International Journal of Computational Methods and Experimental Measurements
https://www.acadlore.com/journals/IJCMEM

Enhancing Real-Time Face Detection Performance through YOLOv11 and Slicing-Aided Hyper Inference

Muhammad Fachrurrozi[1*], Muhammad Naufal Rachmatullah[1], Akhiar Wista Arum[2], Fiber Monado[3]

[1] Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya, 30139 Palembang, Indonesia
[2] Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, 30139 Palembang, Indonesia
[3] Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Sriwijaya, 30139 Palembang, Indonesia

[*] Correspondence: Muhammad Fachrurrozi (mfachrz@unsri.ac.id)

Received: 07-17-2025 **Revised:** 09-03-2025 **Accepted:** 09-09-2025

Citation: M. Fachrurrozi, M. N. Rachmatullah, A. W. Arum, and F. Monado, "Enhancing real-time face detection performance through YOLOv11 and slicing-aided hyper inference," *Int. J. Comput. Methods Exp. Meas.*, vol. 13, no. 3, pp. 520–531, 2025. https://doi.org/10.56578/ijcmem130305.

Abstract: Real-time face detection in crowded scenes remains challenging due to small-scale facial regions, heavy occlusion, and complex illumination, which often degrade detection accuracy and computational efficiency. This study presents an enhanced detection framework that integrates Slicing-Aided Hyper Inference (SAHI) with the YOLOv11 architecture to improve small-face recognition under diverse visual conditions. While YOLOv11 provides a high-speed single-stage detection backbone, it tends to lose fine spatial information through downsampling, limiting its sensitivity to tiny faces. SAHI addresses this limitation by partitioning high-resolution images into overlapping slices, enabling localized inference that preserves structural detail and strengthens feature representation for small targets. The proposed YOLOv11–SAHI system was trained and evaluated on the WIDER Face dataset across Easy, Medium, and Hard difficulty levels. Experimental results demonstrate that the integrated framework achieves Average Precision (AP) scores of 96.33%, 95.87%, and 90.81% for the respective subsets—outperforming YOLOv7, YOLOv5, and other lightweight detectors, and closely approaching RetinaFace accuracy. Detailed error analysis reveals that the combined model substantially enhances small-face detection in dense crowds but remains sensitive to severe occlusion, motion blur, and extreme pose variations. Overall, YOLOv11 coupled with SAHI offers a robust and computationally efficient solution for real-time face detection in complex environments, establishing a foundation for future work on pose-invariant feature learning and adaptive slicing optimization.

Keywords: Face detection, YOLOv11, Slicing aided hyper inference, WIDER face

1 Introduction

The advancement of facial recognition methods and the ease of collecting faces digitally have led to an increase in the use of facial recognition methods [1, 2]. However, the success of a facial recognition application often depends on the success of the face detecting system in an image. In general, there two main challenges in the facial detection process. The first is the identification of faces within images containing diverse backgrounds. These images have differing poses, occlusions, lighting, etc. The second is to identify faces in an image which may occur in varying scales, at different places in the image. Both of these issues are influencing the accuracy and time efficiency, respectively. Because there is a trade-off between performance and time speed, developing a balance both conditions become increasingly difficult.

Early face detection methods primarily relied on traditional approaches involving the extraction of engineered features from images, subsequently utilizing various classifiers to accurately identify facial regions. Notably, the Haar cascade classifier [3] and the Histogram of Oriented Gradients (HOG) paired with a Support Vector Machine (SVM) [4] represent seminal classical techniques in this domain, exemplifying significant historical advancements. Nevertheless, these classical methodologies still exhibit limited detection accuracy when confronted with challenging images containing complex variations, as demonstrated by performance evaluations on the WIDER FACE dataset [5].

https://doi.org/10.56578/ijcmem130305

In the era of deep learning method advancement, researchers are competing to analyze the influence of deep learning to overcome the complexity of face detection cases. Therefore, the use of deep learning methods as face detectors as begun to be developed, such as Face R-CNN [6–11], Face R-FCN [12, 13], single shot detector (SSD) [14–17] which introduces a multiscale mechanism, feature enhancements in the FPN [18–22] architecture and focus loss in the Retinaface architecture [23] have been successfully developed to overcome the challenges in face detection from the WIDER FACE dataset. The architecture has been modified to identify unique patterns of human facial objects and perform well under unconstrained conditions.

Although many existing face detection architectures achieve high accuracy, they often incur substantial computational overhead and long processing times. This is particularly true for two-stage detection pipelines, where the system first generates region proposals and then performs classification and bounding box refinement. While these methods can be precise, their sequential nature limits suitability for real-time applications or deployment on resource-constrained devices. Even single-stage detectors such as SSD and RetinaNet face efficiency challenges: SSD depends heavily on anchor-based multi-scale feature extraction, while RetinaNet's complexity stems from focal loss and feature pyramid networks (FPN). Consequently, these approaches often struggle to meet the requirements of low-latency detection tasks.

In contrast, the You Only Look Once (YOLO) architecture introduced by Redmon et al. [24], perform bounding boxes prediction and classification in a single pass over the image. This architecture is well known for its speed and computational efficiency [25]. The YOLO architecture has been evolved through multiple versions for example YOLOv8 implemented an anchor-free and decoupled head structure, which improved localization precision and stabilized the training process. YOLOv9 incorporated Programmable Gradient Information (PGI) and a Generalized Efficient Layer Aggregation Network (GELAN), resulting in greater feature reuse efficiency and improved accuracy, particularly for small and medium-scale objects. YOLOv10 replaced the traditional Non-Maximum Suppression (NMS) with a dual-label assignment mechanism, which accelerated inference while preserving detection robustness in high-density environments. YOLOv11 focused on refining low-latency feature extraction via lightweight convolutional modules and enhanced feature pyramids, optimizing efficiency for both edge and cloud deployment without compromising precision. Finally, YOLOv12 integrated Area Attention and Residual ELAN modules, enabling superior focus on salient features and fine-grained object structures, thereby achieving higher accuracy for occluded and small targets.

