Review

Image-Based Air Quality Estimation with Deep Learning: A Systematic Review

Mokhammad Parvani Vafa *

Posted Date: 5 May 2026

doi: 10.20944/preprints202604.2212.v1

Keywords: air quality index; AQI; convolutional neural network; deep learning; EfficientNet; haze detection; PM2.5; transfer learning; urban images; Central Asia; Bishkek

 doi:10.20944/preprints202604.2212.v1

Review

Image-Based Air Quality Estimation with Deep Learning: A Systematic Review

Mokhammad Parvani Vafa

Department of Computer Science, Ala-Too International University (AIU), Bishkek, Kyrgyzstan; porvanivafoo@gmail.com

Abstract

The proliferation of urban air pollution—particularly fine particulate matter ($PM_{2.5}$)—demands scalable monitoring approaches that go beyond the sparse networks of reference-grade stations installed in most cities worldwide. Over the past decade, deep learning methods applied to ground-level photographs have emerged as a low-cost complement to conventional sensing: convolutional neural networks (CNNs) and hybrid architectures can extract visually detectable pollution cues—haze opacity, color temperature shifts, and reduced horizon contrast—and map them onto continuous air-quality index (AQI) or $PM_{2.5}$ estimates. This review synthesizes 40 primary studies and several additional supporting sources published between 2020 and 2025 to characterize the state of the art in image-based AQI estimation, identify the key technical and infrastructural limitations, and outline research directions relevant to data-scarce, under-monitored cities such as Bishkek, Kyrgyzstan. Three interlocking themes structure the review: (1) deep learning architectures and training strategies, from single-modality CNNs to multimodal and spatiotemporal hybrid models; (2) dataset characteristics and their decisive influence on regression accuracy; and (3) the monitoring infrastructure gap in low-income and middle-income cities of Central Asia and comparable regions. The evidence consistently shows that positive R^2 values require at least 3,000–5,000 labeled image–pollutant pairs, controlled temporal stratification, and, ideally, auxiliary meteorological inputs. Promising directions include vision transformers, structured state-space models, Grad-CAM interpretability, and cross-city transfer learning. The review concludes with a structured research agenda for image-based air-quality monitoring in Central Asia.

Keywords: air quality index; AQI; convolutional neural network; deep learning; EfficientNet; haze detection; $PM_{2.5}$; transfer learning; urban images; Central Asia; Bishkek

1. Introduction

Air pollution is the single largest environmental risk to human health globally. The GBD 2019 Risk Factors Collaborators (2020) attributed more than 6.7 million premature deaths per year to ambient particulate matter exposure, making $PM_{2.5}$ one of the leading modifiable causes of the global burden of disease. Landrigan et al. (2018) estimated that pollution as a whole—including air, water and chemical pollution—accounts for approximately nine million deaths annually, a figure that rivals malaria and HIV/AIDS combined. The World Health Organization (WHO) (World Health Organization, 2021) revised its annual $PM_{2.5}$ guideline down to 5 $\mu g/m^3$ in 2021, a standard that the vast majority of urban centres worldwide fail to meet.

In high-income countries, dense networks of reference-grade monitors, satellite products, and chemical-transport models provide reasonably complete spatial coverage. In low-income and middle-income countries (LMICs) the picture is starkly different. The World Bank (2023a) noted that $PM_{2.5}$ concentrations in developing-country cities routinely exceed WHO guidelines by factors of five to fifteen, yet monitoring networks capable of capturing the spatial heterogeneity of urban pollution are largely absent. Cities across Central Asia are paradigmatic of this challenge. Rau et al. (2022) documented that Bishkek, the capital of Kyrgyzstan, recorded average winter $PM_{2.5}$ concentrations above 200 $\mu g/m^3$—more than thirteen times the WHO 24-hour guideline—while operating fewer

 doi:10.20944/preprints202604.2212.v1

than eight government-run monitoring stations. United Nations Economic Commission for Europe (2022) and UNICEF (2023) further noted that children in Bishkek suffer elevated rates of respiratory morbidity directly linked to poor air quality.

Against this backdrop, computer vision and deep learning have attracted considerable attention as potentially low-cost complements to fixed sensing. The intuition is straightforward: atmospheric pollution imparts measurable visual signatures on outdoor photographs—hazing of distant features, reduced contrast, bluish-grey sky colouration, and attenuated colour saturation. Convolutional neural networks pre-trained on millions of natural images (ImageNet) can, in principle, learn to decode these signatures and produce quantitative AQI or $PM_{2.5}$ estimates from a single photograph, without any chemical sensor at the prediction location. This is the central hypothesis driving the body of work reviewed here.

Scope.

A central unresolved question in this literature is: under what precise conditions does image-based AQI regression succeed, and what roadmap exists for cities operating under severe data and infrastructure constraints? This review addresses that question directly by synthesising the growing body of work from 2020 to 2025. It is particularly timely given that cities across Central Asia—where seasonal $PM_{2.5}$ peaks rank among the highest in the world—remain almost entirely unstudied in this literature.

The review covers studies published from 2020 to 2025 that (a) use ground-level outdoor photographs or video as a primary input, (b) target $PM_{2.5}$, PM_{10}, or AQI as the output, and (c) employ machine learning or deep learning as the modelling framework. Satellite-only studies and numerical weather prediction studies are included only when they directly inform transfer-learning strategies for the image-based setting.

