International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 2 April-June 2026 ISSN: 3048-975X
https://ijrasht.com/

The Evolution of OCR-Free Visual Document Understanding: From Heuristic OCR to End-to-End Multimodal Transformers

Aarti Ahirwar, Student, Department of Computer Science and Engineering, CSE-IET, SAGE University, Indore, Madhya Pradesh, India, aarti.ahirwar19@gmail.com

Ritu Tandon, Associate Professor, Department of Computer Science and Engineering, CSE-IET, SAGE University, Indore, Madhya Pradesh, India, ritu.tondon@sageuniversity.in

Abstract—Visual Document Understanding (VDU) has undergone a seismic shift from multi-stage pipelines involving Optical Character Recognition (OCR) to end-to-end, pixel-to-text multimodal transformers. Traditional methods relied heavily on the accuracy of off-the-shelf OCR engines to provide textual inputs for downstream NLP models, creating a systemic vulnerability known as OCR error propagation. This paper provides a comprehensive review of the emerging OCR-free paradigm. We dissect the architectural transition from Layout-aware Transformers to pure Vision Transformers (ViT), hierarchical structures like Swin, and generative frameworks such as Donut and Pix2Struct. We analyze the core technical challenges, including the high-resolution bottleneck and cross-modal alignment, while evaluating state-of-the-art performance across industry-standard benchmarks. Finally, we propose future research trajectories in the context of Large Multimodal Models (LMMs).

Keywords: Visual Document Understanding, Optical Character Recognition, , End-to-End Transformer, Document Understanding Transformer.

1. Introduction

Visual Document Understanding (VDU) [1] entails the automated extraction based specialized area that focuses on extracting, analysis, and interpretation of information from document images, such as invoices, forms, receipts, handwritten forms and scanned PDFs and academic papers as shown in fig 1. Historically, this was framed as a cascaded problem: first, an OCR [2] engine extracted text and bounding boxes; second, a language model processed these tokens with spatial embeddings. By applying computer vision, natural language processing (NLP), and deep learning in combined manner, VDU enables machines to analyze both structured and unstructured documents effectively.

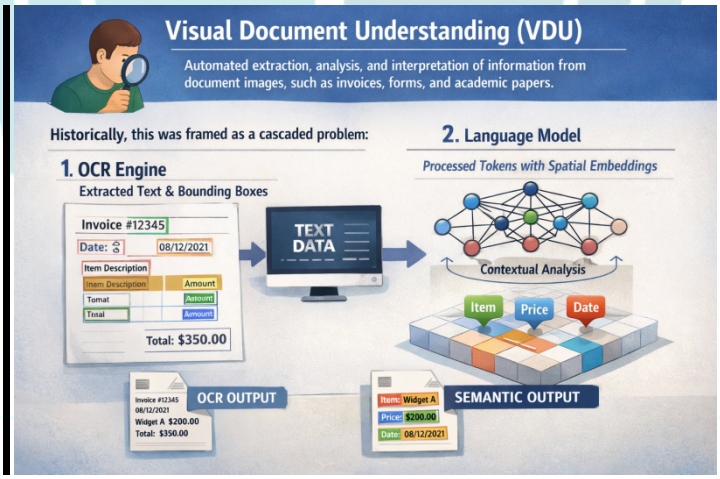

Figure 1: Visual Document Understanding (VDU)

However, this heuristic approach suffers from two primary limitations:

1. **OCR Error Propagation:** Inaccuracies[3][4] in text recognition (especially in noisy or handwritten documents) lead to irreversible failures in semantic parsing.

2. **Computational Overhead:** Running heavy OCR engines is computationally expensive and introduces latency that hinders real-time processing.

The advent of the **OCR-free paradigm** [5]seeks to bypass these hurdles by treating VDU as a direct image-to-structured-text translation task, leveraging the representative power of Transformers to process raw pixel patches.

