Air Pollution Level Detection from Sky Images: A YOLO26n-based Deep Learning Framework

Tanvir Ahmed [1], Brishty Das [1], Abdullah Al Imran [2], Shakila Rahman [1,*], and Jia Uddin [3,*]

[1] Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
[3] AI and Big Data Department, Endicott College, Woosong University, Daejeon, Republic of Korea
Email: 23-51316-1@student.aiub.edu (T.A.); 23-50409-1@student.aiub.edu (B.D.); imranneel114@gmail.com (A.A.I.); shakila.rahman@aiub.edu (S.R.); jia.uddin@wsu.ac.kr (J.U.)
*Corresponding author

***Abstract*—Air pollution is one of the major global challenges affecting environmental quality and human health. The conventional monitoring stations are limited in number and expensive to deploy in highly populated cities. Although fixed sensors give accurate measurements, their limited coverage restricts fine-grained spatial analysis. As a result, visual air quality estimation has gained attention by using image-based cues such as haze density, contrast reduction, and color distortion. However, most existing solutions formulate this problem as image-level classification or regression. In this paper, a lightweight air-pollution detection framework based on the official Ultralytics You Only Look Once (YOLO)26n architecture is introduced. Unlike classical whole-image classification, the problem is formulated as a single-region sky-localization detection task with the model jointly localizing the dominant sky region and identifying the sky pollution level. A systematic data processing pipeline was developed, and it includes duplicate elimination using Message-Digest Algorithm (MD5) hashing, sky area annotation, and six-class pollution level labeling. The proposed model is compared with recent object detectors like YOLOv11n, YOLOv12n, Roboflow Detection Transformer (RF-DETR), Mobile Network V3 Single Shot MultiBox Detector Lite (MobileNetV3-SSDLite), and RetinaNet, under the same experimental conditions. The experimental results demonstrate that the proposed YOLO26n-based model achieves mean Average Precision (mAP)@0.5 of 0.953 and mAP@0.5:0.95 of 0.914 with 2.51 million parameters and 5.8 Giga Floating Point Operations (GFLOPs). Overall, the framework provides a practical balance between accuracy, efficiency, and interpretability to scalable levels of monitoring air pollution.**

***Keywords*—air pollution detection, Ultralytics You Only Look Once (YOLO)26n, sky-region localization, sky image analysis**

I. Introduction

Air pollution is one of the most serious environmental issues in modern society. It impacts on human health, decreases visibility, damages ecosystems, and causes climate-related risks. Conventional air quality surveillance depends primarily on fixed monitoring stations that are highly accurate in measuring the pollutants. However, these systems are costly, difficult to deploy densely, and often unable to capture local or street-level pollution variation. Due to these constraints, there has been an increasing interest in image-based air quality estimation, in which the level of pollution is estimated by visual features such as haze density, contrast loss, and color distortion in outdoor photographs [1–3].

Previous research has demonstrated the ability of deep learning to extract relevant pollution-related patterns directly from images. Convolutional Neural Network (CNN)-based algorithms show that there are helpful cues in natural outdoor images to predict air quality [1, 2, 4], while feature-based pipelines using engineered descriptors and classical classifiers remain informative but are less robust to variations in capture conditions [3]. More recent works have explored the potential for mobile inference and scalable deployment of image-based air quality estimation [5]. In this study, the official Ultralytics You Only Look Once (YOLO)26n model is used for the detection-based formulation. YOLO26 is a real-time vision model family that supports end-to-end native inference and Distribution Focal Loss (DFL)-free box regression with a dual-head detection design [6, 7]. Despite the progress, the majority of current approaches conceive the problem as a global image-level classification or regression, which limits interpretability when the pollution evidence is spatially localized.

This limitation is overcome by object detection frameworks that offer classification and localization in a single model. YOLO introduced a single forward pass, multi-purpose real-time detection method for localization and classification [8], whereas Single Shot MultiBox Detector (SSD) demonstrated effective multi-scale detection using one forward pass [9]. RetinaNet improved dense detection by addressing the imbalance between foreground and background using focal loss [10], while transformer-based detectors like Detection Transformer

Manuscript received May 6, 2026; revised June 5, 2026; accepted July 9, 2026; published November 14, 2026.

(DETR) offered an alternative end-to-end formulation of object detection [11].

This paper proposes a lightweight, real-time, and spatially aware framework for air-pollution-level detection from sky images using Ultralytics YOLO26n. In this work, the official YOLO26n checkpoint is fine-tuned for six air-pollution classes in the sky-image domain. The proposed method aims to balance accuracy, computational efficiency, and spatial interpretability, and includes a comparative analysis of a variety of detection architectures in the same experimental conditions. The key findings of this research can be summed up as follows:

- A novel formulation of the air pollution level estimation as an object detection problem, with locally interpretable and localized predictions.
- A lightweight Ultralytics YOLO26n-based detector fine-tuned from the official checkpoint for six air-pollution classes.
- A pipeline of annotations and dataset refinement on a publicly available dataset to aid detection-based learning.
- A fixed-budget comparative analysis of state-of-the-art detection models using the same datasets, preprocessing, input-size, and training-duration settings.
- The proposed model achieves a competitive mean Average Precision (mAP)@0.5 of 0.953 while maintaining low computational complexity of 5.8 GFLOPs and lower latency than heavier models.

II. Literature Review

The image-based air quality estimation has transitioned from handcrafted feature analysis toward deep learning, multimodal fusion, and spatiotemporal forecasting. Initial studies showed that the atmosphere pollution leaves noticeable traces in outdoor pictures. CNN-based classification methods have shown that haze, loss of contrast, and degradation of the scene can be directly learned using image data to estimate pollution [1]. CNN pipelines based on ResNet have also been used to classify air pollution images, using deep feature representations to enhance discrimination in challenging visual conditions [2]. Simultaneously, feature-based estimation demonstrated that designed visual features with classical classifiers like Support Vector Machine (SVM) can be used to differentiate the level of air quality, although performance may decrease for changing illumination and environmental conditions [3].

As the field evolved, scientists started to investigate more specialized visual learning pipelines. To enhance the ability of the CNN to capture haze-sensitive information, hybrid CNN architectures were suggested to integrate Red, Green, Blue (RGB) features with auxiliary representations that are correlated with atmospheric degradation [4]. These designs demonstrate that physically significant haze indicators can be used to supplement end-to-end feature learning in cases where pollution signals are weak. In addition, optimization-based models such as CNNs which are trained with genetic algorithms prove that training policy and hyperparameter decisions can be critical in terms of generalization across datasets and environmental conditions [12]. Multimodal fusion is one of the main research directions used to understand the limitations of image-only estimation. In reality, air quality cannot be fully visualized on a single picture because factors such as humidity, temperature, and history of pollutants may also affect visual appearance. Mobile multimodal prediction frameworks that are mobile-oriented show that the integration of visual and auxiliary cues enhances reliability and can be scaled to deployment [13]. Likewise, the combination of weather variables and image-based features can be used to stabilize the estimation of particulate matter in varying weather conditions [14]. It has been confirmed in urban studies that visual models do not work in a consistent way in real city scenes and that it is important to consider context variation and that it is hard to generalize beyond a single capture domain [15]. Although multimodal systems are more robust than image-only predictors, they may add complexity to the system and demand concurrent data streams.