Structure.

Section 2 describes the review methodology. Section 3 situates the health burden and monitoring gap. Section 4 reviews the major classes of deep learning architectures used for image-based AQI estimation. Section 5 analyses dataset characteristics as a determinant of prediction accuracy. Section 6 examines results from data-constrained and infrastructure-limited contexts analogous to Bishkek. Section 7 surveys emerging methods. Section 8 synthesises evidence specific to Central Asia. Section 9 provides a comparative discussion. Section 10 concludes with a research agenda.

2. Review Methodology

2.1. Search Strategy

A structured search was conducted in Google Scholar, PubMed, IEEE Xplore, Scopus, ScienceDirect, and arXiv. The primary query strings combined terms from three facets: *image/photo/visual/camera*, *PM2.5/PM10/AQI/ air quality/haze/pollution*, and *deep learning/CNN/convolutional/ transfer learning/neural network*. Secondary searches targeted region-specific strings (*Central Asia*, *Kyrgyzstan*, *Bishkek*, *LMIC*) combined with air-quality terms. All searches were restricted to 2020–2025 except for a small set of seminal foundational papers cited for methodological context.

2.2. Inclusion and Exclusion Criteria

Included: Peer-reviewed journal articles, conference papers, and preprints (arXiv, ESSOAr) that (i) use outdoor, ground-level images or video as at least one model input, (ii) target a continuous or categorical measure of ambient outdoor air quality ($PM_{2.5}$, PM_{10}, AQI, or related indices), and (iii) employ a machine learning or deep learning model.

Excluded: Studies relying exclusively on indoor air quality, remote sensing satellites without any ground-level image component, purely meteorological models without visual inputs, and studies reporting only hardware design without quantitative model evaluation.

 doi:10.20944/preprints202604.2212.v1

2.3. Data Extraction and Quality Assessment

For each included study the following data were extracted: publication year, country/city of data collection, dataset size (number of images), image source (fixed camera, mobile, crowdsourced), target pollutant, architecture type, best reported metric (R^2, RMSE, MAE, accuracy), and any auxiliary inputs used. Study quality was assessed on three dimensions: dataset transparency, reproducibility (random-seed fixation, code availability), and comparability of evaluation (held-out test set vs. cross-validation). Studies with severe methodological concerns were included but flagged.

A total of 40 primary studies were selected after screening, complemented by approximately 15 supporting references on health burden, monitoring infrastructure, and transfer learning methodology.

2.4. Use of Artificial Intelligence Tools

No AI-assisted writing or large language model tools were used in the preparation of this manuscript beyond standard spell-checking software. The authors take full responsibility for all content presented.

3. Health Burden and the Monitoring Infrastructure Gap

3.1. Global Evidence on $PM_{2.5}$ Health Effects

Particulate matter with an aerodynamic diameter of 2.5 μm or less ($PM_{2.5}$) penetrates the alveolar region of the lung and enters the bloodstream, causing systemic inflammation that drives cardiovascular and pulmonary disease (World Health Organization, 2021). Long-term exposure is causally linked to ischaemic heart disease, stroke, chronic obstructive pulmonary disease, lung cancer, and type 2 diabetes (GBD 2019 Risk Factors Collaborators, 2020). Maji et al. (2021) quantified the burden in 130 Chinese cities, finding that PM_{10} and $PM_{2.5}$ together account for tens of thousands to hundreds of thousands of attributable deaths annually per city, depending on population density and ambient concentrations.

The 2021 WHO Air Quality Guidelines reduced the recommended annual $PM_{2.5}$ level from 10 $\mu g/m^3$ to 5 $\mu g/m^3$, a threshold that virtually no urban centre in South Asia, Central Asia, or sub-Saharan Africa achieves. In Bishkek, average annual $PM_{2.5}$ concentrations have been estimated at 44–47 $\mu g/m^3$ (MoveGreen, 2025), while winter daily averages regularly exceed 200 $\mu g/m^3$ (Rau et al., 2022).

3.2. The Spatial Resolution Problem

Even in cities with some monitoring infrastructure, reference stations are too sparse to capture sub-kilometre spatial heterogeneity. Gulia et al. (2022) argued that effective urban air-quality management requires monitoring at spatial scales of 500–1,000 m, whereas typical reference-station networks in developing countries operate at inter-station distances of tens of kilometres. Kumar et al. (2015) reviewed low-cost electrochemical and optical sensors as scalable complements to reference stations, noting persistent challenges with calibration drift, humidity cross-sensitivity, and sensor-to-sensor variability.

The World Bank (2023b) World Bank's *Air Quality Analysis for Bishkek* estimated that the economic cost of air pollution in Bishkek amounts to approximately 1.2% of Kyrgyzstan's GDP, with national-level costs reaching 5.1% of GDP. This macroeconomic framing underscores the argument that scalable, low-cost monitoring is not merely a technical convenience but a prerequisite for evidence-based pollution governance.

3.3. Role of Image-Based Methods in Filling the Gap

Image-based approaches offer a fundamentally different scalability profile compared with sensors. Smartphones and CCTV cameras are ubiquitous even in low-income urban settings, and their data requires no physical deployment or maintenance beyond the model development phase. Singh et al. (2020) demonstrated that photographic visibility can serve as a credible proxy for AQI in East African

 doi:10.20944/preprints202604.2212.v1

cities with minimal monitoring infrastructure, providing an early proof-of-concept for the approach reviewed here.

