Object Detection for Monitoring Workplace Safety

Problem Statement:

Creating an object detection system that can help monitor workplace safety by detecting personal protective equipment such as helmets and vests in videos and images. The system uses a pre-trained model YOLOv11 which is fine tuned on a PPE dataset to make the detections more accurate.

Methodology:

The system works using the pre-trained YOLO model which divides images into grids and predicts bounding boxes and class probabilities in a single forward pass, balancing speed and accuracy.

YOLOv11:

The latest addition in the YOLO series of object detection models is the YOLOv11. The key features of this model is that it can directly predict bounding boxes and class probabilities from the input image using a single neural network pass, thereby minimizing latency. From image input to final predictions, the entire pipeline is end-to-end differentiable. Which means all components of the model can be optimized simultaneously during training which ensures cohesive learning.

The architecture consists of three main components :

1, Backbone: This consists of Darknet which is a modified CNN that focuses on extracting features rather than generating outputs. It also uses C3K2 Blocks that use smaller 3x3 convolution kernels to optimize feature extraction while maintaining computational efficiency. The block processes information through a series of convolutions and bottleneck layers hence enhancing feature representation with few parameters.

2. Neck : Spatial Pyramid Pooling Fast (SPFF) to pool features from different regions of an image at varying scales and improve the model's ability to capture small objects. max-pooling operations are employed to aggregate multi-scale contextual information. A C2PSA Block that applies attention mechanisms to enhance the model's focus on important regions which makes for better detection of small or cluttered objects.

3. Head: The head outputs detection boxes for three different scales using the feature maps generated by the backbone and neck. It processes the features to predict bounding boxes and class probabilities.

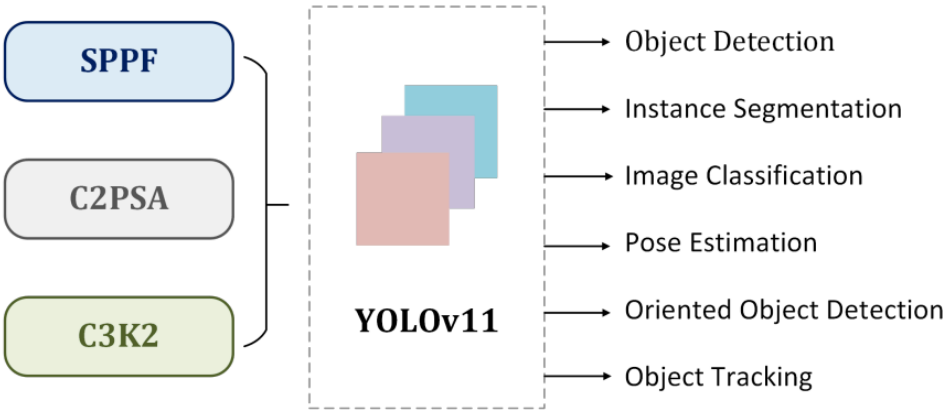

Figure 1 : YOLOv11

Architecture Evolution

YOLOv1 (2015): Introduced a grid based approach where bounding box and class probabilities are predicted by each cell in one pass but struggled with small objects and localization accuracy.

YOLOv2 (2018): Replaced fully connected layers with anchor boxes that improved recall. Added batch normalization and multi scale training

YOLOv3 (2017): Added Darknet-53 architecture which further refined accuracy.

YOLOv4 (2020): Added CSPDarknet53 which improved accuracy and speed over its previous model.

YOLOv5 (2020): Simplified neck design and optimized for speed.

YOLOv6 (2022): Extended efficient layer aggregation for scalable feature extraction. Faster and more efficient real-time object detection.

YOLOv7 (2022): Advanced model scaling and improved backbone design.

YOLOv8 (2023): Incorporated distribution focal loss and optimized gradient flow.Redesigned architecture with dynamic anchor-free detection.

YOLOv9 (2024): Transformer-based feature extraction and multi-scale detection.

YOLOv10 (2024): Quantization-aware training and hardware-friendly design for edge AI applications.

YOLOv11 (2024): Hybrid CNN-transformer models.

Dataset

The Construction PPE Detection Dataset has 1,206 images taken from Roboflow, specifically designed for detecting safety wearables in construction environments. It has five annotated classes: helmet, no-helmet, no-vest, person, and vest, ensuring a comprehensive analysis of worker safety. The dataset is split into training, validation, and test sets, making it suitable for supervised learning tasks. The dataset exhibits a class imbalance, with significantly more instances of person and helmet compared to

no-helmet and no-vest. Bounding box statistics indicate varied object sizes and positions. The dataset was preprocessed and structured in data.yaml for seamless integration with YOLOv11, facilitating efficient object detection and model fine tuning.

