing a cohesive and versatile representation framework suitable for multimodal data processing.

2.1.4 Modeling Approach

The hybridization of Transformer and State Space Models (SSM), particularly the Mamba architecture, has recently emerged as an influential direction in sequence modeling research, promising significant advancements in both natural language processing (NLP) and multimodal applications. The Transformer model, initially proposed by Vaswani et al [84], revolutionized NLP by introducing a purely attention-based mechanism capable of modeling dependencies irrespective of their distance within sequences. This self-attention mechanism enabled Transformers to establish new state-of-the-art results across numerous NLP benchmarks such as machine translation [16, 62], text summarization [43], and sentiment analysis [18]. Building upon the fundamental Transformer architecture, numerous extensions have proliferated, demonstrating diverse strategies to enhance context modeling and computational efficiency. Notably, BERT (Bidirectional Encoder Representations from Transformers) leveraged bidirectional pre-training techniques utilizing masked language modeling and next sentence prediction tasks, significantly enhancing contextual understanding and performance on multiple NLP tasks [18]. The subsequent Generative Pre-trained Transformer (GPT) series, including GPT-2 [67] and GPT-3 [5], expanded the scope by excelling in text generation tasks, employing autoregressive training methods and large-scale datasets. Concurrently, research into multimodal Transformers such as Vision Transformer (ViT) [24] and multimodal Transformers [68, 44] demonstrated the adaptability of the Transformer architecture to visual and multimodal tasks, achieving significant breakthroughs in image classification and cross-modal representation learning. Despite these advancements, Transformers encounter computational bottlenecks due to their quadratic complexity concerning sequence length, prompting the exploration of alternative architectures. State Space Models, historically utilized for modeling temporal dynamics and control systems [38, 23], have re-emerged as promising candidates due to their capability to efficiently capture long-range temporal dependencies with linear complexity. The Structured State Space (S4) model proposed by [30] significantly revitalized interest in state space approaches by presenting a structured parameterization of continuous-time state spaces. This approach exhibited impressive results on sequence modeling benchmarks by efficiently handling long-range dependencies without the computational burden associated with traditional Transformer architectures. Further enhancing the efficacy of state space modeling, the Mamba architecture [28] emerged as an advanced variant building directly upon S4's structured foundations. Mamba incorporates dynamic system principles within a neural network context, utilizing advanced gating mechanisms and specialized parameterizations that allow superior modeling of temporal sequences with notable computational efficiency. Recent empirical evaluations of Mamba across NLP and vision tasks illustrate its capacity to outperform traditional Transformer architectures on long-range dependency tasks, especially those requiring extensive context modeling capabilities [28, 54]. Inspired by these complementary

strengths, a growing body of research has begun investigating hybrid models combining Transformers and Mamba-like SSM architectures. Notably, Jamba [49] interleaves Transformer and Mamba layers to leverage the local attentiveness of Transformers and the global, efficient context modeling of Mamba architectures. By introducing mixture-of-experts (MoE) layers, Jamba further enhances its modeling capacity without significantly inflating computational requirements. This combined modeling approach demonstrated enhanced performance on diverse NLP benchmarks, validating the efficacy of leveraging complementary architectural strengths. Extending this paradigm, recent models such as Block-State Transformer [34] have refined this concept by explicitly separating short-term and long-range contextualization within dedicated sublayers. BST employs block-based Transformer sublayers to capture local short-range dependencies, coupled with dedicated SSM sublayers specifically designed to encode long-range sequential information efficiently. Such explicit separation has proven highly effective, achieving significant performance improvements on sequence modeling and multimodal tasks, indicating a potent area of exploration for future hybrid architectures. Integral to the success of these hybrid architectures is the effective fusion of features from disparate layers and modules. Feature fusion techniques initially rooted in attention mechanisms [84, 16] have evolved significantly, particularly within multimodal learning contexts. Recent advancements propose sophisticated integration methods combining convolutional neural networks (CNNs) with Transformers for enhanced multimodal representation learning. Cross-attention mechanisms within Transformer decoders have become pivotal for effectively merging visual and linguistic features, as evidenced in successful multimodal architectures such as CLIP [68] and ViLT [39]. Given these advancements, current research increasingly focuses on designing architectures that optimally combine Transformer and state space modeling, specifically through Mamba-based frameworks, aiming to harness both local and global sequence processing efficiencies. Hybrid models present a compelling direction for further research, offering significant potential to address the computational and representational limitations encountered by pure Transformer architectures. The development of these hybrid approaches thus underscores a broader trend towards sophisticated, adaptable models capable of integrating and exploiting diverse computational paradigms effectively.

2.2 Approaches towards Hybrid Architecture

The evolution of deep learning in computer vision has been characterized by a constant tension between performance and efficiency. Vision Transformers (ViTs), with their powerful self-attention mechanisms, have achieved state-of-the-art (SOTA) results across diverse visual tasks by modeling global dependencies effectively. However, this performance often comes at the cost of quadratic computational complexity with respect to input size, making ViTs less practical for high-resolution

imagery or resource-constrained environments [31, 55].

Among SSM-based models, Mamba has emerged as a highly efficient alternative, capable of linear-time processing while capturing long-range dependencies. Originally introduced for sequence modeling in NLP, Mamba has since been adapted for vision tasks due to its hardware-friendly design and ability to model temporal or spatial dependencies with selective scanning mechanisms [28]. However, pure Mamba-based vision backbones still face challenges when it comes to preserving spatial richness and global receptive fields, which are crucial for vision tasks such as object detection or semantic segmentation [31].

To bridge this gap, a new class of hybrid architectures has been proposed—most notably *MambaVision* [31] and *Contrast* [11]—which integrate SSM modules like Mamba in early or intermediate layers, and self-attention-based Transformer blocks in deeper layers. This architectural strategy offers the best of both worlds: the linear efficiency and long-range modeling of Mamba, coupled with the rich spatial understanding and global context capture of Transformers. In MambaVision, for instance, CNN-based layers are used in the initial stages for high-resolution feature extraction, followed by a hybrid stack of Mamba mixers and self-attention layers. The inclusion of attention layers at later stages significantly improves global modeling capabilities while maintaining a new Pareto frontier in accuracy vs. throughput trade-offs on ImageNet-1K [31].

Empirical results strongly validate this architectural direction. MambaVision variants outperform competitive baselines like Swin Transformers and ConvNeXt models not only in classification but also in downstream tasks such as object detection on MS COCO and semantic segmentation on ADE20K. For instance, MambaVision-B achieves higher top-1 accuracy with lower GFLOPs and higher throughput compared to VMamba and FastViT variants [31]. Similarly, *Contrast*, a tri-modal hybrid model combining CNN, Transformer, and Mamba blocks, shows strong gains in super-resolution and dense prediction tasks [11].

Notably, this hybrid approach has shown particular strength in robotic vision tasks where long-term temporal understanding and real-time inference are essential. Hybrid Mamba-Transformer backbones deployed in robotic manipulation pipelines, such as grasp detection and scene parsing, demonstrate increased robustness and spatial precision, outperforming standalone Transformer or SSM architectures [82].

From a theoretical perspective, this integration is grounded in how Mamba's implicit recurrence models sequential patterns efficiently, while self-attention explicitly captures all pairwise token interactions. In tasks where localized details (e.g., object boundaries) and global coherence (e.g., table layout) are both vital, neither approach is sufficient alone. The hybrid paradigm addresses this by decoupling early-stage low-level extraction from late-stage semantic fusion.

Key advantages of these hybrid architectures can be summarized as follows:

- **Computational Efficiency:** Mamba's linear-time complexity provides efficient handling of long sequences or high-resolution images, improving over the quadratic scaling of attention [28, 31].
- **Flexible and Rich Feature Representation:** The synergy of sequential modeling (via SSM) and token-level attention enables more expressive and generalizable feature embeddings [11, 82].
- **Cross-Domain Versatility:** These models perform competitively in classification, segmentation, restoration, and robotic control, underscoring their adaptability [31, 82].

In conclusion, the convergence of State Space Models like Mamba and Transformer-based architectures in a hybrid framework represents a pivotal advancement in vision backbone design. This hybridization not only resolves the trade-offs between performance and efficiency but also enables broader applicability across real-world vision tasks. Future research will likely explore further automation in layer-wise composition, architecture search, and broader applications in multimodal settings.

Chapter 3

Tramba: A Hybrid Architecture for Table Understanding

Table detection and table structure recognition are critical tasks in structured document analysis, which enable accurate parsing of complex layouts. Despite the visual clarity of tables to humans, extracting tabular regions and understanding their structural organization—such as rows, columns, and cell relationships—remains a challenging problem for machine learning models. This challenge becomes especially significant in real-world documents where formatting inconsistencies, spanning cells, and missing grid lines are common, yet reliable table understanding is crucial for tasks like data mining, automated form processing, and digital archiving.

Traditional transformer-based architectures have shown strong performance but often struggle with computational efficiency and long-range dependencies. In this work, we propose a hybrid architecture that combines Vision Mamba, a state-space model (SSM) designed for long-range sequence modeling, with a transformer-based approach to enhance table detection and structure recognition. By leveraging the efficiency of Vision Mamba to capture global dependencies and the localized representation power of transformers, our method achieves superior accuracy while maintaining computational efficiency. We evaluate our approach on one of the largest table detection datasets and demonstrate that it outperforms previous state-of-the-art models, including the baseline transformer-based model DETR, achieving higher accuracy in both detection and recognition tasks. Our results highlight the effectiveness of integrating SSMs with transformers for structured document analysis, paving the way for more efficient and scalable solutions in document AI.

3.1 Introduction

Tables are a fundamental modality for conveying structured data across a broad spectrum of document formats, including scientific literature, medical reports, financial statements, and web-based content. Their two-dimensional grid layout facilitates compact and interpretable representations of relational information, making them indispensable in automated document understanding workflows. However, the presence of a visual layout alone does not guarantee machine-readable structure. In many real-world cases, tables are visually well-formed but lack explicit logical annotations such as row-column hierarchies, spanning semantics, or cell-type roles. This disconnect between visual layout and underlying logical structure introduces significant ambiguity for downstream parsing systems.

Figure 1 illustrates such an example, where the table's appearance is visually coherent—clearly demarcating rows, columns, and text regions—but lacks any semantic markup to differentiate between headers, data cells, or complex spanning relationships. Without these logical cues, even state-of-the-art models face difficulty in accurately reconstructing the structure necessary for semantic interpretation.

This challenge is particularly pronounced in large-scale document-related tasks where tables appear in diverse formats with inconsistent layouts, missing grid lines, multi-row headers, and non-rectangular cell groupings. The inability to robustly detect and structurally parse such tables renders much of the embedded information inaccessible to automated pipelines. Consequently, document AI systems risk overlooking critical information or introducing structural noise during extraction.

The core problem this work addresses is the joint task of table detection—localizing tabular regions within complex document images—and table structure recognition, which involves identifying the fine-grained structural relationships between constituent elements (rows, columns, and cells) within each detected table. As noted in the PubTables-1M benchmark dataset [80], accurate structure recognition is a non-trivial and underexplored challenge, particularly when table boundaries are noisy or when visual cues are insufficient to resolve logical relationships.

By focusing on visually structured but semantically unannotated tables, this research targets a real and impactful gap in current document understanding systems—bridging the divide between appearance and structure in complex visual documents.

The process of interpreting and restructuring these visually presented tables into machine-readable formats is known as table extraction (TE). TE generally involves three main subtasks: table detection, structural recognition, and functional analysis [27]. Each of these tasks presents its own challenges, stemming from the wide variability in table design—differences in alignment, spanning cells, headers, styles, and layout configurations [33, 61, 74, 92]. Furthermore, complex visual ambiguities such as misaligned gridlines or inconsistent fonts add additional complexity, making rule-based heuristics unreliable in general scenarios.

