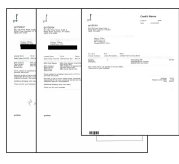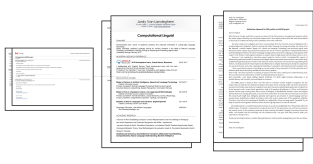# Beyond Document Page Classification: Design, Datasets, and Challenges

Jordy Van Landeghem[1,2]   Sanket Biswas[3]   Matthew Blaschko[1]   Marie-Francine Moens[1]

[1]KU Leuven   [2]Contract.fit   [3]Computer Vision Center, Universitat Autònoma de Barcelona

Covered in public research benchmarks

INPUT TASK	Page f_p	Document f_d	Document bundle f_b	Page stream f_s	Page splits f_m
LABELS	collision form	purchase invoice	email; resume; application letter	wage slip, wage_slip; bank statement; id_back, id_front; wage_slip	ticket_1, ticket_2, ..., ticket_9
USE-CASE	*Insurance claims*	*Robotic accounting*	*HR job screening*	*Loan application*	*Expenditure*

Abstract

This paper highlights the need to bring document classification benchmarking closer to real-world applications, both in the nature of data tested (X: multi-channel, multi-paged, multi-industry; Y: class distributions and label set variety) and in classification tasks considered (f: multi-page document, page stream, and document bundle classification, ...). We identify the lack of public multi-page document classification datasets, formalize different classification tasks arising in application scenarios, and motivate the value of targeting efficient multi-page document representations. An experimental study on proposed multi-page document classification datasets demonstrates that current benchmarks have become irrelevant and need to be updated to evaluate complete documents, as they naturally occur in practice. This reality check also calls for more mature evaluation methodologies, covering calibration evaluation, inference complexity (time-memory), and a range of realistic distribution shifts (e.g., born-digital vs. scanning noise, shifting page order). Our study ends on a hopeful note by recommending concrete avenues for future improvements.

1. Introduction

Visual Document Understanding (VDU) comprises a large set of skills, including the ability to holistically process both textual and visual components structured according to rich semantic layouts. The majority of efforts are directed toward the application-directed tasks of classification and extraction of key information (KIE) in visually-rich documents (VRDs). **Document classification** (DC) is a fundamental step in any industrial VDU pipeline as it assigns a semantically meaningful category, routes a document for further processing (towards KIE, fraud checking), or flags incomplete (e.g., missing scans) or irrelevant documents (e.g., recipe cookbook in a loan application).

Documents are intrinsically multi-paged, explaining (partly) why PDF is one of the most popular universal document file formats.[1] While DC in information management workflows typically involves multi-page VRDs, current public datasets [17, 24] only support single-page images and constitute too simplified benchmarks for evaluating fundamental progress in DC.

With the advent of deep learning, the VDU field has shifted from region-based analysis to whole-page image analysis. This shift led to substantial improvements in processing document images with more complex layout variability, exposing the limitations of template-based methods. Our work highlights the opportunity and necessity of

---

[1]PDF is the 2nd most popular file format on the web (after HTML and XHTML) following detected MIME types in CommonCrawl.

moving *beyond the page* limits toward evaluation on *complete* document inputs, as they prevalently occur (multi-page documents, bundles, page streams, and splits) across various practical scenarios within real-world DC applications, demonstrated in Figure 1 (see *first page*).

The practical task of long document classification [42] is largely underexplored due to challenges in computation and how to efficiently represent large multi-modal inputs. Additionally, the proximity to applications involves a larger community for conducting research, yet innovations may happen in isolation or are kept back as intellectual property, lacking evaluation on public benchmarks [13, 14], consequently hindering reproducibility and fair comparisons.

Existing DC methodology is limited to single-page images, and independently and identically distributed (*iid*) settings. We propose an improved methodology that extends its scope to multi-page images and non-iid settings. We also reflect on evaluation practices and put forward more mature evaluation protocols. To better capture the complexity of real-world document handling, we align DC benchmarking closer to practical applications and task formulations. Our key contributions can be summarized as follows:

- We have redesigned and formalized multi-page DC scenarios to align fragmented definitions and practices.
- We construct and share two novel datasets RVL-CDIP_MP[2] and RVL-CDIP-N_MP[3] to the community for evaluating multi-page DC.
- We conduct a comprehensive analysis of the novel datasets with different experimental strategies, observing the promise from best-case analysis (+6% absolute accuracy) from targeting multi-page document representations and inference.
- We overview challenges stalling DC progress, giving concrete guidelines to improve and increase dataset construction efforts.

2. Problem Formulation

We propose to use formal definitions to better align DC research with real-world document distributions and practices. This will help to standardize DC practices and make it easier to compare different methods.

Let $\mathcal{X}$ denote a space of documents, and let $\mathcal{Y}$ denote the output space as a finite set of discrete labels. Document page classification is a prototypical instance of classification [54], where the goal is to learn an estimator $f : \mathcal{X} \to \mathcal{Y}$ using N supervised input-output pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn *iid* from an unknown joint distribution $P(X, Y)$.

