Deep Learning Based Key Information Extraction from Business Documents: Systematic Literature Review

ALEXANDER MICHAEL ROMBACH, Saarland University, Saarbrücken, Germany and German Research Center for Artificial Intelligence, Saarbrücken, Germany
PETER FETTKE, German Research Center for Artificial Intelligence, Saarbrücken, Germany and Saarland University, Saarbrücken, Germany

Extracting key information from documents represents a large portion of business workloads and therefore offers a high potential for efficiency improvements and process automation. With recent advances in Deep Learning, a plethora of Deep Learning based approaches for Key Information Extraction have been proposed under the umbrella term Document Understanding that enable the processing of complex business documents. The goal of this systematic literature review is an in-depth analysis of existing approaches in this domain and the identification of opportunities for further research. To this end, 130 approaches published between 2017 and 2024 are analyzed in this study.

CCS Concepts: • **Applied computing → Document analysis**; Business process management systems; • **Computing methodologies → Information extraction**; *Neural networks;*

Additional Key Words and Phrases: Key information extraction, document understanding, business documents, deep learning, systematic literature review

ACM Reference Format:
Alexander Michael Rombach and Peter Fettke. 2025. Deep Learning Based Key Information Extraction from Business Documents: Systematic Literature Review. *ACM Comput. Surv.* 58, 2, Article 45 (September 2025), 37 pages. https://doi.org/10.1145/3749369

1 Introduction

The general idea of a paper-free—or at least paperless—office already came up five decades ago [35]. However, to this day, physical paper documents still play an important role in business operations, as they are a key means of communication related to transactions both within and between organizations [99]. The processing of such documents is an essential yet time-consuming task that offers a high potential for automation due to the high workload involved as well as the critical nature of information transfer between different information systems [16, 110]. At the same time, it can be observed that the ongoing digital transformation of business operations is leading to an increase in the digital processing of documents. This trend reinforces the need—but also the potential—for automated document processing, as more and more documents are available in digital form [91].

Author's Contact Information: Alexander Michael Rombach, Saarland University, Saarbrücken, Saarland, Germany and German Research Center for Artificial Intelligence, Saarbrücken, Saarland, Germany; e-mail: alexander_michael.rombach@uni-saarland.de; Peter Fettke, German Research Center for Artificial Intelligence, Saarbrücken, Saarland, Germany and Saarland University, Saarbrücken, Saarland, Germany; e-mail: peter.fettke@dfki.de.

ACM 0360-0300/2025/09-ART45
https://doi.org/10.1145/3749369

Research on document processing is not new and has been conducted for several decades [54]. In fact, the term ”document analysis” can be traced back to the 1980s [118]. In recent years, however, there has been an upsurge in research related to document processing based on **visually-rich documents** (**VRDs**) and business documents, made possible by major advances in **Deep Learning** (**DL**). One of the most studied document processing task in this regard is **Key Information Extraction** (**KIE**) [73], which is concerned with extracting specific named entities from documents in a structured form [99].

Complex business documents pose a significant challenge to KIE systems because they cannot be understood and processed as linear text sequences, as has been the case in most traditional KIE applications [11]. These documents typically contain implicit and explicit cues and complex positional dependencies between certain text segments. In this regard, related research investigates different methods to integrate such cues into the model architectures in order to achieve better extraction results. Business documents also have special characteristics resulting from their connection to business processes. For example, different documents that are being processed as part of the same process run usually contain reoccurring information. These aspects and how they can improve corresponding KIE systems are worth investigating. In general, a process-oriented understanding is critical in order to adequately address the challenge of KIE in real-world contexts.

Although many DL-based KIE approaches for VRDs have been proposed, there is no comprehensive overview of most recent work in this area that focuses on the underlying DL concepts and technical characteristics while also adopting a business process perspective. The aim of this **systematic literature review** (**SLR**) is to fill this gap and provide a detailed overview of this research area and its state of the art. The contribution of this work is threefold:

(1) An SLR based on 130 approaches to provide a concise overview of DL-based KIE methods for business documents.
(2) Categorization and detailed comparison of corresponding methods based on various characteristics.
(3) Dissemination of results, research gaps, and derived potentials for future research.

The remainder of this manuscript is structured as follows: Section two provides background information on key concepts and nomenclature that are relevant to this SLR. In section three, we discuss related work in terms of existing surveys and illustrate, how this study differs. The methodology used for this SLR, and more specifically how relevant literature was identified, is illustrated in section four. Section five provides an overview of the results of the in-depth analysis. Section six serves as a discussion of the key findings and is dedicated to specific aspects of the identified literature. Based on the previous two sections, we propose a research agenda and starting points for follow-up research in section seven. Section eight concludes the manuscript with a summary.

2 Background

2.1 A Business Process Perspective

Although business documents have a value of their own, they are best understood in relation to each other. Whenever they belong to the same process run, the information contained in these documents is also closely related to each other. For example, all documents related to the same run of a purchasing process typically contain the same order ID. Because they are embedded in specific processes that specify what information is relevant, business documents always contain distinct sets of predefined entities [15]. For example, invoices contain many different types of monetary values and unique identifiers such as invoice numbers that need to be identified.

Understanding temporal relationships in business processes is also important. Consider the following example of a simplified purchasing process. First, a customer places an order for a physical good. The company then processes the order, which results in the company sending an order confirmation and, at a later stage, the actual physical goods, a delivery note, and an invoice. Figure SF1 of the electronic supplementary material[1] shows an exemplary run of a corresponding process, including the interactions with internal and external entities such as customers and suppliers. Sequences where business documents are relevant are highlighted in red, indicating parts of the process where **Document Understanding** (**DU**) approaches could play a key role. As can be seen, especially interfaces related to external partners are expressed through document-based exchanges.

In addition to implicit knowledge about predefined entities, the consideration of associated business processes can also be useful in the sense that the information flow between individual process steps could be taken into account. For example, one could consider how the data extracted from a business document is further processed in subsequent process steps. From this, conclusions could be drawn as to how corresponding business documents should be processed. One could also analyze whether—in addition to the documents themselves—there is an additional data flow between the document processing steps that could facilitate the understanding of the documents.

2.2 Document Understanding

DU, based on concepts from **Natural Language Processing** (**NLP**) and **Computer Vision** (**CV**), is an umbrella term covering a wide range of document processing tasks, including KIE, table understanding, **document layout analysis** (**DLA**), document classification, and **visual question answering** (**VQA**) [11]. More recently, novel tasks have been proposed such as ”Key Information Localization and Extraction”, which is an extension of KIE that emphasizes the need to localize key information, e.g., by identifying bounding boxes[2] [99]. For a brief introduction to DU tasks besides KIE that are not covered in this review, we refer to [16]. There are also other terms that are used for this research area, such as Document Analysis, Document AI, or Document Intelligence [16, 75].

DU can also be grouped according to the complexity of the tasks, as suggested by [25]: Perception tasks deal with the recognition of descriptive document elements (e.g., text or entire tables), whereas induction tasks aim at the extraction of enriched information (e.g., document class or named entities) based on the perceived documents. Finally, reasoning represents the most complex subtask, combining perception and induction to enable DU beyond the explicitly contained information, mostly in the context of VQA. An example of a reasoning task is obtaining natural language explanations of figures and diagrams in documents [99].

2.3 Key Information Extraction

KIE investigates methods for the automated extraction of named entities from documents into structured formats [99] and can be further subdivided into individual research areas, namely **Named Entity Recognition** (**NER**), Named Entity Linking, Coreference Resolution, **Relation Extraction (RE)**, Event Detection, and Template Filling [26, 80], with NER and RE being the most prominent ones. The goal of NER is to detect entities in text and assign predefined labels to them, usually solved as a sequence labeling task [6]. Figure 1 visualizes the outlined hierarchy of DU and subordinate research areas such as KIE.[3]

KIE approaches can be divided into three main subgroups based on how they represent the underlying documents, namely graph-based, grid-based, and sequence-based methods. Graph-based systems convert document pages into graph structures that represent the layout and content of the pages. Such graphs typically allow for a flexible structure and can be constructed in a variety

[1] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

[2] A bounding box indicates the coordinates of an element in the document image by its top left corner as well as its width and height.

[3] Note that this hierarchy is not a result of this study, but rather a reflection of the understanding in the literature.

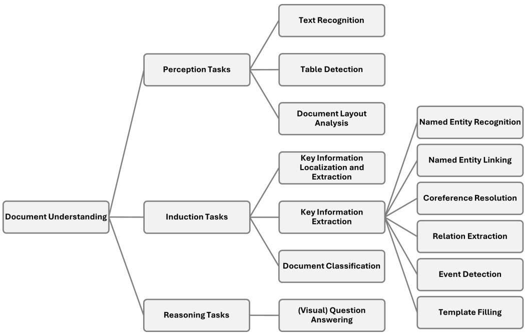
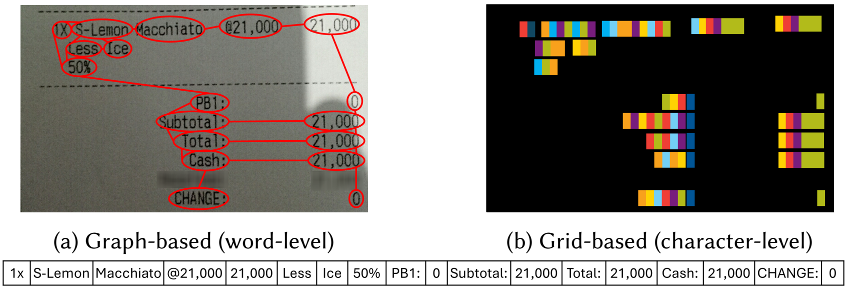


Fig. 1. Overview of Document Understanding and Key Information Extraction.



(a) Graph-based (word-level)

(b) Grid-based (character-level)

1x	S-Lemon	Macchiato	@21,000	21,000	Less	Ice	50%	PB1:	0	Subtotal:	21,000	Total:	21,000	Cash:	21,000	CHANGE:	0

(c) Sequence-based (word-level)

Fig. 2. Main document representations of KIE methods (best viewed in color).

of ways. For example, each word or even character in the document image can be considered as a node in the graph. Different setups are also possible regarding the definition of the edges, for example creating a fully connected graph where every node is connected to every other node. Instead of constructing graphs, the grid-based approaches aim to create more organized grid structures with well-defined connections primarily characterized by rectilinear links—often based on the document image pixels. The grids are usually defined on different granularity levels, which affects which features the individual grid elements will be assigned to. For example, if the grid is defined on character level, all grid elements that are overlapped by the bounding box of a given character in the document image will have a value that is derived from that character (e.g., constant character index). Sequence-based techniques, on the other hand, convert documents into linear input sequences, ideally preserving and incorporating the document layout and other visual cues. These input sequences are then usually processed by Sequence Labeling methods, where each element is assigned to a particular class. For more details, we refer to [94]. The following Figure 2 illustrates the aforementioned paradigms and how they represent a sample document snippet.

2.4 Input Documents

In the context of DU research, input documents are often referred to as “visually-rich documents”. Although there is no universally accepted definition, an aggregated understanding can be derived as follows: VRDs typically represent complex documents in which the simultaneous consideration of text, layout, and visual information is of great importance in order to adequately capture their semantics [16, 30, 48, 122]. Examples of visual features are font properties such as bold text segments with an increased font size that represent a document title or specific keywords that indicate information to be extracted [16]. Thus, all modalities are important for KIE and converting VRDs to linear text sequences would otherwise result in a significant loss of information. This is in contrast to "simpler" documents such as news articles, where representations as linear text sequences are usually sufficient. This is also the reason why advances in DL enabled the processing of such complex input documents, as they allow for the integration and adequate processing of different input modalities.

The term "business document" appears less frequently in related literature and there is also no universal definition of such documents [15]. In general, business documents contain process-relevant details related to the internal and external operations of an organization, as these documents represent a central means of communication [16, 99]. Business documents, like VRDs, pose a significant challenge to automated DU systems for a variety of reasons, as discussed by [75]. For example, corresponding documents exist in a variety of different formats and are often only available in scanned form due to their paper-based distribution. A scanned document is reduced to its image data and therefore requires techniques such as **Optical Character Recognition** (**OCR**) to make the content machine-readable [82]. The layout and overall content of business documents can vary from highly structured to highly unstructured. Furthermore, business documents can have relationships to other documents and/or may consist of multiple documents in a hierarchical fashion. To understand business documents, it is therefore necessary to observe such interrelationships and to understand the temporal relationships in processes.

As mentioned at the beginning, document processing is a central activity in business contexts. Therefore, it is promising to study KIE from a business process perspective and to investigate methods that directly address this challenge. The scope of this work regarding input documents is the intersection of VRDs and business documents. Therefore, we do not consider VRDs that are not typically embedded in business contexts, and at the same time, we do not consider business documents without characteristics of VRDs. An example of the latter is a company's general terms and conditions. Although it can be considered as a business document, it typically appears as a simple text document without enriched visual elements. We use the terms VRDs and business documents interchangeably when addressing this intersection.

3 Existing Reviews on Key Information Extraction

Many surveys are either domain-specific (e.g., healthcare [114]) or cover approaches that deal only with linear text inputs [61]. In recent years, a few surveys have been published that cover DL-based DU methods while also dealing with VRDs. However, not all of them focus on the task of KIE (e.g., [10] cover DLA).

In the following, we discuss the most closely related work. Reference [101] provide a brief KIE section that is limited to discussions on a few relevant aspects. The survey also does not include an in-depth analysis of the approaches. In [7], the authors provide a comprehensive overview of document processing with many different facets. However, it offers limited discussion about KIE. Much refers specifically to NER, where the overall workflow as well as different methods regarding pre-processing and feature extraction are being discussed. The authors discuss some model architectures. Since their chosen search period includes work between 2010 and 2020, a large number of analyzed papers also do not necessarily use DL methods and/or consider VRDs. Reference [16] cover a wide range of DL-based DU topics. KIE, on the other hand, is discussed relatively briefly. The survey discusses aspects related to transformer-based document processing methods from a general point of view. However, details about the employed model architectures and other KIE paradigms are not covered in detail. The survey of [3] focuses on preliminary DU tasks such as text detection

Table 1. Comparison Against Existing Reviews

Distinguishing factor	**[101]**	**[7]**	**[16]**	**[3]**	**[73]**	**[63]**	**[94]**	**[1]**	**Ours**
Search period	n/a	2010-2020	n/a	n/a	2017-2022	n/a	n/a	2014-2023	2017-2024
Detailed search strategy		*****			*****			****	*****
Focus on KIE task	**	***	***	***	**	*****	***	***	*****
Business process perspective		**	*		****				*****
Analysis of DL concepts	*	***	**	**	*	**	**	****	*****
Analysis of key characteristics	**	***	****	***	***	****	****	***	*****
Analysis of KIE datasets	*	*	***	*	*	****	***	*****	*****
Performance comparison		**		**	*	**		****	*****
Consideration of future work		***	**	*	**	***	***	*	*****

and text recognition. KIE is also covered, although only a small number of approaches are included. The authors present some individual approaches in detail, however there is a lack of comprehensive overview. In [73], document processing is positioned in the context of Robotic Process Automation. Although an extensive literature survey is conducted, there is no in-depth technical discussion of KIE approaches for VRDs. An overview of key methods for KIE is provided by [63], albeit at a relatively high level. The authors however provide an extensive outlook on future work. Besides KIE, [94] also cover VQA and document classification. The authors propose a taxonomy for DU along different dimensions and focus on KIE benchmark datasets. Some challenges as well as future work are also discussed. Reference [1] present a review of transformer-based methods for DU tasks. They highlight key paradigms of corresponding models and showcase a few approaches in detail. A major focus of the review is the description of benchmark datasets and related performance comparisons. In general, the survey chooses to discuss a few transformer-based approaches in detail, but does not present a broad overview of DL-based approaches with respect to KIE.

