Vision-Based Estimation of PM2.5 from Surveillance Images

Dipti Mitra and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, I*wate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
email: s231x025@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—The paper proposes a vision-based approach for measuring fine Particulate Matter (PM2.5) concentrations by utilizing environmental images as input. A dataset was created by acquiring surveillance camera-captured images from multiple locations in Japan, forming a pair of collected outdoor images and Ground Truth (GT) PM2.5 observation data obtained from monitoring stations. Two preprocessing steps (image dehazing and semantic segmentation) were used to enhance prediction accuracy under varying atmospheric and meteorological conditions. The dehazing method mitigates visual degradation caused by haze, while semantic segmentation extracts and determines object-level information and the coverage amount of extracted objects relevant to PM2.5 estimation. The proposed image-based system combines the dehazed images and segmentation masks, which are then input into a deep learning-based regression model to predict PM2.5 concentrations. The experimental results demonstrate that integrating dehazed images and segmentation masks reduces prediction errors and produces more consistent estimates compared to using original images and other input configurations alone or in combination. The findings indicate that combining enhanced visual representations and segmented objects with deep learning models can serve as an effective and scalable complement to traditional air quality monitoring systems.**

Keywords-computer vision; dehazing; semantic segmentation; PM2.5 forecasting; regression.

I. INTRODUCTION

Despite substantial improvements in air quality over recent decades, air pollution remains a crucial environmental threat and public health issue in Japan. Rapid industrialization and urban development in the postwar period led to severe air pollution problems. However, strict environmental rules and continuous monitoring have significantly mitigated many conventional pollutants. Nevertheless, an air pollutant, fine Particulate Matter (PM2.5), defined as particles with an aerodynamic diameter of less than 2.5 μm, continues to pose a concern, particularly in urban and industrial areas [1]. Particle components are generally classified as solid and liquid substances released into the air from domestic sources, such as indoor activities, industrial emissions, volcanic eruptions, and others. These particles are transmitted through the air and eventually return to the ground [2]. PM2.5 is small enough to enter the human red blood cell, which is 7.8 μm in diameter, and as a result, it can penetrate deeply into the respiratory system. It is linked to adverse health effects, including cancer, cardiovascular diseases, and asthma. According to the World Health Organization (WHO), seven million people die as o result of PM2.5 particles each year globally [2]. Therefore, it is essential to explore various methods to accurately measure PM2.5 particles. Currently, computer vision technologies are being applied due to their high spatial coverage, scalability, sensitivity to changes in visibility, real-time capabilities, easy deployment, and low cost.

Japan is an island country, where ambient PM2.5 concentrations are routinely monitored using ground-based air quality monitoring stations across its forty-seven prefectures. These stations rely on PM2.5 sensors that measure concentrations of particles. Laser scattering and Infrared Rays (IR) are used to detect particles, which provide accurate point-based observations. Nevertheless, the monitoring stations within each prefecture limit their spatial representativeness as a single or a small number of sensors cannot capture the PM2.5 spatial variability over enormous and diverse geographic locations for detection. Although deploying additional sensors could improve coverage, such an approach is costly and logistically demanding. The sensors cannot provide accurate reading due to different factors and small sensitivity. Therefore, they need to be estimated precisely in different domains of each prefecture as they are scattered around the air.

In this paper, vision-based PM2.5 forecasting is proposed. To identify particulates in the air, we have utilized geographical images of Japan's various prefectures as input. We have performed image dehazing and semantic segmentation, and an AI technique called the regression method to accurately estimate PM2.5 concentration, especially in the presence of haze and objects in the scene. Regression approaches have been applied to visual input to model the relationship between environmental image features and their associated PM2.5 values. When images are used as inputs, the model extracts visual information, reflecting haze density, changes of visibility, contrast degradation, and color attenuation, and regresses these features to predict numerical PM2.5 concentrations. Moreover, it enables quantitative estimation of particulate matter levels, making it suitable for continuous monitoring and forecasting tasks rather than categorical classification. Experiments have been conducted using original images, dehazed images, and segmented objects. A regression model is used to examine the actual relationship between images/masks and their corresponding PM2.5 values. The testing results demonstrate lower correlation in terms of original images, dehazed images, semantic segmentation, and combined input, and an improved correlation when applying dehazed and object segmentation.