Recent studies have increasingly explored YOLO architectures for face detection, with various modifications designed to improve scale robustness, efficiency, and adaptability. Zheng et al. [26] proposed GCSEM-YOLO to emphasize small-scale face detection. This model achieved 94.3% (Easy), 93.0% (Medium), and 84.7% (Hard) AP on the WIDER Face dataset. Pebrianto et al. [27] utilized YOLOv3 for real-time face detection and got reported increase in accuracy (87%) while maintaining low latency. However, the authors provide the declining performance on small and occluded faces. Sufanin Chan et al. [28] analyses the performance YOLOv7 algorithm. They found that the efficiency improved but the AP value dropped to only 83.15% on the Hard subset, indicating limited robustness under extreme conditions. Vemulapalli et al. [29] incorporated landmark detection into YOLOv8. These authors achieving a relatively fast detection process but still underperforming on small and blurred faces. Gao and Yang [30] enhanced Tiny-YOLOv3 with an attention mechanism to reduce false detections. The results show 95.26 (Easy), 89.2 (Medium) and 77.9 (hard) on validation set. Qi et al. [31] introduced YOLO5Face, achieving 93.61% (Easy), 91.54% (Medium), and 80.53% (Hard), demonstrating that even optimized lightweight models still struggle with tiny faces and occlusion.

Despite these advances, the performance of most YOLO-based detectors declines on the Hard subset of the WIDER Face that predominantly includes tiny faces, high occlusion, extreme poses, and poor illumination [32]. This limitation arises from the intrinsic downsampling operations in convolutional backbones, which lead to the loss of fine-grained spatial information critical for small-object detection. To overcome these challenges, this study integrates Slicing Aided Hyper Inference (SAHI) [33] with the YOLOv11 architecture. SAHI improves sensitivity to small-scale and partially visible faces by dividing high-resolution images into overlapping slices and performing inference locally on each region. This approach effectively increases the pixel density in tiny faces and preserves structural detail that would otherwise be lost during global image downscaling. The localized predictions are subsequently merged using Non-Maximum Suppression (NMS), ensuring coherent and duplicate-free detections. By coupling YOLOv11's high-speed inference capability with SAHI's fine-grained spatial analysis, the proposed framework improves accuracy on the Hard subset while maintaining real-time processing performance.

In terms of performance analysis under real-world environment, the majority of previous works primarily emphasize aggregate performance metrics such as mAP or AP50 on the WIDER Face dataset. However, these standard evaluations generally neglect to quantify detection reliability across specific, challenging visual conditions, including varying degrees of blur severity, occlusion, pose orientation, illumination, and facial expression. Consequently, this methodological reliance on headline metrics provides limited insight into the systematic fluctuation of detection accuracy under these specific challenges. This oversight obscures critical factors responsible for real-world failure modes and hinders the comprehensive evaluation of a model's generalization capability in high-density or crowded scenes. Ultimately, this approach conceals systematic architectural weaknesses and impedes meaningful advancements in model performance.

In this study we propose a hybrid framework for face detection that combines YOLOv11 and SAHI, followed by a detailed, attribute evaluation using the WIDER Face dataset. The YOLOv11 architecture provides an efficient and high-accuracy detection backbone suitable for low-latency inference. In accordence with SAHI to divide the input image into overlapping slices to improve detection of small or occluded faces. This approach improves detection accuracy under difficult visual conditions. In summary, the key contributions of this study are as follows.

1. Development of a unified YOLOv11–SAHI framework that enhances detection performance for small, occluded, and densely packed faces in complex scenes.

2. Introduction of an attribute-based evaluation on the WIDER Face dataset, enabling detailed insights on proposed model.

3. Demonstration of real-time processing alongside high detection accuracy, supporting deployment in surveillance and crowd-monitoring applications.

4. Comprehensive validation of the framework's robustness across varied imaging conditions, including poor lighting, motion blur, and high scene density.

2 Material and Method

2.1 Data Collection

This study uses the WIDER Face dataset [4], which is a well-known and large-scale benchmark with 32,203 images containing 393,703 annotated faces in 61 event classes. These images were collected with various real-world conditions like scale, pose, occlusion, blur, illumination and expression make it suitable for face detection benchmark data set.

Following the protocol for official partitioning, the dataset is designated 40%, 10%, and 50% for training, validation, testing respectively. This subset allows for reproducibility and comparison with previous studies. The training set has 12,876 images containing 152,439 annotated faces. The validation set has 3,226 images containing 39,708 annotated faces. Overall, these subsets cover a wide range of visual and environmental conditions that should allow for effective optimization of the model and evaluation of performance on multiple levels of difficulty.

A distinctive feature of WIDER Face is its explicit difficulty partitions—$Easy$, $Medium$, and $Hard$—defined by ranking event classes based on face detection rates under scale, occlusion, and pose constraints. The Hard subset, in particular, contains faces smaller than 32×32 pixels, with heavy occlusion or atypical orientations, making it a stringent benchmark for small-face detection algorithms.

Table 1 present a distribution of the training and validation set under certain condition, and it reveals:

- Size Distribution: Small faces dominate, followed by medium and large faces, confirming the dataset's emphasis on small-scale detection challenges
- Occlusion: While fully visible faces form the largest group, substantial heavy and partial occlusion cases are present, adding complexity to detection.
- Blur: Heavy blur is most frequent, followed by medium blur, with clear faces being least common.
- Illumination: Clear lighting conditions predominate, though some medium degradation cases exist.
- Pose Levels: Most faces exhibit typical poses, with a smaller proportion showing atypical or extreme orientations.
- Expression Levels: Normal expressions dominate, while exaggerated expressions and other variations appear rarely.

