ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction Summary

SROIE Competition

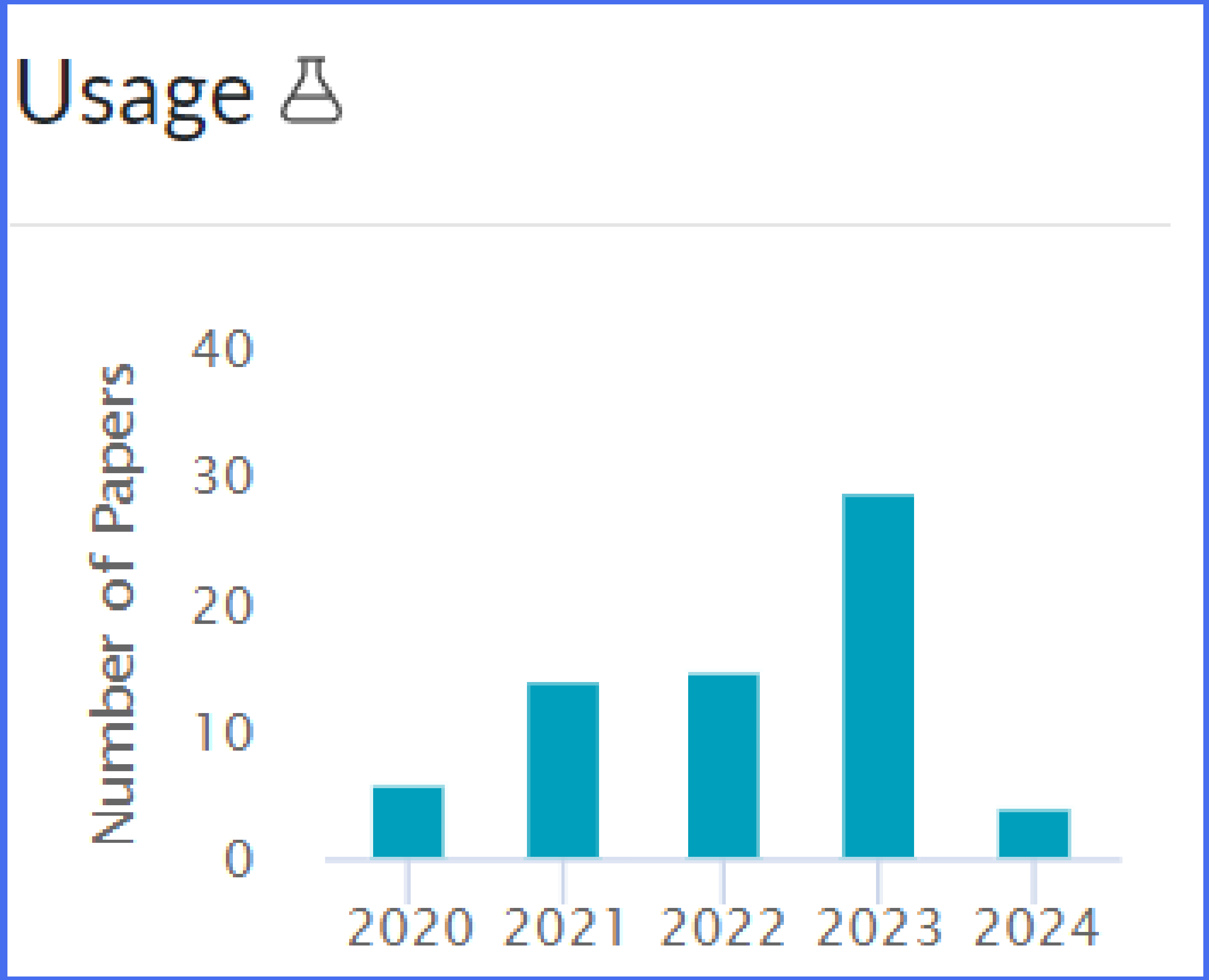
ICDAR 2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction

“Scanned receipts OCR and key information extraction” (SROIE) covers important aspects related to the automated analysis of scanned receipts. The SROIE tasks play a key role in many document analysis systems and hold significant commercial potential. Although a lot of works have been published over the years on administrative document analysis, the community has advanced relatively slowly, as most datasets have been kept private.”

The Dataset

One of the key contributions of SROIE to the document analysis community is to offer a first, **standardized dataset of 1000 whole scanned receipt images and annotations**, as well as an **evaluation procedure for such tasks.** For each receipt you have an **.jpg file of the scanned receipt**, a **.txt file holding OCR information** and a **.txt file holding the key information values.**

"the new dataset has some special features and challenges, e.g., some receipts having poor paper quality, poor ink and printing quality; low resolution scanner and scanning distortion; folded invoices; too many unneeded interfering texts in complex layouts; long texts and small font sizes. To address the potential privacy issue, some sensitive fields (such as name, address and contact number etc) of the receipts are blurred."



The Dataset

(a) (b) (c) (d) (e)

Fig. 2. Examples of scanned receipts for the competition tasks.