The paper is divided into five sections. In Section I, an overview of PM2.5, along with the research problems and objectives, is presented. In Section II, the existing literature on vision-based PM2.5 estimation and prediction approaches is reviewed. In Section III, the proposed methodology, including data collection, dataset construction, pre-processing steps, and the architecture of the regression model, is described. In Section IV, the experimental results are presented, and their implications on PM2.5 prediction are discussed. Lastly, in Section V, the paper is concluded and outlines the directions for future work.

II. LITERATURE REVIEW

Numerous research papers have applied vision-based approaches for forecasting PM2.5 air pollutants as a complementary alternative to traditional sensor-based monitoring systems. Previous studies demonstrated the viability of retrieving air quality information directly from the visual cues present in environmental scenes. For example, the vision-based techniques [3] and [4] rely on extracting haze-related features by quantifying saturation, contrast degradation, color attenuation, and information losses from manually designed statistical images to represent the haze-related visual degradation. These methods provide interpretability and relatively low computational cost. However, their dependence on handcrafted image features limits their robustness in complex outdoor environments under varying illumination, lighting, meteorological conditions, and camera conditions.

With the rapid development of deep learning technologies, researchers started adopting Convolutional Neural Network-based architectures to automatically learn discriminative visual characteristics from the images. Studies, such as a hybrid architecture Deep Neural Network model [5], presented a base model named Convolutional Neural Network (CNN), while an output layer named Long Short-Term Memory (LSTM) network was employed. They demonstrated that the two integrated models can automatically capture spatial features and temporal dependencies within the extracted feature sequences, such as pollution-related patterns, by utilizing the hourly images of the sky and surrounding environment in Bangkok, Thailand. Similarly, another approach, deep learning-based image analysis to predict PM2.5 concentrations [6], leverages existing surveillance infrastructure to achieve wide-area monitoring. In this study, a ResNet-based image analysis method was utilized, turning an existing traffic camera into a PM2.5 sensor. To create a dataset, hourly traffic images and their PM2.5 values were attained over a six-month period from a traffic camera and the nearest monitoring station. In the first phase, the neural network model ResNet50 was used to train the acquired dataset. Moreover, a second phase model, Random Forest, was used, where the outputs of the neural network are utilized as input to predict overall hourly PM2.5 values. While these CNN-based approaches enhance prediction accuracy, they remain sensitive to scene-specific factors, such as camera viewpoints, background complexity, and atmospheric visibility, which can degrade model performance across locations.

Several studies proposed time-series modeling methods to address temporal dependencies in particulate matter dynamics. The methods [7] and [8] combined visual information with temporal learning frameworks, such as Long Short-Term Memory and encoder-decoder architectures, to predict PM2.5. The integrated dual-channel model [7] learned intuitive spatiotemporal features from a series of surveillance images and temporal information from atmospheric conditions, meteorological conditions, and temporal data for precise time-series forecasting of PM2.5 and PM10 concentrations. Likewise, the spatio-temporal model [8] captured and measured feature correlation and loss in particular locations by using an image-like technique at a country-wide level for PM2.5 prediction. These approaches improved the accuracy of prediction by modeling both spatial and temporal correlations. However, these methods often require large, continuous datasets and assume stable visual quality, making them sensitive to haze, clouds, and adverse weather.

More recent studies proposed advanced architecture and multimodal learning to improve the robustness of prediction. A Vision Transformer-based model presented in [9] can effectively process and learn complex data with spatiotemporal dependencies and deep features from image data, even in the absence of extensive labeled data, when leveraging the model. In parallel, the multimodal approach Contrastive Learning–Image Pre-training (CLIP) [10] employs transformers as backbones to learn the combined visual features and contextual information by leveraging 2D image data acquired from mobile devices. The model was trained on a Graphics Processing Unit (GPU) and Single-Board Computer (SBC) to enhance the accuracy and scalability of air quality monitoring. Although the methods demonstrate improved accuracy, they often require greater computational resources and depend on access to various data modalities.

On the contrary, in domain-specific applications [11], computer vision technology and a regression model are used to extract the real-time traffic volume and street-view information from the traffic images and to predict the road concentration of PM2.5, which is trained on meteorological conditions, traffic volume, and building variables. This approach demonstrates the effectiveness of vision-based models in complex urban microenvironments. Although these methods achieve promising results in localized settings, their generalization to diverse geographical areas and broader atmospheric conditions remains a challenge.

