Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

Optimizing YOLO11 for Dense Crowd Counting under Severe Occlusion via Head-Detection Fine-Tuning

Joko Sutrisno*[1], Sri Winarno[2], Affandy[3]

[1,2,3]Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
[1]Indonesian Agency for Meteorology Climatology and Geophysics (BMKG), Indonesia

Email: [1]jokosutrisno96837@gmail.com

Received : Feb 2, 2026; Revised : Mar 6, 2026; Accepted : Mar 7, 2026; Published : Apr 18, 2026

Abstract

Accurate and real-time people counting is essential for crowd management and public safety, yet achieving precision in high-density environments remains a challenge due to severe visual occlusion. While the recently released YOLO11 architecture introduces advanced features such as C3k2 and C2PSA modules, its performance as a pre-trained model for people counting tasks has not been fully explored. This study evaluates the efficacy of a head-detection-based fine-tuning strategy using the YOLO11 model, compared against the default pre-trained baseline. The fine-tuning performance is analyzed across three distinct scenarios: S1 (full fine-tuning at 960 pixels), S2 (partial backbone freezing at 960 pixels), and S3 (partial freezing at 640 pixels). The fine-tuning process was conducted using the CC_Mach_1 dataset from Roboflow Universe, which consists of high-density images annotated for head detection. The results demonstrate that the baseline pre-trained YOLO11, which relies on full-body features, exhibits extremely limited performance with an mAP@0.5 of 0.017 and a Mean Absolute Error (MAE) of 100.3. In contrast, the fine-tuned scenarios achieved substantial improvements, led by S1 which reached the highest accuracy with an mAP@0.5 of 0.682 and reduced the MAE by 62% to 37.8. While S2 remained highly competitive with an MAE of 39.6, the performance in S3 declined to 46.9, confirming that lower input resolutions limit the model's ability to identify small-scale head features. These findings provide empirical evidence that domain-specific fine-tuning for head detection substantially improves the robustness of YOLO11 against occlusion. Beyond technical accuracy, this detection-based approach offers a more computationally efficient alternative to traditional density-map-based methods, making it highly suitable for deployment in real-time surveillance systems for large-scale public monitoring.

Keywords : *Crowd counting, fine-tuning, head detection, visual occlusion, YOLO11.*

1. INTRODUCTION

Accurate and real-time people counting plays a strategic role across various sectors, including security surveillance, crowd management, transportation, and risk mitigation at public events. Precise information on human counts facilitates the early detection of potential hazards and supports rapid decision-making during surges in density [1], [2]. However, achieving high accuracy in crowded environments remains a significant technical challenge, primarily due to visual occlusion, object overlapping, and scale variations [3].

In recent years, crowd counting has been addressed through two major paradigms: density-estimation (density map) methods and detection-based methods. Density-map approaches, such as MCNN [4] and CSRNet [5], estimate per-pixel density and integrate it to produce global counts, which is often effective in extremely dense scenes [4], [5], [6]. Recent surveys report that density estimation remains a dominant direction, while also highlighting persistent issues such as sensitivity to domain shift, annotation design choices, and limited instance-level interpretability for downstream tasks [3], [6], [7], [8], [9], [10]. These limitations motivate continued investigation of detection-based pipelines, particularly when real-time inference and instance-level localization are required in operational settings.

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

Advances in computer vision technology based on deep learning have provided more effective solutions compared to conventional methods. Single-stage object detection models, such as the You Only Look Once (YOLO) family, have become the preferred choice due to their optimal balance of high accuracy and real-time inference speed. Recent literature highlights that the YOLO approach achieves robust performance across various domains, including health protocol monitoring and pedestrian tracking [11], [12], [13], [14]. Despite these successes, standard YOLO variants still exhibit limitations in high-occlusion environments. YOLOv3 often suffers from accuracy degradation in overlapping scenarios, while YOLOv5 and YOLOv8 show instability at extreme densities [15], [16]. In addition, crowded-scene detection is intrinsically challenging because dense overlap increases duplicate predictions and suppression-related misses, which can reduce recall and bias counting results [17], [18].

The release of YOLO11 marks a significant milestone in object detection architecture. Murat and Kiran [19] identify that YOLO11 introduces critical improvements through C3k2 and C2PSA layers, achieving a superior balance between speed and accuracy. He et al. [20] further demonstrate that these modules, alongside Cross-scale Pixel Spatial Attention, sharpen detection boundaries and enhance multi-scale feature extraction. Recent architectural analyses of YOLO11 indicate that C3k2 is designed to strengthen feature extraction efficiency, while C2PSA introduces parallel spatial attention that can improve spatial discrimination under clutter and scale variation [21]. Furthermore, recent studies that extend or adapt YOLO11 for small-object settings report measurable gains in detecting compact targets by enhancing multi-scale representation and attention-driven feature refinement. These properties are theoretically relevant for head detection in crowded scenes because heads are small, frequently partially occluded, and require strong multi-scale cues for consistent localization [3], [22], [23], [24]. Consequently, YOLO11 provides a strong foundation for exploration in more specialized fields like crowd counting.

However, advanced architecture alone may not fully overcome physical occlusion in congested scenes. Research by Ali et al. [25] and Hassan et al. [26] proves that head-detection-based approaches are more resilient to overlapping than full-body detection. By focusing on head features, researchers achieved accuracy levels up to 95.6% even in situations where human bodies were heavily obscured [25]. Hassan et al. [26] also utilized transfer learning across multiple datasets to improve the detection of small-scale head objects. The emergence of recent head-focused benchmarks further supports the view that robust head detection is a key building block for reliable crowd analytics under occlusion [27]. In addition, recent work has explored head counting using newer detector generations, indicating that head-based detection remains an active direction for improving counting robustness in crowded scenes [28].