4. Deep Learning Architectures for Image-Based AQI Estimation

4.1. Single-Modality CNN Regression

The simplest image-based pipeline passes a single outdoor photograph through a CNN backbone and regresses the scalar AQI or $PM_{2.5}$ concentration from the resulting feature vector. Chakma et al. (2020) used a VGG-family architecture on 2,364 real outdoor images, demonstrating that deep features outperform hand-crafted haze descriptors such as transmission maps and depth-based features. Zhang et al. (2020) achieved $R^2 = 0.71$ on over 8,000 outdoor images using a ResNet-based CNN, applying explicit temporal stratification (morning, afternoon, evening) to reduce confounding from diurnal illumination variation.

Gu et al. (2019) trained a lightweight CNN on surveillance images, achieving competitive AQI estimates at low computational cost—an important property for edge-device deployment. Song et al. (2020) proposed ResNet-LSTM for simultaneous $PM_{2.5}$ and PM_{10} estimation from sequential smartphone images, reporting $R^2 > 0.84$ on their test set.

4.2. Multimodal Fusion Architectures

A consistent finding in the literature is that auxiliary numerical inputs—sensor readings, meteorological variables, or image statistics—substantially improve regression accuracy when fused with visual features. Kow et al. (2022) proposed a hybrid CNN–LSTM model that combined VGG/ResNet features with HSV colour statistics from 3,549 hourly labelled samples at a fixed station in Taiwan, achieving $R^2 = 0.94$ and RMSE = 5.38. The use of HSV features captures low-level atmospheric colour cues that are correlated with AQI but not well represented in the final layers of ImageNet-pre-trained networks.

Dong et al. (2021) fused EfficientNet visual features with meteorological metadata (humidity, temperature, wind speed) on a corpus of over 12,000 samples, achieving MAE = 18 $\mu g/m^3$. The authors showed that removing meteorological inputs raised MAE by approximately 25%, quantifying the independent information content of each modality.

J. Zhao et al. (2022) used a multimodal CNN fusion architecture on over 6,000 labelled pairs and reported $R^2 = 0.65$, confirming the general pattern that data fusion consistently outperforms single-modality approaches. Xue et al. (2023a) employed a deep CNN with spatial attention mechanisms on 15,000+ images, achieving $R^2 = 0.78$.

4.3. Spatiotemporal Models: CNN–LSTM Hybrids

Static single-image models discard temporal correlations that carry significant predictive information. Wang et al. (2024) applied a CNN–LSTM architecture to surveillance-camera image sequences, achieving $R^2 > 0.92$ and RMSE < 8.5 for AQI estimation. The LSTM component exploited the temporal continuity of air-quality dynamics: knowing that the previous hour was heavily polluted strongly constrains the current estimate.

Xue et al. (2023b) proposed a 3D-CNN fused with a gated recurrent unit (GRU) and an attention mechanism (3D-CNN-GRU) for multiple-horizon forecasting of air quality from images, reporting that the proposed model outperformed all single-modality baselines. Aslam et al. (2025) demonstrated that video-based AQI estimation using Mamba (a structured state-space model) achieved $R^2 = 0.92$ by exploiting frame-to-frame temporal dependencies in fixed-camera footage—a method well-suited to future deployments in cities with CCTV infrastructure.

4.4. Attention Mechanisms and Interpretability

Attention mechanisms allow the network to weight image regions selectively, prioritising pixels that contain the most pollution-relevant information (sky opacity, horizon contrast) while suppressing irrelevant regions (occluding objects, foreground clutter). Utomo et al. (2024) proposed AQI-Net, a

 doi:10.20944/preprints202604.2212.v1

CNN trained on over 11,000 images from three Indonesian cities augmented with Gradient-weighted Class Activation Mapping (Grad-CAM), achieving 99.81% classification accuracy. Grad-CAM visualisations confirmed that the network attended to sky and horizon regions, providing interpretable scientific validation of the learned features.

Hardini et al. (2024) evaluated ensemble architectures combining Vision Transformers (ViT) with convolutional networks, showing that the complementary global and local feature extraction of the two paradigms yields superior performance under large-data conditions.

4.5. End-to-End Pollutant Prediction from Street-View Images

Hankey et al. (2025) presented an end-to-end pollutant prediction model (E2EPPM) that directly predicts concentrations of ten pollutants from street-level imagery collected via mobile monitoring in Augsburg, Beijing, and Hotan. The model achieved $R^2 > 0.90$ for all pollutants. SHAP analysis identified vegetation, buildings, sky, and vehicles as the dominant visual contributors—confirming that deep networks learn physically meaningful atmospheric cues rather than artefacts.

5. Dataset Characteristics and Their Impact on Accuracy

5.1. Sample Size as the Primary Determinant

The most striking empirical regularity in the literature is the near-monotonic relationship between dataset size and regression accuracy. Table 1 summarises this relationship across 18 representative studies. All studies achieving $R^2 \geq 0.65$ used at least 3,500 labelled image–pollutant pairs; studies working below 1,500 samples consistently report lower or negative R^2 values under uncontrolled real-world conditions.