The existing vision-based PM2.5 forecast approaches demonstrated the efficacy of deep learning and temporal modeling in extracting particulate matter-related visual patterns. However, most prior work relies on raw image data to capture visual information related to atmospheric conditions, such as haze and visibility degradation, meteorological conditions, particulate matter-related predictors, and the development of deep learning technologies to map these visual features to PM2.5 concentrations. These

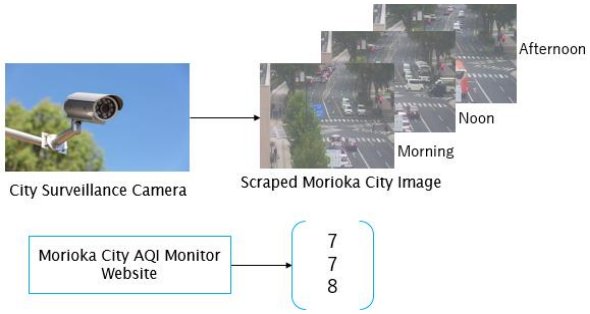

Figure 1. Collecting data using Web-scraping method.

limitations motivate the need for using dehazed images integrated with object segmentation images, which can provide clearer visual cues to atmospheric conditions, while object segmentation enables the model to focus on semantically meaningful regions in the scene for consistent PM2.5 prediction. This proposed approach will be discussed in the following section.

III. VISION-BASED PM2.5 FORECASTING

A. Data Collection Approach

In this study, Web-scraping is employed as a data acquisition method, as shown in Figure 1, to attain environmental scene images along with their associated PM2.5 concentration values. The images are obtained from publicly available city surveillance camera footage from multiple locations, whereas the corresponding PM2.5 observation data are retrieved from official air quality index monitoring websites (www.aqi.in). Data collection is

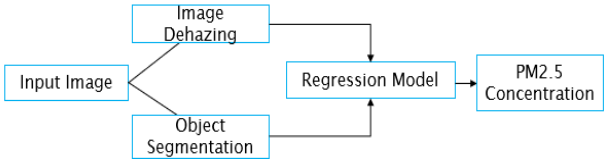

Figure 2. Block diagram of proposed system.

conducted at regular intervals during three daily time periods: morning (9:00 ~ 12:00), noon (12:00 ~ 13:00), and afternoon (14:00 ~ 17:00) from October 2024 to March 2025. The visual data and particulate matter measurements are temporally aligned to ensure consistency between image observations and Ground Truth PM2.5 concentrations.

B. Data Description

The dataset utilized in this work consists of paired environmental images and corresponding PM2.5 values obtained from various urban locations in Japan. Each data sample comprises an outdoor image and its associated PM2.5 concentration, creating a supervised dataset for regression-based particulate matter prediction. In addition to the original images, the dataset is further extended with derived representations, namely dehazed images and their associated segmentation masks, which are used to improve visual quality and extract object-level cues.

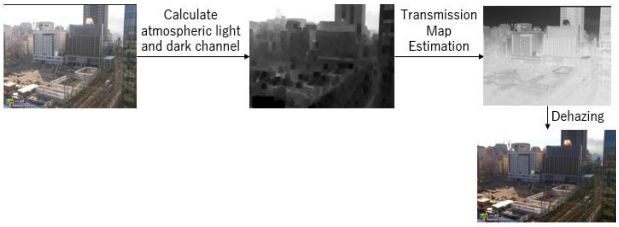

Figure 3. Process of dehazing the Sapporo city image.

The images capture a wide range of environmental conditions, including variations in weather, such as sunny, cloudy, and snowy scenarios. These conditions introduce differences in visibility, illumination, and atmospheric appearance, and these are significant factors for vision-based PM2.5 estimation. The images are collected at fixed resolutions, relying on the camera sources, ensuring that within each location, while preserving sufficient visual details for feature extraction.

