TRIE: End-to-End Text Reading and Information Extraction for Document Understanding

Peng Zhang[1]
Yunlu Xu[1]
Zhanzhan Cheng[21]
Shiliang Pu[1]
Jing Lu[1]
Liang Qiao[1]
Yi Niu[1]
Fei Wu[2]

1. Hikvision Research Institute, Hangzhou, China
2. Zhejiang University, Hangzhou, China

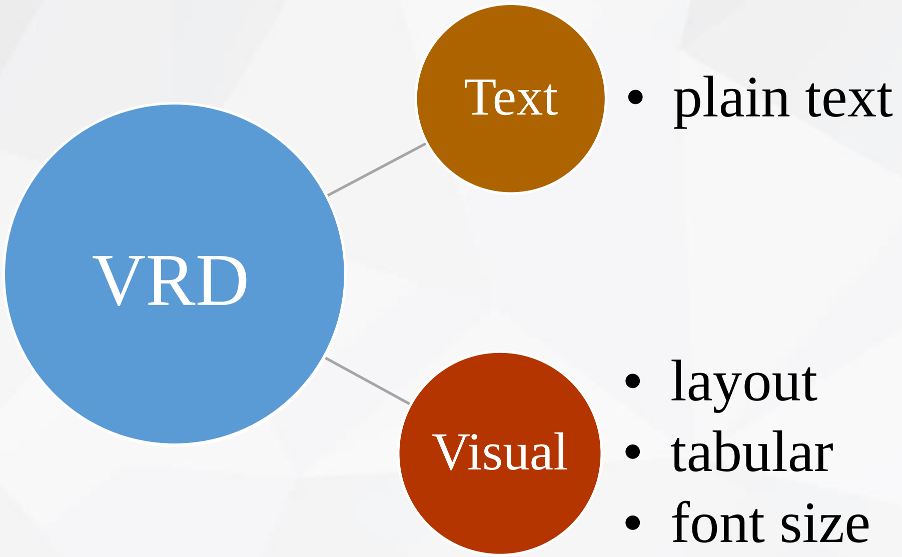
HIKVISION

ZHEJIANG UNIVERSITY 1897

ACM multimedia 2020 Seattle

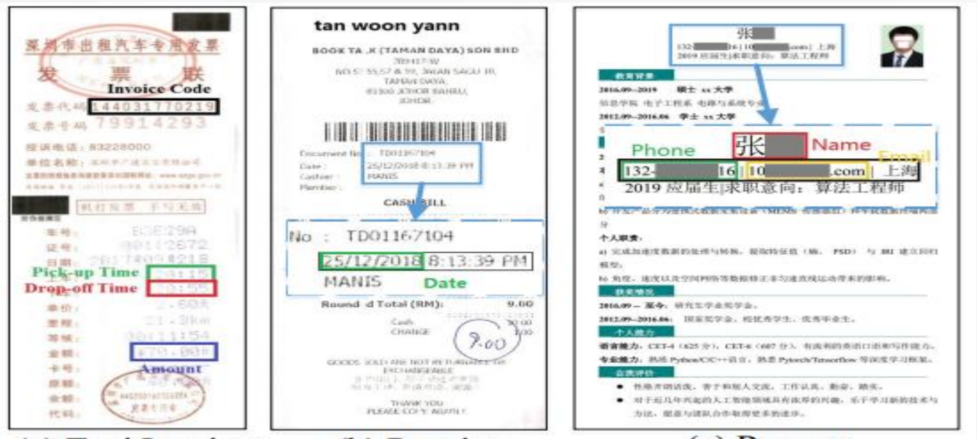
Background

◆ VRD (Visually Rich Document)



