

Receipt OCR solutions for document automation

문서 자동화를 위한 영수증 OCR 솔루션

Presented by 202100509 김민서

Problem Definition



SCANNING

법인카드 사용내역서				
작성자	과장	부장	대표이사	
구 분				
이 름	김○○	부 서 명	마케팅부	
직 급	대리	사용연월	2000년 ○○월분	
법인카드 정보				
카드번호	1234-1234-1234-1234			
카드회사	A카드사			
사용자명	김○○			
법인카드 사용내역				
사용일자	사용처	금 액	사용목적	첨부서류
2000-00-00	A컨퍼런스	300,000	마케팅 세미나 참가비	첨부파일1
2000-00-00	B식당	150,000	팀 회식	
2000-00-00	온라인쇼핑몰	100,000	프로모션 물품 구입비	
2000-00-00	C호텔	200,000	출장(숙박비)	첨부파일2
합계 : 750,000				
** 0629-01-000016 CASHIER: 김경희 **				

TYPING



**Accounting
Management
System ...**

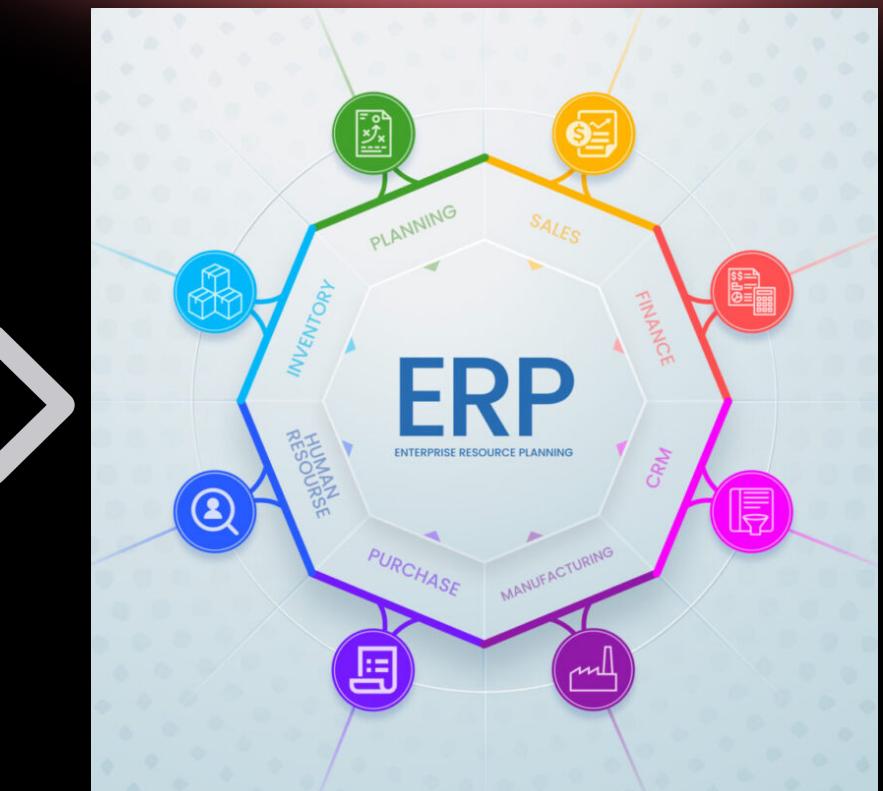
Problem Definition



법인카드 사용내역서				
작성자	과장	부장	대표이사	
구 분				
이 름	김○○	부 서 명	마케팅부	
직 급	대리	사용연월	2000년 ○○월분	
법인카드 정보				
카드번호	1234-1234-1234-1234			
카드회사	A카드사			
사용자명	김○○			
법인카드 사용내역				
사용일자	사용처	금 액	사용목적	첨부서류
2000-00-00	A컨퍼런스	300,000	마케팅 세미나 참가비	첨부파일1
2000-00-00	B식당	150,000	팀 회식	
2000-00-00	온라인쇼핑몰	100,000	프로모션 물품 구입비	
2000-00-00	C호텔	200,000	출장(숙박비)	첨부파일2

**시간과 비용 문제
오류 발생 가능성
비효율적인 데이터 관리**

합 계 750,000



**Accounting
Management
System ...**

Problem Definition



작성자	과장	부장	대표이사	
법인카드 사용내역서				
구 분				
이 름	김○○	부 서 명	마케팅부	
직 급	대리	사용연월	2000년 ○○월분	
법인카드 정보				
카드번호	1234-1234-1234-1234			
카드회사	A카드사			
사용자명	김○○			
법인카드 사용내역				
사용일자	사용처	금 액	사용목적	첨부서류
2000-00-00	A컨퍼런스	300,000	마케팅 세미나 참가비	첨부파일1
2000-00-00	B식당	150,000	팀 회식	
2000-00-00	온라인쇼핑몰	100,000	프로모션 물품 구입비	
2000-00-00	C호텔	200,000	출장(숙박비)	첨부파일2
결제수단별 결제내역				
01. 신용카드 :	125,500			
신용카드 :	125,500			
[카드 번호] 4579-7303-XXXX-8073				
[할부 개월] 일시불				
[카드 사용처] KB국민카드				
[매입사명] KB국민카드				
[가맹점 번호] 00056769491				
[승인번호] 30031535				
[부가세액] 11,408				
[결제금액] 125,500 원				
[발행일자] 2000-01-01				
TYPING				
합계	750,000			

Simplification?



SCANNING

TYPING

OCR (Optical Character Recognition)



Text Detection



RIVERSIDE

"RIVERSIDE"

WALK

"WALK"

Text Recognition

Methodology

1. OCR Model + NLP

OCR모델로 전체 텍스트 추출 + NLP 모델로 주요 정보만 추출

re-MART			
[주문] 2024-09-12 18:59			
상품명	단가	수량	금액
망고주스	3,500	2	7,000
크림파스타	6,000	1	6,000
닭한마리쌀국수	9,000	1	9,000
과세물품			19,800
부가세			2,200
합계			22,000
결제금액			22,000
0010 KB국민	5409***001*23 / 30112340		
카드결제		일시불	/ 22,000
주문번호 : 7372			

re-MART
[주문] 2024-09-12 18:59
상품명 단가 수량 금액
망고주스 3,500 27,000
크림파스타 6,000 16,000
과세물품 19,800
부가세 2,200
합계 22,000
결제금액 22,000
0010 KB국민 5409***001*23 / 30112340
카드결제 일시불 / 22,000
주문번호: 7372