In the dataset, each sample is represented by multiple inputs, including the original image, the dehazed version of the image, and the corresponding segmentation mask, along with the Ground Truth PM2.5 concentration. This multi-representation structure empowers the model to manipulate both enhanced visual information and semantic scene cues. Moreover, the dataset has temporal variation by

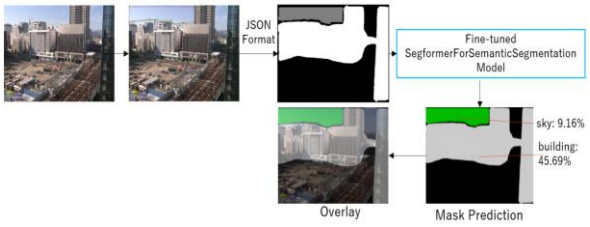

Figure 4. Model fine-tuning and object coverage percentage estimation.

incorporating samples obtained at different times of the day, as well as spatial variability across different cities with distinct urban and atmospheric conditions.

C. Data Pairing

The data pairing is performed to correspond to each environmental image with a Ground Truth PM2.5 value. The corresponding PM2.5 concentration is assigned based on both spatial and temporal alignment for each attained image. Firstly, a camera-to-station correspondence is considered by identifying the nearest air quality monitoring station to each camera location, utilizing Google Maps. The approximate distance between the camera and station is obtained from the map, and only stations located within a threshold distance of 10 kilometers have been considered. This ensures that the selected particulate matter concentrations reasonably reflect the atmospheric conditions seen in the images. Furthermore, temporal alignment is performed by matching each image with the closest available PM2.5 measurement based on its timestamp. This stage verifies consistency between the captured visual scene and the recorded PM2.5 measurement. Finally, each valid image is paired with its associated particulate matter value to create a labeled data sample.

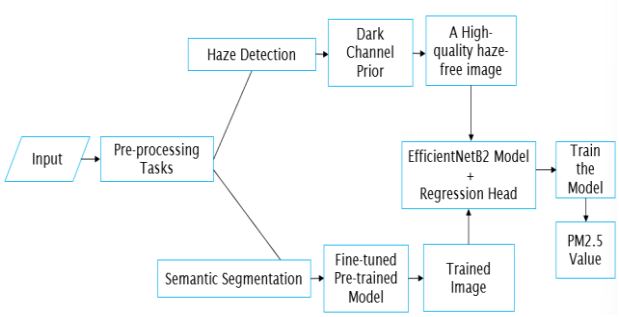

Figure 5. Flowchart of proposed work.

For data cleaning, repeated images and images captured from altered camera positions are excluded to establish consistency in scene representation. Initially, around 200–300 images are obtained per city; after the filtering process, the remaining dataset is reduced to approximately 70–100 images per city. Eventually, the filtered images, along with their corresponding PM2.5 values, are exploited to construct the final dataset.

D. System Model

The proposed image-based system's block diagram is illustrated in Figure 2, where environmental images are used as input to the proposed framework. A dehazing process is applied to enhance image quality by reducing the effects of atmospheric scattering. In addition, object segmentation is subsequently performed to identify the different key scene components and then compute the object coverage amount. These refined visual representations and the extracted object features are fed into the regression model. The model learns the pattern between these two inputs, along with the Ground Truth PM2.5 numerical data, to predict PM2.5 concentrations precisely.

E. Image Dehazing using Dark Channel Prior Algorithm

The first pre-processing step, the dark channel prior approach [12], is employed, as depicted in Figure 3, to remove the influence of atmospheric haze from the retrieved environmental images. This method is based on the observation that at least one color channel exhibits very low intensity values in most non-sky regions of haze-free outdoor images. Based on this assumption, the dark channel for each input image is determined first to estimate the spatial distribution of haze. Afterwards, the atmospheric light is calculated by identifying the brightest pixels in the dark channel, which represent areas with the highest haze concentration. Moreover, a transmission map is estimated by using the determined atmospheric light and dark channel information to illustrate the portions of the scene radiance that reach the camera sensor. Lastly, the dehazed image is recovered by restoring the scene radiance using the estimated parameters, including atmospheric light and transmission map, resulting in a dehazed image that exhibits enhanced visibility and contrast, providing much clearer visual features for further analysis.

