FormLens: From Ink to Insight with Adapting Vision-Language Models for Handwritten Form Digitization

Shaon Bhattacharya, Ajoy Mondal, and C V Jawahar
CVIT, International Institute of Information Technology
Hyderabad, Telangana, India

Abstract

Handwritten forms are still widely used across rural banking, healthcare, and public administration domains. However, their large-scale digitization remains difficult due to handwriting variability and the limitations of conventional OCR-based systems. We introduce **FormLens**, an adaptation of a vision-language model for end-to-end form understanding that directly transforms handwritten forms into structured key-value pairs without requiring explicit region detection or OCR. Built via Low-Rank Adaptation (LoRA) on a multilingual backbone, **FormLens** is designed to handle diverse layouts, noisy backgrounds, and challenging real-world capture conditions. To support evaluation and domain adaptation, we release **Form6000**, a new dataset of 6,000 handwritten English forms collected from 650 individuals under varied conditions. Extensive experiments demonstrate that **FormLens** achieves superior accuracy and robustness compared to both commercial (e.g., *Google Form Parser*, *Azure Form Recognizer*) and open-source baselines, particularly in unconstrained handwritten scenarios. The code and resources are available at: https://formlens.github.io/.

CCS Concepts

• **Applied computing** → **Document analysis**.

Keywords

Handwritten form digitization, OCR, Document Image Analysis, Form Digitization

ACM Reference Format:
Shaon Bhattacharya, Ajoy Mondal, and C V Jawahar. 2025. FormLens: From Ink to Insight with Adapting Vision-Language Models for Handwritten Form Digitization. In *Proceedings of 16th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

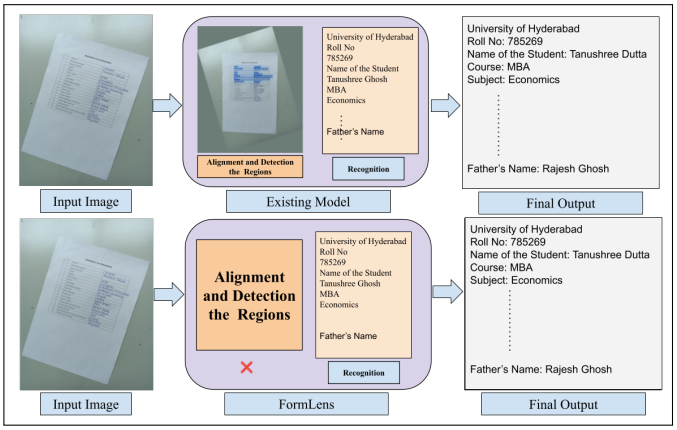


Figure 1: Comparison of form processing approaches. Traditional OCR-based pipelines rely on sequential steps — region detection, text recognition, and post-processing — which are prone to error accumulation. In contrast, FormLens streamlines the process by directly generating structured outputs from full form images in a single end-to-end step, eliminating the need for detection, alignment, or rule-based parsing.

1 Introduction

Handwritten forms remain a cornerstone of information collection in many domains, including rural banks, healthcare centers, government offices, and educational institutions. Picture a crowded rural bank: people stand in line, manually filling out printed forms for basic services. This scene is far from unique. Despite the widespread push for digitization, hand-filled printed forms are still the norm in many settings, particularly those constrained by resources or limited access to digital infrastructure.

Traditionally, these completed forms are physically stored, often in vast record rooms. This method of archiving poses several challenges: retrieving records is time-consuming, errors are common, and physical damage or loss is a constant risk. With the growing importance of data-driven decision-making, governments and organizations are increasingly adopting digitization to improve information access, efficiency, and analytical capabilities [39]. Although electronic forms (e-forms) offer a modern solution, their adoption is uneven. A significant portion of operational workflows, especially in semi-urban and rural regions, still involves scanning manually filled forms and digitizing them by hand. This manual digitization process is slow, labor-intensive, and error-prone, creating bottlenecks in service delivery and limiting the broader benefits of digitization.

To address this gap, we propose an automated form digitization system tailored to handle the specific challenges of handwritten and scanned printed forms. Our system is designed to minimize manual intervention, reduce errors, and accelerate data extraction across a wide range of form layouts.

Despite advances in document understanding through deep learning, handwritten form digitization presents a unique and persistent challenge. Unlike general documents, forms are highly

ICVGIP'25, Mandi, India

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

structured, typically comprising key-value pairs where keys (e.g., "Name", "Address") are pre-printed and users fill in the corresponding values by hand. It introduces a blend of printed and handwritten text, often co-located or overlapping, especially when scanned or captured using mobile devices. Moreover, handwritten inputs often include proper nouns, rare terms, and idiosyncratic spellings that standard language models struggle to recognize owing to limited contextual priors. The diversity in layouts, ranging from tabular structures to free-form compositions, further complicates segmentation, recognition, and extraction tasks [22, 46].

Existing OCR-based approaches [34, 43] typically operate in multiple stages: region detection, text recognition, and post-processing [3, 7] (as shown in Fig. 1). While modular, these systems are brittle, errors in one stage often cascade into others, especially under noisy or distorted real-world inputs. Layout-aware models [43] partially address this by combining spatial and textual cues. However, they still rely on pre-extracted OCR tokens and bounding boxes, making them vulnerable to OCR errors and unsuitable for end-to-end training. These limitations reduce their effectiveness for complex, low-quality, or handwritten documents. While Commercial tools [14, 32] offer out-of-the-box form extraction, they struggle with poor image quality and diverse handwriting in low-resource settings. Their high cost and infrastructure demands further limit practical deployment in real-world scenarios. Recent OCR-free approach [23] translates document images directly into structured outputs. Although promising, these methods have yet to prove robust against the high variability and noise in handwritten, real-world captured forms.

To overcome the limitations of existing approaches, we present **FormLens**, an adaptation of the existing vision-language model, GOT 2.0 [41] for end-to-end form digitization. We introduce a two-stage adaptation strategy: (i) *Stage-1*: adaptation to generic handwritten and printed documents and (ii) *Stage-2*: form-specific adaptation using LoRA [17], while keeping the vision encoder fixed. **FormLens** simplifies form digitization by directly converting entire input images into structured key-value pairs, bypassing traditional region detection, OCR, and rule-based processing. It is built to handle mixed printed and handwritten content, varied layouts, and challenging real-world conditions like poor lighting or tilted captures, making it highly effective for practical form digitization applications. To enable realistic evaluation and domain-specific adaptation, we introduce a benchmark dataset, named **Form6000**, comprising over 6000 handwritten English forms filled by 650 individuals. Captured under diverse lighting, backgrounds, and orientations, this dataset mirrors real-world submission scenarios and can serve as a valuable resource for developing and testing robust handwritten form understanding models. Extensive experiments show that **FormLens** consistently outperforms both commercial and open-source form processing systems, achieving superior accuracy and robustness across diverse handwritten and scanned form inputs.