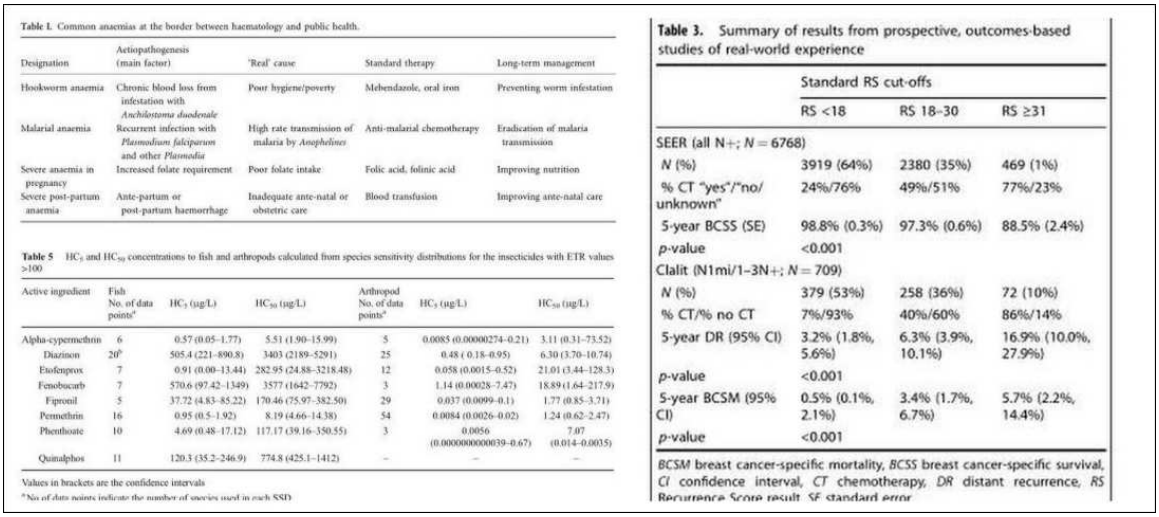

Table I. Common anaemias at the border between haematology and public health.

Designation	Aetiopathogenesis (main factor)	'Real' cause	Standard therapy	Long-term management
Hookworm anaemia	Chronic blood loss from infestation with *Ancylostoma duodenale*	Poor hygiene/poverty	Mebendazole, oral iron	Preventing worm infestation
Malarial anaemia	Recurrent infection with *Plasmodium falciparum* and other *Plasmodia*	High rate transmission of malaria by *Anopheles*	Anti-malarial chemotherapy	Eradication of malaria transmission
Severe anaemia in pregnancy	Increased folate requirement	Poor folate intake	Folic acid, folinic acid	Improving nutrition
Severe post-partum anaemia	Ante-partum or post-partum haemorrhage	Inadequate ante-natal or obstetric care	Blood transfusion	Improving ante-natal care

Table 5 HC_5 and HC_{50} concentrations to fish and arthropods calculated from species sensitivity distributions for the insecticides with ETR values >100

Active ingredient	Fish No. of data points[a]	HC_5 (µg/L)	HC_{50} (µg/L)	Arthropod No. of data points[a]	HC_5 (µg/L)	HC_{50} (µg/L)
Alpha-cypermethrin	6	0.57 (0.05–1.77)	5.51 (1.90–15.99)	5	0.0085 (0.00000274–0.21)	3.11 (0.31–73.52)
Diazinon	20[b]	505.4 (221–890.8)	3403 (2189–5291)	25	0.48 (0.18–0.95)	6.30 (3.70–10.74)
Etofenprox	7	0.91 (0.00–13.44)	282.95 (24.88–3218.48)	12	0.058 (0.0015–0.52)	21.01 (3.44–128.3)
Fenobucarb	7	570.6 (97.42–1349)	3577 (1642–7792)	3	1.14 (0.00028–7.47)	18.89 (1.64–217.9)
Fipronil	5	37.72 (4.83–85.22)	170.46 (75.97–382.50)	29	0.037 (0.0099–0.1)	1.77 (0.85–3.71)
Permethrin	16	0.95 (0.5–1.92)	8.19 (4.66–14.38)	54	0.0084 (0.0026–0.02)	1.24 (0.62–2.47)
Phenthoate	10	4.69 (0.48–17.12)	117.17 (39.16–350.55)	3	0.0056 (0.0000000000039–0.67)	7.07 (0.014–0.0035)
Quinalphos	11	120.3 (35.2–246.9)	774.8 (425.1–1412)	–	–	–

Values in brackets are the confidence intervals

[a]No of data points indicate the number of species used in each SSD

Table 3. Summary of results from prospective, outcomes-based studies of real-world experience

	Standard RS cut-offs		
	RS <18	RS 18–30	RS ≥31
SEER (all N+; $N = 6768$)			
N (%)	3919 (64%)	2380 (35%)	469 (1%)
% CT "yes"/"no/unknown"	24%/76%	49%/51%	77%/23%
5-year BCSS (SE)	98.8% (0.3%)	97.3% (0.6%)	88.5% (2.4%)
p-value	<0.001		
Clalit (N1mi/1–3N+; $N = 709$)			
N (%)	379 (53%)	258 (36%)	72 (10%)
% CT/% no CT	7%/93%	40%/60%	86%/14%
5-year DR (95% CI)	3.2% (1.8%, 5.6%)	6.3% (3.9%, 10.1%)	16.9% (10.0%, 27.9%)
p-value	<0.001		
5-year BCSM (95% CI)	0.5% (0.1%, 2.1%)	3.4% (1.7%, 6.7%)	5.7% (2.2%, 14.4%)
p-value	<0.001		

BCSM breast cancer-specific mortality, *BCSS* breast cancer-specific survival, *CI* confidence interval, *CT* chemotherapy, *DR* distant recurrence, *RS* Recurrence Score result, *SE* standard error

Figure 1: Examples of visually well-structured tables extracted from real-world documents that lack explicit logical annotations.

Historically, TE has been approached using rule-based systems and heuristic algorithms [13, 78], which encoded handcrafted rules specific to visual or layout features. However, such methods tend to be brittle and fail to generalize across diverse domains and noisy data distributions. In response, recent years have witnessed a significant paradigm shift toward data-driven approaches powered by deep learning (DL) techniques [65, 74]. Deep neural networks, particularly those utilizing convolutional and transformer-based architectures, have demonstrated superior robustness and scalability, enabling them to handle a wide range of presentation styles with minimal manual intervention.

Among these approaches, transformer-based architectures have gained significant traction, largely due to their performance in modeling long-range dependencies and their modality-agnostic nature. For instance, methods like Detection Transformer (DETR) propose an end-to-end object detection framework that reframes detection as a set prediction problem, where object queries are directly matched with ground truth objects through bipartite matching, eliminating the need for traditional modules such as anchor generation or non-maximal suppression [7]. These models predict bounding boxes and class labels for all objects simultaneously, streamlining the training process while maintaining competitive accuracy. Autoregressive RNN-based decoding approaches [63, 73, 76, 77, 81] have also been explored in the context of TE, but transformers have largely outpaced them in terms of scalability and modeling capacity.

One of the key advantages of transformers lies in their capacity to treat visual data as sequences of non-overlapping patches, thereby eliminating the reliance on 2D structural inductive biases. This

property makes transformers highly attractive for multimodal learning scenarios where input modalities—such as images, text, or layout metadata—need to be processed jointly [2, 45, 51]. Additionally, transformers excel at large-scale self-supervised pretraining, providing strong visual representations that transfer well across downstream tasks.

Despite these strengths, transformers are not without limitations. Their self-attention mechanism suffers from quadratic time and memory complexity, making it computationally expensive and less feasible for long-sequence modeling or processing high-resolution documents where global context is crucial. In such cases, attention may become bottlenecked, and the network struggles to maintain long-range dependencies over extended inputs, especially in resource-constrained environments.

To address these inefficiencies, the research community has recently turned its attention to alternative architectures. Mamba, a state-space model (SSM) based sequence modeling framework, has emerged as a promising alternative to transformers, particularly for long-context tasks. Originally developed for language modeling, Mamba leverages structured state-space dynamics to model temporal and sequential relationships with linear time complexity. The Structured State-Space Sequence (S4) model introduced by [29] laid the groundwork for Mamba by proposing a parameterization that supports efficient sequence modeling while retaining memory of long-range patterns.

Mamba's primary architectural distinction lies in its implicit recurrence: rather than explicitly modeling pairwise token interactions as in transformers, it propagates state information using convolutional and recurrent mechanisms. This introduces an inductive bias toward sequential structure, enabling strong performance in language modeling tasks with fewer parameters and less computational overhead. Given that many vision problems—particularly those in document understanding such as OCR, table parsing, or layout analysis—also exhibit sequential or grid-structured properties, it is compelling to consider whether these advantages can be transferred from language to vision.

Nonetheless, Mamba also has limitations that constrain its applicability in complex visual settings. Most notably, its unidirectional modeling restricts the ability to reason over bidirectional spatial dependencies, which are essential for comprehending the layout of tables. Furthermore, while transformers employ learned positional embeddings to aid in spatial reasoning, Mamba lacks explicit positional encoding, potentially hampering its ability to localize and align content in structured documents.

To bridge the gap between Mamba's efficiency and transformer's structured reasoning capability, we propose a novel hybrid architecture for table extraction. In this design, we replace the transformer encoder in the DETR framework with a Vision-Mamba encoder, while retaining the original transformer decoder. This architecture harnesses the strengths of both models: Mamba acts as a lightweight and efficient encoder that processes image features with global context and recurrence bias, while the transformer decoder performs object-level matching and alignment via

cross-attention. The result is a model that achieves both computational efficiency and structured precision in a unified pipeline.

The proposed hybrid model offers several key benefits:

- It achieves linear encoding complexity via Mamba while maintaining the transformer decoder's strong object localization performance.
- It introduces recurrence-based inductive biases through Mamba, complementing the transformer's flexible query-based decoding.
- It improves training speed and memory utilization, especially beneficial in document scenarios where input resolution and sequence length are high.

This work aims to demonstrate the effectiveness of this hybrid design in table extraction tasks, where both high-level global understanding and fine-grained object alignment are critical. Through comprehensive experimentation and evaluation, we show that this approach improves over baseline DETR and Vision Transformer models in both accuracy and efficiency, setting a new benchmark for document intelligence tasks.

3.2 Methodology

In this thesis, we propose a novel architecture for object detection by integrating the Vision Mamba (Vim) encoder into the DETR (DEtection TRansformer) framework. This approach aims to enhance computational efficiency, addressing limitations inherent to the traditional Transformer encoder through bidirectional selective state-space modeling (SSM). The core innovation is substituting the original Transformer encoder with Vision Mamba, improving scalability and performance on complex visual datasets.

3.2.1 Algorithmic Summary

Algorithm 1 Hybrid Table Extraction with Vision-Mamba Encoder

Input: Document image I

Output: Predicted table bounding boxes and structure annotations

Step 1: Feature Extraction

Extract visual features: $F_0 = \texttt{CNN\_Backbone}(I)$

Flatten F_0 to a sequence of tokens $X_0 \in \mathbb{R}^{N \times d}$

Step 2: Vision-Mamba Encoding

for $l = 1$ **to** L **do**
$X_l = \texttt{VisionMambaEncoder}_l(X_{l-1})$

end

Set $Z = X_L$

Step 3: Object Query Decoding (Transformer Decoder)

Initialize object queries $Q \in \mathbb{R}^{M \times d}$

Decode predictions: $Y = \texttt{TransformerDecoder}(Q, Z)$

Step 4: Prediction Head

foreach $y_i \in Y$ **do**
Class label: $\hat{c}_i = \texttt{Linear}_{cls}(y_i)$

Bounding box: $\hat{b}_i = \texttt{MLP}_{box}(y_i)$

end

Step 5: Loss Computation

Match $(\hat{c}_i, \hat{b}_i)$ with (c_j, b_j) using bipartite matching

Compute:

$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}(\hat{c}_i, c_j)$

$\mathcal{L}_{\text{box}} = \ell_1(\hat{b}_i, b_j) + \text{GIoU}(\hat{b}_i, b_j)$

$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{box}}$

Step 6: Inference

Retain predictions where $\hat{c}_i \neq$ No-Object and score exceeds threshold

return final predicted bounding boxes and class labels

The overall training and inference procedure includes:

1. Feature extraction via CNN backbone.
2. Projection into token embeddings.
3. Encoding using Vision Mamba blocks.
4. Decoding via learned queries.
5. Object classification and bounding box prediction.
6. Training using the Hungarian set prediction loss.