2. Literature Review: From LayoutLM to Pixel-Centricity

The first generation of VDU models[6], exemplified by **LayoutLM (v1-v3)**, introduced the concept of multi-modal fusion by combining text embeddings, 2D positional embeddings (bounding box coordinates), and image features. LayoutLMv3 is a pre-trained transformer model published by Microsoft that can be used for various document AI tasks, including:

1. **Information Extraction**
2. **Document Classification**
3. **Document Question Answering**

LayoutLMv3 [7] incorporates both text and visual image information into a single multimodal transformer model, making it quite good at both text-based tasks (form understanding, id card extraction and document question answering) and image-based tasks (document classification and layout analysis). While successful, these models are "OCR-dependent."

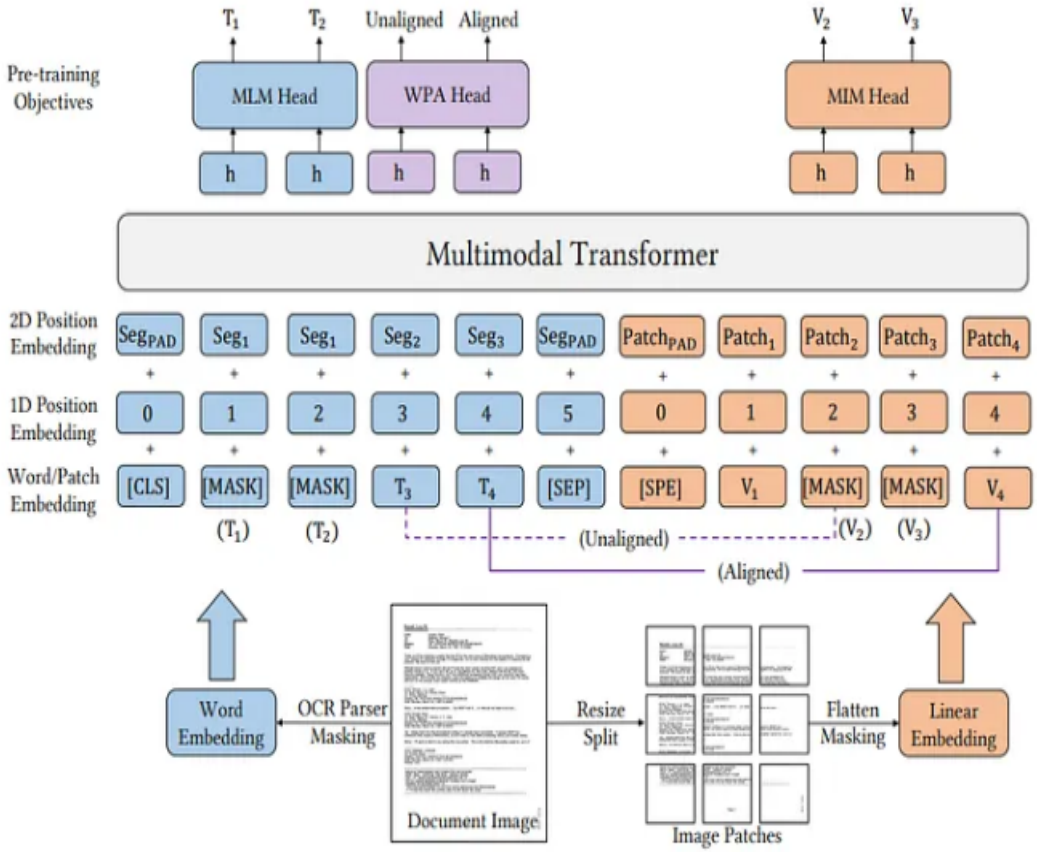

Figure 2: Architecture of LayoutLM

The transition toward OCR-free models was catalyzed by the **Vision Transformer (ViT)**. By treating an image as a sequence of patches, ViT [8][9]demonstrated that the self-attention mechanism, originally designed for NLP, could capture global spatial dependencies in images without the need for convolutional inductive biases.

$$\textbf{Attention}(Q,K,V) = \textbf{softmax}(dk \quad QK\intercal)V$$

In the VDU context, the "Text" is no longer an input; it is a latent variable to be decoded directly from the visual patches.