The conventional PM2.5 estimation techniques are still significant as they offer interpretable baselines. The combination of simple linear regression and image processing can be used to estimate PM2.5 using scene images in limited conditions [16]. Models of photo-based quality assessment based on color, contrast, and structural degradation descriptors also support the idea that image degradation associated with visibility is correlated with the level of pollution [17]. Nevertheless, they are generally sensitive to lighting, weather, and device properties, which limits their portability in a variety of real-world environments. This encourages using learned representations that can select from stable visual evidence rather than details of the scene that are not relevant.

Surveillance imagery and temporal modeling are another significant advancement. Surveillance image-based monitoring demonstrates that persistent image streams have the potential to enhance robustness since pollution is not in a fixed visual form but a dynamically changing one [18]. Combining surveillance imagery with numerical variables in time-series forecasting frameworks are also effective in enhancing the prediction of PM2.5 and PM10 by considering the continuity of time [19]. More sophisticated methods use 3D-CNNs with attention to learn spatiotemporal dependencies to predict multi-horizon [20], and attention-based Long Short-Term Memory (LSTM) show the usefulness of temporal weighting and sequence-aware learning to predict pollutants on large scales [21]. These findings indicate that the temporal context enhances stability, but it may also raise the computational cost and infrastructure demands.

The problem of generalization across cities and devices is an ongoing issue. Scalable transfer learning in mapping hyperlocal air pollution can be used to illustrate the applicability of domain shift and the necessity to adjust to other places and monitoring systems [22]. The other one is few-shot learning, which does not require huge labeled data and can be employed in cases where the training samples are scarce and in the situation of rare severe

pollution [23]. Benchmark and public datasets also provide the ability to evaluate and compare with other image-based air-quality methods in a reproducible manner [24].

Image preprocessing, in particular, dehazing, is directly related to air-quality vision, as haze is one of the indicators of pollution and a visual deformation. Guided transmission-map networks demonstrate that dehazing can enhance visibility restoration in a manner that can be useful to downstream prediction [25]. Recurrent encoder decoder designs that are attentive to nonuniform haze distributions as well as refine restoration iteratively further enhance haze removal [26]. There are also physics-guided alternatives, which involve the use of atmospheric scattering theory, to offer restoration frameworks that are based on the principles of physical modeling [27]. Although these improvements have been made, a significant portion of air-quality pipelines fail to quantitatively estimate the de-hazing impact on consistency of estimates between lighting and weather regimes.

Lastly, urban cues can be used to elucidate air-quality patterns that are not reflected by haze signals. Multimodal models that include traffic analytics show that indicators of urban activity provide complementary information that is useful in prediction [28]. This is consistent with detection-based modeling, in which object detectors are explicit localizers and can be used to extract interpretable proxies of complex scenes; recent detectors like YOLO, SSD, RetinaNet, and DETR provide the technical implementation of efficient localization-conscious inference [8–11]. Mobile-image-based city-specific air quality monitoring systems like AQE-Net also indicate the feasibility of image-based air quality classification in practice, but also the constraints of training on a single city and the limited scope of deployment without adaptation [29].

Although the above developments have been achieved, the majority of image-based air-quality methods are still based on image-level prediction, which compromises spatial interpretability and can be affected by scene clutter and uneven haze distribution [1–4, 15–17]. In order to increase robustness, multimodal and temporal models are used, but they demand more sensor, weather, traffic, or sequence data, which makes computations and deployment more complex [13, 18–21, 28]. End-to-end air-pollution recognition pipelines [25–27] do not always include restoration and dehazing methods, but they are relevant to the improvement of visibility. These gaps encourage the development of a lightweight, localization-aware detection system capable of concentrating on the predominant sky region without compromising the practical real-time performance.

III. Materials and Methods

A. Dataset Collection

The data used in this study were taken from the publicly available Air Pollution Image Dataset on the Kaggle platform [24]. It contains 12,240 images of atmospheric sky conditions from different parts of India and Nepal. The pictures have variations in lighting, weather, and atmospheric visibility, giving a wide range of samples to train the models. To allow distinguishing between the air pollution intensity in terms of visual characteristics, each image was assigned to one of six predetermined air quality classes depending on the sky look. The following classes are the different levels of atmospheric pollution:

- Good—Clear atmosphere with low level of haze and high level of atmospheric visibility.
- Moderate—There is slight haze formation, and the quality of air is fairly acceptable.
- Unhealthy for Sensitive Groups (USG)—Visible haze conditions, which can have a detrimental impact on sensitive people.
- Unhealthy—There is a significant atmospheric pollution with haze that is noticeable and a lack of sky clarity.
- Very Unhealthy—Very high pollution of the atmosphere, extreme smog, and very low visibility.
- Severe—A smog or haze observed in big parts of the sky.

These classes are air pollution levels, and not weather categories such as sunny, cloudy, rainy and foggy. The difference is based on visual atmospheric cues such as haze density, how clear the sky is, how far the vision is reduced, how much contrast is lost, and any color distortion. The intensity of pollution can vary over time and between neighboring classes. For example, Moderate and USG, or Unhealthy and Very Unhealthy, may look similar in appearance under varying light and environmental conditions. Examples of each air-pollution class are shown in Fig. 1.

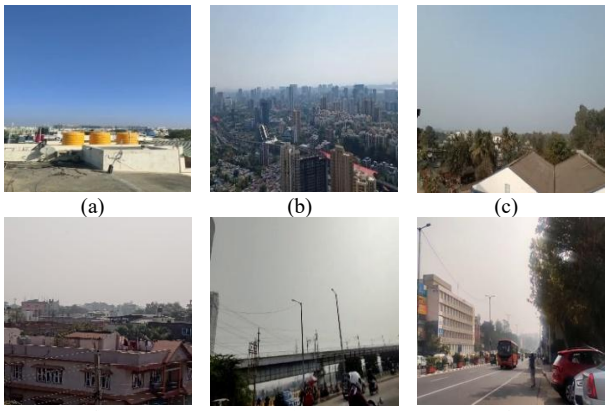


Fig. 1. Visual image of air-pollution class. (a) Good; (b) Moderate; (c) USG; (d) Unhealthy; (e) Very Unhealthy; (f) Severe.

B. Duplicate Detection and Dataset Refinement

Publicly available image datasets can have duplicated samples, which can have adverse impacts on model training and evaluation. The quality of a dataset was improved by applying a duplicate detection algorithm in order to solve this issue. The Message-Digest Algorithm (MD5) Hashing technique was employed to detect duplicate images. Each image was processed to generate a unique hash value on its content. Duplicates were considered to be images with the same hash. The hash values that were generated were compared throughout the

dataset, and duplicate images were eliminated. This ensured that each image was counted once and enhanced the consistency and reliability of the databases. This duplicate detection process detected and removed 126 duplicate images in the original dataset.

A moderate class imbalance was observed in the refined dataset, with USG having the highest count and Severe having the lowest count. For this reason, the results were further investigated in a class-wise manner by the use of the confusion matrix and precision-recall curves. This resulted in the refined data having 12,114 unique skies images shown in Table I on a class-by-class basis. Elimination of the duplicated samples enhances the diversity of the dataset and makes the training data more representative of a range of atmospheric conditions. The following processing, annotation, and training was then performed with a fine-grained dataset of 12,114 unique images.