가맹점명	Re-MART
날짜/시간	2024-09-12 18:59
품목	망고주스, 크림파스타, 닭한마리쌀국수
합계금액	22,000

Methodology

2. Detection + Recognition

먼저 주요 정보 영역 검출 + 텍스트 인식 모델 결합

Detection 모델 학습

re-MART			
[주 문]2024-09-12 18:59			
상품명	단가	수량	금액
망고주스	3,500	2	7,000
크림파스타	6,000	1	6,000
닭한마리쌀국수	9,000	1	9,000
과세물품			19,800
부가세			2,200
합계			22,000
결제금액			22,000
0010 KB국민	5409***001*23 / 30112340		
카드결제		일시불	/ 22,000
주문번호 : 7372			



re-MART			
[주 문]2024-09-12 18:59			
상품명	단가	수량	금액
망고주스	3,500	2	7,000
크림파스타	6,000	1	6,000
닭한마리쌀국수	9,000	1	9,000
과세물품			19,800
부가세			2,200
합계			22,000
결제금액			22,000
0010 KB국민	5409***001*23 / 30112340		
카드결제		일시불	/ 22,000
주문번호 : 7372			



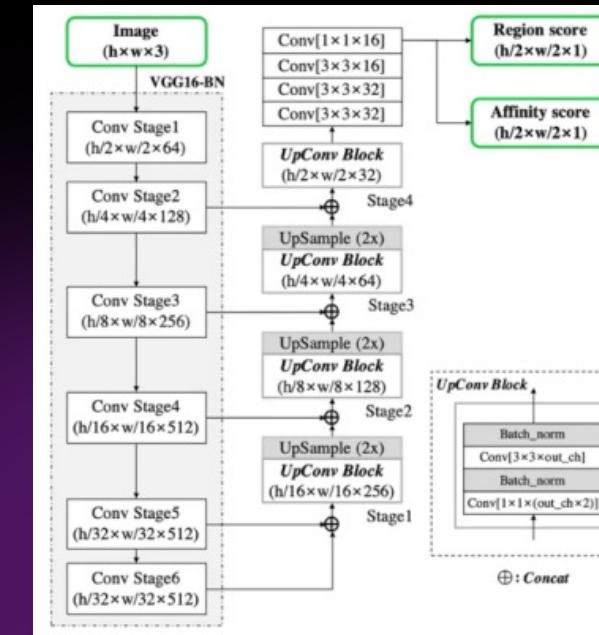
가맹점명	Re-MART
날짜/시간	2024-09-12 18:59
품목	망고주스, 크림파스타, 닭한마리쌀국수
합계금액	22,000

Detection Model



based VGG16
Region score
Affinity score
...

CRAFT (Detection) + CRNN (Recognition)



Character Region Awareness for Text Detection

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee*
Clova AI Research, NAVER Corp.

{youngmin.baek, bado.lee, dongyoon.han, sangdoo.yun, hwalsuk.lee}@navercorp.com

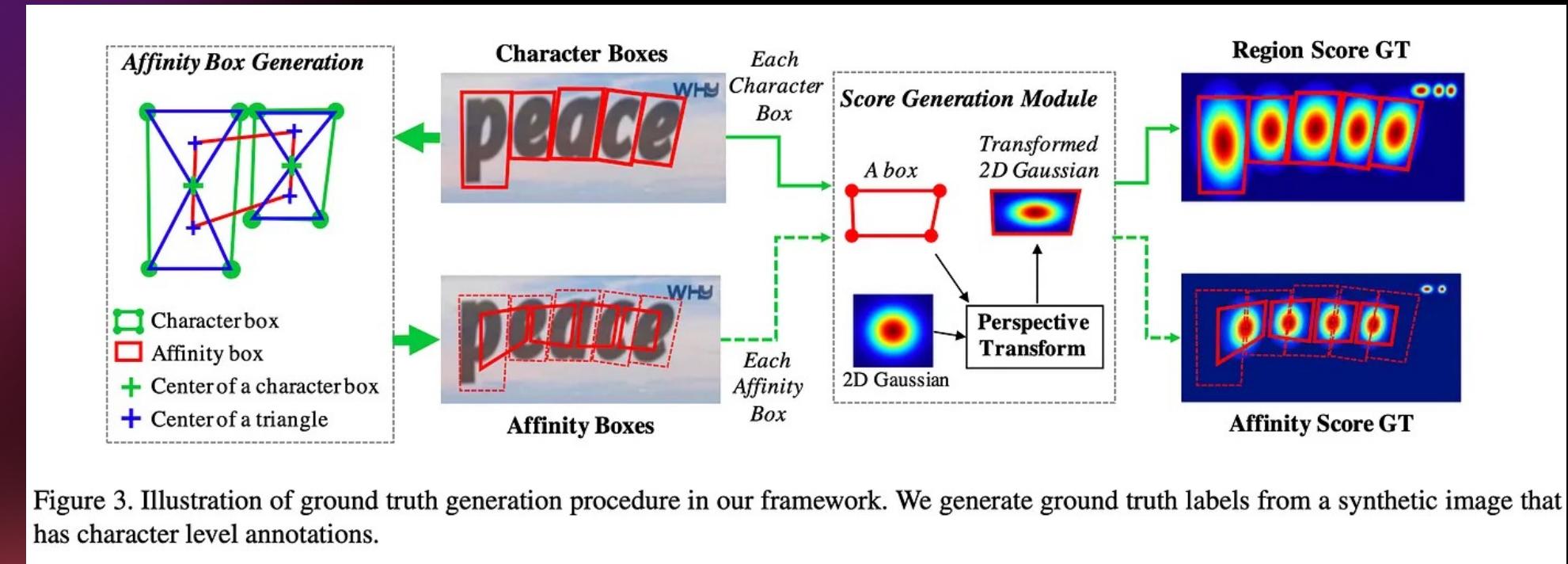
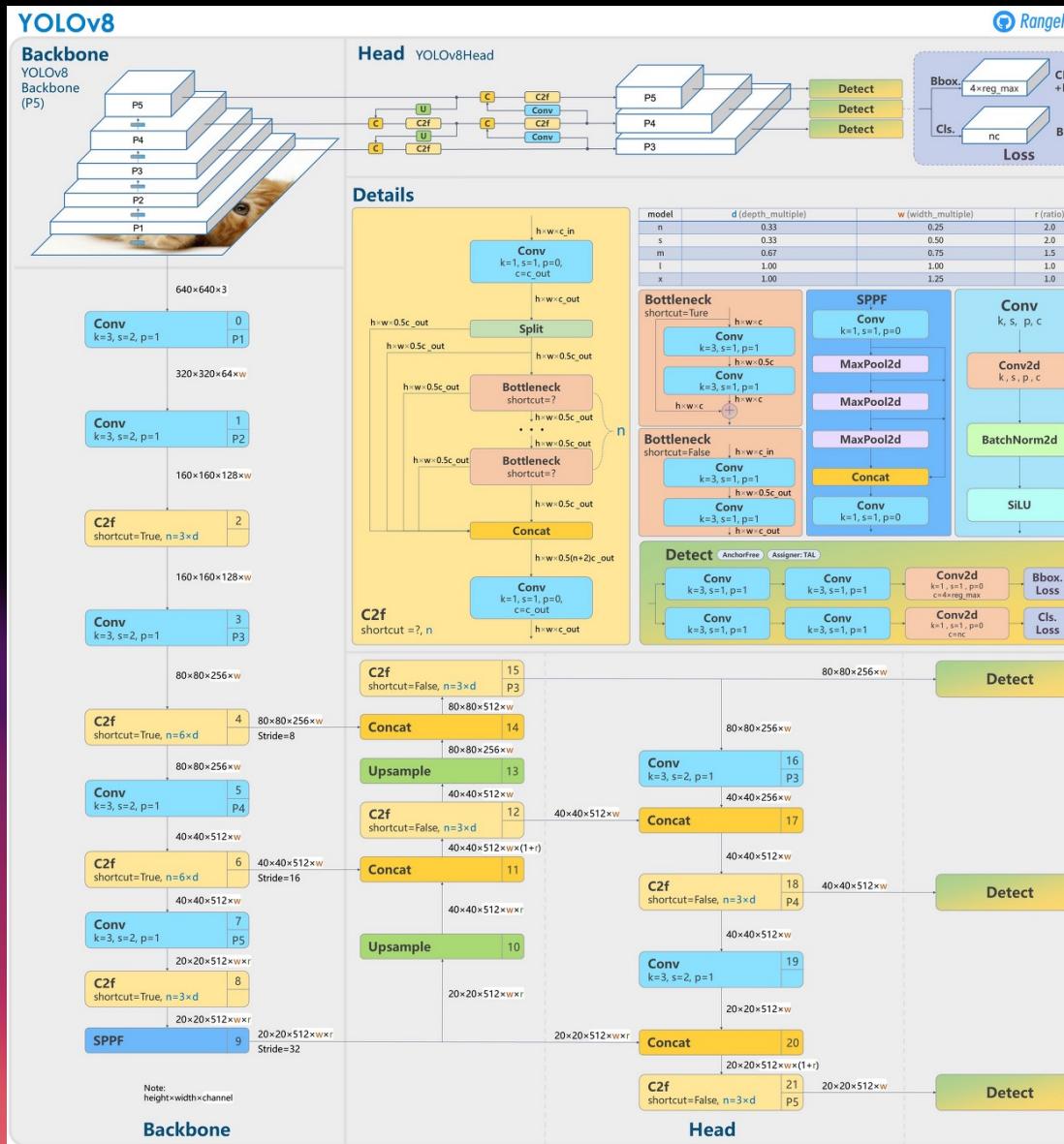


