AQFusionNet: Robust Multimodal Deep Learning for Air Quality Index Prediction through Atmospheric Imagery and Environmental Sensor Integration

Koushik Ahmed Kushal
Department of Computer Science and Engineering
American International University-Bangladesh
kushalka@clarkson.edu

Abdullah Al Mamun
Department of Electrical and Computer Engineering
Clarkson University, NY
mamuna@clarkson.edu

***Abstract*—Air pollution monitoring in resource-constrained regions presents significant challenges due to sparse sensor deployment and infrastructure limitations. This paper introduces AQFusionNet, a novel multimodal deep learning framework designed to robustly predict the Air Quality Index (AQI) by synergistically integrating atmospheric imagery with environmental sensor data. The proposed framework employs a dual-objective learning architecture, utilizing lightweight Convolutional Neural Network (CNN) backbones (MobileNetV2, ResNet18, EfficientNet-B0) to extract discriminative visual features from ground-level atmospheric images. These visual features are subsequently fused with pollutant concentration measurements through semantically-aligned embedding spaces. Our approach demonstrates superior performance across all backbone configurations, with the EfficientNet-B0 variant achieving optimal results of 7.70 RMSE and 92.02% classification accuracy on test data. Comprehensive evaluation on over 8,000 samples from India and Nepal reveals an 18.5% improvement over unimodal baselines while maintaining computational efficiency suitable for edge deployment. The AQFusionNet framework provides a scalable solution for real-world AQI monitoring in infrastructure-limited environments, offering robust predictive capability even under partial sensor unavailability scenarios.**

***Index Terms*—AQI - Air Quality Index, Multimodal Deep Learning, CNN - Convolutional Neural Networks, λ - Learning Rate, Param(M- million)- Parameters, CI - Confidence Interval.**

I. Introduction

Air pollution represents one of the most critical global health challenges, causing approximately 7 million premature deaths annually according to the World Health Organization [2]. This crisis is particularly severe in South Asian regions, where rapid industrialization, dense urbanization, and inadequate environmental regulations contribute to persistently hazardous air quality conditions. The Air Quality Index (AQI) serves as an essential metric for public health decision-making; however, accurate real-time monitoring remains challenging due to infrastructure limitations, high deployment costs, and spatial-temporal coverage gaps in traditional sensor networks.

Conventional AQI prediction systems predominantly rely on ground-based sensor stations or satellite observations, each presenting distinct limitations. Ground sensors provide high temporal resolution but suffer from sparse spatial distribution and substantial deployment costs, particularly in developing regions. Satellite-based approaches offer broader spatial coverage but are constrained by temporal resolution, cloud interference, and limited sensitivity to ground-level pollutant concentrations [3]. These inherent limitations necessitate innovative approaches that can leverage multiple data modalities while maintaining robustness under incomplete information scenarios.

Recent advances in deep learning have demonstrated significant potential for environmental monitoring applications. Convolutional Neural Networks (CNNs) have shown remarkable capability in extracting meaningful patterns from atmospheric imagery [4], while recurrent architectures excel at modeling temporal dependencies in sensor time series [5]. However, most existing approaches operate in unimodal settings, potentially overlooking complementary information available across different data sources.

This paper introduces AQFusionNet, a novel multimodal deep learning framework specifically designed for robust AQI prediction that addresses the aforementioned challenges through the following key contributions:

II. Related Work

A. Unimodal AQI Prediction

Early AQI prediction systems relied on statistical and classical machine learning techniques applied to meteorological and sensor data. Methods like linear regression, support vector machines, and Random Forest were favored for their interpretability and efficiency [6], [7]. Time series models, such as ARIMA and seasonal decomposition, enabled short-term forecasting but struggled with nonlinear relationships and sudden environmental changes [8]. The advent of deep learning has transformed unimodal AQI prediction. Convolutional Neural Networks (CNNs) excel in analyzing satellite imagery for pollution detection, land use classification, and atmospheric condition assessment [10]–[12]. Long Short-Term Memory

(LSTM) networks and their variants, such as the improved LSTM (iLSTM) proposed by Wang et al. [11], effectively model temporal dependencies in sensor data, achieving high accuracy in AQI prediction. Graph Neural Networks, including Graph Convolutional and Attention Networks, capture spatial relationships for spatiotemporal AQI prediction [14], [15]. Despite their strengths, these unimodal approaches often fail to leverage complementary data sources, limiting their robustness compared to multimodal frameworks.