A **page** p is a natural classification input that consists of an image $\boldsymbol{v} \in \mathbb{R}^{C \times H \times W}$ (number of channels, height,

and width, respectively) with T word tokens $\{t_i\}_{i=1}^{T}$ organized according to a layout structure $\left\{\left(x_i^1, y_i^1, x_i^2, y_i^2\right)\right\}_{i=1}^{T}$, typically referred to as bounding boxes, either coming from Optical Character Recognition (OCR) or natively encoded.

Note that in practical business settings, VRDs are presented at inference time to a production VDU system in different forms:

   I. Single page (often scanned or photographed)

  II. Single document

 III. Multiple documents

 IV. Multiple pages (often bulk-scanned to a single PDF)

  V. Single image with multiple localized pages

Classification tasks In a unification attempt, we formalize the different classification inputs and tasks that arise in practical scenarios, as visualized in Figure 1.

Definition 1 [*Page Classification*]. (I) A page (as defined above) is categorized with a single category. When only considering the visual modality, the literature refers to it as 'document image classification' [17]. An estimator for page classification with the input dimensionality ($\mathcal{X}_p$) relative to a page (viz., number of channels, height, and width) is defined as:

$$f_p : \mathcal{X}_p \to \mathcal{Y},$$
where $\mathcal{Y} = [C]$ for C mutually exclusive categories. (1)

Definition 2 [*Document Classification*]. (II) A **document** d contains a fixed number of $L \in [1, \infty)$ pages, which do not necessarily have the same dimensions (height and width). Albeit a design choice, the input dimensionality is normalized across pages (e.g., $3 \times 224 \times 224$). Assuming a fixed input dimensionality ($\mathcal{X}_d$) relative to a document ($L \times C \times H \times W$), a document classifier is defined as:

$$f_d : \mathcal{X}_d \to \mathcal{Y},$$
where $\mathcal{Y} = [K]$ for K mutually exclusive categories. (2)

Note also the difference in label space between the two previous classification tasks, which can have some overlap for document types that are uniquely identifiable from a single page (e.g., an accident statement form).

Definition 3 [*Document Bundle Classification*]. (III) A bundle b can contain a variable number of B documents, each with a potentially different amount of L pages. A bundle classifier models a sequence classification problem over multiple documents:

$$f_b : \mathcal{X}_b \to \mathcal{Y}, \text{ where } \mathcal{Y} \text{ is a product space of } B \text{ documents,}$$
$$\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_B, \text{ with } \{\mathcal{Y}_j = [K] : j \in [B]\}.$$
(3)

Definition 4 [*Document Stream Classification*]. (IV) A page stream s is similar to a document in terms of input (number of pages L), albeit typically more varied in content and page formats. Page streams can implicitly contain many different documents, with pages not necessarily contiguous or even in the right order, as illustrated in Figure 1.

$$f_s : \mathcal{X}_d \to \mathcal{Y}, \text{ where } \mathcal{Y} \text{ is a product space of } L \text{ pages,}$$
$$\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_L, \text{ with } \{\mathcal{Y}_j = [C] : j \in [L]\}. \quad (4)$$

A very concrete example of how the label sets $[C]$ and $[K]$ can differ is in a loan application use-case where national registry proofs need to be sent: If two pages are sent with the front and back of the ID-card, f_s requires two labels (*id_front, id_back*), whereas f_d requires a single document label (*id_card*).

A critical note is due to differentiate page stream segmentation (PSS) [9, 36, 56] and page stream classification as defined above (f_s). PSS treats a page stream as a binary classification task to identify document boundaries, without classifying the identified documents afterward. f_s considers the task in one stage where C is constructed in a way to send atomic units such as a wage slip in Figure 1 for individual downstream processing or it can be combined to a single document label from $[K]$ based on assigned page labels. Two-stage processing is possible by applying PSS as an instance of a f_s classifier with $[C] = \{0, 1\}$ where 1 indicates a document boundary, followed by f_d.

Definition 5 [*Page Splitting*]. (V) A multi-page image m contains multiple page objects of similar types which can have multiple orientations, page dimensions, and often physical overlap from poor scanning [10]. A standard example involves multiple receipts to be analyzed for reclaiming VAT. While a complete approach will consist of localizing pages (using edge/corner detection, object detection, or instance segmentation) and identifying page types, we will only focus on the latter. For instance, multi-page splitting can be defined as a preliminary check on how many page types are present in a multi-page image (with input dimensionality similar to a single page p):

$$f_m : \mathcal{X}_p \to \mathcal{Y}, \text{ where } \mathcal{Y} = \mathbb{Z}^C. \quad (5)$$

Payment proofs such as tickets and receipts more often are packed together due to their compactly printed sizes, which would require splitting the unique documents from within a page to send individually for further processing. Following the national registry example. another rare yet "economical" variation for f_d occurs when a single page contains both the front and back of the ID card stitched together. These edge cases (rightmost example in Figure 1) should be dealt with on a case-by-case basis for how to set up $[K]$ (e.g., specific label: multi-tickets).

The formalisms defined above establishes a taxonomy of DC tasks, which will be retaken in the discussion of challenges to align DC research and applications (Section 5).