Overall, existing surveys either examine KIE from a very broad perspective, cover only a few approaches in detail, or alternatively include a higher number of published work, but at the expense of a detailed discussion of the underlying methods. This work on the other hand provides an in-depth analysis, both quantitatively and qualitatively. One key distinguishing factor is also the adoption of a business process perspective. As discussed in Section 2.1, the consideration of business processes as well as general domain knowledge is of high importance for KIE systems. To this end, only the review by [73] also adopts a practice-oriented perspective during the analysis. Table 1 summarizes the differentiation of the previously mentioned related work along certain distinguishing factors.

4 Methodology

4.1 Research Questions

The overall research question that is to be answered by this SLR is: *"What is the state of the art in Deep Learning based Key Information Extraction from business documents?"*. To this end, eight further research questions have been defined:

- RQ1: Which input modalities are considered and how are they integrated?
- RQ2: Which DL architectures are being used?
- RQ3: According to which criteria can existing approaches be categorized?
- RQ4: Which input documents are considered?

- RQ5: To what extent are practical applications and domain knowledge discussed?
- RQ6: Which are the best performing approaches?
- RQ7: Are there noticeable trends in the proposed approaches and architectures?
- RQ8: What potential for improvement can be formulated for follow-up research?

The first five research questions cover an in-depth analysis of the identified approaches, with the aim of examining the proposed methods and how they differ from each other. This includes key aspects such as input modalities, model architectures, and data bases. Based on this, research questions six to eight adopt a more aggregated view and aim to identify the state of the art and derive recommendations for follow-up research.

4.2 Search Procedure

The following six databases were used to obtain relevant literature: ACM Digital Library (ACM), ACL Anthology (ACL), AIS eLibrary (AIS), IEEE Xplore (IEEE), ScienceDirect (SD), and SpringerLink (SL). The search strings were carefully designed to include the commonly used terms for the research area as well as relevant keywords for the target domain (VRDs) and the application of DL architectures. If supported by the search engine, we also included terms that should not appear to filter out papers that are outside the scope of this work. The search period was limited to the range 2017 to 2024. Considering 2017 as the lower limit is a significant help to avoid false positive results, as KIE related literature typically did not use DL methods before 2017. The final search strings for each database, including the number of results at the time the query was run, can be found in table ST1 of the electronic supplementary material.[4]

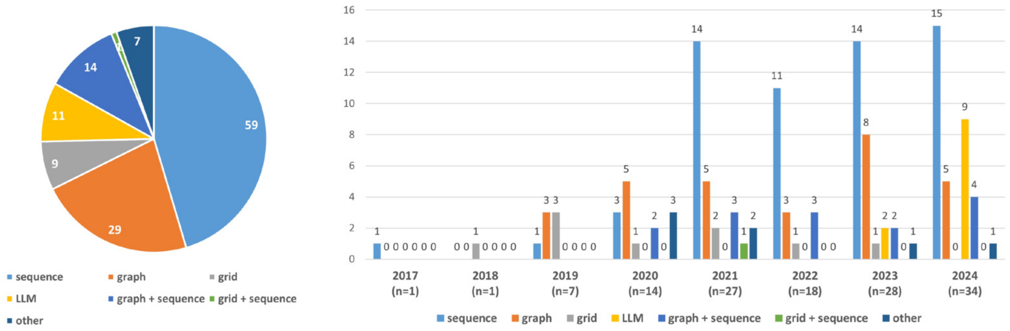
We defined the following inclusion and exclusion criteria as the basis for the literature screening. First, the title, abstract, and conclusion were analyzed. If no violation of the criteria was identified on the basis of these parts, the full texts were analyzed subsequently.

- IN1: The work is peer-reviewed and related to the research area of KIE.
- IN2: The work employs DL concepts and outlines its architecture.
- IN3: The work evaluates the effectiveness of the approach in a quantitative and/or qualitative manner.
- EX1: The work is not written in English and/or was published outside of the defined search period.
- EX2: The work does not propose a new approach for KIE, but rather applies existing approaches to different use cases, conducts a survey of existing approaches, and/or only proposes a novel KIE dataset.
- EX3: The work does not apply the approach to VRDs, but to other domains and/or considers text-only inputs.
- EX4: The work does not focus on KIE, but on other DU tasks and/or focuses only on image pre-processing tasks.
- EX5: The work focuses only on extracting information from very specific document elements such as tables.

The overall search process according to the PRISMA [83] is visualized in figure SF2 of the electronic supplementary material.[5] A total of 130 relevant approaches were identified for this SLR. A list of the identified approaches as well as their numbering, which is also used as a reference in this manuscript, can be found in table ST2 of the electronic supplementary material.

[4] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

[5] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf



(a) Overall distribution

(b) Distribution over time

Fig. 3. Distribution of categories.

5 Results

5.1 Overview

First, we provide a high-level overview of the analyzed work. To this end, table ST3 of the electronic supplementary material[6] shows the results of analysis along different properties, each of which considers different aspects, also with practical applications in mind. Figure 3(a) visualizes the overall distribution of the KIE paradigms. Almost half of the approaches belong to the group of sequence-based methods. Two further common categories are graph-based systems and approaches that combine graph-based and sequence-based concepts. The grid representation is not as widely used as the graph representation, as only nine approaches are purely based on this paradigm. **Large Language Model** (**LLM**) based systems represent an emerging category with 11 identified approaches. Also, only seven of the analyzed papers cannot be assigned to any of the groups, which underlines the dominance of the previously mentioned paradigms for KIE systems. Interestingly, only the work by [113] combines grid-based and sequence-based concepts into one KIE approach. Based on the distribution of the methods over time, as visualized in Figure 3(b), one can see an increased interest in KIE research, especially since 2021. It is also noticeable that the dominance of sequence-based approaches first started in 2021, while graph-based methods were the most common group of approaches before that. The increased popularity of sequence-based methods in this period can probably be attributed to the influential work by [120], which was among the first to show state-of-the-art results using a transformer architecture and extensive pre-training procedures, which in turn led to a lot of follow-up research in this direction. There is also a large increase in graph-based methods in 2023 compared to previous years and compared to the other categories. This category has therefore remained relatively popular over time. Overall, 2024 had the highest number of papers published, showing that there is still a high level of interest in this area of research. In 2022, however, there was a considerable decrease in the number of corresponding KIE approaches. One reason for this observation may be the aforementioned increased interest in sequence-based methods, which typically require extensive pre-training. Therefore, many scholars spent 2022 on developing corresponding methods and on academic write-up, leading to a publication some time later in 2023 or 2024. Given the emergence of LLMs, they also start to appear in KIE research in 2023, with a significant increase in 2024, already becoming the second most common category.

The analysis of the integrated input modalities shows that the vast majority of approaches integrate at least textual and layout-oriented features. This is not surprising, since these elements contain the most relevant information for document processing. Layout modalities are usually derived from

[6] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

bounding boxes, i.e., the coordinates of words in document images. Around 60% of the analyzed approaches integrate visual features obtained from the document images. As discussed in Section 2.4, the integration of such visual cues into KIE systems can be crucial for a proper understanding of complex VRDs. It could be argued that—given this relevance—there are relatively few approaches that integrate image modalities. Nevertheless, the results show that visual cues have been increasingly integrated into the models over time, especially since 2021. In this regard, [81] extensively investigate different methods to obtain and fuse visual modalities into KIE systems. They show that the choice of how visual information is represented depends on the underlying VRDs and that attention-based fusion mechanisms can outperform more basic methods such as concatenations of features vectors. In general, integrating multiple modalities can play an important role for DU systems, as they incorporate cues from different aspects that complement each other. For example, the image modality can provide relevant insights for documents with otherwise limited amounts of text and bounding boxes, while the layout modality is crucial for documents with very complex layouts and distinct structures [123]. Some approaches use hand-crafted input modalities. In most of these cases, they are integrated through custom input features such as boolean flags, whether a word in the document represents specific entities such as dates or monetary values. Another feature type is information about the position of a particular word with respect to the overall reading order. In some papers, hand-crafted inputs are based on syntactic features such as the numbers of characters of a word, details regarding the fonts or encoded character representations. The use of hand-crafted features is also often associated with grid-based approaches. Such methods usually define a custom encoding function that assigns a specific value to each element of the grid. For example, [85] assign a constant value to each pixel in a document image, such as an integer index of a character at each position—or the value 0 for blank parts of the document image. None of the approaches integrate meta data resulting from related business workflows as discussed in Section 2.1, therefore lacking a practice-oriented perspective in this regard. Overall, hand-crafted features provide additional information beyond data that is explicitly contained in documents and can thus improve KIE systems. It must be said, however, that hand-crafted input features can potentially lead to a considerable amount of additional work in terms of labor-intensive data annotation [91], which could also explain the less frequent use of this feature type. One could also argue that the inclusion of such hand-crafted features can potentially result in overfitting to the domain of the training documents—leading to poorer generalization capabilities across different domains. However, there is no work that analyzes this tradeoff between integrating hand-crafted input features and the generalization capabilities across different domains.

In terms of the underlying data basis, the median number of employed documents is 2,172. This is a relatively low amount, especially since the data basis usually has to be split into training and evaluation partitions. As a result, many of the approaches are often only trained on a small document corpus and thus probably with little variety in terms of layouts. Since DL-based DU models usually require large amounts of data for training [91], it is debatable whether the proposed approaches have reached their full potential. Note that these observations do not include the data basis that is being used for pre-training purposes, which will be discussed in the context of sequence-based approaches in Section 5.2.1. Deviating from the average, [53, 84, 93] employ more than one million documents for implementing their KIE system. On the other hand, in some cases, only 199 documents are used for developing the models, while still achieving promising extraction results. This indicates that there are major differences in terms of data requirements of the individual approaches and that some architectures can manage with significantly less data. The vast majority of considered document types are receipts and forms, which are used by at least half of the approaches. The main reason for this is the fact that the most frequently used benchmark datasets are based on these document types. Besides, invoices are most often considered when authors do not use public benchmarks, but rather private in-house datasets.

Regarding the aspect of reproducibility, implemented code is available for 50 approaches. In most cases, authors share their code directly, however there are exceptions where implementations have been made available by external sources through re-implementation. Model weights are even less commonly shared, although this would be particularly helpful for pre-trained models, as improvements and optimizations could be made by external parties on this basis. When either code or model weights are shared, they are mostly provided with a license that allows for commercial use. This is beneficial for organizations that want to integrate corresponding KIE systems into their own business processes, which in turn helps to disseminate the research efforts.

A few approaches are independent of external OCR engines and are therefore either responsible both for text reading and information extraction (end-to-end) or alternatively require no text reading stage at all and map input VRDs immediately to desired outputs (OCR-free) [94]. The system by [4] is only end-to-end during the pre-training phase. Also, [53] propose a system that can be designed in an end-to-end manner, however the authors chose an external OCR engine in their work. These approaches first appeared in 2019, but became more frequent since 2022. The same is true for the 23 identified generative KIE methods, which first appeared in 2021, but now see an increased interest, especially given the rise of generative LLMs. This category of KIE methods is able to output arbitrary texts using autoregressive decoding mechanisms. Key advantages are that corresponding systems are not necessarily influenced by faulty OCR extractions as they can generate OCR-free representations of the target words [13]. They can also be more easily adapted to different DU tasks, for example by using distinct textual prompts [18, 51]. More traditional KIE methods that use a classification head would on the other hand need to be retrained with a different classification layer in order to perform a different task like document classification [18].

Only 10 papers explicitly use domain knowledge. The integration of domain knowledge mostly consists of hand-crafted input features, as discussed before. This again emphasizes the lack of a practical perspective in related literature, as corresponding KIE systems do not consider insights that can be obtained from real-world workflows. Also, only three approaches have been evaluated in real-world industry settings. In these cases, the authors show the impact of the developed KIE methods on real-world document processing tasks—for example in terms of efficiency improvements. Therefore, there is a severe gap between research advancements and real-world applications as the majority of proposed models are not evaluated in practical settings. It is therefore possible that innovations that improve benchmarks may not be immediately suitable for real-world scenarios, e.g., due to extensive data requirements. 16 of the analyzed papers represent an evolution of an existing approach. A prominent example is the LayoutLM family. Based on the original LayoutLM [120], many subsequent refinements and improvements in various aspects have been proposed. Another example is Chargrid [48], which has been the basis for many other grid-based approaches (see Section 5.2.3). A few of the analyzed approaches are implemented using weakly-annotated data. In most cases, this means that besides the document images, only the textual target values are required. Some other approaches only require annotations on segment-level or word-level. Corresponding approaches therefore do not need to be trained on fully annotated documents, which usually involves a labeling of each word including its bounding box and entity. Such approaches can be more suitable for real-world applications where data annotation is costly.

The evolution of model sizes in terms of their number of parameters (in millions) is shown in Figure 4. Note the necessary logarithmic scale of the y-axis due to the approach by [41], which is based on GPT3 and therefore includes 175 billion parameters. The orange line indicates the trend over time. Based on the analyzed approaches, the median number of trainable parameters is 170 million. Most of the approaches tend to have between 20 and 500 million parameters, which did not change significantly over time. However, with the advent of LLMs, more and more approaches with several billion parameters have been proposed, as shown in the figure.

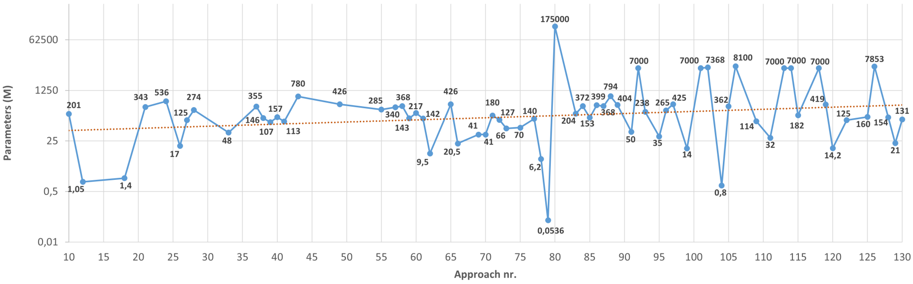


Fig. 4. Evolution of parameter counts.