Hyperparameter Tuning YOLOv11 for PPE Detection

Hyperparameter tuning plays an important role in optimizing the performance of deep learning models, particularly in object detection tasks. In this project we have fine-tuned **YOLOv11** using **automated hyperparameter optimization** to enhance its detection accuracy for PPE-related objects and strictly to classes from dataset only.

1. Tuning Methodology

We used the Ultralytics YOLOv11 tuning module which can systematically adjust hyperparameters over multiple training iterations to identify the best performing parameter values. The model was fine-tuned on 1,206 images from the **Roboflow PPE dataset** using 22 tuning iterations, each evaluated over **30 epochs**.

2. Search Space

The hyperparameter search space was defined to cover factors affecting model performance, including:

- **Learning rate (lr0 and lrf)**: Adjusting the initial and final learning rates for faster convergence.
- **Momentum (momentum)**: Controlling weight updates to avoid oscillations.
- **Weight decay (weight_decay)**: Preventing overfitting through L2 regularization.
- **Warmup settings (warmup_epochs and warmup_momentum)**: Optimizing training stability.
- **Loss weights (box, cls, dfl)**: Fine-tuning loss functions for bounding box regression, classification, and distribution focal loss.
- **Augmentation parameters (degrees, translate, scale, hsv_h, hsv_s, hsv_v, fliplr, mosaic, mixup)**: Enhancing model robustness through data augmentation.

3. Best Hyperparameters Found

After **22 runs** with automated tuning, the best-performing hyperparameters were:

lr0: 0.00919
lrf: 0.01
momentum: 0.86723
weight_decay: 0.00041
warmup_epochs: 2.54
warmup_momentum: 0.731
box: 0.198
cls: 0.427
dfl: 1.23
mosaic: 0.607
fliplr: 0.499

These values yielded the highest **mean Average Precision (mAP@0.5)** of **0.887**, indicating improved detection accuracy across all PPE classes.

4. **Performance Insights**
    - **Learning rate (lr0)** was optimized to **0.00919**, balancing training speed and stability.
    - **Momentum (0.86723)** improved convergence without excessive weight oscillations.
    - **Augmentations (mosaic=0.607, fliplr=0.499)** enhanced generalization without degrading validation performance.
    - **Loss function weights (box=0.198, cls=0.427, dfl=1.23)** were fine-tuned for more accurate object localization.

The final trained model was validated using **precision-recall curves, confusion matrices, and confidence metrics**, demonstrating significant improvements in PPE detection accuracy.

Results:

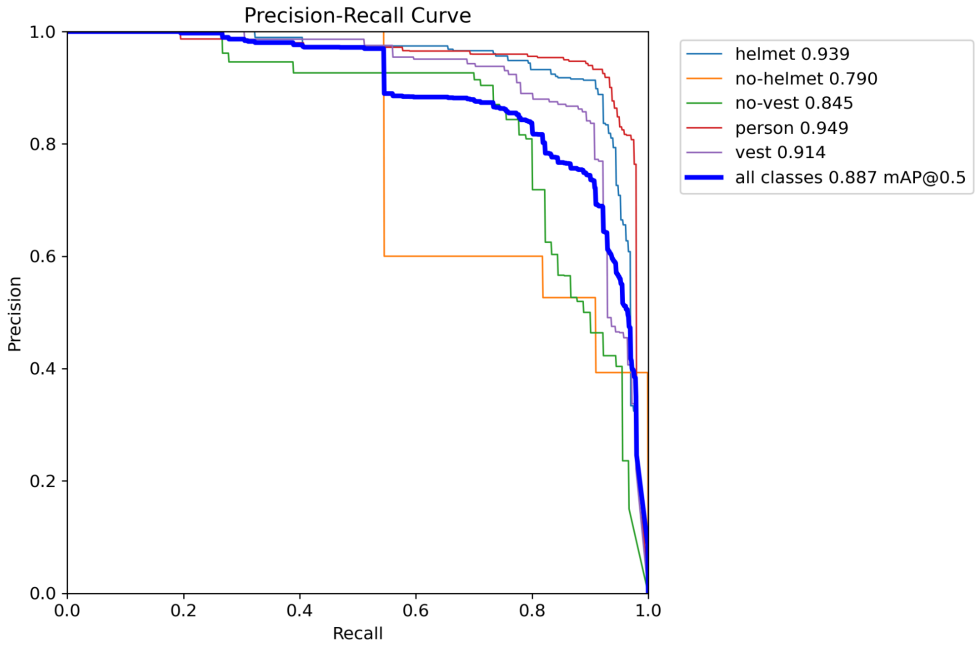

Figure 2: Precision-Recall Curve

The mAP@0.5 of 0.887 indicates strong overall detection performance. A high suggests the model maintains good precision even as recall increases.