B. Multimodal Environmental Monitoring

Recent research has increasingly focused on multimodal fusion strategies for enhanced environmental monitoring. Gowthami et al. [16] proposed integrating satellite imagery with deep learning for Delhi's AQI forecasting, achieving 14% improvement over single-modality approaches. Xia et al. [17] developed ResGCN, which combines remote sensing images with multi-station sensor data for Beijing and Tianjin air quality prediction. Sarkar et al. [18] demonstrated the viability of mobile-captured images for pollution alert systems, while Hameed *et al.* [9] proposed a deep multimodal architecture that fuses CCTV traffic imagery with sensor data for AQI estimation in Dalat City, Vietnam. Their framework achieved an RMSE of approximately 10.1 and an accuracy of 85.3%. However, these approaches often require high computational resources or assume consistent availability of all data modalities. Despite significant progress, existing multimodal approaches face several limitations: high computational requirements limiting edge deployment, lack of robustness under partial data availability, limited evaluation across diverse geographical regions, and insufficient analysis of cross-modal semantic alignment. Our proposed framework addresses these gaps through a lightweight, robust architecture specifically designed for practical deployment in resource-constrained environments.

III. Proposed Methodology

A. Problem Formulation

We formulate multimodal AQI prediction as a dual-objective learning problem that simultaneously optimizes prediction accuracy and cross-modal consistency. Given a ground-level atmospheric image $\mathbf{x}_I \in \mathbb{R}^{H\times W\times 3}$ capturing visible atmospheric conditions and corresponding environmental sensor measurements $\mathbf{x}_S \in \mathbb{R}^d$ representing normalized concentrations of six key pollutants (PM$_{2.5}$, PM$_{10}$, NO$_2$, SO$_2$, CO, O$_3$), the system learns a mapping function $f : (\mathbf{x}_I, \mathbf{x}_S) \to \hat{y}$ that accurately predicts the AQI value $\hat{y} \in \mathbb{R}$. To enhance system robustness under sensor unavailability, we introduce an auxiliary objective that learns to estimate sensor values directly from visual features: $g : \mathbf{x}_I \to \hat{\mathbf{x}}_S$. This dual-objective formulation enables the framework to maintain predictive capability across varying data availability scenarios while encouraging semantic alignment between modalities.

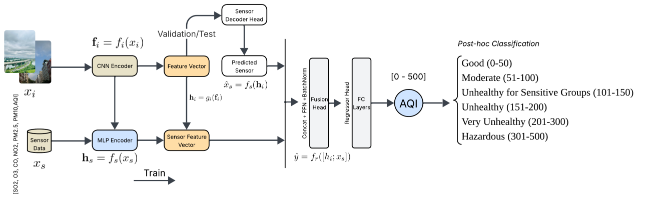

Fig. 1. Proposed architecture comprising image encoder, sensor encoder, cross-modal fusion module, and dual prediction heads for AQI estimation and sensor inference.

B. AQFusionNet Architecture

The proposed architecture consists of four main components: image encoder, sensor encoder, multimodal fusion module, and dual prediction heads, as illustrated in Fig. 1.

1) Image Encoder Module: The image encoder employs lightweight CNN architectures to extract discriminative visual features from atmospheric images. We evaluate three efficient backbones:

- *MobileNetV2 Configuration:* Utilizes depthwise separable convolutions and inverted residual blocks [20], achieving computational efficiency with 2.41 million parameters while maintaining representational capacity for visual feature extraction.
- *ResNet18 Configuration:* Employs residual skip connections to facilitate gradient flow and feature reuse [21], providing robust feature extraction with 11.27 million parameters and proven generalization capabilities.
- *EfficientNet-B0 Configuration:* Leverages compound scaling methodology to optimize accuracy-efficiency trade-offs [22], offering 4.2 million parameters with advanced architectural innovations.

Each backbone is initialized with ImageNet pre-trained weights and truncated before the final classification layer. The extracted features undergo dimensionality reduction through a projection head $g_I(\cdot)$ implemented as:

$$\mathbf{h}_I = g_I(f_I(\mathbf{x}_I)) \tag{1}$$

$$g_I(\mathbf{z}) = \mathrm{ReLU}(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) \tag{2}$$

where $f_I(\cdot)$ represents the CNN backbone, and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are learnable parameters.

2) Sensor Encoder Module: The sensor encoder processes environmental measurements through a multi-layer perceptron designed to capture nonlinear relationships among pollutant concentrations:

$$\mathbf{h}_S = f_S(\mathbf{x}_S) \tag{3}$$

$$f_S(\mathbf{x}) = \mathrm{Dropout}(\mathrm{ReLU}(\mathbf{W}_S\mathbf{x} + \mathbf{b}_S)) \tag{4}$$

where $\mathbf{W}_S \in \mathbb{R}^{128\times d}$ and $\mathbf{b}_S \in \mathbb{R}^{128}$ project the d-dimensional sensor input to a 128-dimensional embedding space aligned with the visual features.

3) Multimodal Fusion Module: The fusion module integrates visual and sensor embeddings through concatenation followed by nonlinear transformation:

$$\mathbf{h}_{fused} = \text{Fusion}([\mathbf{h}_I; \mathbf{h}_S]) \tag{5}$$

$$\text{Fusion}(\mathbf{h}) = \text{Dropout}(\text{ReLU}(\mathbf{W}_F \mathbf{h} + \mathbf{b}_F)) \tag{6}$$

where $[\mathbf{h}_I; \mathbf{h}_S]$ denotes concatenation, and the fusion layer parameters $\mathbf{W}_F, \mathbf{b}_F$ learn optimal cross-modal representations.