Table 1. Summary of representative image-based air-quality estimation studies (2020–2025). Studies are sorted by dataset size. AQ = air quality index or pollutant concentrations.

Study	Location	Architecture	Images	Target	Best metric
Mondal et al. (2024)	Dhaka, BD	DCNN	∼1,000	$PM_{2.5}$	Limited
Parvani Vafa and Khan (2025)	Bishkek, KG	VGG16/EffNetB0	1,014	AQI	$R^2 = -0.28$
Kow et al. (2022)	Taiwan	CNN–LSTM (VGG/ResNet)	3,549	AQI	$R^2 = 0.94$
Wang et al. (2024)	China	CNN–LSTM (VGG16)	7,213	AQI	$R^2 = 0.92$
Zhang et al. (2020)	Multi-city	ResNet CNN	8,000+	AQI	$R^2 = 0.71$
J. Zhao et al. (2022)	China	Multimodal CNN	6,000+	AQI	$R^2 = 0.65$
Xue et al. (2023a)	China	Deep CNN + Attention	15,000+	AQI	$R^2 = 0.78$
Dong et al. (2021)	China	EfficientNet + meta	12,000+	$PM_{2.5}$	MAE=18
Utomo et al. (2024)	Indonesia	AQI-Net + Grad-CAM	11,000+	AQI	Acc=99.8%
Xue et al. (2023b)	China	3D-CNN-GRU-Attention	>10,000	AQI	Improved
Hankey et al. (2025)	Multi-city	E2EPPM (CNN+KAN)	>10,000	Multi-poll	$R^2 > 0.90$
H. Zhao et al. (2025)	Lanzhou, CN	ResNet50+RF+Lasso	2,400	$PM_{2.5}$	Improved
Apte et al. (2024)	Bengaluru, IN	CNN (dashboard cam)	Video	Multi-poll	NRMSE ≤13.7%
Song et al. (2020)	Multi-city	ResNet-LSTM	>8,000	$PM_{2.5}/PM_{10}$	$R^2 > 0.84$
Li et al. (2025)	Global	Multi-backbone	11,114	$PM_{2.5}$	Benchmark
Hardini et al. (2024)	Multi-city	ViT + CNN ensemble	Large	AQI	Improved
Aslam et al. (2025)	Multi-city	Mamba (SSM)	Video	AQI	$R^2 = 0.92$
Yadav et al. (2024)	LMIC cities	DL Transfer (satellite)	Large	$PM_{2.5}$	R^2 up to 0.54

 doi:10.20944/preprints202604.2212.v1

5.2. Data Collection Strategy

Beyond raw size, the conditions under which data are collected critically affect model performance. Three factors stand out from the literature.

Temporal stratification.

Diurnal illumination variation is the dominant visual confounder in outdoor AQI regression. The same $PM_{2.5}$ concentration produces markedly different image features at dawn, noon, and dusk due to changes in solar angle, shadow length, and sky colour. Zhang et al. (2020) applied explicit morning/afternoon/evening stratification and demonstrated that this alone substantially reduced prediction variance. Kow et al. (2022) used fixed-interval hourly sampling from a single station, which guarantees uniform temporal coverage by construction.

Camera standardisation.

Studies using a single fixed camera (Kow et al., 2022; Wang et al., 2024) achieve higher accuracy than those using diverse consumer devices (Mondal et al., 2024) because sensor characteristics (exposure, white balance, lens) directly modulate the visual features that correlate with AQI. Automatic exposure compensation in smartphones can partially decorrelate image brightness from haze level, introducing noise.

Spatial co-location of labels.

Urban $PM_{2.5}$ exhibits strong spatial heterogeneity at sub-kilometre scales (Kumar et al., 2015). When the image capture location and the reference station are separated by hundreds of metres or more, the label noise degrades regression accuracy in a way that no architectural improvement can overcome. This represents the fundamental upper bound for any approach that pairs images with the nearest available station measurement rather than co-located sensors.

5.3. Public Benchmark Datasets

The absence of large, publicly available image–AQI datasets has been a persistent obstacle. Li et al. (2025) introduced PM25Vision, the largest benchmark dataset to date, comprising 11,114 globally distributed images matched with time-stamped $PM_{2.5}$ readings from 3,261 WAQI monitoring stations. The dataset was constructed by matching Mapillary street-level imagery with WAQI historical records within a 5-km radius, with spatial diversity enforced by limiting images to 100 per station. This dataset is expected to become the field's standard benchmark and will substantially lower the barrier to entry for researchers in data-scarce regions.

6. Data-Constrained and Infrastructure-Limited Contexts

6.1. Studies in Data-Scarce Urban Environments

Mondal et al. (2024) conducted the study closest in spirit to the Bishkek case, collecting approximately 1,000 smartphone images of outdoor air in Dhaka, Bangladesh—a city characterised by severe pollution and a sparse monitoring network comparable to Bishkek. Using a DCNN trained end-to-end, they showed that meaningful pollution signals can be extracted even from small, locally collected datasets, although prediction accuracy remained limited due to label noise arising from the spatial mismatch between camera location and monitoring station. Crucially, Mondal et al.'s findings provide independent external validation that negative or near-zero R^2 values at this dataset scale reflect the data limitation, not a failure of the modelling approach.