5.2 Categories

5.2.1 Sequence-Based Methods. Table ST4 of the electronic supplementary material[7] presents the results regarding the analysis of sequence-based approaches. We also include hybrid KIE approaches that are based on multiple paradigms (e.g., a combination of graph-based and sequence-based methods) in this analysis and present the findings regarding the sequence-based subsystem. Most of the methods are defined on word-level, i.e., each word of a document image represents one element in the sequence. In some cases, a segment or sentence-level granularity is chosen, which considers sentences as semantic entities and subsequently calculates aggregated embeddings. Other possible granularities—albeit rarely used—are character-level, token-level or cell-level. The word-level definition is popular because it balances semantic richness with computational efficiency. It also aligns well with the OCR outputs and pretrained transformers like LayoutLM. While more fine-grained definitions can improve structural understanding, they are less common due to increased complexity and alignment challenges. Some authors use multiple granularity types simultaneously, allowing the system to consider both granular and coarse features, which can complement each other.

Around half of the sequence-based approaches do not consider a pre-training of the models and subsequent fine-tuning steps for the KIE task. This is somewhat surprising, as it has been shown that obtaining a general DU through pre-training is a promising research direction. If pre-training is conducted, the average number of documents used is around five million. This highlights the relatively large amount of data that is required to pre-train corresponding models. The largest document corpus for pre-training is used by [12], which consists of 43 million documents. Deviating from this is the approach of [78], which only uses 5,170 documents for pre-training procedures, while still achieving competitive benchmark results. Pre-training is usually very resource-intensive. For example, [120] report that their largest model variant required 170 hours to finish one training epoch of a dataset consisting of six million documents. In around half of the cases, the IIT-CDIP dataset is used. This dataset has been proposed in [59] and includes documents of the Legacy Tobacco Document Library[8] regarding lawsuits against the tobacco industry in the 1990s. Another common dataset is RVL-CDIP, which however represents a subset of IIT-CDIP and therefore includes no additional documents. Also, in seven cases, a private document collection has been used for pre-training procedures. Besides these common datasets, there exist other large datasets such as DocBank, which are however rarely used. Table ST4 also lists the employed pre-training tasks, with the most common being **Masked Visual-Language Modeling** (**MVLM**) proposed by [120]. This task is strongly related to the more general Masked Language Modeling, which was originally

[7] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf
[8] https://www.industrydocuments.ucsf.edu/tobacco/

introduced with BERT [21]. For the sake of MVLM, random tokens of the input sequences are masked where the goal of the model is to predicted the masked tokens given the surrounding context. Importantly, positional information is not masked and thus helps the model with the reconstruction. In this way, the model learns language contexts while also exploiting positional information. Some authors transfer this idea to the image modality and employ Masked Image Modeling, where the model needs to reconstruct masked image tokens. However, a limitation of this task might be the fact that document images for the large part consist of white background and thus the model tends to simply predict white pixels without considering surrounding text and image tokens [98]. Various other related tasks have been proposed over time, which also explicitly include image modalities. This allows KIE models to fully exploit textual, positional and visual information during pre-training. When authors propose end-to-end and/or generative KIE methods, specific pre-training objectives are defined in order to learn text reading and producing outputs in an autoregressive manner. See also Section 6.2 for a further discussion.

Regarding the distribution of integrated encoder and decoder architectures, 12 approaches do not incorporate a dedicated textual encoder at all, while 25 do not use a visual encoder. In [41, 85], neither types of encoders are employed. In most cases, BERT or derived variants such as RoBERTA [69] and SBERT [89] are used for textual encoding. This is not surprising, given the popularity of BERT-based models in NLP. Another common encoding method is the LayoutLM-family, more specifically LayoutXLM, LayoutLM, and LayoutLMv2. This can be explained by the popularity of these models in the context of DU as well as their extensive pre-training, which allows approaches to utilize expressive embeddings. This can also explain why many sequence-based approaches do not employ a dedicated pre-training stage, as they can make use of already pre-trained models. Some authors integrate more traditional embedding methods such as N-gram-models, Word2Vec or FastText instead. Regarding visual encoders, one can see that there is no clear dominance of one particular model. Nonetheless, the most common visual backbone is represented by ResNet models [40], which are also popular in related CV tasks. ResNet50 and ResNet18 are the most often used variants of this architecture, however other variants that have been proposed as improvements such as ResNeXt101-FPN and ConvNeXt-FPN are employed as well. Also, some authors incorporate the more recent Swin Transformer [70] as the visual encoder, which captures local and global image features through sequential processing of non-overlapping image patches and shifted window self-attention mechanisms. Interestingly, only the approach by [37] integrates a **Vision Transformer** (**ViT**). This model also leverages transformers for visual tasks, however compared to Swin Transformers, the input images are processed as sequences of patches without hierarchical divisions. **Document Image Transformer** (**DiT**) [62], which in turn is a self-supervised improvement over ViT, is also only used in the approach by [45]. Transformer-based visual encoders might be less commonly used due to their higher computational and training complexity, despite their potential to capture richer visual features than CNN-based models. The distribution of chosen decoders shows that the majority of approaches use sequence labeling layers for decoding. This usually consist of a linear layer and a Softmax layer, which assigns a probability distribution over all possible fields to each token in the sequence. The task of KIE is then performed by choosing the field with the highest probability according to the probability distribution. Besides, some approaches alternatively use LSTMs [42], BiLSTMs [95], **Conditional Random Fields** (**CRF**) [57] or a combination thereof. BiLSTMs are often used as they effectively capture contextual information in a bidirectional fashion. CRFs are useful, as they model the dependencies between labels in a sequence and incorporate a global context. These more complex models can therefore provide additional information compared to a simple Softmax layer and can potentially increase the extraction performance. When looking at the decoder choice over time, one can see that the two aforementioned variants have not been used as frequently since 2022. Also, especially since 2023, newer decoding methods were examined.

Nonetheless, sequence labeling layers remain popular over time. Some authors also use existing KIE approaches for decoding. For example, [19] use the model by [51] for their decoding step.

Overall, it can be said that there are many different approaches to encoder and decoder mechanisms, with no dominant approach. BERT-models and layout-aware variants like LayoutLM are most commonly used for textual encoding, while ResNet-based models are used as visual backbones. Decoding steps are usually performed either by simple sequence labeling layers or by more complex architectures that allow to incorporate certain dependencies within the token sequences.

5.2.2 Graph-Based Methods. The findings regarding the analyzed graph-based approaches are presented in table ST5 of the electronic supplementary material.[9] A key property in which the approaches differ is how the underlying graph representation is constructed. Most approaches use a fully connect graph, which means that every single node of the graph is connected to each other. As a result, the resulting graphs usually have a very large number of edges. Also common is the k-nearest neighbors algorithm, where each node has k neighboring nodes. The approaches use different values for k, ranging from 4 up to 100. Choosing the value for k is no trivial task and might introduce several biases [9]. Also, increasing k will necessarily lead to more complex graph structures that include too many edges [119]. As a result, KIE approaches may not be able to recognize the actual relationships within a document. Nonetheless, there is no study that investigates these aspects in depth. Other common methods are to considering neighbors in the four major directions *up, down, left,* and *right* or using the β-skeleton graph algorithm, which connects nodes based on their geometric proximity depending on the parameter β. To this end, all approaches choose β=1. A few authors divide documents into distinct areas starting from each node (e.g., in 45 degree angles) and identify closest neighboring nodes in each of these segments. Reference [14] emit 36 rays out of each node and declare all other nodes, whose bounding box is crossed by these rays, as a neighbor. In a few KIE systems, the graph creation is not defined by a heuristic algorithm, but rather iteratively learned by a neural network. The reasoning behind this choice is that the aforementioned heuristic algorithms are not always able to adequately represent VRD layouts.

Regarding different methods, how the information is propagated through the graphs, in most cases, a **Graph Convolutional Network** (**GCN**) is employed. GCNs propagate information between the graph nodes to iteratively learn representations, which capture both local and global contexts. Each iteration (i.e., graph convolution) calculates new embeddings for the nodes and edges. In some papers, a variation of a GCN defined on node-edge-node-triplets is used in conjunction with a **Multilayer perceptron** (**MLP**). In this setting, the features of the node itself, its edges, and all neighboring nodes are fused in order to obtain the new node embedding. Another common method for graph propagation is using **Graph Attention Networks** (**GATs**), either with our without multi-head attention. This architecture utilizes the attention mechanism based on transformers to calculate different weight coefficients for each neighboring node, allowing the model to focus on the most relevant information during graph propagation, and thus improving the overall node representations [108]. Yet, they are less used than GCNs, which might be due to the increased complexity and computational overhead of attention mechanisms, especially when dealing with large graphs. With respect to the graph propagation, one can also analyze the number of propagation layers. The number of layers not only determines the overall system complexity, but also the receptive field of each node – determining how far it can aggregate information from its neighboring nodes. Each layer increases the depth of this receptive field by one. On average, the approaches consist of 3.7 layers. Reference [68] conducted ablation studies regarding different numbers of layers and found that the optimal number of GCN layers is 2, which however can be task-dependent and should be considered on a case-by-case basis. With regard to the GAT-based approaches, one needs to decide

[9] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

on the number of attention heads. Reference [8] conducted a study with different attention heads and came to the conclusion that in their case, 26 attention heads produced the best results. The authors also mention that correlations between the number of attention heads and the number of fields to be extracted might exist. Here, too, it is necessary to examine which variant works best in each individual case.

The approaches differ in the way KIE is ultimately performed. Most of the approaches use a simple node classification, i.e., after the embeddings of each graph node have been obtained, they are used to classify each node into one of the fields to be extracted using a classification layer. Also common are BiLSTMs, CRFs, and their combination in the form of BiLSTM-CRFs. Besides, multiple other methods are used to perform the KIE task. For example, [58] use the Viterbi algorithm, which identifies the most likely sequence of states by iteratively calculating probabilities and backtracking through the graph.

In by far the most cases, the graphs are defined on word-level, which means that every word on a document image represent one node in the constructed graph (see Figure 2(a)). This is also inline with the granularities for the sequence-based methods. Another common way to define graphs is to consider entire segments or sentences. The resulting graphs are therefore rather coarsely defined. Some of the approaches consider entire text lines as nodes, regardless of whether the contained text elements are related to each other or not. A few authors define the graphs on more than one granularity simultaneously, both using fine-granular (e.g., words) and more coarse-granular (e.g., text segments) nodes.

Another central component of graph-based methods is the definition of node features. These features determine, which types of information can be utilized for KIE. In general, the node features usually consist of position-oriented values such as normalized coordinates of the node within the document image. Sometimes, relative distances to neighboring nodes are also part of the node features. Another common feature type is related to the textual content represented by the node, e.g., word embeddings. The approaches differ in terms of which embeddings models are used. In this regard, BERT-based-embeddings are often employed, which was also the observation regarding textual encoders of sequence-based approaches. Some other methods are Byte Pair Encoding [97], BiLSTM, and LayoutLM. In some cases, multiple text embeddings are used simultaneously. For example, [130] employ Word2Vec, BERT, and LayoutLM for the word embeddings. On top of positional and textual embeddings, many approaches also integrate visual features, often obtained from models such as ResNet or LayoutLM. Many authors also consider hand-crafted miscellaneous features. The goal of these features is usually to incorporate certain node characteristics as well as relationships between nodes that should aid the model in performing KIE. Reference [71] for example include various boolean features whether a graph node represents a date, zipcode or known city, among others. The approaches also differ regarding the size of the node feature vectors. Although in many cases the actual number is not explicitly reported in the manuscripts, on average, the features vectors have a length of around 500. In general, there is little discussion about embeddings sizes for node features in related work, from which it can be concluded that these have no significant influence on the performance of the KIE systems. Much more important seem to be the different types of features which are being integrated. Another distinguishing factor of graph-based approaches is, whether they incorporate a global node in the defined graph representation. This global node can contain information about the entire document image and can be connected to the individual nodes of the graph. To this end, only the approaches by [44, 131] use such global nodes. Reference [65] also construct a global node, however it is not used as a information carrier and is rather formally required due to the chosen concept based on document layout trees.

Edges represent the second integral part of the graph structures. It is noticeable that a large portion of work does not incorporate any edge features at all. Consequently, less attention is paid to this type of feature compared to node embeddings. However, [9] show that edge features, especially in terms of geometric data, can improve KIE performance, even more so than geometric information obtained from node features, as it integrates spatial information of the surroundings. The edge-features are usually derived from pairwise positional relationships between the connected nodes. Most often, horizontal and/or vertical distances, both in absolute and relative values, are considered. In this regard, [68] highlight the importance of the visual distances between two nodes. The aspect ratio of the bounding boxes that span the graph nodes is also often included as edge features. In only three of the graph-based approaches are visual embeddings part of the edge features. Regarding the distribution of edge directions, i.e., whether the edges are directed or undirected, there is no clear dominant variant, however in the majority of the cases, the edges are undirected. Additional research is required to appropriately estimate the impact of certain design features such as the edge direction on the extraction performance. To this end, [33] show that both directed edges as well as introducing edge weights can improve extraction results. However, as illustrated before, these considerations are usually dependent on the underlying scenario and generalizing statements can only be made to a limited extent.

5.2.3 Grid-Based Methods. As mentioned in Section 5.1, this paradigm is less popular in related literature and only 10 approaches make use of such grid structures. This suggest that while it offers an intuitive document representation, grid-based systems struggle with flexibility and generalizability across diverse and complex layouts of VRDs. The analysis of the approaches is shown in table ST6 of the electronic supplementary material.[10] An interesting observation is that most approaches represent evolutions of existing grid-based systems, in particular Chargrid [48]. This shows that related research is mostly conducted within a very narrow corridor. Regarding the chosen granularities, a majority of the systems is defined on word level, while three are defined on character level and one method even on token level.

Table ST6 also lists findings regarding the grid dimensions. Almost all methods define the height and width of the grid based on the pixel counts of the input image. The grids are therefore rectangular, as this corresponds to the rectangular format of corresponding VRDs. Refs. [31] and [126] use a fixed value for the height and width, namely 512 and 336 respectively, therefore resulting in square grids. Reference [49] scale down the input dimensions by a factor of 8 in order to reduce the overall model complexity. The work of [113] uses a somewhat different approach and employs a dynamic grid structure, in which the height and width dimensions are defined by the distance between the maximum and minimal coordinates of bounding boxes in vertical and horizontal directions—i.e., the span of actual document content. Another distinguishing factor is the feature vector length. Relatively similar values are chosen here, with 256 and 768 being the most common vector sizes. The approach of [31] uses the smallest feature set with a vector of length 35. The most complex feature vector is defined by [84]. In this case, the feature vector has the dimension 4 × 128 × 103.