3. Balancing Research & Applications

Having established a taxonomy, we further sketch the role of DC in the larger scope of VDU, both in the applications and research context. We point to related VDU benchmarks and describe current DC datasets with their relevant (or missing) properties using the task formalizations. Next, we link to related initiatives in dataset construction and calls for reflection on DU practices. Finally, we introduce the curated DC datasets to support multi-page DC (f_d) benchmarking, which will be used in a further experimental study.

General Benchmarking in VDU: In any *industrial application context* where information transfer and inbound communication services are an important part of the day-to-day processes, a vast number of documents have to be processed. To provide customers with the expected service levels (in terms of speed, convenience, and correctness) a lot of time and resources are spent on categorizing these documents and extracting crucial information. Complex business use cases (such as consumer lending, insurance claims, real estate purchases, and expenditure) involve processing bundles of different documents that clients send via any communication channel. For example, obtaining a loan typically entails sending the following documents to prove solvency: a number of monthly pay stubs, bank statements, tax forms, and national registry proofs. Furthermore, not all documents are born-digital (BD), and as an artifact of the communication channel (bulk scans/photographs, digitization of physical mail), a single client communication can contain an arbitrary amount of document page images in an unknown order, requiring an f_s classifier. Figure 1 provides an overview of the different DC tasks that arise in application scenarios, which are scarcely covered by DC research benchmarks (see Table 2). As RVL-CDIP is the only large-scale non-synthetic DC benchmark, we discuss it in more detail, other dataset descriptions can be found in Supplementary.

Current state-of-the-art DU research based approaches [1, 18, 29] leverage the "pre-train and fine-tune" procedure that performs significantly well on popular DU benchmarks [17, 19, 21, 60] (see Table 1). However, their performance drops significantly when exposed to real-world business use cases mainly due to the following reasons: (1) The models are limited to modeling page-level context due to heavy compute requirements (e.g., quadratic complexity of self-attention [55]), effectively treating each document page as conditionally independent and potentially missing out on essential classification cues. (2) The methods are heavily reliant on the quality of OCR engines

Dataset	Size	Data Source	Domain	Task	OCR	Layout
IIT-CDIP [28]	35.5M	UCSF-IDL	Industry	Pretrain	✗	✗
RVL-CDIP [17]	400K	UCSF-IDL	Industry	DC	✗	✗
RVL-CDIP-N [25]	1K	Document Cloud	Industry	DC	✗	✗
TAB [36]	44.8K	UCSF-IDL	Industry	DC	✗	✗
FUNSD [21]	199	UCSF-IDL	Industry	KIE	✓	✗
SP-DocVQA [35]	12K	UCSF-IDL	Industry	QA	✓	✗
OCR-IDL [6]	26M	UCSF-IDL	Industry	Pretrain	✓	✗
FinTabNet [58]	89.7K	Annual Reports S&P	Finance	TSR	✗	✓
Kleister-NDA [47]	3.2K	EDGAR	US NDAs	KIE	✓	✗
Kleister-Charity [47]	61.6K	UK Charity Commission	Legal	KIE	✓	✗
DeepForm [48]	20K	FCC Inspection	Forms broadcast	KIE	✓	✗
TAT-QA [61]	2.8K	Open WorldBank	Finance	QA	✓	✗
PubLayNet [60]	360K	PubMed Central	Scientific	DLA	✗	✓
DocBank [30]	500K	arxiv	Scientific	DLA	✓	✓
PubTabNet [59]	568K	PubMed Central	Scientific	TSR	✗	✓
DUDE [53]	40K	Mixed	Multi-domain	QA	✓	✗
Docile [45]	106K	EDGAR & synthetic	Industry	KIE	✓	✗
CC-PDF [51]	1.1M	Common-Crawl (2010-22)	Multi-domain	Pretrain	✗	✗

Table 1. **DU Benchmarks** with their significant data sources and properties. Acronyms for tasks DC: Document Classification DLA: Document Layout Analysis KIE: Key Information Extraction QA: Question Answering TSR: Table Structure Recognition

Dataset	Purpose	#d	#p	$|\mathcal{Y}|$	Language	Color depth
NIST [8]	f_s		5590	20	English	Grayscale
MARG [32]	f_s		1553	2	English	RGB
Tobacco-800 [62]	f_s		800	2	English	Grayscale
TAB [36]	f_s		44.8K	2	English	Grayscale
Tobacco-3482 [23]	f_p		3482	10	English	Grayscale
RVL-CDIP [17]	pre-training, f_p		400K	16	English	Grayscale
RVL-CDIP-N [25]	f_p, OOD		1002	16	English	RGB
RVL-CDIP-O [25]	f_p, OOD		3415	1	English/Mixed	RGB
RVL-CDIP_MP	f_d	±400K	$\mathbb{E}[L]=5$	16	English	Grayscale
RVL-CDIP-N_MP	f_d, OOD	1002	$\mathbb{E}[L]=10$	16	English	RGB

Table 2. **Statistical Comparison** of public and proposed extended multi-page DC datasets. OOD refers to out-of-distribution detection. #d and #p refer to number of documents or pages, respectively. For the novel MP datasets, we report the average number of pages.

to extract spatial local information (i.e. mostly at word level) suitable to solve downstream benchmark tasks; but fail to *generalize* well on business documents. (3) Existing datasets used for pre-training [17, 28] are different in terms of domain, content, and visual appearance from many downstream DC tasks (detailed in Section 5.3). Therefore, it can be challenging for industry practitioners to choose a specific model to fine-tune for the DC use cases and task specifics that they commonly encounter.

