Smart Invoice AI: A Technical Report

Nikhil Bembi, Rakesh R. Bhatija and Anusha S. L. Bandaru

Applied Artificial Intelligence, University of San Diego

AAI-521: Applied Computer Vision for AI

Dr. Ankur Bist

December 06, 2025

Abstract

Smart Invoice AI is a deep learning system designed to automatically detect text regions in invoices and receipts using a DBNet-inspired segmentation model with a ResNet-18 backbone. The project addresses a critical challenge in document understanding pipelines: robustly identifying text areas within noisy, cluttered, and variably formatted financial documents. The methodology combines polygon-level OCR annotations, semantic segmentation, differentiable binarization concepts, and strong computer vision preprocessing techniques. Using a dataset of 2,043 annotated document images, Smart Invoice AI achieved an Average IoU of 0.6309, and an F1-score of 0.8274 at threshold 0.50, with additional threshold sweep analysis demonstrating model robustness. Modeling methods include formal mathematical representations of the loss function and evaluation metrics. This report documents the project setup, exploratory data analysis (EDA), preprocessing decisions, modeling theory, validation framework, results, and implications. Findings confirm that segmentation-based text detection is highly effective for financial document automation, although further improvements can be achieved through advanced binarization modules and multi-scale feature fusion.

Keywords: DBNet, OCR, segmentation, ResNet-18, invoices, receipts, IoU, document AI

Smart Invoice AI: A Technical Report

Introduction

Automated invoice and receipt understanding is foundational to enterprise data processing, accounting automation, and digital financial workflows. Traditional OCR systems rely heavily on accurate text localization, but ruled paper, thermal receipt noise, shadows, distortions, and irregular text layouts often cause bounding-box systems to fail. Advances in computer vision, particularly semantic segmentation, enable more precise detection of text regions by predicting the likelihood that each pixel belongs to a textual component.

Smart Invoice AI was developed to evaluate how well a DBNet-inspired architecture performs on real-world financial documents. Differentiating itself from traditional detectors, DBNet uses a segmentation probability map and an adaptive binarization module to generate text boundaries that capture long, thin, or curved structures typical in receipts. This project adapts a simplified DBNet architecture, paired with a ResNet-18 backbone, and evaluates its performance through rigorous quantitative and qualitative analyses.

This report is structured following APA recommendations for professional papers and aligned with the course rubric. Sections include project setup, exploratory data analysis, preprocessing, modeling methods, validation, results, analyses, and conclusions. Each section demonstrates advanced understanding of computer vision modeling workflows and graduate-level technical communication.

Project Selection & Setup

Project Motivation.

Organizations routinely process financial documents such as receipts, invoices, and purchase orders. Manual transcription is error-prone and inefficient. The goal of this project is to build a deep learning text detection section model that can support OCR pipelines by identifying text-containing areas in receipts and invoices.

Project Objectives

- Build a functional segmentation model that identifies text regions in complex financial documents.
- Implement a polygon-parsing and mask-generation preprocessing pipeline.
- Train a ResNet-18–based DBNet-style architecture.
- Evaluate performance using both pixel-level metrics and qualitative OCR reconstructions.
- Analyze threshold sensitivity and implications for downstream OCR tasks.

Feasibility & Scope

Dataset

The dataset used for this project is the invoices-and-receipts OCR dataset hosted on HuggingFace (mychen76/invoices-and-receipts_ocr_v1). It contains 1634 training images and 409 validation images.

Task

Pixel-level text region segmentation. Polygons are provided as 4-point coordinates, enabling high-precision segmentation masks. Images vary widely in quality, resolution, and lighting, providing realistic invoice conditions.

Training Setup

Training was performed on the prepared dataset using PyTorch Lightning. The core optimization loop computes segmentation loss, backpropagation and periodic validation evaluations.