F. Semantic Segmentation

Semantic segmentation constitutes the second pre-processing phase of the proposed system architecture, as described in Figure 4, to obtain fine-grained, object-level cues from the environmental scenes. The aim of this stage is to identify and distinguish semantically meaningful areas or classes, such as sky, buildings, roads, vegetation, water, and other structural components within the different urban scenes that are potentially correlated with PM2.5 concentration. The segmentation process provides further contextual information beyond global image appearance by explicitly modeling the spatial distribution of those objects.

The Ground Truth segmentation masks are constructed from the original images by manually annotating object classes, including sky, building, water, road, etc., exploiting the LabelMe annotation tool. The annotations are converted into JSON format to generate the corresponding segmentation masks.

These masks are utilized to fine-tune a deep learning model based on the SegFormer architecture on the acquired dataset. The model is adopted due to its robust performance in capturing both local and global contextual cues, allowing it to adapt to the visual characteristics of the scenes. After fine-tuning the model, the trained model generates pixel-wise segmentation masks for each input image. Eventually, the predicted masks are subsequently utilized to determine the coverage amount of each segmented object class in the scene, providing quantitative object-level features to accurately estimate PM2.5.

The performance of the segmentation model is evaluated using standard metrics, such as pixel accuracy, mean Intersection over Union (mIoU), F1-score, and recall. The results specify that the model obtains high segmentation performance across all cities. Specifically, Sapporo and Chofugaoka exhibit the highest performance with pixel accuracy and F1-score of 0.99, along with mIoU values of 0.97 and 0.99, respectively. Aomori and Kagoshima also demonstrate strong performance, with pixel accuracy values of 0.98 and mIoU values of 0.93. These results verify that the SegFormer model effectively captures both global and local contextual features of environmental scenes.

G. PM2.5 Prediction

We have employed a backbone named EfficientNet-B2, a Convolutional Neural Network-based regression model, shown in Figure 5, for extracting visual features from the input data. The proposed approach utilizes dehazed images combined with segmentation masks as inputs, where a feature-level fusion strategy is applied. Specifically, the masks are integrated with the dehazed images through channel-wise concatenation to create a unified multi-channel input, enabling the model to simultaneously learn high-quality visual information and learn trained object-level cues derived from semantic segmentation.

These fused inputs are fed into the EfficientNet-B2 backbone to learn patterns and hierarchical visual features, ranging from low-level texture and color information to high-level semantic representations related to atmospheric conditions. The output features are generated by the

TABLE 1. EXPERIMENTAL RESULTS USING FIVE INPUT CONFIGURATIONS

Dataset	Model A			Model B			Model C			Model D			Model E		
	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*
Sapporo	0.16	1.95	1.72	0.29	6.21	1.92	0.01	6.70	1.81	0.22	2.37	1.81	0.18	4.93	1.70
Kagoshima	0.14	5.28	1.99	0.20	6.73	2.09	0.08	7.23	2.20	0.27	3.82	1.58	0.24	5.29	1.79
Aomori	0.19	8.69	2.49	0.26	7.82	2.32	0.04	7.47	2.16	0.43	7.39	2.40	0.31	8.34	2.32
Chofugaoka	-0.21	13.06	3.02	0.18	14.91	3.00	0.04	5.33	1.94	0.15	13.35	2.95	0.11	8.56	2.03

backbone and then are applied to a regression head designed to predict continuous PM2.5 values. The regression head comprised fully connected layers that map the extracted visual cues to a numerical output, enabling end-to-end learning of the relationship between the images and PM2.5 numerical data. For training, the regression head and the backbone network are optimized in an end-to-end manner using a Graphics Processing Unit (GPU). The dataset is split into training and testing sets with an 8:2 ratio. The model is trained utilizing a regression loss function, including Mean Squared Error (MSE), to minimize the difference between predicted and Ground Truth PM2.5 concentrations. An adaptive optimization algorithm named Adam optimizer is used to update the network parameters, with an appropriate learning rate of 0.0003, batch size of 32, and epochs of 30 to ensure convergence and improved prediction stability and generalization performance.

IV. RESULTS AND DISCUSSION

In this section, a comprehensive assessment of the proposed vision-based particulate matter prediction approach has been demonstrated, utilizing datasets from multiple cities in Japan. The experimental results are reported using five phases. Standard regression metrics, namely the coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE), are used to evaluate the model's performance. Additionally, we have conducted comparative analyses to examine the effect of the five stages on prediction accuracy. A detailed discussion of the quantitative results and their implications for PM2.5 prediction is provided in the following subsections.