Regarding which features are chosen for the individual grid elements, the majority of approaches use text embeddings obtained from different models—mostly BERT-based architectures. Another common feature type is a 1-hot encoding of the characters, which means that each character in the word corpus is assigned to a specific index value. A few approaches also integrate visual features based on pixel-wise RGB-values, ResNet18, or Swin Transformer embeddings. Another distinguishing feature is the integration of features for background elements of the grid, i.e., all elements that are not overlapped by contents of the document image. In this regard, almost all approaches use an all-zero feature vector. The aim is to clearly differentiate actual content and backgrounds. Reference [84] use a sparse tensor and therefore discard background elements entirely.

[10] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

Only the work of [49] integrate features for the background elements, namely the RGB-channels of the corresponding coordinates in order to fuse visual information into the architecture.

Most of the methods employ semantic segmentation to perform the KIE task, where each element of the document image is assigned to semantic categories (i.e., fields to be extracted), which results in a segmentation mask of the original document image. Other choices are bounding box regression, where the goal is to predict bounding boxes of semantic entities, which can be helpful to better differentiate individual objects. For example, when processing an invoice, it is helpful to separate each individual line item. This is also the reason why a few approaches combine semantic segmentation with methods for bounding box regression (or line item detection in general) [20, 48, 126].

5.2.4 LLM-Based Methods. With the emergence of LLMs across many research areas, their application has also been studied for KIE in the recent past. Multimodal LLMs in particular have accelerated this process, as they can adequately process VRDs based on different modalities. Due to the nature of the models, KIE is typically formulated as a VQA task. That is, the LLMs receive OCR text as input and are prompted with questions like *"What is the value for the {key}?"*. All LLM-based approaches are generative by design. A key advantage of these approaches lies in the basis of LLMs: they have been pre-trained on extremely large amounts of data, significantly larger than those used in KIE-specialized sequence-based models (see also Section 5.2.1). This allows them to draw on extensive world knowledge, which can result in a higher generalization capability compared to more traditional KIE methods. The analysis of LLM-based approaches is provided in table ST7 of the electronic supplementary material.[11]

More than half of the approaches use an external OCR engine. This shows that robust text recognition remains important and should not be left solely to the LLM itself. This point is underlined by the fact that only [27, 67, 125] operate purely on the image modality. Almost all methods use open-source LLMs as a backbone, with LLama2 being particularly common. In most cases, a visual encoder is used to integrate the visual information. To this end, three of the approaches use LayoutLMv3 for image encoding. The works vary in terms of image resolution. Some use high resolutions such as 2560×2560, while others use lower resolutions like 224×224 but compensated by extracting up to 20 image crops in order to more precisely capture individual document regions. Here, a tradeoff must be made between computational overhead and extraction quality.

Some approaches perform continued pre-training to better adapt the backbone LLMs to VRDs. Tasks used during this phase include text recognition, text spotting, and table understanding. This continued pre-training is often carried out on existing KIE datasets, such as IIT-CDIP. With the exception of [39], all approaches additionally perform **supervised fine-tuning** (**SFT**) for KIE and therefore adjust model weights. The goal is to improve the foundation models' capabilities of processing VRDs and perform KIE. For this purpose, corresponding instruction-tuning datasets are created, often based on popular benchmark datasets such as CORD. This involves converting the document annotations into question-answer pairs similar to the VQA-style question mentioned before. Besides, some authors also include various other SFT tasks such as image captioning where the LLM is prompted to generate textual captions for input documents. In this regard, very few approaches use parameter-efficient fine-tuning methods such as LoRA. Instead, most train all parameters of the LLMs during SFT. This indicates that in order to properly fine-tune the foundation models to VRDs and KIE, it is required to update the weights of the majority of the architecture. However, there is no work that analyzes this aspect in depth.

The approaches also differ in how the different modalities are fed to the LLMs. For example, [39] pass OCR data to the LLM and prompt it to obtain the corresponding class. An example format is: *"Q3:{text: 'TO:', Box:[102 345 129 359]}..., What are the labels for these texts?"*. Incorporating layout

[11] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

information explicitly by adding a list representation of the bounding boxes into the prompt like above is a straightforward and commonly used approach. However, it is associated with limitations. Using this method, a large part of the input sequence is therefore occupied by these tokens, which reduces the available context length for the actual VRD content. Also, the backbone LLMs are typically not (pre-)trained with such prompt formats. This can lead to the model struggling to properly reason about the layout information.

Reference [74] take a different approach by converting documents into an HTML structure and passing it to the LLM in order to obtain a structured JSON output with the extracted information. The authors argue that LLMs are well-suited to processing HTML, as a large portion of their pre-training data consists of HTML. Besides, the common procedure of LLM-based approaches is to obtain image encodings and pass them to the LLM together with the textual prompt to get the extracted value for the respective field. Some works include all fields to be extracted in one prompt (e.g., [74]), while others only consider one field per call (e.g., [125]). The latter requires multiple calls per document, which can increase runtime, but may allow the LLM to better focus on one field and improve extraction quality per field. Authors also experiment with different prompt templates. For example, KIE is sometimes framed as a multiple-choice task: *"{document} What is '{value}' in the document? Possible choices: {keys}"*, where *{keys}* is a randomly selected subset of key names from the dataset [127]. The idea is to leverage the LLM's prior knowledge of field names to better identify the correct value. In general, most works follow similar prompt formulations. However, [132] shows that it is advantageous to create instruction datasets with heterogeneous prompts, as this increases robustness for different layouts of VRDs.

5.2.5 Other Methods. Of the 130 papers analyzed, seven cannot be assigned to any of the paradigms of KIE methods discussed before. In the following, we briefly present key concepts behind those deviating approaches. The approach by [72] is based on learning representations of document snippets and classifying them into fields to be extracted. First, potential candidates are identified for each field based on the data type. The next step is to select the correct candidate for each field. To this end, a representation is obtained for each candidate as well as each field to be extracted. The candidate representation is constructed based on the candidate itself as well as its neighboring segments, utilizing textual and positional information processed by a self-attention mechanism. Finally, a similarity score is calculated for each pair of candidate embedding and field embedding. The similarity scores are then used in a separate module to select the appropriate candidate for each entity, which can be defined in a variety of ways. A trivial method is to select the candidate with the highest similarity score for each entity.

Reference [115] use a hierarchical tree-like structure to represent fragments of the document pages, similar to the graph-based methods. Compared to traditional tree architectures, nodes on the same hierarchy level can also be connected with each other. Parent and child nodes of these trees represent key-value pairs. The KIE task is performed by predicting the relations between the individual fragments, i.e., directions of the edges. Obtaining the most likely parent element of each fragment can be ultimately used for identifying fields to be extracted.

Reference [129] propose an end-to-end KIE system and combine text reading and text understanding. The text reading steps includes a detector model as well as an LSTM-based recognition model which identify text in the document images. The approach also includes a dedicated module to fuse the different input modalities, which is then used by the final KIE module. A BiLSTM coupled with a fully connected network is employed as a decoding mechanism to predict relevant entities for the fields of interest. This approach is therefore closely related to sequence-based KIE methods.

Reference [93] follow concepts of [72] in the sense that for each field to be extracted, a set of candidate spans is identified first. In order to identify candidate spans for a named entity, multiple detector functions are implemented, to some extent based on domain knowledge. The authors then employ an adversarial neural network in order to find local contexts of the identified visual spans. These local contexts represent specific fragments of a document image that contain the spans as well as related relevant context (e.g., a larger paragraph). Based on these identified segments, both global and local context vectors are constructed with textual and visual features. These features are then used by a binary classification model in order to predict, whether a visual span contains a specific field to be extracted or not.

The approach by [104] is another method that is based on generating candidates for each field and scoring them subsequently. Given a document image and fields to be extracted including their data types, the system first identifies candidates in the OCR-output based on third-party detector functions. The identified candidates are then scored according to their likelihood of correctly representing a particular field of interest in a binary classification setting. Similar to [72], the score depends on the similarity between the embeddings of the candidate and the field of interest. Different scoring functions can be integrated, however the authors choose a function that assigns the candidate with the highest similarity score to each field.

Reference [103] propose a novel document representation structure which they refer to as cell-based. This representation is similar to grid-based methods, however the cell-based methodology has no consistent placement of elements in terms of height and width. Instead, the cells are defined depending on the actual document content. For example, different lines or columns in the cell-structure can consist of a different number of elements. The individual cells are also sorted by row and column index respectively, which provides additional information to the KIE system. The obtained cell-based layout is then processed by sequence-based methods such as LayoutLM and therefore follow the typical sequence labeling scenario that respective models use to perform KIE.

Instead of processing entities sequentially or constructing elaborate graph structures, the approach by [66] constructs a token pair representations matrix based on multi-modal features from a pre-trained encoder. It then jointly generates three relation matrices: line extraction (to determine the boundaries of text lines), line grouping (to merge related lines that form multi-line entities), and entity linking (to connect keys with their corresponding values). These predictions are then combined to ultimately obtain extracted key-value pairs.

5.3 Evaluation

5.3.1 Methodology and Setup. Table ST8 of the electronic supplementary material[12] shows the findings regarding the presented evaluation setups. A consistent picture throughout is that authors often do not describe their evaluation procedures in great detail and it is also not always obvious from the manuscript itself. This is the case for around half of the analyzed papers. We declared affected properties as “n/a”. In 19 of the analyzed papers, an element-based evaluation method is chosen, which means that elements like the predicted class of a token are compared with the groundtruth class in case of sequence-based approaches. Another common element-based evaluation is to compare the predicted bounding box with the actual bounding box and to identify a match or mismatch based on the overlap of their respective coordinates. 45 of the analyzed approaches incorporate a string-based evaluation setting. In such cases, the extracted textual values are compared with the groundtruth strings to determine, whether a prediction for a given field was correct or not. A string-based evaluation is particularly relevant for assessing the suitability for real-world applications where the extracted texts are used for further processing (e.g., transfer to other information systems).

In almost all of the manuscripts, the authors present the overall performance, i.e., the aggregated results across all fields of interest and all documents of the test set. However, the field-level

[12] https://dl.acm.org/doi/suppl/10.1145/3749369/suppl_file/csur-2024-0698-File002.pdf

Table 2. Common Datasets for KIE Research

Dataset	Documents	# docs	# classes[a]	Language	# uses[b]	Pre-Training?
FUNSD	Forms	199	4	eng	61	✗
CORD	Receipts	1,000	30	eng	52	✗
SROIE	Receipts	973	4	eng	45	✗
IIT-CDIP	Lawsuits	6,000,000	/	eng	24	✓
XFUND	Forms	1,393	4	zho,jpn,spa,fra,ita,deu,por	15	✗
EPHOIE	Exams	1,494	10	zho	9	✗
RVL-CDIP	Lawsuits	400,000	/	eng	7	✓
Kleister-NDA	NDAs	540	4	eng	6	✗
Kleister-Charity	Reports	2,778	8	eng	6	✗

[a] in terms of entities to extract
[b] as part of analyzed approaches

performance, which shows the performance for each field individually, is only adopted in around 30% of the analyzed papers. One aspect that is almost never presented is the performance on the *Unknown* class. In the context of KIE, *Unknowns* represent all elements of a document that do not belong to any of the fields to be extracted. This can be helpful in determining the extent to which an approach is able to distinguish irrelevant content from relevant parts of a document image. In most cases, the proposed method is compared against existing KIE systems, either by comparing the own results with the evaluation metrics presented in the respective papers or by re-implementation. Thus, it can be stated that a high emphasis is placed on the overall comparability and the legitimization of the own approach. However, especially when authors decide to re-implement existing approaches, comparisons of evaluation results are not always appropriate, since less effort is usually put into training the approaches of third parties.

Considering the employed evaluation metrics, one can see that the F1 score is used most of the time. Precision and Recall, with the F1 score being the harmonic mean of these two metrics, are only explicitly reported in about 25% of the cases. Interestingly, the usually widely used Accuracy metric is only presented by a few authors. One reason for this is that KIE datasets are usually very unbalanced in terms of label distribution, which strongly distorts the overall Accuracy metric [32] – making it less representative for these scenarios. Nonetheless, given that KIE is usually seen as a multi-class classification problem, it is not surprising that authors adopt metrics such as Precision, Recall, and F1, which are very common for such tasks. As mentioned before, many authors present a string-based evaluation. To this end, various different metrics are used, e.g., Word Accuracy Rate, which counts the number of substitutions, insertions, and deletions between groundtruth and predictions [49]. Moreover, it is noticeable that a large majority of the identified evaluation metrics are custom-defined and only used in one particular paper, which prevents an adequate benchmark against existing KIE approaches.

Custom datasets are being used in 54 of the analyzed papers and typically represent in-house document collections that are not publicly available. However, in around 80% of the cases when custom datasets are used, the authors additionally present their results on public benchmark datasets to allow for a comparison with other KIE systems. In addition, there are numerous datasets that were proposed in individual papers, but were not adopted in any other subsequent work. This indicates again that KIE literature strongly focuses on a small set of benchmark datasets.

