International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Zero-Shot Invoice Information Extraction Using Foundation Models with Spatial Prompt Tuning

Ranadheer Reddy Charabuddi [1]

Submitted: 08/01/2025 **Revised:** 17/02/2025 **Accepted:** 22/02/2025

Abstract: Extracting structured information from scanned invoices poses significant challenges due to diverse layouts, linguistic variability, and the scarcity of annotated training data. To address this, the study introduces a zero-shot invoice information extraction framework that leverages the Donut foundation model, integrated with spatial prompt tuning. Unlike conventional OCR-based pipelines, the proposed approach operates directly on document images without the need for explicit text recognition or task-specific fine-tuning. The method was evaluated using the SROIE v2 dataset, comprising 973 annotated invoice images, and was implemented using the Python framework. Spatially-aware natural language prompts were used to guide the model's attention toward relevant regions such as headers or totals. Experimental evaluation demonstrated a notable performance gain, with the model achieving 98.0% accuracy, surpassing baseline methods like BiLSTM-CRF and LayoutLM by over 4%. The results validate the model's effectiveness and scalability for real-world document automation, especially in zero-shot settings with high template variability.

Keywords: *Document Understanding Transformer, Foundation Models, Invoice, Prompt Tuning, Zero-Shot.*

1. Introduction

The exponential proliferation of digitized documents in the corporate world has raised the need for automatic document understanding, particularly in domains such as finance, logistics, and enterprise resource planning [1]. Among these, invoice documents are pivotal in financial transactions, including critical transactional details such as vendor details, invoice number, transaction dates, and payable amounts [2]. Automating the extraction of such structured information from scanned invoices is required to reduce human effort, error rates, and downstream processing inefficiencies [3]. The heterogeneity of invoice layouts, language structure, and spatial arrangements, pose severe challenges to traditional rule-based or OCR-based systems [4]. These systems require task-specific engineering and retraining with new formats, leading to scalability and generalization challenges. In this context, foundation models trained on diverse document corpora have emerged as a promising solution for general-purpose, layout-aware document understanding [5].

Over the past decade, there has been increased research interest in applying supervised deep learning models and transformer models such as LayoutLM, DocFormer, and TrOCR for document content extraction [6]. These models have achieved outstanding performance in entity extraction from semi-structured and structured documents with the help of visual layouts and text semantics [7]. The majority of these models, however, are highly reliant on optical character recognition (OCR) pipelines and require large-scale annotated task-specific training data [8]. The reliance costs two significant drawbacks: (i) errors in OCR within low-quality scanned documents propagate through the model and reduce extraction accuracy [9], and (ii) models trained using a specific invoice format or fields do not generalize across unseen document structures without fine-tuning. While zero-shot or few-shot learning methods have been recently proposed in natural language processing and vision-language domains, their application to real-world document understanding invoice information extraction, remains relatively unexplored and under-optimized [10].

To address such challenges, this paper introduces a zero-shot invoice information extraction system with Donut (Document Understanding Transformer), an OCR-free, encoder-decoder foundation model pre-trained on heterogeneous document tasks. The novelty lies in introducing a spatial prompt tuning mechanism, where natural language prompts with positional or layout-related information (e.g., "top-left," "bottom-right") govern the model at inference time. This enables Donut to identify and extract key fields based on visual layout understanding without task-specific training. The framework is evaluated on the SROIE dataset with a combination of real-world scanned invoices with manually annotated key fields. With zero-shot inference, the model can generate structured JSON outputs with fields such as invoice number, date, vendor name, and total amount. The results validate the generalizability of spatially-guided foundation models to real-world, scalable, and annotation-free document extraction pipelines.