Figure 3. Illustration of ground truth generation procedure in our framework. We generate ground truth labels from a synthetic image that has character level annotations.

Detection Model

YOLOv8 : Object Detection Model

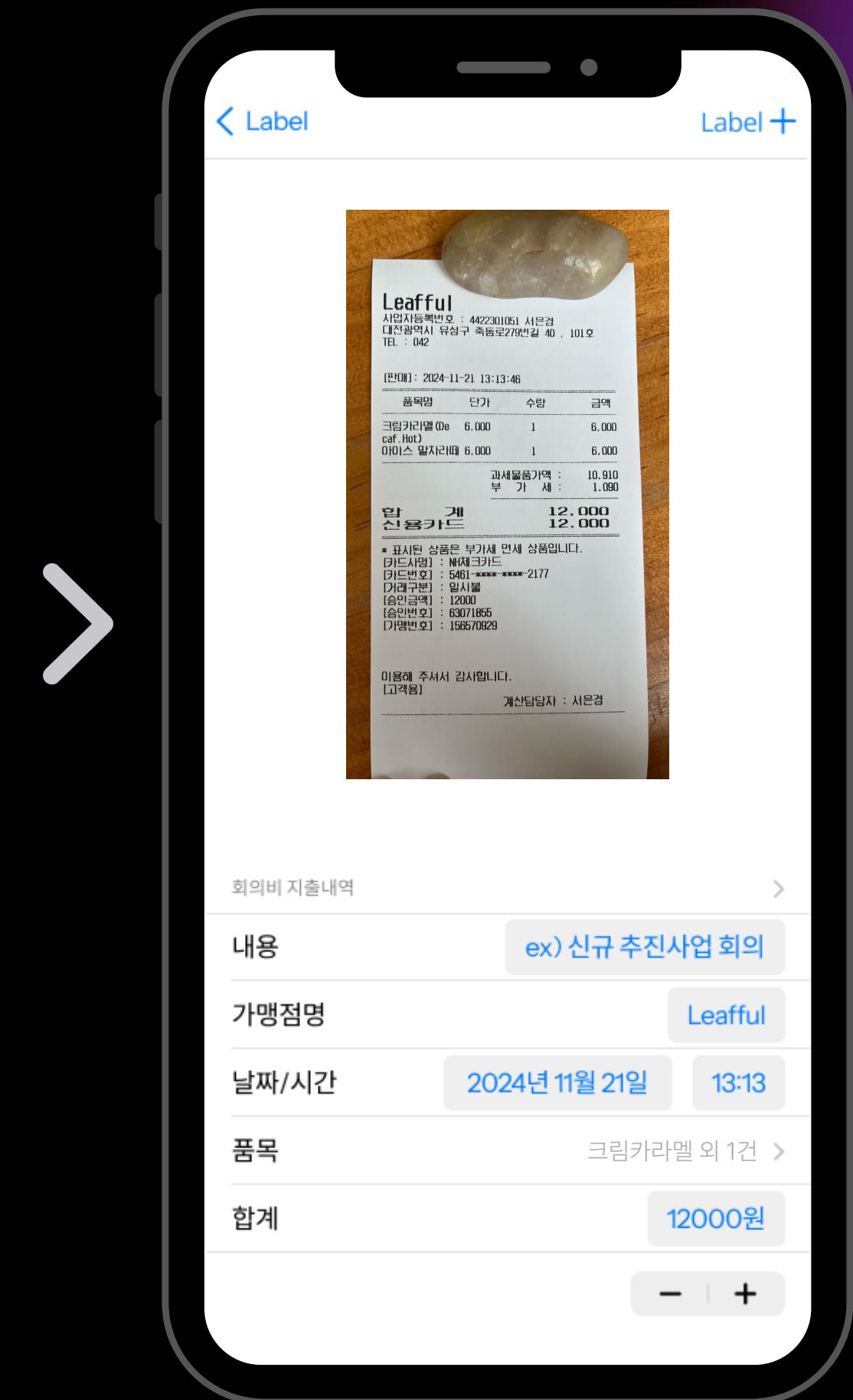
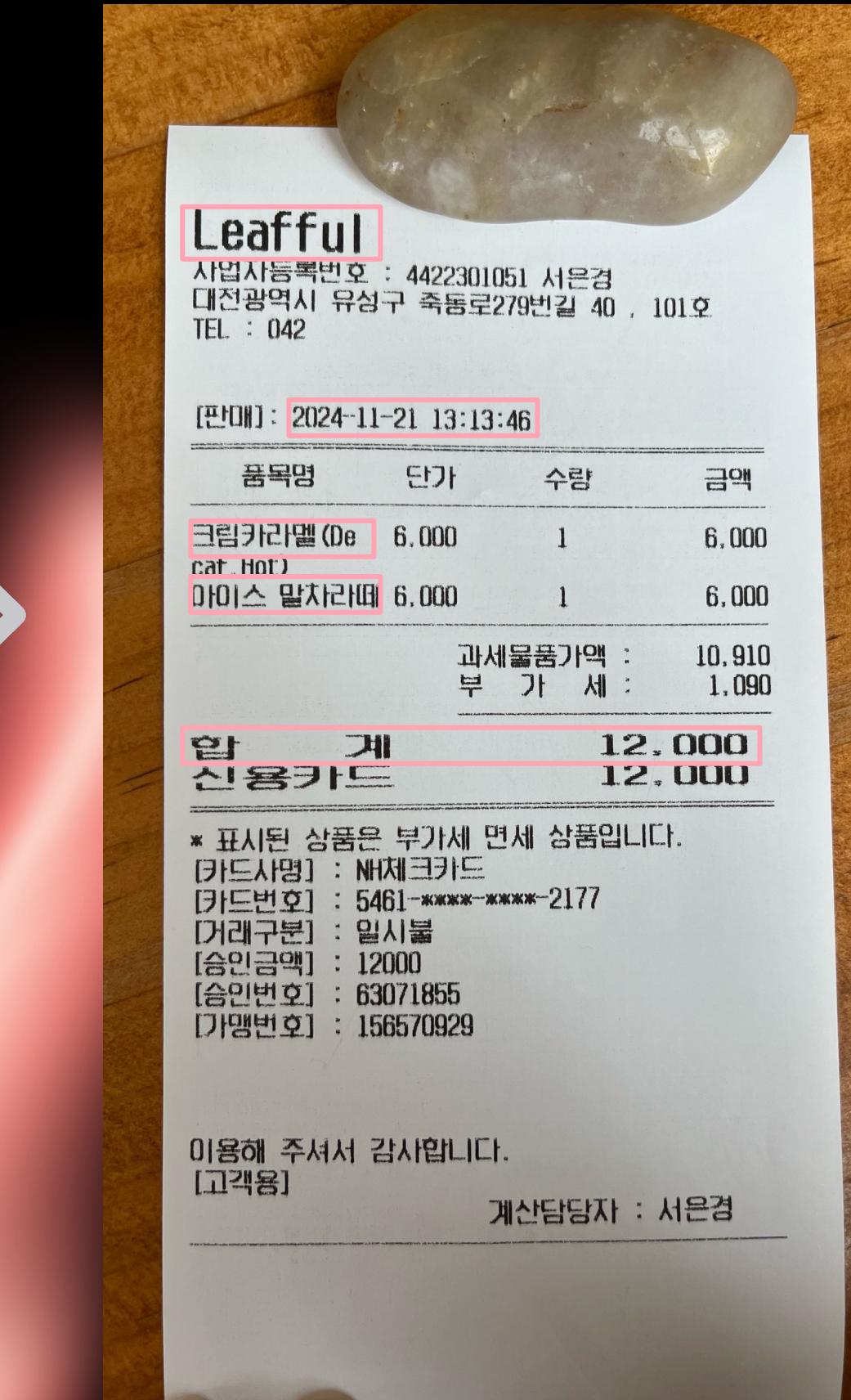


