A Survey on Document Layout Analysis and Related Techniques

www.surveyx.cn

Abstract

Document Layout Analysis (DLA) is a critical field that transforms unstructured documents into structured, machine-readable formats, pivotal for applications like automated data entry, information retrieval, and document management. This survey investigates the integration of deep learning, computer vision, and natural language processing within DLA, emphasizing their collective impact on enhancing document processing capabilities. The survey highlights innovative frameworks such as Paragraph2Graph and LiLT, which address existing limitations through language-independent solutions. Challenges such as data scarcity, model generalization, and computational complexity are explored, with a focus on overcoming these hurdles to advance DLA technologies. The survey also underscores the significance of benchmark datasets and evaluation metrics in driving innovation, facilitating the development of more accurate and efficient models. Applications of DLA across various domains, including automated data entry, information retrieval, and document management, are examined, showcasing its transformative impact. The survey concludes by identifying areas for future research, emphasizing the need for comprehensive datasets and novel methodologies to handle complex document layouts. Through this comprehensive overview, the survey aims to provide insights into the current landscape and future prospects of DLA, highlighting its indispensable role in modern document analysis.

1 Introduction

1.1 Significance of Document Layout Analysis

Document Layout Analysis (DLA) is essential for converting unstructured digital documents into structured, machine-readable formats, facilitating applications in automated data entry, information retrieval, and document management [1]. DLA plays a critical role in addressing challenges in automatic document understanding, especially for historical documents with varied layouts [2]. The demand for effective layout analysis techniques is particularly evident in extracting high-quality text from formatted PDFs, which often exhibit inconsistencies due to diverse publishing tools [3].

In medical imaging, DLA enhances classification and segmentation performance, vital for disease quantification and treatment evaluation [4]. The development of deep learning methods for complex biomedical image segmentation [5] and the application of Transformers in this domain have further propelled advancements [6]. DLA's importance extends to understanding both the physical layout and logical structure of documents, aiding in information retrieval, document summarization, and knowledge extraction [7].

The integration of deep learning into DLA has shown substantial promise in transforming data-driven decision-making across industries [8]. However, challenges such as model generalizability, particularly in scientific literature, persist [9]. Adapting CNN architectures and preprocessing techniques for document images is crucial for enhancing the accuracy and efficiency of DLA systems [10].

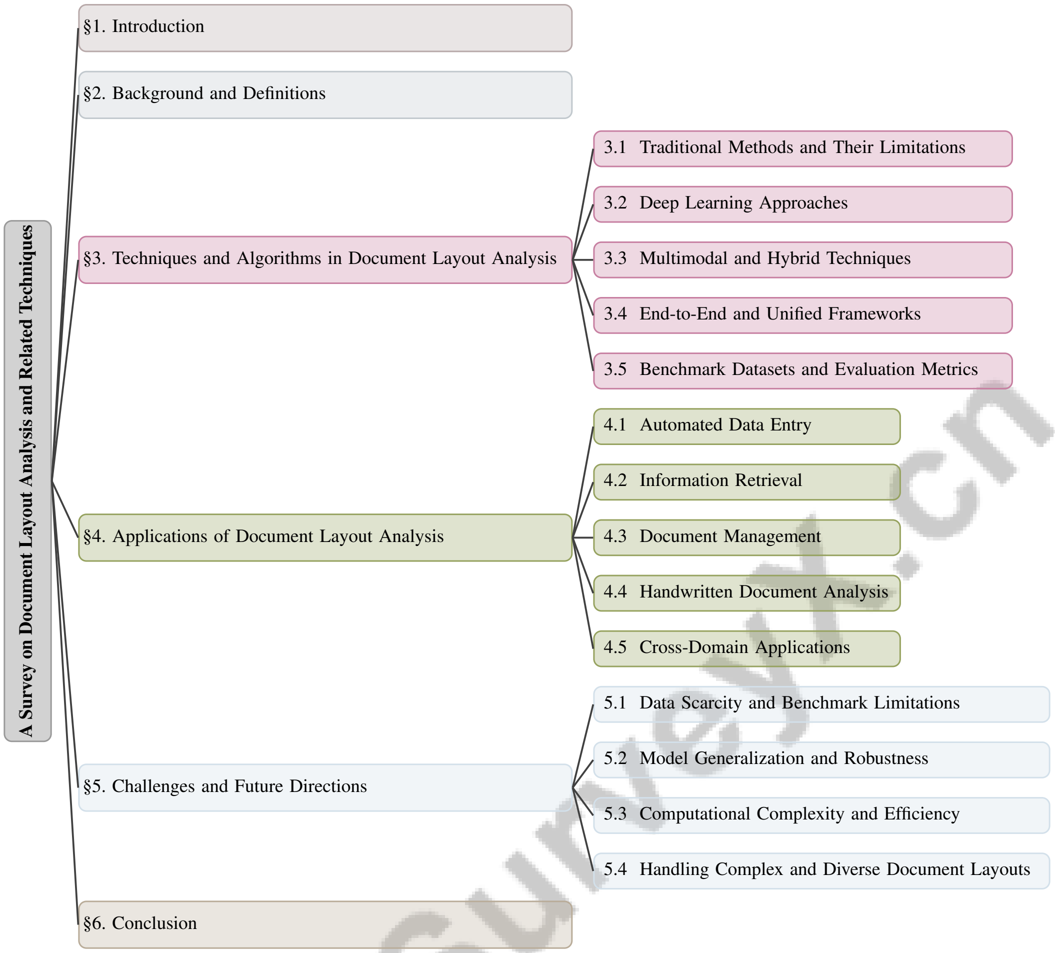

A Survey on Document Layout Analysis and Related Techniques
§1. Introduction
§2. Background and Definitions
§3. Techniques and Algorithms in Document Layout Analysis
3.1 Traditional Methods and Their Limitations
3.2 Deep Learning Approaches
3.3 Multimodal and Hybrid Techniques
3.4 End-to-End and Unified Frameworks
3.5 Benchmark Datasets and Evaluation Metrics
§4. Applications of Document Layout Analysis
4.1 Automated Data Entry
4.2 Information Retrieval
4.3 Document Management
4.4 Handwritten Document Analysis
4.5 Cross-Domain Applications
§5. Challenges and Future Directions
5.1 Data Scarcity and Benchmark Limitations
5.2 Model Generalization and Robustness
5.3 Computational Complexity and Efficiency
5.4 Handling Complex and Diverse Document Layouts
§6. Conclusion

Figure 1: chapter structure

DLA is also significant in forensic applications, where determining the authenticity of digital images is critical due to the prevalence of image manipulation tools [11]. In digital humanities, the increasing interest in historical document analysis has led to a demand for efficient tools [12]. Moreover, while DLA is vital for interpreting complex documents, existing methods often struggle with language dependency and the processing of lengthy documents [13].

Recent literature emphasizes the importance of annotating document images to enhance the generation of high-quality data for recognition tasks [14]. The superiority of deep learning technologies over traditional methods in complex tasks further underscores DLA's significance [15]. The necessity for a novel approach that integrates vision, semantics, and relational understanding is evident, given the limitations of previous methods [16].