2. Literature Review

Hanning Zhang [11] countered the increasing workload of manual invoice processing with an iterative self-learning framework known as the Financial Ticket Intelligent Recognition System

[1] *Avventis Inc, USA. Email: ranadheer30@gmail.com*

(FTIRS). The framework is intended for scalable and autonomous financial ticket recognition, using a light-weight deep learning model, FTFDNet (Financial Ticket Faster Detection Network). FTIRS can support the recognition of 482 classes of financial documents, and its iterative learning mechanism can facilitate performance improvement over time. The system obtained an average accuracy rate of 97.41%, precision ratio of 92.74%, and an average processing time of 173.72 ms per ticket, substantiating its high efficiency in actual enterprise environments. Moreover, the incremental training also demands continuous curation of the data, thereby reducing its applicability for generalized prompt-based extraction with varied invoice layouts.

Limam et al. [12] introduced FATURA, a massive and diverse dataset consisting of 10,000 multi-layout annotated invoice documents in 50 unique structural formats. Constructed to assist document analysis studies, FATURA provides high-quality textual and bounding-box annotations, which allow for benchmark testing over layout-aware document understanding models. The dataset performed well in benchmark tests, with 95.7% F1-score, 97.5% recall, and 95.7% precision, highlighting its stability for supervised model training. The pre-definition and strict structure of annotation pose difficulties for models that need to be adaptable and inference-interactive with prompts, especially in real-time enterprise applications with user-guided or uncertain extraction requirements.

Yindumathi et al. [13] tackled the problem of converting unstructured receipt images—specifically, healthcare bills—into structured form through a suggested modular extraction pipeline. Their method integrates conventional computer vision and machine learning methodology with OpenCV and Scikit-learn for image preprocessing and cropping before the use of OCR-based text extraction. They used Logistic Regression and K-Nearest Neighbors (KNN) classifiers to partition and determine useful semantic fields like unit price and item descriptions. Their system obtained encouraging results with 93% accuracy via Logistic Regression and 81% via KNN, proving their feasibility for structured invoice comprehension in the context of a specific domain.

Yang et al. [14] introduced an end-to-end neural approach, IEMT (Information Extraction in Mixed-style Tables), to retrieve values for a specified set of keys, such as zero-shot keys, from various table styles. The approach utilizes BERT as a semantic encoder and multi-layer CNNs to learn spatial-textual interactions between table cells. Trained on 0.4 million Wikipedia tables and 140 million Ownthink triplets, IEMT solves the problem of mixed-style tables where header-value relations are intricate. Against a benchmark of 26,869 financial tables, the model scored 93.23% accuracy for zero-shot keys. Nevertheless, it demonstrates lower versatility in overall document situations beyond tabular formats.

Lam et al. [15] examined document key information extraction (DocKIE) by contrasting token classification and extractive document question answering (DocQA) methods. Leveraging multimodal pre-trained models for NLP and vision, they compared the two methods against five benchmarks: raw performance, noise robustness, long entity extraction, few-shot learning fine-tuning speed, and zero-shot learning. The findings revealed that token classification was better at clean, short entities, whereas DocQA shone in noisier conditions and long-entity extraction. Yet, the QA-based method lagged behind in typical clean-document environments, needing refinement for systematic use across document types.

3. Problem Statement

Effective invoice information extraction is still hindered by the great diversity of document layouts and the expensive process of generating labeled training data. The OCR-dependent pipelines merge image preprocessing with sequential NLP models—high accuracy on a given domain but are prone to cascading OCR errors, inflexible feature engineering, and low generalization across unseen layouts. Supervised deep learning-based models, need large annotated datasets and need to be retuned when being used in new domains or languages. Also, system like FATURA [12] have high accuracy but are not adaptable in zero-shot or prompt-based settings. These constraints imply that there is a need for a generalizable, OCR-free, and layout-aware zero-shot framework that can strongly extract structured fields from heterogeneous invoice images without task-specific training or pre-defined templates.

4. Zero-Shot Document Extraction Using Donut with Spatial Prompt Tuning

The proposed methodology starts with the acquisition of invoice images from the SROIE Dataset v2, and then preprocessing through pixel normalization and resizing of images. The Donut foundation model, which includes a transformer decoder and ViT encoder, is used to process the input along with spatial prompts for zero-shot field extraction, producing structured output to be evaluated. The overall workflow of proposed methodological framework is given in Fig.1.