These dataset characteristics highlight the significant challenges posed by WIDER Face, particularly in handling small-scale, occluded, blurred, and variably lit faces. Combined with its extensive annotations and standardized Average Precision (AP) evaluation protocol across difficulty levels, WIDER Face provides an ideal testbed for evaluating the robustness of face detection methods. Leveraging this dataset enables rigorous comparison of YOLOv11 with and without SAHI integration, especially in detecting small faces in dense crowd environments. Figure 1 presents sample images from the WIDER Face dataset.

2.2 Proposed Method

The primary objective of this research is to enhance the detection accuracy of faces, particularly small-scale instances, in densely populated images. To achieve this, we propose an integrated detection framework by combining the YOLOv11 architecture with the SAHI method. Our methodology introduces a dual-stage detection pipeline designed to overcome the difficulties of identifying small-scale faces in crowded scenes. This approach, outlined in the flowchart in Figure 2. The proposed method used YOLOv11 as the primary detection backbone while SAHI performing image slicing to handle small faces in crowds.

Table 1. WIDER face distribution across certain condition

Condition	Level	Train	Valid
Image Size	Large ($\geq$ 96^2px)	9,340	2,327
	Medium (32^2 $-$ 96^2px)	32,756	8,606
	Small ($<$ 32^2px)	11,0343	28,775
Blur	Heavy	88,951	23,641
	Medium	40,746	10,381
	None/Clear	22,737	5,686
Illumination	Normal	144,436	37,410
	Extreme	7,998	2,298
Pose	Typical	146,463	38,053
	Atypical	5,971	1,655
Occlusion	None/Clear	93,920	23,812
	Partial	26,021	7,185
	Heavy	32,493	8,711
Expression	Typical	150,601	39,308
	Exaggerate	1,833	400

Figure 1. Example of WIDER face dataset

In order to identify the most effective model configuration, first we benchmarked multiple YOLOv11 variants ranging from the lightweight YOLOv11-nano to the more complex YOLOv11-xlarge (xl) architecture using the WIDER Face training subset. The evaluation considered trade-offs between accuracy and inference speed, with particular emphasis on performance within the Hard subset. Based on these results, the best-performing YOLOv11 variant was selected as the core detection backbone.

In the inference phase, we employ SAHI to address the challenge in detecting small-scale faces. SAHI partitions the input image into overlapping slices which effectively increasing the pixel resolution of small facial regions presented to the detector. This process helps preserve fine-grained facial details that would otherwise be lost in full-image downsampling. Each slice is independently processed by the YOLOv11 model, and the resulting bounding boxes are reprojected from slice-local to global image coordinates. To eliminate redundant predictions from overlapping regions, a confidence-weighted Non-Maximum Suppression (NMS) algorithm is applied. The NMS threshold was carefully optimized to balance recall and precision, particularly in dense visual contexts.

This integrated approach leverages the strengths of both high-speed object detection from YOLOv11 and resolution-aware inference in SAHI, resulting in significantly improved accuracy for small-face detection without sacrificing real-time performance.

2.2.1 You only look once

YOLOv11 represents the latest iteration in the You Only Look Once family, designed to achieve an optimal trade-off between inference speed and detection accuracy. Its architecture incorporates an improved backbone based on Cross Stage Partial Networks (CSP) to reduce computational redundancy, a Path Aggregation Network (PANet) to enhance information flow across feature levels, and a Spatial Pyramid Pooling–Fast (SPPF) module to capture multi-scale contextual information without incurring significant computational overhead as illustrated in Figure 3.

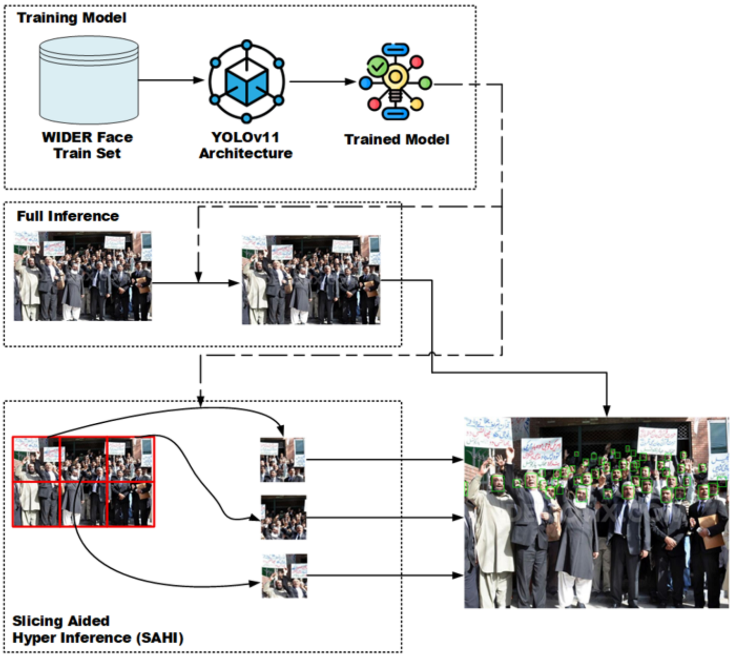


Figure 2. Proposed model

During the feature extraction stage, YOLOv11 processes the input image through the backbone to generate multi-resolution feature maps, which are subsequently fused in the neck for improved feature representation. The head module produces bounding box coordinates, confidence scores, and object class predictions. Although the architecture is optimized for detecting objects of varying scales, the detection of extremely small faces in crowded scenes remains challenging due to the progressive loss of feature resolution during multi-stage down-sampling.