1.2 Interconnectedness of Related Fields

DLA is closely intertwined with deep learning, computer vision, and natural language processing (NLP), forming a synergistic framework that enhances document processing capabilities. This interconnectedness is crucial for developing advanced models that effectively analyze and interpret document structures. The Paragraph2Graph framework exemplifies this fusion by integrating image features and Optical Character Recognition (OCR) data into graph structures, facilitating effective layout analysis across languages [13].

Deep learning serves as a foundational element in this integration, significantly augmenting DLA through various learning paradigms, including supervised, unsupervised, and reinforcement learning [15]. The application of Convolutional Neural Networks (CNNs) for automated feature extraction and architectures like Faster R-CNN and Mask R-CNN illustrate the convergence of DLA with computer vision techniques. Despite challenges such as handling diverse layouts and small-scale text areas, these methods continue to evolve, pushing the boundaries of document analysis.

The mutual inclusion mechanism, exemplified by MIPC-Net, which combines global and local features for precise analysis, further underscores the synergy between deep learning and medical image analysis. This trend towards integrating various data modalities—visual and textual features—enhances analytical precision in DLA. This broader movement towards multimodal frameworks has demonstrated improved performance across diverse applications, including semantic content extraction and handling complex layouts in scientific literature and historical newspapers. By leveraging visual cues and contextual information, researchers achieve significant advancements in model efficiency and generalizability, facilitating more accurate document processing and information retrieval [9, 17, 18, 3, 19].

1.3 Objectives of the Survey

This survey aims to provide a comprehensive overview of current advancements and methodologies in DLA and its related fields. By examining the integration of deep learning, computer vision, and natural language processing, the survey elucidates the interconnected nature of these domains and their collective impact on enhancing document processing capabilities [8]. A particular focus is on innovative frameworks, such as LiLT, which address limitations in existing structured document understanding models through language-independent solutions [20].

Additionally, the survey addresses challenges posed by high-dimensional data in clustering applications, emphasizing the need for improved feature representation to enhance clustering performance [21]. In the context of scientific literature, it highlights challenges in extracting information from lengthy and specialized documents, underscoring the necessity for advanced extraction methods adaptable to academic contexts [22].

Furthermore, the survey bridges knowledge gaps in the application of Transformers compared to traditional CNNs, particularly in capturing global context in medical images [6]. This endeavor aims to facilitate further research and development, contributing to the advancement of DLA technologies across diverse domains.

Finally, the survey emphasizes the importance of specialized tools like Callico, which enhance document annotation quality, thereby advancing research in document analysis, especially within digital humanities [14]. Through these objectives, the survey not only synthesizes current trends and technologies but also identifies areas for future research and development in document layout analysis.

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive exploration of DLA and its interconnected fields. It begins with an **Introduction**, highlighting DLA's significance in modern applications and its relationship with deep learning, computer vision, and natural language processing. This section outlines the main objectives and provides a roadmap for the paper's organization.

Following the introduction, the **Background and Definitions** section defines core concepts and terminologies, offering foundational understanding of key terms such as page segmentation and optical character recognition, while tracing the historical development of DLA and its related fields.

The third section, **Techniques and Algorithms in Document Layout Analysis**, examines various methodologies in DLA, contrasting traditional methods with modern deep learning approaches, exploring multimodal and hybrid techniques, and discussing end-to-end frameworks for comprehensive document analysis. It also highlights benchmark datasets and evaluation metrics crucial for assessing DLA performance.

In the **Applications of Document Layout Analysis** section, the survey explores practical applications across domains, providing real-world examples of DLA enhancing automated data entry, information retrieval, document management, and handwritten document analysis. It also discusses cross-domain applications, illustrating DLA's versatility.

AI-generated, for reference only.

The penultimate section, **Challenges and Future Directions**, identifies current challenges such as data scarcity, model generalization, and computational complexity, while discussing potential future directions to advance DLA.

Finally, the **Conclusion** summarizes key findings and insights, reiterating the significance of automated information extraction in scientific literature and revealing promising avenues for future inquiries and technological innovations, particularly in improving DLA and data extraction methods for scientific PDFs [3, 22]. This structured approach aims to provide a holistic view of the current landscape and future prospects of Document Layout Analysis.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts and Terminologies

Document Layout Analysis (DLA) involves key concepts essential for processing document structures, with *page segmentation* being a fundamental task that partitions documents into regions like text, images, and tables. This segmentation is crucial for *Optical Character Recognition* (OCR), which converts scanned documents into machine-readable text, especially in complex layouts where precise segmentation is vital [23, 2].

Deep learning has revolutionized DLA by enhancing feature extraction and classification through Convolutional Neural Networks (CNNs), which improve object detection and segmentation of intricate layouts [10, 24]. Despite these advancements, the generalizability of models remains a challenge, as they may not perform well across diverse document types [9].

Instance segmentation, as seen in the BaDLAD dataset, emphasizes the need for precise segmentation into meaningful units such as text boxes and tables, crucial for identity document analysis [25, 26]. The MIDV-500 benchmark illustrates the necessity for accurate segmentation in complex layouts.

Semantic structure extraction is another critical aspect, involving the classification of units like tables and paragraphs [27]. Token-level annotations and sequence labeling support tasks like information retrieval and summarization. However, existing benchmarks often fall short in evaluating all aspects of document structure analysis, highlighting performance assessment gaps [7].

The recognition of scripts like Kuzushiji, with multiple forms and limited understanding, showcases the complexities in DLA [28]. The UDIAT dataset addresses the segmentation of semantic regions in ancient manuscripts with complex layouts and degradation [29].

DLA integrates page segmentation, deep learning, OCR, and semantic structure extraction to enhance document layout analysis. By utilizing advanced algorithms and neural networks, DLA evolves to provide innovative solutions for complex document tasks across diverse domains [16].

2.2 Historical Development and Evolution

The evolution of Document Layout Analysis (DLA) has been shaped by technological advancements and increasingly complex document formats. Early DLA methods relied on basic image processing, which were manual and limited in handling diverse document types, lacking the sophistication needed for comprehensive analysis [9].

The introduction of machine learning marked a shift towards more automated and accurate analysis, further advanced by deep learning, which brought robust feature extraction and classification techniques. These developments have significantly enhanced the handling of complex layouts, forming the basis for modern methodologies using neural networks for improved document interpretation [30].

The creation of comprehensive benchmarks has been crucial for DLA's evolution, addressing the historical lack of publicly available datasets for model training and evaluation. The MIDV-500 benchmark addresses identity document recognition challenges in mobile video streams, highlighting the need for real-world variability in datasets [26]. The UDIAT dataset provides a pixel-precise benchmark for layout analysis, supporting both computer scientists and humanities scholars [29].

Challenges persist, particularly regarding training data availability and computational resources for sophisticated implementations. The loss of contextual information in traditional 2D approaches complicates DLA system development, requiring innovative solutions that balance computational efficiency with analytical precision [31].

The ongoing adaptation of DLA methodologies underscores the importance of continuous research and development. The integration of synthetic data generation and heuristic methods with deep learning exemplifies efforts to overcome limitations and enhance model generalizability [9]. As new benchmarks and datasets emerge, DLA is poised for further advancements, ensuring the applicability of its technologies across varied document types and applications.