TABLE I. CLASS WISE IMAGES IN DATASET

Class Name	Count	Total Data
Good	1540	12,114
Moderate	1573	
USG	2859	
Unhealthy	2620	
Very Unhealthy	2195	
Severe	1327	

C. Dataset Upload and Management

After eliminating the duplicated images, the data was uploaded to Roboflow for effective preprocessing. The dataset was prepared to do a six-class object detection task of air pollution. Roboflow ensured consistency in the management of data, controlled augmentation, and reproducible splits, which made it possible to compare models. Its flexibility in exporting also allowed it to be easily combined with other deep learning models.

D. Annotation Process

The images were added to Roboflow and then annotated to be utilized to train an object detection model. The annotation procedure entailed identifying the sky region in each image and assigning the corresponding class label. To show the atmospheric condition in which the image was captured, the area of the sky that seemed to dominate was put in a bounding box.

The examples of the original and annotated images are shown in Fig. 2. One bounding box was annotated on each image because each sky image contains one dominant air-pollution condition. The aim of the bounding box is not to detect multiple pollution objects, because air pollution is an atmospheric condition and not a discrete object. Instead, the box restricts the area of the sky that is most visible where the haze, visibility degradation, contrast loss, and color distortion are most prominent. This makes the task a localization-aware detection problem instead of the usual multi-object detection task. Images were examined where the sky was not clearly visible or was slightly blurred, with care taken to ensure consistency in the annotations. This annotation technique ensures that the model is attuned to atmospheric pollution cues while minimizing the impact of other foreground elements like buildings, trees, roads, and vehicles.



Fig. 2. Original and bounding box annotated images. (a) Original; (b) Annotated; (c) Original; (d) Annotated.

E. Dataset Splitting

Once the annotation was done, the dataset was divided into training, validation, and testing subsets to ensure reliable evaluation and to avoid data leakage. The data was split into: 70% of the data was to be used in training, 20% in validation, and 10% in testing. Model parameters were learned using the training set. The validation set was used to monitor the performance of the model during training and to identify potential overfitting. Final evaluation on unseen data was done on the test set. Fig. 3 shows the whole process of dataset preparation, such as removal of duplicates, annotation, and split of the dataset.

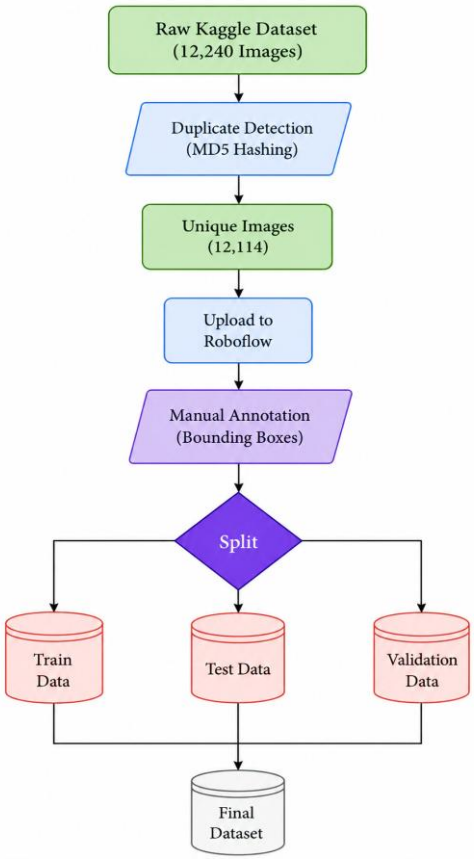


Fig. 3. Data pre-processing diagram.

F. Proposed Pollution Detection Framework

The proposed framework for air-pollution detection in the sky region is shown in Fig. 4, using the official Ultralytics YOLO26n detector. The nano-scale version of the YOLO26 model family is called YOLO26n and is loaded using the public version of the YOLO26 checkpoint, yolo26n.pt, in the Ultralytics/PyTorch framework [6, 7]. The model used in this study is not an internally developed detector and is not a renamed version of a previous architecture of YOLO. Instead, the official YOLO26n detector is trained on six levels of air pollution: Good, Moderate, USG, Unhealthy, Very Unhealthy, Severe. YOLO26 is a modified version of previous YOLO architectures, primarily in terms of its detection head and training/inference design. In the official YOLO26 technical report, a dual-head design is introduced for native end-to-end inference, and DFL is eliminated, which makes the detection head lighter and the regression of boxes simpler [6]. The official Ultralytics documentation also highlights NMS-free inference, DFL-free regression, MuSGD optimization, Progressive Loss, and Small-Target-Aware Label Assignment (STAL) as some of the improvements added to YOLO26 [7]. In the official configuration of YOLO26, the end2end option is enabled with 'True', and the DFL-free regression setting is supported with the reg_max option set to 1 [30].

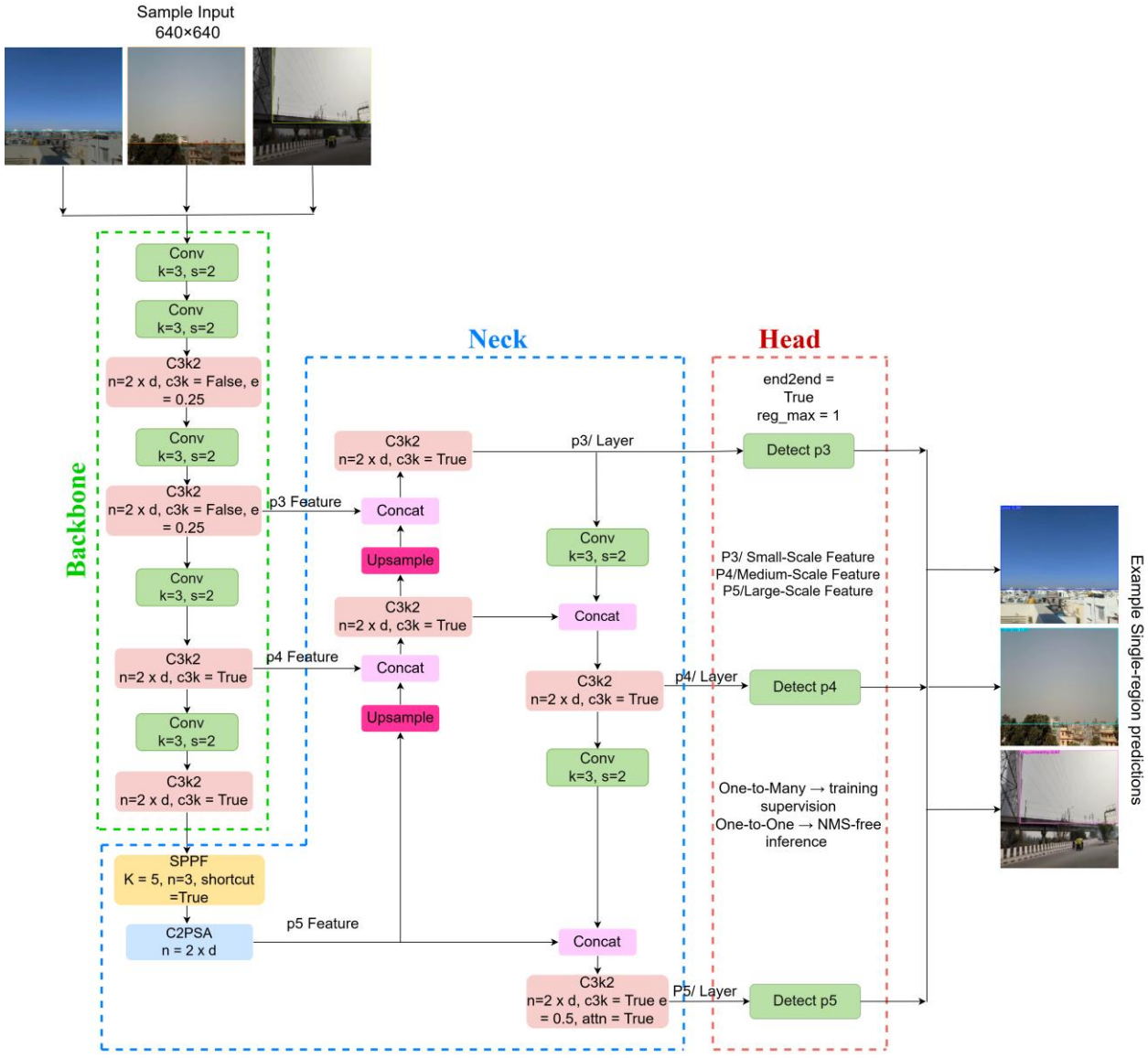


Fig. 4. Proposed YOLO26n-based sky region air-pollution detection framework.