Layout
Detection..??



Field Extraction





CSV,
excel,

...

Dataset

 Sign In Register

TRAINING DATA · UPDATED A YEAR AGO ▲ 52 New Notebook Download ...

OCR Receipts Text Detection - retail dataset

Photos of the receipts and text detection - ocr dataset

Data Card Code (11) Discussion (0) Suggestions (0)

About Dataset

OCR Receipts from Grocery Stores Text Detection - retail dataset

The Grocery Store Receipts Dataset is a collection of photos captured from various **grocery store receipts**. This dataset is specifically designed for tasks related to **Optical Character Recognition (OCR)** and is useful for retail.

For Commercial Usage: To discuss your requirements, learn about the price and buy the dataset, leave a request on [TrainingData](#) to buy the dataset

Each image in the dataset is accompanied by bounding box annotations, indicating the precise locations of specific text segments on the receipts. The text segments are categorized into four classes: **item, store, date_time and total**.



Usability 10.00
License Attribution-NonCommercial-No...
Expected update frequency Never
Tags Business, Retail and Shopping, Image, Object Detection, Image-To-Text

bounding box annotations, indicating the precise locations of specific text segments on the receipts. The text segments are categorized into four classes: **item, store, date_time and total**.



Dataset structure

- **images** - contains of original images of receipts
- **boxes** - includes bounding box labeling for the original images
- **annotations.xml** - contains coordinates of the bounding boxes and detected text, created for the original photo

<https://www.kaggle.com/datasets/trainingdatapro/ocr-receipts-text-detection>

KAGGLE_Receipts Dataset (Only English)

Data Collection



✓ 400 images of English and Korean receipts

Data Labeling

Tool : Roboflow

The Roboflow Workspace interface displays a dataset named "receipt_OCR". The left sidebar contains various project management and data-related tabs like "receipt_OCR", "DATA", "MODELS", and "DEPLOY". The main area shows a grid of 263 receipt images, each with a small purple icon indicating it's been annotated. A search bar at the top allows filtering by filename, class, or image. Buttons for "Generate Version" and "Train a Model" are also present.