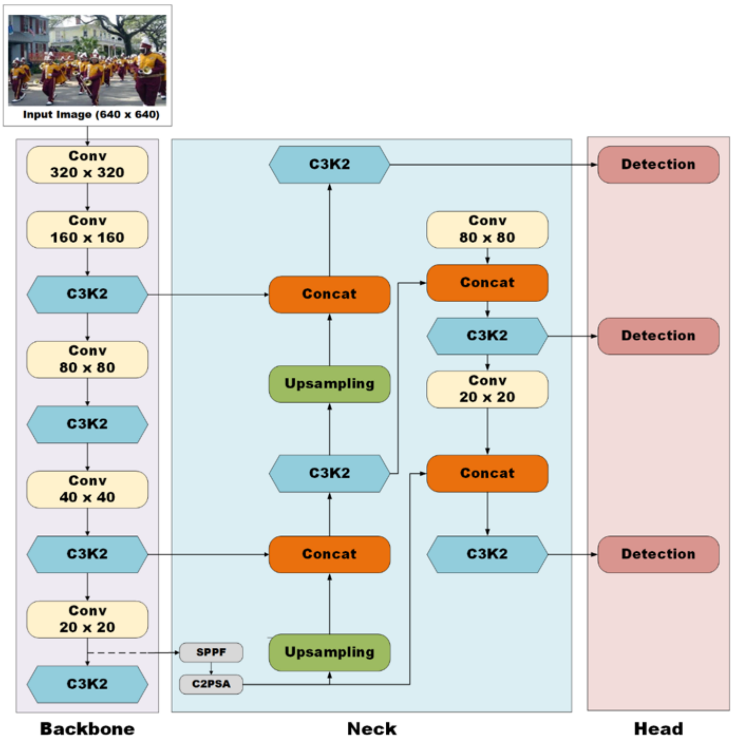


Figure 3. YOLOv11 architecture

2.2.2 Slicing aided hyper inference

Slicing Aided Hyper Inference (SAHI) is a specialized inference strategy designed to improve the accuracy of small-object detection. The method partitions a high-resolution image into smaller, overlapping patches, performs inference on each patch independently, and subsequently merges the results to produce a comprehensive detection output for the original image. Each patch is resized while preserving its aspect ratio before being processed by the object detector. As an optional step, full-image inference may also be conducted to facilitate the detection of larger objects. The outputs from overlapping patches are consolidated into the original image space through Non-Maximum Suppression (NMS), wherein predictions with Intersection over Union (IoU) values exceeding a predefined threshold are considered duplicates. For each duplicate set, detections with confidence scores below the specified threshold are removed, ensuring that only optimal and non-overlapping predictions are retained.

In this study, SAHI is tightly integrated with YOLOv11 to address the challenge of detecting tiny and partially occluded faces in crowded scenes. YOLOv11's multi-scale feature extraction mechanism, supported by its enhanced Path Aggregation Network (PAN) and Cross-Stage Partial (CSP) connections, allows the model to capture hierarchical spatial features across different receptive fields. However, its global downsampling process may still suppress fine details when detecting extremely small faces. SAHI complements this by dividing the input image into overlapping slices, typically with 10–20% overlap, which enables the model to analyze each region at a higher effective resolution. This strategy preserves facial details that would otherwise be lost, while the overlap ensures feature continuity across patch boundaries and prevents object fragmentation.

Each slice is independently processed through the YOLOv11 backbone, generating localized detections that are subsequently reprojected to the original image coordinates. The final stage applies NMS to merge all slice-level detections, eliminating redundant boxes and optimizing overall confidence scores. This slicing-and-merging framework effectively mitigates common detection failures in crowd-dense environments—such as missed tiny faces, mislocalized bounding boxes, or duplicate detections—by ensuring that each small facial region receives focused analysis at an appropriate scale. The integration of YOLOv11's multi-scale representation learning with SAHI's localized high-resolution inference produces a complementary synergy: YOLOv11 ensures efficient global context modeling and robust feature hierarchies, while SAHI enhances spatial granularity for small-object detection. Together, they yield substantially improved precision on the Hard subset of the WIDER Face dataset, where small, occluded, and low-contrast faces dominate. This hybrid inference framework thus provides a computationally efficient yet highly accurate solution for real-world, crowd-intensive face detection applications.

2.2.3 Evaluation metrics

Average Precision (AP) is the primary evaluation metric used to measure the performance of object detection systems, including face detection. AP represents the average of precision values computed across various recall levels, thus reflecting the model's ability to maintain prediction accuracy as the detection coverage expands. The calculation of AP begins with constructing a Precision–Recall (PR) curve, obtained by varying the confidence score threshold for the model's predictions. Eqs.(1) and (2) calculate the precision and recall value from predicted faces

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

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

AP is then computed as the area under the $Precision - Recall$ curve (Area Under Curve, AUC) (Eq.(3)) where $P_{(r)}$ denotes the precision at recall level (r). In practical implementations, this integral is approximated by discrete summation over interpolated precision points.

$$AP = \int_0^1 P_{(r)} dr \tag{3}$$

In this study, AP is calculated at specific Intersection over Union (IoU) thresholds, in accordance with the WIDER Face evaluation protocol. IoU is a metric that quantifies the overlap between the predicted bounding box and the ground truth bounding box, defined in formula 4. A higher AP indicates better capability of the model to consistently produce accurate detections across various recall levels. In the context of small-scale face detection on the hard subset of WIDER Face, a high AP demonstrates that the proposed method can maintain high precision even under challenging conditions, such as object occlusion, poor illumination, or extreme face orientations.

$$IoU = \frac{AreaofOverlap}{AreaofUnion} \tag{4}$$

3 Experimental Result

The experimental evaluation was conducted on the WIDER Face benchmark dataset, which is divided into three difficulty levels: easy, medium, and hard. Performance was measured using Average Precision (AP) at IoU thresholds of 0.5. The proposed YOLOv11 + SAHI framework was compared against the baseline YOLOv11 model without slicing-based inference. In the SAHI configuration, each input image was divided into overlapping slices with a slice height and width of 256 pixels and an overlap ratio of 20%, ensuring that small faces were preserved within at least one slice for improved detection.