The framework mainly consists of a backbone, neck and detection head. The input sky images are scaled to 640 × 640 pixels and, with the help of convolutional and C3k2 blocks, hierarchical features are extracted in the backbone. Spatial Pyramid Pooling-Fast (SPPF) and Cross Stage Partial with Parallel Spatial Attention (C2PSA) modules also enhance representation of the contextual features on a range of spatial scales [30, 31]. Then feature fusion is done in the neck at multiple scales, through the up sampling and concatenation of channels, as well as C3k2 based operations. It generates three feature outputs P3, P4, and P5, passed to the YOLO26 detection head.

The detection head estimates the Air-Pollution Class, Confidence Score, and the Dominant Sky-Region

Bounding Box at the local scale. Detect P3, Detect P4, and Detect P5 are the three scale-specific prediction paths of the unified detection head YOLO26. Each image contains one dominant atmospheric state, which means that only one prediction for the sky-region is expected for each image. So, the proposed framework is not a typical multi-object pollution detector, but it is a single region localization-aware air-pollution detection approach.

IV. Result and Discussion

In this section, the proposed framework based on YOLO26n to detect air pollution is experimentally evaluated. The results of the YOLO26n performance are compared to other object detectors, such as YOLOv11n, YOLOv12n, RF-DETR, MobileNetV3-SSDLite, and RetinaNet. These models are compared based on conventional performance metrics of precision, F1-Score, mean Average Precision at Intersection over Union (IoU) threshold 0.5 (mAP@0.5), and mean Average Precision at several IoU thresholds (mAP@0.5:0.95), and computational complexity in Giga Floating-Point Operations (GFLOPs). The same data splitting, preprocessing pipeline, image size, training time, and evaluation metrics were used for all models. Hence, it is important to understand that the comparison is a fixed-budget comparison, rather than a fully architecture-optimized benchmark for each model.

A. Hyperparameters

The hyperparameters are crucial in determining the behavior of deep learning models both in performance and convergence. The hyperparameters were kept the same in this paper and in all the models in order to provide a comparison. The models were trained with a batch size of 16, which is a trade-off between the efficiency of computing and the generalization of the model. The model was trained for 50 epochs, and this was sufficient to learn without overfitting.

Initialization of the models was done with pre-trained weights (yolo26n.pt), which had been trained on a large-scale dataset. This pre-setting helps to speed up convergence as well as enhance the detection abilities by taking advantage of already trained feature representations. The optimizer was set to the automatic selection (auto) of the Ultralytics framework, which chooses a suitable optimization strategy in training, usually a Stochastic Gradient Descent (SGD) with momentum. The process of updating the weights was controlled by an initial learning rate of 0.01, and the momentum of 0.937 was used to stabilize the gradient updates. In order to eliminate overfitting, regularization of the model was performed with a weight decay of 0.0005.

In order to actively manage the training process, a patience value was set as 100 to carry on training even though an improvement was not seen. Also, Automatic Mixed Precision (AMP) was enabled as well to enhance the performance of calculations and to minimize the memory used. All the experiments were implemented on the Ultralytics framework on Google Colab with an NVIDIA Tesla T4 to ensure the effectiveness of training and the repeatability of the results. A summary of the training hyperparameters of this model that was trained are presented in Table II.

TABLE II. Hyperparameters Used to Train YOLO26n Model

Hyperparameters	Value
Batch size	16
Number of epochs	50
Optimizer	Auto
Pre-trained	yolo26n.pt
Learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Patience	100
Image size	640 × 640
Framework	Ultralytics (PyTorch)
Hardware	NVIDIA Tesla T4 (Google Colab)

B. Model Evaluation

The performance of the proposed YOLO26n-based air-pollution detection model was evaluated using standard metrics, including precision, recall, F1-Score, mAP@0.5, and mAP@0.5:0.95. These metrics provide a complete evaluation of the model for detecting and classifying the atmospheric pollution levels under varied environmental conditions.

The evaluation results demonstrate that the proposed model has a good detection performance, which indicates that it is capable of learning visual patterns that can be associated with different types of pollution. The accuracy of detection is increased under different conditions in the sky by the multi-scale feature representation.

Along with accuracy, there was an analysis of the properties of computation in the model. The lightweight architecture of YOLO26n has 260 layers and about 2.51 million parameters, but only 5.8 GFLOPs which is impressive considering its size.

TABLE III. Parameters of Evaluated Model

Parameters	Value
Model Layers	260
Total parameters	2.51 M
Trainable parameters	2.51 M
GFLOPs	5.8
Model size	5.1 MB
Average total latency	55.41 ms/image
FPS	18.05

Even more importantly, the model takes approximately 5.1 MB of storage and has an inference time of 55.41 ms per image, which translates to 18.05 Frames Per Second (FPS). Up to 200 test images were used, following 30 warm-up images, to test the speed. The latency was measured for batch size 1 and CUDA synchronization. The reported 55.41 ms/image is the total latency for preprocessing, model inference, and postprocessing, and FPS was computed from this mean latency. No explicit AMP autocast was used during this speed benchmark, but it was used during training. These results show that the recommended model provides a reasonable comparison between detection performance and computational power performance and can be used in monitoring air pollution in real-time.

The structural and computational characteristics of the model are given in Table III.

C. Analysis of Results

In this case, the results of different object detection models experimented in this paper, including YOLOv11n, YOLOv12n, YOLO26n (proposed), MobileNetV3-SSDLite, RetinaNet and RF-DETR are compared. The models were tested using standard evaluation measures of precision, F1-Score, mAP@0.5, and mAP@0.5:0.95, with the results presented in Table IV.