Other optimization strategies have been explored to handle the complexity of crowded backgrounds. Khel et al. [29] developed a Hybrid YOLOv4 architecture that integrates a Convolutional Block Attention Module (CBAM) to strengthen relevant feature weights and mitigate invalid data. Similarly, Alhawsawi et al. [13] introduced a Context Enrichment Module (CEM) to capture multi-scale information in complex drone-based imagery. These efforts aim to expand the receptive field and preserve spatial resolution for small target localization. Such developments highlight the ongoing need for architectural adaptations to handle high-density scenarios effectively.

Despite these advancements, a critical research gap remains regarding the application of the latest YOLO iterations. While the effectiveness of head-detection strategies has been validated on older models like YOLOv5, no studies have specifically evaluated YOLO11 for this purpose. Furthermore, as emphasized by Zhuang et al. [30], domain adaptation through transfer learning is essential for achieving optimal performance in specialized tasks. Therefore, this study evaluates the performance of YOLO11 through a comparative analysis between a baseline pre-trained model and task-adapted variant optimized for head-based detection. To strengthen novelty beyond a single pre-trained versus fine-tuned

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

comparison, this study also conducts an ablation study to identify the optimal training configuration for head-detection-based crowd counting. The experiments compare: (S0) default COCO pre-trained YOLO11s without adaptation, (S1) full-layer fine-tuning at 960×960, (S2) partial freezing of the first 10 backbone layers at 960×960, and (S3) the same freezing strategy at 640×640 to evaluate the necessity of high-resolution scaling for small-object head detection. This evaluation aims to investigate how fine-tuning strategy, freezing policy, and input resolution affect counting accuracy and detection reliability under severe occlusion conditions, using established detection metrics [31], [32], [33].

The main contributions of this study can be summarized as follows:

1. To the best of our knowledge, this study provides the first systematic evaluation of YOLO11 for head-detection-based crowd counting under severe occlusion, positioning the approach within both detection-based and density-estimation-based crowd counting literature.
2. It provides an empirical analysis of the limitations of full-body detection when applied to dense crowd scenarios using YOLO11, including the impact of overlap and suppression effects in crowded scenes.
3. It introduces a head-detection-based adaptation strategy for YOLO11 and quantifies, via ablation, the effects of freezing policy and input resolution on detection reliability and counting accuracy, without modifying the original YOLO11 architecture.

2. METHOD

Experiments were conducted using cloud-based infrastructure to handle high-performance computing requirements. The entire implementation was executed within the Google Colaboratory environment utilizing an NVIDIA Tesla T4 Graphics Processing Unit (GPU). High-capacity GPU resources are essential for accelerating the massive parallel computations required by deep learning models. The primary programming language used was Python 3.12.12, and the YOLO model was implemented via the Ultralytics framework. This combination enabled efficient model training and validation without the need for expensive local hardware investments.

2.1. YOLO11 Architectural Framework

The overall architecture of YOLO11 is shown in Figure 1. The model consists of three main subsystems, namely the Backbone, Neck, and Head. The Backbone functions as the primary feature extractor that converts raw input images into hierarchical feature representations. In YOLO11, the C2f blocks used in previous versions are replaced by C3k2 blocks, which implement a customized CSP Bottleneck with adaptive convolution kernel sizes. This modification reduces the number of parameters while preserving the semantic richness of the extracted features. The Backbone also retains the Spatial Pyramid Pooling Fast (SPPF) module to aggregate multi-scale contextual information before passing features to the next stage [19], [34].

The Neck subsystem performs feature fusion across different network depths to balance semantic and spatial information. YOLO11 integrates the C2PSA module after feature fusion to introduce a self-attention mechanism, enabling the model to emphasize important regions and capture long-range spatial dependencies. This mechanism improves detection performance for occluded objects, which are difficult to handle using standard convolution-based approaches [20], [34].

The Head subsystem produces the final predictions, including object classes and bounding box coordinates. YOLO11 adopts a decoupled, anchor-free head architecture, where classification and regression are handled by separate branches. This design improves training convergence and inference accuracy while allowing greater flexibility in detecting objects with large morphological variations, such as human heads observed from different viewing perspectives, without requiring manual anchor design [19].

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

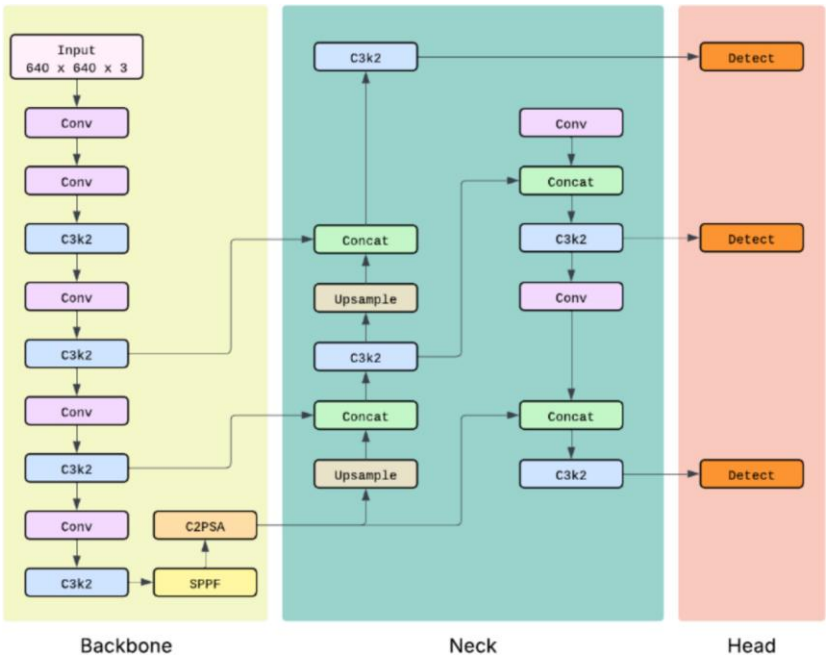

Figure 1. YOLO11 architecture [19].