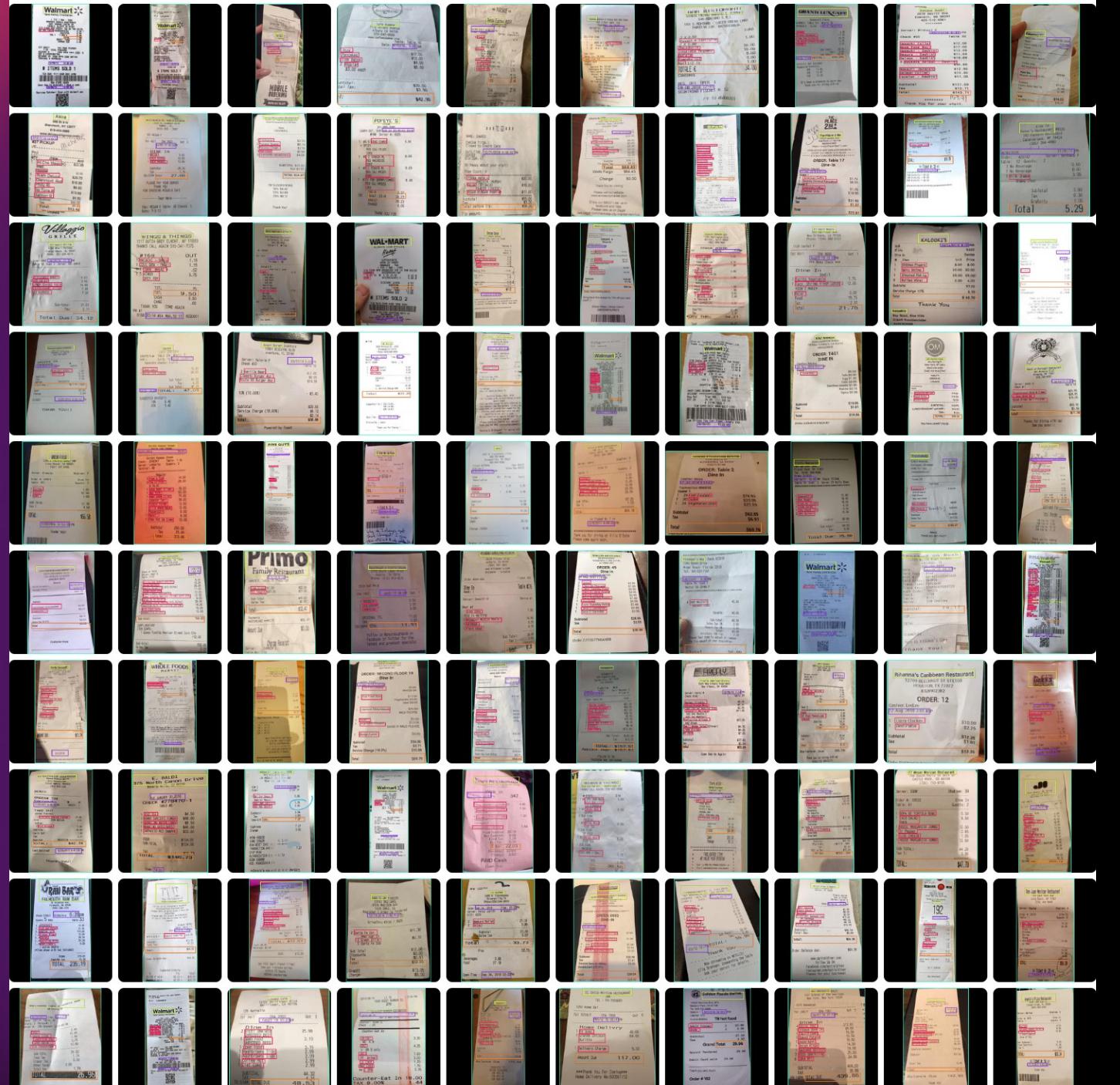
The Roboflow Annotation interface shows a specific image titled "111.jpg" from the "RECEIPT_OCR_FINE_TUNING" project. The left sidebar lists annotations for "date_time", "item", "receipt", "store", and "total". The main view shows a receipt with several items highlighted by red boxes. The receipt details include:

Item	Description	Amount
1	Philadelphia Roll	4.85
1	13. Shrimp Tempura Roll [Roll]	4.25
1	Rich / Famous Roll	4.85
1	16. Spicy Salmon Roll [Roll]	4.25
1	Veg / Tofu Pad Thai Spicy	8.95

Amount : 27.15
TAX(6%) : 1.63
TOTAL : 28.78

✓ Labels: receipt store item date_time total

Data Pre-Processing



- Auto-Rotation
- Dynamic crop : Receipt
- Resize : Fit (black edge) in 640*640

+ Data Augmentation

Model Training (YOLO)

	모델 학습 설정	Fitness	Precision (B)	Recall (B)	mAP50 (B)	mAP50-95 (B)	특징
1	YOLOv8n (Epoch 50) 데이터셋 20장만 사용	0.294	0.525	0.344	0.367	0.286	성능 개선 필요
2	YOLOv8n 데이터셋 200장 (Resize - Fit)	0.347	0.559	0.485	0.463	0.334	이미지 비율 맞춤으로 안정적 성능 향상
3	YOLOv8n (레이블링 수정, 전처리 추가 (회전, Padding))	0.534	0.717	0.687	0.729	0.512	단순 검출 성능 우수, 세부 검출 약점
4	YOLOv8m으로 모델 변경	0.604	0.865	0.715	0.780	0.585	세부 검출 성능 대폭 향상
5	영어+한국어 데이터셋 통합(400장)	0.649	0.882	0.844	0.888	0.623	데이터셋 레이아웃 통일 후 최고 성능 달성
6	데이터 증강 추가	0.6610	0.9013	0.8543	0.9114	0.6331	안정화

클래스별 성능

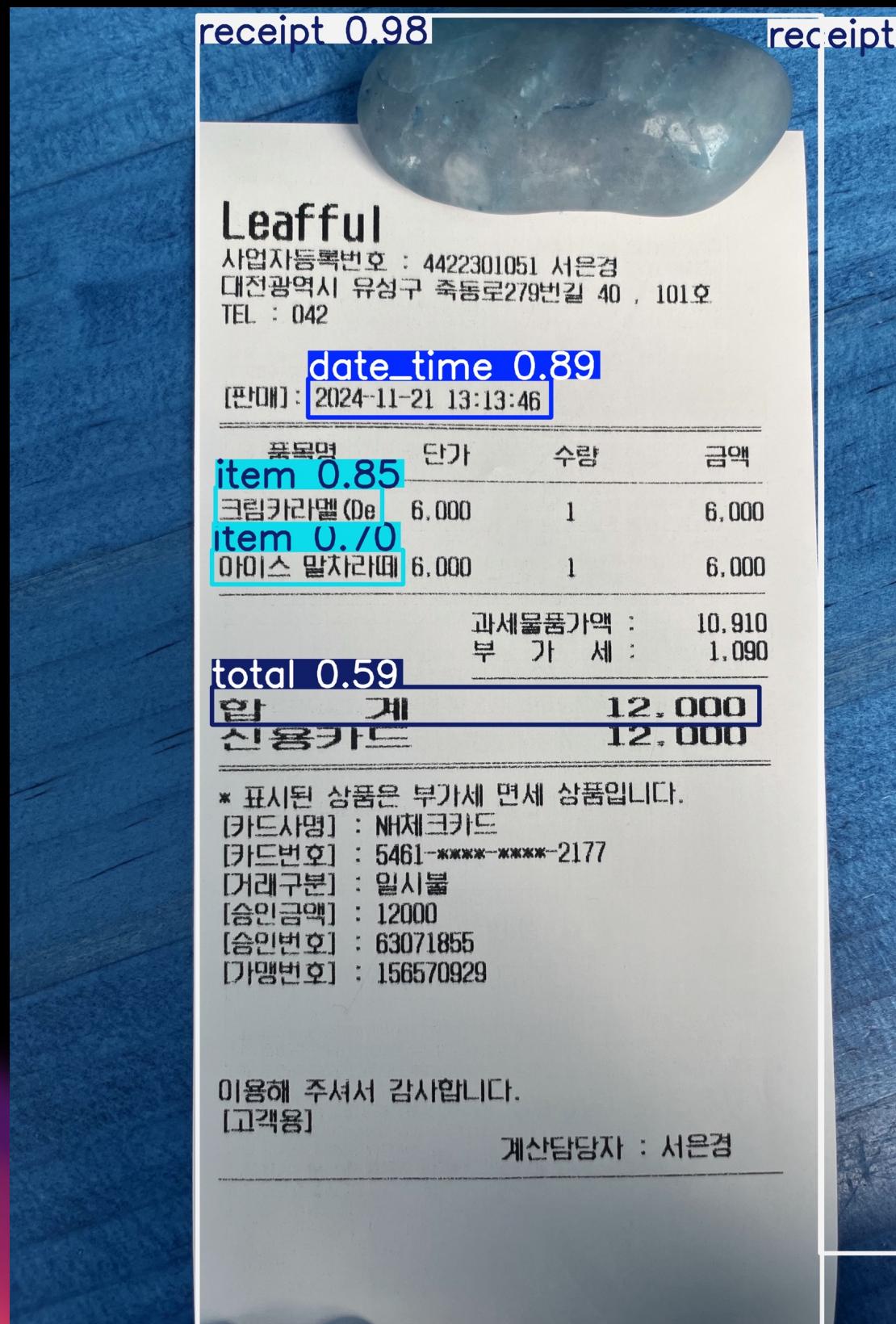
Class Name	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
date_time	0.961 (96.1%)	0.947 (94.7%)	0.981 (98.1%)	0.621 (62.1%)
item	0.978 (97.8%)	0.937 (93.7%)	0.981 (98.1%)	0.663 (66.3%)
receipt	0.985 (98.5%)	1.000 (100%)	0.995 (99.5%)	0.995 (99.5%)
store	0.889 (88.9%)	0.727 (72.7%)	0.823 (82.3%)	0.539 (53.9%)
total	0.915 (91.5%)	0.800 (80.0%)	0.913 (91.3%)	0.639 (63.9%)