RVL-CDIP The Ryerson Vision Lab Complex Document Information Processing [17] dataset used the original IIT-CDIP (The Illinois Institute of Technology dataset for Complex Document Information Processing) [28] metadata to create a new dataset for document classification. It was created as the equivalent of ImageNet in the VDU field, which invited a lot of multi-community (Computer Vision, NLP) efforts to solve this dataset. It consists of low-resolution,

scanned documents belonging to one of 16 classes such as *letter, form, email, invoice*.

Proposed Datasets RVL-CDIP_MP is our first contribution to retrieve the original documents of the IIT-CDIP test collection which were used to create RVL-CDIP. Some PDFs or encoded images were corrupt, which explains that we have around 500 fewer instances. By leveraging metadata from OCR-IDL [6], we matched the original identifiers from IIT-CDIP and retrieved them from IDL using a conversion. However, the same caveats for RVL-CDIP apply. RVL-CDIP_MP-N can serve its original goal as a covariate shift test set, now for multi-page document classification. We were able to retrieve the original full documents from DocumentCloud and Web Search. As no existing large-scale datasets include granular page-level labeling (in terms of [C]) for multi-page documents, we could not create a

benchmark for evaluating f_s. Appendix B points to visualizations from the proposed datasets.

Related Initiatives General benchmarking challenges have driven the `VDU` research community to set the seed for initiatives to create its own document-oriented "ImageNet" [43] challenge, over which multiple long-term grand challenges can be defined (deepdoc2022, scaldoc2023). In another task paradigm, DocuVQA, there have been efforts in the same spirit to redirect focus to multi-page documents [50, 52]. For the task of KIE, [46] launched a similar call for practical document benchmarks closer to real-world applications. While these initiatives demonstrate a similar-looking future direction, our contribution goes beyond introducing novel datasets and seeks to guide the complete methodology of `DC` benchmarking.

4. Experimental Study

To classify a multi-page document, one might ask the question "*Why not just predict based on the first page? What would be the gain of processing all pages? What baseline inference strategies can be applied to classify a multi-page document?*". This prompted us to put these assumptions to the test in a small motivating study[4].

As current public datasets only support page classification, we have extended some existing `DC` datasets to already enable testing a slightly more realistic, yet more complex document classification scenario (f_d).

We have reconstructed the original PDF data of the `DC` datasets in Section 3. The goal of this experiment is to tease some issues and strategies when naively scaling beyond page-level `DC`. Our baseline of choice is the document foundation model DiT-Base [29], which as a visual-only f_p is competitive with more compute-intensive multimodal, OCR-based pipelines [1, 18, 49].

Inference	Strategy	Scope
sample	first	page
	second	page
	last	page
sequence	max confidence	page
	soft voting	page
	hard voting	page
grid	grid	document
document	(not tested)	document

Table 3. **Tested inference methods** to classify multi-paged documents and simulate a true document classifier f_d. Scope refers to the independence assumption taken at inference time.

Table 3 overviews some straightforward inference strategies. Consider the simplest inference strategy is to *sample* a

----

[4]Code provided at: https://huggingface.co/bdpc/src

given page with index $l \in [L]$ (or in our case $\{1, 2, L-1\}$) from $\hat{y}^l = [f_p(x)]^l$. The *sequence* strategies mainly differ in how the final prediction $\hat{y}$ is obtained from predictions per page, assuming a probabilistic classifier $\tilde{f}_p : \mathcal{X}_p \to [0, 1]^K$.

$$\text{MaxConf}(x, y) = \underset{\substack{l \in [L] \\ k \in [K]}}{\arg \max} [\tilde{f}_p(x, y)]_k^l \qquad (6)$$

$$\text{SoftConf}(x, y) = \underset{k \in [K]}{\arg \max} \sum_{l=1}^{L} [\tilde{f}_p(x, y)]^l \qquad (7)$$

$$\text{HardVote}(x, y) = \underset{k \in [K]}{\arg \max} \sum_{l=1}^{L} e_{\hat{y}^l}, \qquad (8)$$

with e a one-hot vector of size K. The *grid* strategy is intuitive as we tile all page images in an equal-sized grid that trades off the resolution to jointly consume all document pages. While results in this experiment with fairly low grid resolution (224 x 224) are poor, variations (with aspect-preserving [27] or layout density-based scaling) deserve to be further explored.