2.2. Research Workflow

The data processing workflow in this study is designed systematically to ensure the validity of the experimental results. These stages are generally visualized in the flowchart shown in Figure 2. Based on the diagram, the main process consists of four stages: dataset collection and characterization, preprocessing, model training through fine-tuning, and comparative evaluation and analysis. The detailed implementation of each stage is further elaborated in the following subsections.

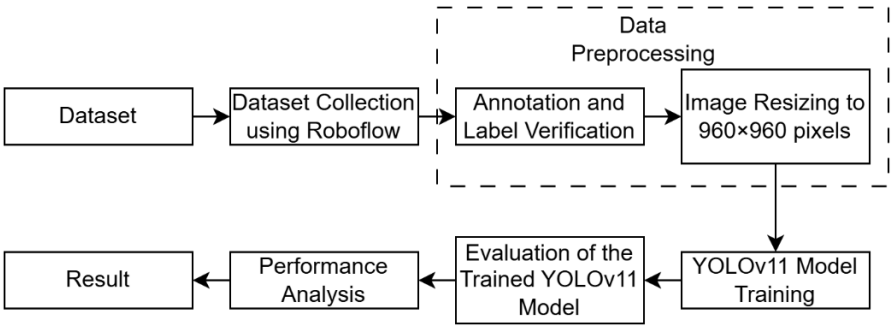

Figure 2. Research flowchart

2.3. Dataset Collection and Characteristics

The primary data source is the CC_Mach_1 dataset acquired from the Roboflow Universe public repository. The dataset was obtained in a format compatible with the YOLO11 architecture. Roboflow serves as an end-to-end data management platform providing curated imagery for computer vision development [35]. As a community-contributed dataset originally designed for human detection projects, CC_Mach_1 remains relatively unexplored in formal large-scale publications. This research serves as an early study to validate its effectiveness for crowd-counting tasks using the YOLO11 architecture [36].

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

Figure 3. Sample from the validation dataset. The green bounding box indicates the region focused on the head area

According to the metadata, the dataset contains 1,774 images, which were automatically divided into three subsets: a training set of 1,397 images, a validation set of 255 images, and a testing set. The visual characteristics include various human crowd scenarios with differing density levels. While the dataset provides a base resolution of 640 pixels, this study utilizes both the original resolution and an upscaled version at 960 x 960 pixels. This dual-resolution approach is implemented to specifically evaluate the impact of image scale on detection precision, particularly for small head objects in congested areas (see Figure 3).

2.4. Data Preprocessing

Before training, the dataset underwent a rigorous preprocessing phase. The first stage involved annotation verification and validation through visual inspection on the Roboflow platform. Manual verification was performed to ensure that all head objects were correctly labeled, with additional bounding boxes added where necessary to maintain the quality of the ground truth.

The subsequent stage involved image standardization and scaling. To facilitate a comparative analysis of the impact of visual detail, two resolution variants were prepared. The standard 640 x 640 pixels resolution was maintained for baseline and comparative ablation tests, while an upscaled target of 960 x 960 pixels was implemented for the proposed high-precision detection models. Resizing was performed using interpolation methods that maintain the aspect ratio. To prevent information loss or object distortion, letterboxing was automatically applied to images with different aspect ratios. Detailed mapping of these resolutions to specific experimental scenarios is provided in Section 2.5.

2.5. YOLO11 Training and Fine-Tuning Configuration

This research employs the YOLO11 small (YOLO11s) variant, which offers an optimal balance between parameter efficiency and accuracy. To establish scientific novelty and determine the optimal configuration for head-detection-based crowd counting, this study implements an ablation study consisting of four distinct experimental scenarios:

- Scenario 0 (S0): Utilizes the default pre-trained YOLO11s weights from the COCO dataset without any further training or domain adaptation.
- Scenario 1 (S1): All network layers are unfrozen (freeze=0) and retrained on the CC_Mach_1 dataset with a high-resolution input of 960 x 960 pixels.
- Scenario 2 (S2): The first 10 layers of the backbone are frozen (freeze=10) to retain generic low-level feature extraction, while the subsequent layers are retrained at 960 x 960 resolution.

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

- Scenario 3 (S3): Uses the same partial freezing strategy as Scenario 2 (freeze=10) but at a standard resolution of 640 x 640 pixels to evaluate the necessity of high-resolution scaling for small object detection.

In the partial freezing scenarios (S2 and S3), the freezing strategy was specifically directed at the initial feature extraction stages spanning layers 0 to 9, which form the primary Backbone. Conversely, all subsequent layers from index 10 through to the final output remained trainable or unfrozen. This unfrozen segment comprises the entire Neck sub-system, which integrates the specialized C3k2 and C2PSA attention modules, along with the decoupled Detection Heads. By making the entire detection assembly trainable rather than limiting updates to the final output layer, the model can recalibrate its feature fusion process and bounding box regression to better match the morphological characteristics of human heads under severe occlusion.

Across all training scenarios (S1–S3), hyperparameters were carefully standardized to ensure a fair comparison and isolate the effects of layer freezing and input resolution. Training was conducted for 50 epochs with a batch size of 4 to fit the GPU memory capacity during high-resolution processing. An early stopping mechanism with a patience value of 15 epochs was activated to prevent overfitting.