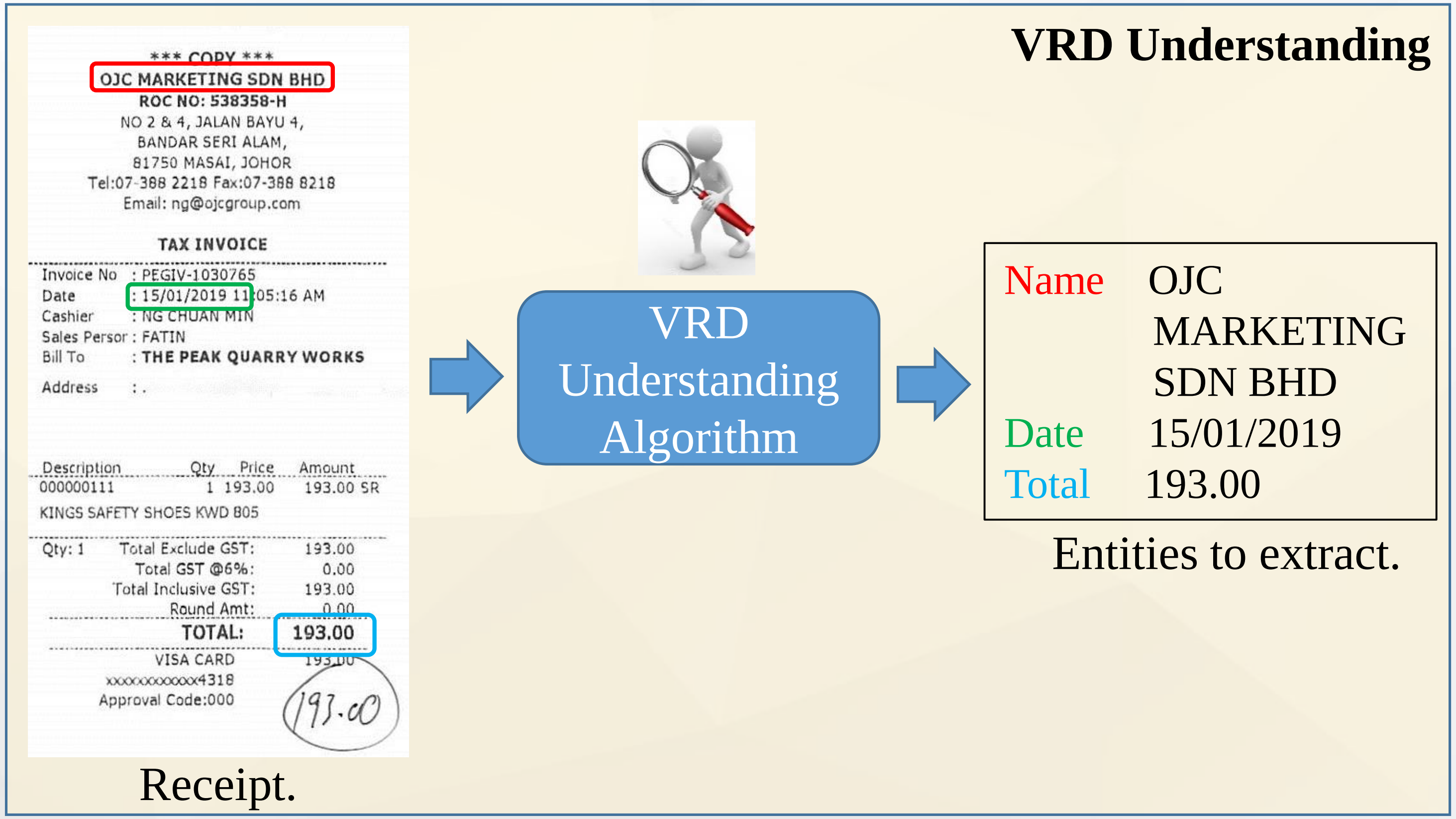
Text
- plain text

Visual
- layout
- tabular
- font size


Invoice Code
Pick-up Time
Drop-off Time
Amount
tan woon yann
No : TD01167104
25/12/2018 8:13:39 PM
MANIS
Date
Name
Phone
Email

(a) Taxi Invoices (b) Receipt (c) Resume

◆ VRD Understanding



VRD Understanding

*** COPY ***
OJC MARKETING SDN BHD
ROC NO: 538358-H
NO 2 & 4, JALAN BAYU 4,
BANDAR SERI ALAM,
81750 MASAI, JOHOR
Tel:07-388 2218 Fax:07-388 8218
Email: ng@ojcgroup.com

TAX INVOICE

Invoice No : PEGIV-1030765
Date : 15/01/2019 11:05:16 AM
Cashier : NG CHUAN MIN
Sales Persor : FATIN
Bill To : THE PEAK QUARRY WORKS
Address : .