전체 성능 지표

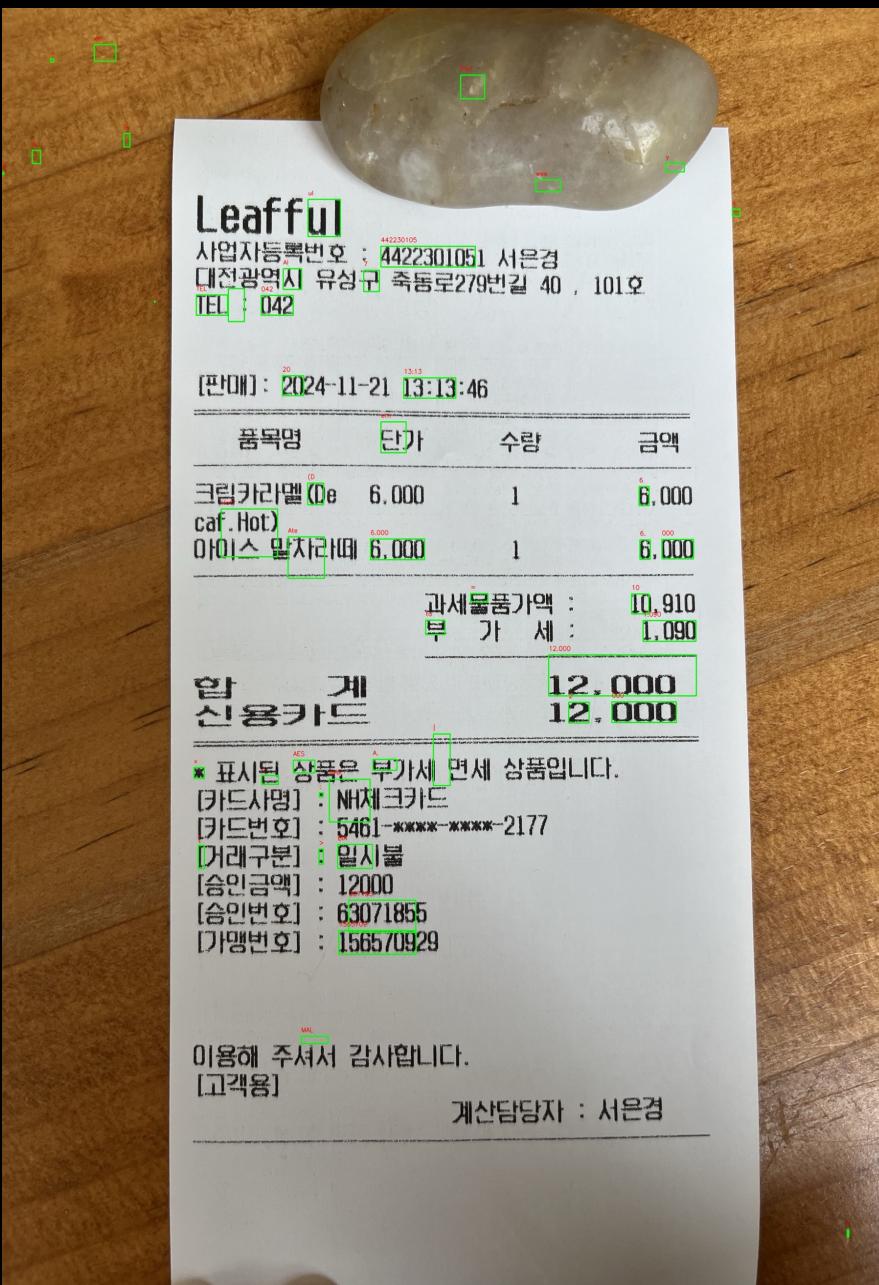
Precision (mAP@0.5):	0.939 (93.9%)
Precision (mAP@0.5:0.95):	0.691 (69.1%)
Overall Fitness:	0.716 (71.6%)

```
# 모델 불러오기
model = YOLO('yolov8m.pt')

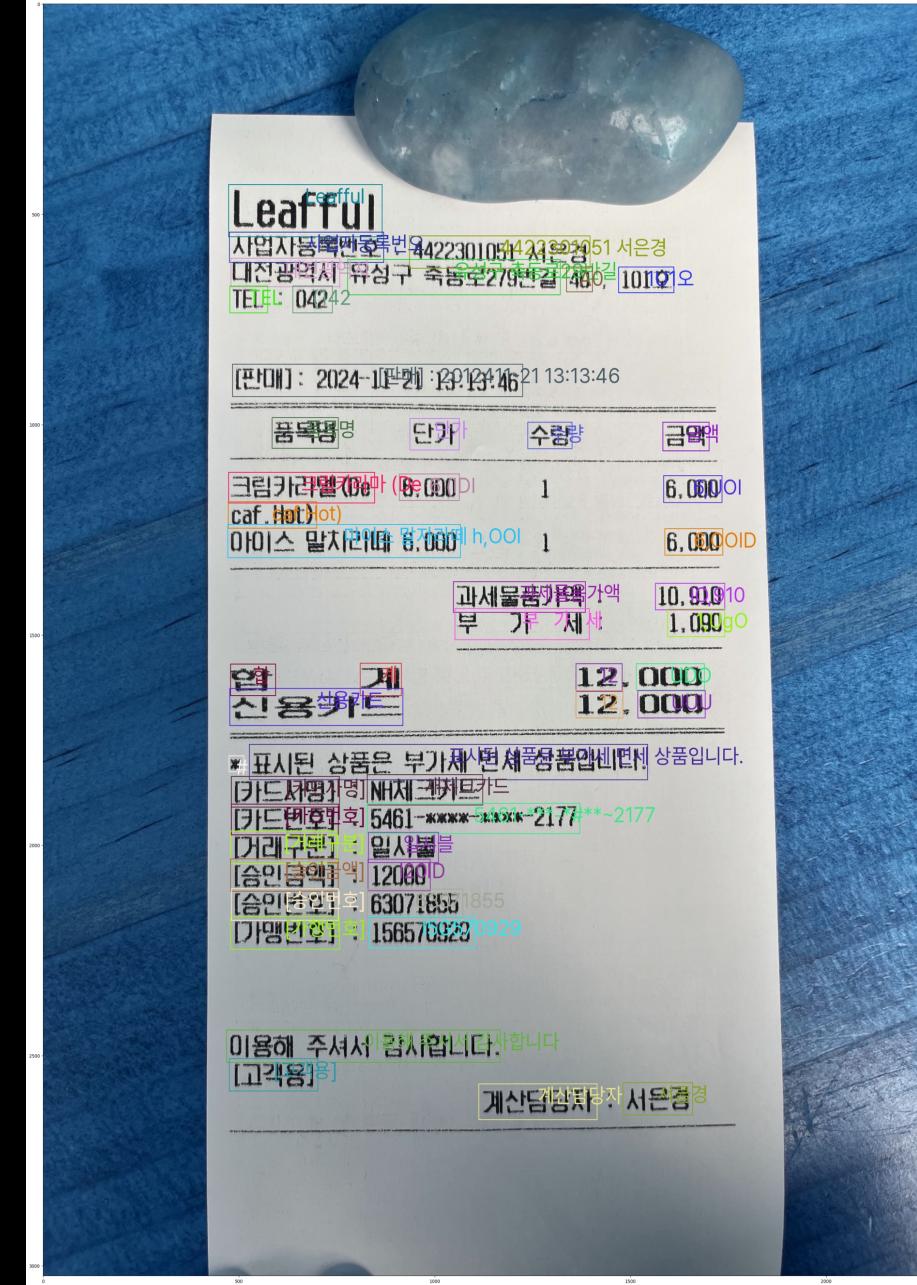
# 데이터 경로 설정 및 학습
model.train(
    data="/content/drive/MyDrive/AISW_project/dataset2/data.yaml",
    epochs=100,
    imgsz=640,
    batch=16,
    name="receipt_detection",
    device=0,
    patience = 20,
    augment=True
)
```