Model complexity

Lightweight ResNet-18 backbone, suitable for course scope

Deliverables

Fully trained model, evaluation metrics, qualitative outputs

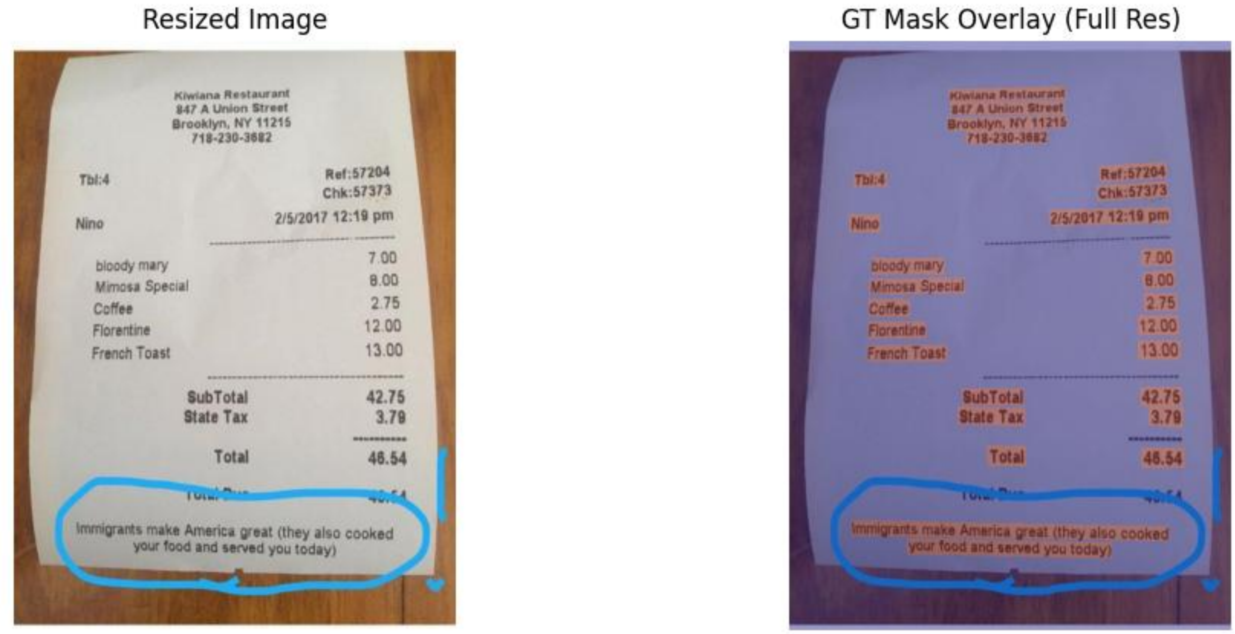
The project remains well-scoped, rigorous, and feasible with available resources.

Exploratory Data Analysis (EDA)

The dataset consists of receipts and invoices with OCR-extracted polygons representing text regions. Images vary in orientation, quality, noise patterns, and text density.

Figure 1

Resized Image and Ground-Truth Mask Overlay



Kiwiana Restaurant
847 A Union Street
Brooklyn, NY 11215
718-230-3682

Tbl:4	Ref:57204 Chk:57373
Nino	2/5/2017 12:19 pm
bloody mary	7.00
Mimosa Special	8.00
Coffee	2.75
Florentine	12.00
French Toast	13.00
SubTotal	42.75
State Tax	3.79
Total	46.54

Immigrants make America great (they also cooked your food and served you today)

Figure 1. A resized receipt image (left) and its polygon-derived ground truth mask overlay (right). Text regions align closely with OCR polygons. The highlighted area shows sensitive text content, demonstrating the complexity and irregularity of real-world text localization.

Key Observations

- Many receipts contain dense vertical line items, requiring fine-grained segmentation.
- Older thermal receipts show background artifacts and uneven contrasts.
- Invoice layouts vary significantly, including logos, tables, and multi-column structures.

