JTE HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Enhancing Workplace Safety: Personal Protective Equipment Detection

Vo Thanh Xuan Le[1], Khac-Chien Nguyen[2*]
[1]*Ho Chi Minh City University of Technology, Vietnam*
[2]*People's Police University, Vietnam*

[*]Corresponding author. *Email: nkchienster@gmail.com*

ARTICLE INFO

Received: 07/08/2024
Revised: 05/09/2024
Accepted: 09/12/2024
Published: 28/08/2025

KEYWORDS

Convolutional Neural Networks (CNN);
Deep learning Architecture;
Personal Protective Equipment (PPE);
You Only Look Once (YOLO);
Machine Learning.

ABSTRACT

Industries such as construction, cold food processing, and the chemical sector are particularly vulnerable to a range of potential hazards. Personal Protective Equipment (PPE) plays a critical role in safeguarding workers in these high-risk environments. However, ensuring the consistent use of PPE and adherence to established safety protocols is a complex task. This complexity arises from factors such as human error, negligence, and inadequate supervision. Traditional methods of monitoring PPE compliance typically involve manual inspections, which are not only labor-intensive but also have demonstrated limited effectiveness in ensuring consistent PPE use. To address these challenges, this study proposes the utilization of the YOLOv8 algorithm to achieve improved accuracy and suitability for a broader range of real-world working environments. In support of this approach, we have developed a new dataset named PPE-AYN, which includes five distinct classes (person, head, hat, glasses, and glove) and comprises a total of 2980 images. The YOLOv8 algorithm represents the latest advancement in the YOLO family of object detection models and is renowned for its rapid and precise detection capabilities. These characteristics make YOLOv8 particularly well-suited for the task of PPE detection, offering a promising solution to enhance safety compliance in various industrial settings. By leveraging this technology, we aim to significantly improve the monitoring and enforcement of PPE usage, thereby reducing the risk of accidents and injuries in hazardous work environments.

Doi: https://doi.org/10.54644/jte.2025.1637

1. Introduction

According to data from the US Bureau of Labor Statistics, in 2021 [1] , the private sector reported a total of 2,607,900 workplace injury cases, with 5,190 cases linked to falls to low levels and 4,830 cases related to exposure to harmful substances or environments. Falls have consistently been the leading cause of work-related fatalities in the construction industry, accounting for 36.4% of all industry fatalities [2]. Incidents involving being struck by flying objects or equipment accounted for 15.4% of fatalities, and electrocutions for 7.2% [2]. There were 2,120 reported eye-related injuries or illnesses in 2020 [3], with material-moving workers accounting for 1,860 cases. Construction laborers represented 31.1% of construction trade workers' eye injury and illness cases, while electricians comprised 22.6% [3]. These statistics highlight the ongoing challenge of ensuring workplace safety, particularly in the construction industry. Previous research has demonstrated the effectiveness of various Personal Protective Equipment (PPE) in reducing injuries [4], [5]. However, managing PPE compliance for the substantial workforce, encompassing around 11 million employees in the construction industry and related sectors such as manufacturing, mining, and oil and gas, is a demanding and inefficiently handled task, representing a notable challenge.

In response to the current need for increased supervision and safety measures, computer vision (CV) technology is a direct and promising solution. To address the stated challenge, we have chosen YOLOv8 as the core architecture for our study. YOLOv8, the latest addition to the You Only Look Once (YOLO) family, represents an anchor-free design with thea potential to improve PPE compliance and worker

safety significantly. Alongside YOLOv8, our research introduced a new PPE-AYN dataset featuring authentic images sourced from construction sites. By harnessing YOLOv8's capacity for rapid and precise detection of the presence or absence of necessary protective gear, we have established a new YOLOv8 real-time detection model trained on our PPE-AYN dataset to ensure steadfast adherence to safety standards. Our research primarily involves integrating YOLO-based technology to assess PPE compliance and non-compliance within construction and cold food processing settings. Our ultimate goal is to reduce risks to human health and safety, mainly when accidents occur due to insufficient protective equipment from inadequate supervision. The strategic application of YOLOv8 holds the promise of transforming workplace safety, providing a proactive approach to risk mitigation while safeguarding the welfare of workers.

In a 2021 study [6], the CHV dataset was utilized and extended to create the CHVG dataset. However, focusing on enhancing the practicality of PPE detection in real-world settings, our efforts have centred on developing and modifying the CHVG dataset with an additional 1,281 images, resulting in a total of 5,074 images after augmentation. To simplify the detection process, we have consolidated four classes of colored helmets into a single "helmet" category. Furthermore, our dataset has been expanded to replicate better the complexities found in genuine construction environments. We have integrated image augmentation techniques to mimic real-world conditions, including rotation, noise, blur, and mosaic effects. Our study has demonstrated that YOLOv8 exhibits superior performance, particularly in challenging environment settings and small object detection. The results are impressive, with the YOLOv8l and YOLOv8x models achieving a remarkable mAP of 0.958, demonstrating their efficiency and accuracy in object detection. Notably, these models excel in challenging real-world scenarios, even when faced with adverse conditions such as blurred images and complex backgrounds. While the study highlights the challenges in accurately detecting the "glove" class, it underscores the YOLOv8l model's resilience and adaptability, making it a compelling choice for applications for PPE detection, for the urge of enhancing worker safety.

The continuation of this paper is as follows: In Section 2 of this study, we delve into the realm of related work, exploring prior research and contributions in PPE detection and computer vision technology. Section 3 focuses on our methodology, where we introduce the YOLOv8 architecture and present the development of our new dataset, PPE-AYN. This section provides insights into the preparation process, including the training environments and specifications. In Section 4, we present the results of our study and offer a comprehensive evaluation of the performance and capabilities of our proposed methodology. Finally, in Section 5, we draw our analysis to a conclusion and discuss the findings, implications, and potential future directions within the domain of PPE detection and worker safety, underlining the significance of our research in the broader context.