Tesseract



EasyOCR



PaddleOCR



정확도	PaddleOCR > EasyOCR >> Tesseract
속도	Tesseract > PaddleOCR > EasyOCR
다국어 지원	Tesseract > EasyOCR > PaddleOCR
레이아웃 처리	PaddleOCR > EasyOCR >> Tesseract
사용편의성	EasyOCR > Tesseract > PaddleOCR

EasyOCR

YOLO 결과 필터링

```
# YOLOv8으로 이미지 처리
results = model(image)
result = results[0] # 단일 이미지 처리 시

# 클래스별로 검출 결과 저장
detections_by_class = {0: [], 1: [], 3: [], 4: []}
confidence_threshold = 0.45

# YOLO 모델 결과에서 바운딩 박스 추출
for box in result.boxes:
    confidence = box.conf[0].item()
    class_id = int(box.cls[0].item())

    if class_id in detections_by_class and confidence >= confidence_threshold:
        x1, y1, x2, y2 = map(int, box.xyxy[0].tolist())
        detections_by_class[class_id].append((confidence, x1, y1, x2, y2))

# 클래스 0, 3, 4: 신뢰도 가장 높은 영역만 선택
selected_detections = []
for class_id in [0, 3, 4]:
    if detections_by_class[class_id]:
        best_box = max(detections_by_class[class_id], key=lambda x: x[0])
        selected_detections.append((class_id, *best_box))

# 클래스 1: 모든 결과 추가
selected_detections.extend([(1, *det) for det in detections_by_class[1]])

# PaddleOCR Reader 초기화
ocr = PaddleOCR(use_angle_cls=True, lang='korean', det_db_box_thresh=0.5)

# 선택된 영역에서 OCR 수행
for detection in selected_detections:
    class_id, confidence, x1, y1, x2, y2 = detection
```