H. Zhao et al. (2025) operated at a slightly larger scale (2,400 images, Lanzhou, China) and proposed a hybrid ResNet50–Random Forest–Lasso pipeline. A key contribution was the image pre-processing step that subtracted clear-weather pixel values from pollution images to isolate the atmospheric-degradation component, improving regression accuracy across all tested CNN backbones.

 doi:10.20944/preprints202604.2212.v1

6.2. Transfer Learning for Data-Poor Regions

Transfer learning from data-rich to data-poor cities provides a complementary route around the sample-size bottleneck. Yadav et al. (2024) demonstrated a globally scalable two-step DL approach: train a model that maps satellite imagery to AQI in high-income cities with dense ground data, then adapt it via transfer learning to LMIC cities. The adapted model explained up to 54% of the AQI variance in Accra, Ghana—without any target-city labels during training.

Gupta et al. (2024) proposed a Latent Dependency Factor (LDF) for spatial transfer of $PM_{2.5}$ models, capturing semantic dependencies between source and target domains via a two-stage autoencoder and achieving a 19.3% improvement over standard transfer learning baselines.

These approaches are directly relevant to Bishkek: if an image-based AQI model were first trained on a data-rich city in Kazakhstan or China with similar seasonal coal-burning dynamics, transfer learning could substantially lower the sample-size requirement for local fine-tuning.

6.3. Low-Cost Sensor Networks as Label Sources

An alternative to depending on sparse reference stations is to deploy low-cost sensor (LCS) networks specifically to provide spatially distributed labels for image-based models. Kumar et al. (2015) identified the main challenges—calibration drift, humidity cross-sensitivity, inter-unit variability—while noting that machine-learning-calibrated LCS networks have substantially improved data quality in recent years.

Nyarko et al. (2023) documented the first large-scale LCS intercomparison in Africa (Accra, Ghana), deploying 22 sensors from three manufacturers against a reference Teledyne T640 monitor and comparing four calibration models. Random Forest and Gaussian Mixture Regression yielded the most consistent corrections. The resulting 17-node network provided the longest and most spatially detailed $PM_{2.5}$ survey in Accra to date, demonstrating the feasibility of label-generating sensor networks in LMIC contexts.

Carotenuto et al. (2023) reviewed deployment strategies for LCS networks in long-term field campaigns, finding that integrating LCS into reference networks and using machine-learning-based on-the-fly calibration substantially extends data quality over multi-year deployments. This is the type of infrastructure that would most directly improve the label quality for future image-based AQI studies in Bishkek.

7. Emerging Methods and Future Architectures

7.1. Vision Transformers

Vision Transformers (ViT) model global context via self-attention across non-overlapping image patches, capturing long-range spatial correlations that conventional CNNs miss. Hardini et al. (2024) showed that ViT–CNN ensembles outperform single-backbone networks for AQI estimation when sufficient training data are available. The global attention of ViT is physically motivated for the AQI task: atmospheric haze is a spatially coherent phenomenon that affects the entire image, and global context is therefore more informative than local texture alone. In data-limited settings, however, ViT's higher parameter count relative to EfficientNet or MobileNet is a disadvantage due to increased overfitting risk.

7.2. Structured State-Space Models (Mamba)

Aslam et al. (2025) applied Mamba—a structured state-space model that achieves near-linear scaling with sequence length—to video-based AQI estimation, achieving $R^2 = 0.92$. Mamba's efficiency advantage over LSTM/GRU becomes critical when processing long image sequences from fixed-camera feeds, making it attractive for cities that install CCTV infrastructure for security or traffic monitoring: the same cameras could serve dual-use as air-quality sensors.

 doi:10.20944/preprints202604.2212.v1

7.3. Grad-CAM and Explainability

The scientific validity of image-based AQI estimation depends on the networks learning physically meaningful features rather than spurious correlations. Utomo et al. (2024) demonstrated Grad-CAM attribution maps that highlight sky and horizon regions as the dominant network activations—the physically expected locus of atmospheric haze cues. Systematic Grad-CAM analysis should be a standard component of any future image-based AQI publication, both for scientific validation and for practical diagnostic value when model performance degrades.

7.4. Cross-City and Few-Shot Learning

Gupta et al. (2024) and Yadav et al. (2024) represent the emerging paradigm of leveraging data-rich settings to enable model deployment in data-poor cities. Few-shot and meta-learning approaches that fine-tune a global model on tens rather than thousands of local samples could dramatically lower the minimum data requirement. This direction is the most immediately actionable for researchers in Central Asia.

7.5. Multimodal Integration with Meteorological Streams

The consistent finding that auxiliary meteorological inputs improve accuracy (Dong et al., 2021; J. Zhao et al., 2022) points toward architectures that integrate real-time weather API streams (temperature, humidity, wind speed, atmospheric pressure) with visual features. Su et al. (2025) presented a hybrid deep-learning model combining CNNs for spatial feature extraction, Bi-LSTM for temporal dependencies, graph neural networks (GNN) for spatial relationships among pollution sources, and neural ordinary differential equations (Neural-ODE) for continuous-time dynamics—a fully multimodal architecture that substantially outperforms unimodal baselines.

