Automated Expiry Date Extraction System

Final Presentation

Joanne Dong
Benedict Ejelonu
Ryo Fukano
Emma Sandidge

Academic Mentors:
M. Ogawa, S. Kano

Industry Mentors:
T. Nakayama, A. Kubo, K. Kitamura, H. Fujimura, H. Sakakura, M. Oh

IHI Team G-RIPS 2025

G-RIPS Sendai 2025

IHI

AIMR

MathCCS

Mathematical Science Center
for Co-creative Society,
Tohoku University

Thank you to our sponsors!

TOHOKU UNIVERSITY

東北大学
TOHOKU UNIVERSITY

IPAM

NSF

TOHOKU FORUM for CREATIVITY

Problem Statement

The goal of the project is to design an algorithm that detects the expiration date from images of cardboard boxes containing food and beverage items.

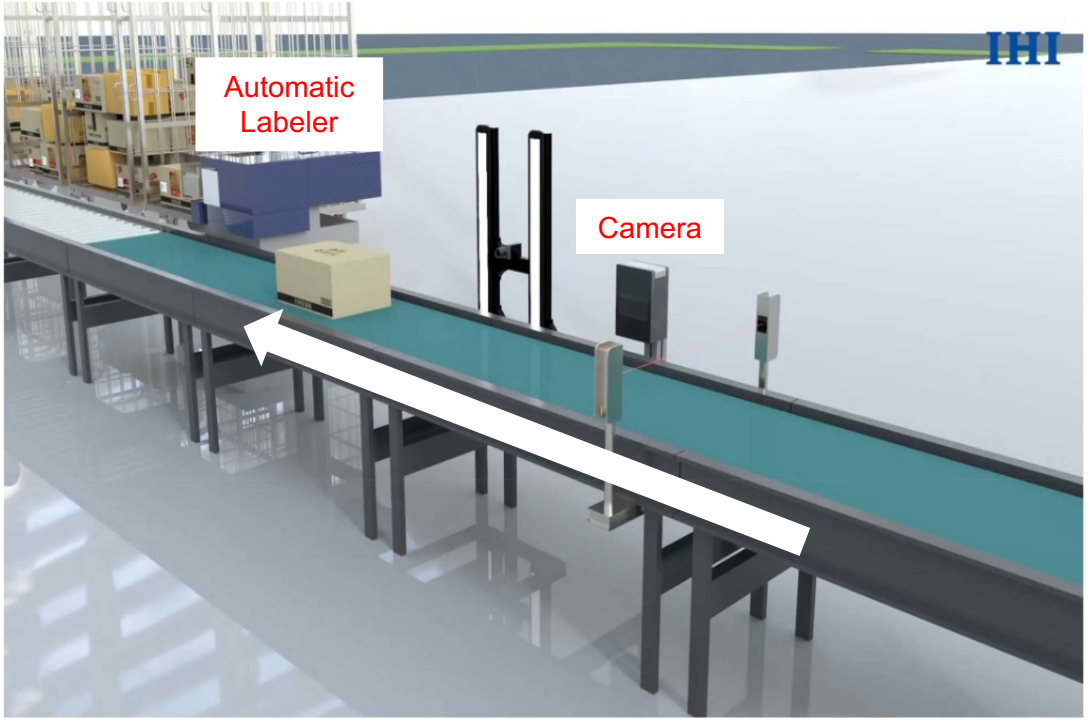
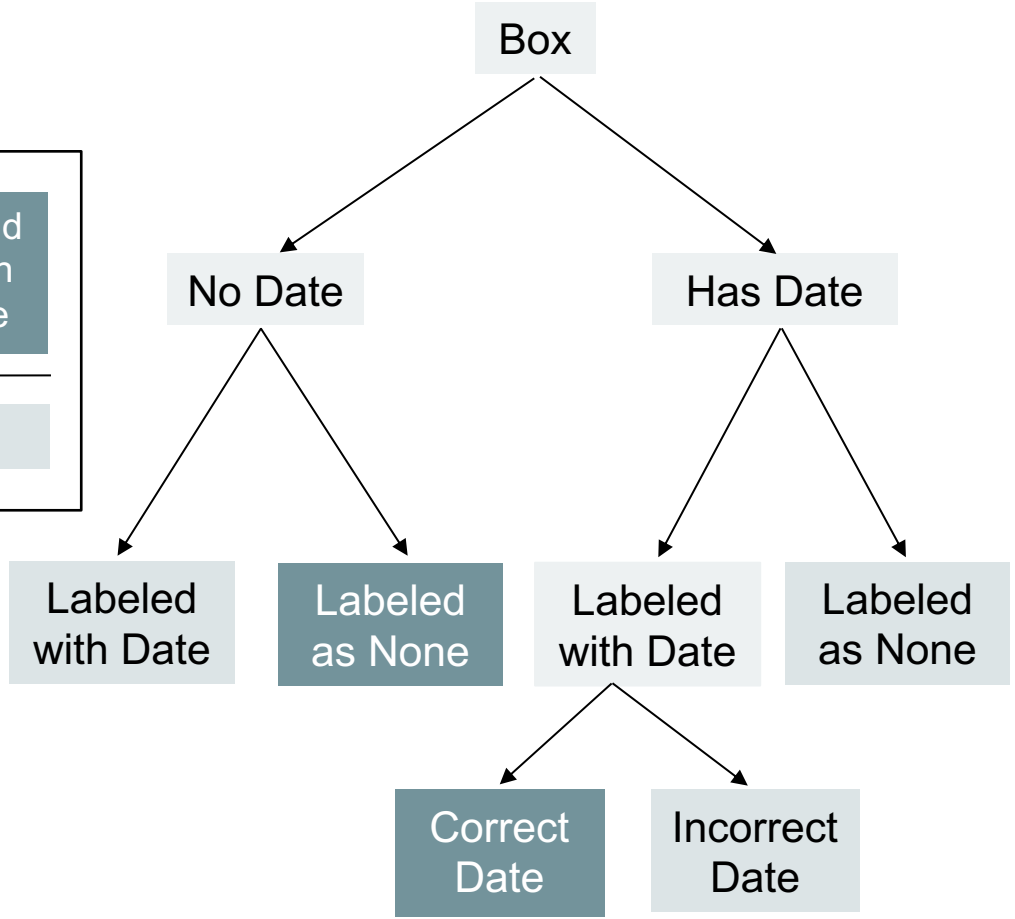