Our key contributions are as follows:

- **FormLens:** We present **FormLens**, an adaptation of the vision-language model GOT 2.0 [41], tailored for end-to-end form digitization. **FormLens** directly transforms full input images into structured key-value pairs, eliminating the need for region detection, OCR, or rule-based processing.
- **Real-World Handwritten Form Dataset:** We release a new benchmark dataset, **Form6000**, comprising 6,000 handwritten forms filled by 650 individuals. Collected under real-world conditions — with diverse layouts, backgrounds, and orientations, this dataset (see Fig. 2) offers a challenging and practical benchmark for advancing handwritten form digitization.
- **Two-Stage Decoder Adaptation Strategy:** We develop a two-stage training pipeline for adapting the decoder on generic and domain-specific handwritten data. This improves robustness to layout variations, handwriting styles, and image degradations typically found in real-world settings.
- **Strong Empirical Performance:** We conduct extensive experiments demonstrating that **FormLens** consistently outperforms both commercial (e.g., *Google Form Parser* and *Azure Form Recognizer*) and open-source form processing systems in accuracy, robustness, and generalization — particularly in unconstrained handwritten scenarios. (see Table 2).

2 Related Work

Document imaging has long been a research focus, with extensive work on core tasks such as text detection. Classical methods like EAST [49], PixelLink [11], CRAFT [4], and TextSnake [31] have proven effective across various document types, including scanned and photographed materials. These models handle complex layouts involving multi-oriented, curved, or dense text and form the backbone of many document analysis pipelines. Printed text recognition has also seen major advances, evolving from feature-engineered approaches [8, 20] to deep learning-based systems such as Tesseract [37], EasyOCR [19], and MMOCR [25]. These modern systems perform well on printed text across different fonts and layouts but often fail under challenging conditions like noisy scans, low-resolution images, or handwritten inputs. Recognizing handwritten text presents an even greater challenge due to high variability in stroke patterns, spacing, and individual writing styles. To advance this area, large-scale datasets such as IAM [15], GNHK [26], and Bentham [35] have been developed as benchmarks. Recent models [9, 12, 27] better combine CNNs, RNNs, and Transformer architectures to capture long-range dependencies and spatial variability in handwriting.

Beyond text recognition in printed and handwritten documents, understanding structured content like tables and forms is crucial for comprehensive document analysis. Table structure recognition has advanced with models like TSRFormer [28], Omniparser [13], and LORE++ [30], aided by various benchmark datasets such as PubTables-1M [38], WTW [29], and FinTabNet [48]. Information extraction (IE) from semi-structured documents has shifted from heuristic-based methods to layout-aware models, leveraging datasets like FUNSD [21] and XFUND [44]. Existing OCR-based approaches (such as *DocTR* [34] and *LayoutLM* [43]) typically operate in multiple stages: region detection, text recognition, and post-processing [3, 7]. While modular, these pipelines are brittle — errors in one stage cascade into subsequent stages, especially under noisy or distorted inputs common in real-world conditions. Multi-modal architectures

like Layout Parser [36], LayoutLM [18, 42, 45] and DocFormer [2] integrate visual, positional, and textual signals through pre-training, enabling tasks such as form field extraction and document classification. However, these models often depend on pre-extracted OCR tokens and bounding boxes, making them sensitive to OCR quality and less effective for noisy handwritten inputs. Recent efforts such as Donut [23] and UDOP [40] move toward end-to-end document understanding by framing it as a vision-language generation task. These methods bypass traditional OCR and parsing but are primarily trained on synthetic or clean printed forms. As a result, they often under-perform on real-world handwritten forms that feature layout irregularities, skewed capture, and diverse writing styles.

Most commercial and open-source systems still follow a multi-stage pipeline: detecting regions or fields, recognizing text within those regions, and applying rule-based parsing. Systems like *Google Document AI* [14], *Azure Form Recognizer* [32], and *Amazon Textract* [1] rely on this approach. While effective for clean, machine-printed forms, they often break down on handwritten forms, poorly scanned documents, or forms captured in uncontrolled environments — where bounding box errors propagate through the pipeline.

In contrast, we propose **FormLens** — a vision-language model specifically adapted for handwritten form digitization. It eliminates intermediate detection and recognition steps by directly predicting structured outputs such as key-value pairs, headers, and footers from full-page images. To our knowledge, it is one of the first unified approaches to robustly handle scanned and camera-captured handwritten forms under real-world conditions.

3 Form6000 Dataset

We introduce a benchmark dataset, **Form6000**, designed for form-specific model adaptation and rigorously evaluate our model's ability to digitize real-world handwritten filled forms. The dataset includes 50 distinct form templates, inspired by real administrative and institutional documents such as school admission forms, railway reservation slips, banking KYC forms, hospital intake sheets, and municipality records. These templates were designed with domain experts to reflect the structural diversity and layout conventions in official documentation.

Characteristic	Count
Number of unique form templates	50
Number of participants (writers)	650
Forms filled per participant	1-2
Total handwritten filled forms	650
Scanned high-resolution forms	650
Captured mobile images (7–10 per form)	5,350
Total dataset size (images)	6,000

Table 1: Summary statistics of our Form6000 dataset, detailing the distribution of forms, fields, and annotation characteristics across varying layouts and handwriting styles.

Each form template was distributed to participants with clear instructions for manual filling. A total of 650 individuals participated in the data collection process. Every participant received a blank version of the assigned form and a filled-out reference form containing synthetically generated yet meaningful content. This strategy ensured that participants did not have to use sensitive or personal information while producing plausible and coherent entries. Participants filled out the forms using their handwriting without any restriction on ink color, writing instrument, or completion speed. It leads to a rich collection of handwriting samples varying in style, alignment, and legibility. The participant demographic covered individuals aged between 16 and 50, with approximately 60% male and 40% female contributors, helping ensure diverse writing characteristics.