The competition is divided into 3 tasks:

1

Scanned Receipt Text Localisation:

Aims to accuratelylocalize textual content within scanned receipts using bounding boxes defined by four vertices.

2

Scanned Receipt OCR:

Aims to accurately recognize the text in a receipt image. No localisation information is provided, or is required.

3

Key Information Extraction from Scanned Receipts:

Aims to extract texts of a number of key fields from given receipts, and save the texts for each receipt image in a json file.

Task 1- Scanned Receipt Text Localisation:

Evaluation Protocol:

- **Mean average precision (mAP) :** computes the average precision value for detected text at different threshold levels of intersection over union (IoU) between the predicted and ground truth bounding boxes.
- **Average recall:** the proportion of actual text bounding boxes that were correctly identified by the model out of all ground truth boxes.

Top Performing Methods:

SCUT-DLVC-Lab

Employs a refinement-based **Mask-RCNN** approach.It iteratively **removes redundant information** to refine bounding box detection.

Ping An & Casualty:

Utilizes an **anchor-free detection** framework, employing FishNet as the backbone.
Prior to detection, images undergo preprocessing with O**penCV's adaptive threshold** to standardize scales.

H&H Lab:

Integrates the **EAST framework with a multi-oriented corner** detection ensemble for robust text detection.

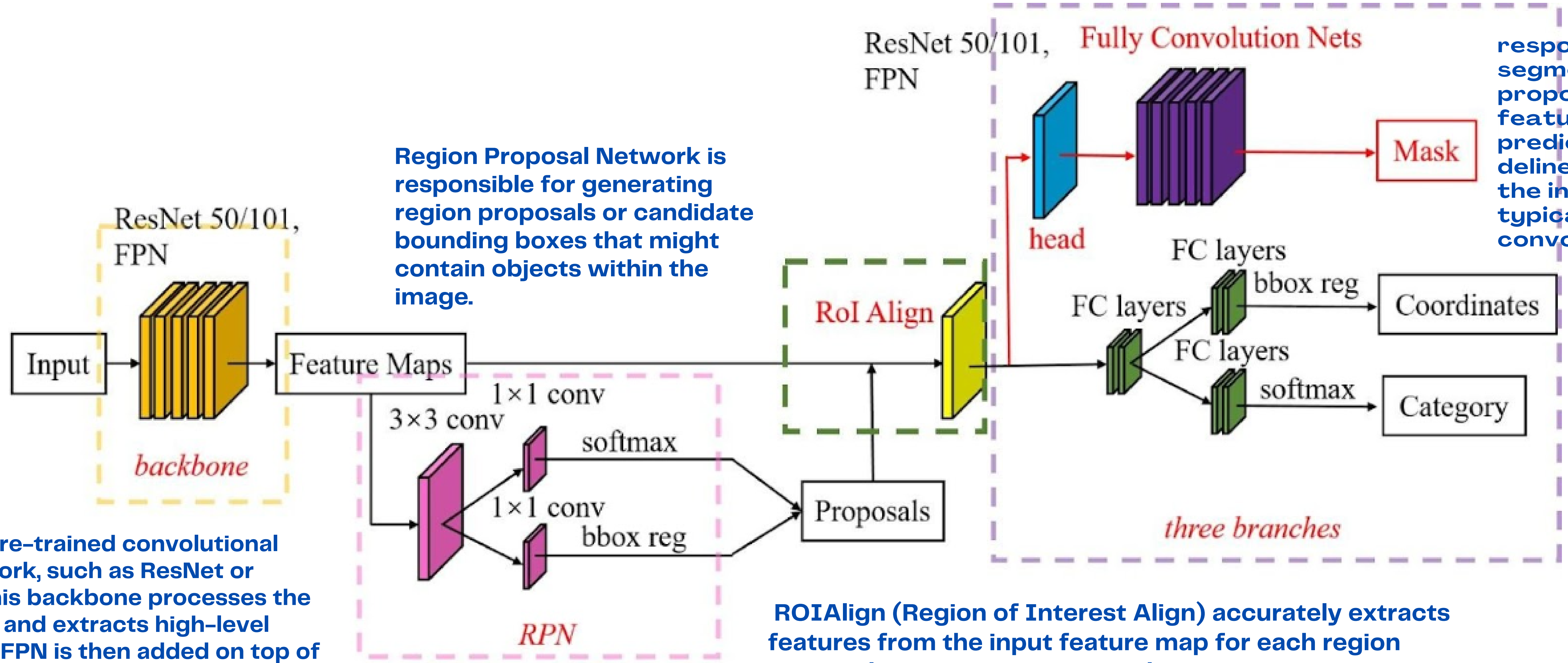
Task 1- Scanned Receipt Text Localisation:

TABLE I
TOP 10 METHODS FOR TASK 1 - SCANNED RECEIPT TEXT LOCALISATION.