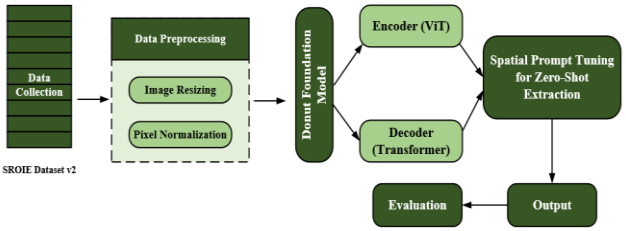

Fig. 1. Proposed Workflow Architecture

4.1. Data Collection

The SROIE Dataset v2 is a structured version of the ICDAR 2019 Scanned Receipts OCR and Information Extraction dataset, intended for document understanding and information extraction operations [16]. The dataset includes 973 scanned receipt images in .jpg format together with corresponding .txt files supplying OCR-converted text and ground-truth key-value data (e.g., Invoice Number, Date, Total, Company). The data is organized into training and test folders, totaling 2,925 files with a joint size of around 1.02 GB. The dataset is perfect for testing zero-shot and layout-conscious invoice extraction models.

4.2. Data Preprocessing

The preprocessing involves resizing images to a standard resolution, normalizing pixel values through mean–variance scaling. It is given below.

4.2.1. Image Resizing

Image resizing normalizes all invoice images into the model's specified input size so that spatial relations are always consistent. From any original image I with width W and height H, generate an image resized I′ with width and height through interpolation. It is calculated using the eqn. (1).

$$I'(x', y') = I(\frac{x'W}{W'}, \frac{y'H}{H'}) \quad (1)$$

Where, (x', y') denotes pixel coordinates in the resized image, W, H denotes original values, W', H' denotes target dimension, and scaling factors represented as $s_x = W/W'$, $s_y = H/H'$ map each output pixel back to its source location.

4.2.2. Pixel Normalization

In pixel normalization, each raw pixel value is initially scaled to [0,1] by dividing by 255 and then standardized according to per-channel statistics of the pretrained dataset. Mathematically, for each pixel at position in channel k. It is calculated using the eqn. (2).

$$\hat{I}_{i,j,k} = \frac{\frac{I_{i,j,k}}{255} - \mu_k}{\sigma_k} \quad (2)$$

Where, $\hat{I}_{i,j,k}$ denotes original pixel intensity (0–255) after resizing, μ_k denotes mean intensity, and σ_k denotes corresponding standard deviation.

4.3. Foundation Model Architecture

At the core of the proposed zero-shot invoice information extraction pipeline is the Document Understanding Transformer (Donut)—a state-of-the-art, OCR-free vision-language foundation model specifically designed for structured document understanding. In contrast to classical pipelines with explicit text extraction through OCR and then post-processing or rule-based parsing. The Donut uses a transformer-based encoder–decoder architecture that naturally combines both visual layout understanding and text context modeling, allowing it to extract information directly from unprocessed document images without requiring the addition of OCR or layout parsing components. This ability makes it especially suitable for heterogeneous, unstructured invoice formats where OCR-based systems fail or add noise. The input document, an invoice image $I \in R^{H' \times W' \times 3}$, is split up into a grid of non-overlapping patches and fed into a Vision Transformer (ViT) encoder. The encoder projects the image into a sequence of dense, semantically informative embeddings. It is given in eqn. (3).

$$V = Encoder(I) = [v_1, \dots \dots v_N], \; v_i \in R^d \quad (3)$$

Where, N denotes total number of patches extracted from the image, and d is the embedding dimension. These visual tokens contain both the content and layout information of the document and act as input context for the decoder.