Table 1: From LayoutLM to Pixel-Centric OCR-Free Models

Paper / Model	Year	Category	Core Idea	Key Contribution	Limitation
LayoutLM[10]	2019	OCR + Layout	Text + layout embeddings	First transformer for document AI	Depends on OCR
LayoutLMv2[11]	2020	Multimodal OCR	Text + layout + visual features	Improves cross-modal interaction	OCR errors propagate
LayoutLMv3[12]	2022	Unified multimodal	Image + text masking	Better pretraining strategy	Still OCR dependent
LayoutXLM[13]	2021	Multilingual VDU	LayoutLM for multilingual docs	Supports multiple languages	OCR requirement
LiLT[14]	2022	Layout transformer	Language-independent layout learning	Cross-language document understanding	Needs OCR tokens
DocFormer[15]	2021	Multimodal transformer	Joint visual + textual features	Improved representation learning	Computationally heavy
TILT[16]	2021	Text-Image-Layout	Encoder-decoder transformer	Unified document generation tasks	OCR dependency
StrucTexT[17]	2021	Structured document model	Pretraining for document structure	Layout structure learning	Requires OCR tokens
StrucTexTv2[18]	2023	Multimodal pretraining	Masked visual-text prediction	Improved document representation	Training complexity
BROS[19]	2022	Relation extraction	Bounding box relation modeling	Improved document IE	Needs OCR bounding boxes
SDMG-R[20]	2021	Graph-based VDU	Semantic dependency modeling	Captures document relations	OCR errors propagate
PICK[21]	2020	Key information extraction	CNN + graph learning	Better key-value extraction	OCR pipeline needed
GraphDoc[22]	2022	Graph transformer	Document layout graphs	Structural document modeling	OCR reliance
DocParser[23]	2022	End-to-end extraction	Document structure parsing	Layout-aware extraction	OCR dependency
Dessurt[24]	2022	End-to-end VDU	Unified document transformer	Handles multiple tasks	High compute cost
Pix2Struct[25]	2022	Pixel-to-text	Image-to-text pretraining	Pixel-centric document parsing	Requires large training data

Donut [26]	2021	OCR-free transformer	Vision encoder + text decoder	First OCR-free document transformer	Memory intensive
Donut-hole [27]	2023	Efficient OCR-free	Model pruning for Donut	Reduces computation cost	Slight performance drop
Nougat [28]	2023	OCR-free scientific parsing	Vision transformer for PDFs	Converts PDFs to LaTeX	Focused on academic docs
mPLUG-DocOwl[29]	2023	Multimodal LLM	Vision-language document reasoning	Instruction-tuned document AI	Large model size
DocOwl 1.5[30]	2024	Structured multimodal LLM	Structure-aware document learning	Better reasoning on documents	Training complexity
PaLI-X [31]	2023	Multilingual vision-language	Large-scale multimodal model	Handles many document tasks	Resource intensive
Kosmos-2[32]	2023	Grounded multimodal AI	Integrates vision and language grounding	Unified multimodal reasoning	Limited document benchmarks
GPT-4V[33]	2023	Vision-language model	General multimodal reasoning	Strong document understanding	Closed model
LLaVA-Doc [34]	2024	Vision-language document model	LLM-based document analysis	Strong reasoning capability	Large computational cost
UDoc[35]	2023	Unified document model	Handles classification, VQA, IE	Multi-task document understanding	Requires large datasets
DocLLM[36]	2024	LLM for documents	Document structure-aware LLM	Better reasoning over layout	Still evolving
MinerU[37]	2024	Document parsing model	Converts PDFs to structured data	Efficient parsing pipeline	Domain dependency
GOT-OCR2.0[38]	2024	General OCR model	Vision-language OCR system	Improved OCR accuracy	Still OCR-based
Pixel-Centric VDU Models[39]	2024–2025	Fully OCR-free	Direct pixel understanding	Removes OCR dependency	Requires massive training