Rank	Method	Recall	Precesion	Hmean
1	SCUT-DLVC-Lab-Refinement	98.64%	98.53%	98.59%
2	Ping An Property & Casualty Insurance Company	98.60%	98.40%	98.50%
3	H&H Lab	97.93%	97.95%	97.94%
4	GREAT-OCR -Shanghai University	96.62%	96.21%	96.42%
5	BOE_IOT_AIBD v5	95.95%	95.99%	95.97%
6	EM_ocr	95.85%	96.08%	95.97%
7	Clova OCR	96.04%	95.79%	95.92%
8	IFLYTEK-textDet_v3	93.77%	95.89%	94.81%
9	A Single-Shot Model for Robust Text Localization	93.93%	94.80%	94.37%
10	SituTech_OCR	93.81%	94.18%	94.00%

Mask R-CNN: Mask Region-based Convolutional Neural Network

A conceptually simple, flexible, and general framework for object instance segmentation. Mask RCNN detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance.
This method extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps.



Region Proposal Network is responsible for generating region proposals or candidate bounding boxes that might contain objects within the image.

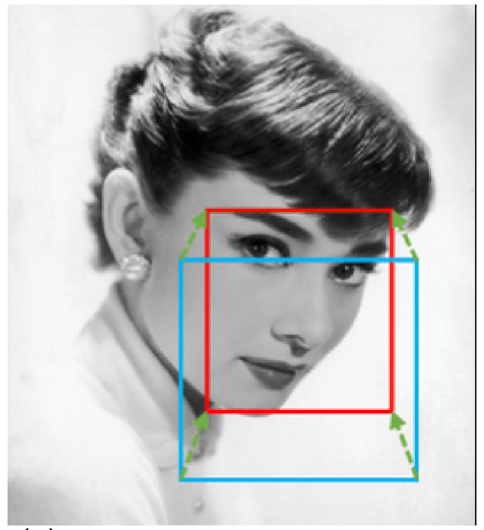
responsible for generating segmentation masks for each region proposal. The head uses the aligned features obtained through ROIAlign to predict a binary mask for each object, delineating the pixel-wise boundaries of the instances. The Mask Head is typically composed of several convolutional layers

typically a pre-trained convolutional neural network, such as ResNet or ResNeXt. This backbone processes the input image and extracts high-level features.An FPN is then added on top of this backbone network to create a feature pyramid.

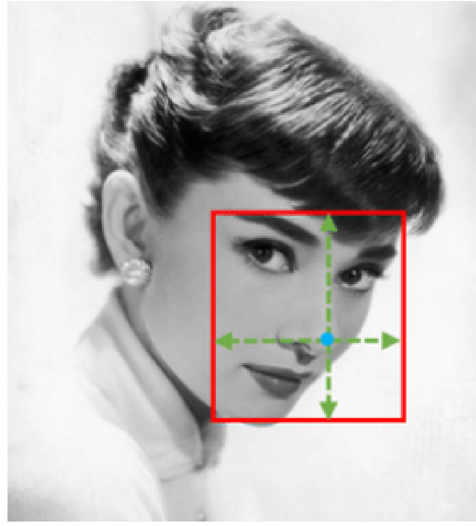
ROIAlign (Region of Interest Align) accurately extracts features from the input feature map for each region proposal, ensuring precise pixel-wise segmentation

Anchor free object detection & OpenCV's adaptive threshold :

Anchor-free object detection is a method in computer vision that locates and identifies objects in an image without relying on predefined anchor points. Anchors are preset boxes of various sizes and ratios that are used as references to detect objects at different scales and orientations. Anchor-free models, on the other hand, predict the presence and the dimensions of objects directly from the image data, making the process simpler and often more flexible because it doesn't need these predefined anchors.

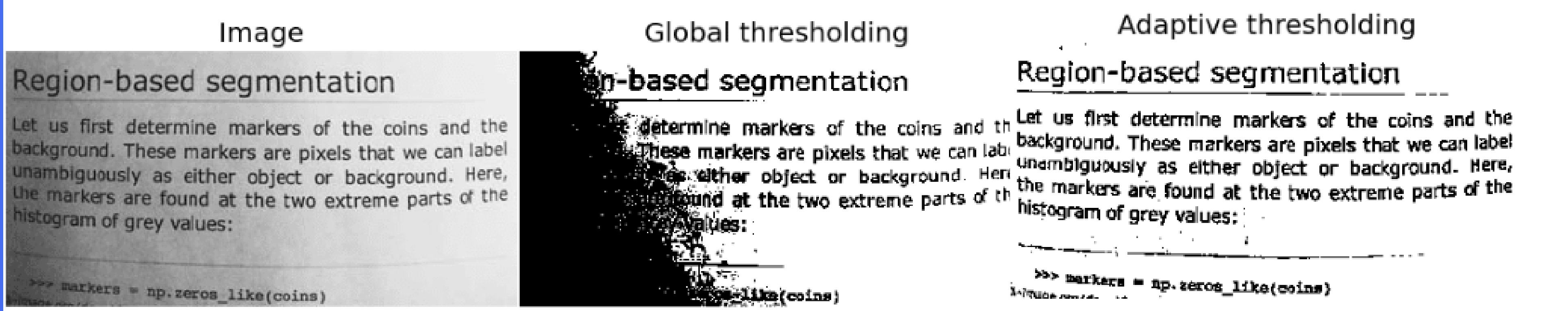


(a) The anchor based methods predict the offsets based on predefined anchor.