3 Techniques and Algorithms in Document Layout Analysis

Category	Feature	Method
Traditional Methods and Their Limitations	User-Centric Information Extraction	PKEM[32]
Deep Learning Approaches	Adaptive Methods	MLST[33], nnU-Net[5]
	Contextual and Spatial Analysis	FRCF[3]
	Image Segmentation Focus	FCN[23]
Multimodal and Hybrid Techniques	Data Combination Techniques	P2G[13]
End-to-End and Unified Frameworks	Integrated Approaches	UD[34], VTLayout[35], MESc[36], UMTD[37], LP[38], DL-GDD[17], SLLD[39], DLAFormer[40], DAT[41], DGA[42], DANIEL[43]
Benchmark Datasets and Evaluation Metrics	Dataset Utilization	DS[44], VSR[16], N/A[14]

Table 1: This table provides a comprehensive summary of various methodologies in Document Layout Analysis (DLA), categorizing them into traditional methods, deep learning approaches, multimodal and hybrid techniques, end-to-end and unified frameworks, and benchmark datasets. Each category is associated with specific features and methods, highlighting the evolution and advancements in DLA techniques. The table also references key studies and frameworks that have contributed to these developments.

Understanding the evolution of techniques and algorithms in Document Layout Analysis (DLA) is essential for appreciating advancements in this field. This section examines various methodologies employed in DLA, starting with traditional methods and their limitations, and transitioning to more advanced strategies that address contemporary document complexities. Table 1 presents a detailed categorization of methodologies employed in Document Layout Analysis, illustrating the progression from traditional techniques to advanced frameworks and their associated features and methods. Table 5 provides a detailed categorization of methodologies employed in Document Layout Analysis, illustrating the progression from traditional techniques to advanced frameworks and their associated features and methods. **??** illustrates the hierarchical categorization of techniques and algorithms in DLA, highlighting the evolution from traditional methods to advanced deep learning, multimodal, hybrid, and unified frameworks. This figure emphasizes the integration of diverse modalities and tasks, as well as the critical role of benchmark datasets and evaluation metrics in advancing DLA methodologies.

3.1 Traditional Methods and Their Limitations

Traditional DLA methods primarily relied on heuristic rule-based approaches, which used manually defined rules for document understanding. These foundational methods often struggled with adapting to diverse and complex document layouts [45]. Their reliance on predefined rules limited flexibility and scalability, making them less effective in handling document variability. Figure 2 illustrates these traditional Document Layout Analysis (DLA) methods, highlighting their heuristic rule-based approaches, limitations in flexibility and adaptability, and recent innovations aimed at improving accuracy and efficiency.

A notable limitation was the separate handling of scene text detection and layout analysis, which restricted efficiency and accuracy by failing to leverage task interconnections [34]. Traditional methods also inadequately recognized categories like Lists and Titles due to their unique physical presentation [35]. The categorization of local features based on detectors and descriptors emphasized robustness in image retrieval tasks but fell short in dealing with historical document layouts, prompting the development of specialized tools like LAREX [46].

Furthermore, traditional keyword extraction techniques often overlooked individual user context, resulting in less relevant information extraction [32]. While traditional methods established a foundation for DLA, their limitations in flexibility, adaptability, and contextual understanding necessitated more advanced approaches. Innovations like a hybrid Transformer-based object detection network have significantly improved accuracy in identifying and classifying document elements, achieving high precision scores across various datasets [47, 48].

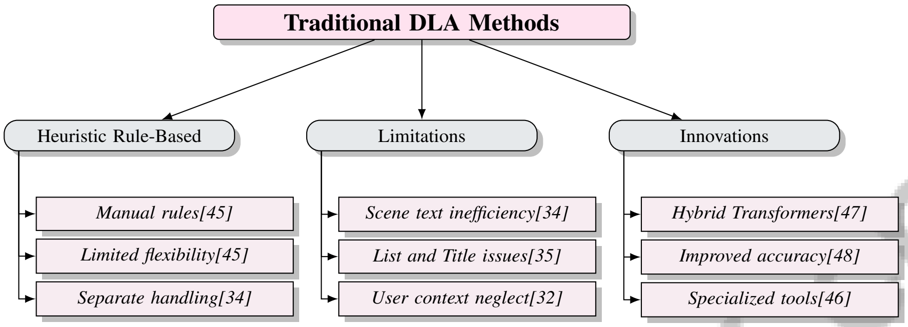


Figure 2: This figure illustrates the traditional Document Layout Analysis (DLA) methods, highlighting their heuristic rule-based approaches, limitations in flexibility and adaptability, and recent innovations to improve accuracy and efficiency.

3.2 Deep Learning Approaches

Method Name	Architectural Innovations	Model Adaptability	Application Domains
nnU-Net[5]	U-Net Architecture	Automates Adaptation	Biomedical Image Segmentation
P2G[13]	Graph Neural Networks	Language-independent	Layout Analysis
MLST[33]	Modular Learning Architectures	Modular Neural Networks	Classification And Segmentation
FRCF[3]	Contextual Information Incorporation	Faster R-CNN Adaptation	Object Detection Segmentation
FCN[23]	Convolutional Architecture	-	Object Detection
DS[44]	Deep Generative Models	Latent Variable Modeling	Document Image Synthesis

Table 2: This table provides a comparative analysis of various deep learning methods applied in document layout analysis (DLA). It highlights the architectural innovations, model adaptability, and application domains of each method, showcasing the diverse capabilities and advancements in the field.

Deep learning has transformed DLA by enhancing precision and scalability in document processing. Neural networks have addressed layout analysis complexities through novel architectures and training strategies. The nnU-Net framework automates deep learning adaptation to diverse datasets, showcasing its potential in DLA [5]. Table 2 presents a detailed comparison of deep learning approaches that have significantly contributed to advancements in document layout analysis by introducing novel architectures and enhancing model adaptability across various application domains.

Transformer-based models, like the Hybrid Transformer-based Document Layout Analysis (HTDLA), employ object detection networks with query encoding and hybrid matching strategies, effectively capturing intricate document structures [6]. The Paragraph2Graph framework uses graph neural networks (GNNs) to incorporate image features and spatial text coordinates, enhancing layout analysis robustness [13].

Deep learning methodologies, especially CNNs, have outperformed traditional approaches in effectiveness and efficiency, as seen in applications like 3D bounding box detection [31]. Modular neural network architectures improve performance by decomposing tasks into manageable components, enhancing adaptability and scalability [33]. The adaptation of Faster R-CNN for visual segmentation of scientific articles illustrates the application of object detection techniques to identify and classify document regions, leveraging contextual information for accuracy [3].

The FCN approach has been pivotal in segmenting historical document images, focusing on foreground pixel predictions [23]. The DocSynth framework synthesizes document images by learning from spatial layouts, facilitating better generalization and performance evaluation [44]. Taye et al.'s survey emphasizes deep learning's effectiveness in managing large, complex datasets, surpassing traditional techniques in handling modern document layouts [15].

3.3 Multimodal and Hybrid Techniques

Multimodal and hybrid techniques in DLA address traditional and single-modality limitations by leveraging diverse data types such as text, images, and structural layout information, enhancing robustness and accuracy. The Paragraph2Graph framework demonstrates the efficacy of combining modalities for comprehensive layout analysis [13].