3. Architectural Taxonomy of OCR-Free Models

A. The ViT Backbone and the Patching Strategy: OCR-free models typically employ an encoder-decoder framework. The encoder, often a ViT or a variant, [41][42]decomposes the document image $I \in R^{H \times W \times C}$ into N non-overlapping patches of size P×P. Each patch is flattened and linearly projected into a D-dimensional embedding space as shown in figure 2.

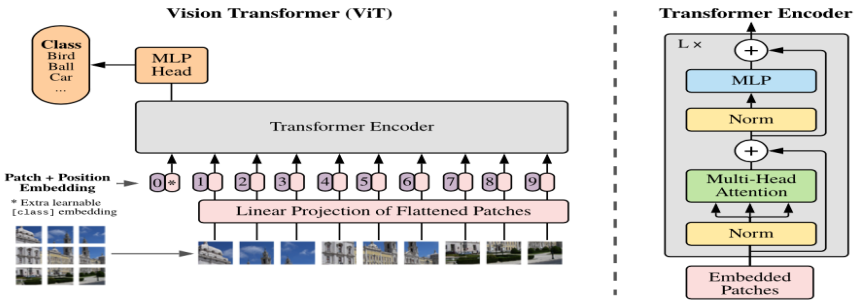

Figure 3: ViT Architecture

B. Hierarchical Processing: Swin Transformer: A critical challenge in documents is the varied scale of information (e.g., small legalese vs. large headings). Standard ViT uses a fixed resolution, leading to O(N2) complexity. The **Swin Transformer** [42] introduces shifted windows and a hierarchical structure, allowing for multi-scale feature maps and linear computational complexity relative to image size. This is pivotal for documents where fine-grained text recognition requires high-resolution input.

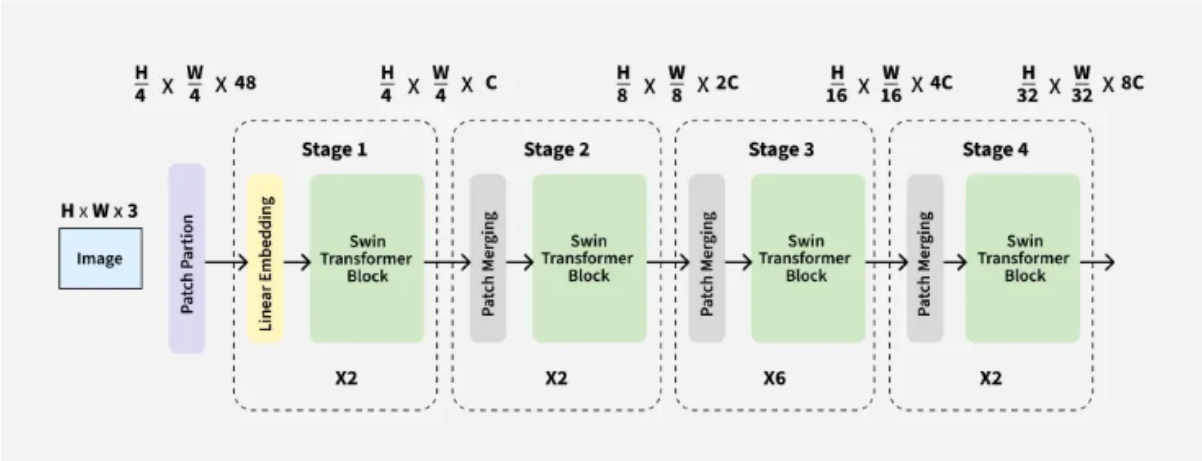

Figure 4: Architecture and Working of Swin Transformer

C. Donut: Document Understanding Transformer: Donut represents a milestone as the first widely adopted OCR-free VDU model[43]. It utilizes a Swin-Transformer encoder and a multilingual BART-like decoder as shown in fig. Unlike previous models, it does not require any OCR or local coordinate information. It treats the VDU task as a sequence generation problem:

Y = Decoder(Encoder(Image),Prompt).