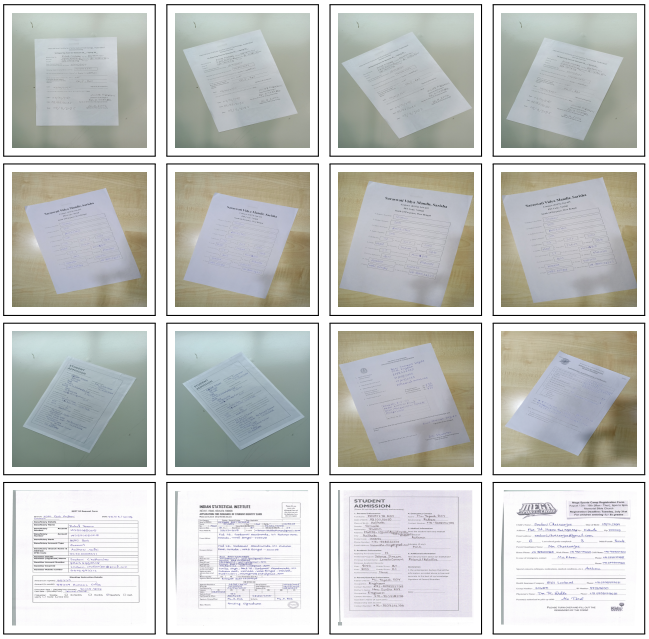


Figure 2: Real handwritten form samples from the Form6000 dataset, used for second-stage, form-specific adaptation and evaluation. These examples reflect user submissions captured under challenging mobile conditions — including skewed angles, uneven lighting, and cluttered backgrounds — enabling the model to adapt to real-world form digitization tasks.

Once hard copy of filled forms collected, these forms underwent in two stages. First, each form was scanned using a flatbed scanner to create clean, high-resolution digital versions. Next, to simulate real-world usage conditions, each form was photographed 7 to 10 times with mobile phone cameras. These photos were taken under various lighting conditions, backgrounds, and angles — some in indoor lighting with shadows, some at slight angles, and others featuring distracting textures in the background. This process resulted in over 5,200 captured images that closely resemble the noisy and imperfect environments in which digitization systems typically operate. Fig. 2 illustrates the diversity of the capture settings.

Each image is paired with a ground truth annotation containing the complete transcription of the handwritten fields. These annotations preserve the layout and structure of the form and are suitable for evaluating multiple sub tasks such as text recognition, key-value extraction, and structure-aware parsing. The dataset thus serves as a comprehensive benchmark for handwritten form digitization

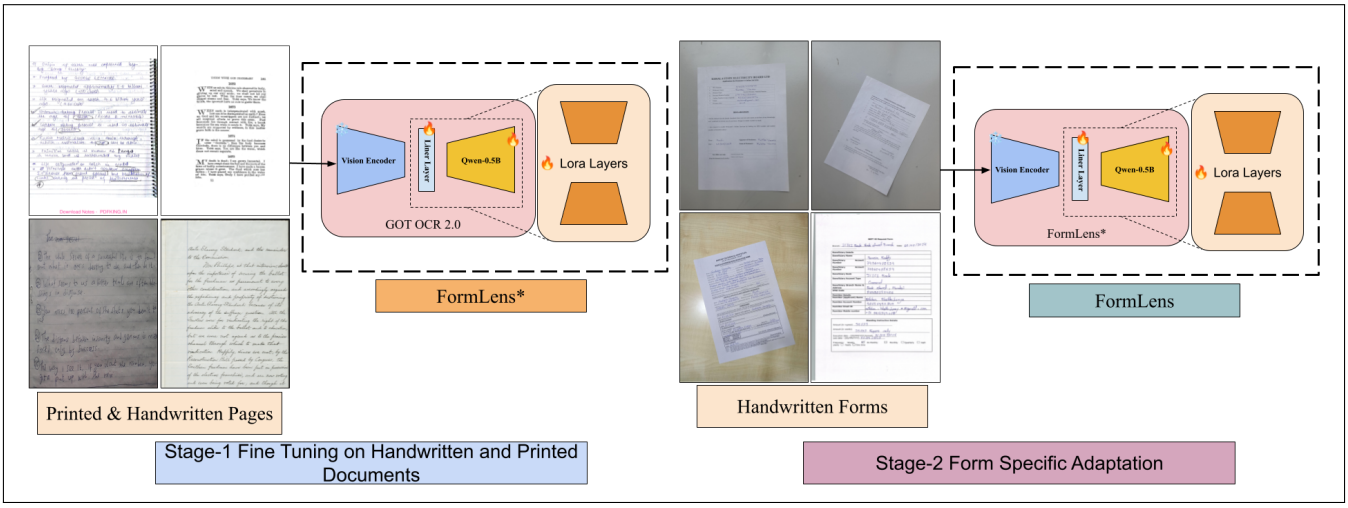


Figure 3: Overview of the two-stage decoder adaptation pipeline in FormLens. *Stage-1* involves adapting the decoder on a diverse mixtures of printed and handwritten document pages with augmentations to build generalization over noisy or degraded inputs. *Stage-2* fine-tunes the decoder using a small set of real-world handwritten forms captured from the target domain, enabling robustness to practical challenges such as tilt, lighting variation, and background clutter, etc.

research. It can help drive progress in practical OCR and document understanding systems. Table 1 presents a detailed breakdown of the dataset. We used 800 forms for form-specific model adaptation and remaining 5200 forms for evaluation purposes.

4 Methodology and Experiments

4.1 FormLens Overview

We built **FormLens** on the pre-trained GOT OCR 2.0 model [41], which was trained on a large-scale corpus of Chinese and English documents [5, 47]. GOT OCR comprises three main components: an *image encoder*, a *linear projection layer*, and an *output decoder*. The training pipeline involves three stages: (i) vision encoder pre-training on diverse OCR tasks, (ii) connecting the encoder to Qwen-0.5B [5] and training on complex multimodal datasets (e.g., mathematics, sheet music, diagrams), and (iii) robustification via multi-crop and multi-page augmentation, while keeping the encoder frozen.

Although the GOT is trained on printed documents, its encoder generalizes handwritten inputs strongly. To adapt the English handwritten form digitization model, we introduce a two-stage decoder fine-tuning strategy using LoRA [17], while keeping the vision encoder fixed throughout the process.

4.2 Decoder Adaptation via LoRA

To efficiently fine-tune the decoder, we employed Low-Rank Adaptation (LoRA), which injects trainable low-rank matrices into the frozen pre-trained weights. Given a weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA modifies it as:

$$W + \Delta W = W + BA, \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, and r is the rank of the matrix. The forward pass becomes:

$$h = Wx + \frac{\alpha}{r}(BA)x. \tag{2}$$

We use $r = 8$, $\alpha = 32$, and apply LoRA to all decoder linear layers, except the final language modeling head. This adaptation enables specialization in handwritten English while leveraging the general-purpose OCR capacity of the frozen encoder.