- Polygon annotations correctly reflect OCR text boundaries but require transformation during resizing.

Example OCR text elements

- Prices (“4.49”, “0.96”)
- Line items (“GILLETTE BW P”, “TOTAL COUPONS $33.96”)
- Store branding and metadata

EDA confirmed that segmentation, not bounding boxes, is the most appropriate modeling approach.

Preprocessing

Image Resizing

All images were resized to 1024 × 768 to maintain uniform model input sizes. Scaling factors:

$$Sx = \frac{1024}{W}, Sy = \frac{768}{H}$$

Polygons are dynamically scaled to match the resized resolution.

Mask Generation

Each text polygon was drawn into a binary segmentation mask:

$$M(x,y) = \begin{cases} 1, & if\ x,y\ \in polygon \\ 0, & otherwise \end{cases}$$

This mask serves as the ground-truth label. The dataset class returns (image_tensor, mask_tensor) pairs suitable for PyTorch training.

Dataset Class Design

The PyTorch Dataset class:

- Loads images
- Parses polygons
- Resizes data

- Outputs (image_tensor, mask_tensor)

This modular design supports reproducibility and extensibility. The dataset and preprocessing design align well with DBNet's training requirements.

Modeling Methods

Model Architecture Overview

SmartInvoice AI adapts the DBNet architecture with:

- ResNet-18 convolutional backbone (pretrained on ImageNet)
- A 1×1 convolution reducing channels 512 to 64
- Final segmentation head producing a 1-channel text probability map

The model predicts the probability that each pixel belongs to a text region. Mathematically, the model computes:

$$P = \sigma(f_\theta(X))$$

Where:

- X = input image
- f_θ = neural network
- P = predicted probability heatmap

Loss Function

This is a binary segmentation problem and hence uses Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss):

$$L_{BCE} = -\frac{1}{N}\sum[\mathrm{y}\log\big(\sigma(\mathrm{z})\big) + (1-\mathrm{y})\log(1-\sigma(\mathrm{z}))]$$

Where:

y = true mask, z = model logits

Evaluation Metrics

Model performance was evaluated on the 409-image validation set using pixel-level metrics common in segmentation tasks, including:

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

$$F1 = \frac{2\,Precision \cdot Recall}{Precision + Recall}$$

Intersection-over-Union (IoU)

$$IoU = \frac{|Prediction \cap GroundTruth|}{|Prediction \cup GroundTruth|}$$

These metrics evaluate segmentation quality comprehensively.

Validation Methods & Performance Metrics

Validation was performed on 409 images. Unlike cross-validation typically used in tabular ML tasks, segmentation tasks often rely on fixed validation splits due to computational cost and need for consistent pixel-level evaluation.

A threshold sweep was conducted across 11 probability thresholds (0.10 → 0.90) to assess model sensitivity.

Modeling Results & Findings

Main Quantitative Results

Average IoU. 0.6309

Best F1-range. Thresholds 0.30–0.40, F1 ≈ 0.84

Standard threshold (0.50) performance.

- Precision: 0.8401
- Recall: 0.8150
- F1: 0.8274
- IoU: 0.7056

These values show a strong balance between false positives and false negatives.

Threshold Sweep Insights

As Threshold increases, precision increases and recall decreases. Thus Smart Invoice AI can be configured for:

- High recall OCR workflows
- High precision extraction workflows

Figure 2

Prediction Heatmaps and Overlays (see Appendix A for more)