The models were optimized using the AdamW optimizer with an initial learning rate of 0.01 and a weight decay of 0.0005 to ensure stable weight convergence. Aggressive data augmentation techniques were implemented to enhance model generalization in crowded scenes, including mosaic (1.0), mixup (0.15), and geometric transformations such as scaling (0.3), shearing (2.0), and rotation (10.0 degrees). Perspective distortion was disabled to maintain the natural proportions of human heads, which is critical for consistent detection under severe occlusion.

2.6. Performance Evaluation and Analysis

The final stage involved evaluating model performance using the validation set. Evaluation metrics were based on standard object detection measures, including Precision, Recall, F1-Score, and mean Average Precision (mAP@0.5). Mean Absolute Error (MAE) was additionally used to quantify the difference between the predicted counts and the ground truth.

The analysis compares a baseline pre-trained YOLO11 model with a fine-tuned variant to assess the impact of task-specific adaptation. The baseline represents the default YOLO11 configuration trained on a generic dataset without domain adaptation for dense crowd scenarios. This comparison is intended to isolate and quantify the effect of head-detection-based fine-tuning and resolution-aware training, rather than to claim superiority over alternative detection architectures.

The accuracy of the detection model is calculated based on the confusion matrix components: True Positive (TP), False Positive (FP), and False Negative (FN). Precision measures the proportion of correct positive detections, while Recall represents the model sensitivity in detecting all target objects. The F1-Score provides the harmonic mean of these two metrics to indicate overall detection balance [31] [32]. These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Furthermore, mAP@0.5 is used to measure the average precision across an Intersection over Union (IoU) threshold of 0.5 [33]. Average Precision (AP) is calculated as the area under the Precision-Recall curve *p(r)*:

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

$$AP = \int_0^1 p(r)\, dr \quad (4)$$

The predicted people count is obtained by summing all detected head bounding boxes per image after applying Non-Maximum Suppression (NMS), a standard post-processing step used to remove duplicate detections and retain a single bounding box for each individual. This detection-based counting approach is commonly adopted in crowd analysis tasks where explicit object localization is required.

To evaluate counting accuracy, the MAE calculates the average absolute error between the predicted number of people (P_i) and the actual ground truth (G_i) across N samples [3]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - G_i| \quad (5)$$

3. RESULT

3.1. Data Preparation and Preprocessing Results

The initial stage of this research involved the collection and standardized preprocessing of the CC_Mach_1 dataset. A total of 1774 high-resolution images were successfully annotated, producing a specialized head-detection dataset. The preprocessing stage resulted in two distinct data configurations: a high-resolution set at 960 x 960 pixels for S1 and S2, and a reduced-resolution set at 640 x 640 pixels for S3. This stage ensured that all bounding boxes were accurately centered on human heads to mitigate the ambiguity of full-body silhouettes in congested scenes. The output of this stage provided the necessary ground truth foundations for the subsequent training and evaluation phases.

3.2. Performance Analysis of the Baseline Pre-trained YOLO11

Figure 4 visualizes the performance of the baseline pre-trained YOLO11 model in estimating human counts against the validation data. Significant extreme discrepancies are observed, where the actual human counts range from 13 to 476 individuals with an average of 105, while the model estimates range only between 0 and 14 individuals, averaging 4.7. The limitations of the model are particularly evident in high-density scenarios, where an actual count of 476 people resulted in a maximum estimation of only 14. This graphical pattern indicates a very weak correlation between the baseline detection results and the validation data.

The subsequent phase involved testing the baseline model's detection capabilities. Figure 5 (a) represents a detection sample in a low-density condition. In this image, the model utilizes green bounding boxes to detect humans as whole entities (full-body). Despite the presence of 24 individuals, the baseline model only successfully detected 9. This inaccurate result highlights the primary weakness of the full-body detection approach, which struggles to overcome visual occlusion in conditions where human figures overlap or are obstructed by other objects. When a significant portion of the body is not visible, the model fails to recognize the object as a human.

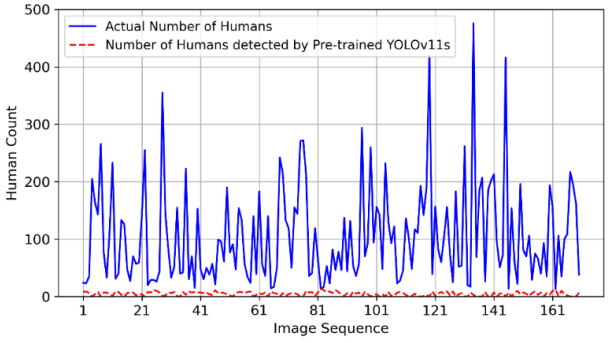

Figure 4. Comparison graph between pre-trained YOLO11 predictions and actual human counts.

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

The constraints and weaknesses of the baseline YOLO11 model were further tested in dense crowd scenarios, as represented in Figure 5 (b). Although the image contains 36 individuals, the model only counted 7. This high prediction error is attributed to the fact that the baseline YOLO11 was trained to recognize full-body human visual features. In dense crowd conditions, extreme visual occlusion results in only the head area being visible. Consequently, a mismatch occurs between the features the model seeks (body silhouettes) and the available visual data (clusters of human heads). These findings demonstrate that full-body detection is inaccurate in crowded conditions, establishing the urgency for retraining the model with a specific focus on head detection.