Table 2 presents the comparative AP@0.5 results between different YOLOv11 model variants—Nano, Small, Medium, Large, and Extra-Large—evaluated with and without SAHI integration on the Easy, Medium, and Hard subsets of WIDER Face. On the Easy subset, YOLOv11 + SAHI achieved performance comparable to the baseline across most variants, with notable gains for the Nano (+1.81%), Medium (+1.18%), and Extra-Large (+1.03%) models, indicating that the slicing strategy maintains or slightly improves detection accuracy for larger and medium-scale faces. On the Medium subset, SAHI consistently boosted AP scores for all variants, with the largest improvements observed in Nano (+2.27%) and Small (+1.91%) models, reflecting enhanced detection for moderately challenging conditions.

The most substantial gains occurred on the Hard subset, where YOLOv11 + SAHI outperformed the baseline by wide margins, particularly in Nano (+4.28%), Small (+3.87%), and Medium (+3.35%) configurations. Across all test cases, the best performance was recorded by the Extra-Large model with SAHI which demonstrate that integrating slicing mechanism substantially enhances the model's ability to detect small-scale faces under challenging crowd conditions without compromising performance on larger faces in less complex scenes.

Table 2. AP comparison between baseline model and with SAHI

Subset	Model	Average Precision (%) @ 0.5				
		Nano	**Small**	**Medium**	**Large**	**Extra-Large**
Easy	Baseline	93.45	94.92	95.55	95.50	95.77
	With SAHI	95.26	93.39	96.73	93.67	**96.33**
Medium	Baseline	92.61	94.12	94.74	94.75	95.02
	With SAHI	94.88	96.03	96.36	95.66	**95.87**
Hard	Baseline	85.68	87.38	88.37	88.38	88.82
	With SAHI	89.96	91.25	91.72	89.71	**90.81**

The most significant improvement was observed on the $hard$ subset, which predominantly contains small and heavily occluded faces. By dividing high-resolution images into overlapping slices, SAHI preserved the relative resolution of small faces during inference, mitigating the loss of spatial detail caused by the down-sampling operations in the YOLOv11 backbone. Figure 4 illustrates a qualitative comparison the baseline YOLOv11 fails to detect several small faces in dense scenes, whereas YOLOv11 + SAHI successfully identifies them. This improvement can be attributed to the fact that each slice effectively acts as a $zoomed - inview$ of the scene, allowing the model to focus on finer details that would otherwise be blurred when the entire image is downscaled.

Precision–Recall (PR) curves (Figure 5) further confirm that the proposed method consistently achieves higher precision across a wide range of recall levels, especially in the high-recall region, indicating that more true faces are correctly detected without a substantial increase in false positives.

4 Discussion

The introduction of SAHI inevitably increases inference time due to multiple passes required for sliced inputs. On average, the proposed method reduced throughput from 62 FPS (baseline YOLOv11) to 25 FPS on an NVIDIA RTX 5090 GPU. However, this trade-off is acceptable in applications where accuracy for small or hard-to-detect faces is prioritized over real-time speed, such as forensic video analysis or high-risk security monitoring. The efficiency of YOLOv11's architecture, particularly its CSP connections and PANet-based feature aggregation, helps mitigate the computational burden of slicing, allowing the system to remain within operational limits.

Despite these gains, error analysis reveals consistent failure modes under extreme conditions. As shown in Figure 6a, the model maintains high success rates on clear (98.42%) and normal blur (98.76%) faces but drops sharply to 84.30% under heavy blur, with nearly 15.7% of faces missed. Similarly, Figure 6b indicates robust performance on non-occluded (95.81%) faces but a marked decline to 78.25% under heavy occlusion. Pose variations remain challenging, with 10% failures for typical and 8.31% for atypical orientations, reflecting the difficulty of extreme head rotations. Illumination (Figure 6c) also plays a role: detection accuracy falls from 94.61% in extreme lighting to 89.79% in normal but inconsistent illumination, suggesting sensitivity to localized shadows or uneven lighting. For facial expressions (Figure 6d), the system performs well on exaggerated expressions (98%) but records 10.01%

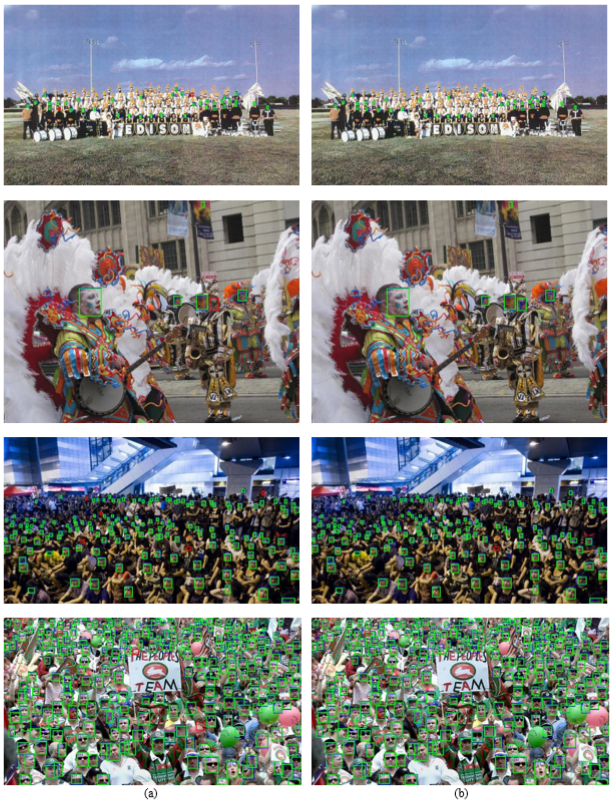


Figure 4. Face detection result (a) YOLOv11 baseline (b) YOLOv11 with SAHI (Green bounding box: True positive, red bounding box: FN)