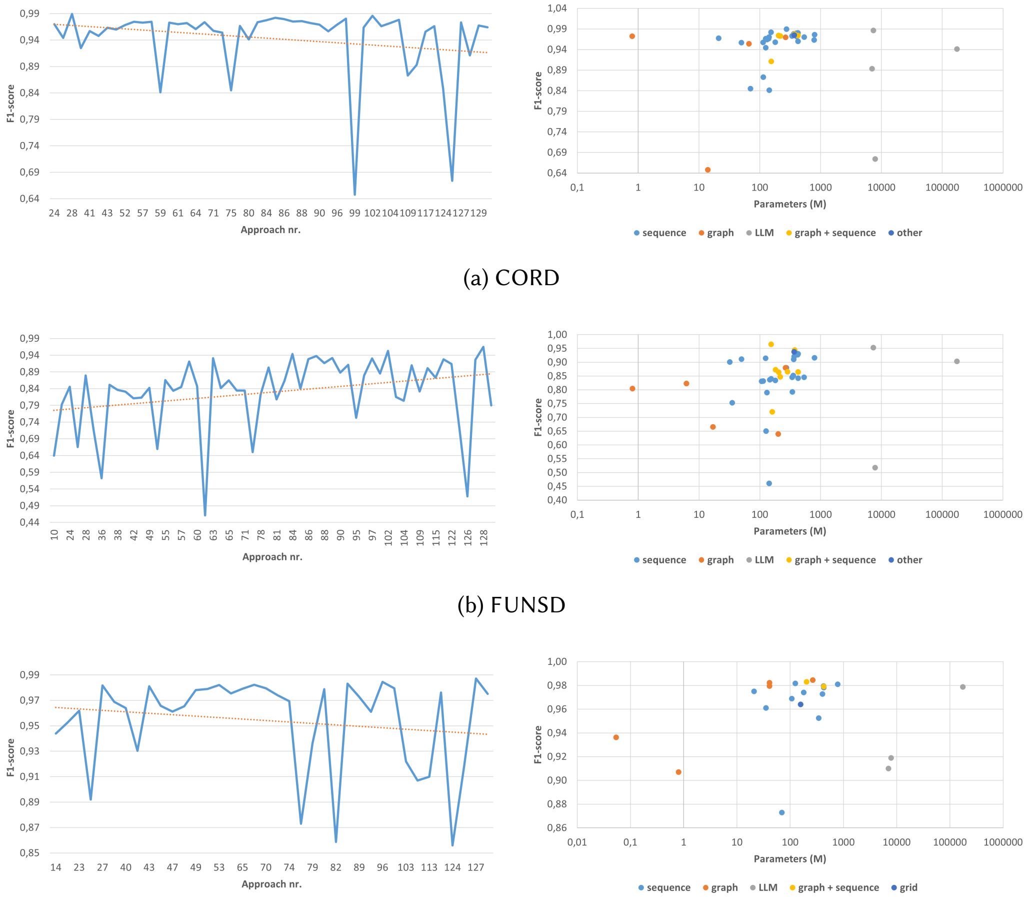
Table 2 provides an overview of the most common datasets adopted in KIE research. Listed are benchmark datasets as well as pre-training datasets. By far the most common datasets in this regard are FUNSD [47], CORD [86], and SROIE [46], all of which are used by at least a third of the papers. FUNSD includes forms from various domains. CORD and SROIE on the other hand contain photographed receipts, mostly from supermarkets and restaurants. There are discrepancies both in terms of dataset size and in the number of keys to be extracted. The datasets with many hundreds of thousands of documents are typically used for the pre-training of corresponding models and not as an evaluation benchmark. Therefore, these datasets are also widely adopted in other DU tasks. Two of the three most used datasets only aim at extracting four fields of interest, which is a relatively low amount compared to the variety of information corresponding documents usually include. CORD on the other hand includes 30 fields to be extracted, which is also the highest among the listed datasets. The key difference between CORD and SROIE is that in case of former, detailed information including individual line item attributes such as their quantity or unit price need to be extracted, while SROIE only requires to extract aggregated information such as the total price. For the most part, the datasets include English documents. One exception is XFUND [121], which specifically intends to investigate multi-lingual capabilities of KIE approaches. Some popular Chinese datasets exist, namely EPHOIE and Ticket, however they are understandably not as widely adopted as most English counterparts that are being used by the international research community.

5.3.2 Quantitative Comparison. To ensure a meaningful and representative quantitative comparison, in the following we focus on the by far most commonly used datasets CORD, FUNSD, and SROIE. We also only consider the F1 score, as it is the most commonly used evaluation metric. These choices allow for the largest possible sample size and a comparability across a wider range of approaches. Note that we do not distinguish between micro, macro or weighted F1 averages,[13] since in many cases it has not been explicitly specified by the authors. The results should therefore be treated with caution.

To this end, Figure 5 shows the results for the three datasets given all papers that present results on the corresponding datasets. A trend line is also displayed. Only FUNSD shows an improvement over time, while for the other two datasets the trend line shows a decline. However, all graphs show a few outliers that performed worse than previous and subsequent work. One reason for this could be a different calculation of the F1 scores, e.g., by using macro averages instead of micro averages. Another possible factor could be the alignment of the models with the specific nature of the task. For example, models that do not take into account the spatial and semantic features of the datasets may perform poorly despite having a large number of parameters. These outliers will therefore negatively affect the trend line. The best performing approach for CORD is proposed by [36], for FUNSD by [119], and for SROIE by [112]. On average, very comparable results are obtained for CORD and SROIE, with an average F1 score of 0.94 and 0.95, respectively. This can be attributed to the fact that both datasets consist mainly of structured receipts, where layout and spatial encoding play a crucial role in performance. In contrast, FUNSD shows significantly worse results, with an average F1 score of only 0.83. One problem that KIE approaches face with FUNSD may be that this dataset was not primarily constructed for the KIE task. The fields to be extracted often span multiple lines of text, which is very different from traditional KIE datasets that aim at extracting key information such as a specific date.

Figure 5 also shows the relationship between parameter count and performance across the three datasets, with the x-axis in logarithmic scale. In all cases, there is a clear clustering with some outliers. The correlation coefficients are -0.02 (p=0.90) for CORD, 0.08 (p=0.60) for FUNSD, and 0.11 (p=0.63) for SROIE, indicating no significant correlation between model size and performance. Notably, the slight negative correlation for CORD suggests that smaller models can outperform larger ones. The figure also distinguishes KIE paradigms by color. Notably, graph-based methods

[13] For an explanation of these aggregated metrics, we refer to [100].



(a) CORD

(b) FUNSD

(c) SROIE

Fig. 5. Evaluation results and correlation between model size and performance.

often appear as outliers. Especially in case of FUNSD, this observation can be attributed to its unique layout, where entities often span multiple lines, making graph representations suboptimal. Hybrid models combining graph and sequence-based approaches tend to outperform pure graph-based models, especially in this scenario. The analysis further shows that increasing model size beyond a certain threshold yields diminishing returns. This suggests that architecture design, tailored to the task, is more critical than model size. This insight is particularly useful in resource-constrained environments, where smaller, task-aligned models can achieve competitive or superior results. The performance spread in FUNSD is wider compared to CORD and SROIE, with some smaller models struggling, likely due to the complex layout and semantic requirements of the dataset. Interestingly, the LLM-based approaches often do not achieve competitive results and in some cases are among the worst performing KIE systems. Only the LLM-based approach by [29] achieves almost state-of-the-art results for CORD and FUNSD. This highlights again that larger models don't always lead to better performance if the architecture is not suitable. On the contrary, the approach by [34] with only 0.8 million parameters achieves promising results and is even one of the best systems in case of CORD.

In conclusion, this analysis emphasizes that model design and task alignment are more important than model size. For structured documents as seen in CORD and SROIE, simpler models with strong layout awareness are effective, while hybrid models combining sequence and graph-based features work best for more complex KIE settings like FUNSD. Architectural choices need to balance semantic complexity with layout structure to optimize KIE performance. It is important to note that the findings are based on datasets with strong domain overlap—in particular CORD and SROIE, which consist of receipts—and that all three analyzed datasets are in English. As a result, the findings may have limited generalizability to other application domains or languages. The outlined performance tradeoffs may be more pronounced for other document types and layouts, where larger models could yield significantly better results.

6 Discussion

6.1 Technical Analysis

KIE approaches vary widely in how they represent and process VRDs, with strengths and weaknesses in handling multi-modal inputs, spatial document structures, and semantic richness. Sequence-based models, such as LayoutLM and its variants, dominate due to their ability to leverage large-scale pre-training and various attention mechanisms. These models incorporate spatial information via positional embeddings and often use linear classification layers to perform NER. However, such classification layers are limited in modeling label dependencies across sequences—unlike CRFs—and struggle with discontinuous entities, which are common in real-world documents [107]. Moreover, while end-to-end models provide a unified extraction pipeline, OCR-dependent systems still achieve superior performance in benchmarks since they decouple the complex tasks of text recognition and entity extraction [96]. Graph-based models excel at capturing document structure through spatially and semantically defined graph structures. At the same time, GCNs and GATs allow for effective layout-aware propagation. However, many approaches rely on heuristic graph construction, which can introduce noise or bias. In contrast, learned graph structures can show improved robustness and generalization at the expense of model complexity [119]. Edge features remain underutilized, despite evidence of their potential to improve spatial reasoning. Grid-based approaches treat documents as grids and apply segmentation-like techniques for KIE. These methods are particularly suited to highly structured VRDs, but are less flexible in adapting to different layouts. They also tend to use limited semantic encoding, with little advancements beyond early design choices such as those proposed by [48]. This is also reflected in the fact that research on grid-based approaches was mostly conducted before 2022. Novel LLM-based methods introduce a generative and instruction-driven paradigm for KIE. While these models offer strong generalization and task flexibility through their use of large foundation models, they have practical limitations. Layout information is often injected via simple prompt engineering, which is not native to the LLM training procedure and can impact efficiency. Furthermore, inference times are significantly higher—under certain circumstances up to 120 times higher than other approaches on comparable hardware [88]—which poses scalability challenges. Although hallucinations (i.e., the generation of text not present in the original document) are rare, occurring less than 1% in some cases [88], dedicated mechanisms may still be required to prevent them.

Almost all KIE approaches have limited capabilities to handle long and multi-page VRDs. This is mainly due to architectural constraints, e.g., transformers have a fixed input sequence length, usually 512 tokens, which is often not even enough to capture even a single page of a VRD. Graph-based and grid-based methods are unable to capture multi-page dependencies because their document representations are defined page by page. LLM-based approaches can overcome this problem to some extent, as the backbone LLMs usually provide longer context lengths (e.g., 4,096 tokens). This aspect is less of a problem when processing VRDs with fewer words (e.g., receipts as seen in CORD and SROIE), but can play a role in documents with more content, such as forms (see FUNSD). All in all, an adequate understanding of long VRDs and maintaining context over long sequences remains a challenge.

Overall, the different categories require tradeoffs between semantic expressiveness, spatial reasoning, and practical constraints. Transformers offer strong contextual understanding but need improvements in sequence modeling and layout perception. Graph-based methods capture document structure explicitly and can benefit from learning the graph connectivity, rather than relying on heuristic algorithms. Grid-based models handle layout well, but lack semantic flexibility. LLM-based approaches allow for a broad adaptability, but remain computationally intensive and less effective at precise spatial reasoning.

6.2 Pre-Training and Fine-Tuning Procedures

An increasingly popular approach in KIE research, in particular when dealing with sequence-based methods, is to split model training into self-supervised pre-training and SFT steps. The aim of the former is to acquire a general DU by the model first, which can then be utilized for different downstream tasks such as KIE as part of subsequent fine-tuning steps. In general, pre-training procedures require large volumes of data to properly let models learn corresponding document representations, which is why most authors use large datasets such as IIT-CDIP, as also discussed in Section 5.2.1. The problem with this dataset in particular is that included documents originate from the 1990s and therefore show poor image quality, noise produced by faulty scans as well as a generally low image resolution. This also means that a large proportion of the proposed approaches are pre-trained on outdated documents and may therefore have difficulties when they are fine-tuned on newer documents as part of real-world use cases. Another aspect that should be considered is the fact that all documents of IIT-CDIP are related to lawsuits against the tobacco industry and therefore lead to a strong domain bias that is being adopted by the models during pre-training. In this respect, it is debatable whether it would not be more effective if pre-training procedures were carried out based on more representative datasets (see also Section 7).

As discussed in Section 5.2.1, the majority of authors adopt MVLM during pre-training. In some cases, it is even the only pre-training objective that is being used. Given the general popularity of MLM in the NLP domain, it is not surprising that the MVLM task, which is derived from it, is also popular in KIE research. It represents an effective method to let models learn useful representations given the surrounding textual and positional contexts. As the adequate processing of VRDs also requires to consider visual cues, newer pre-training objectives have been proposed in order to fuse visual information into KIE systems. This can be done in different ways, for example by letting the model reconstruct local or global document image snippets [4, 106] or predicting the segment length of an image snippet [64]. Other possibilities aim at various matching-tasks between text and document images, for example where the model has to predict whether a sentence describes a document image [4], if a given image and text are part of the same page [122] or if image patches of a word are masked or not [45]. Some authors also define additional pre-training objectives to better exploit positional information, for example by the means of classification problems to estimate a tokens placement into a document image area [60, 111] or to estimate relative positional directions of an element to its neighbors [64]. Research is also conducted related to a better understanding of numerical values and their relationships in business documents [24]. This can be beneficial for processing document types such as invoices, which usually contain various monetary values that are closely related to each other. To conclude, the proposed pre-training objectives are carefully designed to allow for an adequate fusion of textual, positional, and visual inputs, which helps corresponding models to obtain a DU which can then be exploited in downstream tasks like KIE.

Another category of pre-training objectives stems from the generative KIE methods, where corresponding models need to learn both natural language understanding and natural language generation (i.e., generating output sequences conditioned on input sequences) capabilities. To this end, [12, 13] adopt the pre-training tasks defined by [23], namely Unidirectional LM, Bidirectional LM, and Sequence-to-Sequence LM. Importantly, all pre-training tasks are considered simultaneously by using a shared transformer network which can alternate between the three objectives. Reference [106] propose a range of self-supervised and supervised generative pre-training objectives based on task prompts and target outputs in textual form, for example where the model must predicts missing texts and locate them in document images as a structured target sequence. The approach by [18] follows a specific strategy for learning text recognition, DU and generative capabilities. It generally stands out in terms of its pre-training setup, as over 25 different objectives across different document types are being used. Nonetheless, a strong focus is placed on MLM-related tasks. One important pre-training task revolves around the model predicting a structured JSON output for a given document, which is also ultimately used for KIE. This strategy can be helpful in real-world settings, in which a detailed hierarchical output of input documents is required, such as an adequate differentiation between individual line items in invoices. Non-generative KIE methods that for example decode outputs with a sequence labeling layer usually have difficulties in reconstructing such hierarchies and require additional post-processing steps. To let the model learn text recognition, [22, 51] use a pre-training task where the model must predict the next token while considering the previous tokens as well as the document image.

6.3 Practical Perspective

Based on the analysis, there is a clear lack of a practice-oriented perspective that is being adopted in related work. To this end, only a small number of the analyzed approaches integrate domain knowledge. Moreover, in most of these cases, domain knowledge is at most fused into the systems by the means of hand-crafted input features. There is also a lack of evaluating proposed KIE approaches and their impact on real-world scenarios. These observations have also been made in the study by [73].

One work that stands out in terms of its consideration of domain knowledge is the approach by [5]. The authors make use of such knowledge by proposing a hybrid KIE system based on both DL-models and rule-based methods including several post-processing steps to improve the extraction results. This includes, for example, the correction of automatically extracted product codes and product prices. However, this post-processing is considered as a supplement to the DL model in case of incorrect predictions and not an architecture-based adaptation. Another example is the work by [84], which integrates domain-oriented constraints regarding invoices that the total amount is the sum of the subtotal and tax total values, and that the tax total can be computed as the product of subtotal and tax percentage. The authors achieve this by adjusting the loss calculations during training, but report that no significant performance improvements were achieved. At the same time, this emphasizes the need for further research in this area (see also Section 7). While not directly implementing such aspects in the proposed methods, some authors also designed their approaches in a way to allow for integrating domain knowledge. For example, in the work of [104], candidates are first identified for fields to be extracted and subsequently scored. In this regard, one example the authors mention is to define a scoring function that integrates constraints such as the fact that an invoice date must precede its due date.