Multimodal approaches automate feature extraction, improving performance across tasks like image classification and natural language processing. Advanced frameworks integrating Document Layout Generators (GLD), Document Elements Decorators (GED), and Document Style Discriminators (DSD) create sophisticated systems for nuanced document analysis [49, 47, 3, 17].

Hybrid techniques extend beyond modality integration by incorporating advanced architectures like Mask-RCNN for handling overlapping regions and complex layouts. Deep recurrent neural networks (RNNs), with LSTM cells, enhance sequence modeling by addressing the vanishing gradient problem, improving performance in applications like NLP and information retrieval [50, 39, 8, 51, 52].

Modular learning frameworks highlight hybrid techniques' effectiveness, integrating self-training and multi-agent collaboration models to improve information extraction accuracy and adaptability across domains [53, 45, 36, 54, 22]. These approaches facilitate key sample identification and labeling, iteratively updating models for improved performance in dynamic document environments.

The integration of multimodal and hybrid techniques propels DLA evolution, providing innovative solutions for complex document processing tasks. Graph-based models like GLAM and transformer-based approaches like DLAFormer enhance classification and segmentation efficiency, leveraging rich metadata and unifying sub-tasks within a single framework [40, 55, 18]. These advancements improve accuracy and efficiency, expanding DLA's scope and paving the way for further technological progress.

3.4 End-to-End and Unified Frameworks

Method Name	Integration Approach	Task Handling	Technological Application
LP[38]	Unified Toolkit	Layout Detection	Deep Learning
DANIEL[43]	End-to-end Architecture	Layout Analysis, Handwriting Recognition	Convolutional Encoder, Transformer
UD[34]	Unified Approach	Simultaneously Detect Scene	End-to-end Model
VTLayout[35]	Feature Fusion	Dual-stage Process	Transformer-based Methods
UMTD[37]	Unified Framework	Manage Sub-tasks	Deep Learning
DGA[42]	Heterogeneous Features	Layout Recognition	Graph Convolutional
DL-GDD[17]	Cross-domain Analysis	Layout Quality Assessment	Contrastive Learning Methods
SLLD[39]	End-to-end	Layout Detection	Faster R-CNN
DLAFormer[40]	Unified Framework	Multiple Dla Tasks	Transformer-based Model
DAT[41]	Cohesive End-to-end	Multi-task Detection	Transformer Encoder Decoder
MESc[36]	Hierarchical Multi-stage	Layout Recognition, Classification	Transformer-based Model

Table 3: This table presents a comparative analysis of various end-to-end and unified frameworks employed in document layout analysis (DLA). It highlights the integration approaches, task handling capabilities, and technological applications of each method, demonstrating the diversity and effectiveness of these frameworks in enhancing document processing efficiency and accuracy.

End-to-end and unified frameworks in DLA integrate multiple tasks into a single model, enhancing document processing efficiency and accuracy. These frameworks consolidate DLA sub-tasks like layout recognition, text detection, and entity recognition into cohesive systems. LayoutParser exemplifies this approach by combining model accessibility with customizable training for efficient document image analysis [38].

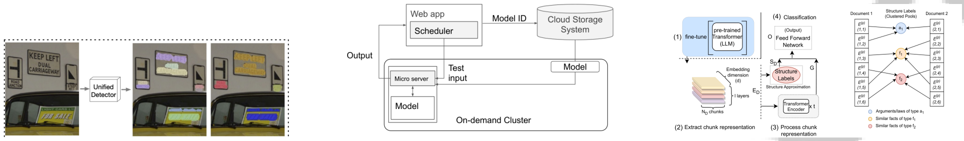
DANIEL integrates layout recognition, handwriting recognition, and named entity recognition within a single architecture, showcasing unified frameworks' potential for handling complex tasks [43]. The Unified Detector model performs scene text detection and layout analysis simultaneously, managing multiple tasks concurrently [34].

VTLayout employs a two-stage model for DLA, localizing document category blocks with a Cascade Mask R-CNN and classifying them through visual and text feature fusion [35]. The Unified Method for Text Detection (UMTD) combines detection techniques to identify text in videos, demonstrating unified methods' versatility [37].

Doc-GCN, a graph-based model, enhances classification by integrating document layout components through graph convolutional networks [42]. DL-GDD generates document layouts guided by style specifications, assessing document quality in creative tasks [17].

In scientific literature, SLLD uses an end-to-end framework to detect and classify layout components, leveraging deep learning for scientific document analysis [39]. DLAFormer, an end-to-end transformer-based model, unifies DLA sub-tasks through relation prediction, highlighting transformers' transformative impact [40].

The DAT framework employs an interactive attention module for improved text instance detection, enhancing textual representation understanding [41]. These frameworks represent significant advancements in DLA, offering comprehensive solutions that improve accuracy, efficiency, and scope across applications. Table 3 provides a comprehensive overview of the different end-to-end and unified frameworks utilized in document layout analysis, illustrating their integration strategies, task management, and technological implementations.



(a) Unified Detector for Traffic Sign Detection in Indoor Environments[34]

(b) On-demand Model Deployment and Testing System[56]

(c) Fine-tuning a pre-trained Transformer for structure labeling[36]

Figure 3: Examples of End-to-End and Unified Frameworks

As shown in Figure 3, end-to-end and unified frameworks have revolutionized document layout analysis by integrating multiple process steps into cohesive systems, enhancing efficiency and accuracy. The "Unified Detector for Traffic Sign Detection in Indoor Environments" exemplifies a unified approach to identifying traffic signs across frames, adapting to environmental changes. The "On-demand Model Deployment and Testing System" highlights a cloud-based solution for model deployment and testing, emphasizing flexibility and scalability. The "Fine-tuning a pre-trained Transformer for structure labeling" demonstrates the application of pre-trained models for structured text data extraction, underscoring end-to-end frameworks' potential in natural language processing tasks. These examples illustrate the transformative impact of unified frameworks in DLA, paving the way for sophisticated, streamlined approaches [34, 56, 36].

3.5 Benchmark Datasets and Evaluation Metrics

Benchmark	Size	Domain	Task Format	Metric
DCIC[10]	1,280,000	Document Image Classification	Image Classification	Accuracy
IlluHisDoc[12]	10,000	Illustration Segmentation	Illustration Segmentation	mIoU
DocBank[57]	500,000	Document Layout Analysis	Sequence Labeling	F1 Score
PubLayNet[58]	360,000	Document Layout Analysis	Document Layout Recognition	Mean Average Precision
ZSL-LLM[59]	4,000	Text Classification	Zero-Shot Classification	F1 Score, Accuracy
SelfDocSeg[60]	81,000	Document Layout Analysis	Document Object Detection	mAP
BaDLAD[61]	40,408	Document Layout Analysis	Segmentation	DICE Score
NewsNet7[62]	3000	Document Layout Analysis	Layout Segmentation	mIoU, mAc

Table 4: This table provides a comprehensive overview of various benchmark datasets used for document layout analysis (DLA), detailing their size, domain, task format, and evaluation metrics. These benchmarks are essential for assessing the performance and effectiveness of DLA methodologies, facilitating the comparison of different approaches and highlighting areas for improvement.