3.2.2 Proposed Architecture

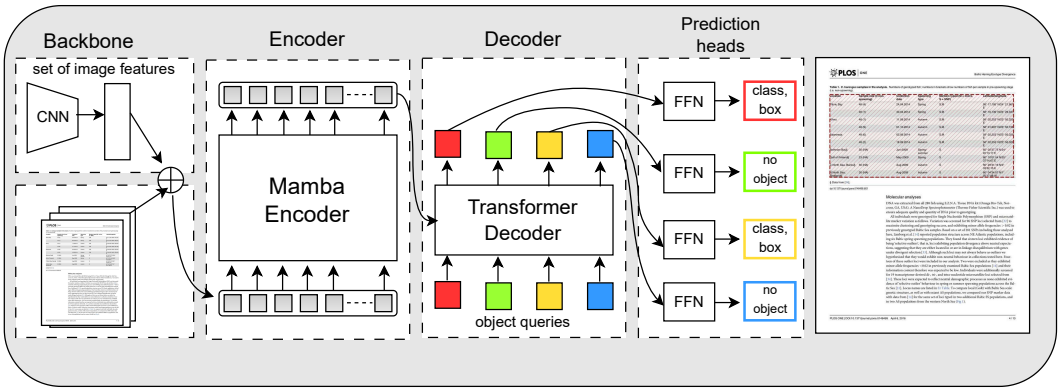

Figure 2: The proposed hybrid architecture comprises three main components: (1) a CNN backbone, (2) a Mamba encoder block, and (3) a Transformer decoder. The CNN backbone learns a 2D representation of an input image. The model flattens it and supplements it with a positional encoding before passing it into the Mamba encoder. The resulting token sequence is passed into the Mamba encoder block, which processes the sequence in both forward and backward directions, incorporating both contextual dependencies. Then, a Transformer decoder takes as input a small fixed number of learned positional embeddings, referred to as object queries. These queries attend to the output of the Mamba encoder through a cross-attention mechanism. Each output embedding from the decoder is then passed through a shared feed-forward network (FFN), which predicts object detections in the form of bounding boxes.

Figure 2 clearly illustrates the integrated hybrid architecture, highlighting the interaction between components and the flow of data through the system. The proposed model is designed to

process documents containing visually complex tables and predict their spatial and structural layout. The input to the system is a document page—typically in PDF format—rendered as an image containing one or more tables with varying structures. The output is the same image annotated with bounding boxes that localize detected tables as well as their fine-grained structural elements, such as rows, columns, and cells.

The hybrid object detection architecture follows a modular pipeline that transforms the input image into structured predictions. First, a CNN backbone extracts a two-dimensional feature map encoding spatial and semantic information from the raw image. This feature representation serves as the foundation for downstream processing. The feature map is then flattened into a sequence of tokens, each corresponding to a spatial location in the image. Positional encodings are added to retain spatial order information, ensuring that the model remains sensitive to the layout of visual elements.

This tokenized sequence is fed into a Mamba encoder block—a bidirectional state space model that efficiently captures both short-range and long-range dependencies within the sequence. The encoder outputs a globally contextualized sequence of embeddings that encode the document's visual content with fine-grained awareness.

A transformer decoder then operates on a fixed number of learned object queries, each serving as a proxy for a potential object in the image. Using a cross-attention mechanism, the decoder enables each query to interact with the encoder outputs and gather relevant contextual features. These refined query embeddings are subsequently passed through a shared feed-forward prediction head, which classifies the presence of an object and regresses its bounding box coordinates.

The final output of the model is a set of object-level predictions corresponding to tables and their internal structural elements, accurately localized within the document image.

3.3 Model Components

This section introduces three state-of-the-art deep learning models that form the core components of our proposed hybrid architecture: Vision-Mamba, DEtection TRansformer (DETR), and ResNet. Each model contributes a distinct functionality within the overall pipeline, and the subsequent subsections provide detailed descriptions of their roles and implementation.

3.3.1 Vision Mamba

Recent advances in sequence modeling have brought renewed attention to State Space Models (SSMs) as a powerful alternative to traditional recurrent and attention-based architectures. Building

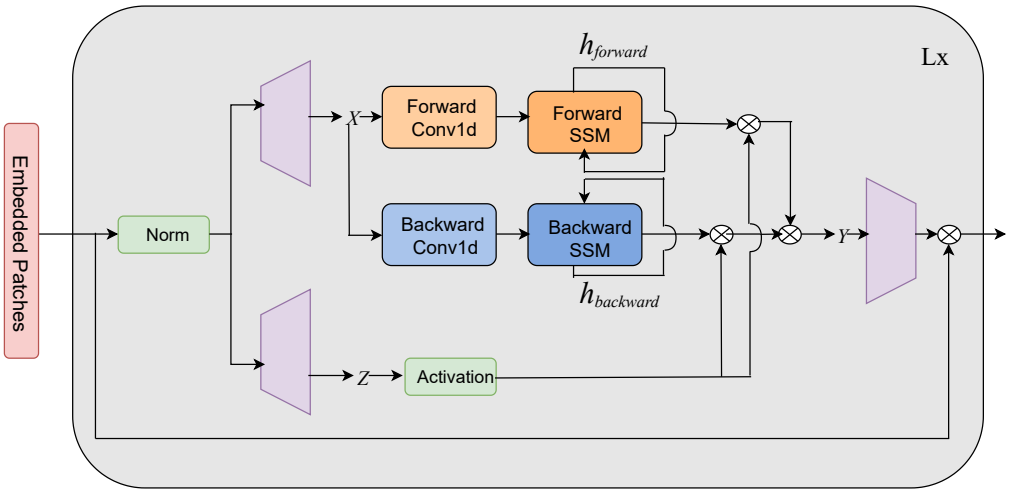

Figure 3: The Vision Mamba encoder processes a sequence of embedded image patches through a structured bidirectional pipeline. The sequence is then split and processed in two directions: forward and backward. In the forward path, the sequence is passed through a Conv1D layer to project the tokens into a latent space, followed by a forward State Space Model (SSM) layer that captures long-range dependencies efficiently in the time-forward direction. Simultaneously, in the backward path, the sequence is also processed via a separate Conv1D layer and a backward SSM, which models context in the reverse temporal direction. The outputs from the forward and backward SSMs—denoted as h_{forward} and h_{backward}—are then aggregated or fused to form the final encoded sequence Z, which contains rich bidirectional contextual information. This figure is from [48]

on foundational work such as the Kalman Filter [10], modern SSMs have demonstrated a remarkable ability to capture long-range dependencies while supporting parallelizable training workflows. Notable contributions in this area include the Linear State-Space Layer (LSSL) [3], the Structured State-Space Sequence Model (S4) [1], the Diagonal State Space (DSS) model [9], and S4D [4]. These models are specifically designed to handle sequential data efficiently across a broad range of modalities and tasks, offering scalable mechanisms to learn extended-range dependencies.

One of the key advantages of SSMs lies in their ability to process extremely long input sequences using near-linear time complexity, often implemented through convolutional operations. This computational efficiency has led to their integration into vision tasks, where two-dimensional data must be modeled with both local and global spatial dependencies. For example, models like ConvSSM [29] successfully combine the structural benefits of SSMs with convolutional and transformer-based components to efficiently model image and video data.

More recently, the introduction of Mamba has further advanced the capabilities of SSMs by introducing time-varying parameterization and a hardware-efficient implementation. Mamba achieves significant improvements in both training and inference speed, positioning itself as a competitive—if not preferable—alternative to transformer architectures for long-range language modeling tasks. However, despite these advancements, the integration of SSMs as a foundational backbone for vision-centric architectures—capable of fully leveraging visual information embedded in both images and videos—remains an open research direction.

In our work, we replace the traditional Transformer encoder with a Vision Mamba Encoder Figure 3, which enhances computational efficiency while preserving global context modeling [48]. Unlike standard attention mechanisms, the Vision Mamba processes token sequences in both forward and backward directions, allowing for bidirectional flow of information. This dual-path architecture improves the encoder's ability to capture spatial dependencies in visual data, making it especially suitable for structured vision tasks such as table detection and layout analysis.

3.3.1.1 Input Preparation and Tokenization

The encoder first applies a CNN backbone to the input image and produces a feature-map

$$f \in \mathbb{R}^{H \times W \times C}.$$

We then partition f into J non-overlapping spatial patches $\{f_p^j\}_{j=1}^J$, each of size $(P \times P \times C)$. Each patch is flattened,

$$t_p^j = \mathrm{vec}(f_p^j) \in \mathbb{R}^{P^2 C},$$

and linearly projected into a D-dimensional embedding via a weight matrix $W \in \mathbb{R}^{(P^2C) \times D}$. We prepend a learnable classification token $t_{\mathrm{cls}} \in \mathbb{R}^D$ and add positional embeddings $E_{\mathrm{pos}} \in \mathbb{R}^{(J+1) \times D}$, yielding

$$T_0 = \left[t_{\mathrm{cls}};\ t_p^1 W;\ t_p^2 W;\ \ldots;\ t_p^J W\right] + E_{\mathrm{pos}}, \tag{1}$$

where

f	CNN feature-map of size $H \times W \times C$,
f_p^j	jth patch of f, flattened to t_p^j,
W	projection matrix,
t_{cls}	learnable "class" token,
E_{pos}	positional embeddings.

3.3.1.2 State Space Model (SSM)

The Vision Mamba encoder utilizes a bidirectional selective SSM, with equations represented as:
Continuous form:

$$h'_t = Ah_t + Bx_t, \quad y_t = Ch_t \tag{2}$$

Discrete form (after Zero-Order Hold discretization):

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t, \tag{3}$$

where A, B, C are learned parameters, and h_t denotes hidden states.

3.3.1.3 Vision Mamba Block

Vision Mamba blocks process sequences through forward and backward convolutions:

$$h^{(forward)} * t + 1 = \sigma(A^{(forward)} h_t + B^{(forward)} x_t), \quad h^{(backward)} * t - 1 = \sigma(A^{(backward)} h_t + B^{(backward)} x_t), \tag{4}$$

The combined output from both directions provides comprehensive visual encoding:

$$y_t = C^{(forward)} h_t^{(forward)} + C^{(backward)} h_t^{(backward)}, \tag{5}$$

where σ is typically the GELU activation function.

3.3.2 Detection Transformer (DETR)

The DEtection TRansformer (DETR) model, introduced by Carion et al. [7], marked a significant paradigm shift in the object detection landscape by framing detection as a direct set prediction problem. Departing from traditional object detection pipelines that rely on a combination of region proposal mechanisms, anchor boxes, and non-maximum suppression (NMS), DETR employs a transformer-based encoder-decoder architecture to directly map an input image to a fixed-size set of objects in a fully end-to-end manner. Its innovative approach eliminates hand-crafted components, leading to a streamlined and conceptually elegant detection framework.

3.3.2.1 Architectural Overview

At a high level, DETR consists of three main components Figure 4: (1) a convolutional backbone for feature extraction, (2) a transformer encoder-decoder for global reasoning and query-based object localization, and (3) a feed-forward prediction head for classification and bounding box regression.

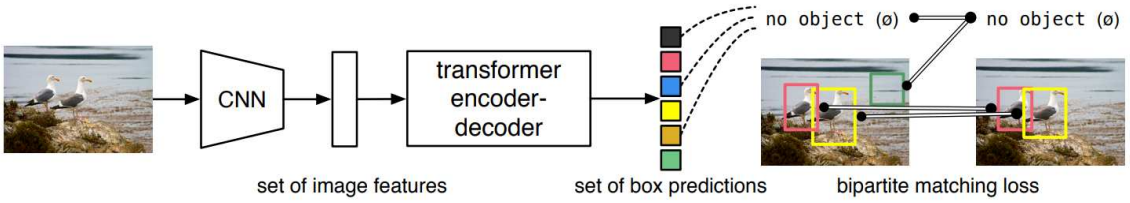

Figure 4: The overview of DETR architecture. This figure is from [7]

1. CNN Backbone. The DETR pipeline begins with a convolutional neural network (typically a ResNet-50 or ResNet-101) [32] to extract high-level visual features from the input image. Let the input image be denoted by $I \in \mathbb{R}^{H \times W \times 3}$. After processing through the CNN backbone, the image is transformed into a lower-resolution feature map $F \in \mathbb{R}^{C \times H' \times W'}$, where $H' = H/32$, $W' = W/32$, and C is typically 256.

To prepare the feature map for the transformer, F is flattened and reshaped into a sequence of N tokens, where $N = H' \times W'$, and each token corresponds to a spatial patch. Positional encodings are added to this sequence to preserve spatial information, which is otherwise absent in the transformer architecture.