As discussed in Section 2.1, adopting a practical perspective and considering domain knowledge can provide valuable insights for an automated extraction system in various ways. However, this aspect has not been properly explored in the literature, even though it has been named as a key challenge for DU in the past [75]. Therefore, future work should consider adopting a business process perspective in order to develop approaches that exploit this unrealized potential. This could also lead to the identification of completely new potentials of KIE systems that are not yet considered in current research, as the current focus often lies only on increasing the performance with respect to established benchmark datasets. These considerations may require a deeper understanding of the associated business processes, which can be achieved by modeling computer-integrated systems including data, behavior, and control flows [28]. The potential of context-aware DL systems has been investigated in other research areas. For example, [117] propose the fusion of information extracted from business documents with traditional event log data in order to enhance predictive process monitoring capabilities. However, the other direction, i.e., feeding context-aware data obtained from process mining techniques into KIE systems, has not yet been considered.

6.4 Trends

One can observe several trends in KIE research since 2017 in a wide variety of facets. One aspect is the dominance of certain paradigms of KIE systems—most notably sequence-based methods, which represent the predominant approach since 2021. To the contrary, grid-based methods have never been able to establish themselves, despite their comparability with graph-based approaches, which have been popular since the beginning. At the same time, this indicates that VRDs with usually complex layouts cannot be appropriately represented as grids with well-defined structures that are less flexible compared to dynamic graphs. As expected, LLM-based systems have emerged in 2023 and quickly became popular. Also, using conjunctions of different paradigms (e.g., sequence-based and graph-based) is not as widely adopted as the focus on one particular paradigm.

Since 2022, an increased focus has been placed on creating OCR-independent and generative KIE approaches. The latter is the consequence of the increasing popularity of generative AI methods such as LLMs in recent years and their influence on the DU domain. The increasing shift toward OCR-independent systems can be explained by the otherwise necessary external OCR engines, which can be error-prone and therefore lead to falsely extracted information, especially in real-world use cases [16].

It can also be observed that over time, there have been different strategies on how to develop and improve KIE systems. While toward the beginning, a lot of research effort has been placed on how to properly integrate the different input modalities, other strategies have also emerged in recent years. For example, some publications highlight and investigate the importance of the reading order. In traditional methods, also as a consequence of the reliance on external OCR engines, a simple top-bottom and left-right reading order is usually adopted. This however can lead to a suboptimal segmentation of complex VRDs. To this end, approaches such as those by [87, 103, 128] investigate more sophisticated methods to obtain an optimized reading order that better suits the actual document layout. Reference [109] also propose evaluation metrics that overcome biases resulting from the reading order. However, what was not or hardly ever considered as a lever for better KIE performance is the model size in terms of trainable parameters. Between 2017 and 2024, there was no significant increase in model sizes besides a more frequent use of LLMs.

It is also positive to note that code and/or model weights are being published more frequently in recent years. Even if the absolute number of implementations that have been shared is relatively low, it has become increasingly more frequent, especially since 2021. This is a positive development, as it can accelerate research progress and research dissemination. This observation also goes hand in hand with the increased popularity of HuggingFace,[14] which is a platform that provides tools, datasets, and pre-trained models to facilitate research in NLP and CV. Some previously discussed KIE approaches are also available, for example LayoutLMv3.[15]

[14] https://huggingface.co/

[15] https://huggingface.co/microsoft/layoutlmv3-base

It can be assumed that these trends and observations will continue in future KIE research. Especially a shift toward generative KIE systems seems evident, as corresponding models become more and more popular across multiple domains. Initial approaches that make use of powerful LLMs exist, however they do not consistently achieve competitive results compared to specialized DU methods yet. In this regard, additional research is required on how to close this performance gap and thus make corresponding approaches more viable, especially with regard to the tradeoff between model size (and therefore hardware requirements) and extraction performance as discussed in Section 5.3.2.

7 Research Agenda

We have identified several aspects that should be considered in future work, which could not only lead to better research results, but also to a better applicability of KIE systems in real-world scenarios.

Novel datasets. Only a small number of public datasets are commonly used. For example, more than half of the sequence-based approaches are pre-trained using (subsets of) IIT-CDIP. One problem with this dataset in particular is the relatively low image quality, which no longer meets today's standards. These properties have a direct impact on the use of corresponding approaches in real-world settings, which typically involve documents with better image quality.

In addition to pre-training datasets, efforts should also be made to construct novel benchmark datasets, ideally based on different document types compared to commonly considered receipts – as also discussed in [63, 99]. Besides different domains, newly created datasets should also be designed to more closely resemble documents found in real-world scenarios. Refs. [116, 125, 128] have shown that existing dominant datasets have numerous shortcomings in this respect. Also, some datasets contain erroneous annotations and inconsistent labeling behavior [66]. Furthermore, common benchmarks such as SROIE and FUNSD show a high degree of layout replication in training and test partitions [56], which distorts the overall validity the reported evaluation results. One possible solution to construct novel datasets can be the generation of synthetic documents, however, there is a risk that synthetic documents do not have the properties of real-world documents in terms of layout variety and complexity [99].

Consistent evaluation. The analysis in Section 5.3 has shown that the papers are very heterogeneous in terms of their evaluation setup. In addition, the authors often do not specify their evaluation methodology in much detail, which raises questions on how exactly the evaluation results were obtained. Therefore, a quantitative comparison of the approaches is not always meaningful. This is especially problematic since the understanding of improving the state of the art in this research area is often associated with an increase in performance on benchmark datasets compared to existing work. Some public competitions, such as SROIE, provide a dedicated evaluation protocol for a consistent ranking of methods.[16] However, this is the exception rather than the rule.

Therefore, it would be desirable to agree on a consistent approach or implement a centralized evaluation toolkit that could be used for different benchmark datasets, where a specific set of metrics (e.g., Precision, Recall, F1) is implemented. Reference [11] have proposed a reference implementation for a DU benchmark, but it is not yet widely used. We also advocate a more frequent use of string-based evaluation metrics, as they allow for a better assessment in real-world applications. However, only one third of the analyzed manuscripts used such string-based evaluation setups. It is also surprising that only about a third of the papers present the results on field-level. In this regard, we advocate that such an evaluation should be presented more frequently in future work, as it gives a good indication of whether an approach faces problems with certain fields, which in turn allows

[16] The leaderboard for the KIE task can be found at https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=3

for a more in-depth analysis, as also discussed by [79]. Last, but not least, authors should adopt entity-level performance measures, as it is more suitable to assess the performance for practical applications where the extracted entities are processed as a whole [128]. The need for performance measures that provide a better indication of real-world performance is also addressed by [50], which propose a novel metric that shall mitigate the aforementioned shortcomings.

Pre-processing methods. There exists a dedicated research area that investigates techniques for pre-processing document images. An example of a sophisticated approach is proposed by [92]. In this context, synergy effects with this research area should be exploited, in particular the integration of corresponding techniques into KIE systems [7]. Related work often does not include pre-processing techniques, or, if they are used, only simple methods such as Deskewing. During the life cycle of a document in real-world settings, there are various points at which image quality can be impaired, for example due to introduced scan artifacts by digitization steps or as a consequence of multiple compression and transmission procedures [2]. Reference [123] show that quality aspects such as image noise or fonts can have significant impacts on the performance of DU systems. Therefore, more emphasis should be put on the adoption of respective methods in DU pipelines.

Tokenization. In particular sequence-based KIE approaches utilize tokenizers for the encoding of input documents into token sequences, which are subsequently supplemented by visual and positional embeddings. Whenever the KIE systems incorporate pre-trained language models such as BERT as their encoder backbone, they usually also adopt the corresponding tokenizer without further adjustments. However, there have been studies highlighting the challenges associated with the adoption of unaltered tokenizers to different domains [77]. To this end, several methods have been proposed to adjust tokenizers to new domains, which improves the performance on downstream tasks. Examples are domain-specific augmentations of the original tokenizers' vocabulary [90, 102], or a careful investigation of training data, pre-tokenization setups, and other changes to the vocabulary [17].

Currently, there is no extensive investigation of the role of tokenizers in the context of KIE. Reference [105] analyze the impact of tokenization regarding NER from biomedical texts, however there is a lack of research considering complex DU tasks on VRDs. Therefore, future work should consider exploring sophisticated methods to allow for an adequate transfer of existing tokenizers to KIE and/or develop novel methods for tokenizing complex documents in a more robust manner. Conceivable first steps could be to initially use existing KIE approaches based on out-of-the-box tokenizers, only focus on the adaptation of the tokenizer and perform benchmarks against the original implementation.

Generalizability. Reference [38] have shown that existing KIE methods lack adequate generalization capabilities, which may be due to the limitations associated with the datasets as mentioned before. Furthermore, the aspect of generalizability is often not considered in detail. Instead, it is usually only implicitly considered when the approaches are evaluated based on multiple benchmark datasets with different document types.

Future work should focus on how to effectively design model architectures as well as novel (pre-)training tasks in order to achieve a higher degree of generalizability across different document layouts, types, and domains. Another possibility is to investigate the modularization of KIE systems more closely in order to reduce interdependencies between individual components and thus obtain more generic approaches [84]. There are some efforts in other domains such as dental image analysis [55]. Research has also been conducted in the context of DLA, where the authors advocate the combination of novel models and curated (synthetic) datasets to address the challenge of generalizability [76]. However, there are not many efforts in the context of KIE.

Domain knowledge. As discussed in Section 6.3, only a few authors integrate domain knowledge into the proposed approaches or adopt a practice-oriented view in general. However, it seems promising to integrate existing domain knowledge into KIE systems, which is also discussed in [16], as it contains valuable information about the corresponding business processes, their documents and how they need to be understood as a whole. The design of novel pre-training tasks could be one possibility for integrating these aspects.

Real-world usability. The relevance of automated document processing in real-world scenarios is indisputable (see also the review by [73]). Because of this great importance, more efforts should be made to improve the overall practicability of KIE approaches. The previously mentioned research directions already address this matter. For example, one could consider pre-processing techniques to remove document image artifacts that are specific to real-world settings, such as removing stamps [124]. The aforementioned focus on generalization capabilities can also have a positive impact on the practicality of KIE systems since deploying a system that can properly process many different types of documents simultaneously could result in lower costs and less maintenance.

On top of that, additional aspects could be considered. For example, [52] propose a method to provide confidence estimates for the extracted data. The authors emphasize that in industrial settings, the primary goal is typically to make decisions based on model predictions rather than the raw extracted data. Research in this direction could greatly assist the collaboration between KIE systems and human administrators during document processing tasks, as it is essential to validate automatically extracted data in real-world settings [43]. Reference [94] also emphasize the importance of real-time KIE in industry applications. In this regard, a focus on lightweight KIE solutions should be considered. This could for example include investigating the tradeoff between model size and extraction performance more closely. Reference [39] have shown that a relatively small model with around 50 thousand parameters can produce competitive results compared to systems that consist of several hundred million parameters.

End-to-end performance. The evaluation has shown that end-to-end approaches usually achieve inferior extraction results compared to systems that use an external OCR engine. In this regard, more research should be conducted on how to improve the text recognition step, as the end-to-end systems have a high potential due to their independence of OCR engines, also highlighted by [94]. End-to-end approaches have received less attention so far, which is also reflected in the number of identified approaches. Nevertheless, there is a relatively strong increase in corresponding methods, especially since 2022.

8 Conclusion

Research in the area of KIE has seen an increased interest in recent years, mainly due to major advances in the field of DL. Nowadays, even visually-rich business documents with very complex layouts and high information-richness can be automatically processed by corresponding systems. To this end, this manuscript represents an SLR covering the research on KIE between 2017 and 2024 including 130 proposed approaches, with the aim of identifying the state of the art in this field and identifying potentials for further research. The identified methods were compared both qualitatively and quantitatively based on various characteristics.

The analysis has shown that related work tends to follow a very narrow corridor in which already proven concepts are being successively refined and improved. In general, the approaches follow the same paradigms for representing document images, namely as sequences, graphs or grids. In detail, the approaches differ in their choice of architectures used for encoding and decoding the input documents, although it can also be seen that certain models are used more frequently (e.g., BERT-based models for textual inputs). Besides, novel concepts have been explored over time, such as OCR-independent and autoregressive methods, which on one hand do not require an external

OCR engine for preliminary text reading steps, and on the other hand can output arbitrary text, making them more flexible in terms of the downstream tasks they support. The authors investigate how different input modalities implicitly and explicitly contained in complex documents can be optimally integrated into the model architectures. In particular, visual cues from document images are increasingly incorporated into the models in order to improve their performance for more complex use cases. Much effort is also put into learning a general DU through DL models, which can then be used for KIE. This is usually done by extensive and innovative pre-training tasks followed by specific fine-tuning steps. Another general observation is that the complexity of the corresponding models in terms of the number of parameters does not play a significant role and that even lightweight models can achieve promising extraction results.

The research area is strongly characterized by the fact that an improvement of the state of the art is achieved by obtaining better extraction results on established benchmark datasets. However, a quantitative comparison of the presented results is not always meaningful, since the evaluation setups used by the authors are very heterogeneous. Furthermore, for most of the benchmarks, very good results have already been achieved (F1-scores above 0.97). The research area should therefore move away from focusing on improving benchmarks results and instead investigate innovations along other dimensions. Future work could focus on even more lightweight models in order to improve their practical applicability. It could also be investigated how the models can be designed in order to require less data for training procedures. We identified several other starting points for follow-up research based on the findings. These include proposing novel and more diverse datasets as well as consistent evaluation setups that allow for an adequate quantitative comparison. We also advocate that more attention should be paid to the real-world usability of corresponding approaches. This includes the integration of domain knowledge, since on one hand document processing tasks play a key role in daily business workloads and, on the other hand, offer a special perspective on KIE with unique requirements, but also possibilities.

Acknowledgments

This research was funded by the German Federal Ministry of Education and Research under grant number 01IS23064.