TABLE IV. PERFORMANCE OF YOLOV11N, YOLOV12N, YOLO26N, MOBILENETV3-SSDLITE, RETINANET AND RF-DETR

Model	Params (M)	mAP@0.5	mAP@0.5:0.95	Precision	F1-Score	GFLOPs
YOLOv11n	2.58	0.950	0.909	0.955	0.915	6.3
YOLOv12n	2.56	0.949	0.912	0.968	0.923	6.5
Proposed YOLO26n	2.51	0.953	0.914	0.967	0.919	5.8
RF-DETR	30.1	0.955	0.912	0.959	0.947	136
MobileNetV3-SSDLite	3.79	0.92	0.863	0.981	0.953	2.15
RetinaNet	32.27	0.93	0.883	0.762	0.825	128.31

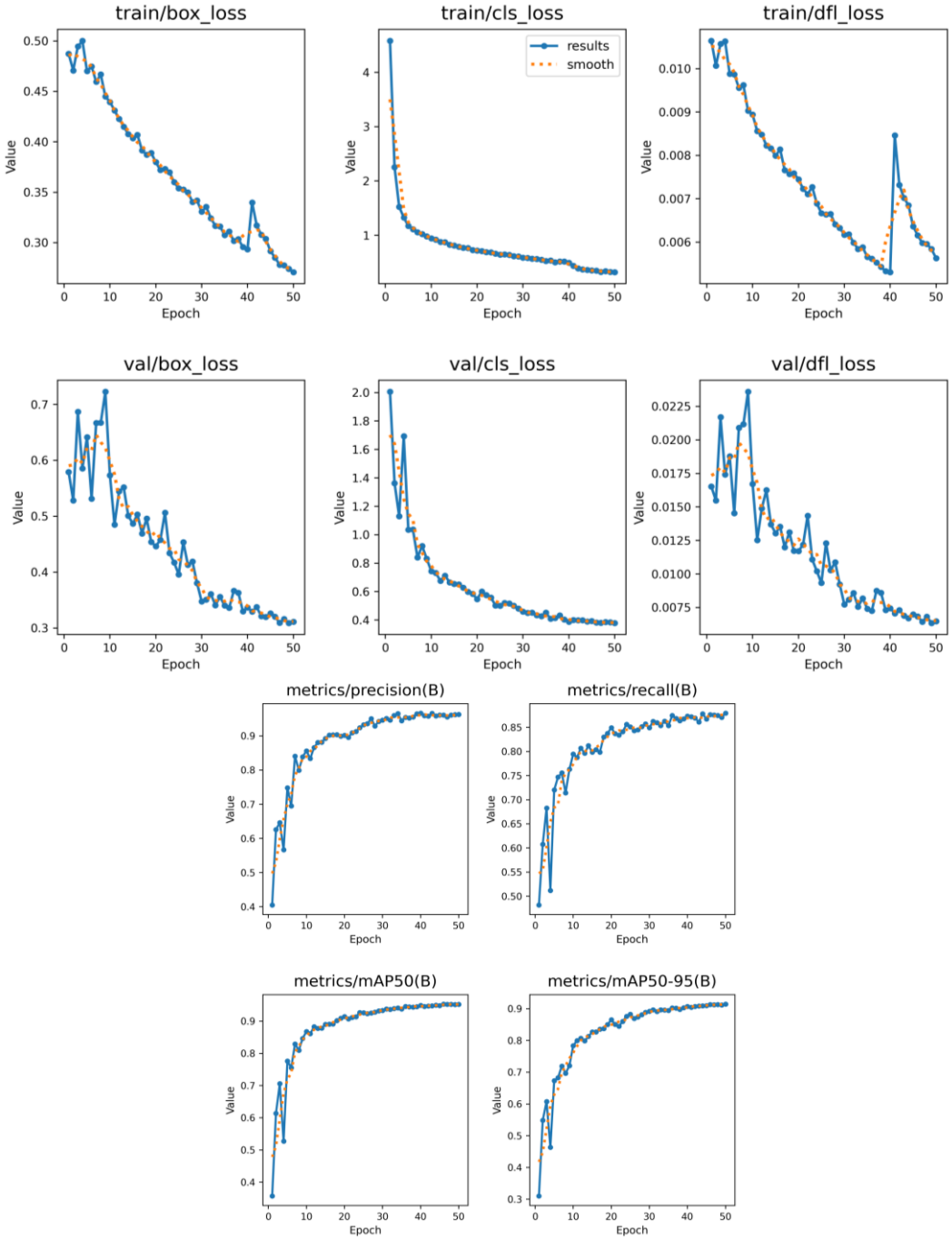


Fig. 5. YOLO26n based training and validation performance curves.

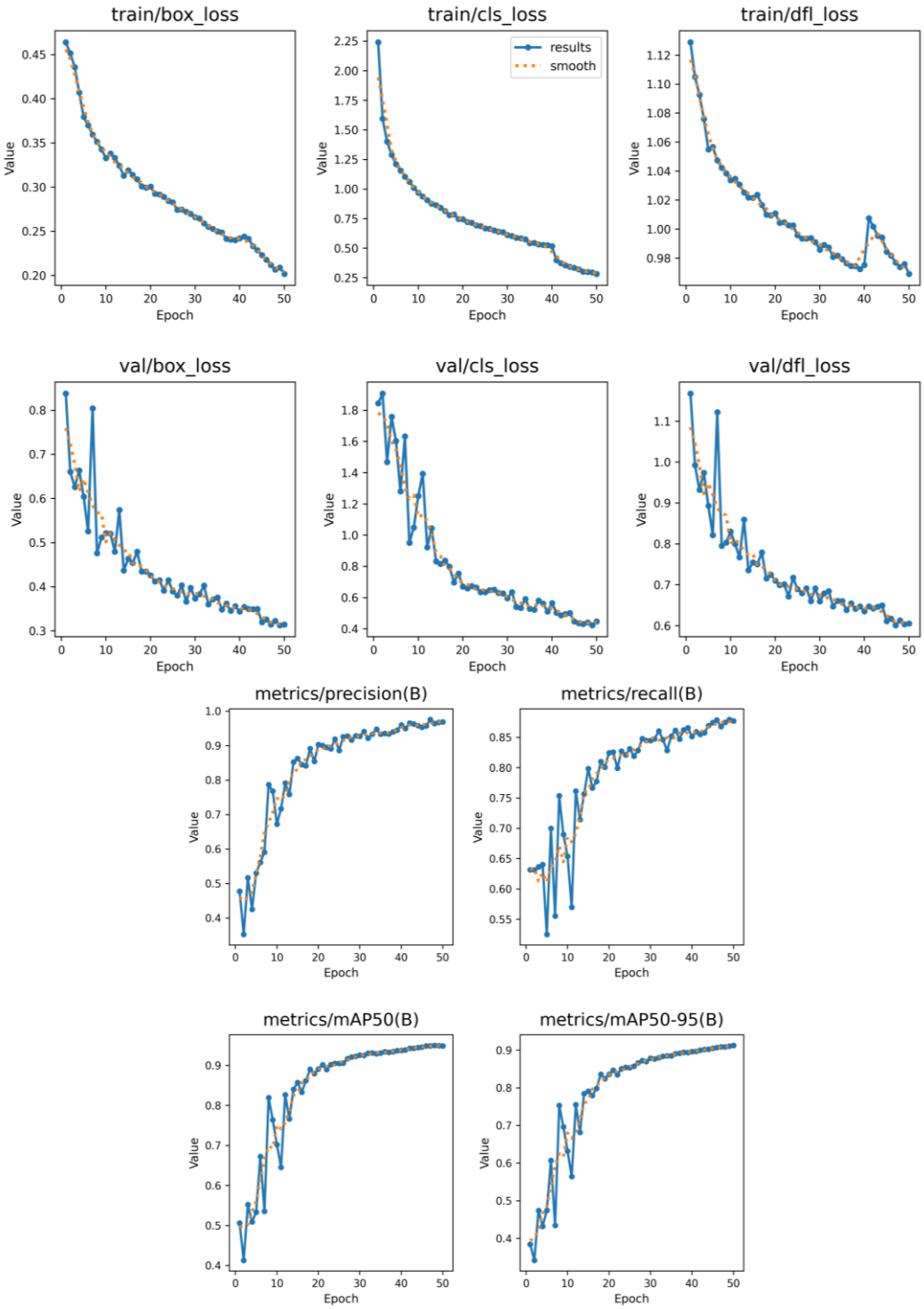


Fig. 6. YOLOv12n based training and validation performance curves.