Benchmark datasets and evaluation metrics are crucial for advancing DLA, providing standardized platforms for assessing methodologies' performance and effectiveness. PubLayNet, with its extensive coverage of 335,703 training and 11,245 validation images, facilitates comprehensive DLA evaluations [44]. The RVL-CDIP and ANDOC datasets, consisting of scanned office documents and genealogical records, respectively, are pivotal for model training and evaluation [10].

The IlluHisDoc dataset offers a diverse collection for evaluating illustration segmentation, crucial for distinguishing textual and non-textual document elements [12]. Additionally, the Article Regions, PubLayNet, and DocBank datasets serve as benchmarks for comparing methodologies like the VSR framework using standard metrics [16].

Evaluation metrics like precision, recall, IoU, and mAP measure DLA models' accuracy and reliability in detecting and classifying document elements. Advanced frameworks incorporate BLEU scores and CIDEr for benchmarking natural language processing components within DLA. Annotation platforms like Callico are evaluated using metrics such as annotation task completion time and accuracy, demonstrating their superiority over existing tools [14].

Integrating diverse datasets and advanced metrics is crucial for advancing DLA, facilitating comprehensive assessments that support the development of precise, efficient models. This integration consolidates sub-tasks like text region detection, logical role classification, and reading order prediction into unified frameworks like DLAFormer, demonstrating superior performance over traditional architectures. Innovative strategies like Human-in-the-Loop (HITL) and Key Samples Selection (KSS) methods enable models to learn from minimal data while improving accuracy, evidenced by substantial benchmark performance gains [40, 55]. These benchmarks facilitate approach comparison, highlight improvement areas, and advance document layout analysis. Table 4 presents a detailed summary of key benchmark datasets and their associated evaluation metrics, which are instrumental in advancing the field of document layout analysis.

Feature	**Traditional Methods and Their Limitations**	**Deep Learning Approaches**	**Multimodal and Hybrid Techniques**
Adaptability	Low Adaptability	High Adaptability	Medium Adaptability
Methodology Type	Rule-based	Neural Networks	Hybrid Models
Task Integration	Separate Tasks	Integrated Tasks	Multimodal Integration

Table 5: Comparison of methodologies in Document Layout Analysis highlighting the adaptability, methodology type, and task integration across traditional, deep learning, and multimodal/hybrid approaches. This table illustrates the evolution of techniques from rule-based methods to advanced neural networks and hybrid models, emphasizing their integration capabilities and adaptability to complex document structures.

4 Applications of Document Layout Analysis

Document Layout Analysis (DLA) has significantly broadened its applications, underscoring its essential role in effective document processing across various domains. This section delves into the diverse applications of DLA, beginning with its transformative impact on automating data entry processes, thereby enhancing efficiency and accuracy in handling unstructured documents.

4.1 Automated Data Entry

DLA is vital in automating data entry, converting unstructured documents into structured, machine-readable formats and reducing manual labor. This automation is particularly beneficial in legal contexts, where it substantially cuts costs and processing time [63]. Advanced models, such as those by Wan et al., unify multiple text detection tasks, streamlining data entry by enhancing document processing efficiency [41].

The MSoS approach's segmentation of web pages into coherent blocks exemplifies DLA's utility in automating data entry for multi-device distribution [64]. This facilitates relevant information extraction, optimizing data entry processes. In human resources, DLA automates resume processing, enhancing HR efficiency [65].

Frameworks like nnU-Net and VTLayout illustrate DLA's potential in automated data entry, offering state-of-the-art segmentation without extensive expert knowledge. These frameworks are advantageous in processing structured documents, such as grant applications and scientific literature, where precision and speed are crucial [9, 53, 3, 36, 22].

DLA integration in automated data entry, supported by datasets and advanced models like DLAFormer and DL-DSG, enhances document digitization accuracy and efficiency. DLAFormer consolidates sub-tasks like text region detection and logical role classification, while DL-DSG generates high-quality samples that bridge style gaps [18, 66, 40, 55, 67]. By minimizing manual intervention, DLA streamlines data entry and broadens its applicability, paving the way for further advancements in document processing technologies.

4.2 Information Retrieval

DLA enhances information retrieval systems by efficiently extracting and organizing information from diverse document formats. It enables the construction of dynamic pages from search engine results, personalizing content based on user profiles to improve relevance and user experience [68].

Innovative frameworks like AutoIE advance key information extraction from complex documents, improving scientific information accessibility and usability [22]. Benchmark datasets like DCQA support developing models requiring complex reasoning, enhancing retrieval models' cognitive capabilities [69].

Deep learning has significantly improved information retrieval systems, outperforming traditional methods and enabling systems to handle complex document structures with greater accuracy and efficiency [70]. Techniques like DLAFormer and GLAM streamline analysis and reduce the need for extensive labeled datasets, enhancing DLA's accessibility and applicability across domains [40, 55, 67, 18]. The continuous evolution of DLA techniques, supported by deep learning and benchmarks, promises further enhancements in information retrieval systems across diverse applications.

4.3 Document Management

DLA is crucial for managing documents by detecting and classifying semantic content into categories, converting them into structured formats, and enhancing retrieval and categorization. GLAM leverages metadata from PDFs for efficient DLA through graph segmentation and classification. HITL methodologies improve DLA by incorporating human input for key sample labeling, enhancing model performance with minimal data [55, 18].

Integrating DLA with advanced machine learning models, including Transformer-based and graph neural networks, improves document management systems' accuracy and efficiency. These models interpret spatial relationships and content types, facilitating automatic categorization and indexing [47, 48, 18, 3, 1]. Hybrid models combining CNNs and NLP techniques extract semantic information for effective document classification and retrieval.

DLA automates workflows by extracting metadata and content, triggering automated actions in document management systems. This reduces manual data entry reliance and improves processing speed and accuracy. Frameworks like AutoIE enhance data management in scientific literature, while Faster R-CNN leverages contextual information for faster processing [3, 22]. DLA supports regulatory compliance by ensuring accurate document categorization and retrieval for audits.

Benchmark datasets and evaluation metrics accelerate DLA technology progress in document management, enabling comprehensive model performance assessments and optimization. This advancement enhances document understanding accuracy and supports applications in NLP and computer vision [40, 45, 18]. Benchmarks foster innovation and the development of robust document management solutions.

DLA significantly enhances document management systems' organization and accessibility, streamlining workflows and improving operational efficiency across industries. As DLA technologies advance, they promise to enhance document management practices through HITL collaborative intelligence, end-to-end transformer models like DLAFormer, and graph-based approaches like GLAM, improving information extraction accuracy and content understanding [40, 55, 18].

4.4 Handwritten Document Analysis

Handwritten Document Analysis (HDA) faces challenges due to handwriting variability and layout complexity. Recent DLA advancements have improved handwritten document processing, enhancing information extraction accuracy and efficiency. The DANIEL framework exemplifies state-of-the-art performance in HDA, surpassing existing models in speed and accuracy [43].

Multitask learning approaches, like those by Quirs et al., enhance DLA systems' capability to process complex handwritten layouts, improving segmentation and text recognition [71]. Deep learning techniques in HDA accelerate progress in document analysis, visual information extraction, and information retrieval, leveraging hierarchical representations and predictive modeling capabilities [51, 70, 45]. CNNs and RNNs, including LSTM networks, capture handwritten text's sequential nature, improving recognition accuracy.