Photo credit: IHI Logistics & Machinery Corporation

Accuracy

$$Accuracy = \frac{\text{No Date and Labeled as None} + \text{Has Date and Labeled with Correct Date}}{\text{Total Number of Boxes}}$$



Dataset

Bounding Box Coordinates in (x,y,w,h) format:

```
1 0.3345189296207956 0.5552776191330846 0.20351175621678036 0.027518116074985385
0 0.547505929548074 0.4782751715197483 0.13018160923988376 0.05503623214997063
2 0.7839750572232466 0.5210006675309086 0.1425406349587628 0.13855612830738234
0 0.7489643720136934 0.7738711402630623 0.16602270663208935 0.06371784010757174
```

Labeling Strategies

Not line-separated

Line-separated

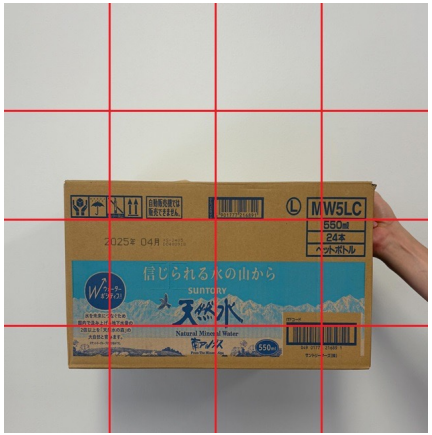
YOLO models for object detection

- YOLO (You Only Look Once) : State-of-the-art **object detection** and image segmentation model built on a deep learning framework.
- Key trade-off : Model size vs. accuracy performance.
- Model variants tested : YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv11m.
- Training approaches: From scratch vs. **transfer learning** methods.
- Dataset: 2500 images with 70% train / 20% validation / 10% test.

YOLO Architecture

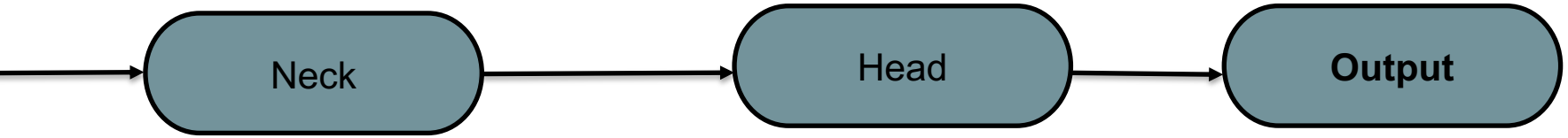
Input: Image

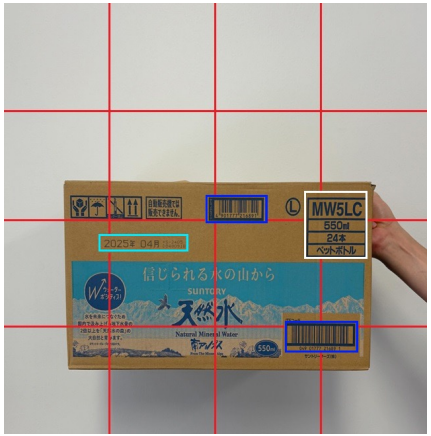
Backbone



- Convolutional neural network.
- Extracts features from input image.