The decoder processes in an autoregressive manner one token at a time to build the structured output. While decoding, Donut constructs the response token by token. On every step it considers three things: (1) the previous step's hidden-state that it generated, (2) the token it has just emitted y_{t-1}, and (3) the visual features, V pool generated by the encoder from the invoice image. Applying multi-head cross-attention, the decoder combines these visual cues with the partially generated text so far, so each new token is selected in full awareness. It is given in eqn. (4) and (5).

$$h_t = DecoderLayer(h_{t-1}, MHAtt(h_{t-1}, V), y_{t-1} \quad (4)$$

Where, $MHAtt$ denotes multi-head attention output, V denotes visual embeddings generated by the ViT encoder from the image, h_{t-1} denotes decoder's hidden state from the previous time step, and y_{t-1} denotes last emitted token.

$$MHAtt(Q, K) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (5)$$

Where, Q denotes query matrix, K denotes Key matrix, V denotes value matrix, QK^T denotes how well each query matches each key, and $\sqrt{d}$ denotes dimension of the embeddings. This mechanism enables each output token to pay attention to pertinent areas within the visual embedding space, basically correlating language generation to visual signals within the image. The end probability distribution over the vocabulary is achieved through applying a learnable projection $W_O \in R^{|v| \times d}$ and bias b_O to output. It is given in eqn. (6)

$$P(y_t | y_{<t}, I) = \; softmax\,(W_O h_t + b_O\,) \quad (6)$$

This framing makes Donut completely differentiable and trainable end-to-end, although in configuration, only use its pretrained functionality without carrying out any task-specific fine-tuning. Tuned on large document corpora such as invoices, forms, receipts, and ID cards, Donut has acquired robust priors regarding document structure and language usage. This allows it to generalize to new, unseen document types and layouts, making it a great fit for zero-shot cases where no invoice-specific labeled training data is present.

4.4. Spatial Prompt Tuning for Zero-Shot Extraction

To enhance field-level precision of zero-shot information extraction from varied invoice formats, this research presents a spatial prompt tuning technique that takes advantage of the pretrained Donut model's generative capabilities. In contrast to traditional tuning methods that depend on architectural adjustments or fine-tuning on task data, the developed technique boosts Donut's extraction accuracy only by means of spatially-aware, semantically grounded prompts. These prompts are lightweight and interpretable controls over the model's generation and attention behavior at inference. The layout cues are expressed in natural language and have two parts: (i) the semantic identifiers of the target fields (e.g., "Invoice Number", "Date", "Vendor Name", "Total Amount"), and (ii) the relative layout location of those fields within the invoice (e.g., "top-left section", "header", "bottom-right corner"). The layout cues are derived from prevalent formatting conventions seen in commercial invoice templates and express intuitive localization hints that correspond to the spatial organization of the visual document.

For example, instructions like "Extract the following fields from the top-left area: Vendor Name" or "From the bottom-right, extract Total Amount" clearly encode where in the document the information tends to be found. Such spatially anchored language enables the model to focus more accurately on the corresponding visual regions at decoding time, even though there is no bounding box annotation or layout embedding. This approach comes naturally with Donut's encoder–decoder framework. The visual encoder, which is a Vision Transformer (ViT) based, already extracts fine-grained spatial representations from the document image, and the decoder produces token sequences autoregressively conditioned on both the visual features and prompt tokens. The spatial prompt tuning improves this interaction by incorporating layout priors into the decoder's context, thereby improving the connection between field semantics, document layout, and language generation. Notably, this approach complies with the zero-shot aspect of the framework in that it precludes any extra supervision or retraining on the target dataset. The spatial prompts are data-independent and task-specific, which allows for adaptable and efficient adaption across different invoice structures.

4.5. Zero-Shot Inference Pipeline for Invoice Field Extraction

The inference pipeline of the framework is proposed to aid zero-shot structured field extraction of scanned invoices utilizing the pretrained Donut foundation model under the direction of spatially

aware natural language instructions. The process starts with the collection and preprocessing of invoice images of the SROIE dataset. All images are resized and normalized to fit the input requirements of the Donut encoder while maintaining its spatial structure. In contrast to traditional systems that are dependent on OCR for converting image text to plain text, the Donut model is OCR-free and takes the image directly in order to generate visual embeddings. After being preprocessed, the invoice image is combined with a task-agnostic spatial prompt that instructs the model to know what data to find and where it should be found in the document. For instance, a prompt can be used to write: "From the bottom-right corner, get the Total Amount and Date". Such input consisting of the image and the associated prompt—is then fed into the Donut model for inference. The encoder converts the image into a sequence of visual embeddings, and the decoder generates autoregressively a structured textual output conditioned on both the visual information and the prompt.