Description	Qty	Price	Amount
000000111	1	193.00	193.00 SR
KINGS SAFETY SHOES KWD 805			

Qty: 1 Total Exclude GST: 193.00
Total GST @6%: 0.00
Total Inclusive GST: 193.00
Round Amt: 0.00
TOTAL: 193.00
VISA CARD 193.00
xxxxxxxxxxxx4318
Approval Code:000

193.00

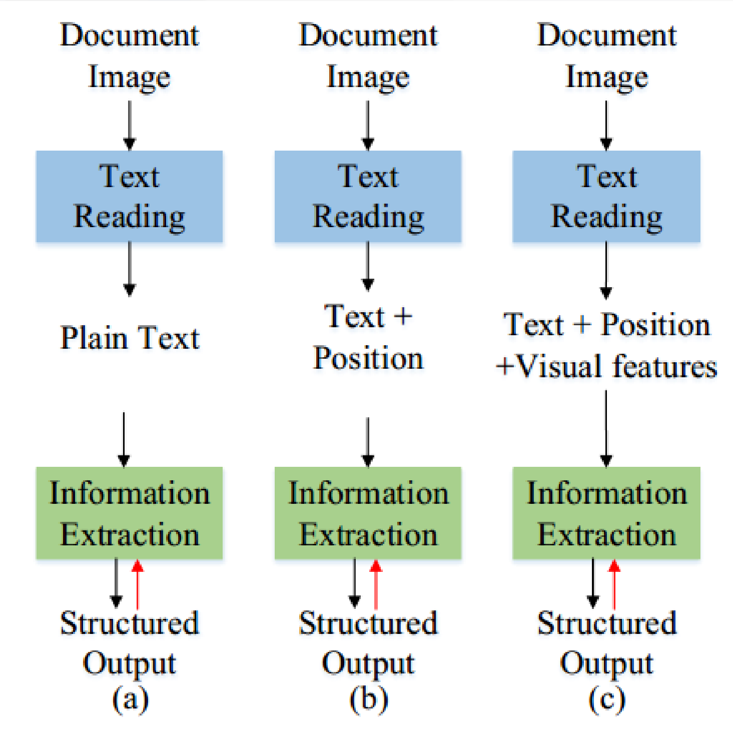
Receipt.

Name	OJC MARKETING SDN BHD
Date	15/01/2019
Total	193.00

Entities to extract.

Background

◆ Problem of current framework



Limitation 1: **Limitted visual features in IE.**

Keep:

$x_0,y_0...,x_3,y_3$, '15/01/2019'
$x_0,y_0...,x_3,y_3$, '193.00'

Lost:

Font, Color, Layout etc.

Limitation 2: **Ignoring relations between OCR & IE.**

◆ Motivation

Advantage 1: **Multimodal fusion in IE.**

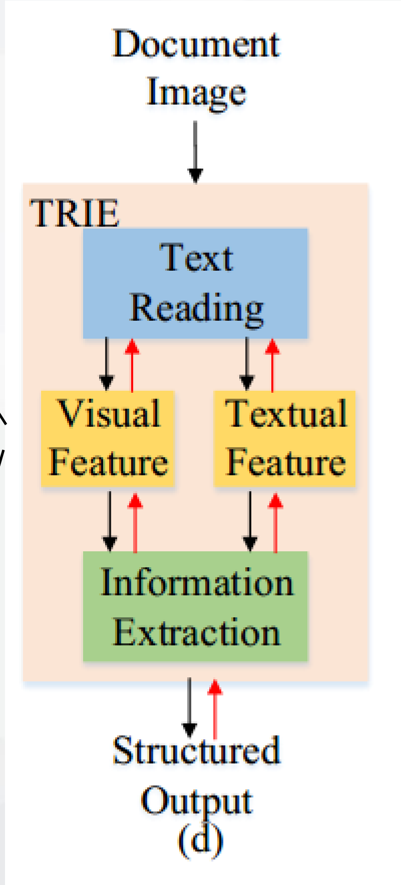
Keep:

$x_0,y_0...,x_3,y_3$, '15/01/2019'
$x_0,y_0...,x_3,y_3$, '193.00'

Font, Color, Layout etc.