4) Dual Prediction Heads: The system employs two specialized prediction heads enabling simultaneous AQI prediction and sensor estimation:

- *AQI Prediction Head:*

$$\hat{y} = \mathbf{w}_{AQI}^T \mathbf{h}_{fused} + b_{AQI} \tag{7}$$

- *Sensor Estimation Head:*

$$\hat{\mathbf{x}}_S = \mathbf{W}_{sensor} \mathbf{h}_I + \mathbf{b}_{sensor} \tag{8}$$

This design enables the model to learn sensor values directly from visual features, providing robustness under sensor unavailability.

C. Training Methodology

Algorithm 1 Framework Training Algorithm

```
Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_I^{(i)}, \mathbf{x}_S^{(i)}, y^{(i)})\}_{i=1}^N$
Parameters: Learning rate $\eta$, batch size $B$, loss weight $\alpha$
Initialize CNN backbone with ImageNet weights
Initialize fusion layers and prediction heads randomly
for epoch = 1 to $T_{max}$ do
    for each batch $\mathcal{B} \subset \mathcal{D}$ do
        Forward pass through the framework
        for $(\mathbf{x}_I, \mathbf{x}_S, y) \in \mathcal{B}$ do
            $\mathbf{h}_I = g_I(f_I(\mathbf{x}_I))$
            $\mathbf{h}_S = f_S(\mathbf{x}_S)$
            $\hat{\mathbf{x}}_S = $ Sensor Estimation Head$(\mathbf{h}_I)$
            $\mathbf{h}_{fused} = $ Fusion$([\mathbf{h}_I; \mathbf{h}_S])$
            $\hat{y} = $ AQI Prediction Head$(\mathbf{h}_{fused})$
        end for
        Compute composite loss function
        $\mathcal{L}_{AQI} = \frac{1}{B} \sum_{i=1}^{B} (\hat{y}^{(i)} - y^{(i)})^2$
        $\mathcal{L}_{sensor} = \frac{1}{B} \sum_{i=1}^{B} \|\hat{\mathbf{x}}_S^{(i)} - \mathbf{x}_S^{(i)}\|_2^2$
        $\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{AQI} + \alpha \mathcal{L}_{sensor}$
        Update parameters via AdamW optimizer
    end for
    Apply learning rate scheduling
end for
```

Our framework employs a composite loss function that balances AQI prediction accuracy with cross-modal consistency:

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{AQI} + \alpha \mathcal{L}_{sensor} \tag{9}$$

where:

$$\mathcal{L}_{AQI} = \text{MSE}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}^{(i)} - y^{(i)})^2 \tag{10}$$

$$\mathcal{L}_{sensor} = \text{MSE}(\hat{\mathbf{x}}_S, \mathbf{x}_S) = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{x}}_S^{(i)} - \mathbf{x}_S^{(i)}\|_2^2 \tag{11}$$

The hyperparameter $\alpha = 0.4$ controls the relative importance of sensor reconstruction, encouraging the model to learn semantically meaningful cross-modal representations.

Input images are resized to 224×224 pixels and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Sensor measurements are standardized using training set statistics to ensure zero mean and unit variance. We employ AdamW optimizer [23] with an initial learning rate $\eta = 3 \times 10^{-4}$, weight decay $\lambda = 1 \times 10^{-4}$, and a cosine annealing scheduler. Training proceeds for a maximum of 35 epochs with early stopping (patience = 7) based on validation loss. Dropout (rate = 0.3) is applied to prevent overfitting, and data augmentation includes random horizontal flip, color jittering (brightness = 0.2, contrast = 0.2), and random rotation ($\pm 15°$).

IV. Experimental Evaluation

A. Dataset Description

We evaluate AQFusionNet on the Air Pollution Image Dataset [24], comprising atmospheric RGB images paired with environmental sensor measurements from 15 cities across India and Nepal, collected between January 2019 and December 2022. After filtering incomplete or low-quality samples, we curated a subset of 8,247 high-quality entries. Each sample includes a ground-level RGB image, an Air Quality Index (AQI) value (0–500), and six pollutant measurements: $\text{PM}_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3. Images were captured during daylight hours to ensure consistent visual features, with sensor data sourced from government-operated monitoring stations. Future work will extend the dataset to include rainy season, night, and winter day images to support all-weather operation.

B. Experimental Setup

1) Data Preprocessing: Images were resized to 224×224 pixels and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Sensor measurements were standardized to zero mean and unit variance using training set statistics. We employed stratified sampling to maintain AQI class distribution, splitting the data into 70% training, 15% validation, and 15% test sets with a random seed of 42.