Strategy	Acc↑	F1↑	F1$_M$ ↑	ECE↓	AURC↓
f_p\$ [29]	93.345	93.351	93.335	0.075	0.010
first	91.291	91.286	91.271	0.073	0.014
second	87.295	87.305	87.277	**0.070**	0.029
last	85.091	85.060	85.028	0.072	0.038
MaxConf	**91.407**	**91.453**	**91.344**	0.124	0.006
SoftVote	91.220	91.185	91.236	0.134	**0.004**
HardVote	85.995	86.182	85.781	0.085	0.018
grid	72.642	72.045	73.266	0.109	0.042

Table 4. Base classification accuracy of DiT-base [29] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_MP` per baseline f_d strategy. Best results per metric are boldfaced. \$ refers to our reproduction of results.

Strategy	Acc↑	F1↑	F1$_M$ ↑	ECE↓	AURC↓
f_p [25]	78.643	81.947	60.564	0.105	0.076
first	**78.760**	**75.316**	**60.801**	0.144	**0.025**
second	64.939	58.741	50.773	0.132	0.071
last	64.228	58.192	48.859	0.128	0.074
MaxConf	76.321	72.855	57.470	0.180	0.042
SoftVote	73.984	69.163	56.486	0.183	0.039
HardVote	67.480	63.188	52.235	0.110	0.088
grid	47.755	40.645	38.584	**0.102**	0.170

Table 5. Base classification accuracy of DiT-base [29] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_N_MP` per baseline f_d strategy. Best results per metric are boldfaced.

Following similar calls (discussed infra. Section 5.4) in the `VDU` literature [53] to establish calibration and confi-

dence ranking as default evaluation metrics, we include Expected Calibration Error (ECE) [16,37,38] to evaluate top-1 prediction miscalibration and Area-Under-Risk-Coverage-Curve (AURC) [11, 20] to measure selective (proportion of test set%) accuracy.

Results in Tables 4 and 5 demonstrate that classifying by only the first page is a solid strategy, with performance dropping when considering only later pages. Maximum confidence and soft voting require L (pages) times more processing, yet attain similar performance as the best single-page prediction. However, this could be attributed to two factors: i) dataset creation bias since [17] constructed RVL-CDIP from a page of each original .tiff file, for which the label was kept if it belonged to one of the 16 categories, whereas RVL-CDIP-N [25] consistently chose the first-page; ii) documents are fashioned in a summary-detail or top-down content structure over pages. To confirm the validity of the latter hypothesis, more robust experiments on more fine-grained labeled DC are needed.

The results from Table 4 and Table 5 can be interpreted as an upper bound (iid) and a loose lower bound (non-iid, yet related), respectively. For the former, MaxConf is the most accurate, yet compared to SoftVote has worse AURC, potentially making SoftVote a better candidate for industry use where controlled risk is more valued. While this trend is not reproduced in RVL-CDIP_N_MP, it can be explained by the more consistent first-page labeling, adding distracting classification cues from later pages.

Dataset	Strategy	Acc↑	Δ
RVL-CDIP_MP	first+second[*]	93.795	2.504
	first+last[*]	93.675	2.384
	second+last[*]	89.709	−1.583
	first+second/last[*]	**94.454**	3.163
RVL-CDIP_N_MP	first+second[*]	83.638	4.878
	first+last[*]	83.130	4.370
	second+last[*]	71.545	−7.215
	first+second/last[*]	**84.553**	5.793

Table 6. Best-case classification accuracy indicated with [*] when combining 'knowledge' over different pages. Δ refers to the absolute difference with the first page only.

To answer what can be gained from processing a multi-page document in a single shot, Table 6 reports a best-case error analysis, where a page prediction is counted as correct if the model would have had access to the other pages. This is calculated by using a bit-wise OR operation between the one-hot vectors ($\mathbb{I}[y == \hat{y}]$) expressing correctness for each strategy model. As a proof of concept, this shows that targeting multi-page document representations and inference is a promising avenue to improve DC.

5. Challenges and Guidelines

Following the introduced task formalizations of Section 2, we claim that the distribution on which document classification is currently evaluated publicly and the real-world distributions have heavily diverged. Additionally, our experimental validation on the novel datasets demonstrated the potential of multi-page DC, empirically reinforcing our call to action on improving DC methodologies. Let $P^A(X, Y)$ and $P^R(X, Y)$ denote those two distinct distributions, *real-world applications* and *research* respectively. Further, we will characterize the specific divergences with concrete examples and suggestions for better alignment.

5.1. Divergence of tasks: f

The challenge of directly processing multi-paged documents is typically avoided by current DC models which only support single-page images [1,15,18,22,27,31,41,49]. Whenever a new DU model innovation happens, the impact for document classification is publicly only measured on the first task scenario (e.g., f_p on RVL-CDIP), whereas production DU systems more often need to deal with the other settings (II,III,IV,V) in Figure 1. Moving beyond the limited page image context will test models' ability to sieve through potentially redundant and noisy signals, as the classification can be dependent on very local cues such as a single title on the first page or the presence of signatures on the last page. Without any datasets to test this ability, we also cannot blindly assume that we can simply scale f_p classifiers to take in more context or that aggregating isolated predictions over single pages is a future-proof (performant and efficient) strategy, as our experiments have shown.