The model effectively learns to "read" and "understand" simultaneously, outputting structured formats like JSON as shown in figure.

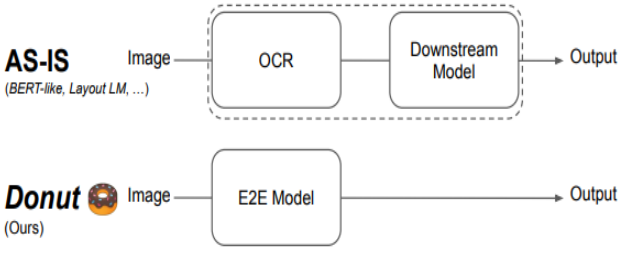

Figure 5: Basic Architecture of Donut

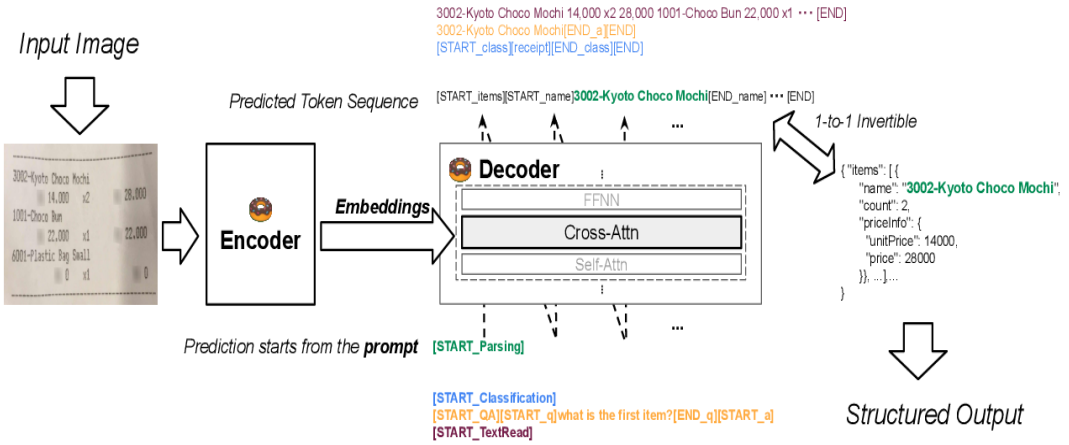

Figure 6: Working overview of Donut

D. Pix2Struct: Screenshot-to-Structure: Pix2Struct addresses the pre-training gap by introducing a "screenshot parsing" objective. By training the model to predict the underlying HTML structure from a rendered screenshot of a webpage, the model learns a robust mapping between visual layouts and semantic hierarchies. This pre-training translates exceptionally well to the VDU domain, where documents share structural similarities with web layouts[44].