Advantage 2: **Bridging OCR & IE,**

Forward: OCR boost IE
Backward: IE boost OCR



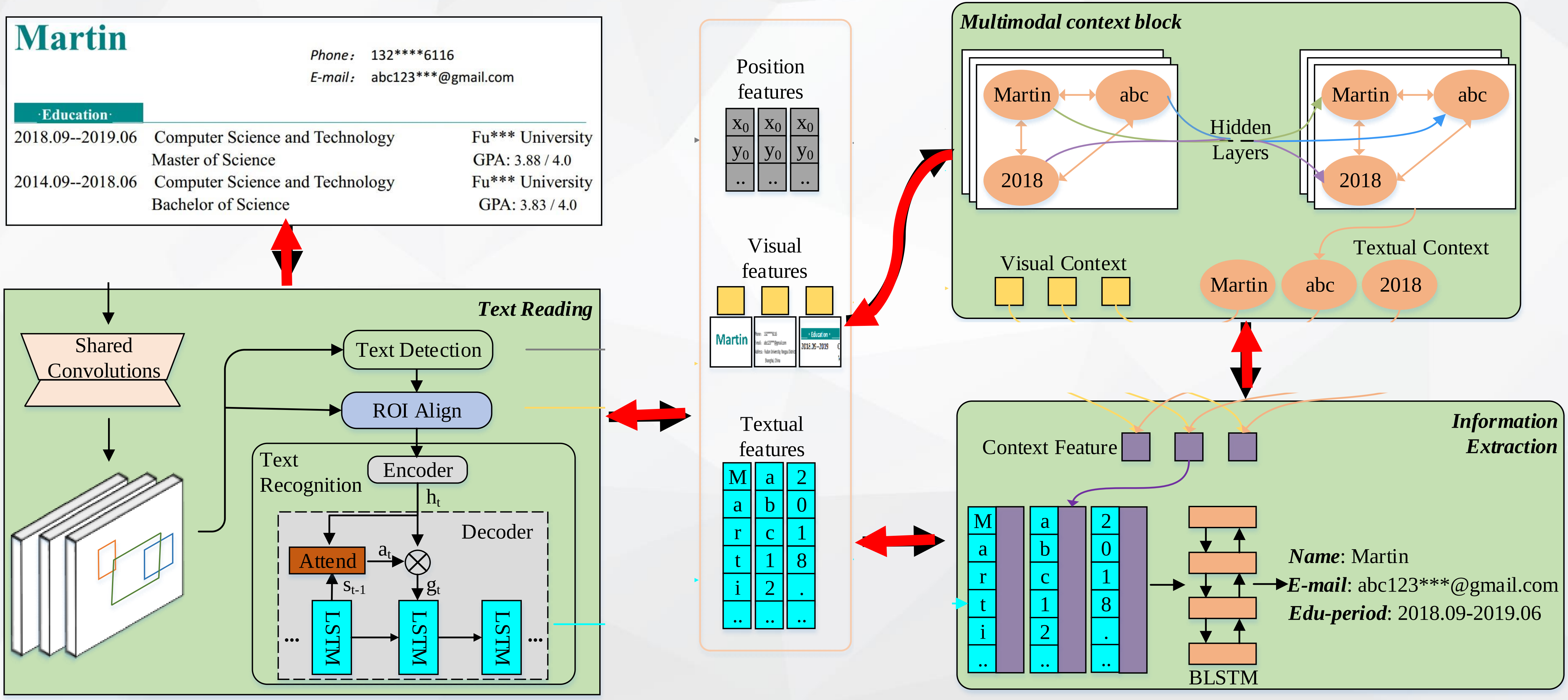
Method

HIKVISION

◆ Overall Architecture.