4.3 Stage-1: Adaptation to General Handwritten and Printed Documents

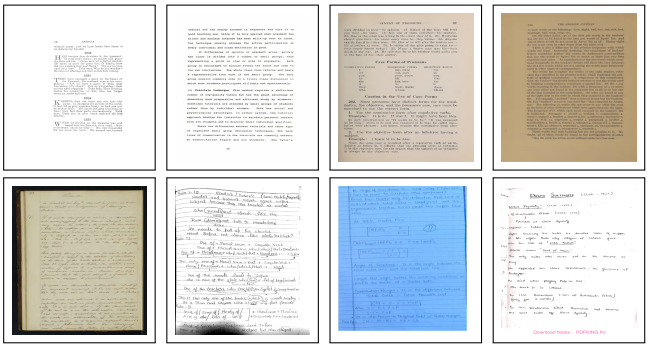


Figure 4: Sample pages of curated dataset from the internet and other sources used for first-stage adaptation, illustrating a diverse collection of printed and handwritten materials, including handwritten notes, ruled manuscripts, and unstructured free-form documents. The dataset covers various handwriting styles, font types, and document layouts.

To adapt our base OCR model for improved handwritten and printed content recognition, we curated various images from the

internet and other sources, comprising handwritten and printed documents. The handwritten subset includes a broad mix of real-world documents, such as classroom notes, informal scribbles, personal annotations, and scanned answer scripts. These samples reflect various handwriting styles collected from individuals spanning different age groups and professions. The printed subset consists of pages from English books, academic materials, typed documents and administrative documents.

These documents were scanned using flatbed scanners or photographed using mobile devices to simulate practical capture scenarios. Scanned and photographed images often include mild artifacts such as skew, blur, low contrast, or uneven lighting. This variation helps make the OCR backbone robust to quality fluctuations commonly seen in handwritten paperwork. The primary goal of this fine-tuning step is to allow the model to learn from neat and degraded textual representations and prepare it for the downstream digitization task. Fig. 4 illustrates a few samples from this collection, showcasing the variation in handwriting, fonts, and visual quality.

In the initial adaptation stage (see Fig. 3), we fine-tuned the decoder using a carefully curated dataset that included printed and handwritten documents. This dataset features samples with various augmentations such as Gaussian noise, motion blur, rotation, and compression artifacts. It aims to equip the model with generalization capabilities across various handwriting styles and degraded input conditions commonly encountered in practical OCR pipelines. Because the base GOT model was initially trained on clean, printed text, this adaptation introduced robustness to distortions and diverse character morphologies associated with cursive, slanted, or loosely written scripts. Notably, the vision encoder remains frozen throughout, allowing the decoder to specialize in transcribing non-ideal inputs.

After sufficient training iterations, the model was evaluated on our English handwritten form dataset (refer to Fig. 2), where it demonstrated high recognition accuracy on cleanly scanned documents and moderate success on some mobile-captured forms. However, in extreme blur, shadowing, or unconventional lighting (often found in real-world mobile captures), the model underperformed, highlighting a residual domain gap that needs to be addressed.

4.4 Stage-2: Form-Specific Adaptation

To address the domain gap identified in Stage-1, we conducted a second fine-tuning phase using a small but carefully chosen subset (800 images) of real-world images from our **Form6000** dataset (see Fig. 2). Although the size of selected subset is limited, these samples capture genuine noise factors, such as skew, lighting variations, blurriness, and perspective distortion, which are challenging to replicate synthetically.

Using the same LoRA-based decoder adaptation setup from Stage-1, we applied minimal training updates, leveraging the prior knowledge learned from general handwriting adaptation. Targeted fine-tuning significantly boosts the performance of complex mobile-captured forms without overfitting, showcasing the strength of LoRA in low-resource transfer learning scenarios.

This two-stage decoder adaptation strategy begins with general handwriting and advances to domain-specific forms captured in the target environment. This approach allows **FormLens** to effectively and reliably handle both controlled and wild-document digitization scenarios. The entire adaptation process is illustrated in Fig. 3.

4.5 Training Process

The overall training pipeline includes two stages of fine-tuning: (i) *Stage-1*: adaptation to a diverse set of generic handwritten and printed documents, and (ii) *Stage-2*: form-specific adaptation using a small subset of our **Form6000** dataset. Both stages utilize Low-Rank Adaptation (LoRA) [17] for parameter-efficient fine-tuning of the decoder while keeping the vision encoder frozen in order to preserve robust visual features.

Stage-1: Adaptation to Generic Handwritten and Printed Documents. The first stage involved fine-tuning the decoder on an extensively curated dataset comprising 155,650 printed and 135,830 handwritten English document images. Fig. 4 shows the sample pages from this dataset. It includes handwritten classroom notes, informal scribbles, personal annotations, scanned answer scripts, printed English books, academic materials, typed documents, and administrative documents.

The decoder fine-tuning was performed over 970,000 steps using LoRA with a rank of 8 and an alpha value of 32, following the original setup to adhere to computational constraints. The model was fine-tuned on two NVIDIA GPUs with a total batch size of 16 (8 samples per GPU). We utilized the Adam optimizer with a learning rate of 0.0001, weight decay of 0.1, and cosine learning rate scheduling. This process enables the model to generalize effectively across noisy, unconstrained handwritten inputs and establishes a strong foundation for understanding layout and text.

Stage-2: Form-specific Adaptation. In the second stage, we conducted lightweight fine-tuning using 800 form images, which is a subset of our **Form6000** dataset (Fig. 2). To enhance our training, we applied several augmentation techniques to these images, generating additional samples that reflected real-world scenarios. These augmentations are specifically designed to address domain-specific challenges such as variations in orientation, lighting conditions, and alignment of fields in documents captured under real-world conditions. This stage involved training for an additional 15,000 steps using the same LoRA configuration and hardware setup established for *Stage-1* adaptation. The focused nature of this stage allows the model to quickly adapt to the structure and semantics of handwritten forms, ensuring robustness across inputs from both scanned documents and those captured using mobile devices. Together, these two stages empower the model to effectively manage a diverse range of handwritten document inputs while requiring minimal supervision specific to individual forms.