2. Related Works

Multiple research endeavors have delved into the application of CV technology for the purpose of detecting compliance with PPE regulations. Researchers across various disciplines have sought to harness the potential of CV techniques to enhance safety in industrial settings by developing systems capable of effectively identifying and enforcing compliance with PPE requirements. These studies collectively underscore the growing significance of computer vision in addressing PPE compliance, with each contributing unique insights and innovative solutions to this critical safety concern. In 2018 [7], a study proposed the CAHD algorithm, which harnessed the power of computer vision and deep learning models to accurately detect proper hard hat utilization. The study emphasized fine-tuning neural networks to streamline computational demands while upholding precise recognition capabilities. The study yielded a significant mean Average Precision (mAP) score of 54.6%. Nonetheless, it's worth noting that CAHD primarily focuses on hard hat detection and may not fully accommodate the diverse range of Personal Protective Equipment (PPE) used across various construction sites. Moreover, the achieved mAP, while noteworthy, falls relatively short when compared to more recent algorithms in the field.

In 2021, a study [8] trained eight deep learning detectors using YOLOv5 architectures, targeting six PPE classes, with an emphasis on real-world applicability, was introduced. YOLO v5x emerged as the

top performer, achieving an 86.55% mAP, while YOLO v5s demonstrated a rapid 52 FPS on a GPU. However, the study identified some limitations that offer potential avenues for future improvements. These limitations included occasional inaccuracies in predicting helmet colors and mistaking regular green T-shirts for vests, which could be mitigated by expanding the dataset. Additionally, the used YOLO models faced challenges detecting small instances from a distance, suggesting the need for architectural adjustments. Finally, the study highlighted the potential for further PPE classes, such as masks, glasses, and gloves, to be incorporated into future work.

Another notable study from 2022 [9] introduces an extension to the CHV [6] dataset to create a new CHVG dataset. This study adopts an anchor-free training mechanism-based computer vision architecture called YOLOX-m. YOLOX-m achieves the highest mAP of 89.84%, surpassing other state-of-the-art methods. The authors are confident that the proposed model is highly suitable for real-time deployment and industrial applications. However, it's worth noting that false detections still occur when dealing with smaller objects and reducing the image size.

In a recent investigation carried out inhe primary focus lies on creatiismmediate detection system to identify infringements about the use of Personal Protective Equipment (PPE). Utilizing the YOLOv8 model for object detection and the versatile Django web-based user interface framework, this system's core objective is to oversee and enforce compliance with PPE regulations actively. The classification of PPE infractions into four distinct categories based on bounding box characteristics, encompassing aspects like helmets and safety vests, resulted in an impressive 82.3% average accuracy across a dataset of 230 test samples. Additionally, the system exhibited a mAP50 value of 81.6%, a precision rate of 90.3%, and a recall rate of 75.1%. It's important to note that a detection error factor is attributed to lighting and camera specifications.

Acknowledging the critical need for enhanced precision in PPE detection across a broad spectrum of work environment conditions like blurring due to rain or steam and haze or noise from the surveillance system, and the need for more variety of objects in detection, this study introduces an innovative model powered by YOLOv8, with 5 classes of objects that can cover the need for a supervisor among all the basic settings. This model can be readily integrated into various surveillance systems or web-based applications tailored to the specific needs of users. YOLOv8 represents one of the latest advancements in deep-learning object identification models within the YOLO series [10], [11], [12], [13], [14], [15], [16], [17] emphasizing speed, size, and accuracy, distinguishing itself from prior versions.

3. Methodology

3.1. Baseline method YOLOv8 Algorithm

YOLOv8, released in January 2023 by Ultralytics, introduced anchor-free detection, reducing box predictions and speeding up non-maximu m suppression (NMS). It adopted mosaic augmentation but disabled it for the last ten epochs. YOLOv8 offers command-line and PIP package support, multiple integrations, and five scaled versions of n (nano), s (small), m (medium), l (large), and x (extra). The YOLOv8 architecture mainly consists of a backbone, neck, and head, as shown in Figure 1 [18].

3.1.1. Backbone

YOLOv8 employs a modified CSPDarknet53 [12] as its backbone network, and it undergoes a down-sampling process five times to produce five distinct scale features labelled B1 to B5. The configuration of the backbone network is visually represented in Figure 1a. Instead of the original Cross Stage Partial (CSP) module, YOLOv8 incorporates the C2f module, whose structure is illustrated in Figure 1f (where 'n' represents the number of bottlenecks). The C2f module introduces a gradient shunt connection to enhance the flow of information within the feature extraction network while keeping the network lightweight.

Additionally, the CBS module performs a sequence of operations, starting with convolution on the input data, followed by batch normalization, and culminating in activating the information flow using the SiLU function to yield the output result, as delineated in Figure 1g. The backbone network integrates the spatial pyramid pooling fast (SPPF) module to create a fixed-size output to pool the input feature maps. This differs from the structure of spatial pyramid pooling (SPP) [19] by reducing computational

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

overhead and lowering latency, accomplished by sequentially linking three maximum pooling layers, as depicted in Figure 1d.

3.1.2. Neck

Taking inspiration from PANet [20], YOLOv8 has been structured with a PAN-FPN architecture in the neck, as illustrated in Figure 1b. Compared to the neck structure of YOLOv5 and YOLOv7 models, YOLOv8 introduces a notable change by eliminating the convolution operation after up-sampling within the PAN structure. This adjustment manages to preserve the model's original performance while significantly reducing its weight. In the PAN structure, two distinct scales of features are labelled P4-P5 and N4-N5, while in the FPN structure, they are denoted as F4-F5. Traditional FPN approaches employ a top-down method to convey deep semantic knowledge. While FPN effectively enhances the semantic content of features by combining B4-P4 and B3-P3, it does have a drawback in that it may lose some object localization details in the process. To address this issue, PAN-FPN incorporates PAN into FPN. This addition significantly improves the model's ability to learn location-specific information by fusing P4-N4 and P5-N5 to facilitate top-down path enhancement. PAN-FPN effectively constructs a network structure combining top-down and bottom-up approaches, achieving a harmonious synergy between shallow positional data and deep semantic insights through feature fusion. This results in greater feature diversity and completeness in the model.

In the detection component of YOLOv8, a distinctive decoupled head structure is adopted, as showcased in Figure 1e. This specialized decoupled head structure employs two distinct branches to handle object classification and the prediction of bounding box regression. Notably, separate loss functions are applied to these two types of tasks. For the classification task, the binary cross-entropy loss (BCE Loss) is utilized, while the prediction of box bounding regression tasks makes use of distribution focal loss (DFL) [21] and CIoU (X. Li et al., 2020).