A. Experimental results

In this study, we train a computer vision model, EfficientNet-B2, as a feature extractor along with the regression head on Japan's different cities datasets. The experiments are conducted using environmental images of various cities, including Sapporo, Kagoshima, Aomori, and Chofugaoka. The prediction results for each city are presented in Table 1 using five input configurations. Model A utilizes original environmental images as input to the regression model. Model B utilizes dehazed images to assess the effect of visual enhancement on prediction performance. Model C uses semantic segmentation masks derived from the original images to verify the contribution of object-level scene cues. Model D integrates dehazed images and their associated masks, indicating the proposed method that combines both improved visual features and semantic information. Finally, Model E uses the combination of all inputs.

As shown in Table 1, all alternative configurations have demonstrated varying changes in prediction accuracy across cities, compared to Model A. The dehazed images enhance R^2 values compared to the original images across all cities, demonstrating the effectiveness of visibility improvement in capturing pollution-related features. On the other hand, segmentation masks exhibit inconsistent performance, with generally lower correlation values, indicating that segmentation is insufficient solely for reliable prediction.

Model D, which integrates dehazed images and segmentation masks without original images, exhibits more consistent and competitive performance across multiple datasets. It achieves the strongest R^2 values, which are 0.27 and 0.43, respectively, in Kagoshima and Aomori, while also maintaining relatively lower MAE values of 1.58, particularly for Kagoshima. These results suggest that combining dehazed visual information with semantic features can significantly capture pollution-related patterns even in the absence of raw images.

In comparison, Model E, which incorporates original images along with dehazed images and segmentation masks, does not consistently outperform Model D. Although Model E achieves lower MAE in some cases, including 1.70 and 2.03 for Sapporo and Chofugaoka, respectively. Its R^2 values are lower than those of Model D in key datasets, including Kagoshima and Aomori. This suggests that the inclusion of original images does not necessarily lead to enhanced prediction.

Overall, the comparative analysis demonstrates that Model D provides a balanced and robust performance across cities.

B. Discussion

The experimental results show that employing dehazed images and semantic segmentation masks improves PM2.5 prediction performance across most cities, yielding a higher R^2 value and lower error metrics compared to the original image results. In addition, the performance of the regression model is influenced by dataset characteristics, including sample size, segmented predictors, weather conditions, camera resolution, and distance between the camera and station.

The performance of the Sapporo dataset achieves relatively stable performance across configurations. This can be attributed to the moderate dataset size of 72 samples, the extraction of semantic features, such as sky and building, consistent sunny weather conditions, and a comparatively short camera-station distance of 3.4 kilometers, which ensures reliable spatial alignment. The uniform camera resolution of

960 x 540 provides consistent feature extraction, further enabling enhanced prediction accuracy.

In Kagoshima, although the dataset contains 79 samples, the semantic information sky and mountain, appearance of sunny and cloudy weather, and a longer camera-station distance of 4.5 kilometers add additional variability. Despite this, Model D achieves the best performance, indicating that the integration of dehazing and semantic segmentation helps reduce the effects of environmental variation.

For Aomori, Model D obtains the highest R^2 value of 0.43, indicating a strong correlation between predicted and observed PM2.5 values. This indicates that the integration of dehazed images and segmentation masks effectively captures the overall pollution trends in this dataset. However, the error metrics MSE and MAE remain relatively higher compared to other cities, suggesting that although the model captures the general pattern well, prediction deviations still exist. This behavior may be attributed to snowy conditions, which introduce visual complexity and weak semantic feature cues, thereby limiting precise prediction. While the camera-to-station distance is 3.5 kilometers, which supports reasonable spatial alignment, the challenging environmental conditions limit accurate estimation.

For Chofugaoka, despite having a relatively larger dataset, which contains 93 samples and the shortest distance between the camera and station is 2 kilometers, the performance is comparatively lower. It indicates that factors, including scene complexity and variability in sunny and cloudy weather, and the presence of sky and water, have a stronger impact than spatial proximity alone. The higher error values specify that visual obscurity and weak feature correlations limit the model's effectiveness.