(b) The anchorfree methods directly estimate the offsets of a point to its outside boundaries.

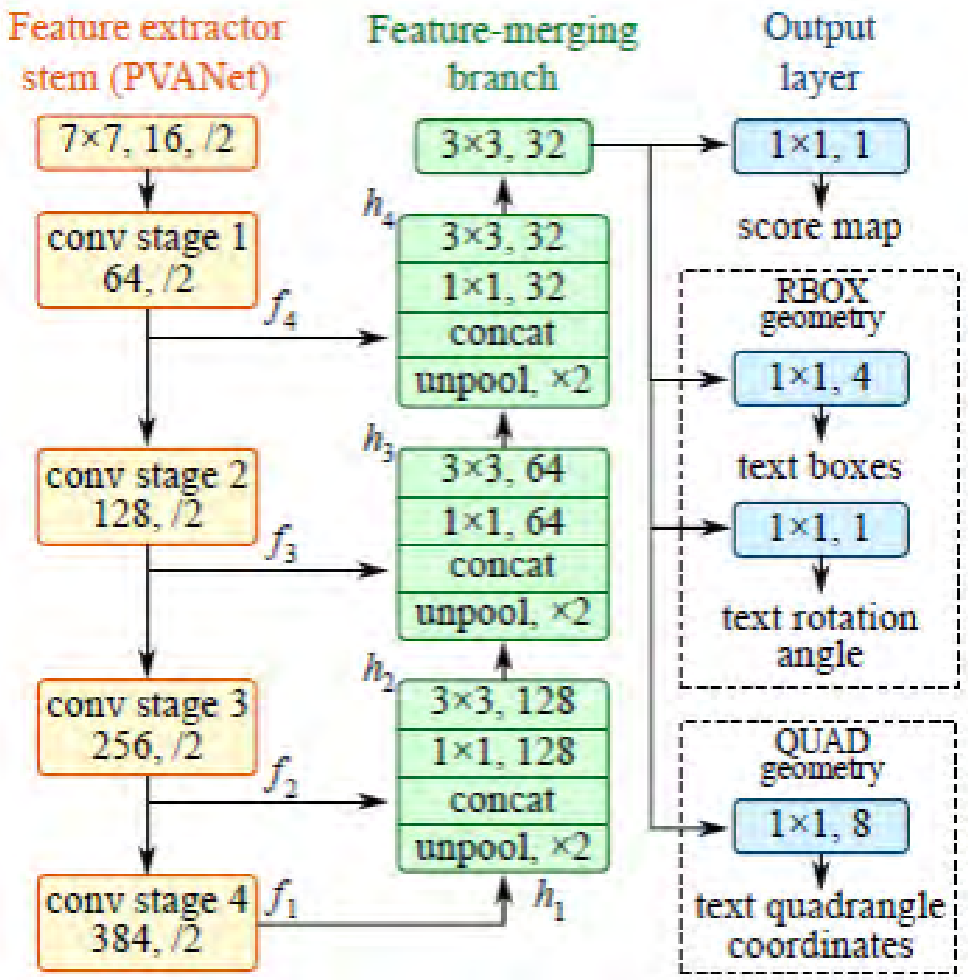
Adaptive thresholding is a **preprocessing technique** used in image processing to convert a grayscale image into a binary image, where the pixels are marked as either black or white.



- **Obtain better segmentation than using global thresholding methods, such as basic thresholding and Otsu thresholding**
- **Avoid the time consuming and computationally expensive process of training a dedicated Mask R-CNN or U-Net segmentation network**

EAST & multi oriented corner network

EAST was introduced in the paper "EAST: **An Efficient and Accurate Scene Text Detector**" . This algorithm is designed to tackle text detection in natural scenes, where **text can appear in various sizes, orientations, and perspectives.**



EAST employs a single neural network to generate predictions directly at the word or line level from entire images. This method significantly surpasses previous approaches by delivering superior accuracy and efficiency, ensuring rapid and precise text recognition. Additionally, it incorporates PVANET :Deep but Lightweight Neural Networks for Real-time Object Detection, which enhances its performance without compromising speed, **making it an ideal choice for applications requiring immediate text detection results.**

Model	Computation cost (MAC)				Running time		mAP
	Shared CNN	RPN	Classifier	Total	ms	x(PVANET)	(%)
PVANET+	7.9	1.3	27.7	37.0	46	1.0	82.5
Faster R-CNN + ResNet-101	80.5	N/A	219.6	300.1	2240	48.6	83.8
Faster R-CNN + VGG-16	183.2	5.5	27.7	216.4	110	2.4	75.9
R-FCN + ResNet-101	122.9	0	0	122.9	133	2.9	82.0

- **Robust to Text Variability**
- **Efficiency: can be deployed in real-time applications.**
- **Accuracy: achieves state-of-the-art performance in text detection tasks.**

Task 2- Scanned Receipt OCR:

Evaluation Protocol:

- **Matching Words: The words detected by the OCR system are matched against the ground truth.**
- **Precision and Recall: Precision is calculated as the number of correct matches over the number of detected words.**
- **F1 Score: The harmonic mean of precision and recall.**

Top Performing Methods:

H&H Lab:

They primarily used a **CRNN** architecture. The CNN structure within was modified **to resemble PVANet**, which is known for being lightweight and efficient, and they used **multiple Gated Recurrent Unit (GRU) layers** to better capture dependencies in the sequence data.