Under the fixed-budget experimental setting, the proposed YOLO26n model shows a strong trade-off between detection performance and computational efficiency. YOLO26n has a mAP at 0.5 of 0.953 and a mAP at 0.5:0.95 of 0.914, which is competitive or better than most of the models considered. Though RF-DETR marginally wins over YOLO26n in mAP at 0.5 (0.955), it is much more resource-intensive, having more than 30 million parameters and 136 GFLOPs, and is therefore not as efficient in real-time usage.

The proposed YOLO26n, compared to the previous versions of YOLO v11n and v12n, has a higher mAP at 0.5:0.95, which is a measure of the generalization of the model at various IoU thresholds. Also, YOLO26n has an equal or even higher F1-Score, which validates its capacity to effectively strike the balance between precision and recall.

Lightweight models like MobileNetV3-SSDLite are more precise but have lower mAP scores, meaning that they are limited by their detection accuracy in general.

Conversely, more advanced models like RetinaNet offer average performance but are much more expensive to compute, and thus not as useful to run in practice.

While an ablation of the internal YOLO26n component is not conducted, the fixed-budget comparative evaluation delivers a model-level ablation-related analysis. The same dataset split, input size, training time, and evaluation metrics are used for all of the detectors, which allows for the evaluation of the impact of using YOLO26n for detection accuracy and computational efficiency. The results indicate that YOLO26n has a strong balance between mAP performance and model complexity for the sky-image air-pollution detection task. One of the most significant features of the suggested YOLO26n model is that it is a lightweight model with around 2.51 million parameters and does not need more than 5.8 GFLOPs, despite being competitive in terms of accuracy.

To illustrate the training behavior and the stability of convergence, Figs. 5 and 6 are provided for the proposed YOLO26n model and the baseline YOLOv12n. The training loss and validation loss curves, as well as the precision, recall, and mAP curves, are plotted against epochs. They are employed to determine if the models learned in a similar fashion, if the validation metrics settled, and if there was a clear overfitting trend in training. The YOLO26n curves show decreasing loss and gradually stable detection metrics, suggesting stable learning behavior under the fixed-budget training setup. The YOLOv12n curves are provided as a comparative baseline to demonstrate the performance of another light model, YOLO, under the same experimental circumstances.

The training behavior is also an illustration of the advantages of the proposed model. The reduction in jagged curves and the decrease in fluctuations observed in YOLO26n also suggest that the system is more optimally stable and more efficient to learn. These properties can help to ensure more dependable performance, especially in demanding conditions when there are minor variations in atmospheric conditions.

Also, the proposed YOLO26n model has an acceptable balance of detection accuracy and computational efficiency, compared to the other architecture in question. Part of these models only achieve small improvements on some of these metrics yet are much more computationally expensive. On the other hand, the lightweight architecture of YOLO26n and its good performance is more feasible to use.

Overall, the results show that the proposed YOLO26n model provides a more appropriate trade-off between detection performance, model complexity, and computation efficiency, which implies that the proposed solution is an appropriate and efficient one for real-time air pollution detection systems.

D. Visualization

In this section, qualitative and visual analysis of the proposed pollution detection model based on YOLO26n is provided. The pictorial findings offer a better understanding of the classification behavior of the model, the consistency of detection, and the overall performance of the model at varying levels of pollution.

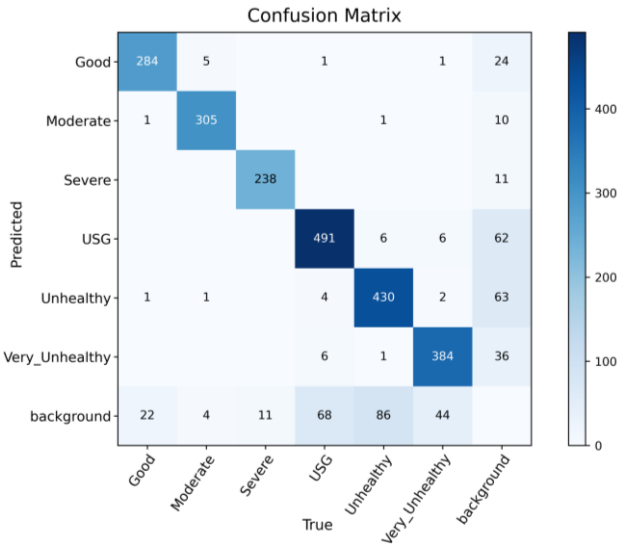


Fig. 7. Confusion matrix of YOLO26n.

Fig. 7 shows the confusion matrix of the classification in six categories of six types of pollution and the background of the classes. The model is quite well-developed in terms of diagonal dominance; in other words, the majority of the classes show a high number of correct predictions. Specifically, there is high classification accuracy in the USG, Unhealthy, and Very Unhealthy classes. There are, however, certain misclassifications in the transition between adjacent pollution levels, including Moderate and Severe, which is anticipated since the levels are visually similar. Additionally, a portion of the samples is approximated to be background, meaning that it is to some extent challenging to distinguish low-contrast sky regions.

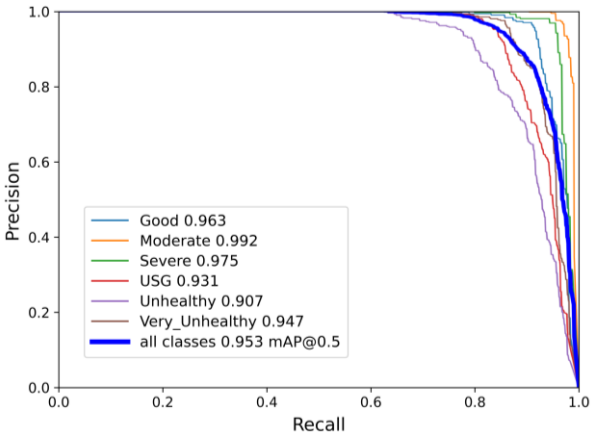


Fig. 8. Precision recall curve of all classes for YOLO26n.

Fig. 8, which shows the Precision-Recall curve, also shows the strength of the proposed model. All the curves are close to the upper-right corner, which implies high precision and recall scores. The overall mean Average Precision (mAP) of 0.5 of about 0.953, illustrating the high detection performance of the model. Although the performance of most classes is always high, there is a slight decline in the performance of the Unhealthy class, as this could be due to slight visual variations and environmental variations.

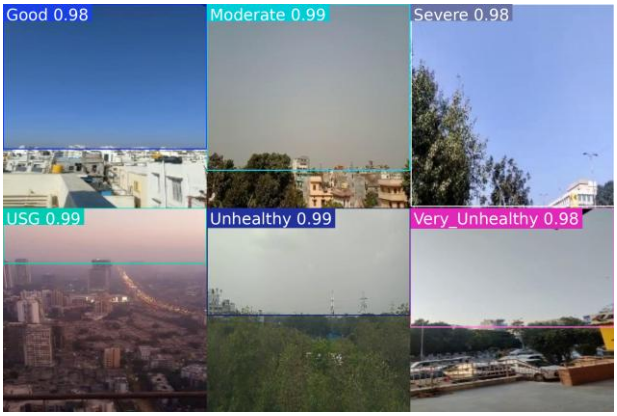


Fig. 9. Class-wise prediction of the proposed model.