4.6 Evaluation Metrics

We employed two standard evaluation metrics: Character Recognition Rate (CRR) and Word Recognition Rate (WRR) to quantify the model's performance. These are defined as follows:

$$\text{CRR} = \frac{N_p^c}{N_g^c}, \quad \text{WRR} = \frac{N_p^w}{N_g^w}. \tag{3}$$

Where N_p^c and N_g^c denote the number of correctly predicted characters and the total number of characters in ground truth, respectively. Similarly, N_p^w and N_g^w represent the number of correctly recognized words and the total number of words in the ground truth. Both metrics ranged from 0 to 1, with higher values indicating better recognition performance. CRR captures fine-grained character-level accuracy, whereas WRR reflects the model's ability to correctly recognize complete words, complementing them for evaluating OCR quality.

We also evaluated the performance of the model using Precision (P), Recall (R), and F1-score (F1), defined as:

$$\mathrm{P} = \frac{TP}{TP + FP}, \ \mathrm{R} = \frac{TP}{TP + FN}, \ \text{and F1} = \frac{2 \cdot P \cdot R}{P + R}, \tag{4}$$

Here, TP, FP, and FN refer to the number of true positives, false positives, and false negatives measured over character spans. Precision reflects the proportion of correctly recognized characters among all predicted characters, whereas Recall indicates the proportion of ground-truth characters.

5 Result Analysis

5.1 Comparison with Existing Form Processors

Method	WRR	CRR	P	R	F1
Google Form Parser	92.14	96.38	88.90	90.45	89.67
Azure Form Recognizer	<u>93.29</u>	<u>97.25</u>	<u>91.12</u>	<u>92.60</u>	<u>91.85</u>
PaddleOCR	35.23	64.03	52.21	48.70	50.40
DocTR	32.22	65.44	50.93	46.28	48.50
Donut	65.33	70.12	66.45	67.21	66.83
Naugat	76.13	83.39	77.90	79.15	78.52
FormLens (ours)	**95.44**	**98.33**	**94.12**	**95.31**	**94.71**

Table 2: The performance (all metrics in %) comparison of various methods on scanned and mobile-captured handwritten forms from the Form6000 dataset. The bold and underline values indicate the best and second-best results, respectively.

To evaluate the performance of our approach, we benchmark **FormLens** against leading commercial and open-source OCR systems on a held-out set of 5,200 scanned and mobile-captured handwritten forms from the **Form6000** dataset. The comparison includes *Google Form Parser*, *Azure Form Recognizer*, *PaddleOCR* [10], and *DocTR* [33], with results summarized in Table 2.

Among the commercial systems, *Azure Form Recognizer* achieves the highest WRR at 93.29% and a CRR of 97.25%, with a Precision of 91.12% and Recall of 92.60%, closely followed by *Google Form Parser* (WRR: 92.14%, CRR: 96.38%, Precision: 88.90%, and Recall: 90.45%). These systems perform well on clean scanned forms, but their performance degrades noticeably on mobile-captured handwritten inputs, particularly under challenging conditions like tilted orientations, cluttered backgrounds, and uneven lighting. Open-source tools such as *PaddleOCR* (WRR of 35.23%, CRR of 64.03%, Precision of 52.21%, and Recall of 48.70%) and *DocTR* (WRR of 32.22%, CRR of 65.44%, Precision of 50.93%, and Recall of 46.28%) struggle to produce good accuracy, which is limited by their inability to handle such diverse handwritten inputs or produce structured outputs reliably.

In contrast, **FormLens** achieves the best overall results, with a WRR of 95.44%, CRR of 98.33%, a Precision of 94.12%, and Recall of 95.31%, outperforming all baselines. Its end-to-end architecture allows for processing entire form images holistically, without requiring intermediate layout detection or rule-based parsing. This end-to-end design makes it significantly more robust to visual noise, handwriting variability, and structural inconsistencies. Although the accuracy gains over commercial systems may appear modest in absolute terms, **FormLens** consistently demonstrates superior reliability in real-world scenarios.

We evaluate two **OCR-free baselines**, *Donut* [24] and *Nougat* [6], on the **Form6000** dataset. *Donut* achieved 65.33/70.12% (WRR/CRR) and 66.83% F1, while *Nougat* obtained 76.13/83.39% and 78.52% F1. Despite good layout generalization, both under perform compared to **FormLens** (96.16/98.56%), highlighting its superior handling of handwritten forms and structural consistency.

Qualitative results in Fig. 5 further highlights its strength in maintaining semantic alignment and preserving form structure across diverse capture settings[1].

5.2 Results on Existing Benchmark

Method	WRR	CRR	P	R	F1
UNIT [50]	64.52	69.33	65.11	69.22	67.14
Google Form Parser	90.44	<u>95.88</u>	92.44	<u>93.72</u>	93.07
Azure Form Recognizer	<u>91.26</u>	95.66	<u>93.51</u>	93.32	<u>93.41</u>
FormLens (Ours)	**91.32**	**96.70**	**94.22**	**93.59**	**93.90**

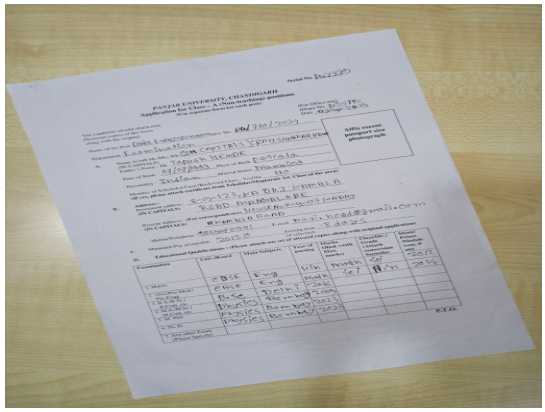
Table 3: Comparison of model performance (all metrics in %) on the FUNSD dataset, illustrating generalization to unseen semi-structured forms. Bold and underlined values represent the best and second-best results, respectively.

To evaluate the generalization capability of **FormLens** beyond our primary dataset (**Form6000**), we assess its performance on the publicly available **FUNSD** [21] benchmark, which comprises unseen, semi-structured form layouts. We compare **FormLens** against both commercial OCR systems (*Google Form Parser* and *Azure Form Recognizer*) and a recent research model (*UNIT* [50]). As reported in Table 3, **FormLens** achieves the highest scores across all metrics — WRR (91.32%), CRR (96.70%), Precision (94.22%), Recall (93.59%), and F1-score (93.90%) — surpassing all baselines. These results highlight **FormLens**'s strong cross-domain adaptability and robust performance with minimal dataset-specific tuning.