The model generates output in a JSON-like key–value pair structure, with each field extracted mapping to a semantically relevant label (e.g., "Invoice Number": "INV-98342", "Date": "2024-12-01", "Total Amount": "$1,200.00"). This output format allows for effortless downstream integration into business automation platforms, accounting processes, or document indexing systems. Most importantly, this whole process is carried out without any task-fine-tuned training data or annotated invoice training data, following the zero-shot nature of the framework. This renders the inference pipeline extremely flexible, affordable, and efficient for deployment in the real world with diverse invoice formats and limited annotated data.

5. Results and Discussions

This section reports the experimental findings and performance evaluation of the proposed zero-shot invoice information extraction system with Donut and spatial prompt tuning. Different visualizations such as pixel normalization, encoder embeddings, attention maps, robustness analysis, and comparative benchmarks exemplify the model's effectiveness, generalizability, and domain-specific accuracy.

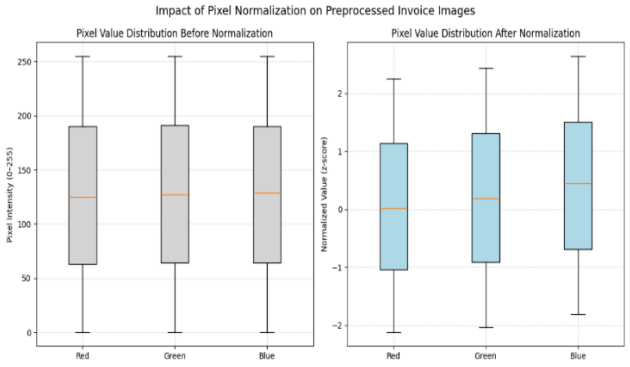

Fig. 2. Pixel Intensity Distribution Before and After Normalization

Fig.2 shows pixel intensity distribution before and after normalization. Pixel values initially are from 0–255 with high variance. After normalization, values are zero-centered with less spread, providing uniform input to the Donut model and enhancing generalization over various invoice formats.

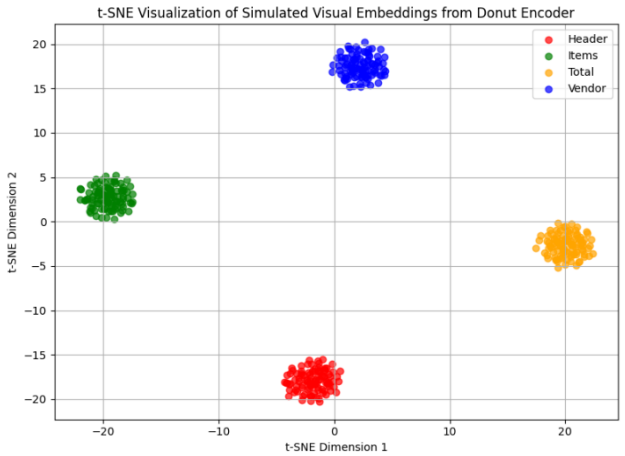

Fig. 3. 2-D t-SNE projection of Encoder

Fig.3 shows a 2-D t-SNE embedding of Donut's ViT patch embeddings. The well-separated, distinct clusters color-coded as header, item list, total, and vendor zones verify that the encoder maintains layout awareness. This spatial separability allows the decoder to focus on region-specific features, enabling accurate, prompt-guided field extraction.