YOLO Architecture



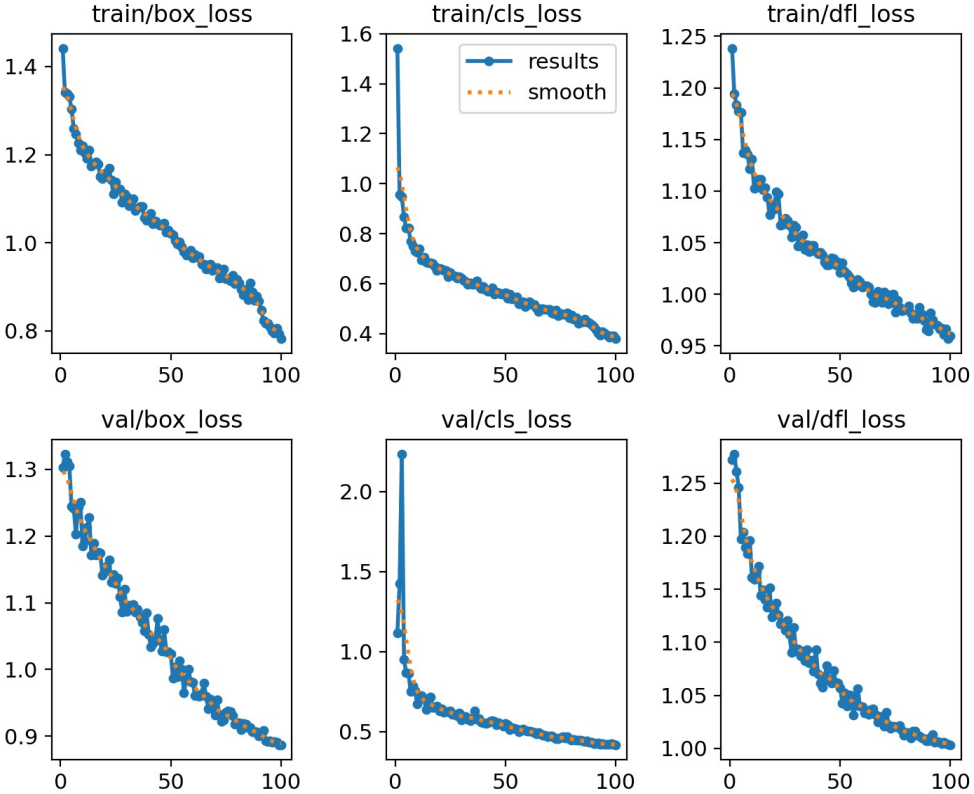


- Merges feature maps to gather information at different scales.
- Uses a C2f module.

- Makes predictions for bounding boxes.
- Classifies bounding boxes.

- Bounding boxes for detected objects.
- Class for each object detected.
- Objectness loss.
- Classification loss.
- Distributed focal loss.

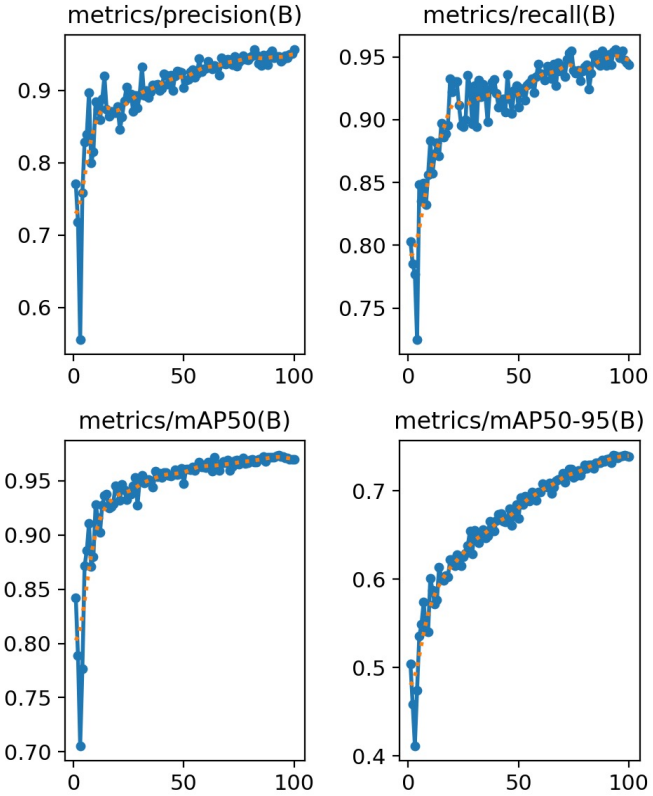
YOLOv11m Metrics



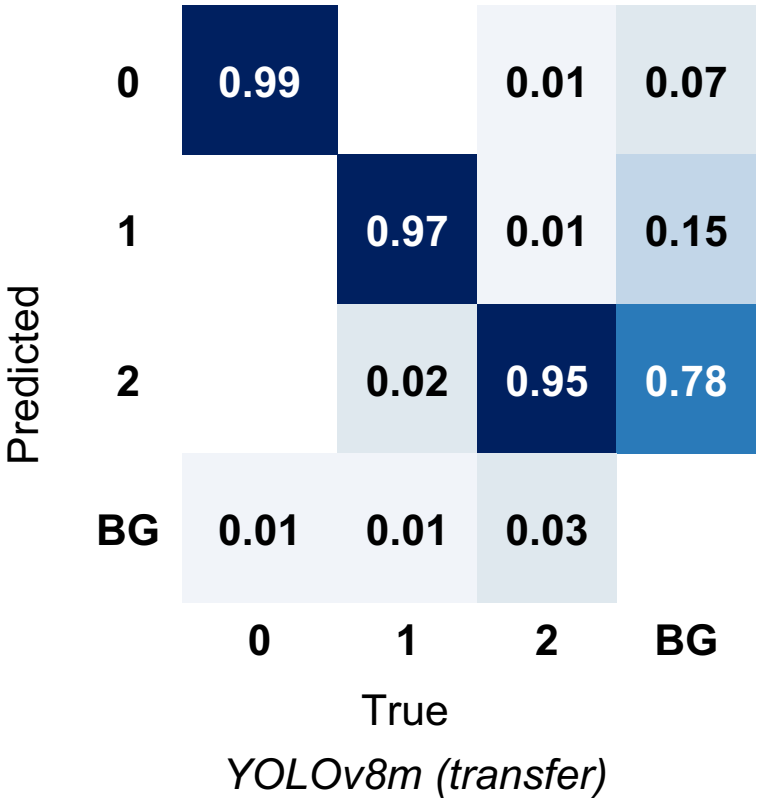
- **Loss Minimization**
- 100 Epochs
- Total Loss = λ_1 × Box Loss + λ_2 × Class Loss + λ_3 × DFL Loss
- Where λ_i is a loss weight to balance the components