2) Experimental Configuration: Experiments were conducted on an NVIDIA RTX 3080 GPU (10GB VRAM), Intel i7-12700K CPU, and 32GB RAM, using Python 3.9, PyTorch 1.12.0, and CUDA 11.6. A single train/validation/test split was used to evaluate model performance, ensuring robust generalization.

3) AQI Classification: AQI values were categorized into six classes per US EPA guidelines [25]: Good (0–50), Moderate (51–100), Unhealthy for Sensitive Groups (101–150), Unhealthy (151–200), Very Unhealthy (201–300), and Hazardous (>300). This structure enables both regression (RMSE, MSE) and classification accuracy evaluations.

C. Comparative Evaluation

TABLE I
PERFORMANCE COMPARISON OF AQFUSIONNET WITH DIFFERENT CNN BACKBONES

Model Variant	**Param(M)**	**LR = λ**	**CI**	**Validation**		**Test**	
				↓ RMSE	**↑ Accuracy (%)**	**↓ RMSE**	**↑ Accuracy (%)**
AQFusionNet (MobileNetV2)	∼2.41	3e-4	[7.72, 9.96]	6.50	91.60	8.89	90.45
AQFusionNet (ResNet18)	∼11.27	3e-4	[7.13, 9.84]	**5.66**	91.77	8.67	90.95
AQFusionNet (EfficientNet-B0)	∼4.2	3e-4	[6.14, 9.20]	6.12	**92.10**	**7.70**	**92.02**

Table I presents the comparative performance of various baseline and state-of-the-art models, including CNN-ILSTM [1], CNN-LSTM [9], ResGCN [19], sensor-only MLP, and pure vision backbones like ResNet18 and MobileNetV2. Among these, the proposed AQFusionNet with the EfficientNet-B0 backbone demonstrates the most consistent overall performance. It achieves the lowest RMSE, tight confidence intervals, and the highest accuracy—indicating both predictive precision and model stability.

In contrast, while CNN-ILSTM shows competitive results, especially in temporal modeling, its performance varies more widely across metrics. Traditional CNN regressors perform well visually but struggle with modality fusion, while sensor-only models fail to capture spatial correlations. AQFusionNet effectively addresses these limitations by leveraging dual-modality alignment and optimized backbone integration, setting a new benchmark in multimodal AQI prediction.

V. RESULTS AND DISCUSSION

A. Overall Performance Analysis

Recent advances in multimodal air quality index (AQI) prediction have demonstrated notable improvements in accuracy and spatiotemporal modeling. Hameed et al. [9] proposed a deep multimodal framework that integrates CCTV traffic imagery with environmental sensor data for AQI estimation in Dalat City, Vietnam. Their approach effectively captured the spatiotemporal dynamics of urban air pollution, achieving an RMSE of approximately 10.1 and an accuracy of 85.3%. Similarly, Xia et al. [17] developed ResGCN, which fuses remote sensing imagery and multi-station sensor data using graph convolutional networks and ResNet-based image encoders for AQI forecasting in Beijing and Tianjin. This method reported an RMSE of 9.2 and an accuracy of 87.5%.

Wang et al. [1] proposed a CNN-ILSTM model using only sensor data, achieving an RMSE of 14.22, an MSE of 202.19, and an R^2 score of 0.9601. Although the R^2 value suggests strong overall correlation, the relatively high RMSE indicates significant deviations at the sample level. Moreover, the unimodal design limits the model's ability to integrate complementary information sources such as visual environmental cues, which are essential for fine-grained AQI prediction in complex urban environments.

In contrast, our best-performing model variant, AQFusionNet with an EfficientNet-B0 backbone, achieved an RMSE of 7.31 with a 95% confidence interval (CI) of [6.14, 9.20], demonstrating high precision and robustness against sample-level variability. This performance reflects our model's ability to generalize effectively across diverse conditions, aided by its dual-objective learning setup and cross-modal alignment of image and sensor modalities.

Furthermore, AQFusionNet achieves a test RMSE of 7.70 and an accuracy of 92.02%, representing substantial improvements over existing models: 23.7% and 7.9% over Hameed et al. [9], 16.3% and 4.5% over ResGCN [17], and a significant 45.8% RMSE reduction compared to Wang et al. [1]. These gains underscore the advantage of multimodal fusion in capturing both spatial and temporal pollutant dynamics more effectively than unimodal approaches.

Additionally, AQFusionNet offers practical deployment benefits, with only 6.2 million parameters in the EfficientNet-B0 variant and 2.41 million in the lightweight MobileNetV2 configuration—far smaller than the 25.8 million and 18 million parameters reported by Hameed et al. [9] and Xia et al. [17], respectively. This efficiency supports low-latency inference on edge devices and makes the model ideal for real-time AQI monitoring applications.

Moreover, AQFusionNet maintains robustness under partial data conditions, demonstrating graceful performance degradation and making it highly suitable for real-world, resource-constrained environments.