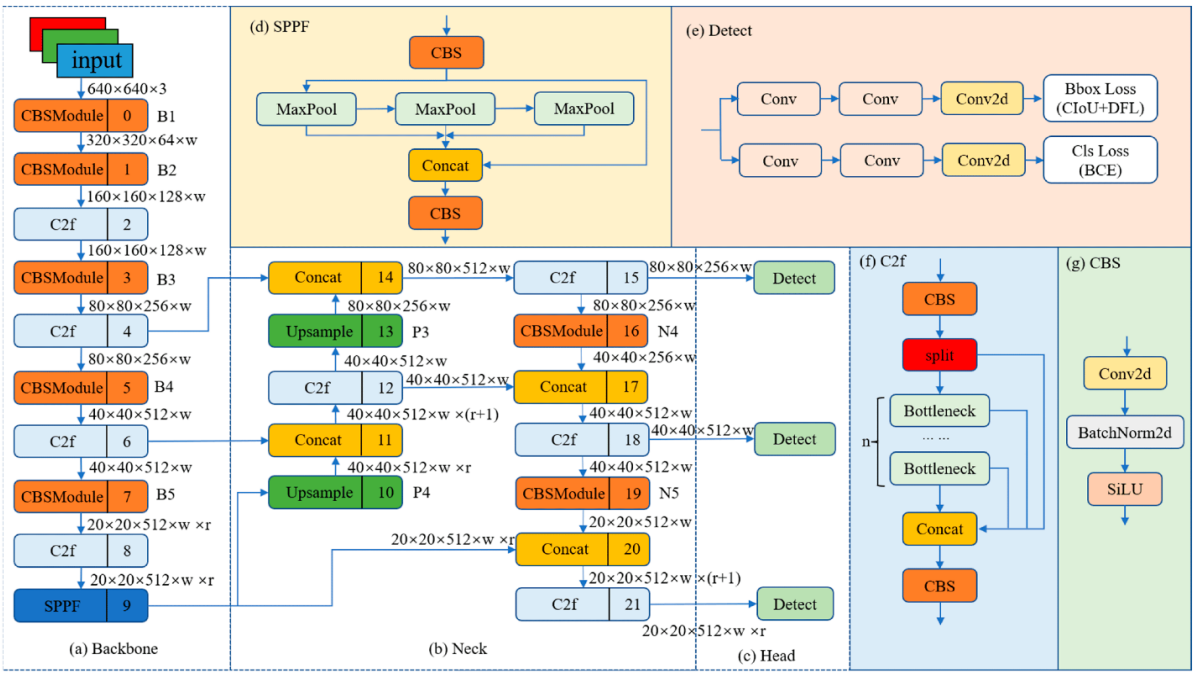

Figure 1. *YOLOv8 structure.*

Implementing this detection structure yields substantial enhancements in detection accuracy and expedites the convergence of the model. It's imperative to recognize that YOLOv8 is characterized as an anchor-free detection model that precisely defines positive and negative samples. Furthermore, it incorporates the Task-Aligned Assigner [22] for the dynamic assignment of samples, notably elevating the model's detection accuracy and overall resilience.

3.2. Dataset preprocessing process

To propose a reliable and high-quality dataset for detection models, a thorough preprocessing process was followed.The initial phase of dataset processing began with the selection of a batch of 16 images from the CHVG dataset for a thorough review. During this examination, it was observed that one image lacked proper annotations for the 'Helmet', 'Head', 'Glove', and 'Glasses' classes, as demontrated in Figure 3. To maintain the integrity of the dataset, all previously assigned labels for this image, and any potentially inconsistent annotations within the batch, were removed for throughout annotation process.

Specific annotation rules were established to maintain uniformity across the dataset. For each object, the bounding box was drawn tightly around the object, ensuring that the edges closely aligned with the visible boundaries of the object without cutting off any part of it. The center of the bounding box was consistently positioned at the geometric center of the object. In cases where objects overlapped, careful attention was given to ensure that each object was distinctly labeled, even when boundaries were partially obscured. These rules aimed to provide precise and consistent object representation in the dataset.

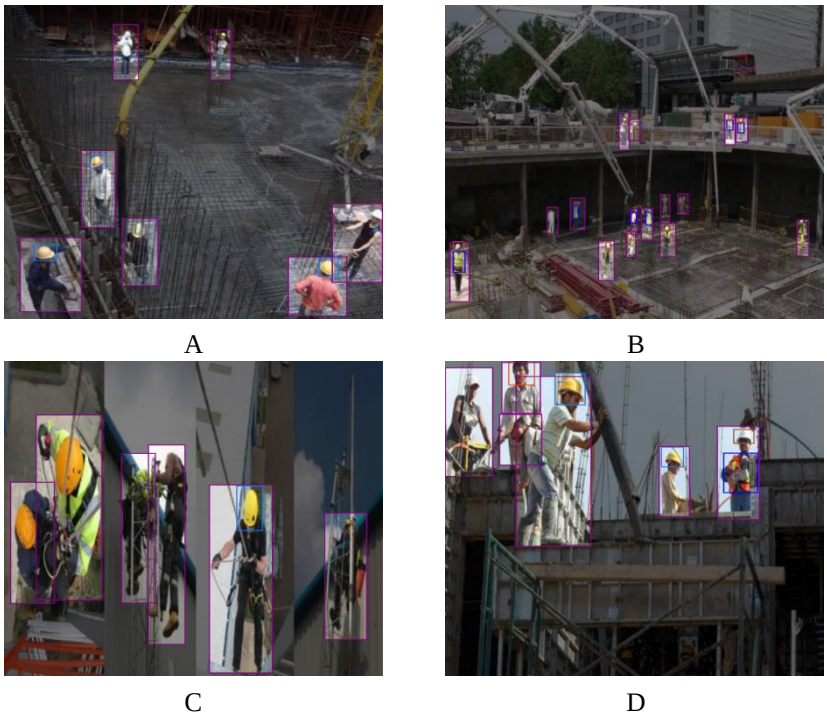

Figure 2. *Images with missing annotations in CHVG. Missing annotations for the 'Helmet' and 'Head' classes are shown in (A), (B), (C), (D). Missing annotations for the 'Vest' class are shown in (B), (C).*

Before proceeding with automatic annotation, all images were resized to a fixed resolution of 640x640 pixels. This resizing step was critical in standardizing the dataset, ensuring uniformity in image dimensions for model training, and optimizing the detection model's performance. Following the resizing process, each image underwent automatic orientation correction. This step was implemented to prevent unintentional flipping or rotation of images during subsequent processing stages, which could otherwise introduce biases or errors in the model's learning. By ensuring that all images were properly oriented, we minimized the likelihood of mislabeling objects due to image distortion.