INTSIG-HeReceipt

Their method was grounded on **combining CNNs with RNNs,** leveraging the strengths of both in feature extraction and sequence modelling.

Model Ensemble: They trained multiple models with varied backbones and recurrent structures .

Ping An & Casualty:

Encoder-Decoder with Attention: This method is based on an encoder-decoder framework with an attention mechanism.

Data Synthesis: They synthesized 2 million text lines against the backgrounds of receipts, ranging from one to five words per line.

Task 2- Scanned Receipt OCR:

TABLE II

TOP 10 METHODS FOR TASK 2 - SCANNED RECEIPT OCR.

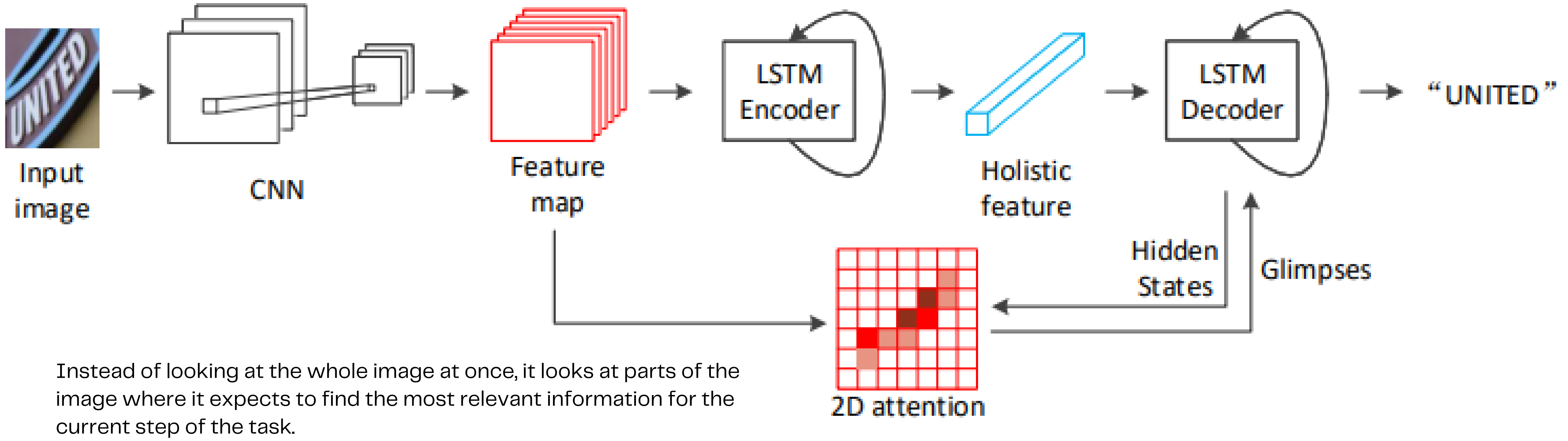
Rank	Method	Recall	Precesion	Hmean
1	H&H Lab	96.35%	96.52%	96.43%
2	INTSIG-HeReceipt-Ensemble	94.56%	95.10%	94.82%
3	Ping An Property & Casualty Insurance Company	94.48%	94.86%	94.67%
4	CLOVA OCR	94.30%	94.88%	94.59%
5	SCUT-DLVC-Lab-Lexicon	94.18%	94.88%	94.53%
6	DenseNet-Attention Recognition	94.29%	94.58%	94.44%
7	CITlab Argus Text Recognition	93.55%	93.61%	93.58%
8	Unet followed by CRNN with CTC	88.58%	87.30%	87.93%
9	BOE_IOT_AIBD T2 V5	87.84%	86.66%	87.24%
10	CRNN after UNet Segmentation	85.77%	86.48%	86.12%

The Decoder-Encoder architecture with attention

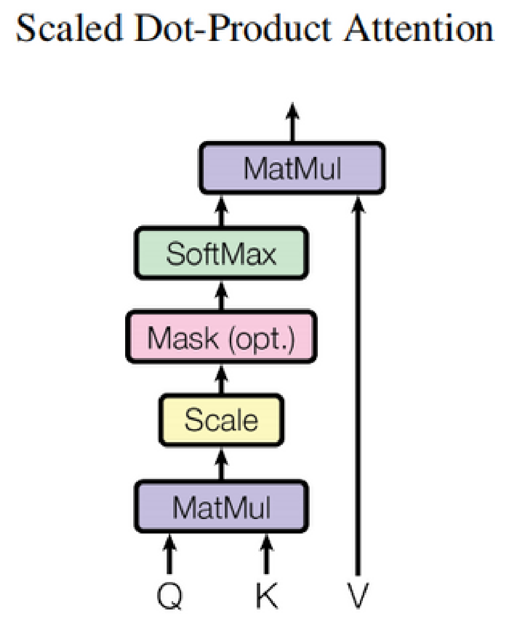
The model uses a 2D attention mechanism to focus on different parts of the feature map while decoding the sequence. Instead of looking at the whole image at once, it looks at parts of the image where it expects to find the most relevant information for the current step of the task.