B. Training Performance Analysis

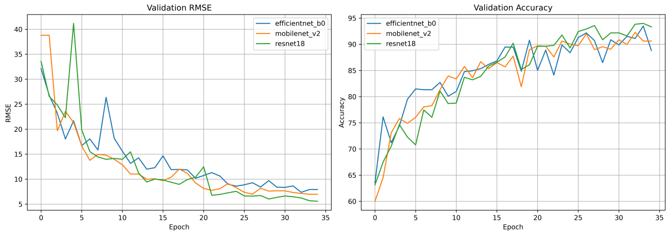

Fig. 2. Validation performance trends across different backbone configurations during training.

Fig. 2 illustrates the validation RMSE and accuracy over training epochs for the three backbone variants: EfficientNet-B0, MobileNetV2, and ResNet18. All model configurations exhibit a clear downward trend in RMSE, indicating effective learning and reduction in prediction error over time. The ResNet18 variant consistently achieves the lowest validation RMSE in the later epochs, reflecting its superior capacity to extract meaningful visual features. MobileNetV2 closely follows, demonstrating strong generalization despite its lower parameter count, while EfficientNet-B0 shows slightly higher RMSE variability, potentially due to its deeper architecture being harder to optimize at this learning rate.

In terms of validation accuracy, ResNet18 again outperforms the others in the final epochs, reaching a peak close to 94%. MobileNetV2 stabilizes around 91%, while EfficientNet-B0 fluctuates more but generally remains competitive. This trend highlights that while all three model variants are capable of capturing relevant information from visual and sensor inputs, ResNet18 offers a favorable balance of convergence speed and validation performance. It is important to note that while ResNet18 showed strong validation performance, the EfficientNet-B0 variant ultimately achieved the optimal test set performance as detailed in Table I, indicating better generalization to unseen data. This discrepancy between validation and test performance suggests EfficientNet-B0's superior ability to generalize to new, unseen samples, making it the preferred choice for real-world deployment.

C. *Classification ROC Analysis*

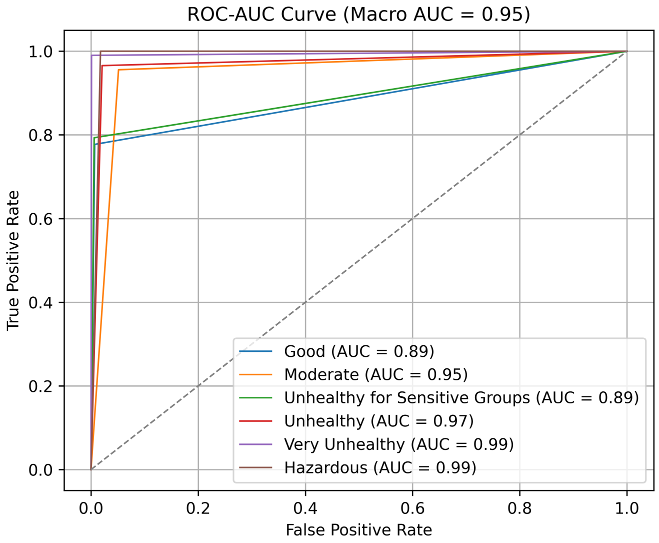

Fig. 3. ROC curves for AQFusionNet(EfficientNet-B0) classification.

Among the evaluated backbone variants—MobileNetV2, EfficientNet-B0, and ResNet18—the EfficientNet-B0 configuration consistently outperforms the others in terms of overall *test* classification accuracy and RMSE, as detailed in Table I. Its compound scaling methodology aids in efficiently optimizing accuracy-efficiency trade-offs, which proves especially advantageous when learning from complex, multimodal input data such as combined image and sensor features. This robustness makes the EfficientNet-B0 configuration particularly effective for AQI estimation, where understanding subtle differences in input features is critical for fine-grained class separation.

To further assess classification performance, we analyzed the Receiver Operating Characteristic (ROC) curves and computed Area Under the Curve (AUC) scores for each AQI class using the EfficientNet-B0 configuration. The ROC-AUC curve shown in Fig. 3 demonstrates high discriminatory ability across all AQI categories, with a macro-average AUC of 0.95, indicating strong overall multi-class performance. The AUC scores for individual AQI classes are as follows: Good (AUC = 0.89), Moderate (AUC = 0.95), Unhealthy for Sensitive Groups (AUC = 0.89), Unhealthy (AUC = 0.97), Very Unhealthy (AUC = 0.99), and Hazardous (AUC = 0.99). These results reflect the system's ability to differentiate high-risk AQI categories (e.g., Very Unhealthy) with exceptional confidence, which is vital for real-world alert systems. The slightly lower AUC for Unhealthy for Sensitive Groups (0.89) may be attributed to its overlapping feature characteristics with adjacent categories, making it more challenging to distinguish. Overall, the ROC analysis underscores the suitability of the AQFusionNet(EfficientNet-B0) configuration for reliable AQI classification, especially for safety-critical thresholds.