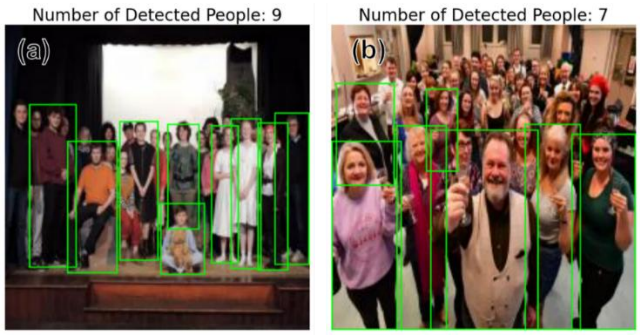

Figure 5. Sample of human detection results using pre-trained YOLO11 in (a) low-density conditions, and (b) high-density conditions (source: validation dataset).

3.3. Performance Analysis of the Fine-tuned YOLO11

The retraining process of the YOLO11 model was monitored via the loss graphs shown in Figure 6. Analysis of these graphs indicates a successful convergence across all configurations. In S1, the training and validation loss decreased consistently without intersecting until the final epoch, suggesting a stable learning process for the full fine-tuning approach. Conversely, both S2 and S3 exhibited a crossover point at epoch 43, where the validation loss began to increase and surpass the training loss. This transition indicates that the models in S2 and S3 entered an overfitting phase after epoch 43, suggesting that the optimal training duration for these partial freezing strategies is slightly shorter than for full fine-tuning.

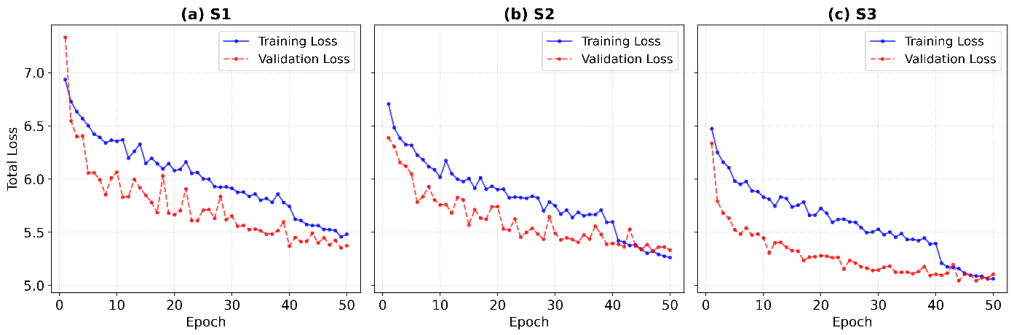

Figure 6. Training and validation loss curves of the YOLO11 model for (a) S1, (b) S2, and (c) S3

Following the training process, the fine-tuned YOLO11 model was re-evaluated. Figure 7 displays the comparison between the actual human counts and the predictions for S1, S2, and S3. Visually, all three models demonstrated a strong ability to track the fluctuations and trends of the ground truth data. Quantitatively, the evaluation shows that S1 and S2 achieved comparable mean predictions of 68.92 and 67.34, respectively. In contrast, S3 showed a notable decrease in performance with a mean prediction of 59.06.

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

A critical observation in Figure 7 is the systematic underestimation in extremely dense scenes, where the models saturated at a maximum prediction of 300 despite a ground truth of 476 individuals. This performance degradation at extreme density levels is primarily attributed to several visual factors. In highly congested regions, severe occlusion causes human heads to appear as a continuous texture rather than individual objects, leading the model to lose the discriminative circular contours required for detection. Furthermore, variations in head orientation, such as individuals facing sideways or downwards, significantly reduce the availability of recognizable facial features. The effectiveness of the C2PSA spatial attention module also diminishes in background regions with low contrast, where individuals in the distance are often misinterpreted as background noise, resulting in a plateau where the model can no longer distinguish new targets as density increases.

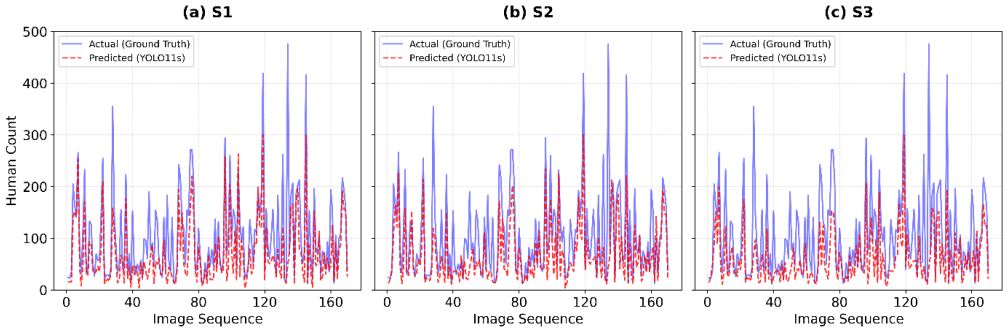

Figure 7. Comparison graph between fine-tuned YOLO11 predictions and actual human counts for (a) S1, (b) S2, and (c) S3.

The effectiveness of each strategy was further tested using specific samples representing different density levels. In low-density conditions with a ground truth of 24 individuals, S2 achieved perfect accuracy by detecting exactly 24 people, as illustrated in Figure 8. S1 showed a slight overcount with 25 detections, while S3 underestimated the count with only 22 detections. These findings suggest that the partial freezing strategy at a 960-pixel resolution (S2) provides the most precise localization when occlusion is minimal. By focusing on the head area at high resolution, the model can successfully distinguish individuals without significant ambiguity.

Figure 8. Sample of human detection results using fine-tuned YOLO11 in low-density conditions for (a) S1, (b) S2, and (c) S3 (source: validation dataset).