5.3 Ablation Study

Impact of Form-Specific Adaptation: To assess the impact of form-specific adaptation, we compare two variants of our model: **FormLens***, trained on a broad set of generic handwritten and printed documents, and **FormLens**, fine-tuned with just 800 forms from the **Form6000** dataset. While **FormLens*** performs reasonably on clean, scanned inputs, it struggles with mobile-captured forms exhibiting noise, skew, and clutter — highlighting the limitations of generic adaptation for structured form understanding. In

[1]Additional visual results and detailed analysis are provided in the supplementary material.



(a) Input Image

Serial No.: PU7329
PANJAB UNIVERSITY, CHANDIGARH
Application for Class- A (Non-teaching) positions
(Use separate form for each post)
The candidate should attach nine
Photostat copies of this form
along-with the original
(For Office use)
(Diary No): D54786
Date: 03/06/2025
Name of the Post: Data Entry Opertor
Advt. No.: PU/710/2024
Department: Examination
A. Name in full Mr/Mrs/Ms.(IN CAPITALS): RAVI SHANKER HEGDE
Father's Name: Mr. Tapash HEGDE
Date of Birth: 07/07/1993
Place of Birth: Pataila
Nationality: Indian
Marital Status: Married
Member of Scheduled Caste/Backward Class: No
(If yes, please attach certificate from Tehsildar/Magistrate 1st Class of the area)
B. Address
Permanent address(IN CAPITALS): HOUSE NO. 8-9-123, , KADRI KAMBLA ROAD MANGALORE
Present Address (For correspondence) (IN CAPITALS): HOUSE NO. 8-9-123, KADRI KAMBLA ROAD
Mobile/Telephone: 9845678901
E-mail: ravi.hegde@gmail.com
C. Minimum pay acceptable: 20120
Joining time (if selected): 8 days
D. Educational Qualifications (Please attach one set of attested copies along-with original application)
Examination Univ./Board, Main Subjects Year of passing, Marks Obtd.(with Max. marks), Class/Div/Grade(Attach conversion formula), Merit/Prizes/Medals won, if any
1. Matric: CBSE, Eng, Hin, Math, Sci, 2012
2. 10+2/Pre-Med./Pre-Engg: CBSE, Eng,Math,Sci,Hin, 2014,
3. B.A/B.Sc./BCom. etc: B.Sc,Delhi, 2016,
4. M.A./M.Sc/M.Com. etc: M.Sc.,Physics, Bombay, 2019
5. M. Phil: Physics, Bombay, 2023, -, -, -
6. Ph.D: Physics, Bombay, 2023, -, -, -
7 Any other Exam (Please Specify):
P.T.O

(b) FormLens (Ours)

Serial No.: PU732g
PANJAB UNIVERSITY, CHANDIGARH
Application for Class- A (Non-teaching) positions
(Use separate form for each post)
The candidate should attach nine
Photostat copies of this form
along-with the original
(For Office use)
(Diary No): D54786
Date: 03/06/2025
Name of the Post: Data Ently Opertor
Advt. No.: PU/710/2024
Department: Examination
A. Name in full Mr/Mrs/Ms.(IN CAPITALS): RAVI SHANKAR HEDGE
Father's Name: Mr. Tapash HEGDE
Date of Birth: 07/07/1993
Ptace of Birth: Pataila
Nationality: Indian
Marital Status: Married
Member of Scheduled Caste/Backward Class: No
(If yes, please attach certificate from Tehsildar/Magistrate 1st Class of the area)
B. Address
Permanent address(IN CAPITALS): HOUSE NO. 8-9-123, , KADRI KAMBLA ROAD MANGALORE
Present Address (For correspondence) (IN CAPITALS): HOUSE NO. 8-9-123, KADRI KAMBLA ROAD
Mobile/Telephone: 9845678901
E-mail: ravi.hegde@gmail.com
C. Minimum pay acceptable: 20120
Joining time (if selected): 8 days
D. Educational Qualifications (Please attach one set of attested copies along-with original application)
Examination Univ./Board, Main Subjects Year of passing, Marks Obtd.(with Max. marks), Class/Div/Grade(Attach conversion formula), Merit/Prizes/Medals won, if any
1. Matric: CBSE, Eng, Hin, Math, Sci, 201Z
2. 10+2/Pre-Med./Pre-Engg: CBSE, Eng,Math,Sci,Hin, 201A
3. B.A/B.Sc./BCom. etc: B.Sc,Delhi, 2016,
4. M.A./M.Sc/M.Com. etc: M.Sc.,Physics, Bombay, 2019
5. M. Phil: Physics, Bombay, 2023, -, -, -
6. Ph.D: Physics, Bombay, 2023, -, -, -
7 Any other Exam (Please Specify):
P.T.O

(c) Google Form Parser

Serial No.: PU7329
PANJAB UNIVERSITY, CHANDIGARH
Application for Class- A (Non-teaching) positions
(Use separate form for each post)
The candidate should attach nine
Photostat copies of this form
along-with the original
(For Office use)
(Diary No): D54786
Date: 03/06/2025
Name of the Post: Data Enrty Operator
Advt. No.: PU/710/2024
Department: Examination
A. Name in full Mr/Mrs/Ms.(IN CAPITALS): RAVI SHANKAR HEGD
Father's Name: Mr. Tapsh HEGDE
Date of Birth: 07/07/1993
Place of Birth: Patiala
Nationality: Indian
Marital Status: Married
Member of Scheduled Caste/Backward Class: No
(If yes, please attach certificate from Tehsildar/Magistrate 1st Class of the area)
B. Address
Permanent address(IN CAPITALS): HOUSE NO. 8-9-123, , KADRI KAMBLA ROAD MANGALORE
Present Address (For correspondence) (IN CAPITALS): HOUSE NO. 8-9-123, KADRI KAMBLA ROAD
Mobile/Telephone: 9845678901
E-mail: ravi.hegde@gmail.com
C. Minimum pay acceptable: 20120
Joining time (if selected): 8 days
D. Educational Qualifications (Please attach one set of attested copies along-with original application)
Examination Univ./Board, Main Subjects Year of passing, Marks Obtd.(with Max. marks), Class/Div/Grade(Attach conversion formula), Merit/Prizes/Medals won, if any
1. Matric: CBSE, Eng, Hin, Math, Sci, 2012
2. 10+2/Pre-Med./Pre-Engg: CBSE, Eng,Math,Sci,Hin, 2014
3. B.A/B.Sc./BCom. etc:B.Sc,Delhi, 2016,
4. M.A./M.Sc/M.Com. etc: M.Sc.,Physics, Bombay, 2019
5. M. Phil: Physics, Bombay, 2023, -, -, -
6. Ph.D: Physics, Bombay, 2023, -, -, -
7 Any other Exam (Please Specify):
P.T.O

(d) Azure Form Recognizer

Figure 5: Qualitative comparison of handwritten form digitization outputs across systems. FormLens demonstrates markedly better transcription quality on mobile-captured forms, particularly under challenging conditions such as handwriting variation, low lighting, and rotated layouts. Compared to *Google Form Parser* and *Azure Form Recognizer*, it more accurately handles irregular scripts, maintains field alignment, and infers field boundaries even in noisy or cluttered inputs. Wrongly recognized characters are highlighted in red.