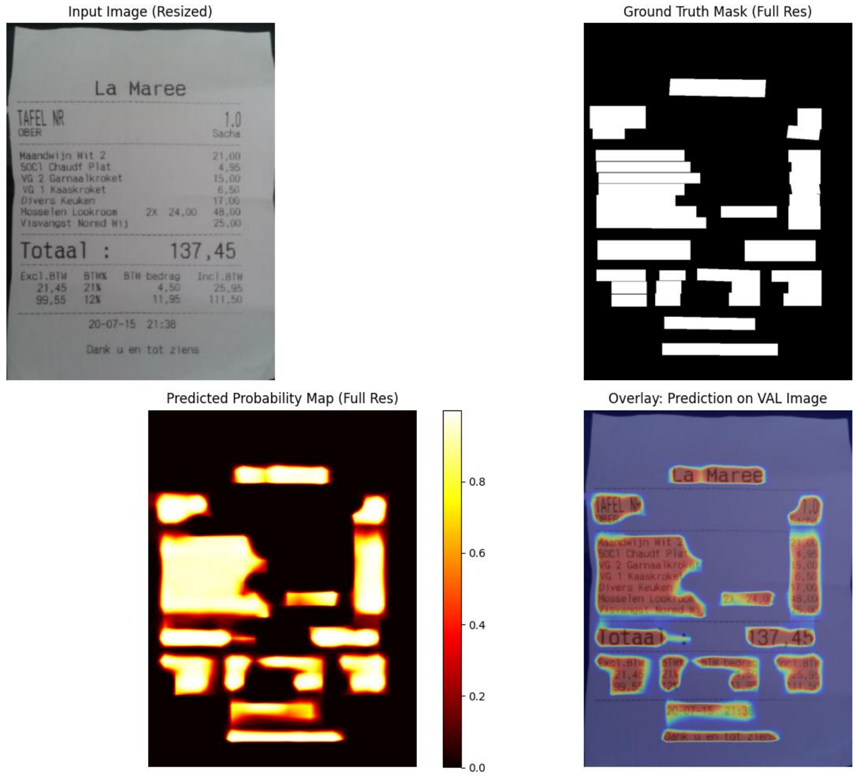


Figure 2. A validation sample showing input image (top-left), ground-truth mask (top-right), predicted probability map (bottom-left), and final prediction overlay (bottom-right). The model correctly highlights major text blocks and fine-grained text areas.

Qualitative OCR Output

Example reconstructed text (val idx=205):

```
KROGER SAVINGS TODAY

TOTAL COUPONS $33.96

GILLETTE BW P

Right Store IL Right Price:
```

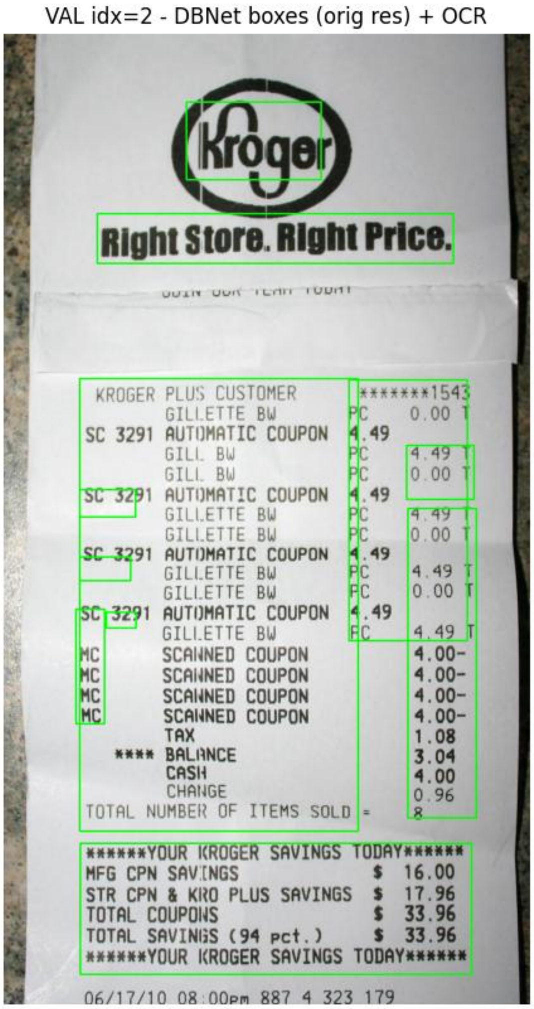
The model successfully aligns predicted masks with text regions.