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

The inputs to the attention function are queries (Q), keys (K), and values (V). These are all matrices where each row represents a word in the input sentence, and columns correspond to the features or dimensions representing those words. The raw scores are scaled down by a factor of the square root of the dimension of the key vectors to stabilize gradients during training. Then a softmax function is applied to the scaled scores to obtain the attention weights. This normalizes the scores so they are positive and add up to 1. The output is computed as the weighted sum of the value vectors, with weights being the softmax-normalized attention scores.

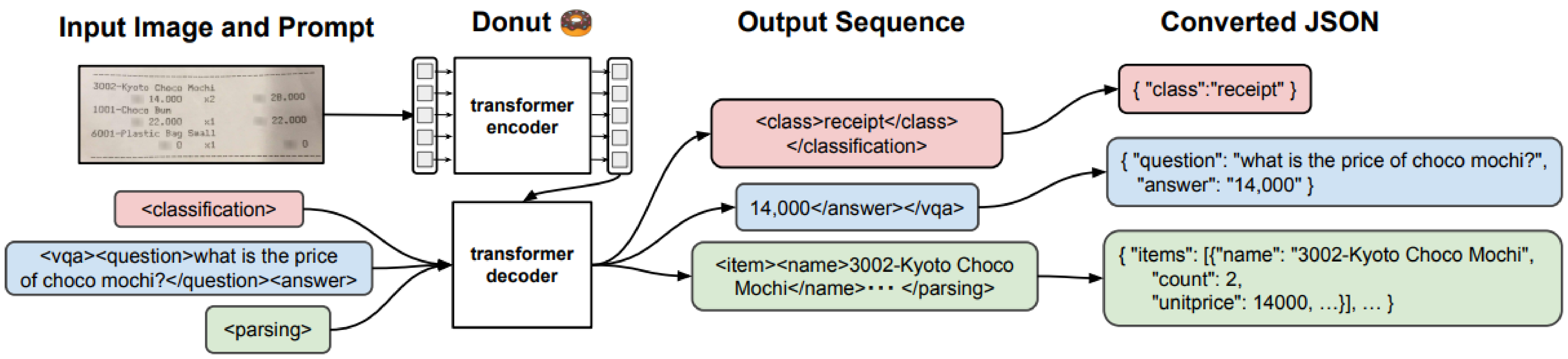


Instead of looking at the whole image at once, it looks at parts of the image where it expects to find the most relevant information for the current step of the task.



Donut Encoder Decoder (OCR Free)

The encoder takes a document image and splits it into patches, much like breaking up the image into smaller, manageable pieces. It then uses a **Swin Transformer**, which captures different aspects of the image, such as shapes and patterns, that are important for understanding the text and layout.



The decoder is based on BAR and uses the embeddings generated by the encoder to generate a sequence of tokens, which make up the words and numbers found in the document.

Task 3- Key Information Extraction from Scanned Receipts:

Evaluation Protocol:

- **mean Average Precision (mAP)**
- **Recall**
- **F1 score**

Top Performing Methods:

Ping An & Casualty:

utilized a **lexicon** built from the training dataset to autocorrect and refine their extraction results. **Regular expressions (RegEx)** were employed to identify and extract patterns corresponding to key fields.

Entity Detection:

The method integrates **content parsing** strategies with an **entity-aware detection system**, utilizing the **EAST.** Post detection, a text classifier with **RNN** embedding categorizes the extracted text into predefined classes

H&H Lab:

This method uses a hybrid neural network combining **Bidirectional Gated Recurrent Units (BiGRU)-CNNs and Conditional Random Fields (CRF)** to extract and classify text from receipt images.

Task 3- Key Information Extraction from Scanned Receipts:

TABLE III
TOP 10 METHODS FOR TASK 3 - KEY INFORMATION EXTRACTION FROM SCANNED RECEIPTS.