YOLOv11m Metrics

- **Metric Maximization**
- mAP : mean Average Precision



Confusion Matrices for YOLOv8m and YOLOv11m



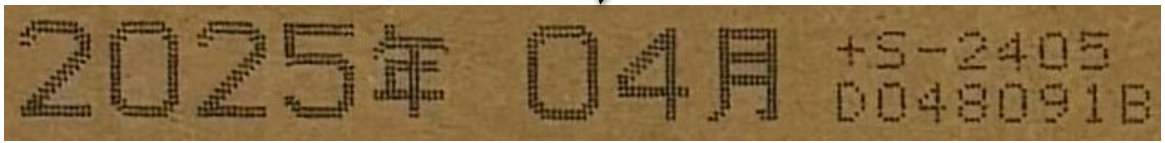
YOLOv8m (transfer)

YOLO v11m (transfer)

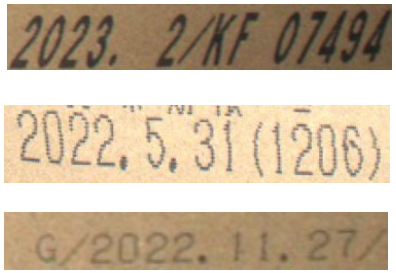
Legend	
0	BarCode
1	DateInfo
2	ProdInfo
BG	Back-ground

Text Extraction

Crop ROI from detections before reading text from the image.



Optical Character Recognition



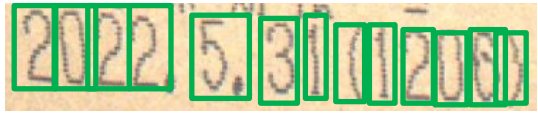
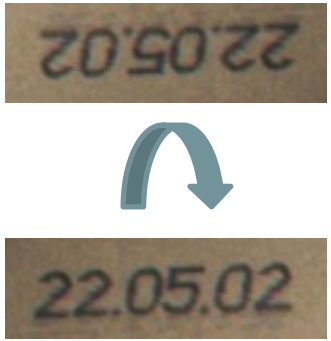


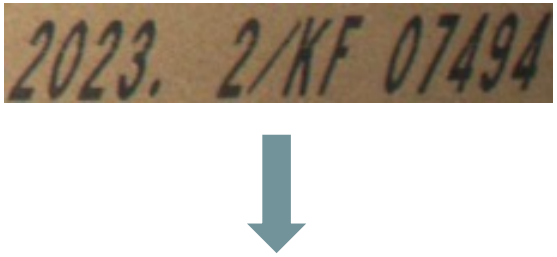
Image → Label Text → Detect Text

- Cropped ROI from YOLO

- PPOCR Label
- StyleText

- DB
- EAST
- SAST





['2023.', '2/KF', '07494']

{
"rec_texts":
['2023.', '2/KF', '07494'],
"rec_score":
[0.91, 0.98. 0.88]
}

Classify Direction → Recognize Text → Output

- Identify text orientation

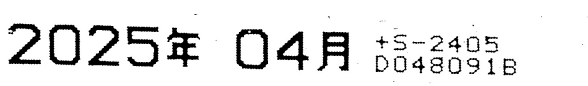
- Rosetta
- StarNet
- CRNN
- RARE
- SRN

- Detected texts
- Confidence scores

Reproduced from: https://2021.desosa.nl/projects/paddleocr/posts/paddleocr-e2/

EasyOCR vs. PaddleOCR

- Our first pipeline using EasyOCR plateaued at around 70% accuracy.
- OCR Engine Switching: **EasyOCR → PaddleOCR.**

	Raw	Preprocessed
Input Image		
EasyOCR	)1/) . ::2:: 7- . 年 (.	2025年04月 64398
PaddleOCR	2025年04月 +5-2405 D048091B	2025年04月 +5-2405 D048091B

Ground Truth : 2025年04月 +S-2405 D048091B

Post-processing: regex patterns

Dates with Separators

2024○8○1

- 2-4 digits (year)
- Separator*
- 1-2 digits (month)
- *(Optional)* Separator*
- *(Optional)* 1-2 digits (day)

Chinese Dates

2025年04月

- 2-4 digits (year)
- 年 or 4
- 1-2 digits (month)
- 月
- *(Optional)* 1-2 digits (day)

Dates without Separators

21122

- 5-8 digits

* ○ is a placeholder for space, comma, midpoint(·), period(.) , slash(/), and hyphen (-).