Testing in high-density scenarios provided further insights into the robustness of each configuration. In a sample containing 36 individuals, both S1 and S2 predicted 38 people, resulting in a slight overcount of two false positives, as visualized in Figure 9. This error likely stems from extreme crowding where background noise or overlapping features are misinterpreted as additional head targets. Meanwhile, S3 detected 35 individuals, which represents a slight undercount. Although S3 is

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

numerically closer to the ground truth in this specific instance, the result reflects its consistent tendency to underestimate counts due to limited resolution.

Figure 9. Sample of human detection results using fine-tuned YOLO11 in high-density conditions for (a) S1, (b) S2, and (c) S3 (source: validation dataset).

The overall performance across the entire dataset, where S3 showed a significant drop in the mean count compared to S2, confirms that reducing the resolution to 640 pixels limits the model's ability to identify small-scale head features in more congested environments. This suggests that while all fine-tuned models demonstrate a drastic improvement over the baseline, the combination of partial backbone freezing and high-resolution input remains the most effective strategy for maintaining sensitivity. To provide a more comprehensive evaluation of these trends, the following subsection presents a detailed analysis of the accuracy metrics, including mAP and Mean Absolute Error.

3.4. Evaluation of Accuracy Metrics

To quantitatively evaluate model performance, the experimental scenarios were evaluated using standard object detection and crowd counting metrics, as presented in Table 1. The findings reveal a highly significant performance gap between the pre-trained YOLO11 model and the fine-tuned variants. The baseline YOLO11, which utilized default weights from the COCO dataset, demonstrated an inability to perform the head detection task, with an mAP@0.5 of only 0.017 and a MAE of 100.3. These near-zero values indicate that the baseline model almost entirely failed to detect humans correctly. This failure underscores the necessity of domain-specific adaptation, as the pre-trained features were insufficient for identifying small-scale human heads in dense environments.

Table 1. Performance Comparison of YOLO11 Training Scenarios.

Accuracy Metrics	S0 (Pre-trained)	S1	S2	S3
mAP@0.5	0.017	0.682	0.667	0.656
Precision	0.033	0.823	0.799	0.828
Recall	0.003	0.511	0.496	0.460
F1-score	0.006	0.630	0.612	0.591
MAE	100.3	37.8	39.6	46.9

Analysis of the fine-tuning strategies reveals that S1, which involved full unfreezing of all network layers at a 960 x 960 pixels resolution, achieved the highest overall accuracy. This configuration produced an mAP@0.5 of 0.682 and the lowest MAE of 37.8. S2, which implemented partial backbone freezing at the same high resolution, remained highly competitive with an mAP@0.5 of 0.667 and an MAE of 39.6. These findings suggest that while full fine-tuning allows for maximum weight optimization, the partial freezing strategy successfully preserves fundamental feature extraction capabilities with only a marginal reduction in counting accuracy.

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

The role of input resolution is demonstrated in the comparison between S2 and S3. Despite utilizing the same partial freezing policy, S3, which was trained at a standard 640 x 640 resolution, showed a significant degradation in performance, as evidenced by its higher MAE of 46.9. A critical observation can be made regarding the trade-off between Precision and Recall in this scenario. While S3 achieved the highest Precision of 0.828, its Recall dropped to 0.460. This indicates that at a lower resolution, the model becomes more conservative, correctly identifying clearly visible targets but failing to detect a substantial portion of the population due to the loss of fine-grained spatial information.

The overall results confirm that counting accuracy in dense crowd scenarios is heavily dependent on the model's ability to maintain a high Recall rate. The reduction in MAE from 100.3 in the baseline to 37.8 in S1 represents an improvement of approximately 62%. This advancement validates that the combination of task-specific fine-tuning and high-resolution scaling is essential for reliable human estimation under conditions of severe occlusion.

4. DISCUSSIONS

The performance improvement achieved by the fine-tuned YOLO11 model is consistent with prior studies demonstrating that task-specific adaptation of YOLO-based architectures can enhance detection accuracy in complex environments. Similar findings were reported by Su et al. [37], who showed that targeted modifications to YOLOv7 improved pedestrian detection performance under challenging visual conditions. In this study, the observed performance gap should therefore be interpreted as the effect of domain-specific fine-tuning, rather than an absolute measure of YOLO11's capability in its default configuration.

The ablation study provides critical insights into the relationship between backbone freezing and input resolution. S1 (Full Fine-tuning) achieved the highest mAP@0.5 of 0.682. However, the competitive performance of S2 (Partial Freezing) at 0.667 suggests that freezing the initial ten backbone layers is a viable strategy to preserve generic features while reducing training overhead.

The comparison between S2 and S3 highlights the necessity of high-resolution scaling for small-object detection. Reducing the resolution from 960 to 640 pixels in S3 led to a significant increase in MAE from 39.6 to 46.9. Although S3 showed the highest Precision at 0.828, its Recall dropped to 0.460. This confirms that lower resolutions cause the model to become overly conservative, correctly identifying only a few prominent targets while missing a large portion of the occluded population [4].

Table 2. Comparison with existing crowd counting studies

Study	Method	Representation	Dataset Type	Key Metric
Li et al. [5]	CSRNet	Density Map	Dense Crowd	MAE: 66.2
Ali et al. [25]	YOLOv5	Head	High Density	mAP@0.5: 0.956
Hassan et al. [26]	YOLOv5	Head	Medium Density	mAP@0.5: 0.78
Su et al. [37]	YOLOv7	Full Body	Pedestrian	mAP@0.5: 0.822
This Study (S2)	YOLO11s	Head	Low to High Density	MAE: 39.6

To contextualize these findings within the broader crowd counting landscape,

Table 2 summarizes the performance of the proposed strategy alongside existing research. The competitive performance and high precision achieved in this study align with the theoretical improvements of the YOLO11 architecture. These results suggest that the integration of spatial attention mechanisms provides a robust foundation for handling severe occlusion, positioning YOLO11 as a highly capable successor to previous YOLO generations typically utilized for high-density counting tasks. While direct numerical comparisons are limited by dataset variations, the observed stability in

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

detection and the reduction in counting errors support the efficacy of the C3k2 and C2PSA modules in preserving discriminative features within congested environments [19], [21].