With the images resized and oriented correctly, the DINO model, known for its robustness in self-supervised learning and object detection, was employed for automatic labeling. A confidence threshold of 35% per class was applied to ensure that only predictions with a high degree of certainty were retained. This threshold was strategically chosen to strike an optimal balance between precision and recall, capturing a sufficient number of positive samples while minimizing the risk of false positives.

Once the DINO model had completed the automatic labeling, a comprehensive manual verification process was undertaken. Although automated labeling provides a strong foundation, it is not immune to errors, especially in complex environments with small or overlapping objects. Therefore, each image and its annotations were meticulously reviewed by human annotators to ensure accuracy. This manual rechecking phase was crucial for refining the dataset, as it allowed for the correction of any erroneous labels or missed objects, thus enhancing the overall quality and reliability of the annotations.

3.3. Dataset argurementations

To further enrich and expand the PPE-AYN dataset, we applied a range of augmentations using Roboflow as the primary tool for this process. These augmentations were designed to not only simulate real-world conditions but also to create an enlarged dataset, providing additional variety and complexity for training. This approach aimed to improve the model's adaptability to diverse, practical scenarios where Personal Protective Equipment (PPE) detection is crucial.

One of the key augmentations involved introducing controlled noise, affecting approximately 5% of the pixels in each image. This was implemented to imitate environmental factors such as dust, steam, or debris, which can obscure objects in real-world industrial environments. Additionally, brightness levels were adjusted within a range of -30% to +30%, simulating the variable lighting conditions often encountered in outdoor settings and surveillance footage. These adjustments accounted for dynamic changes in sunlight and shadows, as well as varying indoor lighting conditions typical of factory settings.

Mosaic augmentation was also applied, which involved combining multiple images into a single composite. This technique not only enriched the dataset but also replicated scenarios where objects overlap or are partially visible—common occurrences in real-life PPE compliance checks. The augmentation process led to the creation of an enlarged and more diverse PPE-AYN dataset, consisting of a total of 5,074 images. This enhanced dataset ensured that machine learning models trained on it would be better equipped to handle the complexities of real-world applications, making the models more robust and versatile for PPE detection tasks.

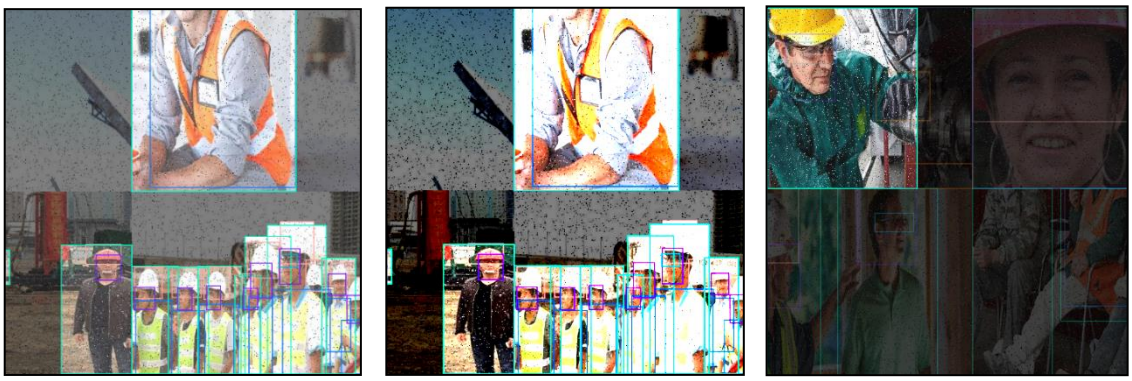

Figure 3. *Image noise, mosaic, bright and dim light are applied during data argurementation process.*

4. Experimental

4.1. Dataset collection

In an earlier study [9], a comparison between the CHVG and CHV datasets revealed some similarities, with the CHVG dataset featuring 369 more images than the CHV dataset. However, the CHVG dataset introduced two additional classes, specifically safety glass and a head without a hard hat, a decision driven by the recognition that machine learning thrives on challenges presented by an increased number of classes. The CHVG dataset comprised 1,699 images, each accompanied by corresponding annotations. The dataset encompassed a total of 11,604 objects. In order to meet the demand of complexity and corresponding in real-time detection, we have taken a significant step forward by introducing a PPE-AYN dataset, a considerably expanded version. We've streamlined the hard hat category into a single 'hat' class to facilitate more versatile general hat detection. Additionally, we've introduced a new 'glove' class to enhance the comprehensiveness of the dataset, with detailed class

distribution and additional information provided in an upcoming table. To further enrich our dataset, we've incorporated 1,281 images, extending beyond the original CHV dataset. This expansion underscores our commitment to addressing the pressing need for enhanced PPE detection capabilities in various industrial settings, providing a robust foundation for our study's objectives. The dataset detail is illustrated in Table 1 below.

This dataset comprises 2,980 images, each richly annotated with 16,673 annotations spread across six distinct classes. With an average of 5.6 annotations per image, it offers a comprehensive foundation for PPE detection tasks. The dataset's images consistently maintain an average size of 0.41 megapixels, and they all fall within this same size range. The predominant image resolution within the dataset is 640x640 pixels, making it a valuable resource for various computer vision applications, particularly those related to PPE detection.

Table 1. *PPE-AYN Dataset description*

Class	Number of objects
Person	4,960
Hat	3,798
Head	2,540
Gloves	2,385
Vest	2,198
Glasses	792

4.2. Experimental setup

We began by configuring a Python environment, with a preference for Google Colab for its cloud-based resources. This provided the foundational infrastructure for our study, supported by the installation of critical packages, including PyTorch and CUDA. The hardware platform implemented in this study provided by Google Colab Pro and the environment parameters used during the training phase are shown in Table 2.