Model D demonstrates more consistent performance across cities, specifically under varying environmental conditions. Notably, it reaches competitive or superior results without utilizing original images. In contrast, Model E does not consistently enhance performance, suggesting that raw images may introduce redundancy or noise rather than useful information.

V. CONCLUSION AND FUTURE WORK

The aim of this study is to develop an image-based system framework for predicting particulate matter concentrations using geographical images collected from various geographical locations of Japan. The urban-level datasets are constructed by pairing outdoor city images with the associated Ground Truth PM2.5 observations obtained from the cities' monitoring stations. We have employed image dehazing techniques to enhance the reliability of visual information under varying atmospheric conditions for reducing haze effects. Semantic segmentation is proposed to extract meaningful object-level characteristics from the urban scenes. These proposed pre-processing stages aim to emphasize visual features, such as sky conditions, buildings, and water, that are closely related to air pollution, thus improving the input data quality for prediction.

The proposed system integrates two complementary inputs, such as dehazed images and object segmentation masks, along with PM2.5 measurements, which are fed into a regression method and trained to predict PM2.5 concentrations. Experimental results demonstrate satisfactory performance in terms of using dehazed images and segmentation masks in vision-based air quality estimation compared to the other input configurations.

In future work, we would like to apply our proposed method to other existing PM2.5 vision datasets to evaluate its performance under diverse environmental conditions.

ACKNOWLEDGMENT

We would like to convey our special thanks to Swannack Raymond Amaki for providing continuous support in the data collection process.

REFERENCES

[1] T. Ohara and M. Ono, "An Overview of PM2.5 Pollution Research Conducted in ERTDF Projects since 2011," Global Environmental Research, vol. 22, pp. 3–12, Dec. 2018.

[2] D. Mitra and A. Saha, "IoT-Based Air Pollution Detection, Monitoring and Controlling System," Journal of Discrete Mathematical Sciences and Cryptography, vol. 25, pp. 2173–2182, Dec. 2022, doi: 10.1080/09720529.2022.2133254.

[3] K. Zhang, Z. Chen, and Y. Xiang, "Vision-Based Particulate Matter Estimation," Deep Learning Applications, pp. 3–17, 2023, doi: 10.1142/9789811266911_0001.

[4] G. Wang et al., "Vision-Based PM2.5 Concentration Estimation with Natural Scene Statistical Analysis," IEEE Transactions on Artificial Intelligence, vol. 5, pp. 2805–2815, Oct. 2023, doi: 10.1109/TAI.2023.3324892.

[5] S. Laohakiat, S. Klerkkidakan, and N. Wiwatwattana, "Visually Estimating and Forecasting PM2.5 Levels Using Hybrid Architecture Deep Neural Network," Current Applied Science and Technology, vol. 24, p. e0258074, Dec. 2023, doi: 1055003/cast.2023.258074.

[6] Y. Liu et al., "Applying Traffic Camera and Deep Learning-Based Image Analysis to Predict PM2.5 Concentrations," Science of The Total Environment, vol. 912, p. 169233, Dec. 2024, doi: 10.1016/j.scitotenv.2023.169233.

[7] 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, pp. 95–113, Dec. 2024, doi: 10.3390/s25010095.

[8] N. Sirisumpun, K. Wongwailikhit, P. Painmanakul, and P. Vateekul, "Spatio-Temporal PM2.5 Forecasting in Thailand Using Encoder-Decoder Networks," IEEE Access, vol. 11, pp. 69601–69613, Jul. 2023, doi: 10.1109/ACCESS.2023.3293398.

[9] T. Zhao and M. Qu, "VDMS: An Improved Vision Transformer-based Model for PM2.5 Concentration Prediction," Applied Sciences, vol. 15, pp. 7346–7362, Jun. 2025, doi: 10.3390/app15137346.

[10] 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, pp. 4053–4076, Jun. 2025, doi: 10.3390/s25134053.

[11] Z. Fan et al., "Enhancing Urban Real-Time PM2.5 Monitoring in Street Canyons by Machine Learning and Computer Vision Technology," Sustainable Cities and Society, vol. 100, p. 105009, Jan. 2024, doi: 10.1016/j.scs.2023.105009.

[12] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, pp. 2341–2353, Dec. 2011, doi: 10.1109/TPAMI.2010.168.