References

[1] Abdelrahman Abdallah, Daniel Eberharter, Zoe Pfister, and Adam Jatowt. 2024. Transformers and language models in form understanding: A comprehensive review of scanned document analysis. *CoRR* abs/2403.04080, (2024). Retrieved from https://doi.org/10.48550/arXiv.2403.04080

[2] Alireza Alaei, Vinh Bui, David Doermann, and Umapada Pal. 2023. Document image quality assessment: A survey. *ACM Computing Surveys* 56, 2 (Feb 2023), 1–36. DOI: https://doi.org/10.1145/3606692

[3] Jason Antonio, Aditya Rachman Putra, Harits Abdurrohman, and Moch Shandy Tsalasa Putra. 2022. A Survey on Scanned Receipts OCR and Information Extraction. 0–26 pages. DOI: https://doi.org/10.13140/RG.2.2.24735.84643

[4] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 973–983. DOI: https://doi.org/10.1109/ICCV48922.2021.00103

[5] Roberto Arroyo, Javier Yebes, Elena Martínez, Héctor Corrales, and Javier Lorenzo. 2022. Key information extraction in purchase documents using deep learning and rule-based corrections. In *Proceedings of the 1st Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*. International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 11–20. arXiv:2210.03453. Retrieved from http://arxiv.org/abs/2210.03453

[6] Krisztian Balog. 2018. *Entity-Oriented Search*. The Information Retrieval Series, Vol. 39. Springer International Publishing, Cham. DOI: https://doi.org/10.1007/978-3-319-93935-3

[7] Dipali Baviskar, Swati Ahirrao, Vidyasagar Potdar, and Ketan Kotecha. 2021. Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access* 9 (2021), 72894–72936. DOI: https://doi.org/10.1109/ACCESS.2021.3072900

[8] Djedjiga Belhadj, Yolande Belaïd, and Abdel Belaïd. 2021. Consideration of the word's neighborhood in GATs for information extraction in semi-structured documents. In *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12822 LNCS, Springer International Publishing, Cham, 854–869. DOI: https://doi.org/10.1007/978-3-030-86331-9_55

[9] Nil Biescas, Carlos Boned, Josep Lladós, and Sanket Biswas. 2024. GeoContrastNet: Contrastive key-value edge learning for language-agnostic document understanding. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 294–310. DOI: https://doi.org/10.1007/978-3-031-70533-5_18

[10] Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Computing Surveys* 52, 6 (Nov 2019), 1–36. DOI: https://doi.org/10.1145/3355610

[11] Łukasz Borchmann, Michal Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał P. Turski, Karolina Szyndler, and Filip Gralinski. 2021. DUE: End-to-end document understanding benchmark. In *Proceedings of the Conference on NeurIPS Datasets and Benchmarks*. Retrieved from https://openreview.net/forum?id=rNs2FvJGDK

[12] Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. 2022. Query-driven generative network for document information extraction in the wild. In *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. ACM, New York, NY, USA, 4261–4271. DOI: https://doi.org/10.1145/3503161.3547877

[13] Haoyu Cao, Jiefeng Ma, Antai Guo, Yiqing Hu, Hao Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. GMN: Generative multi-modal network for practical document information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3768–3778. arXiv:2207.04713 DOI: https://doi.org/10.18653/v1/2022.naacl-main.276

[14] Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot text field labeling using attention and belief propagation for structure information extraction. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. ACM, New York, NY, USA, 340–348. DOI: https://doi.org/10.1145/3394171.3413511

[15] Matteo Cristani, Andrea Bertolaso, Simone Scannapieco, and Claudio Tomazzoli. 2018. Future paradigms of automated processing of business documents. *International Journal of Information Management* 40, C (2018), 67–75. DOI: https://doi.org/10.1016/j.ijinfomgt.2018.01.010

[16] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, models and applications. In *Proceedings of the Workshop on Document Images and Language at ICDAR 2021*, Vol. abs/2111.0.

[17] Gautier Dagan, Gabriel Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org.

[18] Brian Davis, Bryan Morse, Brian Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2023. End-to-End document recognition and understanding with Dessurt. In *Computer Vision - ECCV 2022 Workshops*, Leonid Karlinsky, Tomer Michaeli, and Ko Nishino (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 13804 LNCS, Springer-Verlag, Berlin, 280–296. DOI: https://doi.org/10.1007/978-3-031-25069-9_19

[19] Xinrui Deng, Zheng Huang, Kefan Ma, Kai Chen, Jie Guo, and Weidong Qiu. 2023. GenTC: Generative transformer via contrastive learning for receipt information extraction. In *Artificial Neural Networks and Machine Learning - ICANN 2023*, Lazaros Iliadis, Antonios Papaleonidas, Plamen Angelov, and Chrisina Jayne (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 14259 LNCS, Springer Nature Switzerland, Cham, 394–406. DOI: https://doi.org/10.1007/978-3-031-44223-0_32

[20] Timo I. Denk and Christian Reisswig. 2019. BERTgrid: Contextualized embedding for 2D document representation and understanding. In *Proceedings of the Workshop on Document Intelligence at NeurIPS 2019*. arXiv:1909.04948. Retrieved from http://arxiv.org/abs/1909.04948

[21] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[22] Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. 2023. DocParser: End-to-end OCR-free information extraction from visually rich documents. In *Document Analysis and Recognition - ICDAR 2023*, Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 14191 LNCS, Springer Nature Switzerland, Cham, 155–172. DOI: https://doi.org/10.1007/978-3-031-41734-4_10

[23] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in*

Neural Information Processing Systems. H. Wallach, H. Larochelle, A Beygelzimer, F. d'Alché-Buc, E. Fox, and R Garnett (Eds.), Vol. 32, Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper.pdf

[24] Thibault Douzon, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. 2022. Improving information extraction on business documents with specific pre-training tasks. In *Document Analysis Systems*, Seiichi Uchida, Elisa Barney, and Véronique Eglin (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 13237 LNCS, Springer International Publishing, Cham, 111–125. DOI: https://doi.org/10.1007/978-3-031-06555-2_8

[25] Qinyi Du, Qingqing Wang, Keqian Li, Jidong Tian, Liqiang Xiao, and Yaohui Jin. 2022. CALM: Commen-sense knowledge augmentation for document image understanding. In *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. ACM, New York, NY, USA, 3282–3290. DOI: https://doi.org/10.1145/3503161.3548321

[26] Jacob Eisenstein. 2019. *Introduction to Natural Language Processing*. The MIT Press, Cambridge, Massachusetts. 519 pages.

[27] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. DocPedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences* 67, 12 (Dec 2024), 220106. DOI: https://doi.org/10.1007/s11432-024-4250-y

[28] Peter Fettke and Wolfgang Reisig. 2021. Modelling service-oriented systems and cloud services with heraklit. In *Advances in Service-Oriented and Cloud Computing*. Christian Zirpins, Iraklis Paraskakis, Vasilios Andrikopoulos, Nane Kratzke, Claus Pahl, Nabil El Ioini, Andreas S. Andreou, George Feuerlicht, Winfried Lamersdorf, Guadalupe Ortiz, Willem-Jan den Heuvel, Jacopo Soldani, Massimo Villari, Giuliano Casale, and Pierluigi Plebani (Eds.), Springer International Publishing, Cham, 77–89. DOI: https://doi.org/10.1007/978-3-030-71906-7_7

[29] Masato Fujitake. 2024. LayoutLLM: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.), ELRA and ICCL, Torino, Italia, 10219–10224. Retrieved from https://aclanthology.org/2024.lrec-main.892/

[30] Rinon Gal, Shai Ardazi, and Roy Shilkrot. 2020. Cardinal graph convolution framework for document information extraction. In *Proceedings of the ACM Symposium on Document Engineering, DocEng 2020 (DocEng'20)*. ACM, New York, NY, USA, 1–11. DOI: https://doi.org/10.1145/3395027.3419584

[31] Rinon Gal, Nimrod Morag, and Roy Shilkrot. 2019. Visual-Linguistic methods for receipt field recognition. In *Computer Vision - ACCV 2018*, C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11362 LNCS, Springer International Publishing, Cham, 542–557. DOI: https://doi.org/10.1007/978-3-030-20890-5_35

[32] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 4 (Jul 2012), 463–484. DOI: https://doi.org/10.1109/TSMCC.2011.2161285

[33] Hamza Gbada, Karim Kalti, and Mohamed Ali Mahjoub. 2024. Information extraction from visually rich documents using directed weighted graph neural network. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 248–263. DOI: https://doi.org/10.1007/978-3-031-70552-6_15

[34] Hamza Gbada, Karim Kalti, and Mohamed Ali Mahjoub. 2024. Multimodal weighted graph representation for information extraction from visually rich documents. *Neurocomputing* 573, C (Mar 2024), 127223. DOI: https://doi.org/10.1016/j.neucom.2023.127223

[35] Vincent E. Giuliano. 1975. The office of the future. *Business Week* 2387, 30 (1975), 48–70.

[36] Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. UniDoc: Unified pretraining framework for document understanding. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin P. S. Liang, and J. Wortman Vaughan (Eds.). Curran Associates, Inc., 39–50. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2021/file/0084ae4bc24c0795d1e6a4f58444d39b-Paper.pdf

[37] Mouad Hamri, Maxime Devanne, Jonathan Weber, and Michel Hassenforder. 2023. Enhancing GNN feature modeling for document information extraction using transformers. In *Reproducible Research in Pattern Recognition*, Bertrand Kerautret, Miguel Colom, Adrien Krähenbühl, Daniel Lopresti, Pascal Monasse, and Benjamin Perret (Eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 14068 LNCS, Springer Nature Switzerland, Cham, 25–39. DOI: https://doi.org/10.1007/978-3-031-40773-4_2

[38] Jiabang He, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. Do-GOOD: Towards distribution shift evaluation for pre-trained visual document understanding models. In *Proceedings of the 46th International ACM*

SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23). ACM, New York, NY, USA, 569–579. DOI: https://doi.org/10.1145/3539618.3591670

[39] Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: In-Context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE International Conference on Computer Vision*. 19428–19437. DOI: https://doi.org/10.1109/ICCV51070.2023.01785

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[41] Shaojie He, Tianshu Wang, Yaojie Lu, Hongyu Lin, Xianpei Han, Yingfei Sun, and Le Sun. 2023. Document information extraction via global tagging. In *Chinese Computational Linguistics*, Maosong Sun, Bing Qin, Xipeng Qiu, Jiang Jing, Xianpei Han, Gaoqi Rao, and Yubo Chen (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 14232 LNAI, Springer Nature Singapore, Singapore, 145–158. DOI: https://doi.org/10.1007/978-981-99-6207-5_9

[42] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (Nov 1997), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[43] Constantin Houy, Maarten Hamberg, and Peter Fettke. 2019. Robotic process automation in public administrations. In *Digitalisierung von Staat und Verwaltung*. Michael Räckers, Sebastian Halsbenning, Detlef Rätz, David Richter, and Erich Schweighofer (Eds.), Gesellschaft für Informatik e.V., Bonn, 62–74.

[44] Yuan Hua, Zheng Huang, Jie Guo, and Weidong Qiu. 2020. Attention-based graph neural network with global context awareness for document understanding. In *Chinese Computational Linguistics Maosong Sun*, Sujian Li, Yue Zhang, Yang Liu, Shizhu He, and Gaoqi Rao (Eds.). Lecture Notes in Computer Science. 45–56. DOI: https://doi.org/10.1007/978-3-030-63031-7_4

[45] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. ACM, New York, NY, USA, 4083–4091. DOI: https://doi.org/10.1145/3503161.3548112

[46] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1516–1520. DOI: https://doi.org/10.1109/ICDAR.2019.00244

[47] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2, IEEE, 1–6. DOI: https://doi.org/10.1109/ICDARW.2019.10029

[48] Anoop R. Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4459–4469. DOI: https://doi.org/10.18653/v1/d18-1476

[49] Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2021. VisualWordGrid: Information extraction from scanned documents using a multimodal approach. In *Document Analysis and Recognition - ICDAR 2021 Workshops*, Elisa H. Barney Smith and Umapada Pal (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12917 LNCS, Springer International Publishing, Cham, 389–402. DOI: https://doi.org/10.1007/978-3-030-86159-9_28

[50] Minsoo Khang, Sang Chul Jung, Sungrae Park, and Teakgyu Hong. 2025. KIEval: Evaluation metric for document key information extraction. arXiv:2503.05488 [cs.CL]. Retrieved from https://arxiv.org/abs/2503.05488

[51] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeong Yeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *Computer Vision - ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 13688 LNCS, Springer Nature Switzerland, Cham, 498–517. DOI: https://doi.org/10.1007/978-3-031-19815-1_29

[52] Juhani Kivimäki, Aleksey Lebedev, and Jukka K. Nurminen. 2023. Failure prediction in 2D document information extraction with calibrated confidence scores. In *Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 193–202. DOI: https://doi.org/10.1109/COMPSAC57700.2023.00033

[53] Shachar Klaiman and Marius Lehne. 2021. DocReader: Bounding-box free training of a document information extraction model. In *Document Analysis and Recognition - ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12821 LNCS, Springer International Publishing, Cham, 451–465. DOI: https://doi.org/10.1007/978-3-030-86549-8_29

[54] Bertin Klein, Stevan Agne, and Andreas Dengel. 2004. Results of a study on invoice-reading systems in Germany. In *Document Analysis Systems VI*, Marinai Simone and Andreas R. Dengel (Eds.). Lecture Notes in Computer Science

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 3163, Springer Berlin Heidelberg, Berlin, 451–462. DOI: https://doi.org/10.1007/978-3-540-28640-0_43

[55] Joachim Krois, Anselmo Garcia Cantu, Akhilanand Chaurasia, Ranjitkumar Patil, Prabhat Kumar Chaudhari, Robert Gaudin, Sascha Gehrung, and Falk Schwendicke. 2021. Generalizability of deep learning models for dental image analysis. *Scientific Reports* 11, 1 (Mar 2021), 6102. DOI: https://doi.org/10.1038/s41598-021-85454-5

[56] Seif Laatiri, Pirashanth Ratnamogan, Joël Tang, Laurent Lam, William Vanhuffel, and Fabien Caspani. 2023. Information redundancy and biases in public document information extraction benchmarks. In *Document Analysis and Recognition - ICDAR 2023.* Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (Eds.), Springer Nature Switzerland, Cham, 280–294. DOI: https://doi.org/10.1007/978-3-031-41682-8_18

[57] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[58] Chen Yu Lee, Chun Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. FormNet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics, Dublin, Ireland, 3735–3754. DOI: https://doi.org/10.18653/v1/2022.acl-long.260

[59] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06).* ACM, New York, NY, USA, 665–666. DOI: https://doi.org/10.1145/1148170.1148307

[60] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural pre-training for form understanding. In *ACL-IJCNLP 2021 - Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 6309–6318. DOI: https://doi.org/10.18653/v1/2021.acl-long.493

[61] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (Jan 2022), 50–70. DOI: https://doi.org/10.1109/TKDE.2020.2981314

[62] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised Pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22).* ACM, New York, NY, USA, 3530–3539. DOI: https://doi.org/10.1145/3503161.3547911

[63] Yangchun Li, Wei Jiang, and Shouyou Song. 2023. Review of semi-structured document information extraction techniques based on deep learning. In *Proceedings of the 2023 2nd International Conference on Machine Learning, Cloud Computing, and Intelligent Mining, MLCCIM 2023.* IEEE Computer Society, Los Alamitos, CA, USA, 112–119. DOI: https://doi.org/10.1109/MLCCIM60412.2023.00022

[64] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui DIng. 2021. StrucTexT: Structured text understanding with multi-modal transformers. In *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia (MM'21).* ACM, New York, NY, USA, 1912–1920. DOI: https://doi.org/10.1145/3474085.3475345

[65] Haofu Liao, Aruni Roychowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, and Vijay Mahadevan. 2023. DocTr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE International Conference on Computer Vision.* IEEE, 19527–19537. DOI: https://doi.org/10.1109/ICCV51070.2023.01794

[66] Zening Lin, Jiapeng Wang, Teng Li, Wenhui Liao, Dayi Huang, Longfei Xiong, and Lianwen Jin. 2024. PEneo: Unifying line extraction, line grouping, and entity linking for end-to-end document pair extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24).* ACM, New York, NY, USA, 5171–5180. DOI: https://doi.org/10.1145/3664647.3680931