Table 2. *Training and testing platform descriptions*

Component	Specification
CPU	Intel(R) Xeon(R) 2.30GHz
GPU	Tesla V100-SXM2
GPU memory size	16 GB
Memory	13 GB
Disk	167 GB
Deep learning architecture	Ultralytics YOLOv8.0.200 + Python-3.10 + torch-2.1.0+cu118

The dataset was divided into three subsets: 70% for training, 20% for validation, and 10% for testing on unseen data, ensuring comprehensive evaluation on previously unlearned images. We trained YOLOv8 models, including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x, alongside EfficientDet, and Detectron. The training process involved fine-tuning key hyperparameters to optimize model performance. Specifically, we adjusted the learning rate to 0.00001, set Non-Maximum Suppression (NMS) to 0.938, and applied a weight decay of 0.00001. All models were trained for 100 epochs, with the mosaic augmentation technique disabled during the last 10 epochs. This adjustment allowed the models to focus on refining their ability to generalize without the additional complexity introduced by mosaic transformations, ensuring robust performance across a variety of detection scenarios.

This fine-tuning ensured that the models achieved optimal performance, with adjustments made throughout the training process based on validation results. The models were then rigorously tested on the remaining 10% of the dataset, providing an accurate evaluation of their ability to detect PPE in real-world scenarios. The combination of carefully chosen parameters and dataset splitting ensured robust model training and testing, setting the foundation for accurate and reliable performance metrics.

Here, APi refers to the AP value for category index i, and N represents the total number of categories in the training dataset (in our case, N is 10). Additionally, we computed mAP0.5 and mAP0.5:0.95, which denote the average accuracy at different IoU thresholds, with the latter spanning a range from 0.5 to 0.95 at 0.05 intervals.

4.3. Evaluation Indicators

We employed a set of evaluation metrics to assess our model's detection capabilities. These metrics included precision, recall, mean average precision (mAP) [23] at IoU [24] thresholds of 0.5 and 0.5:0.95, the number of model parameters, model size, and detection speed, . In the context of these metrics, we utilized specific parameters, such as True Positives (TP), False Positives (FP), and False Negatives (FN), to gauge the model's performance. These metrics were instrumental in quantifying the models' accuracy and effectiveness. The results, as depicted in Tables 4 and 5, make use of the mAP value, Precision and Recall, as illustrated in Figure 3. These metrics serve to emphasize the competence of our models and provide a comprehensive overview of their performance.

Precision (P), a key metric, quantifies the ratio of positive samples correctly predicted by the model to all detected samples. It is calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

Recall (R), that another essential metric expresses the ratio of positive samples correctly predicted by the model to the total positive samples in existence:

$$R = \frac{TP}{TP + FN} \tag{2}$$

Average Precision (AP) is a measure derived from the area under the precision-recall curve, signifying the model's precision-recall performance:

$$AP = \int_0^1 P(R)d(R) \tag{3}$$

Mean Average Precision (mAP) is an aggregate metric that takes the weighted average of AP values across all sample categories. It provides a comprehensive evaluation of the model's performance across all categories, and it is calculated as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

4.4. Experimental Results

In addition to YOLOv8, we also trained and tested the EfficientDet and Detectron models. Upon completing the training and evaluation across two datasets, the results are summarized in Table 4. The YOLOv8l and YOLOv8m models consistently outperformed all others, achieving the highest overall mAP scores for all classes. Specifically, on dataset (1), the YOLOv8l and YOLOv8m models achieved a total mAP of 0.925, while on dataset (2), they reached a total mAP of 0.924.

In dataset (1), the "yellow" class achieved an exceptional mAP score of 0.952 with the YOLOv8x model. Similarly, in dataset (2), the "person" class achieved a notable mAP score of 0.964, also with the YOLOv8x model. These results highlight the robust performance of YOLOv8l and YOLOv8m models, particularly in achieving high mAP values across various classes, establishing them as top-performing models for detection tasks.

Table 3. *YOLOv8 performance on different datasets. 1: CHVG dataset; 2: PPE-AYN dataset*

Model	red	white	yellow	blue	glass	gloves	hat	head	person	vest	All
1-YOLOv8n	0,883	0,902	0,942	0,907	0,79	_	_	0,905	0,953	0,925	0,901
2-YOLOv8n	_	_	_	_	0,854	0,88	0,932	0,819	0,954	0,929	0,895
1-YOLOv8s	0,908	0,902	0,951	0,923	0,859	_	_	0,928	0,961	0,913	0,918
2-YOLOv8s	_	_	_	_	0,896	0,911	0,937	0,853	0,961	0,942	0,917
1-YOLOv8m	0,922	0,91	0,938	0,929	0,878	_	_	0,932	0,958	0,93	0,925
2-YOLOv8m	_	_	_	_	0,917	0,916	0,941	0,853	0,963	0,954	0,924
1-YOLOv8l	0,912	0,924	0,944	0,934	0,861	_	_	0,921	0,96	0,943	0,925
2-YOLOv8l	_	_	_	_	0,915	0,919	0,945	0,857	0,961	0,947	0,924
1-YOLOv8x	0,909	0,923	0,952	0,947	0,862	_	_	0,899	0,959	0,942	0,924
2-YOLOv8x	_	_	_	_	0,904	0,913	0,947	0,862	0,964	0,947	0,923
1-EfficientDet	0,770	0,731	0,797	0,799	0,811	-	-	0,835	0,807	0,897	0,805
2-EfficientDet	-	-	-	-	0,821	0,791	0,899	0,812	0,812	0,914	0,846
1-Detectron	0,874	0,887	0,815	0,897	0,859	-	-	0,892	0,856	0,864	0,868
2-Detectron	-	-	-	-	0,904	0,901	0,912	0,903	0,89	0,892	0,902

For further investigating, training was conducted on the arguremented PPE-AYN dataset—designed to capture real-life industrial scenarios. The training process on this dataset was particularly crucial as it reflected real-world challenges faced in construction, manufacturing, and high-risk environments where accurate PPE detection is critical for safety. To further assess the robustness of the models, dataset 2 was augmented with challenging transformations, such as applying 2.5x blur to simulate degraded image quality, 15-degree rotations along the X and Y axes to mimic different camera orientations, and brightness adjustments ranging from -30% to +30% to reflect dynamic lighting conditions often found in both indoor and outdoor industrial settings. Mosaic augmentation, which involves combining multiple images into one, was also employed to replicate real-world scenarios where objects may overlap or be partially obscured. These augmentations ensured that the models were exposed to a wide variety of conditions, allowing them to generalize better when detecting PPE in real-life settings.