2. Transformer Encoder-Decoder. DETR adopts the standard transformer architecture [84] to globally model relationships between all spatial tokens. The encoder processes the entire feature sequence in parallel using self-attention and feedforward layers, enabling it to capture rich contextual information across the image.

The decoder is the most distinctive element of DETR. Rather than relying on region proposals or sliding windows, the decoder takes as input a fixed set of M learnable object queries, denoted $Q \in \mathbb{R}^{M \times d}$, where d is the model's hidden dimension (typically 256), and M is the maximum number of objects the model can predict (usually set to 100). Each object query is expected to specialize in detecting one object.

Through multiple layers of multi-head attention, the decoder learns to associate each query with an object present in the image or to assign it as a "no object" prediction. The decoder attends to both the encoder outputs and the object queries, yielding a sequence of decoder outputs $Y \in \mathbb{R}^{M \times d}$.

3. Prediction Heads. Each decoder output y_i is passed through a shared prediction head comprising a linear layer for classification and a multi-layer perceptron (MLP) for bounding box regression. Specifically:

$$\hat{c}_i = \text{Linear}_{\text{cls}}(y_i), \quad \hat{b}_i = \text{MLP}_{\text{box}}(y_i)$$

where $\hat{c}_i$ is the class probability distribution (including a "no object" class), and $\hat{b}_i \in [0,1]^4$ represents normalized bounding box coordinates in the format (center_x, center_y, width, height).

3.3.2.2 Loss Function and Set-Based Matching

A key innovation in DETR is the use of bipartite matching between predicted and ground-truth objects. Traditional detectors rely on IoU-based heuristics and NMS to resolve overlapping boxes, which are brittle and non-differentiable. In contrast, DETR leverages the Hungarian algorithm to perform a one-to-one matching between predictions and targets based on a matching cost that combines classification and localization errors.

The loss function is defined over the matched pairs and includes:

- Cross-entropy loss for object classification.
- ℓ_1 loss and Generalized IoU (GIoU) [75] for bounding box regression.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{CE}} + \lambda_{\ell_1} \cdot \mathcal{L}_{\ell_1} + \lambda_{\text{giou}} \cdot \mathcal{L}_{\text{GIoU}}$$

This set-based loss formulation ensures that each object query is trained to specialize in a unique detection task, resulting in diverse and non-redundant predictions.

3.3.2.3 Efficiency and Benefits

While DETR was initially criticized for slow convergence (e.g., requiring 500 epochs on COCO with ResNet-50), subsequent work has introduced various improvements to address this issue. Notably, Deformable DETR [97] and Conditional DETR [59] accelerate training through dynamic attention mechanisms and better query initialization.

Despite the early drawbacks, DETR offers several compelling advantages:

- **Simplicity**: The model eliminates region proposal networks, anchor generation, and NMS, reducing engineering complexity.
- **End-to-End Optimization**: The entire model is trained jointly with a single loss function, making it easier to tune and more stable to optimize.
- **Global Context Reasoning**: Transformers allow each token (patch) to attend to all others, leading to better object disambiguation in complex scenes.
- **Fixed-Size Output**: DETR always predicts a fixed number of object slots, simplifying post-processing and making it suitable for structured output tasks.

These properties make DETR particularly suitable for structured vision tasks like table detection, document layout analysis, and scene parsing, where spatial reasoning and semantic alignment are crucial.

3.3.2.4 Relevance to Our Work

In our proposed hybrid architecture, we retain the DETR decoder and prediction head components while replacing the original transformer encoder with a Vision-Mamba encoder. This design allows us to maintain the structured set-based decoding strengths of DETR while improving encoder-side efficiency and long-range dependency modeling via state-space representations. As a result, we achieve better computational scalability and training stability, especially for high-resolution document images with complex tabular structures.

By integrating the DETR decoding mechanism with a more efficient encoder, our model preserves the architectural elegance of DETR while adapting it to modern efficiency-optimized modules, making it well-suited for document understanding and table extraction tasks.

3.3.3 ResNet

Our model initially employs a CNN backbone, such as ResNet-18, to extract spatial features from input images. ResNet-18 is a member of the Residual Network (ResNet) family, introduced by He et al. [32], which addresses fundamental challenges in training deep convolutional neural networks, particularly the degradation of performance with increased depth.

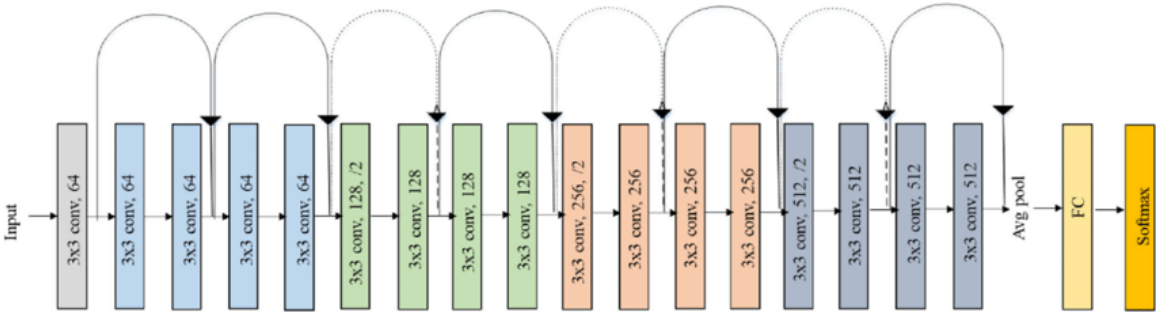

Figure 5: Architectural overview of ResNet-18. This figure is from [70]

Unlike traditional convolutional networks where deeper models may suffer from increased training error, ResNets leverage residual learning by introducing shortcut connections that bypass one or more layers Figure 5. This allows the network to learn residual mappings instead of attempting to directly learn unreferenced transformations, facilitating more stable and effective gradient flow during backpropagation. In the case of ResNet-18, the architecture is composed of 18 layers including

convolutional blocks, batch normalization, and ReLU activations, structured with residual connections after every two convolutional layers. Formally, given an image $I \in \mathbb{R}^{3\times H_0\times W_0}$, the backbone outputs feature maps $f \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channels, and H, W indicate spatial dimensions.

One of the main advantages of using ResNet-18 is its balance between model complexity and representational capacity. Compared to deeper variants like ResNet-50 or ResNet-101, ResNet-18 is significantly lighter in terms of parameters and computational cost, making it particularly suitable for document-level vision tasks where high throughput and efficient processing are essential. Despite its relatively shallow depth, ResNet-18 is capable of capturing rich hierarchical features necessary for downstream tasks such as object detection, segmentation, and table structure recognition.

Furthermore, its modular design and widespread use in vision tasks make it a robust and well-understood choice for use as a backbone in hybrid architectures. When integrated with modern architectures like Vision-Mamba and DETR, ResNet-18 provides a strong and efficient feature base, enabling the higher-level components to focus on long-range context modeling and structured reasoning.

3.4 Conclusion

This section has presented the core methodology underlying our hybrid architecture for table detection and structure recognition. The proposed model builds upon the DETR framework by integrating a Vision Mamba encoder in place of the traditional transformer encoder, paired with a ResNet-18 convolutional backbone for efficient visual feature extraction. This design aims to leverage the strengths of state-space modeling and transformer-based decoding in a unified, end-to-end trainable architecture optimized for document image understanding.

We began by revisiting the architectural challenges inherent in traditional transformer-based vision models, particularly those related to computational inefficiencies and the quadratic complexity of self-attention in handling high-resolution inputs or long-range dependencies. To address these limitations, we introduced Vision Mamba, a recent state-space model that offers bidirectional processing capabilities with near-linear computational complexity. Its ability to model sequential data efficiently and retain long-range spatial dependencies makes it a compelling encoder alternative for vision tasks, particularly in structured document analysis.

The encoder's output is consumed by the original DETR decoder, which remains a central component of the architecture. With its fixed set of learnable object queries and cross-attention mechanisms, the decoder facilitates set-based prediction and object-level alignment without relying on region proposals or non-maximum suppression. The DETR decoder, in this context, plays a

crucial role in learning structured associations between encoded visual representations and semantic table components.

To support lightweight yet expressive feature extraction, we employ ResNet-18 as the backbone. Its residual blocks enable stable gradient flow, even in deep architectures, and serve as a reliable module for producing spatially rich token embeddings suitable for Mamba-based sequence modeling. The simplicity and efficiency of ResNet-18 also ensure that the overall pipeline remains computationally tractable, making it viable for high-resolution document processing scenarios.

Together, these components form a hybrid system that balances architectural efficiency with modeling precision. The combination of Mamba's recurrence-style encoding, ResNet's lightweight feature abstraction and DETR's structured decoding allows our model to process complex table layouts while maintaining scalability and generalizability.

This section serves to bridge the conceptual underpinnings of our proposed architecture with the experimental results that follow. In the next section, we detail the implementation workflow, training setup, and evaluation methodology. We also present comparative results against established baselines, offering insights into the model's real-world applicability and its capacity to generalize across diverse document formats.

Chapter 4

Experiments

4.1 Introduction

This section presents a comprehensive evaluation of our proposed hybrid architecture, which integrates a Vision-Mamba encoder into the DETR framework for table structure recognition and detection. The goal of these experiments is to rigorously assess the effectiveness, robustness and efficiency of the model under various conditions, and to compare its performance against state-of-the-art methods.

We begin by outlining the experimental setup, including dataset usage, training schedule, and evaluation protocol. This is followed by a summary of the datasets, with a focus on PubTables-1M as our primary benchmark. We then present the baseline models used for comparison, briefly describing their relevance to our task.

Next, we provide key implementation details, including model configuration, hyperparameters, and the training environment, to support reproducibility. Finally, we describe our loss functions, optimization strategy, and inference process used throughout the experiments.

Following this, we present experimental results on a variety of metrics and conditions. We explore the impact of different hyperparameters and a head-to-head comparison with transformer-based baselines. These results are analyzed both quantitatively and qualitatively to understand how our hybrid model handles challenges such as long-range dependencies, complex table layouts, and ambiguous structures.

Finally, we conclude this section by distilling key insights gained from the experiments. These include an analysis of trade-offs between model complexity and accuracy, the efficiency gains achieved through Vision-Mamba encoding, and the overall effectiveness of the hybrid design in balancing structured reasoning with computational scalability.

4.2 Experimental Setup

Dataset: In this work, we utilize the PubTables-1M dataset to train and evaluate our proposed hybrid architecture for table detection and structure recognition. It was introduced by Microsoft Research to address the limitations of previous datasets, which were either small in size, lacked detailed annotations, or were limited in domain diversity.

Dataset	Input Modality	# Tables	Cell Topology	Cell Content	Cell Location	Row & Column Location	Canonical Structure
TableBank [46]	Image	145K	✓		✓		
SciTSR [13]	PDF*	15K	✓	✓	✓		
PubTabNet [92, 93]	Image	510K	✓	✓	✓	✓	
FinTabNet [92]	PDF*	113K	✓	✓	✓	✓	
PubTables-1M	**PDF***	**948K**	✓	✓	✓	✓	✓

Table 1: Comparison of crowd-sourced datasets for table structure recognition. This data was taken from [80]

PubTables-1M is currently one of the largest and most comprehensive datasets for document-based table extraction tasks Table 1. It provides high-quality annotations for both structural and functional aspects of tables, enabling end-to-end training and evaluation of table extraction pipelines across multiple subtasks. The dataset comprises over 947,000 annotated tables extracted from approximately 460,000 document pages, making it one of the most expansive and reliable benchmarks for evaluating deep learning models in document understanding.

Property	**Value**
Total document pages	460,000+
Total annotated tables	947,000+
Total cells (including empty)	35 million+
Tables with structural annotations	100%
Tables with functional labels	100%
Average rows per table	9.2
Average columns per table	5.8
Average non-empty cells per table	37.1

Table 2: Summary statistics of the PubTables-1M dataset. This data was taken from [80]

Given the reliance of our model on capturing both global layout and fine-grained cell structure, PubTables-1M enables a realistic and challenging testbed to validate the architecture's ability to handle high variability in table styles, as found across academic documents Table 2.