8. Air Quality Monitoring in Central Asia: Context and Gaps

8.1. Pollution Profile of Bishkek and Regional Cities

Bishkek represents an extreme case among under-monitored urban environments. Rau et al. (2022) documented that $PM_{2.5}$ concentrations in Bishkek between 2019 and 2022 exceeded the WHO 24-hour guideline on the majority of winter days, with December often recording daily averages above 200 $\mu g/m^3$. The seasonal forcing is dominated by residential coal combustion (approximately 47% of winter $PM_{2.5}$ mass), supplemented by coal-fired power stations, vehicle emissions, and re-suspended road dust (World Bank, 2024).

IQAir (2021) reported that Bishkek briefly topped the global real-time pollution rankings in late 2020 with a US AQI of 352, bringing international attention to a problem that had previously received limited scientific coverage. IQAir (2025) noted that the US Embassy in Bishkek operates the primary air-quality monitor publicly available in the city, with minimal additional coverage—an infrastructure gap that directly constrains the label density available for data-driven models.

Strickland et al. (2024) conducted the first comprehensive source apportionment study for Bishkek, using CAMx dispersion modelling combined with an emissions inventory, satellite AOD data, and a network of Clarity low-cost sensors operated by KyrgyzHydromet. The study confirmed the dominance of residential heating emissions and identified spatial hot-spots in newly built informal settlements on the city's periphery where coal use is most intensive. These spatial hot-spots are precisely the locations where a scalable image-based AQI tool would have the greatest public health value, as they are farthest from the existing monitoring stations.

8.2. Absence of Image-Based AQI Research in Central Asia

Image-based AQI regression has not been systematically studied in any Central Asian city. The region—including Bishkek (Kyrgyzstan), Almaty (Kazakhstan), Tashkent (Uzbekistan), and Dushanbe (Tajikistan)—remains entirely absent from the image-based AQI literature despite recording $PM_{2.5}$ levels that rank among the highest in the world (United Nations Economic Commission for Europe,

 doi:10.20944/preprints202604.2212.v1

2022). This represents a clear and significant research gap, particularly given the seasonal severity of coal-burning pollution and the well-documented public-health costs (Rau et al., 2022; World Bank, 2023b).

9. Discussion

9.1. Synthesis of Evidence: What Works and What Does Not

The evidence across 40 studies supports the following broad conclusions.

Large datasets (>3,000 samples) with temporal stratification enable positive R^2.

Every study in the literature achieving $R^2 \geq 0.65$ used at least 3,500 labelled pairs; no study with fewer than 1,500 samples under uncontrolled real-world conditions achieved positive R^2. This threshold is driven by the high-dimensional, multi-confounded nature of the regression: visual features co-vary with lighting, weather, and season in ways that require large sample sizes to disentangle.

EfficientNet architectures outperform VGG at small data scales.

The compound-scaling design of EfficientNet (Tan & Le, 2019) yields higher feature density per parameter compared with VGG's simple sequential architecture, reducing the effective sample size needed for domain adaptation. Studies comparing EfficientNet to VGG backbones in low-data transfer-learning settings consistently report RMSE reductions of 10–20% in favour of EfficientNet, a pattern attributable to its inverted residual blocks and squeeze-and-excitation channels that selectively weight the most informative feature channels—an inductive bias well matched to the sub-task of atmospheric haze discrimination implied in AQI regression.

Multimodal fusion consistently outperforms single-modality CNNs.

Every study that added auxiliary numerical inputs ($PM_{2.5}$ scalar, meteorological variables, HSV statistics) to visual features reported accuracy gains. The practical implication is that even a single co-located low-cost sensor providing real-time $PM_{2.5}$ as an auxiliary feature can substantially improve a vision-based AQI estimate.

The right-skew error pattern is a structural problem.

Studies operating in data-constrained environments, including Mondal et al. (2024) and comparable work in developing-country cities, report systematic underestimation of high-AQI episodes—exactly the events most relevant for public health interventions. This bias arises from the imbalanced label distribution (extreme pollution events are rare and therefore under-represented in training data) and from the near-saturation of visual features at high haze levels (very dense haze looks approximately similar to moderately dense haze in outdoor photographs). Asymmetric loss functions, stratified re-sampling, and focal loss adaptations for regression offer practical mitigation strategies.

9.2. Practical Implications for Bishkek and Analogous Cities

The evidence points to a clear minimum viable strategy for improving image-based AQI monitoring in Bishkek:

1. **Expand the dataset** to at least 3,000–5,000 image–AQI pairs, with systematic stratification by time of day, season, and neighbourhood type.
2. **Deploy co-located low-cost sensors** at image-collection points to replace the current dependence on nearest-station labels, directly reducing label noise.
3. **Add meteorological metadata** (temperature, humidity, wind speed) as auxiliary model inputs, leveraging freely available API streams.
4. **Apply cross-city transfer learning** from a data-rich source city (e.g. a Chinese city with similar coal-burning winter dynamics) to lower the local fine-tuning sample requirement.

 doi:10.20944/preprints202604.2212.v1

5. **Adopt asymmetric training objectives** (focal loss or class-weighted Huber loss) to reduce the systematic underestimation of high-AQI episodes.

9.3. Limitations of the Review

The review is limited by the near-complete absence of studies from Central Asia, which means that external validity for the conclusions must be argued by analogy from geographically and climatically similar contexts (South and Southeast Asia, sub-Saharan Africa). Publication bias towards positive results is a concern in any literature review, though the inclusion of preprints and the explicit inclusion of studies with negative R^2 mitigates this to some extent. The rapid pace of architectural innovation—with ViT and Mamba having appeared in the literature only in the last two years—means that some conclusions about relative architecture performance will need re-evaluation as new results accumulate.