Table 4. *Performance of YOLOv8 in the noise, blur, rotation, and mosaic*

Model	glass	gloves	hat	head	person	vest	all
YOLOv8n	0,855	0,935	0,953	0,868	0,967	0,96	0,923
YOLOv8s	0,907	0,962	0,961	0,9	0,978	0,971	0,947
YOLOv8m	0,926	0,968	0,966	0,912	0,979	0,979	0,955
YOLOv8l	0,929	0,968	0,97	0,918	0,979	0,98	0,958
YOLOv8x	0,926	0,97	0,971	0,92	0,978	0,981	0,958

The F1-Confidence curves demonstrated in Figure 3 provide valuable insights into the performance of the models. As depicted across all the classes—glass, gloves, head, hat, person, and vest—the models generally show an increase in the F1-score as confidence increases, reaching an optimal point before sharply declining. This trend suggests that as the model becomes more confident, it initially performs better in balancing precision and recall. However, when the confidence threshold exceeds the optimal range, recall diminishes, leading to a decrease in the F1-score. Notably, the "person" class achieves consistently higher F1-scores across the models, indicating its reliable detection across different confidence levels. On the other hand, classes such as "gloves" and "glasses" show a relatively lower peak F1-score, suggesting more challenges in detecting these objects accurately. These curves demonstrate the models' ability to perform optimally within a specific confidence threshold, beyond which their ability to maintain a balance between precision and recall declines.

In the Precision-Recall curves illustrated in Figure 4, a similar evaluation is presented by plotting precision against recall, further highlighting the model's performance across different classes. The curves illustrate that the "person" class continues to outperform others, maintaining a high precision and recall across most thresholds. However, classes like "gloves" and "glasses" experience a slight drop-off in precision at higher recall levels, indicating some degree of false positives when the model attempts to capture a higher number of true positives. Overall, these curves reflect the robustness of the models across the majority of classes, with "person" being the most reliably detected, while some variability is observed in the detection of smaller or less distinguishable objects such as gloves and glasses.

The YOLOv8 testing results are shown in Table 5, all models achieve impressive average mAP scores exceeding 0.92. Notably, the YOLOv8x model consistently demonstrates superior efficiency by attaining higher accuracy than the other models across all classes, with a mAP reaching 0.958, which it shares with YOLOv8l. This highlights these models' effective object detection capabilities, even in suboptimal conditions. However, it is worth noting that despite achieving higher mAP scores and confidence levels, there are instances where YOLOv8x and other models struggle to detect objects like gloves, presenting a new challenge. In contrast, YOLOv8l successfully detect the "gloves" class, establishing its suitability among all models while maintaining the same average mAP as YOLOv8x and consistently performing well across different classes.

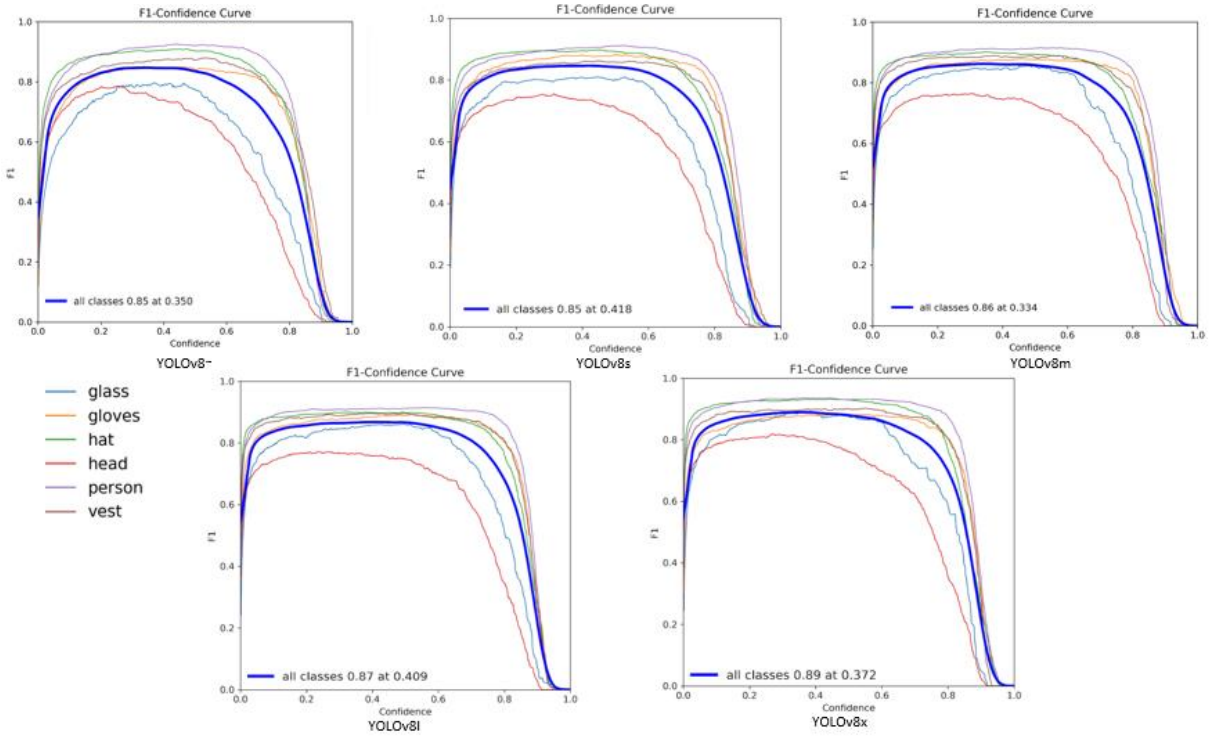

Figure 4. *F1- Confidence curve results of YOLOv8.*

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

While all models are proficient in detecting most object classes with high confidence, as illustrated in Figure 5, the "glove" class posed a unique challenge, resulting in missed detections across the board. However, the YOLOv8l model shone as the exception, successfully detecting all classes, including the elusive "glove" class. This exceptional performance underscores the "YOLOv8l" model's adaptability and resilience in challenging image conditions, positioning it as a compelling choice for applications where robust object detection in adverse scenarios is imperative.