Baseline Methods: To evaluate the effectiveness of our proposed hybrid architecture, we compare it against two strong baseline methods commonly used in table detection and structure recognition tasks:

Faster R-CNN and DETR. Faster R-CNN is a two-stage object detection framework that first proposes candidate regions using a Region Proposal Network (RPN), followed by a second-stage classifier and regressor for object recognition and bounding box refinement.

DETR (DEtection TRansformer) is a transformer-based end-to-end object detection model that replaces hand-crafted components with a set-based prediction approach. It eliminates the need for region proposals and post-processing steps such as non-maximum suppression by learning a direct mapping between object queries and targets through bipartite matching.

DETR serves as a natural baseline for our hybrid model, as we retain its decoder while replacing the transformer encoder with a Vision-Mamba encoder. These baselines enable a fair comparison in terms of both detection accuracy and computational efficiency, allowing us to isolate the contributions of our encoder modification.

Implementation Details: All experiments were conducted on a Linux workstation equipped with two NVIDIA RTX A4500 GPUs (20 GB VRAM each) and 256 GB of system RAM. To ensure a fair comparison, we adhered to the official dataset splits and experimental configurations used by baseline models. The PubTables-1M dataset is partitioned into three subsets: a training set containing approximately 900K tables extracted from 420K pages, a validation set with 24K tables, and a test set comprising 23K tables.

Our hybrid model was implemented using the PyTorch framework. The encoder consists of two layers of Vision-Mamba, each with a hidden dimension of 256. We trained the model for 20 epochs using the AdamW optimizer [57], with a learning rate of 1e-4 and a weight decay of 1e-4. A dropout rate of 0.1 was applied uniformly across all datasets and model layers.

For supervision, we employed a combination of loss functions: cross-entropy loss for classification and a composite loss for bounding-box regression, consisting of L1 loss and generalized intersection over Union (GIoU) loss. These losses are combined using the set-based matching strategy from DETR, enabling effective end-to-end optimization.

4.3 Experiments and Results

Motivation. While transfer learning via pre-training and fine-tuning has become a widely adopted strategy in computer vision and document understanding tasks, its effectiveness is often contingent on the alignment between the pre-training domain and the downstream task. In the context of table structure recognition, many existing pre-trained models are either trained on natural image datasets

(e.g., ImageNet) or general document layouts, which may not capture the unique structural patterns, layout semantics, and visual grammar inherent in scientific tables extracted from PDF documents.

The PubTables-1M dataset provides a large-scale annotated corpus specifically curated for table structure recognition. Its high-quality annotations include cell-level boundaries, content, and logical structure—offering a comprehensive supervision signal tailored to the nuances of tabular data. Training our model architecture from scratch on this domain-specific dataset ensures that the model learns inductive biases and feature representations directly from the task-relevant distribution, without being constrained by the potentially misaligned pre-training objectives or architectures designed for unrelated domains.

Moreover, our proposed hybrid architecture introduces novel design elements—such as the integration of state space models with vision-specific encoders—that differ significantly from conventional transformer-based or CNN-based backbones. Relying on pre-trained weights from fundamentally different architectures may impede convergence or lead to suboptimal performance. By training from scratch, we allow the model to fully exploit the design space of our architecture and adapt it specifically to the characteristics of the PubTables-1M dataset, leading to better generalization on table parsing tasks.

Ultimately, training from scratch offers the dual advantage of architecture-aligned representation learning and dataset-specific optimization, enabling a fair and principled evaluation of our proposed model's capabilities in extracting table structure information from complex, real-world documents.

Metrics. Since our task is table detection and table structure recognition, we use the metrics Average Precision **(AP)** and Average Recall **(AR)** to evaluate our model's performance.

In the context of table detection, these metrics assess how accurately the model can identify and localize entire tables within a document. AP50 serves as a tolerance-based indicator of correct detection, while AP75 requires stricter alignment, ensuring that the predicted bounding boxes closely match the ground truth.

For *structure recognition*, where the task involves identifying and localizing fine-grained components such as rows, columns, and cells, **AR** becomes particularly important. A high recall indicates that the model successfully captures the majority of table structure elements, which is crucial for downstream tasks like table parsing and reconstruction.

Using AP and AR metrics allows us to comprehensively evaluate both the precision and completeness of detections across varying levels of strictness, enabling fair comparison with existing models and robust benchmarking on PubTables-1M and similar datasets.

Average Precision (AP)

Average Precision (AP) measures the area under the precision-recall curve, summarizing the trade-off between precision and recall across varying confidence thresholds. It is defined as:

$$\text{AP} = \int_0^1 p(r)\, dr$$

where $p(r)$ is the precision as a function of recall r. In practice, AP is often approximated using discrete recall levels and computed using either 11-point interpolation or COCO-style evaluation, which averages precision at multiple Intersection over Union (IoU) thresholds.

AP@50 (AP50) and AP@75 (AP75)

AP@50 and AP@75 are specific instances of AP computed at fixed IoU thresholds:

- **AP@50** evaluates the average precision at an IoU threshold of 0.50, indicating a moderate match between predicted and ground-truth boxes.
- **AP@75** uses a stricter threshold of 0.75, requiring more precise localization of detected regions.

Formally, for a given IoU threshold τ:

$$\text{AP@}\tau = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i \quad \text{where IoU}_i \geq \tau$$

Average Recall (AR)

Average Recall (AR) measures the mean recall across a range of IoU thresholds and detection limits. It provides insight into the model's ability to detect all relevant objects. It is calculated as:

$$\text{AR} = \frac{1}{T} \sum_{\tau \in T} \text{Recall}(\tau)$$

where T is the set of IoU thresholds, typically from 0.50 to 0.95 in increments of 0.05. Higher AR values indicate the model's effectiveness in capturing all relevant structural components within tables.

4.3.1 Performance Analysis

4.3.1.1 Quantitative Results: Table Detection Task

The table detection task forms a foundational step in document parsing pipelines, as accurate localization of tabular regions is a prerequisite for downstream structure recognition and content extraction. In this section, we present a detailed quantitative evaluation of our proposed hybrid model against two widely adopted baselines—DETR and Faster R-CNN—using standard detection metrics: Average Precision (AP), AP at IoU thresholds 0.5 (AP50) and 0.75 (AP75), and Average Recall (AR). All models are evaluated on the PubTables-1M benchmark dataset under identical conditions to ensure fairness.

Model	AP	AP50	AP75	AR
Faster R-CNN	0.825	0.985	0.927	0.866
DETR	**0.966**	**0.995**	0.988	**0.981**
Tramba	0.962	**0.995**	**0.992**	0.977

Table 3: Performance comparison on the table detection task.

Metric-Wise Comparison Table 3 summarizes the performance of the three models on the table detection task. Our model achieves an overall AP of 0.962, which is on par with DETR (0.966) and significantly outperforms Faster R-CNN (0.825). While DETR slightly edges out our model on the AP metric, the difference is marginal (0.004 absolute). Notably, at AP75—a stricter criterion for localization precision—our model achieves the highest score of 0.992, surpassing both DETR (0.988) and Faster R-CNN (0.927). This highlights our model's superior ability to tightly align predicted bounding boxes with ground truth annotations.

In terms of recall, our model records an AR of 0.977, again closely matching DETR (0.981) and outperforming Faster R-CNN (0.866) by a significant margin. High AR scores indicate the model's capability to detect nearly all table instances with minimal false negatives, a critical factor for ensuring completeness in document parsing pipelines.

Interestingly, all models converge at AP50, achieving a near-perfect score of 0.995 (DETR and our model) and 0.985 (Faster R-CNN). This suggests that while coarse localization is handled well across models, the differentiating factor lies in fine-grained precision, where our model demonstrates competitive, if not superior, performance.

Model Behavior under IoU Thresholds The behavior of a detection model under varying Intersection over Union (IoU) thresholds is indicative of its localization robustness. A model that performs well at lower thresholds may still struggle with precise alignment, whereas a high AP75 implies that predictions are tightly bound to the actual table regions.

In this regard, our model, Figure 6 demonstrates exceptional performance, with an AP75 of 0.992—the highest among all three. This suggests that our Vision Mamba-based encoder contributes positively to modeling long-range visual dependencies and spatial context, enabling more precise boundary estimation. DETR also performs well at AP75 (0.988), benefiting from its transformer-based global reasoning. However, Faster R-CNN trails notably, reinforcing that region-based methods may falter in complex document layouts where contextual cues play a larger role.

The negligible gap between AP and AP75 in our model (0.962 vs. 0.992) is particularly noteworthy. It implies that a large fraction of our detections are not just correct, but also highly accurate

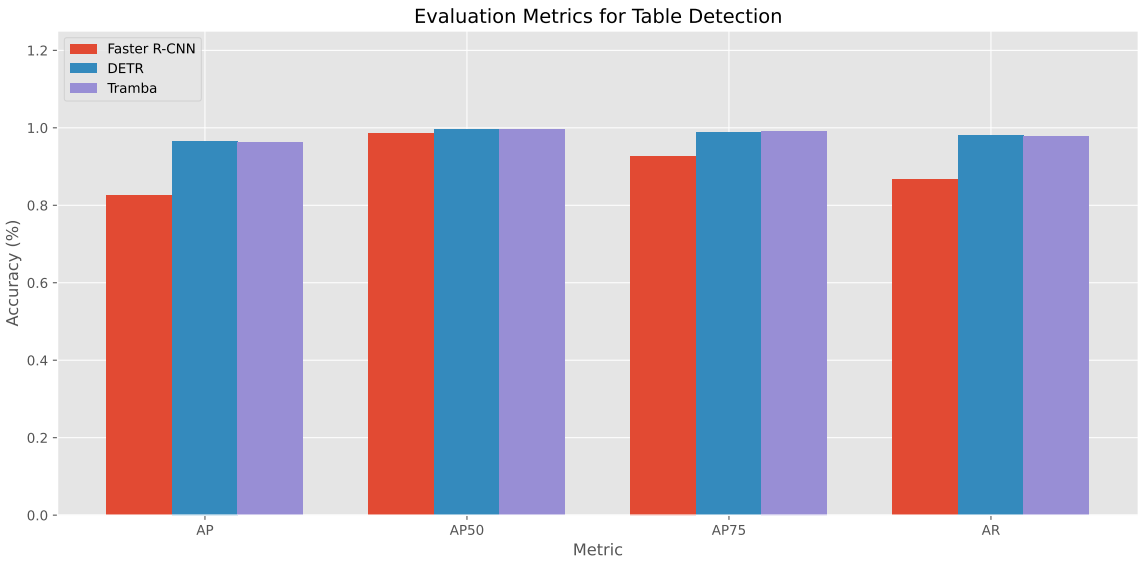

Figure 6: Evaluation metrics on detection task

in terms of spatial overlap—a desirable trait when tight box alignment is critical for downstream structural segmentation.

Efficiency Level Comparison In addition to accuracy, the efficiency of a model—both in terms of training convergence and architectural complexity—is essential for real-world deployment. Table 4 presents an ablation of our model's performance across different training epochs. At just 10 epochs, our model already achieves an AP of 0.950 and an AR of 0.970, with strong AP75 performance (0.988). By 20 epochs, the model converges to 0.956 AP and 0.977 AR—values that are nearly indistinguishable from the results of full baseline models.

Model (Epoch)	**AP**	**AP50**	**AP75**	**AR**
Tramba (10 epochs)	0.950	0.985	0.988	0.970
Tramba (15 epochs)	0.952	0.990	0.990	0.973
Tramba (20 epochs)	**0.962**	**0.995**	**0.992**	**0.977**

Table 4: Epoch-wise convergence of our model on the detection task.

What makes this convergence particularly significant is the architectural configuration. While DETR and Faster R-CNN were trained with 6 encoder and 6 decoder layers, our model was trained using only 2 encoder and 2 decoder layers, yet it manages to achieve near-equivalent or superior

detection performance. This not only indicates better learning efficiency but also reflects the architectural advantages introduced by the Vision Mamba encoder, which leverages bidirectional State Space Models for efficient long-range sequence modeling.

The rapid convergence of our model within just 20 epochs and with reduced model depth demonstrates a favorable accuracy-efficiency trade-off. It implies that our approach can be trained faster, deployed with lower memory overhead, and scaled more easily—all without sacrificing detection quality. In high-throughput document processing systems, where inference time and compute cost are bottlenecks, such lightweight yet high-performing models are invaluable.