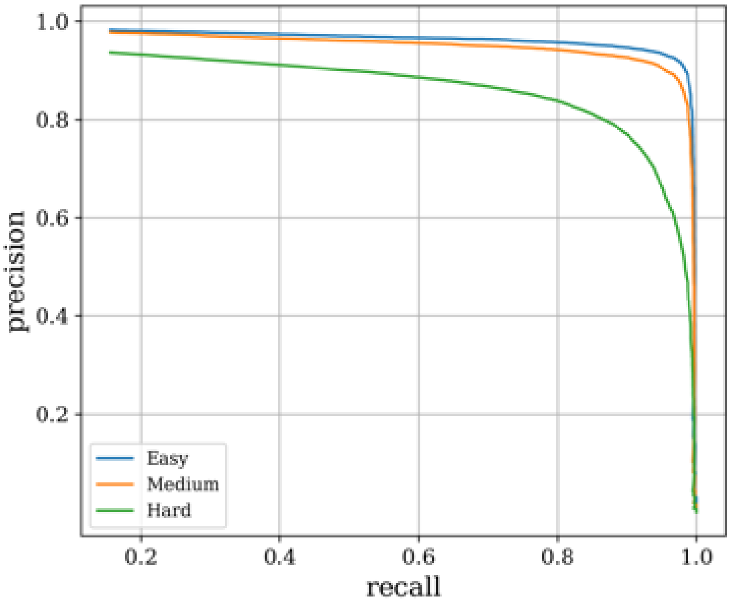
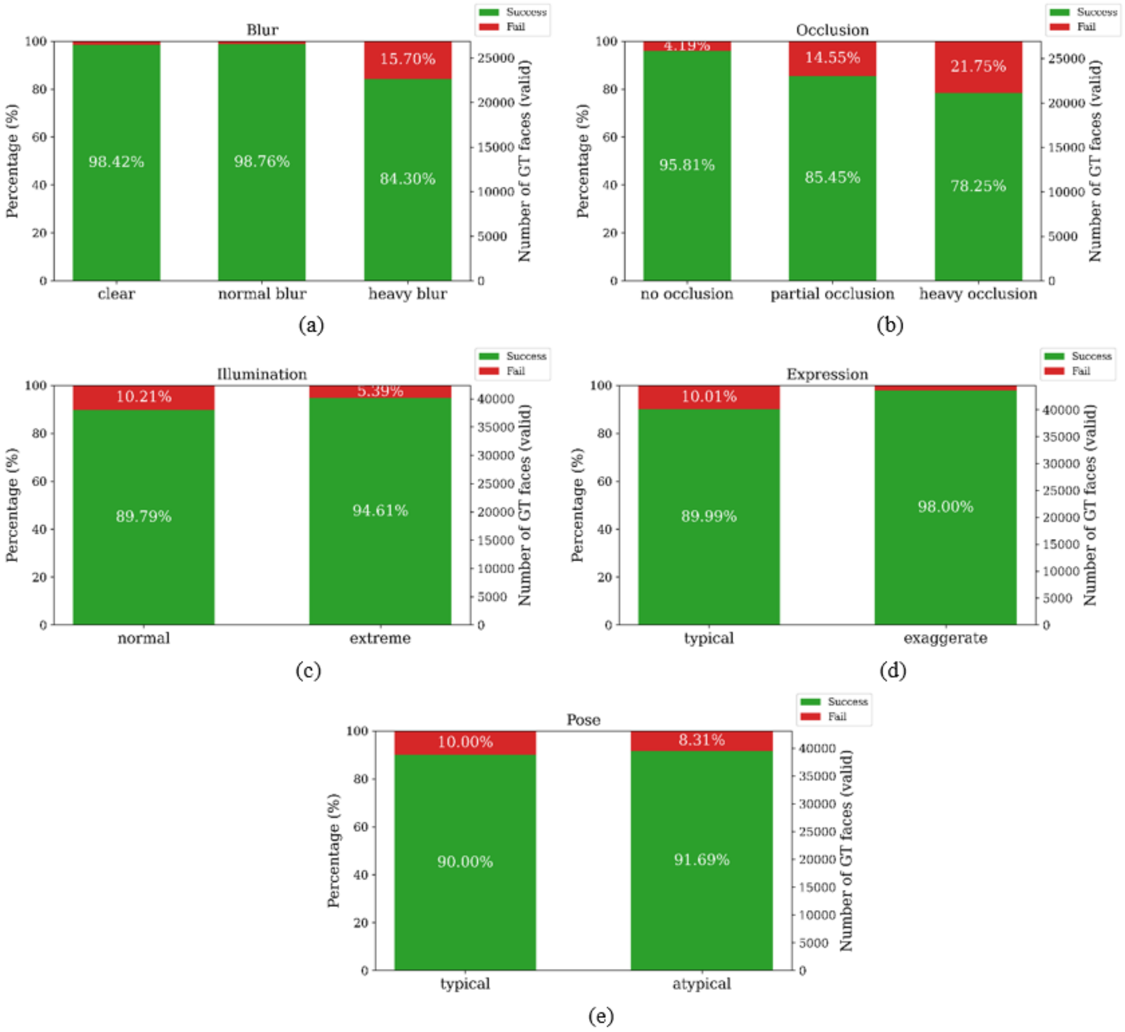


Figure 5. PR-Cure of YOLOv11-X with SAHI

failures on typical expressions, likely due to subtler variations in mouth and eye regions. Lastly, for facial pose (Figure 6e) typical poses achieved a success rate of 90.00% while the atypical poses increased slightly to 91.69% which indicates that the model performs consistently well across both categories.