Post-processing: regex patterns

Dates with Separators

2024◦8◦1

2024-08-01

Chinese Dates

2025年04月

2025-04-30

Dates without Separators

21122

2021-01-22
2021-12-02

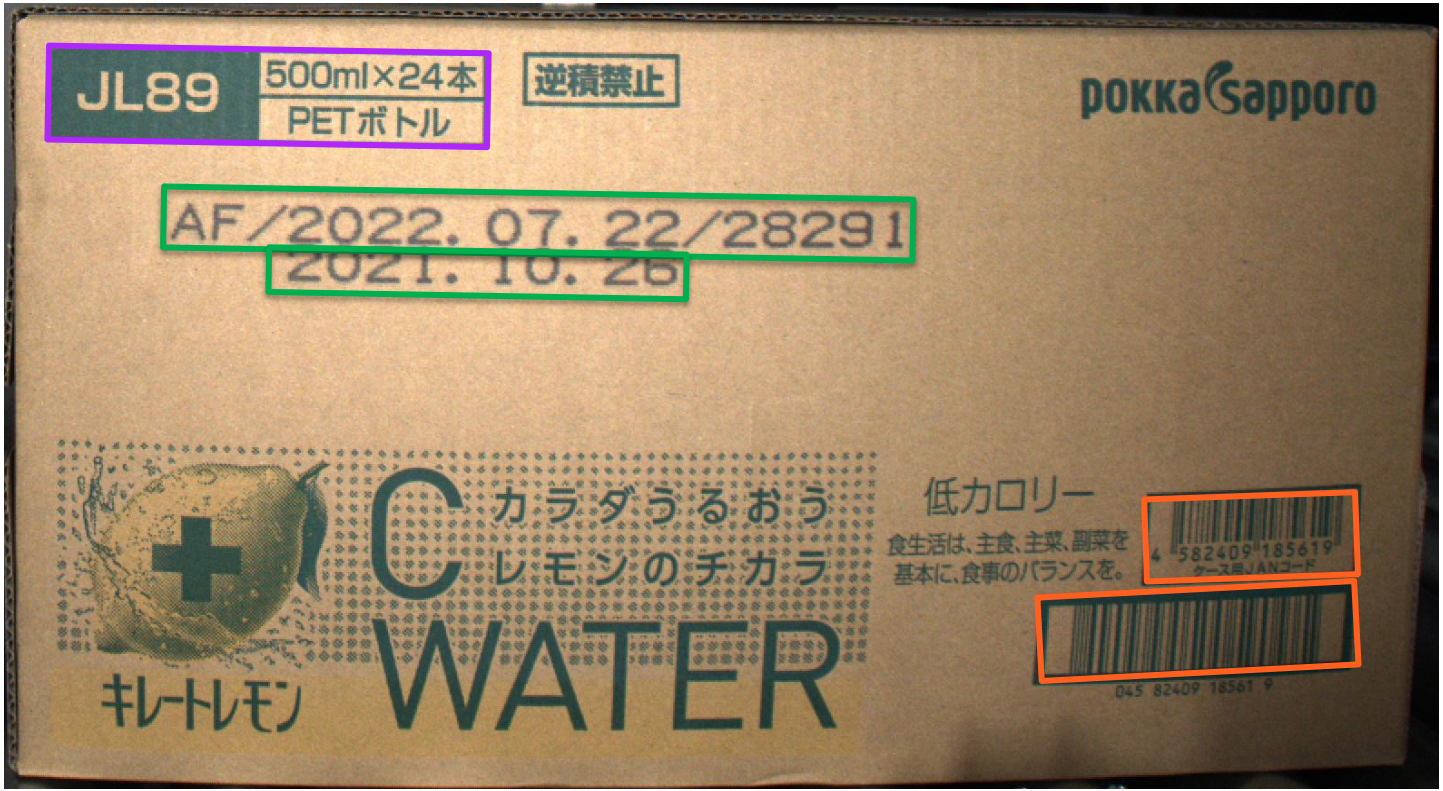
* ◦ is a placeholder for space, comma, midpoint(·), period(.) , slash(/), and hyphen (-).

Input Image

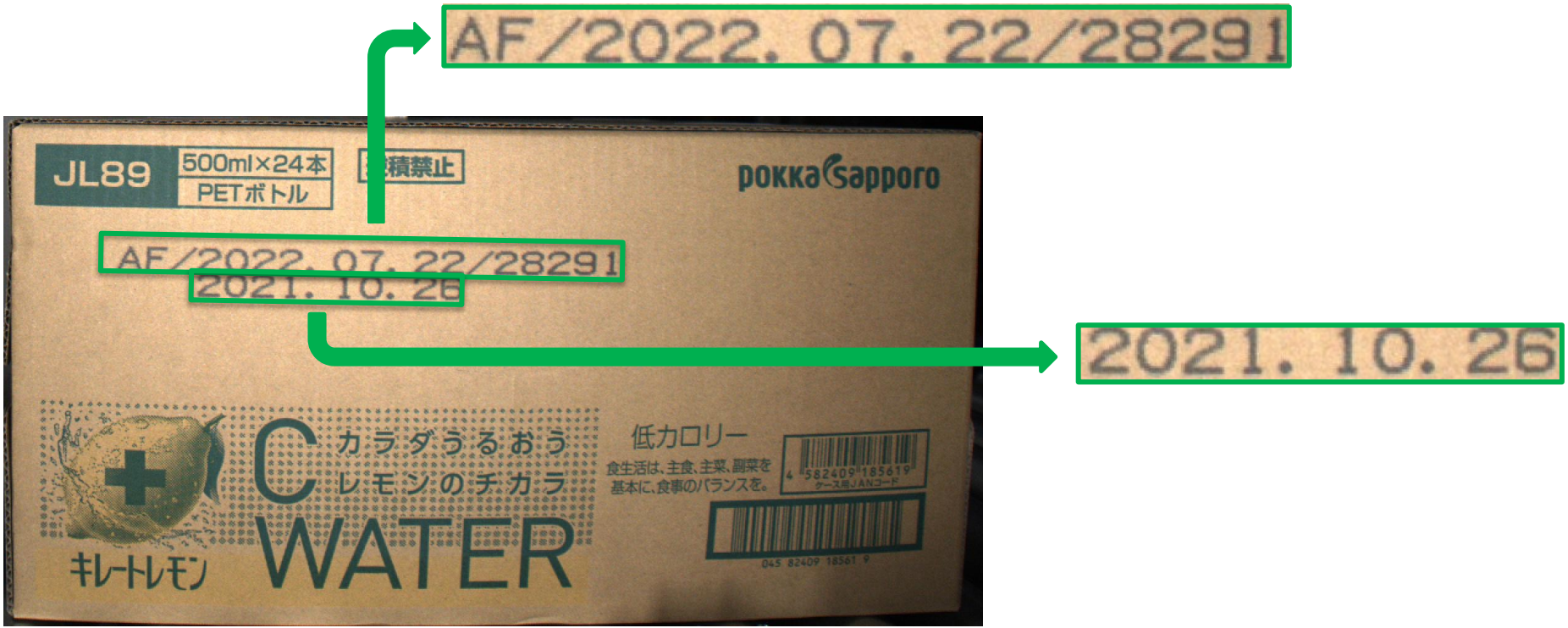

JL89
500ml×24本
PETボトル
逆積禁止
pokka sapporo
AF/2022. 07. 22/28291
2021. 10. 26
C
カラダうるおう
レモンのチカラ
WATER
キレートレモン
低カロリー
食生活は、主食、主菜、副菜を
基本に、食事のバランスを。
4 582409 185619
ケース用JANコード
045 82409 18561 9