End-to-end frameworks in HDA enhance analysis efficiency by consolidating tasks like text detection, segmentation, and recognition into cohesive models. Approaches like DLAFormer and DAT unify sub-tasks, streamlining workflows and improving accuracy across benchmarks [40, 3, 41, 34]. This integration enhances handwritten document processing efficiency and information extraction accuracy.

DLA advancements significantly improve handwritten document analysis, tackling handwriting variability and layout complexity challenges. These techniques facilitate meaningful information extraction and expand applications across domains like historical document preservation, legal analysis, and educational assessment. Innovations in layout analysis, such as multi-task analysis using neural networks and few-shot learning frameworks, enrich information retrieval from complex handwritten texts. Fast CNNs streamline layout analysis, ensuring rapid processing and high accuracy for modern cognitive computing applications [71, 72, 73, 1].

4.5 Cross-Domain Applications

DLA transcends traditional applications, adapting to various document types and formats across domains. Frameworks like RPC-Attention demonstrate DLA's robustness against data corruption and adversarial attacks, enhancing document processing tasks in dynamic environments where data integrity and security are crucial [74].

The M 6Doc dataset showcases DLA's cross-domain applicability, providing a comprehensive collection of annotated document images that facilitate models capable of handling diverse formats and content types [75]. Such datasets train models to generalize across domains, ensuring effectiveness in varied contexts.

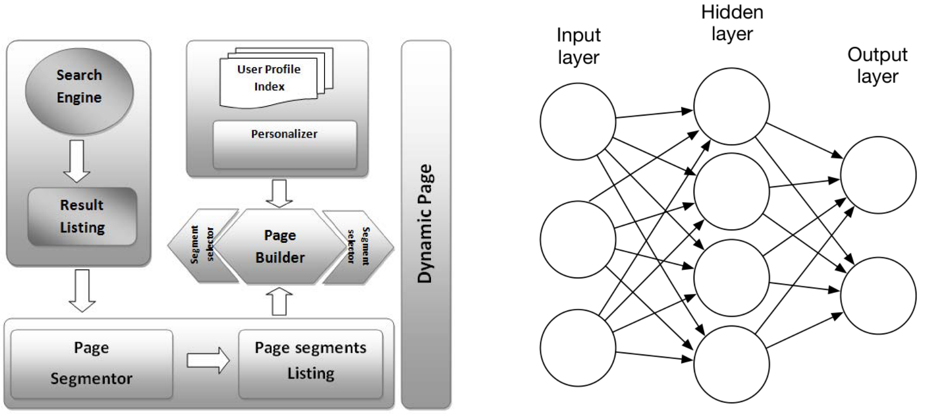
Culturally relevant datasets, like those by Clanuwat et al., highlight DLA's interdisciplinary potential by encouraging research that spans cultural and academic boundaries, fostering technologies sensitive to cultural nuances and historical contexts [28]. This approach enriches document analysis and enhances its applicability in digital humanities, where understanding cultural document significance is essential.

DLA's cross-domain applications are vast, driven by adaptability to different document types and challenges. Ongoing dataset and framework advancements enhance DLA technologies, enabling innovative solutions that harness DLA's strengths, such as classifying and segmenting document elements into structured formats. Models like GLAM leverage metadata for superior performance, while frameworks like DRFN and unsupervised cross-domain methods address training data limitations and document style variability, broadening DLA applications in real-world challenges like optical character recognition and document retrieval [17, 18, 76].

As shown in Figure 4, document layout analysis showcases versatile applications across domains, enhancing technological processes. The "Search Engine Personalization Process Flowchart" illustrates DLA's role in optimizing search engines by personalizing results. This involves processing user queries, filtering and ranking results, and tailoring them using a personalizer module. The "Neural Network Architecture" example demonstrates DLA's application in designing and understanding complex neural networks, such as a multilayer perceptron (MLP) with interconnected nodes for refined outputs. These examples underscore DLA's cross-domain applicability, emphasizing its role in personalizing user experiences and advancing machine learning methodologies [68, 51].

5 Challenges and Future Directions

Advancements in Document Layout Analysis (DLA) require overcoming challenges such as data scarcity, benchmark limitations, model generalization, computational complexity, and diverse document layout management. The following sections explore these issues and propose future directions to enhance DLA technologies.



(a) Search Engine Personalization Process Flowchart[68]

(b) A Neural Network Architecture[51]

Figure 4: Examples of Cross-Domain Applications

5.1 Data Scarcity and Benchmark Limitations

Data scarcity and benchmark limitations significantly impede DLA progress. The lack of comprehensive annotated datasets, crucial for training deep learning models, particularly affects specialized areas like historical document analysis [12]. Annotation challenges further exacerbate issues in OCR and HTR tasks [14]. Current benchmarks often require extensive annotations for specific tasks, limiting their generalizability across diverse document types [12]. Additionally, the lack of effective annotation tools complicates acquiring high-quality labeled data necessary for training DLA models [14], while high computational demands associated with large datasets hinder real-time performance and scalability.

Creating diverse and realistic synthetic datasets is vital for preserving complex document layouts. However, existing methods struggle to generate images accurately reflecting real-world document variability [44]. Limitations in current benchmarks, including high computational costs and insufficient 3D training data, further impede advanced DLA techniques capable of leveraging contextual information. In applications like named entity recognition in resumes, restricted datasets and challenging annotation processes hinder accurate results [65]. Addressing these challenges requires creating inclusive datasets that effectively capture various document layout complexities, enhancing DLA technologies' generalization and applicability across real-world scenarios.

5.2 Model Generalization and Robustness

Model generalization and robustness pose significant challenges in DLA, particularly in adapting to diverse document types and layouts. Complex model architectures necessitate extensive datasets for effective training and generalization across various formats [15]. Optimization is complicated by the high dimensional space of hyperparameters and design choices, affecting model robustness [5]. The reliance on large labeled datasets, often scarce in multilingual and specialized domains, limits generalization due to constraints in input sequence lengths and available data [13]. Additionally, determining deep learning models' applicability to specific problems challenges managing vast information and ensuring fault tolerance [8].

The VSR framework's requirement for both document images and text positions may restrict its generalization compared to unimodal methods [16]. Developing comprehensive perturbation taxonomies and novel evaluation metrics is essential for assessing model resilience against diverse perturbations. Enhancing DLA technologies' adaptability and robustness will significantly improve their capability to process various document types while leveraging rich metadata from electronically generated PDFs. Innovations like the Graph-based Layout Analysis Model (GLAM) and Human-in-the-Loop (HITL) methodologies show promise in improving DLA task efficiency and performance, achieving state-of-the-art results with reduced training data [55, 17, 18, 76].

5.3 Computational Complexity and Efficiency

Computational complexity and efficiency significantly challenge DLA techniques, impacting scalability and applicability. Complex modules, such as those in MIPC-Net, increase computational demands, potentially hindering real-time applications [77]. This complexity is exacerbated by the need for large-scale datasets with fine-grained annotations, as seen in benchmarks like DocBank [27]. Models like DiT highlight the reliance on large datasets for effective learning, underscoring challenges related to computational demands and efficiency [78]. While nnU-Net automates the design process to minimize resource demands, high-quality annotated datasets remain a limiting factor affecting model generalization across different modalities [4].