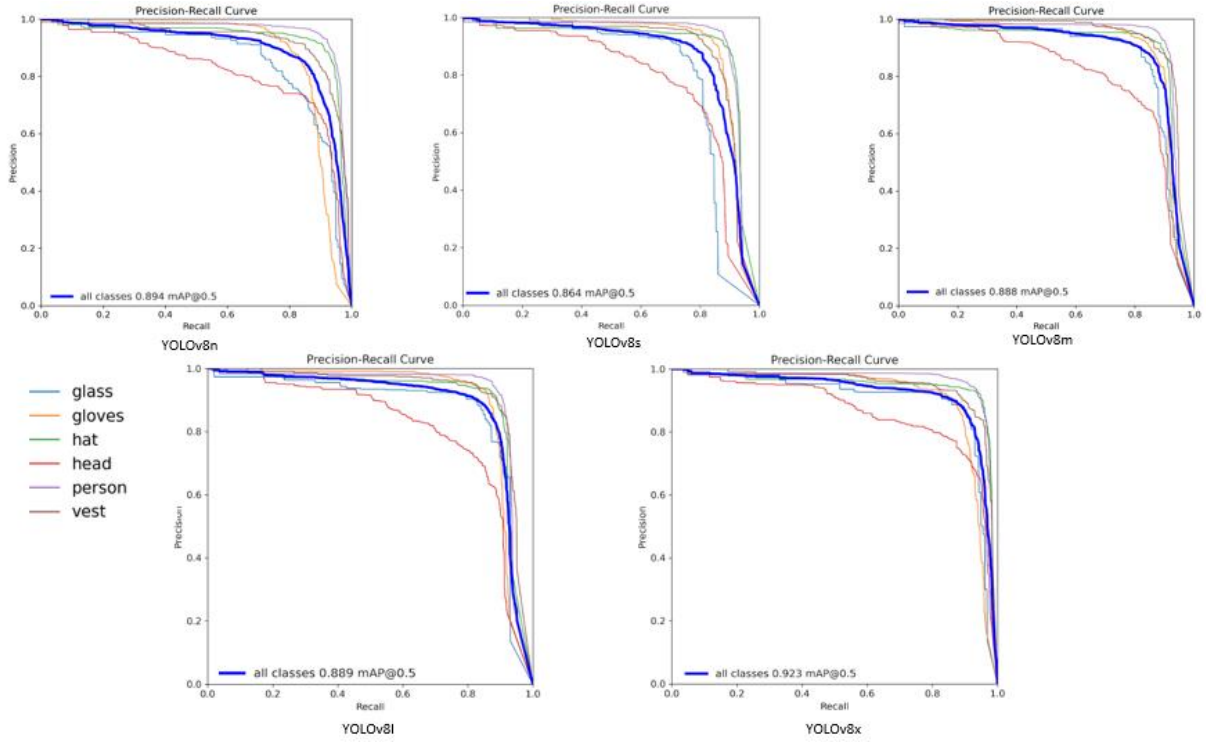

Figure 5. *Percision-Recall curve results of YOLOv8.*

Figure 6. *Testing result of YOLOv8. (a) YOLOv8n, (b) YOLOv8s, (c) YOLOv8m, (d) YOLOv8l, (e) YOLOv8x.*

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

A closer examination revealed that when objects were positioned at far-away distances, as illustrated in Figure 5, the YOLOv8l model's performance experienced a decline, especially in the “glove” class. This decline can be attributed to the inherent challenges of recognizing objects with reduced visibility and detail at far distances. This contrast in performance can be attributed to the nature of the "glove" class itself, which may present inherent complexities, such as varying textures and colors, that render it more challenging to identify accurately in complex backgrounds. Nonetheless, the overall exceptional performance of the YOLOv8l model in navigating diverse real-world scenarios reinforces its suitability for applications where robust object detection across varying distances and intricate backgrounds is paramount, notwithstanding the inherent complexity of certain object classes.

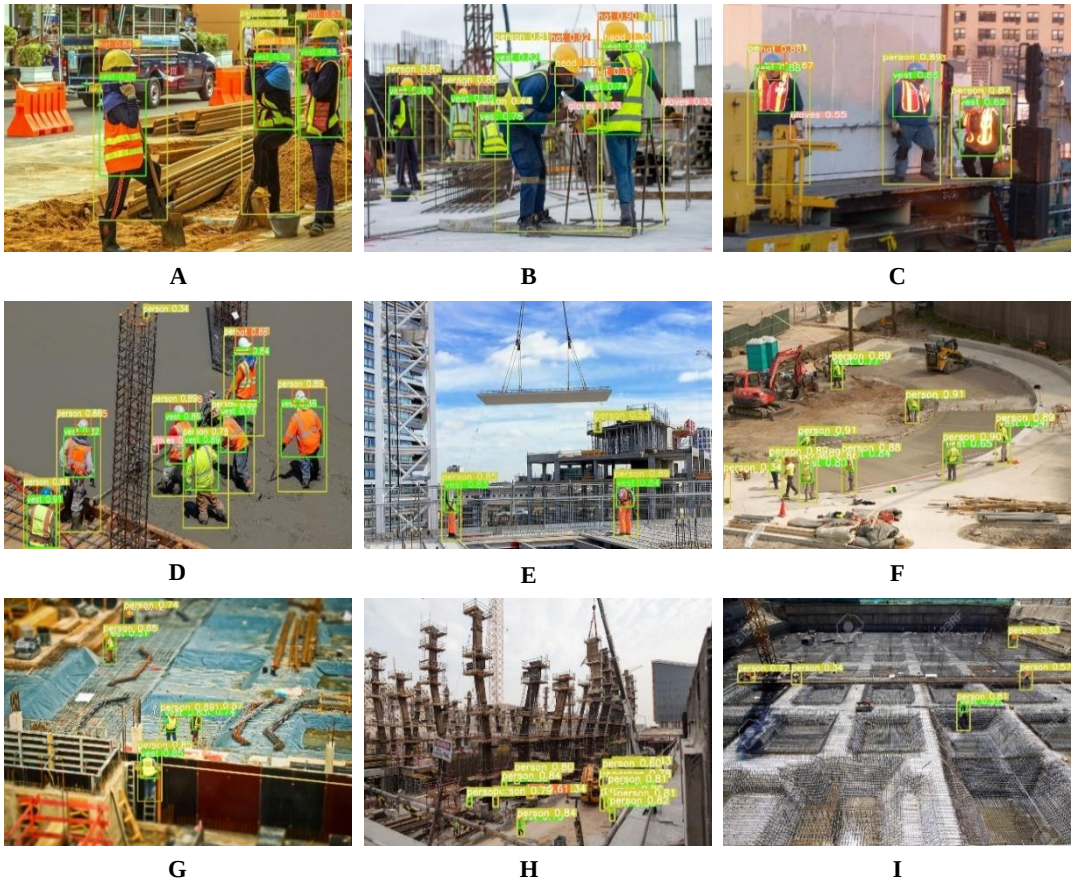

Figure 7. *Testing result of YOLOv8l in different background and distances. (A, B, C) Objects in close distance, (D, E, F) Objects in medium distance, (G) Objects in far distance, (H, I) Object in far distance with complex background.*

5. Conclusion and Discussion

5.1. Conclusion

In conclusion, evaluating the YOLOv8x, YOLOv8l, and YOLOv8m models on our new PPE-AYN dataset has demonstrated their remarkable efficiency and high accuracy in object detection. Notably, YOLOv8l and YOLOv8x outperformed other models with an impressive mAP of 0,958.