4.3.1.2 Quantitative Results: Table Structure Recognition Task

Table structure recognition poses a considerably greater challenge than table detection. While detection is primarily concerned with identifying and localizing bounding boxes around tabular regions, structure recognition requires fine-grained understanding of the internal organization of these tables — including rows, columns, and cell boundaries. This task becomes more complex when dealing with varied document layouts, cell spanning behavior, and implicit structural cues (e.g., alignment, whitespace, visual dividers). In this section, we evaluate the performance of our hybrid model on this task, comparing it with two established baselines: Faster R-CNN and DETR, using Average Precision (AP), AP at IoU thresholds of 0.5 (AP50) and 0.75 (AP75), and Average Recall (AR) as evaluation metrics.

Model	**AP**	**AP50**	**AP75**	**AR**
Faster R-CNN	0.722	0.815	0.785	0.762
DETR	**0.912**	**0.971**	**0.948**	**0.942**
Tramba	0.774	0.940	0.852	0.845

Table 5: Performance comparison on the table structure recognition task.

Structural AP/AR Metrics As presented in Table 5, our model achieves an AP of 0.774 and an AR of 0.845 on the structure recognition task. This is a clear improvement over Faster R-CNN, which achieves an AP of 0.722 and AR of 0.762. Our model also demonstrates significantly stronger localization precision, recording an AP75 of 0.852 — notably higher than Faster R-CNN's 0.785. This gap suggests that our architecture is better suited for capturing nuanced table structures, especially in documents with complex or less regular layouts.

However, DETR remains the top-performing model in this task, achieving an AP of 0.912 and an AR of 0.942. Its high AP75 (0.948) indicates strong capability in precise boundary alignment for

structure components. Our model, while competitive, does fall short of matching DETR's absolute performance. This performance gap is attributable to several important factors. Most significantly, our model was trained with **only 2 encoder and 2 decoder layers**, while DETR was trained with **6 encoder and 6 decoder layers**, giving it a much deeper capacity for feature abstraction and spatial reasoning.

Furthermore, the table structure recognition dataset used in this task is considerably larger and more complex than the detection dataset. This magnifies the impact of architectural depth and training resources. While our model demonstrates strong generalization and outperforms a traditional detection-based pipeline like Faster R-CNN, it does not yet fully close the performance gap with DETR in structure-level parsing.

Model Strengths in Fine-Grained Detection Despite being constrained in architectural depth, our model shows considerable strength in localizing fine-grained table components. The Vision Mamba encoder, with its bidirectional state-space modeling capability, contributes significantly to this performance by efficiently capturing long-range contextual dependencies in both forward and backward directions. This allows the model to better understand structural patterns in documents, such as aligned rows, equally spaced columns, and hierarchical cell groupings.

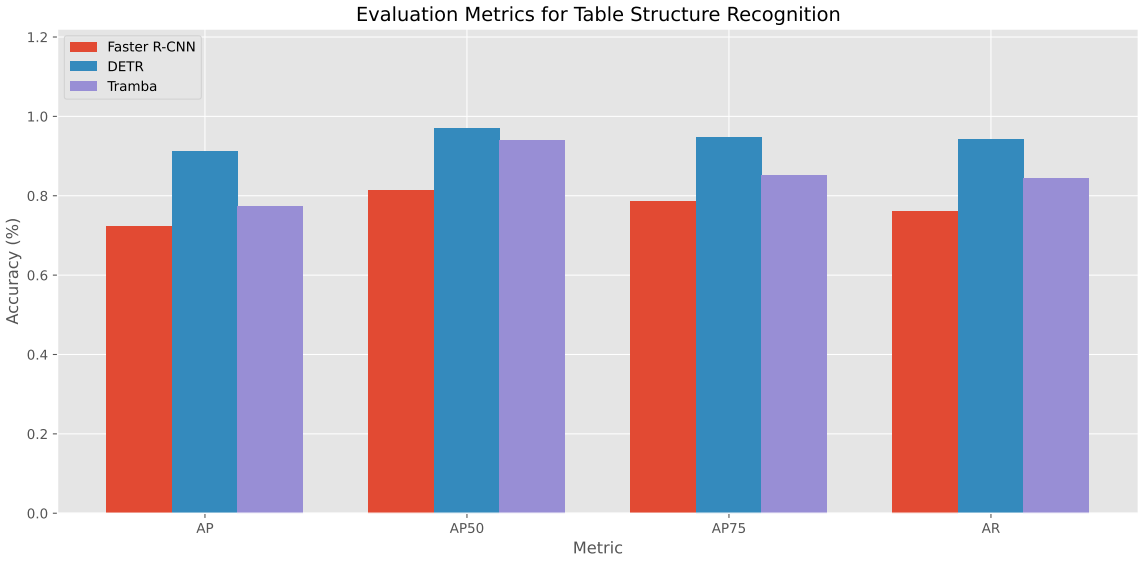

Figure 7: Evaluation metrics on structure recognition task

This capability becomes particularly valuable in cases involving spanning cells, nested tables, or soft layout cues (e.g., white space-based cell separation). Figure 7 the AP75 of 0.852 — substantially higher than Faster R-CNN — reflects this ability to localize structure elements with high precision. Additionally, the relatively small drop from AP50 (0.940) to AP75 (0.852) further suggests that a

majority of the structure predictions made by our model are not just correct in terms of coverage but also tightly aligned with ground truth annotations.

Even though DETR outperforms our model across all structure metrics, our results affirm that Vision Mamba's sequence modeling architecture holds promise for fine-grained layout understanding, especially given that our model achieves these results using one-third of the layers and a fraction of the training compute.

Cross-Task Performance and Correlation The relationship between detection quality and structure recognition performance is also evident in our results. In the previous section, we showed that our model performs strongly in table detection (AP = 0.962), indicating that it provides reliable input regions for structure parsing. This high detection accuracy clearly contributes to robust performance in the structure task, with AR reaching 0.845.

That said, structure recognition does not solely depend on detection quality. While good bounding boxes are essential, structure recognition requires additional modeling of intra-table relationships. Our results suggest that even with high detection performance, limitations in model depth restrict the ability to fully parse structural components, especially in long or irregular tables. Nonetheless, the consistency in AP/AR progression across both tasks indicates that the Vision Mamba backbone is capable of learning transferable features that benefit both detection and structure segmentation.

Moreover, the fact that our model performs relatively well despite having significantly fewer parameters and shallower architecture demonstrates that performance improvements in structure recognition are not purely tied to model size. This opens up an important avenue for further exploration: scaling Vision Mamba-based models with deeper encoders and decoders, or hybridizing Mamba with attention-based modules to better handle varying table configurations.

Training Convergence and Efficiency To understand the convergence behavior of our model, we present its epoch-wise performance in Table 6. At just 10 epochs, the model achieves an AP of 0.685 and AR of 0.720. By 15 epochs, the AP climbs to 0.772 and AR to 0.846 — effectively nearing the final 20-epoch performance. At 20 epochs, the model plateaus at an AP of 0.774 and AR of 0.845.

Epoch	**AP**	**AP50**	**AP75**	**AR**
Tramba (10 Epochs)	0.685	0.782	0.754	0.720
Tramba (15 Epochs)	0.772	0.938	0.846	0.846
Tramba (20 Epochs)	**0.774**	**0.940**	**0.852**	**0.845**

Table 6: Epoch-wise convergence of our model on structure recognition.

This rapid convergence reinforces the strength of the underlying Mamba encoder in learning document structure patterns with relatively few training iterations. Importantly, even by epoch 15, the model is already outperforming Faster R-CNN across all metrics, showing that with sufficient architectural design, shallow models can outperform deeper but less context-aware baselines.

The relatively early convergence also points to computational efficiency, making our approach suitable for deployment in low-resource environments or large-scale document processing systems. Given that the model achieves near-peak performance within 15 epochs and without extensive architectural depth, it can serve as a strong foundation for further fine-tuning, multi-task integration, or semi-supervised extensions.

In summary, our hybrid Vision Mamba-based model achieves robust performance in the table structure recognition task, outperforming Faster R-CNN significantly and showing promising results when compared to DETR. The bidirectional context modeling of Mamba enables strong structural alignment, especially at higher IoU thresholds. However, due to resource limitations, we trained a smaller model which limited the full realization of its potential.

The current findings suggest that even a shallow Vision Mamba encoder can deliver competitive structure recognition performance. With deeper configurations and additional training compute, it is plausible that this architecture could surpass DETR in structure-level tasks. Future work will focus on scaling the model depth, introducing architectural enhancements like multi-scale fusion or attention-SSM hybrids, and evaluating performance on more diverse real-world document collections.

4.3.1.3 Qualitative Analysis and Visual Comparisons

Visual Examples To better illustrate the comparative behavior of our model and the DETR baseline, we present a set of qualitative results on a shared document sample. The examples include visualizations from both the table detection and table structure recognition tasks.

In the DETR output image Figure 8, we observe that the model performs strong table localization, correctly identifying the bounding box around the tabular region with tight alignment. Also, the structure recognition part classifies the complex table structures correctly, making it a strong benchmark for such comparisons.

In our work, the table detection output Figure 9, from our model, shows equally accurate boundary localization. The detected bounding box tightly encloses the table, with high spatial fidelity to the ground truth. Also, it can successfully detect multiple tables on a page.

More importantly, our model's structure recognition output, Figure 10, demonstrates cleaner segmentation. Rows are consistently aligned, and columns are distinctly partitioned without visible overflow. The model captures spanning cells effectively, preserves hierarchical structure, and avoids collapsing neighboring cells. This suggests that the bidirectional state-space modeling offered by

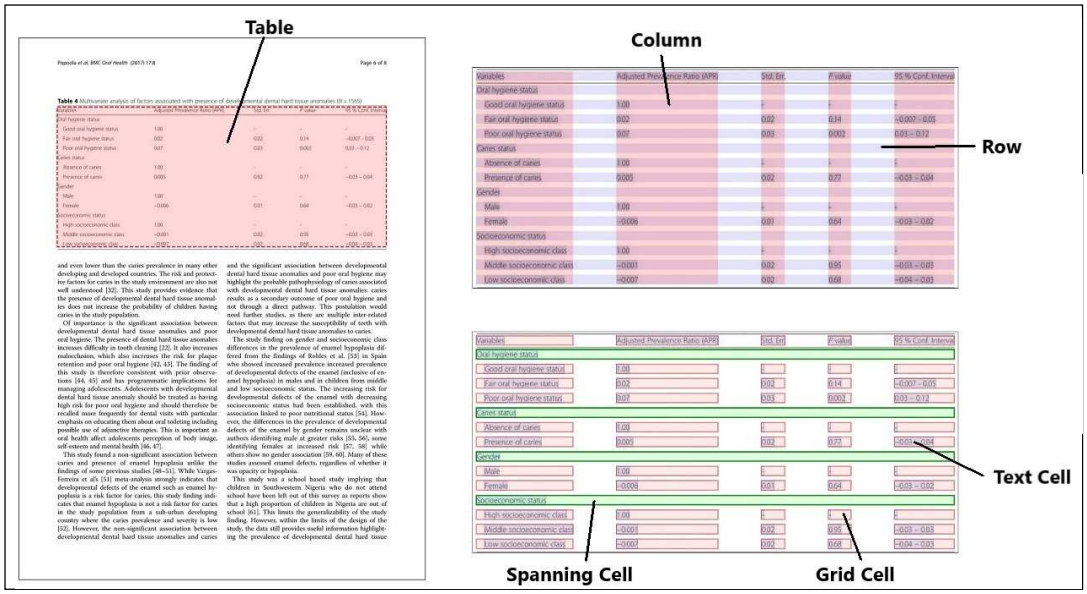

Figure 8: DETR results on real-world cases. This figure was taken from [80]

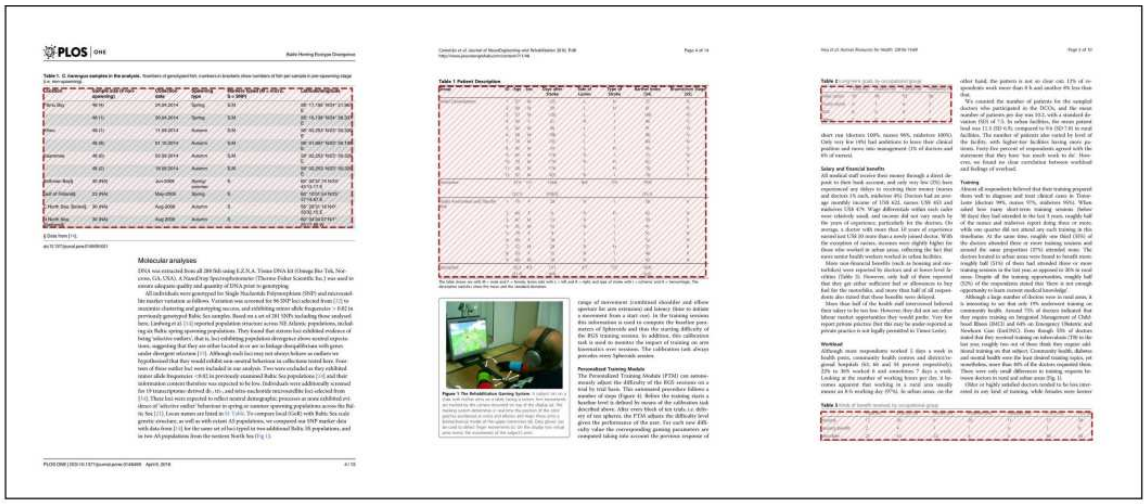

Figure 9: Detection results of our model on real-world cases

the Vision Mamba encoder helps encode both fine-grained local cues and long-range structural dependencies.