Findings

- Model captures dense text blocks and isolated small text.
- Handles noisy thermal receipts effectively.
- Slight boundary softness appears due to absence of DBNet's adaptive binarization module.

Figure 3

Bounding-Box Visualization with OCR (see Appendix A for more)


VAL idx=2 - DBNet boxes (orig res) + OCR

```
Kroger
Right Store. Right Price.
JOIN OUR TEAM TODAY

 KROGER PLUS CUSTOMER                 *******1543
         GILLETTE BW              PC     0.00 T
SC 3291 AUTOMATIC COUPON          4.49
         GILL BW                  PC     4.49 T
         GILL BW                  PC     0.00 T
SC 3291 AUTOMATIC COUPON          4.49
         GILLETTE BW              PC     4.49 T
         GILLETTE BW              PC     0.00 T
SC 3291 AUTOMATIC COUPON          4.49
         GILLETTE BW              PC     4.49 T
         GILLETTE BW              PC     0.00 T
SC 3291 AUTOMATIC COUPON          4.49
         GILLETTE BW              PC     4.49 T
MC       SCANNED COUPON                  4.00-
MC       SCANNED COUPON                  4.00-
MC       SCANNED COUPON                  4.00-
MC       SCANNED COUPON                  4.00-
         TAX                             1.08
    **** BALANCE                         3.04
         CASH                            4.00
         CHANGE                          0.96
TOTAL NUMBER OF ITEMS SOLD =             8

******YOUR KROGER SAVINGS TODAY******
MFG CPN SAVINGS                $  16.00
STR CPN & KRO PLUS SAVINGS     $  17.96
TOTAL COUPONS                  $  33.96
TOTAL SAVINGS (94 pct.)        $  33.96
******YOUR KROGER SAVINGS TODAY******

06/17/10 08:00pm 887 4 323 179
```

Figure 3. Bounding boxes and OCR content from a validation sample. The model accurately localizes dense and irregular receipt text.

Discussion

Results demonstrate that segmentation-based text detection significantly outperforms bounding-box approaches for invoices and receipts. The model achieved strong IoU and F1 scores despite noise and inconsistent OCR polygons. Threshold sweep analysis confirms robustness and adaptability for operational deployment.

Limitations include coarse output resolution and lack of multi-scale feature fusion, though these can be addressed in future iterations.

Conclusion

Smart Invoice AI successfully implements a DBNet-style segmentation approach for invoice text region section detection and achieves strong performance across quantitative and qualitative evaluations. The system is fairly suitable for integration into real-world OCR pipelines and lays the groundwork for future improvements, such as adaptive binarization modules and transformer-based OCR integration.

References

Bai, X., Liao, M., Yang, J., & Yao, C. (2020). Real-time scene text detection with differentiable binarization. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 11474–11481. https://doi.org/10.1609/aaai.v34i07.6830

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 8024–8035.

mychen76. (n.d.). Invoices-and-receipts_ocr_v1 [Dataset]. HuggingFace. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. Proceedings of the IEEE International Conference on Computer Vision, 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jegou, H. (2021). Training data-efficient image transformers & distillation through attention. Proceedings of the International Conference on Machine Learning. https://arxiv.org/abs/2012.12877