Fig. 9 shows the sample outputs of the model in various categories of pollution in order to have a clearer idea of the class-wise predictions. The model is able to determine and categorize sky regions with high confidence levels, indicating that the model can generalize under different atmospheric conditions. The localization of the bounding boxes is precise to the pertinent sky areas, and the predicted labels are close to visual expectations.

Fig. 10 shows a mosaic visualization of the predictions from several test images; it is not multiple detections in an image or frame. A single predicted bounding box for the dominant sky region and predicted pollution class is contained in each sub-image. The figure is meant to reflect the behavior of the model's single-region prediction given a wide range of different sky appearances and lighting conditions, and backgrounds.

Lastly, Fig. 11 gives a comparison between the ground truth annotations and model predictions. The labels are highly localized, with the estimated bounding boxes equally corresponding to the labeled areas. In a few cases, particularly in low contrast or hazy ones, there are little deviations, but the general fit is the same. This indicates that the model has learnt useful spatial and visual properties of the data.

The qualitative results presented in Figs. 10 and 11 corroborate the quantitative results, as the model can localize the relevant sky area and can classify the sky pollution level in various visual conditions. Furthermore, the results show that the proposed localization-aware approach is indeed useful since the predicted bounding boxes contain the region of the image used for the air-pollution-level recognition, and not just a single class label for the entire image.

Overall, the visualization findings confirm the usefulness of the proposed YOLO26n model to correctly identify and classify the air-pollution levels from sky images. The qualitative results are stable, and the high-level of quantitative results confirms the ability of the model to be applied in real-world environmental monitoring applications.

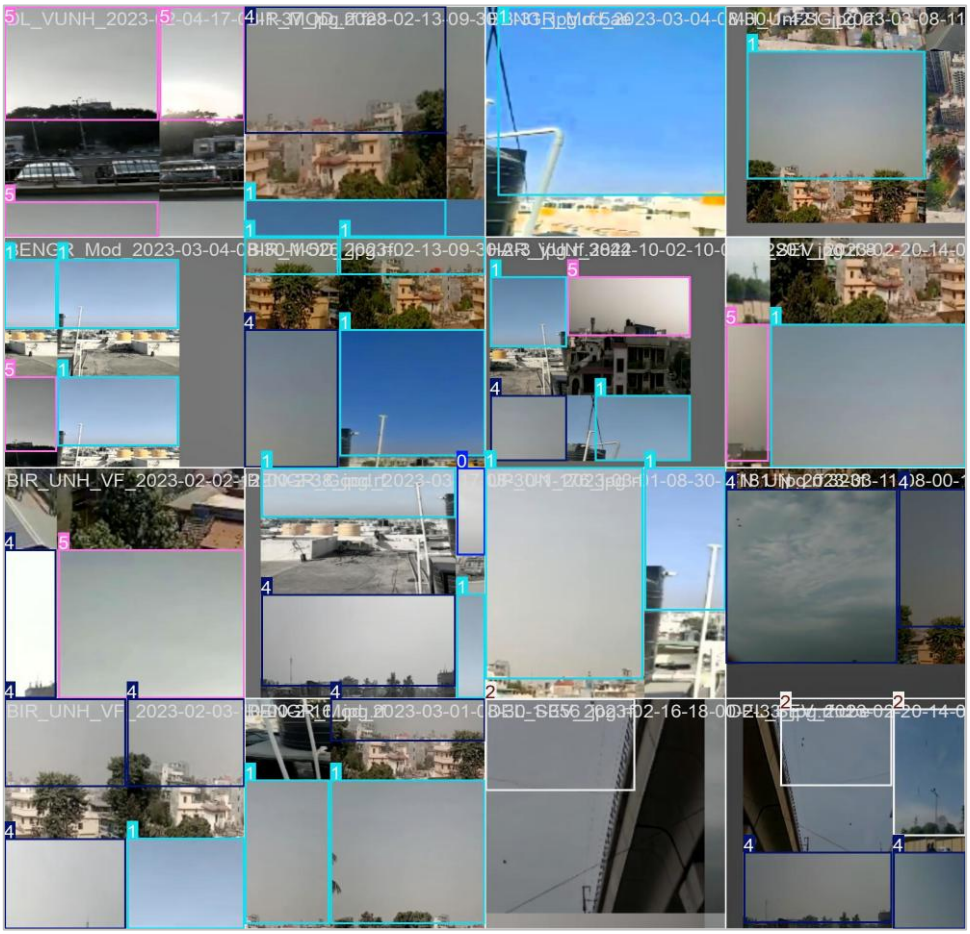


Fig. 10. Mosaic visualization of single sky-region predictions across multiple independent test images using the YOLO26n-based model.



Fig. 11. comparison of annotations and model predictions using YOLO26n-based model: (a) Label data; (b) Model prediction.

V. Conclusion

This study proposed a framework for detecting air-pollution levels from sky images using Ultralytics YOLO26n. The proposed method is based on a localization-aware detection problem to localize the dominant sky region and classify it into one of six pollution categories, as opposed to conventional whole-image classification methods. The formulation increases the interpretability as it shows the region of an image used to predict the class label rather than just a label for the whole image.

The experimental results indicate that the YOLO26n-based model achieves a strong detection performance with low computational cost. With the fixed-budget experimental setting, the model was able to obtain competitive mAP metrics while keeping the model structure as light as possible (low parameters, low GFLOPs). This makes it suitable for possible practical applications like mobile monitoring, edge computing, and camera-based environmental observation on a large scale. The visual results also confirm this practical approach—the model yields meaningful predictions of the sky regions, which are associated with the pollution level decision with the corresponding atmospheric region in the picture.

It is important to recognize that this study has a few limitations, however. Changes in lighting, weather, scene composition, or low-contrast sky conditions can affect the model. Additionally, the data is moderately imbalanced, with a class imbalance in the Severe category that could

affect recognition of minority pollution levels. Furthermore, the annotation scheme consists of a single dominant sky region bounding box for each image, and hence the framework is based on assessment of a single sky region sky-localization detection, not multi-region atmospheric pollution mapping. Note also that the cross-model comparison was performed using a constant training budget, meaning that the reported results may not be the best possible performance of each baseline architecture.

Future work will focus on improving robustness by using multimodal information including temporal image sequence, low-cost sensor readings, and meteorological parameters. Further studies could also explore a more comprehensive set of class-balancing techniques, a wider geographic set of data, and more detailed component-wise ablation tests to decouple the impact of each architectural and annotation decision. Overall, the proposed framework shows that lightweight detection models can provide an interpretable and scalable direction for vision-based air-pollution monitoring.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Tanvir Ahmed: Performed the experimental setup, dataset preparation, model implementation, data analysis, and wrote the initial manuscript. Brishty Das: Contributed to the literature review, data pre-processing, annotation, and assisted in result analysis. Abdullah Al Imran: performed dataset management and wrote the final manuscript. Shakila Rahman: Served as the research supervisor, provided continuous guidance, critical feedback on every step, and reviewed the manuscript. Jia Uddin: funded the research, reviewed the final manuscript, and contributed to the visualization. All authors have read and approved the final version of the manuscript.