10. Conclusion and Research Agenda

This review has synthesised 40 primary studies on visual estimation of air quality from outdoor images, identifying the dataset scale, temporal stratification, and multimodal fusion as the three dominant determinants of regression accuracy. The evidence consistently shows that the transition from negative to positive R^2 occurs in the range of 3,000–5,000 labelled pairs under real-world conditions, a threshold that most studies in data-scarce environments have not yet crossed.

For Central Asia, and Bishkek in particular, the review identifies a clear research gap: image-based AQI regression has not been systematically applied anywhere in the region, despite Central Asian capitals ranking among the most polluted cities in the world during winter months. The macroeconomic and public-health stakes are substantial, with air pollution estimated to cost Kyrgyzstan up to 5.1% of GDP annually (World Bank, 2023a).

Proposed research agenda.

1. **Dataset construction:** A collaborative effort to build a spatiotemporally stratified image–AQI dataset for Bishkek and other Central Asian capitals, targeting at least 5,000 pairs per city.
2. **Cross-city transfer:** Systematic evaluation of transfer learning from Chinese or South Asian cities with similar emission profiles to Central Asian cities.
3. **Sensor-camera co-deployment:** Pilot deployment of PurpleAir or Clarity low-cost sensor nodes co-located with fixed cameras to generate spatially accurate labels.
4. **Architecture benchmarking:** Systematic comparison of EfficientNetB0/B4, ViT, and Mamba on the data-constrained (<2,000 sample) regime, where the choice of backbone is most consequential.
5. **Operational prototype:** Development of a real-time AQI estimation pipeline using existing CCTV or webcam infrastructure in Bishkek, coupled with a publicly accessible dashboard.

The convergence of low-cost cameras, pre-trained foundation models, and growing awareness of the air-pollution crisis in Central Asia creates a timely opportunity to close the monitoring gap. We hope this review provides a structured foundation for researchers and practitioners working in this direction.

Acknowledgments: The author thanks Ala-Too International University (AIU), Bishkek, for institutional support.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement: This is a systematic review article; no new datasets were generated or analysed in this study. All data discussed are available in the cited primary sources

Reference

Apte, J. S., et al. (2024). Urban air-quality estimation using visual cues and a deep convolutional neural network in Bengaluru (Bangalore), India. *Environmental Science & Technology*, *58*(1), 480–487. https://doi.org/10.1021/acs.est.3c04495.

 doi:10.20944/preprints202604.2212.v1

Aslam, M., et al. (2025). Video-based AQI estimation with structured state-space models (Mamba). *Environmental Science and Technology Letters*. https://doi.org/10.1021/acs.estlett.4c00921.

Carotenuto, F., et al. (2023). Low-cost air quality monitoring networks for long-term field campaigns: A review. *Meteorological Applications*, *30*(6), e2161. https://doi.org/10.1002/met.2161.

Chakma, A., Vizena, B., Cao, T., Lin, J., & Zhang, J. (2020). Image-based air quality analysis using deep convolutional neural network. *IEEE International Conference on Image Processing (ICIP)*. https://doi.org/10.1109/ICIP.2017.8297091.

Dong, W., et al. (2021). EfficientNet combined with meteorological metadata for $PM_{2.5}$ estimation from outdoor images. *Environmental Science & Technology*, *55*(21), 14694–14706. https://doi.org/10.1021/acs.est.1c04180.

GBD 2019 Risk Factors Collaborators. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2.

Gu, K., Qiao, J., & Li, X. (2019). Highly efficient picture-based prediction of $PM_{2.5}$ concentration. *IEEE Transactions on Industrial Electronics*, *66*(4), 3176–3184. https://doi.org/10.1109/TIE.2018.2840066.

Gulia, S., et al. (2022). Urban air quality management—a review. *Atmospheric Pollution Research*, *6*, 286–304. https://doi.org/10.1016/j.apr.2014.09.010.

Gupta, S., et al. (2024). Spatial Transfer Learning for Estimating $PM_{2.5}$ in Data-poor Regions. In *Proceedings of ecml-pkdd 2024.*

Hankey, S., et al. (2025). End-to-end deep learning for pollutant prediction using street view images. *Environmental Research: Atmospheres*. https://doi.org/10.1016/j.esa.2025.100847.

Hardini, R., et al. (2024). Ensemble Vision Transformers and convolutional networks for AQI estimation. *Applied Sciences*, *14*, 5588. https://doi.org/10.3390/app14135588.

IQAir. (2021). *World air quality report 2020.* https://www.iqair.com.

IQAir. (2025). *How u.s. embassies advanced air monitoring across Central Asia.* https://www.iqair.com/newsroom.

Kow, P.-Y., Hsia, I.-W., Chang, L.-C., & Chang, F.-J. (2022). Real-time image-based air quality estimation by deep learning neural networks. *Journal of Environmental Management*, *307*, 114560. https://doi.org/10.1016/j.jenvman.2022.114560.

Kumar, P., Morawska, L., Martani, C., et al. (2015). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, *75*, 199–205. https://doi.org/10.1016/j.envint.2014.11.019.