While p is a natural processing unit for humans, acquiring supervised annotations for every single page can be more expensive than attaching a single content-based label (from $[K]$) to a multi-page document. However, fine-grained labeling with f_s could allow for more targeted and constrained KIE, as knowing a certain page l has label $y^l = \text{id\_front} \in [C]$ will allow you to focus on specific entities such as *national registry number, date/place of birth*. Ultimately, these classification task formulations can also help one consider how to set up f directly and annotate document inputs, depending on the DC use-case.

5.2. Divergence of label space: Y

Current benchmarks often use simplified label sets that are difficult to reconcile with industry requirements. While RVL-CDIP is the de facto standard for measuring performance on f_p DC, recent research [26] has revealed several undesirable characteristics. It supports only 16 labels that pertain to a limited yet generic subset of business documents, which is far from the 1K classes in ImageNet on whose image it was modeled. Real-world DC use cases

typically support a richer number of classes ($K \sim$ 50-400). RVL-CDIP suffers from substantial label noise, estimated to be higher than current state-the-art f_p error rates (see [26] for a detailed analysis) which are overfit to noise. Due to the absence of original labeling guidelines, the labels in RVL-CDIP can be ambiguous, containing disparate subtypes (e.g., business cards in the *resume* category), and inconsistencies between classes (cheques present in both *budget* and *invoice*). Other errors include (near-)duplicates causing substantial overlap between train and test distributions, corrupt documents, and plain mislabeling. However, many common CV benchmarks are plagued by similar issues [4] and would benefit from relabeling campaigns [57] to maintain their relevance.

Considering the above, multi-label classification (not covered explicitly in Section 5.1) could be a solution to resolve label ambiguities, yet this requires absolute consistency in label assignments, which when lacking introduces even more label noise. The highest labeling quality could arise from consistent labeling at the page level and hierarchically aggregating page labels ($C \rightarrow K$), yet granular annotations are more expensive to obtain. Alternatively, it may be better to follow the mutually exclusive and collectively exhaustive (MECE) principle [7] to construct label sets at the document level.

Finally, an overlooked aspect of current benchmarks is that label sets $[K]$ can be constructed based on some business logic, where a very local cue can lead to a class assignment such as some checked box on page 26. Admittedly, this does conflate the tasks of document object detection, KIE, and DC within a single label set. However, the current focus on classes with plenty of evidence across a document, with more global classification cues, should be balanced with document types that rely on local cues.

Taking the above issues into account, the community should work together towards developing more effective and realistic DC datasets that better align with the needs of industry practitioners. While tackling the challenge of Y divergence was out-of-scope for the contributed datasets, the next Subsection gives systematic recommendations for obtaining better future DC benchmarks.

5.3. Divergence of input data: X

We offer suggestions for future benchmark construction efforts such that they take into account what properties are currently unaccounted for, organically improving on our first pursuit towards multi-page DC benchmarking.

We argue that current VDU benchmarks fail to account for many real-world document data complexities: multiple pages, the distinction between born-native, (mobile) scanned documents, accounting for differences in quality, orientation, and resolution. Additionally, the UCSF Industry Document Library (and in consequence all DC datasets
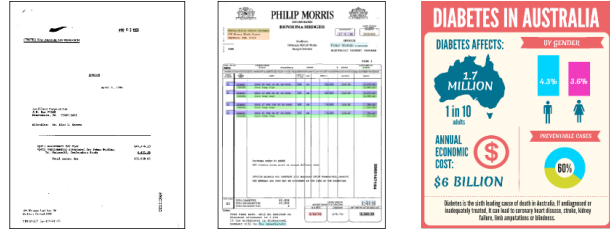

Figure 2. **Divergence of input data.** The first image is an example from DC benchmark RVL-CDIP [17], the second one from Docile [45] for KIE, while the third one comes from Info-VQA [34], illustrating the visual-layout richness of modern VRDs vs. the monotonicity of most DC document data.

drawn from this source) contains mostly old (estimated period 1950s to 2002), type-written black and white documents, while in reality, modern documents can have multiple channels, colors, and (embedded) fonts varying in size, typeface, typography. Recently, there have been efforts to collect more modern VRD benchmarks for tasks such as DocVQA [34,53], KIE [45], DLA [39]. Modern VRDs contain visual artifacts such as logos, checkboxes, barcodes, and QR codes; geometric elements such as rectangles, arrows, charts, diagrams, ..., all of which are not frequently encountered with the same variety in current benchmarks. Future DC benchmarks should incorporate modern VRDs to bring more diversity and variability in input data.

When developing DU models, it is therefore important to consider the role of vision, language, and layout and how these are connected to the classification task. For example, current datasets are based on tobacco industry documents containing very domain-specific language, which a less robust classifier can overfit (e.g., the spurious cue of a particular cigarette brand indicates an invoice). We highlight that document data can be multi-lingual, and code-switching is fairly common in document-based communications. For instance, an email may be in one language while the attachment is in another language.

In summary, future benchmarks must contain multi-page, multi-type, multi-industry (e.g., retail vs. medical invoice) , multi-lingual documents with a wide range of document data complexities to build and test generic DC systems.

The community should explore potential solutions to the lack of adequate datasets for testing DC models such as i) leveraging public document collections, ii) synthetic generation, and iii) anonymization.