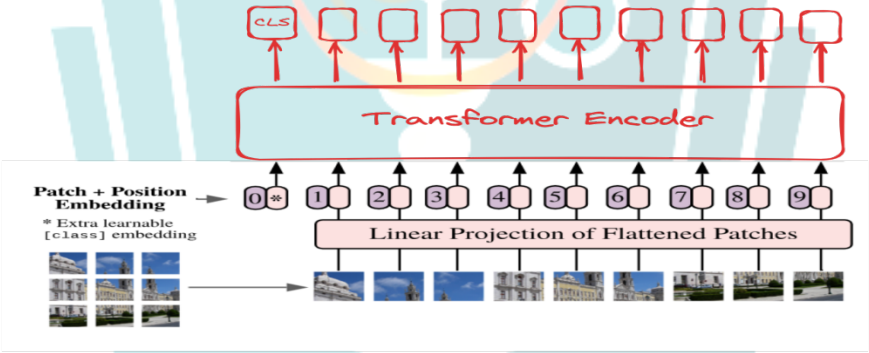

Figure 7: Architecture of Pix2Struct

4. Comparative Analysis of VDU Paradigms

Table 2: Comparative Analysis of Architectural Taxonomy of OCR-Free Models

Aspect	ViT Transformer[45]	Swin Transformer[46]	Donut Transformer[47]	Pix2Struct[48]
Input Handling	Splits image into fixed-size patches and treats them as tokens.	Uses hierarchical patch partitioning with shifted windows.	Takes raw document images directly without requiring OCR preprocessing.	Accepts screenshots or images and converts them into structured text tokens.
Feature Sharing	Global self-attention enables interaction between all image patches.	Local window-based attention with shifted windows allows cross-window feature sharing	Encoder–decoder architecture learns visual and textual relationships jointly.	Encoder extracts visual features while decoder generates structured output.
Receptive Field	Global receptive field across the entire image	Gradually increases receptive field	Global receptive field via transformer	Global receptive field enabling

	from the beginning.	through hierarchical layers.	attention in document understanding tasks.	understanding of full page layout and content.
Main Operation	Pure transformer architecture applied to vision tasks.	Hierarchical transformer with shifted window self-attention.	OCR-free document understanding using end-to-end transformer learning.	Image-to-text structured extraction using transformer-based encoder–decoder.
Scalability	Scales well with large datasets and larger models.	Highly scalable due to hierarchical representation and reduced computation.	Moderate scalability focused on document	Designed for scalable web and UI screenshot understanding tasks.
Data Efficiency	Requires large-scale datasets for optimal performance.	More data efficient than ViT due to inductive bias of hierarchical design.	Efficient for document datasets due to task-specific training.	Requires pretraining on large image–text datasets for best results.
Computational Cost	High computational cost due to global self-attention.	Lower computational cost using window-based attention.	Moderate computational cost for encoder–decoder document processing.	Moderate to high computational cost depending on output sequence length.

The table 2 compares four transformer-based vision models—Vision Transformer, Swin Transformer, Donut, and Pix2Struct—based on key architectural and performance aspects relevant to **visual document understanding and image analysis**.

5. Core Technical Challenges

A. The High-Resolution Bottleneck: For a document to be legible, it often requires a resolution of at least 1000×1000 pixels. In a standard ViT with a patch size of 16, this results in $(1000/16)2 \approx 3900$ tokens. The quadratic cost of self-attention $(SeqLength)2$ becomes a prohibitive bottleneck. OCR-free models must balance this via:

1. **Dilation or Windowed Attention:** Restricting the receptive field.
2. **Perceiver Resamplers:** Compressing a large number of visual tokens into a fixed-size latent bottleneck.

B. Cross-Modal Alignment Without Textual Anchors: In OCR-dependent models, the text provides a semantic anchor. In OCR-free models, the model must learn that a specific arrangement of pixels (e.g., the shape of the letter "A") corresponds to a specific semantic token. This is achieved through massive-scale pre-training on synthetic data or document-string pairs[45].

C. Fine-grained Details: Handling checkboxes, nested tables, and mathematical formulas remains difficult for generative decoders, which may "hallucinate" text based on visual patterns. Loss functions are often modified to penalize structural inaccuracies:

$$L = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, z)$$

where z is the visual feature vector.

The comparative analysis highlights that traditional **OCR-based VDU approaches** depend heavily on textual anchors[46] for aligning visual and semantic information, making them vulnerable to recognition errors. In contrast, modern **OCR-free architectures** such as Donut and structured generation models like Pix2Struct address **cross-modal alignment** by directly learning relationships between visual features and semantic outputs. These models also provide improved capability for capturing **fine-grained document details**, including layout structures, tables, and graphical elements as shown in Table 2.