Forward Process

Backward Process



Martin

Phone： 132****6116

E-mail： abc123***@gmail.com

·Education·

2018.09--2019.06	Computer Science and Technology Master of Science	Fu*** University GPA: 3.88 / 4.0
2014.09--2018.06	Computer Science and Technology Bachelor of Science	Fu*** University GPA: 3.83 / 4.0

Name: Martin

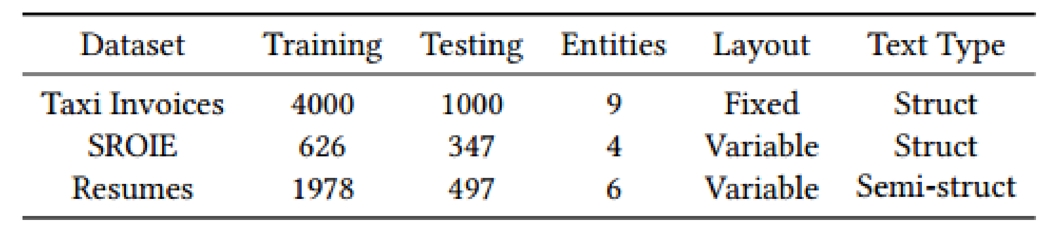
E-mail: abc123***@gmail.com

Edu-period: 2018.09-2019.06

Experiment

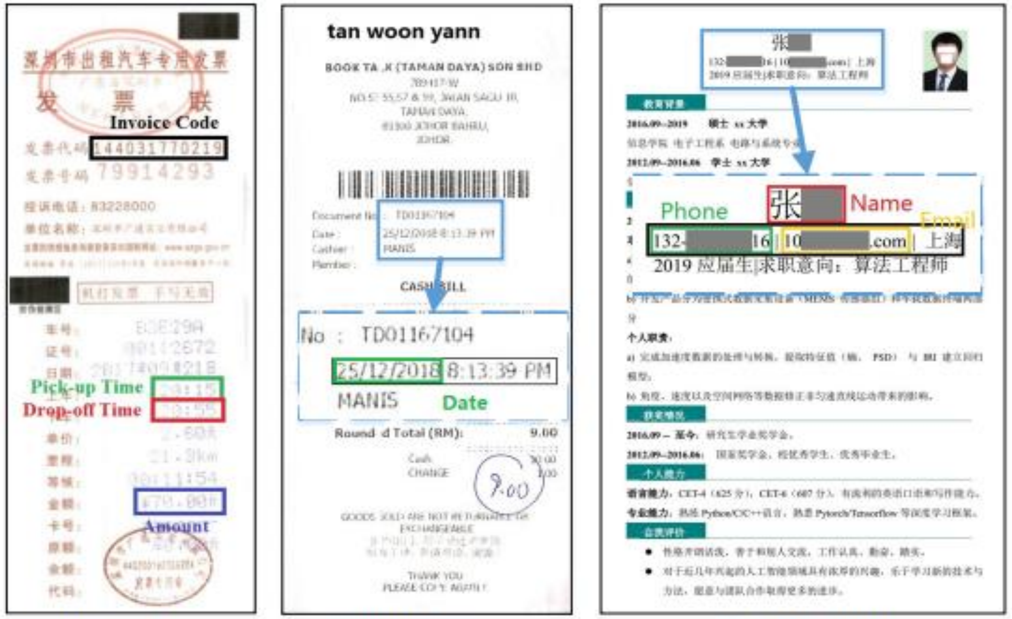
HIKVISION

◆ Datasets



Dataset	Training	Testing	Entities	Layout	Text Type
Taxi Invoices	4000	1000	9	Fixed	Struct
SROIE	626	347	4	Variable	Struct
Resumes	1978	497	6	Variable	Semi-struct