Current studies face challenges related to data imbalance, overfitting, and deep learning models' interpretability, hindering practical implementation [8]. Synthetic approaches, as demonstrated in historical document segmentation, show competitive performance but require considerable computational resources [12]. Methods achieving processing speeds significantly faster than traditional text-based techniques illustrate advancements in computational efficiency, with some demonstrating a 14-fold increase in speed [3]. However, extensive processing resources, particularly in models requiring significant GPU memory and longer training cycles, persist [41].

To tackle computational complexity and efficiency challenges, innovative solutions must harmonize model complexity with practical deployment constraints. Leveraging advancements such as graph-based representations, human-in-the-loop methodologies, and end-to-end transformer architectures can enhance efficiency and accuracy while minimizing resource requirements. The Graph-based Layout Analysis Model (GLAM) has shown over five times the efficiency of traditional models, while DLAFormer integrates multiple DLA sub-tasks into a single framework, streamlining analysis [40, 55, 79, 18]. Optimizing existing frameworks and advancing algorithmic efficiency can ensure DLA technologies' scalability and applicability across diverse real-world scenarios.

5.4 Handling Complex and Diverse Document Layouts

Handling complex and diverse document layouts in DLA presents significant challenges due to real-world document structures' variability and intricacy. Traditional deep learning systems often struggle with tasks requiring generalization, open-ended inference, and hierarchical structure understanding, crucial for managing varied layouts [15]. The generalization of models trained on existing benchmarks is hampered by simplistic document layouts and background styles, as highlighted by challenges in datasets like DCQA, which may not adequately represent real-world complexity [24].

Datasets such as MIDV-500, targeting identity document analysis in video streams, address some complexities in handling diverse layouts [26]. However, datasets like UDIAT still face class imbalance and segmentation difficulties in ancient manuscript layouts [14]. Current datasets often lack the diversity needed to capture the full spectrum of document variations, especially in historical contexts [2].

The DocSynth framework demonstrates layout-guided synthesis's potential, utilizing deep generative models to enhance generated document images' realism and diversity [44]. Nevertheless, limitations in addressing complex components like tables and figures persist, necessitating further refinement of methods such as Paragraph2Graph [13]. Future research should focus on expanding datasets to encompass a wider variety of document types and styles while integrating multimodal approaches that leverage different data modalities for improved DLA. The complexity of medical data and the diversity of layouts in scientific documents underscore the need for effective segmentation methods capable of managing intricate structures. Enhancing datasets like PubLayNet to include relationships between layout elements and improving annotation quality could significantly bolster DLA models' robustness and generalizability [3].

Addressing complex and diverse document layout challenges requires ongoing research and developing methodologies that adapt to real-world document structures' intricacies. By leveraging innovative techniques and expanding datasets, the field can advance, enhancing document layout analysis' accuracy and efficiency across various applications. Future research should also explore creating more interpretable models, hybrid architectures, and strategies to address data quality and quantity challenges in training deep learning systems [15].

AI-generated, for reference only.

6 Conclusion

Document Layout Analysis (DLA) stands at the forefront of modern technological advancements, significantly transforming document processing through its integration with deep learning, computer vision, and natural language processing. These technologies collectively enhance the ability to analyze and interpret complex document structures, as demonstrated by innovative frameworks like VSR and nnU-Net. Such advancements not only elevate performance in biomedical image segmentation but also set new standards for future research endeavors. The emergence of tools like DocSynth and the Paragraph2Graph framework highlights the progress in generating realistic document images and achieving efficient layout analysis with minimal parameters. Despite these advancements, challenges in model training and interpretability persist, necessitating continued exploration and improvement. The application of Transformers, particularly in modeling long-range dependencies, has shown remarkable promise in medical imaging, further broadening the horizons of DLA applications. The generation of synthetic documents has proven effective in enhancing document element extraction, achieving high-quality results across varied datasets. Moreover, advancements in named entity recognition models like RoBERTa and ELECTRA provide robust solutions for information extraction in technology domains. These developments underscore the vital role of DLA in advancing document processing and its potential for future innovations, urging a continued pursuit of alternative strategies to overcome current limitations.

References

[1] Dario Augusto Borges Oliveira and Matheus Palhares Viana. Fast cnn-based document layout analysis. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1173–1180. IEEE, 2017.

[2] Mélodie Boillet, Christopher Kermorvant, and Thierry Paquet. Multiple document datasets pre-training improves text line detection with deep neural networks, 2021.

[3] Carlos Soto and Shinjae Yoo. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3464–3470, 2019.

[4] M Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel RS Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, et al. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings*, volume 10553. Springer, 2017.

[5] Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[6] Transformers in medical imaging:.

[7] Jiawei Wang, Kai Hu, Zhuoyao Zhong, Lei Sun, and Qiang Huo. Detect-order-construct: A tree construction based approach for hierarchical document structure analysis, 2024.

[8] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. A review of deep learning with special emphasis on architectures, applications and recent trends, 2019.

[9] Jill P. Naiman. Generalizability in document layout analysis for scientific article figure caption extraction, 2023.

[10] Chris Tensmeyer and Tony Martinez. Analysis of convolutional neural networks for document image classification. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 388–393. IEEE, 2017.

[11] Ruiting Shao and Edward J. Delp. Forensic scanner identification using machine learning, 2020.

[12] Tom Monnier and Mathieu Aubry. docextractor: An off-the-shelf historical document element extraction, 2020.

[13] Shu Wei and Nuo Xu. Paragraph2graph: A gnn-based framework for layout paragraph analysis, 2023.

[14] Christopher Kermorvant, Eva Bardou, Manon Blanco, and Bastien Abadie. Callico: a versatile open-source document image annotation platform, 2024.

[15] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.

[16] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: A unified framework for document layout analysis combining vision, semantics and relations, 2021.

[17] Xingjiao Wu, Luwei Xiao, Xiangcheng Du, Yingbin Zheng, Xin Li, Tianlong Ma, Cheng Jin, and Liang He. Cross-domain document layout analysis using document style guide, 2024.

[18] Jilin Wang, Michael Krumdick, Baojia Tong, Hamima Halim, Maxim Sokolov, Vadym Barda, Delphine Vendryes, and Chris Tanner. A graphical approach to document layout analysis, 2023.

[19] Raphaël Barman, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan. Combining visual and textual features for semantic segmentation of historical newspapers, 2020.

[20] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*, 2022.

[21] Gang Chen. Deep learning with nonparametric clustering, 2015.

[22] Yangyang Liu and Shoubin Li. Autoie: An automated framework for information extraction from scientific literature, 2024.

[23] Christoph Wick and Frank Puppe. Fully convolutional neural networks for page segmentation of historical document images, 2018.

[24] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey, 2020.

[25] Md Ataullha, Mahedi Hassan Rabby, Mushfiqur Rahman, and Tahsina Bintay Azam. Bengali document layout analysis with detectron2, 2023.