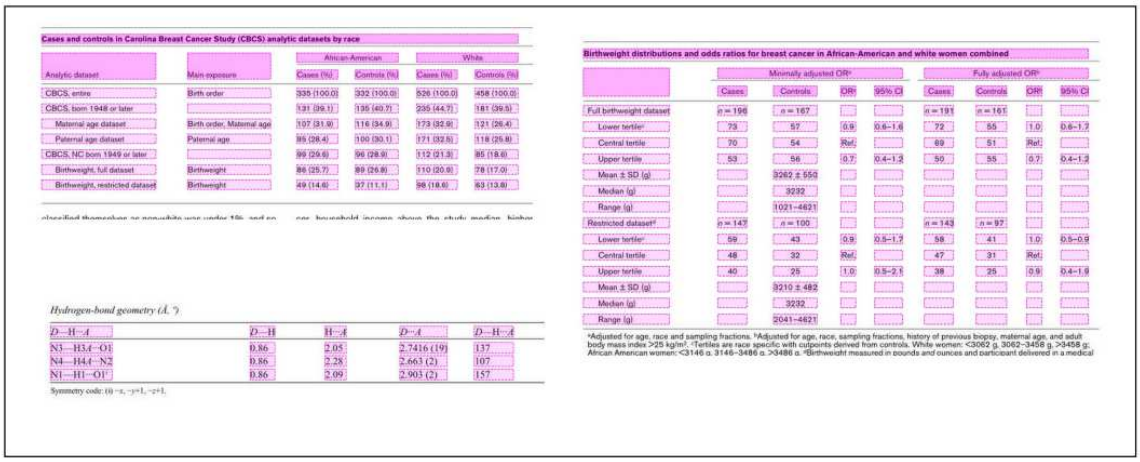

Figure 10: Structure recognition results of our model on real-world cases

These qualitative examples affirm that despite having a shallower architecture, our model is capable of producing cleaner and more coherent structural segmentation in certain real-world cases, especially where layout complexity and cell variability are high.

Failure Cases While our model demonstrates strong performance across both detection and structure recognition tasks, it is not without limitations. A closer examination of certain failure cases, Figure 11, reveals areas where the model underperforms, particularly in the structure recognition task. As illustrated in the examples above, one consistent pattern involves the model's difficulty in accurately localizing column headers — especially in tables with complex multi-row headers or irregular alignment.

In the first failure case, although the overall table is correctly detected and most cells are properly segmented, the column header rows are either partially missed or merged into a single bounding box. This results in downstream mislabeling of column positions and can significantly impact table interpretation, especially in scientific documents where headers convey categorical hierarchies.

In contrast, DETR Figure 8 appears to handle such scenarios more gracefully. Its transformer-based encoder, with deeper contextual modeling, is able to separate column header regions with better granularity. This allows DETR to assign cleaner and more distinct boundaries to multilevel headers, which our model sometimes merges due to its reduced depth and lack of explicit multi-scale reasoning.

These cases highlight the limitations of shallow architectures when dealing with hierarchical

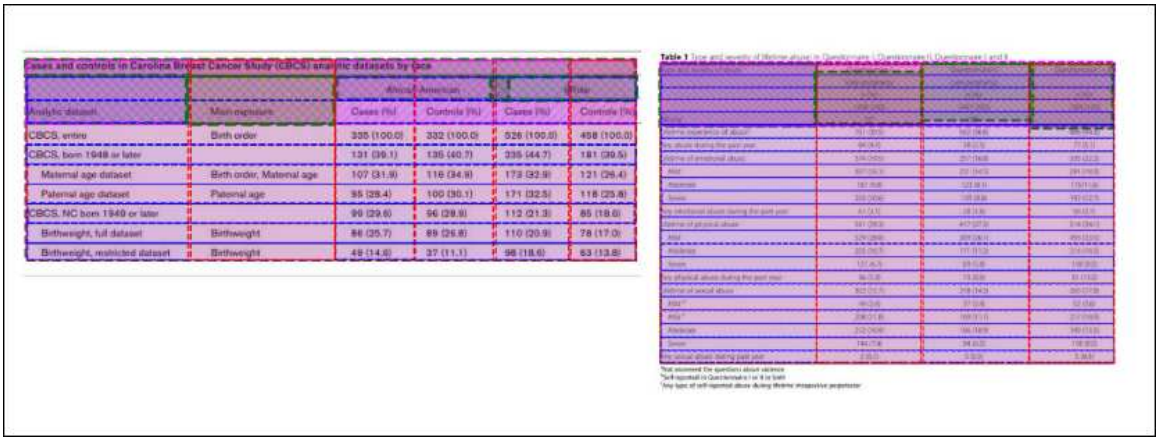

Figure 11: Failed cases in some real-world scenarios

structure. They suggest a potential future direction involving adaptive scaling or specialized header-aware modules to better handle structured semantic zones within tables.

4.3.2 Overall Summary and Experimental Outcome

Our experimental analysis provides a multifaceted understanding of how architectural modifications — specifically the integration of a Vision Mamba encoder into a DETR-style pipeline — influence performance across document-level table detection and structure recognition tasks. The results reveal a nuanced performance landscape shaped by trade-offs between model depth, structural precision, and training efficiency.

From a holistic standpoint, our model demonstrates that substantial performance gains can be achieved even with shallower architectures, as long as the inductive biases are closely aligned with the structure of the task. By leveraging the bidirectional sequence modeling capabilities of Vision Mamba, our hybrid architecture delivers high-quality table detection and structure recognition results. Notably, it surpasses the performance of Faster R-CNN and achieves results competitive with full DETR models, while utilizing only **one-third** of the encoder-decoder layers and requiring significantly fewer training epochs. This efficiency highlights the model's suitability for resource-constrained environments where deploying full-scale transformer stacks is impractical, without compromising on accuracy or generalization.

In structure recognition, while DETR maintains its advantage due to greater representational depth, our model consistently outperforms traditional region-based frameworks such as Faster R-CNN. More importantly, it achieves this while remaining robust to layout noise and variable structural patterns — an essential trait for real-world deployment. This outcome affirms the architectural

value of Mamba's state-space formulation, especially in modeling the hierarchical and long-range relationships intrinsic to table layouts.

Qualitative results further support this view, highlighting the model's strengths in preserving row-column integrity, while also exposing areas such as column header parsing that remain challenging. These insights do not point to fundamental flaws but instead serve as informed boundaries — indicating where future research can yield the most impact.

Taken together, our experiments underscore the viability of moving beyond attention-centric models in vision-based document understanding. By coupling lightweight yet expressive encoders with modular detection-decoder pipelines, we open up a design space that balances precision, generalization, and efficiency. The strong empirical foundation laid by our work provides ample direction for future research, particularly in scaling Mamba-based hybrids, integrating hierarchical priors, and advancing table parsing toward end-to-end semantic extraction.

Chapter 5

Conclusion and Future Work

In this chapter, we summarize the contributions made by the thesis, analyze its limitations, and outline potential avenues for future research. The research presented in this thesis addresses the growing demand for computationally efficient yet effective deep learning architectures for structured document understanding. By exploring hybridization between transformer-based models and state-space models like Vision Mamba, the proposed work demonstrates a promising direction in rethinking the backbone of modern vision architectures. The conclusions drawn are based on extensive empirical evaluation and architectural innovations, and they collectively form a foundation for further studies in this evolving space.

5.1 Contributions of the Thesis

This thesis proposes and evaluates a novel hybrid architecture for table understanding tasks, specifically targeting table detection and table structure recognition. The central contribution lies in the integration of Vision Mamba into the Detection Transformer (DETR) pipeline, replacing the conventional transformer encoder with a Mamba-based encoder-decoder framework. Through this architectural shift, we examine the potential of state-space models in structured vision tasks traditionally dominated by attention mechanisms.

To assess the effectiveness of the proposed framework, a series of experiments were conducted on the PubTables-1M dataset. The experiments demonstrate that the Mamba-based model achieves performance that rivals transformer-based counterparts like DETR while significantly outperforming classical architectures such as Faster R-CNN. For the table detection task, our model trained with only 10 epochs achieves an Average Precision (AP) of 0.950 and Average Recall (AR) of 0.970, exceeding the performance of Faster R-CNN (AP: 0.825, AR: 0.866) and closely matching DETR (AP: 0.966, AR: 0.981). Furthermore, with 20 epochs of training, the AP and AR metrics for

the Mamba model reach 0.956 and 0.977, respectively—essentially indistinguishable from DETR's results.

In the table structure recognition task, although the Mamba-based model does not surpass DETR, it still yields competitive results while outperforming Faster R-CNN. Specifically, a 20-epoch Mamba model achieves an AP of 0.774 and AR of 0.845, in comparison to DETR's AP of 0.912 and AR of 0.942. This suggests that even with reduced model depth and limited training epochs, the proposed hybrid architecture remains highly competitive.

Importantly, our Mamba-enhanced DETR architecture utilizes only 2 encoder and decoder layers, in stark contrast to the 6 layers employed by the original DETR configuration. This structural simplification, along with faster convergence observed during training, signifies the computational efficiency gained through the Vision Mamba integration. Notably, our architecture achieved near-saturation performance with just 10 training epochs, suggesting that Mamba's input-dependent and hardware-efficient design can enable quicker training and inference cycles.

Beyond empirical results, this thesis contributes to the theoretical understanding of integrating state-space models into transformer-style pipelines for vision tasks. We demonstrate that Mamba's ability to model long-range dependencies through linear-time recurrence mechanisms can be synergistically combined with the decoder-based object query formulation from DETR. This hybrid approach leverages the strengths of both paradigms—efficient sequence modeling from Mamba and global object reasoning from transformers.

The research also presents a new perspective on designing scalable vision architectures. Rather than relying exclusively on deep and complex transformer stacks, our work showcases that shallow Mamba-based encoders can achieve comparable results, indicating a promising direction for building more efficient vision models with fewer parameters and lower computational costs.

5.2 Limitations

Despite the encouraging outcomes, the proposed framework has several limitations that merit critical reflection. These limitations not only delineate the boundaries of our current implementation but also illuminate directions for future work.

First, due to limited computational resources, our model was trained using only 2 encoder and decoder layers, whereas the standard DETR architecture typically utilizes 6 layers in both encoder and decoder stacks. While the results with 2 layers were competitive, it is reasonable to hypothesize that performance could further improve with deeper architectures, particularly in the structure recognition task where DETR still holds a significant advantage.

Second, although our model converged faster—achieving strong performance within 10 epochs—the

inability to conduct extended training schedules due to computational constraints might have hindered the architecture's full potential. Training the model for more epochs or incorporating a more extensive hyperparameter sweep could lead to better generalization and improved performance across tasks.

Third, the evaluation was limited to the PubTables-1M dataset and focused solely on table detection and structure recognition tasks. While these tasks are representative of structured document understanding, the generalizability of the proposed hybrid architecture to other visual document understanding tasks (e.g., form field extraction, key-value pair detection) remains unverified. Future work must evaluate the adaptability of the proposed model across diverse document understanding benchmarks to validate its robustness and versatility.