This study provides a significant empirical validation that contributes to the shift from density-estimation methods to instance-level detection in crowd analytics. Traditionally, dense crowd counting has been dominated by density-map regression, such as MCNN [4] and CSRNet [5], which often lack individual-level interpretability. By demonstrating that a single-stage detector like YOLO11 can achieve competitive accuracy, this research proves that detection-based pipelines are viable alternatives for high-density scenarios. This advancement is vital for Informatics applications requiring granular data for downstream tasks, including real-time tracking and behavior analysis, which global density maps cannot provide [22], [38].

From a practical perspective, the near-real-time inference capability of the fine-tuned YOLO11 enables its deployment in real-world crowd monitoring applications. Reliable head-based detection can support operational decision-making in scenarios such as early congestion detection, crowd density control, and adaptive access management during large-scale public events [22], [39]. In urban environments, accurate crowd estimation may further assist authorities in optimizing pedestrian flow and infrastructure planning based on spatiotemporal density patterns [40].

Despite these advantages, challenges remain in maintaining high recall under extreme crowd density and prolonged occlusion. Recent studies suggest that integrating detection with multi-object tracking frameworks can improve temporal consistency and robustness in such conditions [39]. Future research is therefore encouraged to explore hybrid detection and tracking approaches, such as combining YOLO-based detectors with tracking algorithms like DeepSORT, to further enhance counting accuracy in highly congested environments [38]. Ethical considerations, including data anonymization and privacy protection, should also be addressed when deploying such systems in real-world settings.

5. CONCLUSION

This study concludes that task-specific fine-tuning on a specialized head-detection dataset significantly improves the performance of YOLO11 for crowd counting in congested environments. Quantitative results reveal that Scenario 1 achieved a substantial improvement, reducing the Mean Absolute Error (MAE) by 62% from 100.3 in the baseline to 37.8. The ablation study further indicates that Scenario 2 remains highly competitive with an MAE of 39.6, while Scenario 3 confirms that a standard 640-pixel resolution is insufficient for dense scenes, resulting in a higher MAE of 46.9. These findings emphasize that the combination of partial backbone freezing and high-resolution input at 960 pixels provides the most balanced configuration for maintaining detection sensitivity and counting accuracy under severe occlusion.

From a broader informatics perspective, this research validates the use of single-stage detectors as a robust and interpretable alternative to traditional density-map regression methods. By utilizing modern attention-driven modules such as C3k2 and C2PSA, YOLO11 enables precise instance-level localization, which is essential for advanced analytical tasks like pedestrian tracking and spatial density mapping. This methodological shift offers a more granular approach to public safety and urban monitoring systems compared to global density estimation. Future work should investigate the integration of these detection capabilities with multi-object tracking frameworks, such as DeepSORT, to enhance temporal consistency and further reduce counting errors in real-world deployments.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] S. Yi, H. Li, and X. Wang, "Pedestrian Behavior Modeling From Stationary Crowds With Applications to Intelligent Surveillance," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4354–4368, Sep. 2016, doi: 10.1109/TIP.2016.2590322.

[2] S. Koswatte, K. McDougall, and X. Liu, "Crowd-Assisted Flood Disaster Management," 2022, pp. 39–55. doi: 10.1007/978-3-031-14096-9_3.

[3] L. Deng, Q. Zhou, S. Wang, J. M. Górriz, and Y. Zhang, "Deep learning in crowd counting: A survey," *CAAI Trans. Intell. Technol.*, vol. 9, no. 5, pp. 1043–1077, Oct. 2024, doi: 10.1049/cit2.12241.

[4] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.

[5] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 1091–1100. doi: 10.1109/CVPR.2018.00120.

[6] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018, doi: 10.1016/j.patrec.2017.07.007.

[7] H. F. Elsepae, H. M. El-Hoseny, E. K. I. Hamad, and E.-S. M. El-Rabaie, "Deep learning for crowd counting in complex environments: challenges and novel trends," *Discov. Comput.*, vol. 29, no. 1, p. 101, Feb. 2026, doi: 10.1007/s10791-026-09928-8.

[8] M. Wang, X. Zhou, and Y. Chen, "A comprehensive survey of crowd density estimation and counting," *IET Image Process.*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/ipr2.13328.

[9] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "A survey of deep learning methods for density estimation and crowd counting," *Vicinagearth*, vol. 2, no. 1, p. 2, Feb. 2025, doi: 10.1007/s44336-024-00011-8.

[10] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, Feb. 2022, doi: 10.1016/j.neucom.2021.02.103.

[11] A. Nugroho, F. Indaryanto, and A. F. Suni, "Distance and people counting app based on YOLO as a Covid-19 health protocol," 2023, p. 040024. doi: 10.1063/5.0141518.

[12] P. Ren, L. Wang, W. Fang, S. Song, and S. Djahel, "A novel squeeze YOLO-based real-time people counting approach," *Int. J. Bio-Inspired Comput.*, vol. 16, no. 2, p. 94, 2020, doi: 10.1504/IJBIC.2020.109674.

[13] A. N. Alhawsawi, S. D. Khan, and F. U. Rehman, "Enhanced YOLOv8-Based Model with Context Enrichment Module for Crowd Counting in Complex Drone Imagery," *Remote Sens.*, vol. 16, no. 22, p. 4175, Nov. 2024, doi: 10.3390/rs16224175.