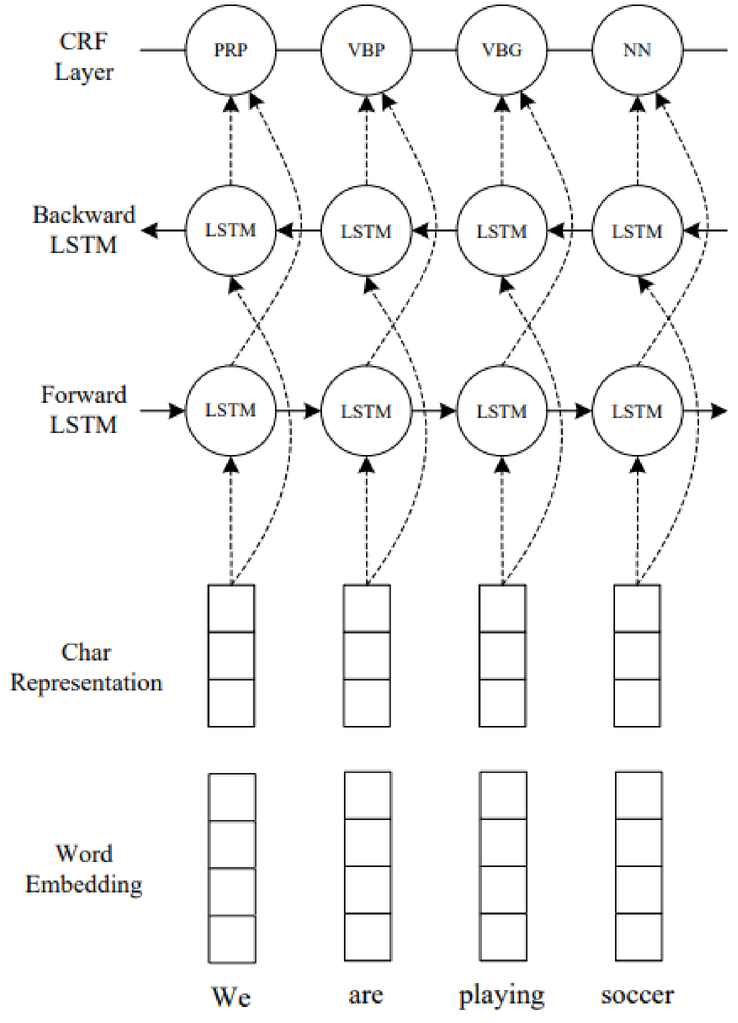
Rank	Method	Recall	Precesion	Hmean
1	Ping An Property & Casualty Insurance Company	90.49%	90.49%	90.49%
2	Enetity detection	89.70%	89.70%	89.70%
3	H&H Lab	89.63%	89.63%	89.63%
4	CLOVA OCR	89.05%	89.05%	89.05%
5	INTSIG-HeReceipt-withoutRM	83.00%	83.24%	83.12%
6	BOE_IOT_AIBD_v3	82.71%	82.71%	82.71%
7	PATECH_CHENGDU_OCR	81.70%	82.29%	82.00%
8	NER with spaCy model	78.96%	79.02%	78.99%
9	CITlab Argus Information Extraction (positional & line features, enhanced gt)	77.38%	77.38%	77.38%
10	A Simple Method for Key Information Extraction as Character-wise Classification with LSTM	75.58%	75.58%	75.58%

- **A lexicon** is essentially a dictionary or a database that the system uses to understand words or phrases. It contains words along with their meanings, usage, and related linguistic information. It can be used to correct misspellings or variations in text extracted from the receipts, improving the accuracy of the information extraction.
- **Pattern Recognition:** Regular expressions are used to identify patterns in text. For instance, dates, phone numbers, and amounts usually follow specific patterns that can be effectively captured using RegEx.

Combining a lexicon with RegEx allows the system to quickly identify and extract key information from unstructured text, making the process more efficient and reliable.

Combining BiGRU, CNNs, and CRFs:

- This combination allows for a robust model that can extract features from images (CNNs), understand the sequence and context of these features (BiGRU), and make accurate predictions about the nature of text blocks or sequences in the context of their neighboring elements (CRFs).