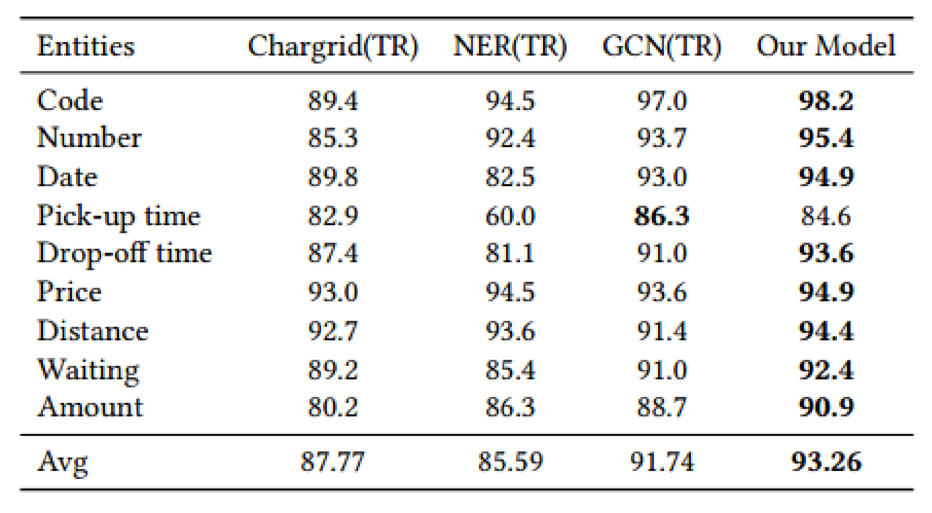
Dataset Statics


Invoice Code
Pick-up Time
Drop-off Time
Amount
tan woon yann
Date
Phone
Name
Email

(a) Taxi Invoices (b) Receipt (c) Resume

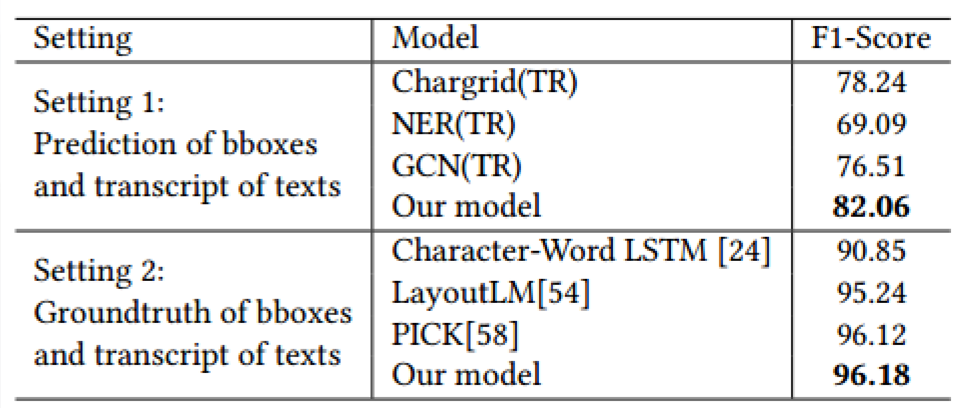
Dataset Examples

◆ Performance Summary



Entities	Chargrid(TR)	NER(TR)	GCN(TR)	Our Model
Code	89.4	94.5	97.0	**98.2**
Number	85.3	92.4	93.7	**95.4**
Date	89.8	82.5	93.0	**94.9**
Pick-up time	82.9	60.0	**86.3**	84.6
Drop-off time	87.4	81.1	91.0	**93.6**
Price	93.0	94.5	93.6	**94.9**
Distance	92.7	93.6	91.4	**94.4**
Waiting	89.2	85.4	91.0	**92.4**
Amount	80.2	86.3	88.7	**90.9**
Avg	87.77	85.59	91.74	**93.26**

Taxi Invoices Dataset



Setting	Model	F1-Score

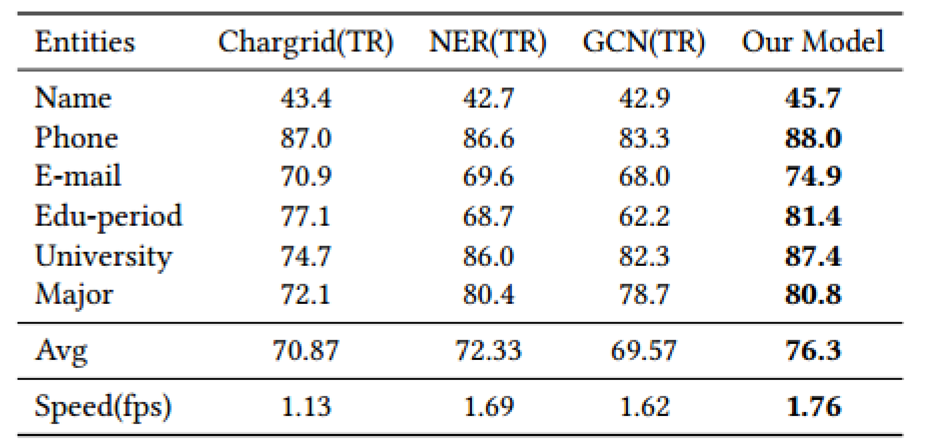
Setting 1:
Prediction of bboxes and transcript of texts
	Chargrid(TR)
	NER(TR)
	GCN(TR)
	Our model
		78.24
		69.09
		76.51
		82.06