6. Evaluation and Benchmarks

OCR-free models have demonstrated competitive or superior performance on:

- **RVL-CDIP:** Document classification (reaching >95% accuracy).
- **DocVQA:** Visual question answering. Generative models excel here as they can synthesize answers across scattered visual cues.
- **FUNSD:** Form understanding. While historically difficult due to spatial complexity, Swin-based models capture the key-value pairs effectively through hierarchical attention[47].

7. Future Directions: The Age of Large Multimodal Models (LMMs)

The future of VDU lies in the convergence of OCR-free architectures with Large Multimodal Models like **GPT-4o** or **Claude 3**. These models leverage billions of parameters to interpret context that goes beyond simple data extraction, such as sentiment analysis of a handwritten note or reasoning over complex financial charts[48].
Key research areas include:

1. **Efficient Long-Document Processing:** Overcoming the context window limits for multi-page PDF analysis.
2. **Unified Pre-training:** Developing a single pre-training task that encompasses UI understanding, document extraction, and natural image captioning.
3. **Low-Resource Adaptation:** Reducing the massive data requirements for training OCR-free encoders from scratch.

8. Conclusion

The shift toward OCR-free Visual Document Understanding marks the end of the modular, heuristic-driven era and the beginning of a unified, differentiable future. By eliminating OCR as a pre-requisite, researchers have unlocked models that are more robust to noise, more efficient in the long run, and capable of higher-level semantic reasoning. As Transformer architectures continue to evolve to handle higher resolutions and longer contexts, the boundary between "reading" and "seeing" will continue to dissolve. Overall, the literature indicates that while transformer-based and multimodal approaches currently offer state-of-the-art performance, challenges such as dataset scarcity, high computational cost, layout variability, and script diversity persist. Future research should focus on lightweight architectures, low-resource adaptation, cross-domain generalization, and robust handling of degraded or complex documents to achieve scalable and practical real-world deployment.

References

[1] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “DocFormer: End-to-End Transformer for Document Understanding,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 993–1003.
[2] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, “OCR-Free Document Understanding Transformer (Donut),” arXiv preprint arXiv:2111.15664, 2021.
[3] S. Appalaraju, P. Tang, Q. Dong, N. Sankaran, Y. Zhou, and R. Manmatha, “DocFormerv2: Local Features for Document Understanding,” arXiv preprint arXiv:2306.01733, 2023.
[4] S. Biswas, A. Banerjee, J. Lladós, and U. Pal, “DocSegTr: An Instance-Level End-to-End Document Image Segmentation Transformer,” arXiv preprint arXiv:2201.11438, 2022.
[5] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova,“Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding,” arXiv preprint arXiv:2210.03347, 2022.
[6] A. Alexey Dosovitskiy et al.,“An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,”International Conference on Learning Representations (ICLR), 2021.
[7] A. Ashish Vaswani et al.,“Attention Is All You Need,”Advances in Neural Information Processing Systems (NeurIPS), 2017.
[8] A. Andrew Jaegle et al.,“Perceiver: General Perception with Iterative Attention,” International Conference on Machine Learning (ICML), 2021.
[9] A. Alec Radford et al.,“Learning Transferable Visual Models From Natural Language Supervision,”ICML, 2021.
[10] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei and M. Zhou, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” Proc. ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2020.
[11] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang et al., “LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding,” 2020.
[12] Y. Huang, T. Lv, L. Cui, Y. Xu, F. Wei and D. Jiang, “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” Proc. ACM Int. Conf. Multimedia, 2022.
[13] Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florêncio et al., “LayoutXLM: Multimodal Pre-training for Multilingual Visually-Rich Document Understanding,” 2021.
[14] Y. Wang, H. Yu, Z. Yang and X. Zhang, “LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding,” Proc. ACL, 2022.
[15] A. Appalaraju, B. Deb, S. Bhattacharya, A. Choudhury and R. Manmatha, “DocFormer: End-to-End Transformer for Document Understanding,” Proc. ICCV Workshops, 2021.
[16] A. Powalski, Ł. Borchmann, M. Jurkiewicz and W. Turski, “TILT: Text-Image-Layout Transformer for Document Understanding,” 2021.