Furthermore, all models successfully detected most object classes in challenging practical scenarios, including blurred images. All models successfully detected all classes, but on the other hand, the "glove" class posed a unique challenge, in exception, YOLOv8l has successfully detected this class. This adaptability highlights the proficiency of daset as well as it varieties in challenging image conditions.

5.2. Discussion

While YOLOv8 models performed exceptionally well, especially YOLOv8l and YOLOv8x, challenges remain in detecting smaller objects like gloves and distant objects in complex environments. To address these issues, future research will focus on modifying the model architecture by adding additional CNN layers to improve the detection of small objects and removing redundant layers to enhance speed without compromising accuracy. We will also explore incorporating techniques like Fuzzy Logic to handle uncertainties in object detection, particularly in challenging conditions.

Moreover, advanced techniques such as attention mechanisms and feature fusion will be considered to help the model focus on critical regions in an image. However, there is a necessary trade-off between model performance and efficiency, as adding layers may slow down inference, while removing layers could reduce accuracy. Our future work will aim to optimize this balance, ensuring that the model remains both efficient and robust for real-time PPE detection in industrial settings.

REFERENCES

[1] "IIF Latest Numbers : U.S. Bureau of Labor Statistics." Accessed: Oct. 31, 2023. [Online]. Available: https://www.bls.gov/iif/latest-numbers.htm
[2] "Construction Statistics | NIOSH | CDC." Accessed: Oct. 31, 2023. [Online]. Available: https://www.cdc.gov/niosh/construction/statistics.html
[3] "Workers suffered 18,510 eye-related injuries and illnesses in 2020 : The Economics Daily: U.S. Bureau of Labor Statistics." Accessed: Oct. 31, 2023. [Online]. Available: https://www.bls.gov/opub/ted/2023/workers-suffered-18510-eye-related-injuries-and-illnesses-in-2020.htm
[4] D. Hardison, A. Dickerson, B. Sylcott, and K. Lee, "Evaluating the Effectiveness of Worker Safety Vests on Drivers' Visual Attention," in *Construction Research Congress 2020*, pp. 105–113, doi: 10.1061/9780784482872.012.
[5] A. Hulme, M. Nigel, and G. A, "Industrial head injuries and the performance of helmets," Sep. 1995.
[6] M. Ferdous and S. M. M. Ahsan, "PPE detector: a YOLO-based architecture to detect personal protective equipment (PPE) for construction sites," *PeerJ Comput. Sci.*, vol. 8, p. e999, Jun. 2022, doi: 10.7717/peerj-cs.999.
[7] Z. Xie, H. Liu, Z. Li, and Y. He, "A convolutional neural network based approach towards real-time hard hat detection," in *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2018, pp. 430–434, doi: 10.1109/PIC.2018.8706269.
[8] M. I. B. Ahmed *et al.*, "Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach," *Sustainability*, vol. 15, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/su151813990.
[9] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, "Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches," *Sensors*, vol. 21, no. 10, p. 3478, 2021, doi: 10.3390/s21103478.
[10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *CoRR*, vol. abs/1506.02640, 2015, [Online]. Available: http://arxiv.org/abs/1506.02640
[11] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
[12] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *CoRR*, vol. abs/1804.02767, 2018, Accessed: Oct. 31, 2023. [Online]. Available: http://arxiv.org/abs/1804.02767
[13] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *CoRR*, vol. abs/2004.10934, 2020, Accessed: Oct. 31, 2023. [Online]. Available: https://arxiv.org/abs/2004.10934
[14] "Comprehensive Guide to Ultralytics YOLOv5 - Ultralytics YOLOv8 Docs." Accessed: Oct. 31, 2023. [Online]. Available: https://docs.ultralytics.com/yolov5/#citation
[15] C. Li *et al.*, "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," Sep. 07, 2022, doi: 10.48550/arXiv.2209.02976.
[16] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 06, 2022, doi: 10.48550/arXiv.2207.02696.
[17] G. Jocher, A. Chaurasia, and J. Qiu, *YOLO by Ultralytics*. (Jan. 2023). Python. Accessed: Oct. 31, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[18] G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, "UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios," *Sensors*, vol. 23, p. 7190, 2023, doi: 10.3390/s23167190.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *CoRR*, vol. abs/1406.4729, 2014, Accessed: Oct. 31, 2023. [Online]. Available: http://arxiv.org/abs/1406.4729
[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *CoRR*, vol. abs/1803.01534, 2018, Accessed: Oct. 31, 2023. [Online]. Available: http://arxiv.org/abs/1803.01534
[21] X. Li *et al.*, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," *CoRR*, vol. abs/2006.04388, 2020, Accessed: Oct. 31, 2023. [Online]. Available: https://arxiv.org/abs/2006.04388
[22] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned One-stage Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3490–3499, doi: 10.1109/ICCV48922.2021.00349.
[23] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit," *Electronics*, vol. 10, no. 3, 2021, doi: 10.3390/electronics10030279.
[24] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," presented at *the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society*, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.

Vo Thanh Xuan Le is studying in Computer Science from Information Technology Faculty in Ho Chi Minh University of Technology, Ho Chi Minh City since 2021. His research interests are focused on Computer vision and Data mining.
Email: thanhxuan8054@gmail.com. ORCID: https://orcid.org/0009-0001-9957-3188

Khac-Chien Nguyen received the master degree in Computer Science from the University of Natural Sciences - Vietnam National University, Ho Chi Minh City in 2009, and his PhD in Communication Engineering from the Post and Telecommunication Institute of Technology Hanoi in 2019. His research interests include: Cloud computing and Data mining.
Email: nkchienster@gmail.com and nk.chien@hutech.edu.vn. ORCID: https://orcid.org/0009-0008-1035-3359