Comparative analysis indicates that heavy occlusion (21.75% missed detections) represents the most severe limitation, followed by heavy blur (15.7% drop) and extreme pose shifts (10% failures). These conditions significantly degrade feature consistency, making them harder to compensate for even with slicing. In contrast, illumination and expression variations have relatively smaller impacts, suggesting that SAHI and YOLOv11 effectively leverage texture cues in such scenarios.

Overall, these findings confirm that SAHI substantially enhances the detection of small faces, particularly in dense crowd scenarios, yet persistent challenges remain under heavy occlusion, motion blur, and extreme pose variations. When compared against other state-of-the-art detectors (Table 3), the proposed YOLOv11 + SAHI framework demonstrates highly competitive performance. Specifically, it achieves 96.33% AP on Easy, 95.87% on Medium, and 90.81% on Hard, which is comparable to RetinaFace (96.71%, 96.08%, 91.44%) and surpasses popular YOLO variants such as YOLOv7 (96.11%, 95.12%, 88.14) and YOLOv5 (96.67%, 95.08%, 86.55) on the Hard subset. Moreover, the method markedly outperforms earlier architectures, including YOLOv3 Tiny (77.9% Hard AP), MTCNN (60.7% Hard AP), and Faceness (42.4% Hard AP), highlighting its robustness in detecting small-scale and occluded faces. Importantly, this level of accuracy is achieved with lower architectural complexity and greater backbone flexibility compared to traditional multi-stage frameworks such as DSFD and cascade-based methods. Future research should focus on integrating pose-invariant feature learning and adaptive tiling strategies based on face density to mitigate redundant slicing while maintaining high recall and efficiency.



(a) (b) (c) (d) (e)

Figure 6. Error analysis on several condition (a) Blur, (b) Occlusion, (c) Illumination, (d) Expression, (e) Pose

5 Conclusions

This study presented an enhanced face detection framework that integrates YOLOv11 as the primary detection backbone with Slicing Aided Hyper Inference (SAHI) to address the challenge of detecting small-scale faces in crowded scenes. The results demonstrate that the proposed method significantly improves the detection of small and crowded faces, achieving 96.33% AP on the Easy subset, 95.87% on Medium, and 90.81% on Hard. Compared with state-of-the-art detectors, YOLOv11 + SAHI performs competitively with RetinaFace and consistently outperforms YOLOv7, YOLOv5, and other lightweight architectures such as YOLOv3 Tiny and MTCNN, especially on the Hard subset, where small-scale faces, occlusion, and pose variations dominate.

Table 3. Comparison of the proposed method with state-of-the-art detectors

Network Model	AP (Easy)	AP (Medium)	AP (Hard)	Mean AP
Proposed Method	96.33	95.87	90.81	94.34
YOLOv8n [34]	-	-	-	67.60
YOLOv7 [28]	96.11	95.12	88.14	93.12
YOLOv5 [31]	96.67	95.08	86.55	92.77
YOLOv3 Tiny [30]	95.26	89.2	77.9	87.45
Multiscale Cascade [21]	69.1	63.4	34.5	55.67
DEFace [22]	87.2	85.6	75.4	82.73
Faceness [35]	71.3	66.4	42.4	60.03
MTCNN [36]	85.1	82.0	60.7	75.93
RetinaFace [23]	96.71	96.08	91.44	94.74
Li et al. [37]	87.4	84.1	67.3	79.60

The improvement is primarily attributed to SAHI's ability to preserve the relative resolution of small faces by processing overlapping high-resolution slices, allowing YOLOv11 to capture fine-grained spatial details that are often lost in conventional full-image inference. Although the slicing process introduces additional computational overhead, the resulting trade-off between speed and accuracy remains acceptable for applications prioritizing detection robustness over real-time performance.

Nevertheless, performance declines persist under severe occlusion, heavy blur, and extreme pose variations, and fixed slicing introduces redundancy in sparse images. Future work should focus on pose-invariant representation learning, occlusion-aware modeling, and adaptive slicing strategies to further improve robustness and efficiency across diverse real-world conditions.

Funding

The research funding by National Grants for Applied Research from the Ministry of Education, Culture, Research, and Technology, Indonesia (Grant No.: 090/E5/PG.02.00.PL/2024).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electron.*, vol. 9, no. 8, p. 1188, 2020. https://doi.org/10.3390/electronics9081188

[2] M. Fatchan, P. Andono, A. Affandy, and A. Fanani, "Hybrid deep autoencoder and adaboost for robust facial expression recognition," *Int. J. Comput. Methods Exp. Meas.*, vol. 13, no. 1, pp. 141–147, 2025. https://doi.org/10.18280/ijcmem.130115

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. I–I. https://doi.org/10.1109/CVPR.2001.990517

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177

[5] S. Yang, P. Luo, C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.