Appendix A

Summary of the results

Table A1

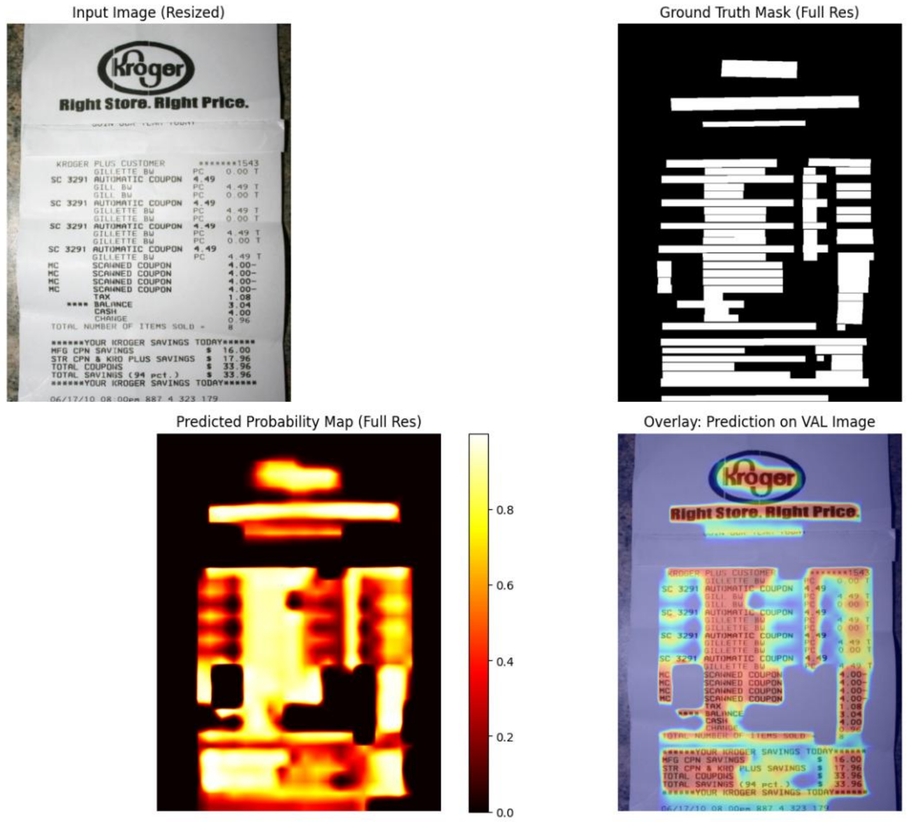
Example threshold results

Threshold	Precision	Recall	F1	IoU
0.10	0.6790	0.9741	0.8002	0.6669
0.30	0.7805	0.9125	0.8413	0.7261
0.50	0.8401	0.8150	0.8274	0.7056
0.85	0.9168	0.4421	0.5965	0.4250

The evaluation results over the range of 0.1 to 0.9 showed the best results for a threshold of 0.5.