Setting 2: Groundtruth of bboxes and transcript of texts	Character-Word LSTM [24] LayoutLM[54] PICK[58] Our model	90.85 95.24 96.12 **96.18**

ICDAR2019 SROIE Dataset

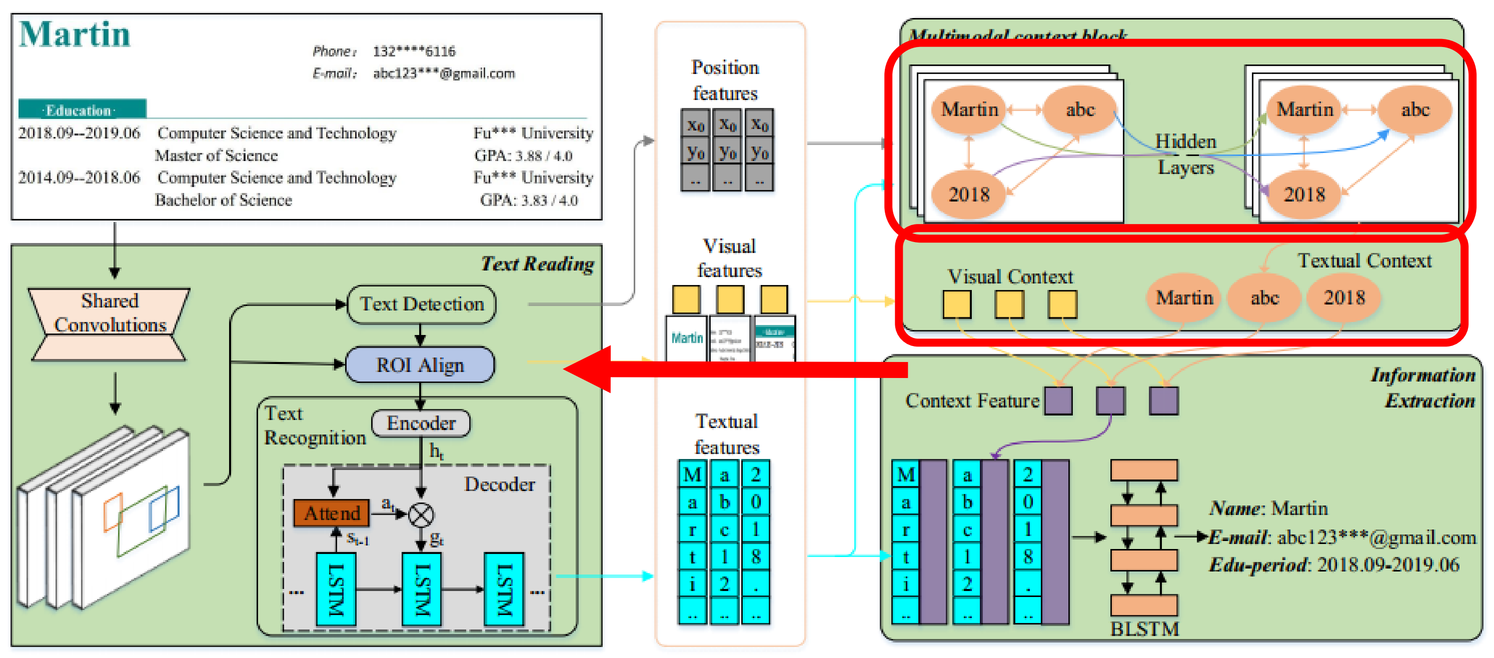
Experiment

◆ Performance Summary



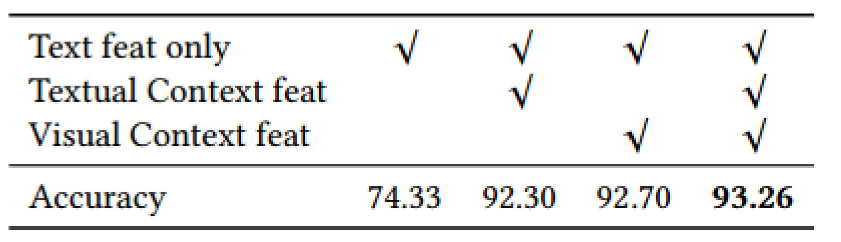
Entities	Chargrid(TR)	NER(TR)	GCN(TR)	Our Model
Name	43.4	42.7	42.9	**45.7**
Phone	87.0	86.6	83.3	**88.0**
E-mail	70.9	69.6	68.0	**74.9**
Edu-period	77.1	68.7	62.2	**81.4**
University	74.7	86.0	82.3	**87.4**
Major	72.1	80.4	78.7	**80.8**
Avg	70.87	72.33	69.57	**76.3**
Speed(fps)	1.13	1.69	1.62	**1.76**

Resumes Dataset



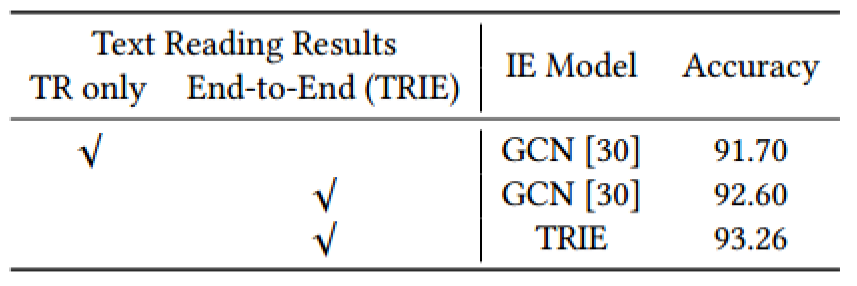
Algorithm Architecture.

◆ Discussion