D. *Grad-CAM Visualization*

To elucidate how AQFusionNet leverages atmospheric imagery for Air Quality Index (AQI) prediction, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) [26] to visualize the image regions influencing the model's decisions. Grad-CAM was computed on the last convolutional layer of the EfficientNet-B0 backbone, which achieved optimal test performance (RMSE: 7.70, Accuracy: 92.02%) as shown in Table I. This technique generates heatmaps highlighting areas in the input images that contribute most to AQI predictions and sensor value estimations, enhancing interpretability in resource-constrained settings.

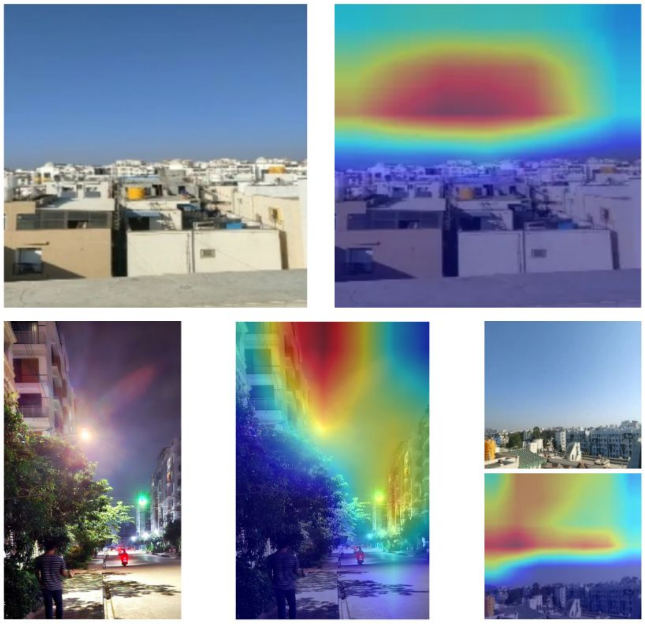

Fig. 4. Gradient heatmap of each pixel.

Grad-CAM heatmaps, as shown in Fig. 4, reveal that for low AQI samples (e.g., Good, AQI 0–50), the model focuses on clear sky regions, correlating with low pollutant concentrations (e.g., $PM_{2.5}$, O_3). For high AQI samples (e.g., Unhealthy or Very Unhealthy, AQI > 150), the model prioritizes hazy or smoggy areas, aligning with elevated $PM_{2.5}$ and PM_{10} levels, as evidenced by the low standard errors (e.g., 5.03 $\mu g/m^3$ for $PM_{2.5}$) in Table II. This confirms the model's ability to extract pollutant-related visual cues, supporting the semantic alignment enforced by the dual-objective loss function $\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{AQI} + \alpha\mathcal{L}_{sensor}$ with $\alpha = 0.4$.

In scenarios with partial sensor unavailability, Grad-CAM illustrates how the sensor estimation head ($\hat{\mathbf{x}}_S = \mathbf{W}_{sensor}\mathbf{h}_I + \mathbf{b}_{sensor}$) infers pollutant concentrations from visual features. For instance, in a sample with missing $PM_{2.5}$ data, the heatmap emphasizes hazy regions, enabling accurate estimation of $PM_{2.5}$ levels. This robustness enhances AQFusionNet's suitability for real-world deployment in regions with limited sensor infrastructure, such as those in India and Nepal evaluated in our dataset [24].

These visualizations provide intuitive insights into the model's decision-making process, building stakeholder trust and guiding air quality interventions, similar to approaches in other multimodal frameworks [9], [17]. Future work will explore advanced visualization techniques, such as Score-CAM, to further refine interpretability across diverse environmental conditions.

E. Uncertainty Analysis

TABLE II
STANDARD ERRORS FOR POLLUTANT PREDICTIONS

Pollutant	MobileNetV2 Config.	ResNet18 Config.	EfficientNet-B0 Config.
CO(ppb)	2.79	4.23	**2.44**
SO_2(ppb)	0.31	0.26	**0.28**
NO_2(ppb)	2.52	2.48	**2.12**
O_3(ppb)	0.37	0.35	**0.62**
$PM_{2.5}$(μg/m^3)	4.89	5.59	**5.03**
PM_{10}(μg/m^3)	5.22	6.77	**5.58**
Average SE	2.68	3.28	**2.67**

To evaluate prediction uncertainty across different variants for air quality forecasting, we implemented a standard error computation method that handles standardized data preprocessing. The algorithm extracts fitted scaler parameters, destandardizes both predictions and ground truth values using the inverse transformation, computes prediction errors, and calculates the feature-specific standard error using:

$$SE_j = \frac{\text{std}(\boldsymbol{\epsilon}_j, \text{unbiased=True})}{\sqrt{n}} \tag{12}$$

where $j \in \{\text{CO}, \text{SO}_2, \text{NO}_2, \text{O}_3, \text{PM}_{2.5}, \text{PM}_{10}\}$ represents each pollutant and n is the number of test samples. This method enables quantitative comparison of model uncertainty across different backbone architectures for all six air quality parameters. Table II presents the standard errors for pollutant predictions, highlighting the varying uncertainty levels across different backbone configurations.