[6] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face R-CNN," *arXiv preprint*, 2017. https://arxiv.org/abs/1706.01061

[7] O. Cakiroglu, C. Ozer, and B. Gunsel, "Design of a deep face detector by mask R-CNN," in *Proceedings of the Signal Processing and Communications Applications Conference (SIU)*, 2019, pp. 1–4. https://doi.org/10.1109/SIU.2019.8806447

[8] W. Wu, Y. Yin, X. Wang, and D. Xu, "Face detection with different scales based on faster R-CNN," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 4017–4028, 2018. https://doi.org/10.1109/TCYB.2018.2859482

[9] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*. Springer, 2017, pp. 57–79. https://doi.org/10.1007/978-3-319-61657-5_3

[10] X. Sun, P. Wu, and S. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018. https://doi.org/10.1016/j.neucom.2018.03.030

[11] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2017, pp. 650–657. https://doi.org/10.1109/FG.2017.82

[12] Q. Chen, F. Shen, Y. Ding, P. Gong, Y. Tao, and J. Wang, "Face detection using R-FCN based deformable convolutional networks," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 4165–4170. https://doi.org/10.1109/SMC.2018.00706

[13] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," *arXiv preprint*, 2017. https://arxiv.org/abs/1709.05256

[14] B. Ye, Y. Shi, H. Li, L. Li, and S. Tong, "Face SSD: A real-time face detector based on SSD," in *Proc. Chin. Control Conf.*, 2021, pp. 8445–8450. https://doi.org/10.23919/CCC52363.2021.9550294

[15] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 951–959. https://arxiv.org/abs/1612.04402

[16] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Li, "Faceboxes: A cpu real-time face detector with high accuracy," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 1–9.

[17] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Li, "S3FD: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 192–201.

[18] J. Nirmaladevi, S. Poornachandra, P. Jayavardhini, R. Ragsana, and R. Mahalakame, "Deep learning based FPN and MT-CNN face mask detection system," in *Proceedings of the International Conference on Science, Technology, Engineering and Mathematics (STEM)*, 2024, pp. 1–5. https://doi.org/10.1109/ICONSTEM60960.2024.10568713

[19] X. Tang, D. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 812–828. https://doi.org/10.1007/978-3-030-01240-3_49

[20] J. Zhang, X. Wu, S. Hoi, and J. Zhu, "Feature agglomeration networks for single stage face detection," *Neurocomputing*, vol. 380, pp. 180–189, 2020. https://doi.org/10.1016/j.neucom.2019.10.087

[21] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170

[22] T. Hoang, G. Nam, J. Cho, and I. Kim, "Deface: Deep efficient face network for small scale variations," *IEEE Access*, vol. 8, pp. 142 423–142 433, 2020. https://doi.org/10.1109/ACCESS.2020.3012660

[23] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211. https://doi.org/10.1109/CVPR42600.2020.00525

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

[25] N. Jegham, C. Koh, M. Abdelatti, and A. Hendawi, "YOLO evolution: A comprehensive benchmark and architectural review of YOLOv12, YOLO11, and their previous versions," *arXiv preprint*, 2024. http://arxiv.org/abs/2411.00201

[26] X. Zheng, Z. Zhou, and C. Photong, "Gcsem-YOLO small scale enhanced face detector based on YOLO," *Edelweiss Appl. Sci. Technol.*, vol. 9, no. 3, pp. 840–855, 2025. https://doi.org/10.55214/25768484.v9i3.5356

[27] W. Pebrianto, P. Mudjirahardjo, and S. Pramono, "YOLO method analysis and comparison for real-time human face detection," in *Proceedings of the Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, 2022, pp. 333–338. https://doi.org/10.1109/EECCIS54468.2022.9902919

[28] A. Chan, M. Abdullah, S. Mustam, F. Poad, and A. Joret, "Face detection with YOLOv7: A comparative study of YOLO-based face detection models," in *Proceedings of the International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, 2024, pp. 105–109. https://doi.org/10.1109/GECOST60902.2024.10475115

[29] N. Vemulapalli, P. Paladugula, G. Prabhat, and S. Abhishek, "Face detection with landmark using YOLOv8," in *Proceedings of the International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, 2023, pp. 1–5. https://doi.org/10.1109/ICEFEET59656.2023.10452204

[30] J. Gao and T. Yang, "Face detection algorithm based on improved TinyYOLOv3 and attention mechanism,"

Comput. Commun., vol. 181, pp. 329–337, 2022. https://doi.org/10.1016/j.comcom.2021.10.023

[31] D. Qi, W. Tan, Q. Yao, and J. Liu, “YOLO5Face: Why reinventing a face detector,” in *European Conference on Computer Vision (ECCV)*, 2022, pp. 228–244. https://doi.org/10.1007/978-3-031-25072-9_15

[32] N. Ali, A. Alsafo, H. Ali, and M. Taha, “An effective face detection and recognition model based on improved YOLOv3 and VGG 16 networks,” *Int. J. Comput. Methods Exp. Meas.*, vol. 12, no. 2, pp. 107–119, 2024. https://doi.org/10.18280/ijcmem.120201

[33] F. Akyon, S. Altinuc, and A. Temizel, “Slicing aided hyper inference and fine-tuning for small object detection,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 966–970. https://doi.org/10.1109/ICIP46576.2022.9897990

[34] I. Al Amoudi and D. Ramli, “YOLOv7-tiny and YOLOv8n evaluation for face detection,” in *International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Robotics, Vision and Simulation (ROBOVIS)*, 2021, pp. 477–483. https://doi.org/10.1007/978-981-99-9005-4_60

[35] S. Yang, P. Luo, C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3676–3684. https://doi.org/10.1109/ICCV.2015.419

[36] J. Xiang and G. Zhu, “Joint face detection and facial expression recognition with mtcnn,” in *Proceedings of the International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 424–427. https://doi.org/10.1109/ICISCE.2017.95

[37] X. Li, Z. Yang, and H. Wu, “Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks,” *IEEE Access*, vol. 8, pp. 174 922–174 930, 2020. https://doi.org/10.1109/ACCESS.2020.3023782

Nomenclature

TP	True Positive
FP	False Positive
FN	False Negative
AP	Average Precision
IoU	Intersection over Union
AP	Average Precision