Funding

This research is funded by Woosong University Academic Research 2026.

References

[1] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis using deep convolutional neural network," in *Proc. 2017 IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3949–3952. doi: 10.1109/ICIP.2017.8297023

[2] V. Lavanya, N. Sunny, V. Jyothika, and T. Ravindra Babu, "Air pollution detection using Resnet-50," in *Proc. 2023 7th Int. Conf. Computing Methodologies and Communication (ICCMC)*, 2023, pp. 845–849. doi: 10.1109/ICCMC56507.2023.10083837

[3] Q. Li and B. Xie, "Image-based air quality estimation," in *Proc. Pattern Recognition and Computer Vision (PRCV 2019)*, LNCS 11859. Cham, Switzerland: Springer, 2019, pp. 161–171. doi: 10.1007/978-3-030-31726-3_14

[4] J. Ma, K. Li, Y. Han, and J. Yang, "Image-based air pollution estimation using hybrid convolutional neural network," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 471–476. doi: 10.1109/ICPR.2018.8546004

[5] J. J. Mondal *et al.*, "Uncovering local aggregated air quality index with smartphone captured images leveraging efficient deep convolutional neural network," *Sci. Rep.*, vol. 14, 1627, 2024. doi: 10.1038/s41598-023-51015-1

[6] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, "Ultralytics YOLO26: Unified real-time end-to-end vision models," arXiv Preprint, arXiv:2606.03748, 2026.

[7] Ultralytics. (2026). Ultralytics YOLO26. *Ultralytics Doc*. [Online]. Available: https://docs.ultralytics.com/models/yolo26/

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91

[9] W. Liu, D. Anguelov, D. Erhan *et al*., "SSD: Single shot multibox detector," in *Proc. Computer Vision–ECCV 2016*, LNCS 9905. Cham, Switzerland: Springer, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.324

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Computer Vision—ECCV 2020*, LNCS 12346. Cham, Switzerland: Springer, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13

[12] A. A. Khan, M. A. Majid, and A. Dandoush, "Image-based air quality estimation using convolutional neural network optimized by genetic algorithms: A multi-dataset approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 3, 2025. doi: 10.14569/IJACSA.2025.01603113

[13] H. Madokoro and S. Nix, "Multimodal particulate matter prediction: enabling scalable and high-precision air quality monitoring using mobile devices and deep learning models," *Sensors*, vol. 25, no. 13, 4053, 2025. doi: 10.3390/s25134053

[14] Z. Zhu, R. Zhao, J. Ni, and J. Zhang, "Image and spectrum based deep feature analysis for particle matter estimation with weather information," in *Proc. 2019 IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 3427–3431. doi: 10.1109/ICIP.2019.8803713

[15] A. Feldman, S. Kendler, J. Marshall, M. Kushwaha, V. Sreekanth, A. R. Upadhya, P. Agrawal, and B. Fishbain, "Urban air-quality estimation using visual cues and a deep convolutional neural network in Bengaluru (Bangalore), India," *Environ. Sci. Technol.*, vol. 58, no. 1, pp. 480–487, 2024. doi: 10.1021/acs.est.3c04495

[16] J.-J. Liaw, Y.-F. Huang, C.-H. Hsieh, D.-C. Lin, and C.-H. Luo, "PM2.5 concentration estimation based on image processing schemes and simple linear regression," *Sensors*, vol. 20, no. 8, 2423, 2020. doi: 10.3390/s20082423

[17] K. Sun, L. Tang, S. Huang, and J. Qian, "A photo-based quality assessment model for the estimation of PM2.5 concentrations," *IET Image Process.*, vol. 16, no. 4, pp. 1008–1016, 2022. doi: 10.1049/ipr2.12201

[18] X. Wang, M. Wang, X. Liu, Y. Mao, Y. Chen, and S. Dai, "Surveillance-image-based outdoor air quality monitoring," *Environ. Sci. Ecotechnol.*, vol. 18, 100319, 2024. doi: 10.1016/j.ese.2023.100319

[19] Y. Wu, X. Wang, M. Wang, X. Liu, and S. Zhu, "Time-series forecasting of PM2.5 and PM10 concentrations based on the integration of surveillance images," *Sensors*, vol. 25, no. 1, 95, 2025. doi: 10.3390/s25010095

[20] K. Elbaz, W. M. Shaban, A. Zhou, and S.-L. Shen, "Real time image-based air quality forecasts using a 3D-CNN approach with an attention mechanism," *Chemosphere*, vol. 333, 138867, 2023. doi: 10.1016/j.chemosphere.2023.138867

[21] S. Sekar and Z. Wei, "Air quality prediction-based big data analytics using Hebbian concordance and attention-based long short-term memory," *Sci. Rep.*, vol. 15, 28719, 2025. doi: 10.1038/s41598-025-09508-8

[22] Z. Yuan, J. Kerckhoffs, H. Li, J. Khan, G. Hoek, and R. Vermeulen, "Hyperlocal air pollution mapping: A scalable transfer learning LUR approach for mobile monitoring," *Environ. Sci. Technol.*, vol. 58, no. 32, pp. 14372–14383, 2024. doi: 10.1021/acs.est.4c06144

[23] D. C. Pham, T. D. Ngo, and H. N. Vu, "Image-based air quality estimation by few-shot learning," *Comput., Mater. Continua*, vol. 84, no. 2, pp. 2959–2974, 2025. doi: 10.32604/cmc.2025.064672

[24] A. Rouniyar. (2023). Air pollution image dataset from India and Nepal. *Kaggle*. [Online]. Available:

https://www.kaggle.com/datasets/adarshrouniyar/air-pollution-image-dataset-from-india-and-nepal
[25] L.-A. Tran, S. Moon, and D.-C. Park, "A novel encoder-decoder network with guided transmission map for single image dehazing," arXiv Preprint, arXiv:2202.04757, 2022.
[26] S. Yin, Y. Wang, and Y.-H. Yang, "Attentive U-recurrent encoder-decoder network for image dehazing," *Neurocomputing*, vol. 437, pp. 143–156, 2021. doi: 10.1016/j.neucom.2020.12.081
[27] L. Hu, P. Wang, G. Gao, C. Wang, and Z. Zheng, "A PDE-based image dehazing method via atmospheric scattering theory," arXiv Preprint, arXiv:2506.08793, 2025.
[28] S. Hameed, A. Islam, K. Ahmad, S. B. Belhaouari, J. Qadir, and A. Al-Fuqaha, "Deep learning based multimodal urban air quality prediction and traffic analytics," *Sci. Rep.*, vol. 13, 22181, 2023. doi: 10.1038/s41598-023-49296-7
[29] M. Ahmed, Y. Shen, M. Ahmed, Z. Xiao, P. Cheng, N. Ali, A. Ghaffar, and S. Ali, "AQE-Net: A deep learning model for estimating air quality of Karachi city from mobile images," *Remote Sens.*, vol. 14, no. 22, 5732, 2022. doi: 10.3390/rs14225732
[30] Ultralytics. (2026). YOLO26 model configuration file: yolo26.yaml. *GitHub Repository*. [Online]. Available: https://github.com/ultralytics/ultralytics/blob/main/ultralytics/cfg/models/26/yolo26.yaml
[31] P. Hidayatullah and R. Tubagus, "YOLO26: A comprehensive architecture overview and key improvements," arXiv Preprint, arXiv:2602.14582, 2026.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).