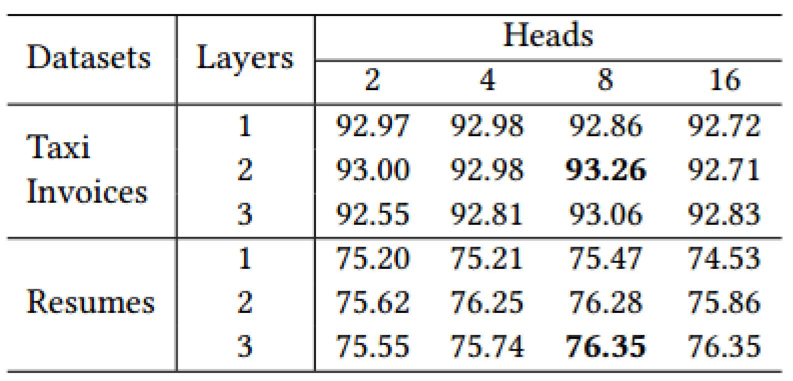
Text feat only	√	√	√	√
Textual Context feat		√		√
Visual Context feat			√	√
Accuracy	74.33	92.30	92.70	**93.26**

Effects of multimodal features on IE.



Text Reading Results		IE Model	Accuracy
TR only	End-to-End (TRIE)		
√		GCN [30]	91.70
	√	GCN [30]	92.60
	√	TRIE	93.26

Effects of E2E framework on text reading.



Datasets	Layers	Heads			
		2	4	8	16
Taxi Invoices	1	92.97	92.98	92.86	92.72
	2	93.00	92.98	**93.26**	92.71
	3	92.55	92.81	93.06	92.83
Resumes	1	75.20	75.21	75.47	74.53
	2	75.62	76.25	76.28	75.86
	3	75.55	75.74	**76.35**	76.35

Effects of layers and heads in textual context block.

HIKVISION

见远 行更远

See Far, Go Further