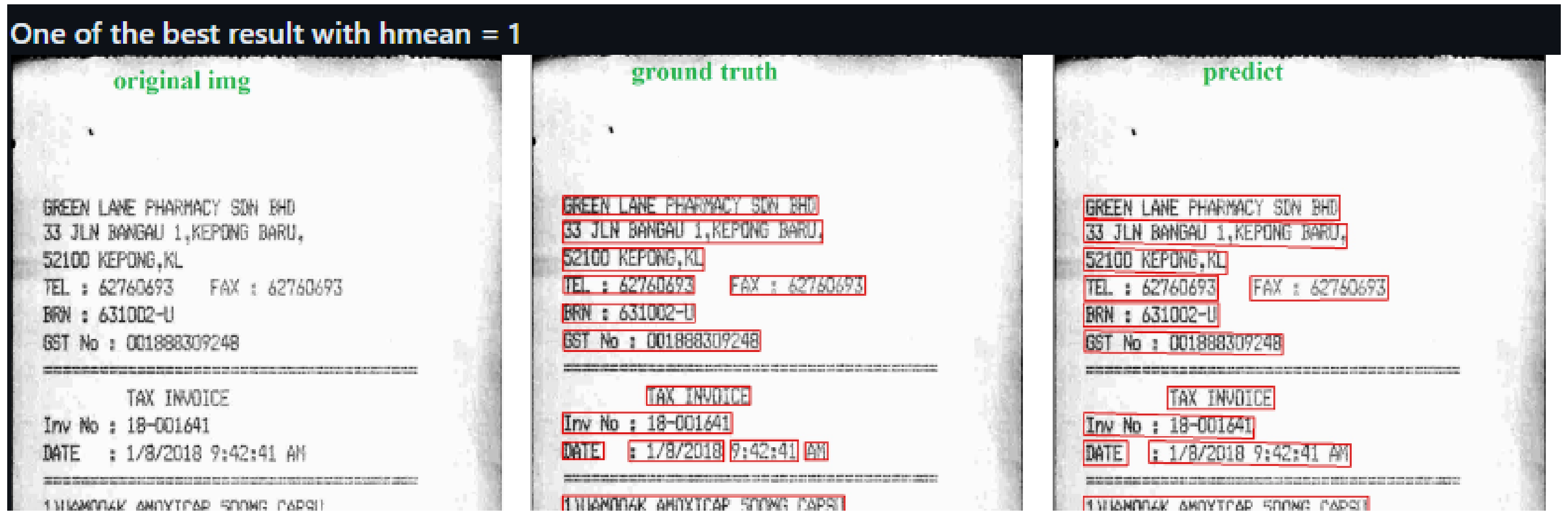


Available Code

ICDAR-2019-SROIE : CTPN (connectionist text proposal network)

Task	Recall	Precision	Hmean	Evaluation Method
Task 1	85.23%	88.73%	86.94%	Deteval
Task 2	26.33%	72.53%	38.63%	OCR
Task 3	75.58%	75.58%	75.58%	/

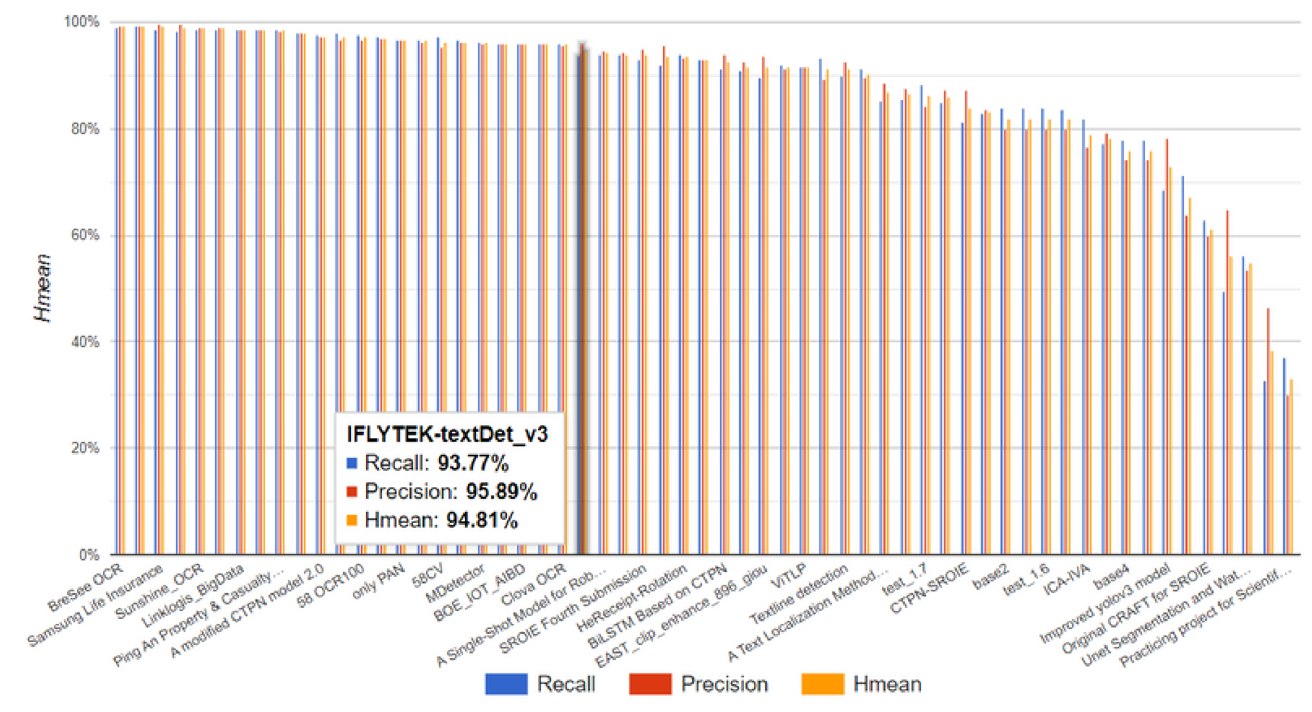
Scanned-Text-Receipt_Text-Localization: anchor-free text detectors - task 1



One of the top 3 competitors published two articles in CVPR 2017/2018 and provided their code:

EAST: An Efficient and Accurate Scene Text Detector
code
Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation
code

and much more!



THANK YOU ^^