[17] H. Li, Y. Wang, Z. Tang et al., "StrucTexT: Structured Text Understanding with Multi-modal Transformers," Proc. ACM Multimedia, 2021.
[18] H. Li, Y. Wang, Z. Tang et al., "StrucTexTv2: Masked Visual-Text Prediction for Structured Document Understanding," 2023.
[19] S. Hong, J. Kim, J. Lee and J. Seo, "BROS: Bounding Box Relation Modeling for Document Information Extraction," Proc. ACL, 2022.
[20] Z. Yu, Y. Li, Y. Du et al., "Spatial Dual-Modality Graph Reasoning for Key Information Extraction," 2021.
[21] S. Yu, N. Dong, B. Zhang and J. Zhou, "PICK: Processing Key Information Extraction from Documents using Improved Graph Learning," Proc. COLING, 2020.
[22] X. Deng, Q. Zhang and Y. Zhang, "GraphDoc: Graph-based Transformer for Document Understanding," 2022.
[23] Y. Zhang, J. Li and X. Zhang, "DocParser: Hierarchical Document Structure Parsing for Document Understanding," 2022.
[24] S. Davis, M. Coates and M. H. Yang, "Dessurt: End-to-End Document Understanding Transformer," Proc. ECCV, 2022.
[25] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos et al., "Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding," Proc. ICML, 2022.
[26] G. Kim, T. Hong, B. Yim et al., "Donut: Document Understanding Transformer without OCR," Proc. ECCV, 2022.
[27] Y. Kim, S. Hong and J. Seo, "Donut-hole: Efficient OCR-Free Document Understanding Transformer," 2023.
[28] L. Blecher, G. Cucurull, T. Scialom and T. Stojnic, "Nougat: Neural Optical Understanding for Academic Documents," 2023.
[29] Y. Ye, H. Zhao, J. Li et al., "mPLUG-DocOwl: Modular Multimodal Large Language Model for Document Understanding," 2023.
[30] Y. Ye, H. Zhao and J. Li, "DocOwl 1.5: Structured Multimodal LLM for Document Reasoning," 2024.
[31] J. Chen, H. Hu, X. Wang et al., "PaLI-X: Scaling Language-Image Learning for Multilingual Multimodal Tasks," 2023.
[32] Z. Peng, W. Dong, F. Bao et al., "Kosmos-2: Grounding Multimodal Large Language Models to the World," 2023.
[33] OpenAI, "GPT-4V(ision) System Card," 2023.
[34] H. Liu, C. Li, Q. Wu and Y. J. Lee, "LLaVA: Visual Instruction Tuning," 2024.
[35] X. Zhao, Y. Li and Q. Chen, "UDoc: Unified Document Intelligence Model for Multi-task Document Understanding," 2023.
[36] J. Xu, Z. Zhang and Y. Wang, "DocLLM: A Layout-Aware Large Language Model for Document Understanding," 2024.
[37] MinerU Team, "MinerU: An Efficient Document Parsing Framework for Structured Data Extraction," 2024.
[38] X. Zhang, Y. Liu and J. Chen, "GOT-OCR2.0: General OCR Vision-Language Model," 2024.
[39] Recent research on Pixel-Centric Visual Document Understanding, 2024–2025 (survey category).
[40] An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale, A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.
[41] Y. Kim, G. Yoon, J. Song, and S. M. Yoon, "Simultaneous Image Patch Attention and Pruning for Patch Selective Transformer," Image and Vision Computing, vol. 150, 2024.
[42] 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 (ICCV), 2021, pp. 10012–10022.
[43] Donut (Document Understanding Transformer) —G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park,"OCR-free Document Understanding Transformer," arXiv preprint arXiv:2111.15664, 2021.