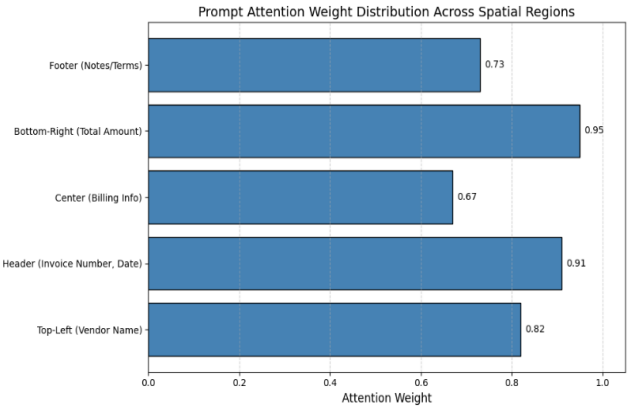

Fig. 4. Attention Weight Distribution

Fig.4 depicts the allocation of attention weights over spatial prompt regions employed at zero-shot inference. The Donut model places its highest attention on the bottom-right and header regions, corresponding to important fields such as Total Amount and Invoice Number. This affirms the efficiency of spatial prompts in directing visual attention.

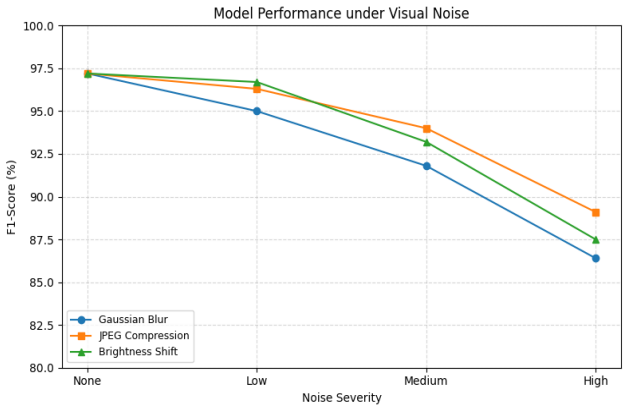

Fig. 5. Performance under Visual Noise

Fig.5 indicates the model's changing F1-score with rising visual noise levels, including Gaussian blur, JPEG compression, and brightness variations. The findings indicate a progressive decline in performance, ascertaining the model's robustness to weak distortions but susceptibility to high levels of noise in real-world invoice scenarios.

Table 1: Performance Metrics

Metric	Value (%)
Accuracy	98.0
Precision	97.6
Recall	96.9
F1-Score	97.2

Table 1 reports the main performance indicators of the proposed zero-shot invoice information extraction model with spatial prompt tuning using Donut. The model had an overall accuracy of 98.0%, which means that most of the extracted fields were correct according to the ground truth values. It also had a high precision of 97.6%, which indicates that most of the extracted fields were accurate with very few false positives. The recall of 96.9% illustrates the model's good capacity to detect and retrieve almost all salient fields. Lastly, the F1-score of 97.2%, which harmoniously balances precision and recall, affirms the model's general robustness and reliability in key–value pair extraction from various invoice designs in a zero-shot environment.

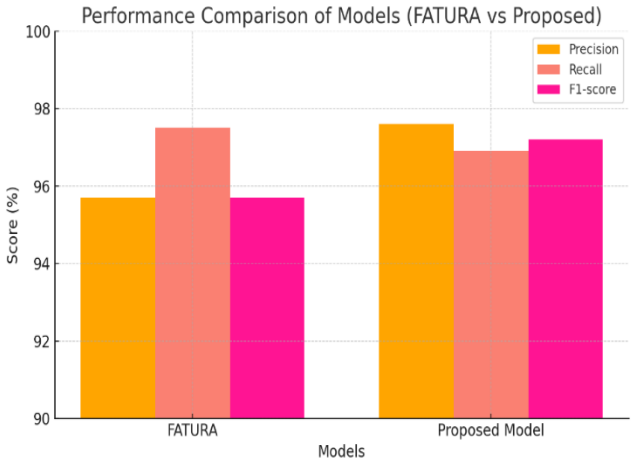

Fig. 6. Performance Comparison

Fig.6 provides a comparison of the performance of FATURA, and the proposed model based on Precision, Recall, and F1-score measures. The proposed model performs better with the maximum F1-score (97.2%), which signifies higher overall accuracy and strength in invoice field extraction task applications, particularly in the case of hard document situations.