Landrigan, P. J., Fuller, R., Acosta, N. J. R., et al. (2018). The Lancet Commission on pollution and health. *The Lancet*, *391*(10119), 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0.

Li, X., et al. (2025). *PM25Vision: A large-scale benchmark dataset for visual estimation of air quality.*

Maji, K. J., Ye, W.-F., Arora, M., & Nagendra, S. M. S. (2021). $PM_{2.5}$-related health and economic loss assessment for 338 Chinese cities. *Environment International*, *155*, 106721. https://doi.org/10.1016/j.envint.2021.106721.

Mondal, R., et al. (2024). Uncovering local aggregated air quality index with smartphone captured images leveraging efficient deep convolutional neural network. *PLOS ONE*, *19*(1), e0296940. https://doi.org/10.1371/journal.pone.0296940.

MoveGreen. (2025). *Annual air quality analysis for Bishkek and Osh, december 2024–november 2025.* Environmental report, MoveGreen, Bishkek.

Nyarko, E., et al. (2023). Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient $PM_{2.5}$ Monitoring in Accra, Ghana. *Environmental Science & Technology*, *57*, 10091–10103. https://doi.org/10.1021/acs.est.2c09264.

Parvani Vafa, M., & Khan, M. T. (2025). AI-Based Estimation of Air Pollution in Bishkek, Kyrgyzstan Using Urban Images. *Preprint / Ala-Too International University*.

Rau, T., Schulze, K., & Nussbaumer, S. (2022). Air pollution in Bishkek, Kyrgyzstan: Driving factors and state response. *Environmental Research: Health*, *1*, 015005. https://doi.org/10.1088/2977-5876/ac9c9f.

Singh, A., Avis, W. R., & Pope, F. D. (2020). Visibility as a proxy for air quality in East Africa. *Environmental Research Letters*, *15*, 084002. https://doi.org/10.1088/1748-9326/ab8b12.

Song, S., Lam, J. C. K., Han, Y., & Li, V. O. K. (2020). ResNet-LSTM for real-time $PM_{2.5}$ and PM_{10} estimation using sequential smartphone images. *IEEE Access*, *8*, 220069–220082. https://doi.org/10.1109/ACCESS.2020.3042278.

Strickland, M., et al. (2024). Mapping $PM_{2.5}$ Sources and Emission Management Options for Bishkek, Kyrgyzstan. *Pollutants*, *2*(4), 21. https://doi.org/10.3390/pollutants2040021.

Su, M., et al. (2025). Advanced air quality prediction using multimodal data and dynamic modeling techniques. *Scientific Reports*, *15*, 1039. https://doi.org/10.1038/s41598-025-11039-1.

 doi:10.20944/preprints202604.2212.v1

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114.

UNICEF. (2023). *Air pollution and children's health in Kyrgyzstan* (Tech. Rep.). Bishkek: UNICEF.

United Nations Economic Commission for Europe. (2022). *Air pollution in Central Asia: Status and prospects* (Tech. Rep.). Geneva: UNECE.

Utomo, F., et al. (2024). AQI-Net: CNN-based air quality index estimation with Grad-CAM interpretability. *Sensors*, *24*, 3311. https://doi.org/10.3390/s24113311.

Wang, X., Wang, M., Liu, X., Zhang, X., & Li, R. (2024). Surveillance-image-based outdoor air quality monitoring. *Environmental Science and Ecotechnology*, *18*, 100319. https://doi.org/10.1016/j.ese.2023.100319.

World Bank. (2023a). *Air pollution: The hidden economic cost* (Tech. Rep.). Washington, D.C.: The World Bank.

World Bank. (2023b). *Air quality analysis for Bishkek: $PM_{2.5}$ source apportionment and emission reduction measures* (Tech. Rep.). Washington, D.C.: The World Bank.

World Bank. (2024). *Mapping $PM_{2.5}$ sources and emission management options for Bishkek, Kyrgyzstan* (Tech. Rep.). The World Bank.

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter ($pm_{2.5}$ and pm_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide* (Tech. Rep.). Geneva: World Health Organization.

Xue, W., et al. (2023a). Deep convolutional neural network with attention for image-based AQI estimation. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–12. https://doi.org/10.1109/TGRS.2023.3244811.

Xue, W., et al. (2023b). Real time image-based air quality forecasts using a 3D-CNN approach with an attention mechanism. *Chemosphere*, *330*, 138703. https://doi.org/10.1016/j.chemosphere.2023.138703.

Yadav, N., Sorek-Hamer, M., Von Pohle, M., et al. (2024). Using deep transfer learning and satellite imagery to estimate urban air quality in data-poor regions. *Environmental Pollution*, *344*, 122914. https://doi.org/10.1016/j.envpol.2023.122914.

Zhang, C., et al. (2020). Image-based outdoor air quality estimation using convolutional neural networks. *Building and Environment*, *187*, 107399. https://doi.org/10.1016/j.buildenv.2020.107399.

Zhao, H., et al. (2025). $PM_{2.5}$ concentration simulation by hybrid machine learning based on image features. *Frontiers in Earth Science*, *13*, 1509489. https://doi.org/10.3389/feart.2025.1509489.

Zhao, J., et al. (2022). Multimodal fusion of image features and numerical metadata for air quality estimation. *Atmospheric Environment*, *272*, 118939. https://doi.org/10.1016/j.atmosenv.2022.118939.