Public document collections: There are increasingly more (non-profit) organizations (e.g., DocumentCloud), governments (SEC EDGAR, financial institutions (World Bank Documents & Reports), and charities (Guidestar) that make business-related documents publicly available for transparency in their operations and archival/research purposes. These collections provide datasets that are closer

to real-world scenarios. However, these documents are typically unlabelled, although annotations could be crowd-sourced through combined funding from interested parties. Since most document data sources restrict automated crawling or document scraping, future dataset constructions will require some cooperation and creativity, whilst fulfilling licensing, ethical, and legal requirements. A specific high-lighted initiative is CC-PDF [51], which collected modern, multi-lingual VRDs from CommonCrawl for future use.

Data synthesis: This alternative was suggested by prior work on KIE [3, 46] and DLA [5]for generating business and scientific documents. [45] followed up on this, delivering a large-scale KIE dataset with 6K real documents annotated and 100K synthetic examples. However, synthetic generation can be challenging to simulate real-world documents with similar data and classification complexity.

Anonymization can be a viable option to construct a DC dataset without compromising ethical guidelines and privacy regulations. This process involves removing, masking, replacing, or obfuscating data so that document content can no longer be attributed to an individual or entity. For example, one should remove names, addresses, and identifying information such as social security numbers or replace it with a textual tag ([SOCIAL-SECURITY-NUMBER]) or similar pattern (e.g., Faker). While this process is not viable for creating KIE datasets, KIE can play a big role in semi-automatically anonymizing documents [12,40]. Companies may be hesitant to make document collections public due to concerns about privacy, confidentiality and GDPR compliance. While anonymization can be an effective method, it should be approached with caution as potential risks of re-identification can make someone with originally good intentions legally liable. A potential side-step can be investing in privacy-preserving federated learning (e.g., PFL-DocVQA) to allow access to private industry document data.

5.4. Maturity of evaluation methodology

Most DC models are evaluated using predictive performance metrics such as accuracy, precision-recall, and F1-score on *iid* test sets. However, in user-facing applications, calibration can be as important as accuracy [16, 37, 38]. Even more so, when the confidence estimation of a DC is used to triage predictions to either an automated flow or manual processing by a human. Once a DC is in production, the *iid* assumption will start to break, which would recommend a priori testing of robustness against various sources of noise (OCR, subtle template changes, wording or language variations, ...) and expected distribution shifts (born-digital-scanning artifacts, shifting page order, page copies, irrelevant or out-of-scope documents, novel document classes, concept drift, ...).

Nevertheless, we observe only a few applications in DC (only reported on f_p) of more mature evaluation protocols [20] beyond predictive performance. Notable exceptions include covariate shift detection from document image augmentations [33], sub-class shift and generalization in [25, RVL-CDIP-N], out-of-distribution detection [25, RVL-CDIP-O], and cross-domain generalization [2, (RVL-CDIP ↔ Tobacco-3482)]. However, the results on the latter can be misleading as both datasets are drawn from a similar source distribution. Another gap in DC benchmarking concerns evaluating selective classification [11, 20], which is closer to the production value evaluation of how many documents can be automated without any human assistance.

Another interesting evaluation protocol concerns *out-of-the-box* performance or how data-hungry/sample-efficient a certain model is. In practice, few-shot learning from minimal annotations is a highly valued skill. This few-shot learning evaluation protocol has been applied in [44] with different data regimes. Finally, inference complexity (time-memory) has been brought back to the attention of OCR-free models [22], which we believe will be the key to measuring when scaling solutions to multi-page documents.

6. Closing Remarks

Our work represents a pivotal step forward in establishing multi-page DC by proposing a comprehensive bench-marking and evaluation methodology. Thereby, we have addressed longstanding challenges and limitations Section 5 that have hindered progress in the field. As motivated in our experimental study, we have proven the need to advance multi-page document representations and inference.

Following up on this, we provide recommendations for future DC dataset construction efforts pertaining to the type and nature of document data, variety in and quality of the classification label set, with a focus on particular DC scenarios closer to applications, and finally how future progress should be measured. Nonetheless, we are hopeful that the VDU community can come together on these shortcomings and apply the lessons from this reality check. Extending the applicability of current state-of-the-art models in VDU to multi-page documents needs further exploration, which will go hand in hand with benchmark creation initiatives or incorporating multiple DC task annotation layers on a single dataset.

Acknowledgements

References

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 3, 5, 6

[2] Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. Vlcdoc: Vision-language contrastive pre-training model for cross-modal document classification. *Pattern Recognition*, 139:109419, 2023. 8

[3] Oliver Bensch, Mirela Popa, and Constantin Spille. Key information extraction from documents: Evaluation and generator. In *European Semantic Web Conference (ESWC 2021) and 2nd International Workshop, in conjunction with ESWC 2021: Workshop: Deep Learning meets Ontologies and Natural Language Processing*, 2021. 8

[4] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 7

[5] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. Docsynth: a layout guided approach for controllable document image synthesis. In *International Conference on Document Analysis and Recognition*, pages 555–568. Springer, 2021. 8

[6] Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-idl: Ocr annotations for industry document library dataset. *arXiv preprint arXiv:2202.12985*, 2022. 4