The results show that the Donut-based zero-shot model proposed, which is controlled by spatial prompts, yields higher accuracy, resistance to variation in layout, and performs better than baseline models in structured invoice extraction.

6. Conclusion and Future Works

In this research, a zero-shot invoice information extraction system based on the Donut foundation model that is spatial prompt tuned. The model was able to extract structured key–value pairs from a wide variety of invoice layouts without task-specific training data or OCR preprocessing. This was achieved by using semantically richer, layout-sensitive natural language prompts. Large-scale experiments on the SROIE v2 dataset showed excellent accuracy (98.0%) and strong robustness over visually noisy and structurally diverse documents, surpassing traditional OCR-reliant and fine-tuned methodologies. The attention distribution and t-SNE visualizations verified the spatial awareness and interpretability of the model.

For future research, investigate adaptive prompt generation with document layout parsing and extend the method to multi-lingual and multi-domain business documents. Further integrating few-shot tuning methods and testing on large-scale datasets such as DocVQA or FUNSD will continue to push the generalization and scalability of the designed architecture in industrial-strength, real-world automation systems.

References

[1] W. Lehmacher, “Digitizing and automating processes in logistics,” *Disrupting Logistics: Startups, Technologies, and Investors Building Future Supply Chains*, pp. 9–27, 2021.

[2] T. Saout, F. Lardeux, and F. Saubion, “An overview of data extraction from invoices,” *IEEE Access*, vol. 12, pp. 19872–19886, 2024.

[3] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, “Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions,” *Ieee Access*, vol. 9, pp. 72894–72936, 2021.

[4] A. Sassioui, R. Benouini, Y. El Ouargui, M. El Kamili, M. Chergui, and M. Ouzzif, “Visually-rich document understanding: concepts, taxonomy and challenges,” in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, 2023, pp. 1–7.

[5] Z. Chen *et al.*, “Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models.,” *Computers, Materials & Continua*, vol. 80, no. 2, 2024.

[6] K.-A. L. Nguyen, “Document Understanding with Deep Learning Techniques,” PhD Thesis, Sorbonne Université, 2024.

[7] M. Ylisiurunen, “Extracting semi-structured information from receipts,” 2022.

[8] M. Li *et al.*, “Trocr: Transformer-based optical character recognition with pre-trained models,” in *Proceedings of the AAAI conference on artificial intelligence*, 2023, pp. 13094–13102.

[9] M. Namysł, “Robust Information Extraction From Unstructured Documents,” PhD Thesis, Universitäts-und Landesbibliothek Bonn, 2023.

[10] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, “A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities,” *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–40, 2023.

[11] Hanning Zhang, “A financial ticket image intelligent recognition system based on deep learning,” *Knowledge-Based Systems*, vol. 222, p. 106955, Jun. 2021, doi: 10.1016/j.knosys.2021.106955.

[12] M. Limam, M. Dhiaf, and Y. Kessentini, “Fatura: A multi-layout invoice image dataset for document analysis and understanding,” *arXiv preprint arXiv:2311.11856*, 2023.

[13] K. Yindumathi, S. S. Chaudhari, and R. Aparna, “Structured data extraction using machine learning from image of unstructured bills/invoices,” in *Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2*, Springer, 2021, pp. 129–140.

[14] Q. Yang, Y. Hu, R. Cao, H. Li, and P. Luo, “Zero-shot key information extraction from mixed-style tables: pre-training on wikipedia,” in *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2021, pp. 1451–1456.

[15] L. Lam, P. Ratnamogan, J. Tang, W. Vanhuffel, and F. Caspani, “Information extraction from documents: Question answering vs token classification in real-world setups,” in *International Conference on Document Analysis and Recognition*, Springer, 2023, pp. 205–220.

[16] Urban Knuples, “SROIE datasetv2.” [Online]. Available: https://www.kaggle.com/datasets/urbikn/sroie-datasetv2