[14] W. Farhat, O. Ben Rhaiem, H. Faiedh, and C. Souani, "Pedestrian detection and tracking using an enhanced YOLOv9 model for automotive vehicles," *Measurement*, vol. 254, p. 118009, Oct. 2025, doi: 10.1016/j.measurement.2025.118009.

[15] L. Wu, X. Li, P. Ma, and Y. Cai, "Research on a Dense Pedestrian-Detection Algorithm Based on an Improved YOLO11," *Futur. Internet*, vol. 17, no. 10, 2025, doi: 10.3390/fi17100438.

[16] M. L. Ali and Z. Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," *Computers*, vol. 13, no. 12, 2024, doi: 10.3390/computers13120336.

[17] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in Crowded Scenes: One Proposal, Multiple Predictions," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2003.09163

[18] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS — Improving Object Detection with One Line of Code," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 5562–5570. doi: 10.1109/ICCV.2017.593.

[19] A. A. Murat and M. S. Kiran, "A comprehensive review on YOLO versions for object detection," *Eng. Sci. Technol. an Int. J.*, vol. 70, p. 102161, Oct. 2025, doi: 10.1016/j.jestch.2025.102161.

[20] L. He, Y. Zhou, L. Liu, W. Cao, and J. Ma, "Research on object detection and recognition in

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

remote sensing images based on YOLOv11," *Sci. Rep.*, vol. 15, no. 1, p. 14032, Apr. 2025, doi: 10.1038/s41598-025-96314-x.

[21] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," vol. 2024, pp. 1–9, 2024, [Online]. Available: http://arxiv.org/abs/2410.17725

[22] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Mach. Vis. Appl.*, vol. 19, no. 5–6, pp. 345–357, Oct. 2008, doi: 10.1007/s00138-008-0132-4.

[23] Z. Xu, H. Zhao, P. Liu, L. Wang, G. Zhang, and Y. Chai, "SRTSOD-YOLO: Stronger Real-Time Small Object Detection Algorithm Based on Improved YOLO11 for UAV Imageries," *Remote Sens.*, vol. 17, no. 20, p. 3414, Oct. 2025, doi: 10.3390/rs17203414.

[24] X. Gong, J. Yu, H. Zhang, and X. Dong, "AED-YOLO11: A small object detection model based on YOLO11," *Digit. Signal Process.*, vol. 166, p. 105411, Nov. 2025, doi: 10.1016/j.dsp.2025.105411.

[25] M. A. Ali, A. J. Hussain, and A. T. Sadiq, "Detection And Count of Human Bodies In a Crowd Scene Based on Enhancement Features By Using The YOLO v5 Algorithm," *Iraqi J. Comput. Commun. Control Syst. Eng.*, pp. 125–134, Jun. 2022, doi: 10.33103/uot.ijccce.22.2.11.

[26] M. Hassan, F. Hussain, S. D. Khan, M. Ullah, M. Yamin, and H. Ullah, "Crowd counting using deep learning based head detection," *Electron. Imaging*, vol. 35, no. 9, pp. 293--1-293–6, Jan. 2023, doi: 10.2352/EI.2023.35.9.IPAS-293.

[27] M. Abubaker, Z. Alsadder, H. Abdelhaq, M. Boltes, and A. Alia, "RPEE-Heads Benchmark: A Dataset and Empirical Comparison of Deep Learning Algorithms for Pedestrian Head Detection in Crowds," *IEEE Access*, vol. 13, no. April, pp. 73451–73467, 2025, doi: 10.1109/ACCESS.2025.3563311.

[28] R. V Vadavadagi, S. E. N, A. Marlinganavvar, A. Hurkadli, K. Bhoomraddi, and U. Kulkarni, "Head Counting in Crowded Scenes Using YOLOv10: A Deep Learning Approach," in *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025)* -, 2025, pp. 611–618.

[29] M. H. K. Khel *et al.*, "Realtime Crowd Monitoring—Estimating Count, Speed and Direction of People Using Hybridized YOLOv4," *IEEE Access*, vol. 11, pp. 56368–56379, 2023, doi: 10.1109/ACCESS.2023.3272481.

[30] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[31] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.16061

[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[33] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, IEEE, Jul. 2020, pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130.

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

[35] F. Ciaglia, F. S. Zuppichini, P. Guerrie, M. McQuade, and J. Solawetz, "Roboflow 100: A Rich, Multi-Domain Object Detection Benchmark," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.13523

[36] G. A. Capstone, "CC_Mach_1 Dataset," Jul. 2023, *Roboflow*. [Online]. Available: https://universe.roboflow.com/ga-capstone-3f9vu/cc_mach_1

[37] J. Su, F. Wang, and W. Zhuang, "An Improved YOLOv7 Tiny Algorithm for Vehicle and Pedestrian Detection with Occlusion in Autonomous Driving," *Chinese J. Electron.*, vol. 34, no. 1, pp. 282–294, Jan. 2025, doi: 10.23919/cje.2023.00.256.

[38] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962.

[39] D. Helbing and A. Johansson, "Pedestrian, Crowd and Evacuation Dynamics," in *Extreme*

Vol. 7, No. 2, April 2026, Page. 1871-1885
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5699

Environmental Events, New York, NY: Springer New York, 2011, pp. 697–716. doi: 10.1007/978-1-4419-7695-6_37.

[40] Z. Asadi-Shekari, M. Moeinaddini, and M. Zaly Shah, "Pedestrian safety index for evaluating street facilities in urban areas," *Saf. Sci.*, vol. 74, pp. 1–14, Apr. 2015, doi: 10.1016/j.ssci.2014.11.014.