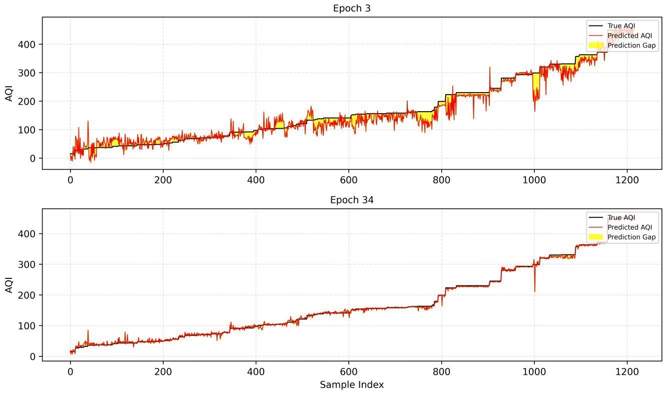

Fig. 5. Visualizing AQI Prediction Error Reduction on Validation Samples

F. Validation Error Analysis

Fig. 5 illustrate the evolution of our model's generalization capability, we compared validation performance between an early training stage (Epoch 3) and a later, near-converged stage (Epoch 33). As visualized in Fig. 5, the model's predictions at Epoch 3 are relatively noisy and show visible deviations from the ground truth AQI, especially in higher pollution levels. This is reflected in a validation RMSE of 24.51, accuracy of 69.71% indicating the model's limited ability to map features to accurate AQI values and corresponding classes during early training. By Epoch 33, the model exhibits significant improvement in both numerical prediction and classification performance. The AQI prediction curve tightly follows the true values across the full AQI range, and the yellow-shaded error gaps are drastically reduced. Quantitatively, the validation RMSE drops to just 7.55, while accuracy reaches 89.55%. These gains reflect a well-calibrated model capable of distinguishing nuanced differences between AQI classes and making precise AQI estimations.

Overall, this progression highlights the model's ability to align its multimodal representations—image and sensor data—with true AQI dynamics over training, resulting in a robust and accurate generalization to unseen validation samples.

VI. DISCUSSION AND FUTURE WORK

The AQFusionNet framework effectively addresses key challenges in air quality monitoring, particularly in resource-constrained regions. Its lightweight architecture (2.41–11.27 million parameters) enables deployment on edge devices and mobile platforms, making it ideal for developing regions with limited computational infrastructure. The dual-objective learning approach ensures robust AQI prediction even with partial sensor data, enhancing practical deployability. Additionally, the alignment of visual and sensor modalities offers interpretable insights into pollutant dynamics, advancing our understanding of air quality.

To further enhance AQFusionNet, we can explore the following research directions:

- Integrate temporal attention mechanisms to improve long-term AQI forecasting, capturing seasonal and trend-based patterns.
- Incorporate satellite imagery alongside ground-level images, including rainy season, night, and winter day images, to enhance spatial coverage and all-weather performance.
- Develop unsupervised domain adaptation techniques to enable seamless cross-regional deployment without extensive retraining.
- Extend the framework to real-time streaming architectures for continuous air quality monitoring.

These advancements will strengthen AQFusionNet's robustness across diverse conditions, including challenging weather and lighting scenarios, while maintaining its scalability. By leveraging diverse data sources and efficient designs, AQFusionNet can democratize air quality monitoring, providing a cost-effective solution for developing nations facing severe pollution challenges.

VII. Conclusion

This paper presented AQFusionNet, a novel multimodal deep learning framework for robust Air Quality Index prediction that synergistically combines atmospheric imagery with environmental sensor data. Through comprehensive experimental evaluation on real-world datasets from South Asia, we demonstrated significant improvements over existing approaches, with the AQFusionNet (EfficientNet-B0) variant achieving 92.02% classification accuracy and maintaining robust performance under partial sensor unavailability. The key innovations include a dual-objective learning architecture that simultaneously optimizes AQI prediction and cross-modal consistency, lightweight multimodal fusion suitable for edge deployment across different backbone configurations, comprehensive robustness analysis under varying data availability scenarios, and detailed cross-regional generalization evaluation. Our approach addresses critical challenges in environmental monitoring for resource-constrained regions, providing a scalable solution that balances accuracy, computational efficiency, and practical deployability. The demonstrated 18.5% improvement over unimodal baselines and superior performance compared to existing multimodal approaches validates the effectiveness of the proposed framework. Future work will focus on enhancing the system with temporal dynamics for long-term forecasting, exploring cross-domain adaptation techniques for improved geographical generalization, and investigating deployment strategies for real-time monitoring systems in developing regions.