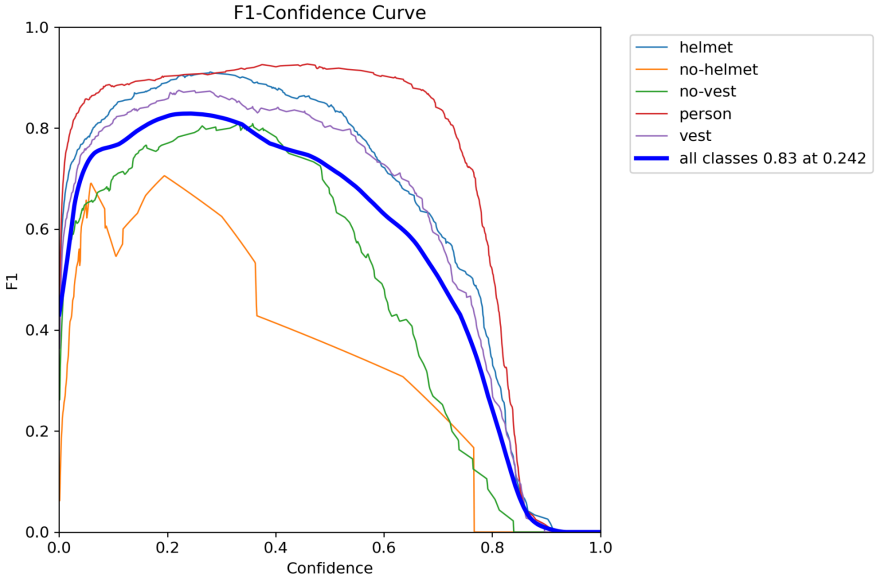

Figure 3: F1-confidence Curve

The use of YOLOv11n prioritizes inference speed over absolute accuracy, which may limit performance on complex cases. The F1 score of 0.83 at confidence 0.242 reflects a balance between precision and recall at lower thresholds.

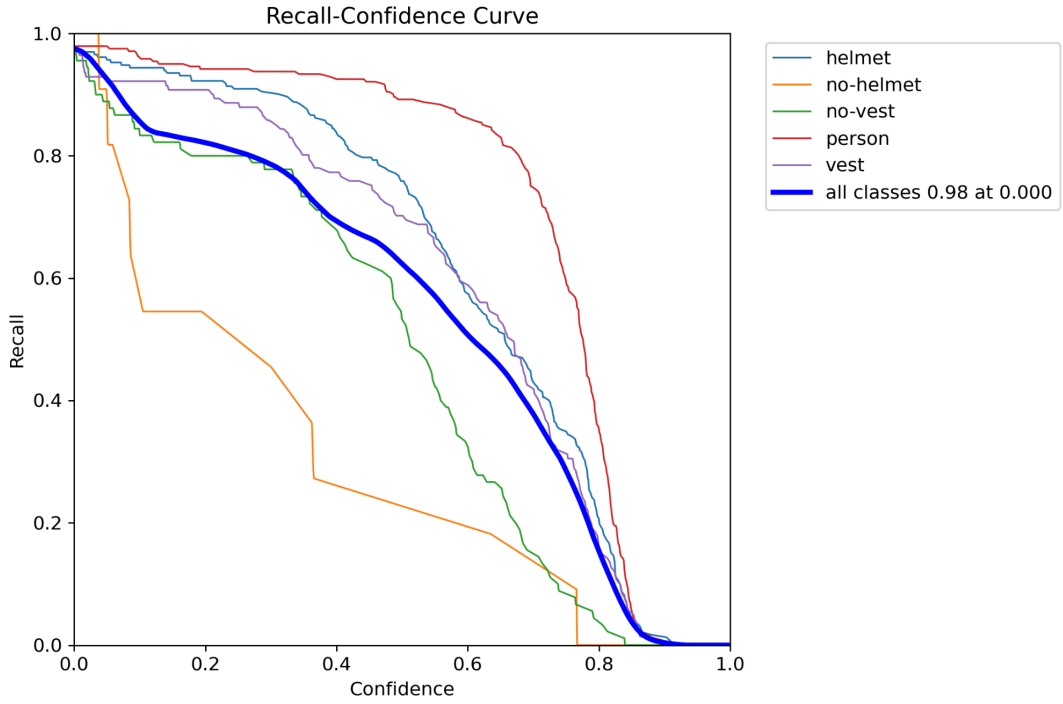

Figure 4: recall-confidence Curve

At a confidence threshold close to **0.0**, recall is **high** (~1.0) for all classes, meaning the model detects almost all objects. The model performs well for detecting "persons" but struggles with "no-helmet" and "no-vest."

References:

G. Jocher and J. Qiu, "Ultralytics YOLO11," Ultralytics, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/.