Model	Fine-tune	WRR	CRR	P	R	F1
FormLens*	GA	75.22	80.19	77.43	78.22	77.82
FormLens	FSA	**95.66**	**98.76**	**96.91**	**97.45**	**97.18**

Table 4: Illustrates impact of form-specific adaptation on FormLens performance (all metrics in %) for handwritten form digitization. GA indicates generic adaptation and FSA indicates form-specific adaptation. The bold and underlined values indicate the best and the second-best results.

contrast, **FormLens** achieves significantly higher accuracy, with a WRR of 95.66% and CRR of 98.76%, Precision of 96.91% and Recall of 97.45%, as shown in Table 4, demonstrating that even minimal domain adaptation leads to substantial gains in robustness and recognition performance under real-world conditions. These findings underscore the importance of targeted adaptation for reliable form digitization in low-resource settings.

Impact of Imaging Conditions on Model Performance: We assess **FormLens** under four input conditions to reflect real-world

Imaging Condition	WRR	CRR	P	R	F1
Scanned (Clean)	**97.12**	**99.02**	**96.88**	**97.55**	**97.21**
Camera Captured (standard)	95.87	98.51	95.01	96.12	95.56
Captured + RT & BG	95.43	98.27	94.65	95.72	95.18
Captured + S + LV	93.33	97.51	92.26	93.81	93.03

Table 5: Performance (all metrics in %) of FormLens across varied input conditions, highlighting its robustness to real-world capture challenges. RT, BG, S and LV indicate rotation, background, scale, and lighting variation, respectively. The bold and underlined values indicate the best and the second-best results.

form submission scenarios: (i) clean scanned forms, (ii) standard mobile camera captures, (iii) mobile captures with added rotation and background clutter, and (iv) captures exhibiting scale changes and lighting variations. As reported in Table 5, the model achieves the highest accuracy on clean scanned forms, with a WRR of 97.12%

and CRR of 99.02%. Importantly, **FormLens** maintains strong performance across increasingly challenging settings, achieving over 93% WRR and 97% CRR even under Precision degradation and Recall drops caused by compounded distortions like skew, lighting imbalance, and cluttered backgrounds. These results underscore the model's robust generalization to diverse imaging conditions, enabled by its lightweight form-specific adaptation.

Type	Condition	WRR	CRR	P	R	F1
Perspective	Bright	**90.12**	**94.55**	**89.72**	**90.11**	**89.91**
	Colored	89.65	94.22	89.10	89.76	89.43
	Shadow	85.32	90.12	84.60	85.20	84.90
Curled	Bright	**83.21**	**88.55**	**82.75**	**83.10**	**82.92**
	Shadow	80.12	88.12	79.40	80.05	79.72
Folded	Fewfold	**92.11**	**96.33**	**91.72**	**92.08**	**91.90**
	Multifold	89.22	93.22	88.90	89.15	89.02
Crumpled	Easy–Colored	91.22	96.77	90.85	91.15	91.00
	Easy–Bright	**91.39**	**98.12**	**90.95**	**91.32**	**91.13**
	Hard–Colored	78.12	85.49	77.60	78.05	77.82
	Hard–Shadow	76.53	84.78	75.85	76.40	76.12

Table 6: Performance (all metrics in %) of FormLens on the *Inv3d* dataset, demonstrating robustness across diverse lighting and deformation scenarios. The bold and underlined values indicate the best and second-best results within each category.

Impact of Model Robustness on Lighting and Distortion Conditions. We assess **FormLens**'s real-world generalization on the *Inv3d* dataset [16], comprising 360 smartphone-captured invoices with varied lighting and geometric distortions. As shown in Table 6, **FormLens** achieves up to 90.12% WRR and 94.55% CRR under perspective distortion, maintains over 89% F1 under lighting variations, and records 80–96% WRR/CRR for curled and folded forms. Even on crumpled inputs, it attains 91% F1 (easy) and 76–78% (hard), demonstrating strong robustness across real-world conditions.

Noise Type	WRR	CRR	P	R	F1
Gaussian Noise (sigma=5)	94.91	97.96	93.41	94.52	93.96
Median Noise (p=5)	**95.10**	98.14	93.88	95.33	94.60
Gaussian Blur (3,3)	94.72	97.90	93.19	94.30	93.74
Average Blur (3,3)	95.01	**98.33**	**94.21**	**95.88**	**95.04**

Table 7: Performance (all metrics in %) of FormLens under synthetic image degradations simulating real-world noise and quality variations. The bold and underlined values indicate the best and the second-best results.

Impact of Noisy Inputs on Model Performance: To evaluate **FormLens** under real-world noise conditions, we apply four synthetic distortions: Gaussian noise, Median noise, Gaussian blur, and Average blur, mimicking common capture issues like low-light grain, motion blur, and lens softness. As shown in Table 7, **FormLens** remains highly robust, with WRR above 94.72% and CRR above 97.90% across all cases. The model shows minimal degradation, particularly under Precision and Recall with median and average blur. This resilience stems from its end-to-end vision-language design, which avoids the error propagation of multi-step OCR pipelines, and its LoRA-based fine-tuning, which enables efficient adaptation to noisy, real-world data. These results highlight FormLens's suitability for deployment in unconstrained environments.

Field-Level Evaluation. To further illustrate **FormLens**'s practical performance, we assess field-level metrics in the *Ablation Study*. Beyond word and character recognition, we compute field-level exact match and per-field F1 on the *Form6000* dataset, achieving 92.36% and 95.13%, respectively. These results confirm that **FormLens** not only ensures accurate word recognition but also reliably extracts complete field values — crucial for robust end-to-end form digitization in real-world settings.

Model	WRR	CRR	P	R	F1
FormLens	**95.71**	**98.92**	**96.21**	**95.71**	**95.96**
Google Form Parser	94.85	98.10	95.35	94.85	95.10

Table 8: Performance (all metrics in %) on the IAM dataset (page-level handwritten documents). Bold and underlined values indicate the best and second-best results, respectively.