[67] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024. HRVDA: High-resolution visual document assistant. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 15534–15545. DOI: https://doi.org/10.1109/CVPR52733.2024.01471

[68] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL HLT 2019 - Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 32–39. DOI: https://doi.org/10.18653/v1/n19-2005

[69] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from http://arxiv.org/abs/1907.11692

[70] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Los Alamitos, CA, USA, 9992–10002. DOI: https://doi.org/10.1109/ICCV48922.2021.00986

[71] D. Lohani, Abdel Belaïd, and Yolande Belaïd. 2019. An invoice reading system using a graph convolutional network. In *Computer Vision - ACCV 2018 Workshops*, Gustavo Carneiro and Shaodi You (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11367 LNCS, Springer International Publishing, Cham, 144–158. DOI: https://doi.org/10.1007/978-3-030-21074-8_12

[72] Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James B. Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 6495–6504. DOI: https://doi.org/10.18653/v1/2020.acl-main.580

[73] A. Martínez-Rojas, J. M. López-Carnicer, J. González-Enríquez, A. Jiménez-Ramírez, and J. M. Sánchez-Oliva. 2023. *Intelligent Document Processing in End-to-End RPA Contexts: A Systematic Literature Review*. Vol. 335. Springer Nature Singapore, Singapore, 95–131. DOI: https://doi.org/10.1007/978-981-19-8296-5_5

[74] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. 2024. Embedding layout in text for document understanding using large language models. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 280–293. DOI: https://doi.org/10.1007/978-3-031-70533-5_17

[75] Hamid Motahari, Nigel Duffy, Paul Bennett, and Tania Bedrax-Weiss. 2021. A report on the first workshop on document intelligence (DI) at NeurIPS 2019. *ACM SIGKDD Explorations Newsletter* 22, 2 (Jan 2021), 8–11. DOI: https://doi.org/10.1145/3447556.3447563

[76] Jill P. Naiman. 2023. Generalizability in document layout analysis for scientific article figure & caption extraction. arXiv:2301.10781 [cs.DL]. Retrieved from https://arxiv.org/abs/2301.10781

[77] Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words. In *Proceedings of the 1st Workshop on Insights from Negative Results in NLP*. Anna Rogers, João Sedoc, and Anna Rumshisky (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 1–5. DOI: https://doi.org/10.18653/v1/2020.insights-1.1

[78] Tuan Anh D. Nguyen, Hieu M. Vu, Nguyen Hong Son, and Minh Tien Nguyen. 2021. A span extraction approach for information extraction on visually-rich documents. In *Document Analysis and Recognition – ICDAR 2021 Workshops*, Barney Smith, Elisa H. Pal, and Umapada (Eds.). Springer International Publishing, 353–363. DOI: https://doi.org/10.1007/978-3-030-86159-9_25

[79] Armineh Nourbakhsh, Sameena Shah, and Carolyn Rose. 2024. Towards a new research agenda for multimodal enterprise document understanding: What are we missing?. In *Findings of the Association for Computational Linguistics ACL 2024*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 14610–14622. DOI: https://doi.org/10.18653/v1/2024.findings-acl.870

[80] Berke Oral, Erdem Emekligil, Seçil Arslan, and Gülşen Eryiĝit. 2020. Information extraction from text intensive and visually rich banking documents. *Information Processing and Management* 57, 6 (Nov 2020), 102361. DOI: https://doi.org/10.1016/j.ipm.2020.102361

[81] Berke Oral and Gülşen Eryiğit. 2022. Fusion of visual representations for multimodal information extraction from unstructured transactional documents. *International Journal on Document Analysis and Recognition* 25, 3 (Sep 2022), 187–205. DOI: https://doi.org/10.1007/s10032-022-00399-3

[82] Ismail Oussaid, William Vanhuffel, Pirashanth Ratnamogan, Mhamed Hajaiej, Alexis Mathey, and Thomas Gilles. 2022. Information extraction from visually rich documents with font style embeddings. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 1657–1663. DOI: https://doi.org/10.1109/ICPR56361.2022.9956120

[83] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan,et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372 (Mar 2021). DOI: https://doi.org/10.1136/bmj.n71

[84] Rasmus Berg Palm, Florian Laws, and Ole Winther. 2019. Attend, copy, parse end-to-end information extraction from documents. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. IEEE, 329–336. DOI: https://doi.org/10.1109/ICDAR.2019.00060

[85] Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. CloudScan - A configuration-free invoice analysis system using recurrent neural networks. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, Vol. 1. IEEE, 406–413. DOI: https://doi.org/10.1109/ICDAR.2017.74

[86] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A consolidated receipt dataset for post-OCR parsing. In *Proceedings of the Workshop on Document Intelligence at NeurIPS 2019*. Retrieved from https://openreview.net/forum?id=SJl3z659UH

[87] Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. ERNIE-Layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3744–3756. DOI: https://doi.org/10.18653/v1/2022.findings-emnlp.274

[88] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, et al. 2024. LMDX: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics ACL 2024*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 15140–15168. DOI: https://doi.org/10.18653/v1/2024.findings-acl.899

[89] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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)*. Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 3980–3990. DOI: https://doi.org/10.18653/v1/D19-1410

[90] Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing*. Nafise Sadat Moosavi, Iryna Gurevych, Angela Fan, Thomas Wolf, Yufang Hou, Ana Marasović, and Sujith Ravi (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 155–165. DOI: https://doi.org/10.18653/v1/2021.sustainlp-1.16

[91] Saifullah Saifullah, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2023. Analyzing the potential of active learning for document image classification. *International Journal on Document Analysis and Recognition (IJDAR)* 26, 3 (Sep 2023), 187–209. DOI: https://doi.org/10.1007/s10032-023-00429-8

[92] Saifullah Saifullah, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2023. ColDBin: Cold diffusion for document image binarization. In *Document Analysis and Recognition - ICDAR 2023*. Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (Eds.), Springer Nature Switzerland, Cham, 207–226. DOI: https://doi.org/10.1007/978-3-031-41734-4_13

[93] Ritesh Sarkhel and Arnab Nandi. 2021. Improving information extraction from visually rich documents using visual span representations. *Proceedings of the VLDB Endowment* 14, 5 (Jan 2021), 822–834. DOI: https://doi.org/10.14778/3446095.3446104

[94] Abdellatif Sassioui, Rachid Benouini, Yasser El Ouargui, Mohamed El Kamili, Meriyem Chergui, and Mohammed Ouzzif. 2023. Visually-rich document understanding: Concepts, taxonomy and challenges. In *Proceedings of the 10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023*. IEEE, 1–7. DOI: https://doi.org/10.1109/WINCOM59760.2023.10322990

[95] M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681. DOI: https://doi.org/10.1109/78.650093

[96] Anna Scius-Bertrand, Atefeh Fakhari, Lars Vögtlin, Daniel Ribeiro Cabral, and Andreas Fischer. 2024. Are layout analysis and OCR still useful for document information extraction using foundation models?. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 175–191. DOI: https://doi.org/10.1007/978-3-031-70546-5_11

[97] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Katrin Erk and Noah A. Smith (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 1715–1725. DOI: https://doi.org/10.18653/v1/P16-1162

[98] Huang Siyuan, Yongping Xiong, and Wu Guibin. 2024. LayoutPointer: A spatial-context adaptive pointer network for visual information extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Kevin Duh, Helena Gomez, and Steven Bethard (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 3737–3748. DOI: https://doi.org/10.18653/v1/2024.naacl-long.207

[99] Matyáš Skalický, Štěpán Šimsa, Michal Uřičář, and Milan Šulc. 2022. Business document information extraction: Towards practical benchmarks. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Dgli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro (Eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 13390 LNCS, Springer International Publishing, Cham, 105–117. DOI: https://doi.org/10.1007/978-3-031-13643-6_8

[100] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (Jul 2009), 427–437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002

[101] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2021. A survey of deep learning approaches for OCR and document understanding. arXiv:2011.13534 [cs]. https://doi.org/10.48550/arXiv.2011.13534

[102] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Trevor Cohn, Yulan He, and Yang Liu (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 1433–1439. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.129

[103] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2023-June. IEEE, Los Alamitos, CA, USA, 19254–19264. DOI: https://doi.org/10.1109/CVPR52729.2023.01845

[104] Sandeep Tata, Navneet Potti, James B. Wendt, Lauro Beltrão Costa, Marc Najork, and Beliz Gunel. 2021. Glean: Structured extractions from templatic documents. *Proceedings of the VLDB Endowment* 14, 6 (Feb 2021), 997–1005. DOI: https://doi.org/10.14778/3447689.3447703

[105] Christos Theodoropoulos and Marie-Francine Moens. 2023. An information extraction study: Take in mind the tokenization!. In *Fuzzy Logic and Technology, and Aggregation Operators*. Sebastia Massanet, Susana Montes, Daniel Ruiz-Aguilera, and Manuel González-Hidalgo (Eds.), Springer Nature Switzerland, Cham, 593–606. DOI: https://doi.org/10.1007/978-3-031-39965-7_49

[106] Yi Tu, Ya Guo, Huan Chen, and Jinyang Tang. 2023. LayoutMask: Enhance text-layout interaction in multi-modal pre-training for document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.), Vol. 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 15200–15212. DOI: https://doi.org/10.18653/v1/2023.acl-long.847

[107] Yi Tu, Chong Zhang, Ya Guo, Huan Chen, Jinyang Tang, Huijia Zhu, and Qi Zhang. 2024. UNER: A unified prediction head for named entity recognition in visually-rich documents. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*. ACM, New York, NY, USA, 4890–4898. DOI: https://doi.org/10.1145/3664647.3681473

[108] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rJXMpikCZ

[109] David Villanova-Aparisi, Solène Tarride, Carlos-D. Martínez-Hinarejos, Verónica Romero, Christopher Kermorvant, and Moisés Pastor-Gadea. 2024. Reading order independent metrics for information extraction in handwritten documents. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 191–215. DOI: https://doi.org/10.1007/978-3-031-70536-6_12

[110] Joris Voerman, Ibrahim Souleiman Mahamoud, Aurélie Joseph, Mickael Coustaty, Vincent Poulain D'Andecy, and Jean-Marc Ogier. 2021. Toward an incremental classification process of document stream using a cascade of systems. In *Document Analysis and Recognition – ICDAR 2021 Workshops*. Elisa H. Barney Smith and Umapada Pal (Eds.), Springer International Publishing, Cham, 240–254. DOI: https://doi.org/10.1007/978-3-030-86159-9_16

[111] Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. LiLT: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics, Dublin, Ireland, 7747–7757. DOI: https://doi.org/10.18653/v1/2022.acl-long.534

[112] Jiapeng Wang, Zening Lin, Dayi Huang, Longfei Xiong, and Lianwen Jin. 2025. LiLTv2: Language-substitutable layout-image transformer for visual information extraction. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 3 (March 2025), 1–27. DOI: https://doi.org/10.1145/3708351

[113] Jiapeng Wang, Tianwei Wang, Guozhi Tang, Lianwen Jin, Weihong Ma, Kai Ding, and Yichao Huang. 2021. Tag, copy or predict: A unified weakly-supervised learning framework for visual information extraction using sequences. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence*. 1082–1090. DOI: https://doi.org/10.24963/ijcai.2021/150

[114] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* 77 (Jan 2018), 34–49. DOI: https://doi.org/10.1016/j.jbi.2017.11.011

[115] Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020. DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 898–908. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.80

[116] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. VRDU: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'23)*. ACM, New York, NY, USA, 5184–5193. DOI: https://doi.org/10.1145/3580305.3599929

[117] Sven Weinzierl, Kate Revoredo, and Martin Matzner. 2019. Predictive business process monitoring with context information from documents. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

[118] K. Y. Wong, R. G. Casey, and F. M. Wahl. 1982. Document analysis system. *IBM Journal of Research and Development* 26, 6 (Nov 1982), 647–656. DOI: https://doi.org/10.1147/rd.266.0647

[119] Chun-Bo Xu, Yi-Ming Chen, and Cheng-Lin Liu. 2024. EntityLayout: Entity-level pre-training language model for semantic entity recognition and relation extraction. In *Document Analysis and Recognition - ICDAR 2024*. Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng (Eds.), Springer Nature Switzerland, Cham, 262–279. DOI: https://doi.org/10.1007/978-3-031-70533-5_16

[120] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 1192–1200. DOI: https://doi.org/10.1145/3394486.3403172

[121] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3214–3224. DOI: https://doi.org/10.18653/v1/2022.findings-acl.253

[122] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *ACL-IJCNLP 2021 - Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2579–2591. DOI: https://doi.org/10.18653/v1/2021.acl-long.201

[123] Hsiu-Wei Yang, Abhinav Agrawal, Pavlos Fragkogiannis, and Shubham Nitin Mulay. 2024. Can AI models appreciate document aesthetics? An exploration of legibility and layout quality in relation to prediction confidence. arXiv:2403.18183. Retrieved from http://arxiv.org/abs/2403.18183

[124] Xinye Yang, Dongbao Yang, Yu Zhou, Youhui Guo, and Weiping Wang. 2023. Mask-Guided stamp erasure for real document image. In *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1631–1636. DOI: https://doi.org/10.1109/ICME55011.2023.00281

[125] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 2841–2858. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.187

[126] Arsen Yeghiazaryan, Khachatur Khechoyan, Grigor Nalbandyan, and Sipan Muradyan. 2022. Tokengrid: Toward more efficient data extraction from unstructured documents. *IEEE Access* 10 (2022), 39261–39268. DOI: https://doi.org/10.1109/ACCESS.2022.3164674

[127] Chang Oh Yoon, Wonbeen Lee, Seokhwan Jang, Kyuwon Choi, Minsung Jung, and Daewoo Choi. 2024. Language, OCR, form independent (LOFI) pipeline for industrial document information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Franck Dernoncourt, Daniel Preotiuc-Pietro, and Anastasia Shimorina (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 1056–1067. DOI: https://doi.org/10.18653/v1/2024.emnlp-industry.79

[128] Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *EMNLP 2023 - Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 13716–13730. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.846

[129] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. TRIE: End-to-end text reading and information extraction for document understanding. In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. ACM, New York, NY, USA, 1413–1422. DOI: https://doi.org/10.1145/3394171.3413900

[130] Yue Zhang, Bo Zhang, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity relation extraction as dependency parsing in visually rich documents. In *EMNLP 2021 - Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2759–2768. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.218

[131] Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2023. Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia* 25 (2023), 6743–6755. DOI: https://doi.org/10.1109/TMM.2022.3214102

[132] Ran Zmigrod, Pranav Shetty, Mathieu Sibue, Zhiqiang Ma, Armineh Nourbakhsh, Xiaomo Liu, and Manuela Veloso. 2024. "What is the value of templates?" Rethinking document information extraction datasets for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.), Association for Computational Linguistics, Stroudsburg, PA, USA, 13162–13185. arXiv:2410.15484 DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.770

Received 23 July 2024; revised 2 May 2025; accepted 11 July 2025