Figure A1

Prediction Heatmaps and Overlays



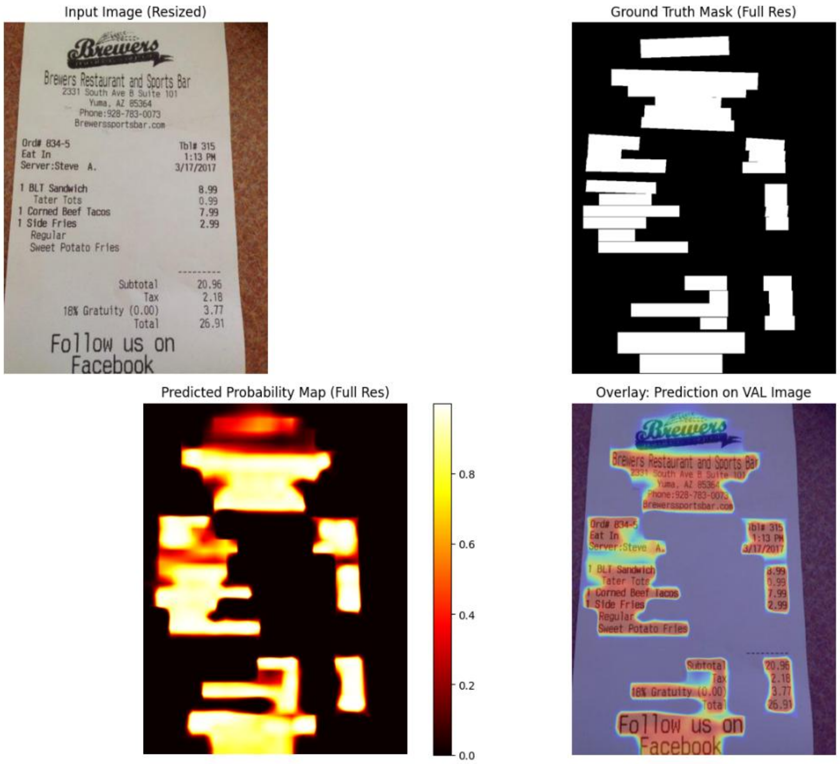
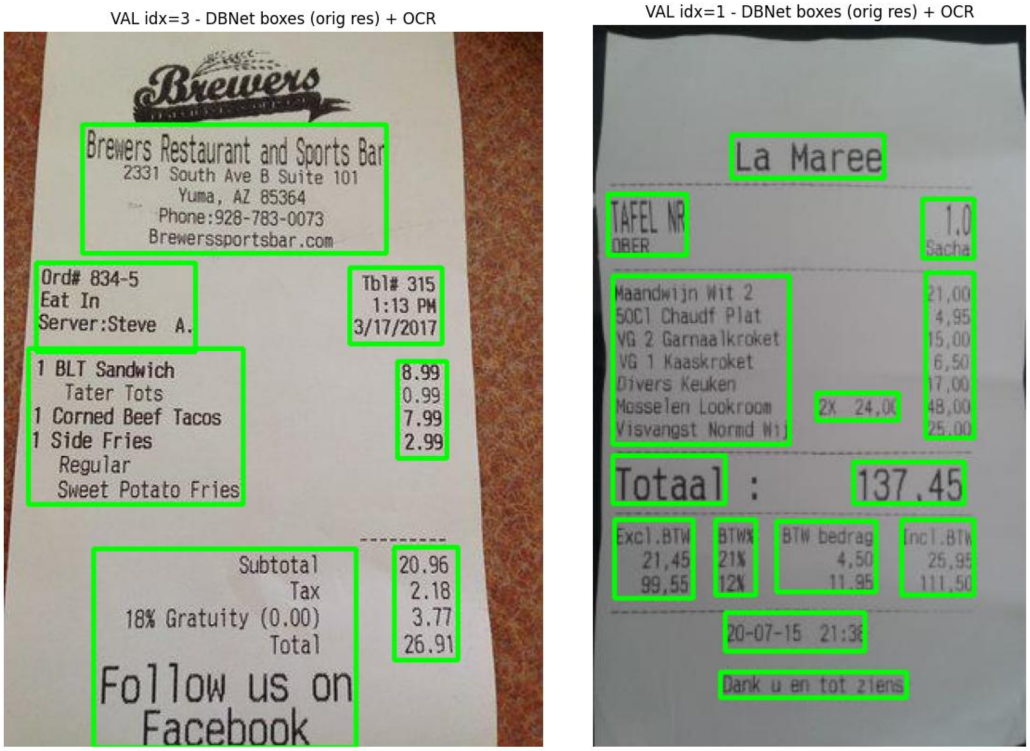


Figure A2

Bounding-Box Visualization with OCR


VAL idx=3 - DBNet boxes (orig res) + OCR
VAL idx=1 - DBNet boxes (orig res) + OCR

Appendix B

Repository Information

The full source code used in this project is available at the following repository:

Bembi, N., & Bhatija, R. & Bandaru, A., (2025). SmartInvoice-AI [Source code]. GitHub. https://github.com/rbhatija/SmartInvoice-AI

This repository contains all scripts for data preprocessing, analysis, and visualization. Users can access the latest updates and documentation in the repository's README file.