YOLO Detections

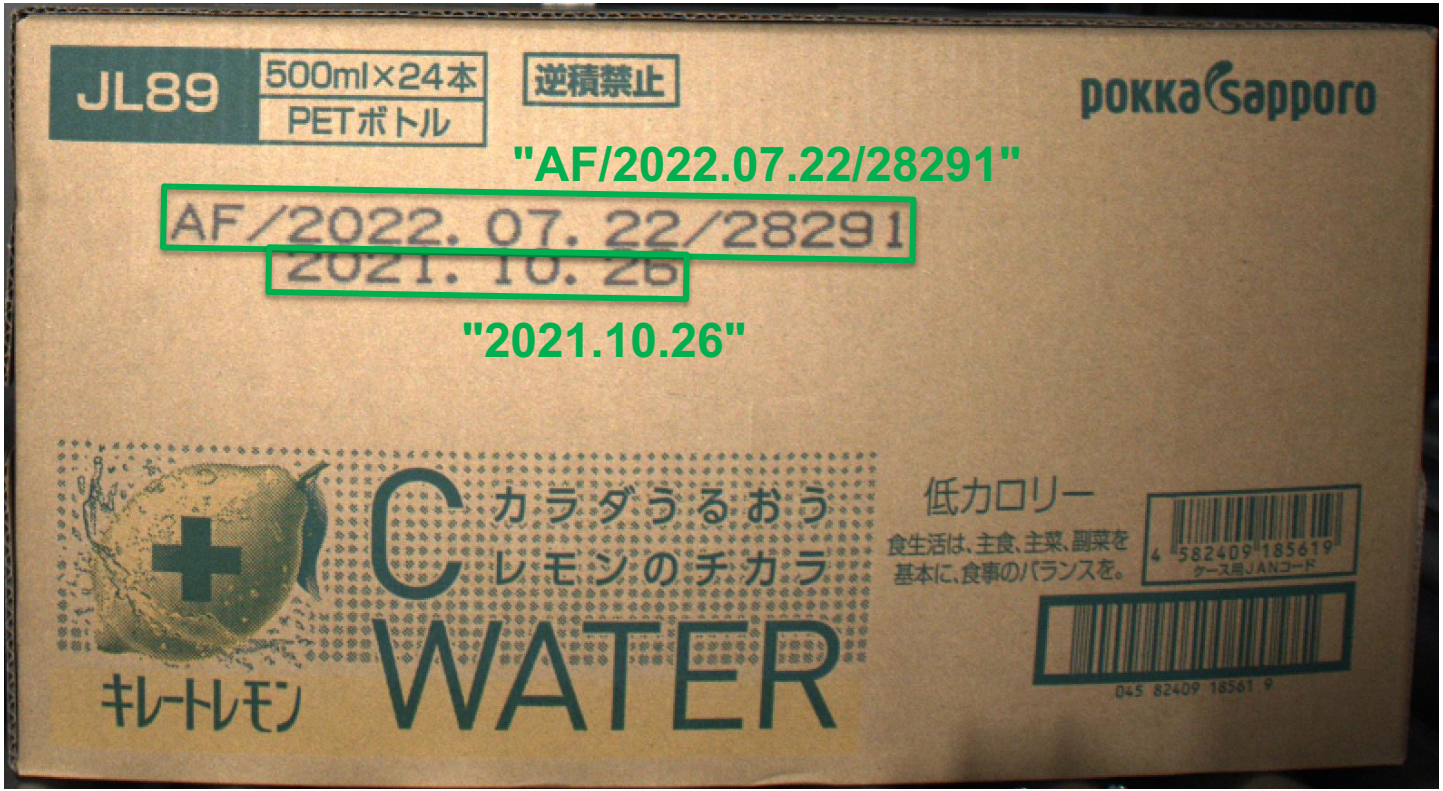
BarCode
DateInfo
ProdInfo



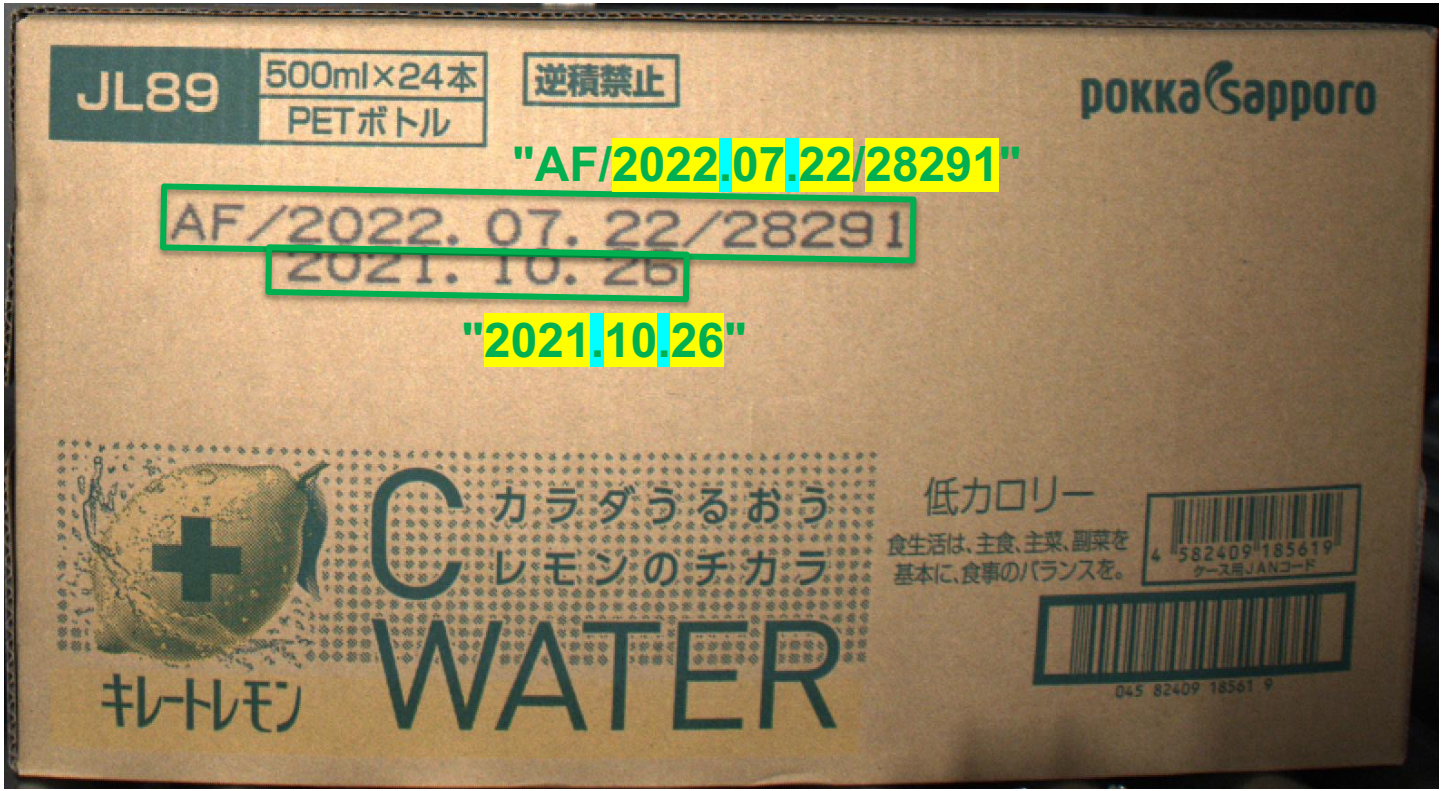
Cropped Detections



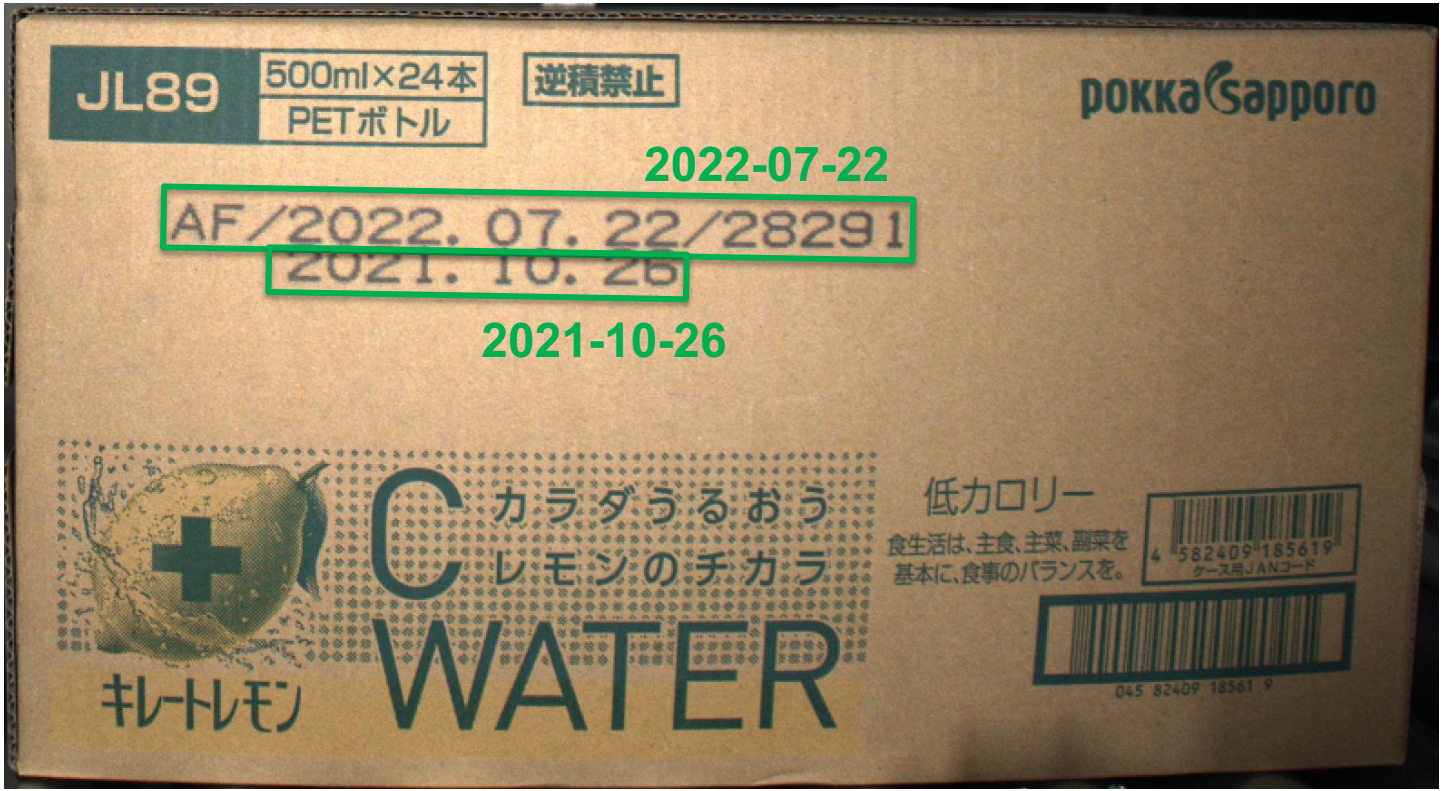
Extracted Texts from OCR



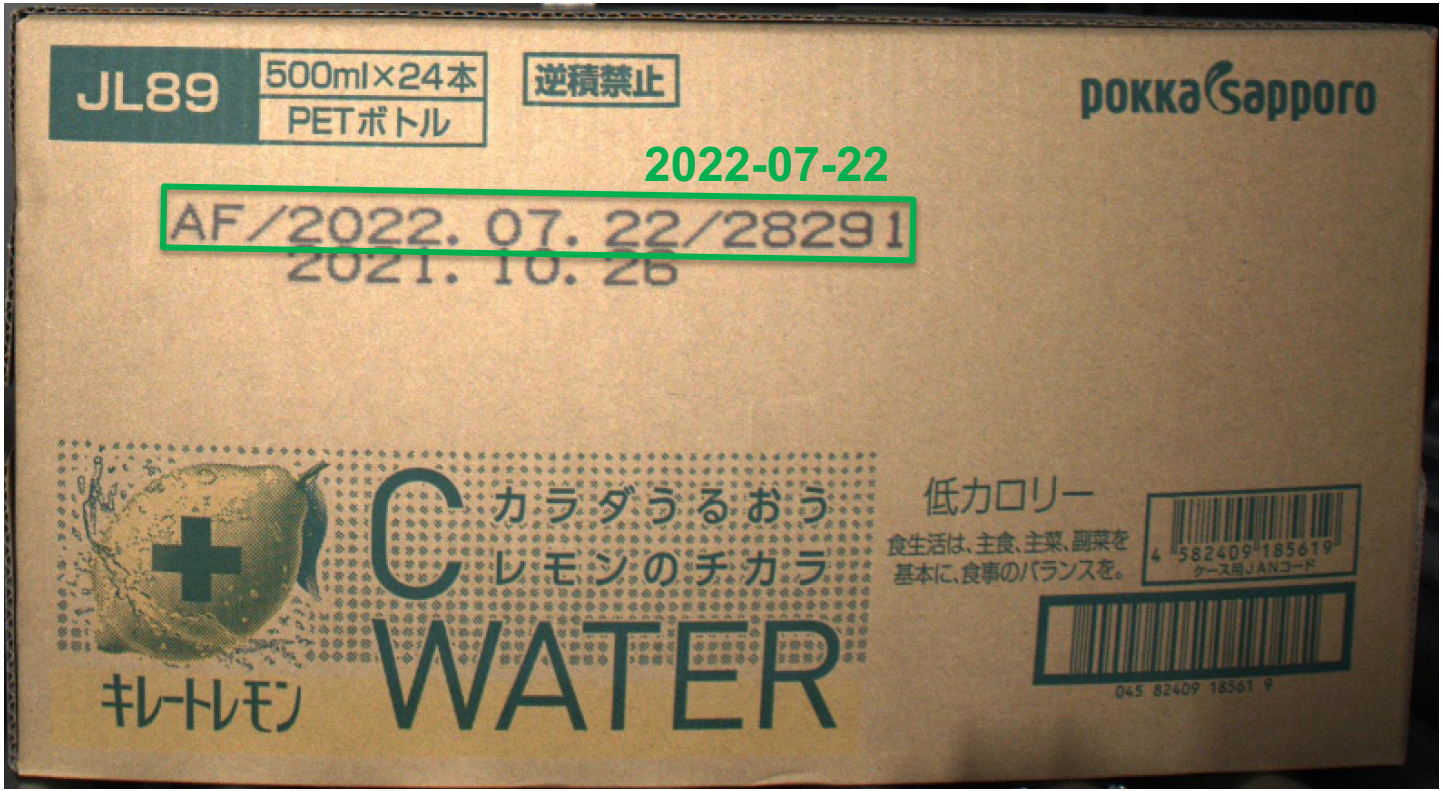
Regex Pattern Matching



Date Parsing



Final Output



Overall Performance

Pipeline Version	OCR Engine	Year-Month-Day Accuracy [%]	Year-Month Accuracy [%]
First	EasyOCR	70.8	76.2
Latest	PaddleOCR	96.7	98.1

Accuracies calculated from a test set of 1000 images.

Future Works

- Extraction of other kinds of information: **ITF-14 Barcode**.
- Resource Optimization: CPU or lower tier GPUs than A100.
- Processing speed: Improving real-time processing speed.
- Improving the **robustness** of the **regex** operation.
- Robustness to various image conditions: YOLO retraining with expanded data diversity.

Thank You

IHI Team G-RIPS 2025