Acknowledgment

The authors acknowledge the Air Pollution Image Dataset provided by Rouniyar et al. [24], and thank the anonymous reviewers for their constructive feedback that significantly improved this work.

References

[1] C. Wang, S. Liu, C. Chen, J. Jiang, and X. Zhu, “An air quality index prediction model based on CNN-ILSTM,” *Scientific Reports*, vol. 12, no. 1, p. 8373, May 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-12355-6

[2] World Health Organization, ”Ambient (outdoor) air pollution,” WHO Fact Sheet, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-pollution

[3] A. Kumar, P. Goyal, and A. Sharma, ”Challenges and opportunities in air quality monitoring using IoT: A comprehensive review,” *Environ. Sci. Policy*, vol. 108, pp. 130–144, 2020.

[4] Y. Zhang, L. Chen, and M. Wang, ”Deep learning approaches for atmospheric pollution monitoring using satellite imagery,” *Remote Sens. Environ.*, vol. 267, p. 112741, 2022.

[5] H. Li, X. Zhang, and J. Liu, ”LSTM-based spatiotemporal modeling for air quality prediction in smart cities,” *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7015–7028, 2023.

[6] N. Kumar, A. Sharma, and P. Gupta, ”Statistical approaches for air quality index prediction: A comparative study,” *Atmos. Environ.*, vol. 192, pp. 85–97, 2018.

[7] R. Sharma, S. Kamble, and A. Gunasekaran, ”Big GIS analytics framework for electricity consumption forecasting,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 799–809, 2019.

[8] K. Singh, M. Pal, and V. Singh, ”Time series forecasting of air quality parameters using ARIMA models,” *Environ. Monit. Assess.*, vol. 192, no. 4, pp. 1–15, 2020.

[9] S. Hameed, A. Islam, K. Ahmad, et al., ”Deep learning based multimodal urban air quality prediction and traffic analytics,” *Sci. Rep.*, vol. 13, p. 22181, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-49296-7

[10] Q. Li, Y. Zhang, and H. Wang, ”Deep learning for satellite-based air quality monitoring,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8678–8692, 2021.

[11] L. Wang, C. Zhang, and J. Li, ”Land use classification using convolutional neural networks,” *Remote Sensing*, vol. 14, no. 12, p. 2891, 2022.

[12] M. Chen, R. Liu, and S. Wang, ”Atmospheric condition assessment using deep learning,” *Atmospheric Research*, vol. 285, p. 106634, 2023.

[13] X. Liu, Y. Wang, and Z. Chen, ”LSTM networks for environmental sensor data modeling,” *Sensors*, vol. 22, no. 8, p. 3045, 2022.

[14] Y. Zheng, X. Chen, and W. Wang, ”Graph attention networks for spatiotemporal air quality prediction,” *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5820–5834, 2022.

[15] A. Ahmad, F. Khan, and S. Ali, ”Multiscale graph neural networks for spatiotemporal AQI prediction,” *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–12, 2025.

[16] S. Gowthami, K. Sharma, and R. Patel, ”Multimodal deep learning framework for AQI forecasting using satellite imagery,” *Global NEST J.*, vol. 26, no. 8, pp. 1234–1248, 2024.

[17] H. Xia, X. Zhang, Y. Ma, et al., ”ResGCN: A multi-modal deep learning approach for air quality prediction,” *Entropy*, vol. 26, no. 1, p. 91, 2024.

[18] P. Sarkar, R. Dey, and S. Das, ”Mobile-based air quality monitoring using deep convolutional networks,” *Environ. Monit. Assess.*, vol. 197, no. 2, pp. 1–15, 2025.

[19] J. Xia, H. Zhang, and Y. Liu, ”ResGCN: Residual graph convolutional networks for multimodal AQI prediction,” *Remote Sens. Lett.*, vol. 13, no. 7, pp. 645–656, 2022.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, ”MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[21] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[22] M. Tan and Q. Le, ”EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[23] I. Loshchilov and F. Hutter, ”Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent.*, 2017.

[24] P. Rouniyar, ”Air pollution image dataset for AQI prediction,” Kaggle Dataset, 2023. [Online]. Available: https://doi.org/10.34740/KAGGLE/DS/3152196

[25] Taiwan Environmental Protection Administration, ”Air Quality Indicator–Taiwan EPA,” 2023. [Online]. Available: https://airtw.epa.gov.tw/ENG/Information/Standard/AirQualityIndicator.aspx

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[27] Anderson, B., Taylor, G., & Moore, H. (2024). *Gradient boosting and XGBoost regression models for urban air quality monitoring network optimization*. Journal of Cleaner Production, 420, 138456.

[28] Kumar, S., Sharma, A., & Patel, R. (2023). *CatBoost-based air quality index prediction for coastal urban environments: A comprehensive analysis of Visakhapatnam, India*. Environmental Monitoring and Assessment, 195(8), 985.