store, date_time, total : 1 Bounding Box
with Highest Confidence

Item : All Bounding Box with Confidence > 0.5

EasyOCR



Leaffful

2024-11-21 13:13:46

크림카라멜 (De

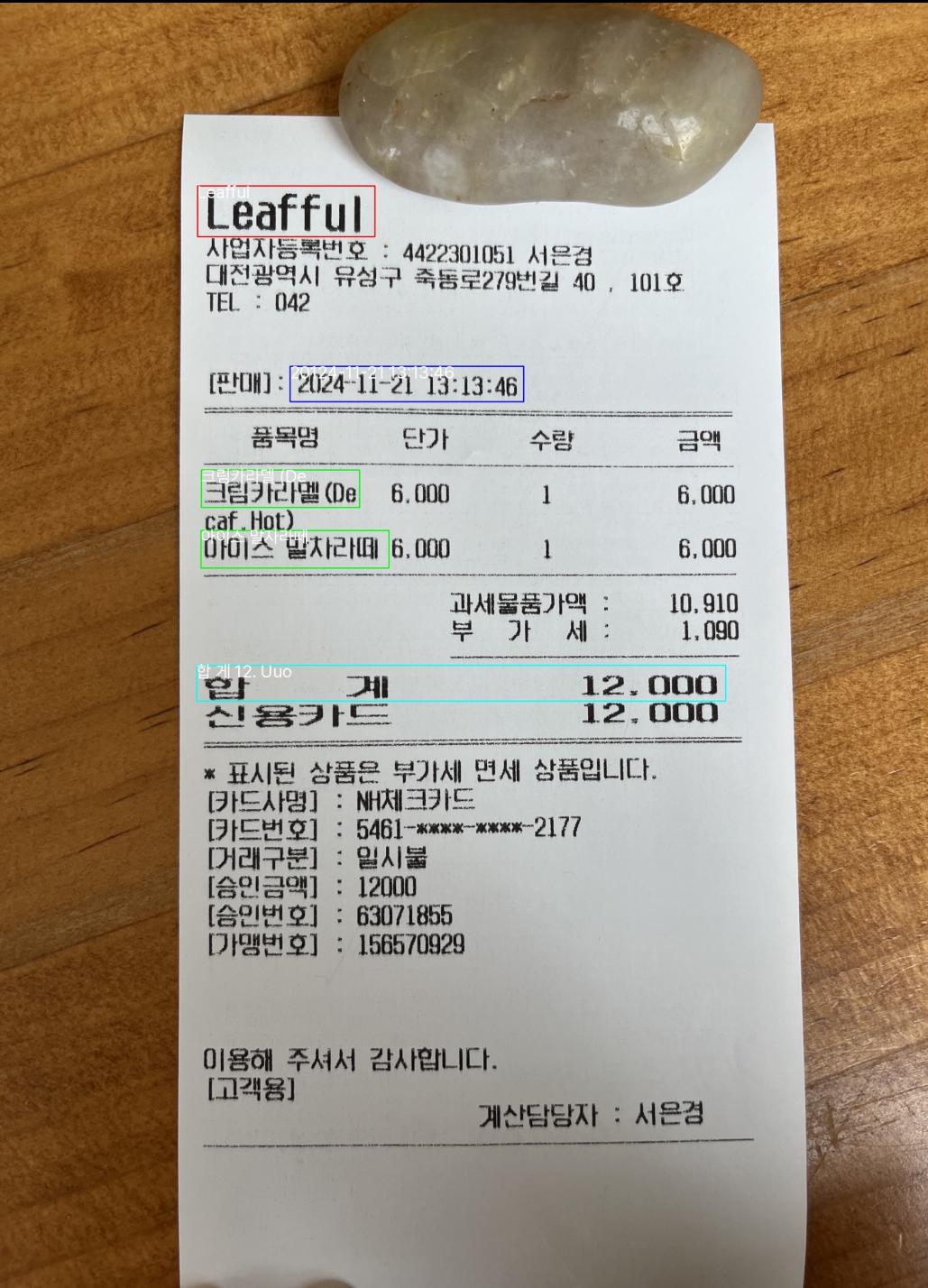
아이스 말자라떼

합계

12,000

store	Leaffful
date_time	20214-11-21 13:13:46
item	크림카라멜 (De , 아이스 말자라떼
total	합계 12.Uuo

EasyOCR



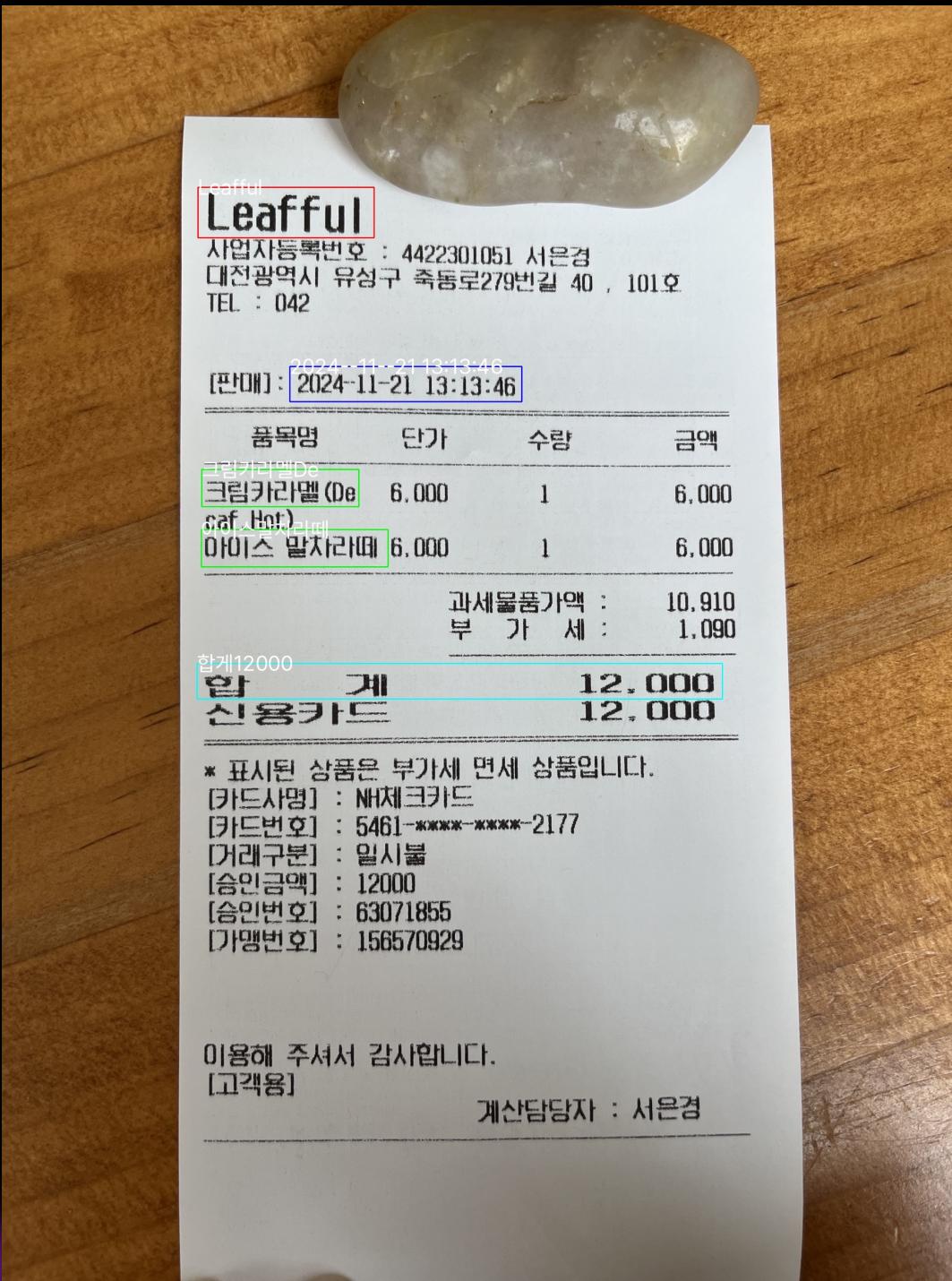
PaddleOCR



store	Leafful
date_time	20214-11-21 13:13:46
item	크림카라멜 (De , 아이스 말차라떼)
total	합계 12.Uuo

store	Leafful
date_time	2024—11--21 13:13:46
item	크림카라멜De , 아이스 말차라떼
total	합계 12000

PaddleOCR



store	Leaffful
date_time	2024—11--21 13:13:46
item	크림카라멜De , 아이스 말차라떼
total	합계 12000

날짜 (Date): 날짜 형식으로 변환.

상호명 (Store Name): 텍스트 그대로 저장.

품목 (Items): 여러 품목을 리스트 형태로 정리.

금액 (Total): 숫자만 추출해 금액 필드로 저장.



Datetime Library 정규표현식



구조화된 결과: {'상호명': 'Leaffful',
'금액': 20100,
'품목': ['크림카라멜', '아이스 말차라떼'],
'날짜': '2024-11-21'}

THANK YOU