Impact on IAM, a Generic Handwritten Dataset: To further evaluate robustness, we tested **FormLens** on the IAM dataset [15], comprising complex, page-level handwritten documents with diverse writing styles. As shown in Table 8, **FormLens** without fine-tuning with the IAM dataset achieves a WRR of 95.71%, CRR of 98.92%, Precision of 96.21%, Recall of 95.71%, and an F1-score of 95.96%, outperforming *Google Form Parser* across all metrics. It confirms the model's ability to generalize beyond structured form data, effectively handling unconstrained handwriting and underscores its suitability for large-scale handwritten document digitization in real-world scenarios.

6 Conclusion

We introduced **FormLens**, an adaptation of the vision-language model GOT 2.0, designed for end-to-end handwritten form digitization. It directly converts full-page inputs into structured key-value pairs — bypassing traditional region detection, OCR, and rule-based processing. To support realistic evaluation and training, we release **Form6000**, a challenging benchmark dataset of 6,000 handwritten forms collected from 650 individuals under diverse real-world conditions, including varied layouts, backgrounds, and capture orientations. Our two-stage decoder adaptation — combining generic and form-specific data — enhances robustness to layout variability, handwriting styles, and image degradations. Extensive experiments show that **FormLens** outperforms commercial and open-source baselines in accuracy and robustness, setting a new benchmark for low-resource handwritten form digitization. Future work includes multilingual support and improved structured field extraction.

Acknowledgments

This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

References

[1] Amazon Web Services. [n.d.]. Amazon Textract: Automatically extract text, handwriting, and data from documents. https://aws.amazon.com/textract/.

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In *ICCV*. 993–1003.

[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *CVPR*. 9365–9374.

[4] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. (2019), 9365–9374.

[5] Jinze Bai and others. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[6] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418* (2023).

[7] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. 2018. Rosetta: Large scale system for text detection and recognition in images. In *ACM SIGKDD*. 71–79.

[8] Roberto Brunelli. 2008. Template matching techniques in computer vision. *Theory and Practice* (2008), 25–28.

[9] Denis Coquenet, Clément Chatelain, and Thierry Paquet. 2022. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE* 45 (2022), 508–524.

[10] Cheng et al. Cui. 2025. PaddleOCR 3.0 Technical Report.

[11] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. 2018. Pixellink: Detecting scene text via instance segmentation. (2018).

[12] Daniel Hernandez Diaz, Siyang Qin, Reeve Ingle, Yasuhisa Fujii, and Alessandro Bissacco. 2021. Rethinking text line recognition models. *arXiv preprint arXiv:2104.07787* (2021).

[13] Wan et al. 2024. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *CVPR*. 15641–15653.

[14] Google Cloud. [n.d.]. Form Parser | Document AI. https://cloud.google.com/document-ai/docs/form-parser. Accessed on [Current Date].

[15] Samuel Grieggs, Bingyu Shen, Greta Rauch, Pei Li, Jiaqi Ma, David Chiang, Brian Price, and Walter J Scheirer. 2021. Measuring human perception to improve handwritten document transcription. *IEEE* 44 (2021), 6594–6601.

[16] Felix Hertlein, Alexander Naumann, and Patrick Philipp. 2023. Inv3D: a high-resolution 3D invoice dataset for template-guided single-image document unwarping. *International Journal on Document Analysis and Recognition (IJDAR)* (2023).

[17] Hu and others. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1 (2022).

[18] 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 *ACMM*. 4083–4091.

[19] JaidedAI. Year. EasyOCR. https://github.com/JaidedAI/EasyOCR.

[20] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence* 22 (2000), 4–37.

[21] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*. 1–6.

[22] Mahsa Kholghi, Xiaoxiao Shao, Subhojeet Ghosh, et al. 2021. FormX: A comprehensive benchmark for visual information extraction from forms. *arXiv preprint arXiv:2106.11363* (2021).

[23] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free Document Understanding Transformer. In *ECCV*. 498–517.

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

[25] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, et al. 2021. MMOCR: a comprehensive toolbox for text detection, recognition and understanding. In *ACMM*. 3791–3794.

[26] Alex W. C. Lee, Jonathan Chung, and Marco Lee. 2021. GNHK: A Dataset for English Handwriting in the Wild. In *ICDAR*. 399–412.

[27] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *AAAI*. 13094–13102.

[28] Weihong Lin, Zheng Sun, Chixiang Ma, Mingze Li, Jiawei Wang, Lei Sun, and Qiang Huo. 2022. TSRFormer: Table structure recognition with transformers. In *ACM MM*. 6473–6482.

[29] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. 2021. Parsing table structures in the wild. In *ICCV*. 944–952.

[30] Rujiao Long, Hangdi Xing, Zhibo Yang, Qi Zheng, Zhi Yu, Fei Huang, and Cong Yao. 2025. LORE++: Logical location regression network for table structure recognition with pre-training. *PR* (2025).

[31] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. Textsnake: A flexible representation for detecting text of arbitrary shapes. (2018), 20–36.

[32] Microsoft Azure. [n.d.]. Azure Form Recognizer documentation. https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence/. Accessed on [Current Date].

[33] Mindee. 2021. docTR: Document Text Recognition. https://github.com/mindee/doctr.

[34] docTR Mindee et al. 2022. docTR: a Document Text Recognition toolbox. *GitHub repository* (2022). https://github.com/mindee/doctr

[35] Isaac Sanchez and Enrique Vidal. 2016. Bentham Dataset R0. https://doi.org/10.5281/zenodo.44519

[36] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. In *ICDAR*. 131–146.

[37] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *ICDAR*. 629–633.

[38] Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *CVPR*. 4634–4642.

[39] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534* (2020).

[40] 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 *CVPR*. 19254–19264.

[41] Haoran Wei and others. 2024. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. *ArXiv* (2024).

[42] 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 *ACM SIGKDD*. 1192–1200.

[43] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Linjun Shao. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *ACM KDD*. 1192–1200.

[44] 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 *ACL*. 3214–3224.

[45] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

[46] Le Yang, David Burdick, and Yann LeCun. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *CVPR*. 5315–5324.

[47] Susan Zhang and others. 2022. OPT: Open Pre-trained Transformer Language Models. *ArXiv* (2022).

[48] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In *WACV*. 697–706.

[49] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *CVPR*. 5551–5560.

[50] Yi Zhu, Zhou Yanpeng, Chunwei Wang, Yang Cao, Jianhua Han, Lu Hou, and Hang Xu. 2024. Unit: Unifying image and text recognition in one vision encoder. *Advances in Neural Information Processing Systems* 37 (2024), 122185–122205.