[7] Arnaud Chevallier. *Strategic thinking in complex problem solving*. Oxford University Press, 2016. 7

[8] DL Dimmick, MD Garris, and CL Wilson. Nist special database 6. structured forms database 2. Technical report, Technical report, National Institute od Standards and Technology. Advanced . . . , 1992. 4, 12

[9] Ignazio Gallo, Lucia Noce, Alessandro Zamberletti, and Alessandro Calefati. Deep neural networks for page stream segmentation and classification. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2016. 3

[10] Siddharth Garimella. Identification of receipts in a multi-receipt image using spectral clustering. *International Journal of Computer Applications*, 155(2), 2016. 3

[11] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 6, 8

[12] Ingo Glaser, Tom Schamberger, and Florian Matthes. Anonymization of german legal court rulings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 205–209, 2021. 8

[13] Albert Gordo and Florent Perronnin. A bag-of-pages approach to unordered multi-page document classification. In *2010 20th International Conference on Pattern Recognition*, pages 1920–1923. IEEE, 2010. 2

[14] Albert Gordo, Marçal Rusinol, Dimosthenis Karatzas, and Andrew D Bagdanov. Document classification and page stream segmentation for digital mailroom applications. In *2013 12th International Conference on Document Analysis and Recognition*, pages 621–625. IEEE, 2013. 2

[15] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021. 6

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330, 2017. 6, 8

[17] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 1, 2, 3, 4, 6, 7

[18] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 3, 5, 6

[19] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 3

[20] Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *International Conference on Learning Representations*, 2023. 6, 8

[21] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019. 3, 4

[22] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 2021. 6, 8

[23] Jayant Kumar and David Doermann. Unsupervised classification of structurally similar document images. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1225–1229. IEEE, 2013. 4, 12

[24] Jayant Kumar, Peng Ye, and David Doermann. Structural similarity for document image classification and retrieval. *Pattern Recognition Letters*, 43:119–126, 2014. 1

[25] Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. Evaluating out-of-distribution performance on document image classifiers. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4, 5, 6, 8

[26] Stefan Larson, Gordon Lim, and Kevin Leach. On evaluation of document classification with rvl-cdip. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023. 6, 7

[27] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova.

Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 5, 6

[28] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 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*, pages 665–666, 2006. 4

[29] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022. 3, 5

[30] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis, 2020. 4

[31] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 6

[32] L Rodney Long, Sameer K Antani, and George R Thoma. Image informatics at a national research center. *Computerized Medical Imaging and Graphics*, 29(2-3):171–193, 2005. 4, 12

[33] Samay Maini, Alexander Groleau, Kok Wei Chee, Stefan Larson, and Jonathan Boarman. Augraphy: A data augmentation library for document images. *arXiv preprint arXiv:2208.14558*, 2022. 8

[34] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 7

[35] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020. 4

[36] Thisanaporn Mungmeeprued, Yuxin Ma, Nisarg Mehta, and Aldo Lipani. Tab this folder of documents: page stream segmentation of business documents. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, pages 1–10, 2022. 3, 4, 12

[37] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 6, 8

[38] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 625–632, 2005. 6, 8

[39] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751, 2022. 7

[40] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, 2022. 8

[41] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *ICDAR*, 2021. 6

[42] Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. Towards a multi-modal, multi-task learning based pretraining framework for document representation learning. *arXiv preprint arXiv:2009.14457*, 2020. 2

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[44] Clément Sage, Thibault Douzon, Alex Aussem, Véronique Eglin, Haytham Elghazel, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. Data-efficient information extraction from documents with pre-trained language models. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 455–469. Springer, 2021. 8

[45] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. *arXiv preprint arXiv:2302.05658*, 2023. 4, 7, 8

[46] Skalicky, Matyas and Simsa, Stepan and Uricar, Michal and Sulc, Milan. Business document information extraction: Towards practical benchmarks, 2022. 5, 8

[47] Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In *ICDAR*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer, 2021. 4

[48] J Stray and S Svetlichnaya. Deepform: extract information from documents (2020). 4

[49] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023. 5, 6

[50] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022. 5

[51] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. *arXiv preprint arXiv:2304.14953*, 2023. 4, 8

[52] Jordy Van Landeghem, Lukasz Borchmann, Rubèn Tito, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Sta-

nisławek. ICDAR 2023 Competition on Document Under-stanDing of Everything (DUDE). In *Proceedings of ICDAR 2023*, 2023. 5

[53] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew B. Blaschko, Marie-Francine Moens, and Tomasz Stanislawek. Document Understanding Dataset and Evaluation (DUDE). In *International Conference on Computer Vision*, 2023. 4, 5, 7

[54] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992. 2

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[56] Gregor Wiedemann and Gerhard Heyer. Multi-modal page stream segmentation with convolutional neural networks. *Language Resources and Evaluation*, 55:127–150, 2021. 3, 12

[57] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021. 7

[58] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021. 4

[59] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 564–580. Springer, 2020. 4

[60] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 3, 4

[61] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, oct 2022. 4

[62] Guangyu Zhu and David Doermann. Automatic document logo detection. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 864–868. IEEE, 2007. 4, 12