[26] Vladimir Viktorovich Arlazarov, Konstantin Bulatovich Bulatov, Timofey Sergeevich Chernov, and Vladimir Lvovich Arlazarov. Midv-500: a dataset for identity document analysis and recognition on mobile devices in video stream. , 43(5):818–824, 2019.

[27] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[28] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[29] Silvia Zottin, Axel De Nardin, Emanuela Colombi, Claudio Piciarelli, Filippo Pavan, and Gian Luca Foresti. U-diads-bib: a full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts, 2024.

[30] Chang Liu, Yuanhe Tian, and Yan Song. A systematic review of deep learning-based research on radiology report generation, 2024.

[31] Daria Kern and Andre Mastmeyer. 3d bounding box detection in volumetric medical image data: A systematic literature review, 2020.

[32] K. S. Kuppusamy and G. Aghila. A model for personalized keyword extraction from web pages using segmentation, 2012.

[33] Nosseiba Ben Salem, Younes Bennani, Joseph Karkazan, Abir Barbara, Charles Dacheux, and Thomas Gregory. Modular neural network approaches for surgical image recognition, 2023.

[34] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis, 2022.

[35] Shoubin Li, Xuyan Ma, Shuaiqun Pan, Jun Hu, Lin Shi, and Qing Wang. Vtlayout: Fusion of visual and text features for document layout analysis, 2021.

[36] Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents, 2024.

[37] Sauradip Nag, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Michael Blumenstein. A new unified method for detecting text from marathon runners and sports players in video, 2020.

AI-generated, for reference only.

[38] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 131–146. Springer, 2021.

[39] Huichen Yang and William H. Hsu. Vision-based layout detection from scientific literature using recurrent convolutional neural networks, 2020.

[40] Jiawei Wang, Kai Hu, and Qiang Huo. Dlaformer: An end-to-end transformer for document layout analysis, 2024.

[41] Xingyu Wan, Chengquan Zhang, Pengyuan Lyu, Sen Fan, Zihan Ni, Kun Yao, Errui Ding, and Jingdong Wang. Towards unified multi-granularity text detection with interactive attention, 2024.

[42] Siwen Luo, Yihao Ding, Siqu Long, Josiah Poon, and Soyeon Caren Han. Doc-gcn: Heterogeneous graph convolutional networks for document layout analysis, 2022.

[43] Thomas Constum, Pierrick Tranouez, and Thierry Paquet. Daniel: A fast document attention network for information extraction and labelling of handwritten documents, 2024.

[44] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. Docsynth: A layout guided approach for controllable document image synthesis, 2021.

[45] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications, 2021.

[46] Larex a semi-automatic open-so.

[47] Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. A hybrid approach for document layout analysis in document images, 2024.

[48] Sotirios Kastanas, Shaomu Tan, and Yi He. Document ai: A comparative study of transformer-based, graph-based models, and convolutional neural networks for document layout analysis, 2023.

[49] Tianlong Ma, Xingjiao Wu, Xin Li, Xiangcheng Du, Zhao Zhou, Liang Xue, and Cheng Jin. Document layout analysis with aesthetic-guided image augmentation, 2021.

[50] Siwei Lai. Word and document embeddings based on neural network approaches, 2016.

[51] Ye Zhang, Md Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. Neural information retrieval: A literature review, 2017.

[52] Sarvesh Patil. Deep learning based natural language processing for end to end speech translation, 2018.

[53] Jinghong Li, Wen Gu, Koichi Ota, and Shinobu Hasegawa. Object recognition from scientific document based on compartment refinement framework, 2024.

[54] Linda Studer, Michele Alberti, Vinaychandran Pondenkandath, Pinar Goktepe, Thomas Kolonko, Andreas Fischer, Marcus Liwicki, and Rolf Ingold. A comprehensive study of imagenet pre-training for historical document image analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 720–725. IEEE, 2019.

[55] Xingjiao Wu, Tianlong Ma, Xin Li, Qin Chen, and Liang He. Human-in-the-loop document layout analysis, 2021.

[56] Nham Le, Tuan Lai, Trung Bui, and Doo Soon Kim. Autonlu: An on-demand cloud-based natural language understanding system for enterprises, 2020.

[57] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis, 2020.

[58] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.

[59] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. Large language models are zero-shot text classifiers, 2023.

[60] Subhajit Maity, Sanket Biswas, Siladittya Manna, Ayan Banerjee, Josep Lladós, Saumik Bhattacharya, and Umapada Pal. Selfdocseg: A self-supervised vision-based approach towards document segmentation, 2023.

[61] Kazi Reyazul Hasan, Mubasshira Musarrat, Sadif Ahmed, and Shahriar Raj. Framework and model analysis on bengali document layout analysis dataset: Badlad, 2023.

[62] Wenzhen Zhu, Negin Sokhandan, Guang Yang, Sujitha Martin, and Suchitra Sathyanarayana. Docbed: A multi-stage ocr solution for documents with complex layouts, 2022.

[63] Hsiu-Wei Yang and Abhinav Agrawal. Extracting complex named entities in legal documents via weakly supervised object detection, 2023.

[64] Mira Sarkis, Cyril Concolato, and Jean-Claude Dufourd. Msos: A multi-screen-oriented web page segmentation approach, 2015.

[65] Ege Kesim and Aysu Deliahmetoglu. Named entity recognition in resumes, 2023.

[66] Alejandro Peña, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Marcos Grande, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. Document layout annotation: Database and benchmark in the domain of public affairs, 2023.

[67] Xingjiao Wu, Luwei Xiao, Xiangcheng Du, Yingbin Zheng, Xin Li, Tianlong Ma, Cheng Jin, and Liang He. Cross-domain document layout analysis using document style guide. *Expert Systems with Applications*, 245:123039, 2024.

[68] K. S. Kuppusamy and G. Aghila. Segmentation based approach to dynamic page construction from search engine results, 2012.

[69] Anran Wu, Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Zisong Zhuang, Nian Xie, Cheng Jin, and Liang He. Dcqa: Document-level chart question answering towards complex reasoning and common-sense understanding, 2023.

[70] Nicholas G. Polson and Vadim O. Sokolov. Deep learning, 2018.

[71] Lorenzo Quirós. Multi-task handwritten document layout analysis, 2018.

[72] Lorenzo Quirós and Enrique Vidal. Evaluation of a region proposal architecture for multi-task document layout analysis, 2021.

[73] Axel De Nardin, Silvia Zottin, Matteo Paier, Gian Luca Foresti, Emanuela Colombi, and Claudio Piciarelli. Efficient few-shot learning for pixel-precise handwritten document layout analysis, 2022.

[74] Rachel S. Y. Teo and Tan M. Nguyen. Unveiling the hidden structure of self-attention via kernel principal component analysis, 2024.

[75] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M^{6}doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis, 2023.

[76] Xingjiao Wu, Ziling Hu, Xiangcheng Du, Jing Yang, and Liang He. Document layout analysis via dynamic residual feature fusion, 2021.

[77] Yizhi Pan, Junyi Xin, Tianhua Yang, Teeradaj Racharak, Le-Minh Nguyen, and Guanqun Sun. A mutual inclusion mechanism for precise boundary segmentation in medical images, 2024.

[78] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer, 2022.

[79] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis, 2023.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.