Fourth, although Mamba offers linear-time complexity in theory, its implementation remains nontrivial and still exhibits considerable computational overhead in practice. Specifically, the operations involved in selective scanning and input-dependent parameterization, while more efficient than self-attention, are still resource-intensive when scaled to high-resolution inputs. This challenges the perception of Mamba as a lightweight model and calls for further engineering optimization to make it truly deployable in low-resource or real-time environments.

Additionally, the current model does not address dynamic or streaming visual inputs. Vision Mamba and DETR were originally designed for static images, and extending this work to handle temporally evolving data (e.g., document sequences or videos) would require architectural modifications and temporal consistency mechanisms. This represents another avenue where the model's applicability is currently constrained.

Interpretability also remains a concern. Like many deep learning architectures, the decision-making process of the proposed hybrid model lacks transparency. Understanding which components contribute most to prediction outcomes, and under what circumstances, remains a challenge. This is particularly important for sensitive applications in domains like finance, healthcare, or legal analysis, where model decisions must be interpretable and auditable.

Finally, while this work successfully demonstrates a proof-of-concept for integrating Vision Mamba into the DETR pipeline, it stops short of establishing a full theoretical framework or formal analysis. A more rigorous theoretical characterization of how Mamba's state-space dynamics interact with the attention mechanisms of the DETR decoder could uncover deeper insights and lead to even more optimized hybrid architectures.

These limitations, while notable, do not undermine the value of the proposed research. Instead, they offer fertile ground for future exploration and refinement. As discussed in the next section, extending the model's depth, breadth of evaluation, and architectural sophistication holds promise

for advancing both the performance and applicability of hybrid vision models in document understanding.

5.3 Future Work

While the proposed hybrid architecture demonstrates promising results in table detection and structure recognition, several directions remain unexplored that could further enhance its capabilities and broaden its applicability. Future work will focus on integrating vision-language modeling through Large Language Models (LLMs), extending the architecture to complex domains like medical document processing, and deploying the model in real-world applications that demand multimodal reasoning and robust generalization. These avenues offer the potential to elevate both the semantic understanding and practical impact of hybrid architectures. Several interesting research directions, motivated by this thesis, are discussed below:

5.3.1 Hybrid Vision-Language Architectures with LLM Integration

The convergence of computer vision and large language models (LLMs) has given rise to a new class of hybrid architectures capable of performing complex multimodal reasoning. These models combine the visual perception capabilities of deep vision encoders with the generative and semantic understanding strengths of LLMs, enabling systems to process, interpret, and respond to richly structured visual and textual data. This fusion has proven especially beneficial in domains like document analysis, robotics, medical diagnostics, and scientific literature understanding, where high-level reasoning must be grounded in visual evidence.

Building upon the modularity of our proposed DETR-Mamba hybrid architecture, integrating an LLM component into the decoding pipeline presents a promising extension. In this configuration, the Vision Mamba encoder could extract rich spatial and structural representations from input documents or images, while the LLM could generate textual outputs such as captions, summaries, answers to queries, or relational inferences. This architecture would allow for natural language interfaces over structured visual content, enabling intuitive and interpretable downstream applications such as visual question answering (VQA), document grounding, or regulatory compliance checking.

Beyond enhanced expressiveness, the modular hybrid design offers advantages in training efficiency and adaptability. Vision components can be fine-tuned on domain-specific visual tasks, while the language model remains pretrained and fixed, reducing compute cost and overfitting risk. Moreover, aligning transformer-free models like Mamba with transformer-based LLMs opens new opportunities for latency-aware deployment, especially in edge or cloud environments where compute constraints vary. This synergy sets a compelling precedent for developing scalable, intelligent, and

language-accessible AI systems.

5.3.2 Medical Document Processing and Multimodal Reasoning

Recent advances in vision-language modeling have significantly improved the capacity to extract structured information and enable domain-specific reasoning from complex multimodal data. Notably, the Med-R1 framework introduces a reinforcement learning-enhanced VLM capable of high generalization across diverse medical imaging modalities such as CT, MRI, and Ultrasound, and across tasks like lesion grading and anatomy identification. Instead of relying on large-scale supervised fine-tuning, Med-R1 employs Group Relative Policy Optimization (GRPO), enabling scalable reward-guided learning even in the absence of high-quality Chain-of-Thought annotations. This approach reduces reliance on costly expert annotations while boosting clinical coherence and interpretability.

This line of work aligns well with the possibilities explored in our research. Specifically, our hybrid DETR-Mamba-based architecture provides a foundation for extending structured information extraction beyond visual tasks into multimodal domains like healthcare, where visual content (e.g., medical scans, histopathology images) must be fused with textual descriptions or EHR narratives. Integrating reinforcement-driven optimization into our framework could enhance generalization in clinical settings, particularly for downstream tasks such as form parsing, patient case triaging, and anomaly detection in longitudinal medical records.

Furthermore, by drawing inspiration from Med-R1's strategy of "No-Think" inference—which emphasizes direct, high-confidence predictions without verbose reasoning—we may adapt our system for medical scenarios where interpretability and reliability are paramount. Combining our encoder-decoder vision architecture with lightweight RL-driven language modules opens a pathway for efficient multimodal document parsing and VQA in digital health, providing a compelling direction for real-world medical AI systems.

5.3.3 Applications

The hybrid architecture proposed in this thesis opens up several promising avenues for real-world deployment and integration across various industries. One significant direction is the extension of this model into a vision-language architecture by incorporating large language models (LLMs). This integration would enable multimodal understanding of complex documents, facilitating tasks such as key information extraction, natural language summarization of tabular data, and end-to-end document question answering systems. Such capabilities are especially valuable in legal, financial, and healthcare sectors where structured and unstructured data coexist.

Moreover, the lightweight yet powerful nature of the proposed architecture makes it suitable for processing large volumes of industrial documents at scale. Enterprises dealing with massive data streams—such as manufacturing companies, logistics providers, and data centers—can leverage this model for automated document classification, invoice parsing, and intelligent form analysis, thereby improving operational efficiency.

Other potential applications of this work include integration into edge devices for on-site document processing, deployment within smart OCR pipelines in enterprise-level software, and real-time document analytics in cloud-based platforms. In collaboration with **ERA Environmental Software Solutions**, this architecture is being explored for industrial use in automating complex document workflows, such as regulatory compliance and material safety data extraction. With further refinement, the proposed hybrid model has the potential to serve as a foundational component for the next generation of document AI systems—offering a compelling combination of scalability, efficiency, and cross-modal reasoning. We also plan to publish this work in a computer vision or document intelligence conference, further contributing to the academic and applied research communities.

Bibliography

[1] Ethan Baron, Itamar Zimerman, and Lior Wolf. 2-d ssm: A general spatial layer for visual transformers. *arXiv preprint arXiv:2306.06635*, 2023.

[2] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

[3] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tasırlar. Introducing our multimodal models, 2023. *URL https://www. adept. ai/blog/fuyu-8b*, 2, 2023.

[4] Alan F Blackwell. The reification of metaphor as a design tool. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(4):490–530, 2006.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[9] Hu Chen, Mingzhe Sun, and Eckehard Steinbach. Compression of bayer-pattern video sequences using adjusted chroma subsampling. *IEEE transactions on circuits and systems for video technology*, 19(12):1891–1896, 2009.

[10] Jieshan Chen, Mulong Xie, Zhenchang Xing, Chunyang Chen, Xiwei Xu, Liming Zhu, and Guoqiang Li. Object detection for graphical user interface: Old fashioned or deep learning or a combination? In *proceedings of the 28th ACM joint meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1202–1214, 2020.

[11] Xin Chen, Liang Xu, et al. Contrast: A hybrid cnn-transformer-mamba network for image super-resolution. *arXiv preprint*, 2024.

[12] Xiuwei Chen, Sihao Lin, Xiao Dong, Zisheng Chen, Meng Cao, Jianhua Han, Hang Xu, and Xiaodan Liang. Transmamba: Fast universal architecture adaption from transformers to mamba. *arXiv preprint arXiv:2502.15130*, 2025.

[13] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019.

[14] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. URL `https://arxiv.org/abs/1904.10509`.

[15] Krzysztof Choromanski, Haoxian Chen, Han Lin, Yuanzhe Ma, Arijit Sehanobish, Deepali Jain, Michael S Ryoo, Jake Varley, Andy Zeng, Valerii Likhosherstov, Dmitry Kalashnikov, Vikas Sindhwani, and Adrian Weller. Hybrid random features, 2022. URL `https://arxiv.org/abs/2110.04367`.

[16] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

[17] 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. Ieee, 2009.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[19] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[20] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

[21] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12124–12134, 2022.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford, 2012. ISBN 9780199641178. URL `https://books.google.ca/books?id=fOq39Zh0olQC`.

[24] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pages 2286–2296. PMLR, 2021.

[25] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, June 2023.

[26] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

[27] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. A methodology for evaluating algorithms for table understanding in pdf documents. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 45–48, 2012.

[28] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[29] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[30] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 951–967, 2022.

[31] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jianying Hu, Ramanujan S Kashi, Daniel P Lopresti, and Gordon Wilfong. Table structure recognition and its evaluation. In *Document Recognition and Retrieval VIII*, volume 4307, pages 44–55. SPIE, 2000.

[34] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR, 2024.

[35] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[36] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023.

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning*

Research, pages 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

[38] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

[39] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[40] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020. URL https://arxiv.org/abs/2001.04451.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[42] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[43] Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33:18470–18481, 2020.

[44] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[45] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[46] M Li, L Cui, S Huang, F Wei, M Zhou, and Z TableBank Li. Table benchmark for image-based table detection and recognition. arxiv 2019. *arXiv preprint arXiv:1903.01949*.

[47] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1922–1960, 2023.

[48] Zhu Lianghui, Liao Bencheng, Zhang Qian, Wang Xinlong, Liu Wenyu, and Wang Xinggang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv. Org*, pages arXiv–org, 2024.

[49] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

[50] 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.

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[52] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), January 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL `https://doi.org/10.1145/3560815`.

[53] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[54] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[55] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[56] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[58] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[59] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.

[60] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

[61] Ermelinda Oro and Massimo Ruffolo. Trex: An approach for recognizing and extracting tables from pdf documents. In *2009 10th international conference on document analysis and recognition*, pages 906–910. IEEE, 2009.

[62] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018.

[63] Eunbyung Park and Alexander C Berg. Learning to decompose for object detection and instance segmentation. *arXiv preprint arXiv:1511.06449*, 2015.

[64] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[65] David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242, 2003.

[66] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

[67] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[69] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

[70] Farheen Ramzan, Muhammad Usman Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, 44, 12 2019. doi: 10.1007/s10916-019-1475-2.

[71] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[72] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[73] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6656–6664, 2017.

[74] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[75] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[76] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 312–329. Springer, 2016.

[77] Amaia Salvador, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i Nieto. Recurrent neural networks for semantic instance segmentation. *arXiv preprint arXiv:1712.00617*, 2017.

[78] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1162–1167. IEEE, 2017.

[79] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[80] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.

[81] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[82] Liang Sun, Xin Li, et al. Hybrid state-space and attention architectures for robotic vision. *arXiv preprint arXiv:2409.00410*, 2024.

[83] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[85] Dong Wang, Zixiang Wang, Ling Chen, Hongfeng Xiao, and Bo Yang. Cross-parallel transformer: Parallel vit for medical image segmentation. *Sensors*, 23(23):9488, 2023.

[86] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[87] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[88] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[89] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8239–8249, 2024.

[90] Juntao Zhang, Shaogeng Liu, Kun Bian, You Zhou, Pei Zhang, Wenbo An, Jun Zhou, and Kun Shao. Vim-f: Visual state space model benefiting from learning in the frequency domain. *arXiv preprint arXiv:2405.18679*, 2024.

[91] Long Zhang and Yi Wan. Tranmamba: a lightweight hybrid transformer-mamba network for single image super-resolution. *Signal, Image and Video Processing*, 19(5):1–12, 2025.

[92] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021.

[93] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020.

[94] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[95] Tao Zhou, Yuhu Du, Jingjing Mao, Caiyue Peng, Hongwei Wang, and Zhongwei Zhou. Parallel attention multi-scale mandibular fracture detection network based on centernet. *Biomedical Signal Processing and Control*, 95:106338, 2024.

[96] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[97] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.