Advancing Environmental Applications through Machine Learning and Computer Vision: Modeling, Algorithms, and Real-World Implementations

by

Tony Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Robert P. Dick, Chair
Professor Stuart Batterman
Professor Mingyan Liu
Professor Qin Lv

Tony Zhang
ttzhan@umich.edu
ORCID iD: 0000-0003-3755-3349

ACKNOWLEDGMENTS

This thesis was completed under the advice and guidance from my advisor, Professor Robert P. Dick. I want to thank Professor Dick for the continuous support he provided during my doctoral studies. I am grateful for the opportunity I had to collaborate with him. I appreciate all his contributions to make my doctoral experience productive.

Besides my advisor, I would like to thank the members of my dissertation committee, Professor Stuart Batterman, Professor Mingyan Liu, and Professor Qin Lv for their time and feedback. Your suggestions greatly improved various aspects of this thesis.

I also like to acknowledge my labmates for their advice and support, as well as other collaborators at other departments and institutes, including the University of Colorado Boulder and Zhejiang University at Hangzhou, China. Professor Yun Xiang from Zhejiang University have provided unique insights and great suggestions. Prof. Qin Lv, Prof. Michael Hannigan, and Prof. Daven Henze, all from University of Colorado Boulder, were collaborators that helped me in my first year as a doctoral student. Therefore, I want to thank them for their weekly inputs and discussions. I feel grateful for the opportunity to work with them. In particular, my research was supported, in part, by the University of Michigan and National Science Foundation under grants CBET-1240584, CC-1836230, and CNS-2008151.

Moreover, it is an honor to attend the University of Michigan, which is a world-class institution. The Electrical and Computer Engineering department here is one of the top engineering programs in the world. I also like to recognize the people I met at the University of Michigan, and I appreciate that I got to interact with bright and knowledgeable individuals. Without them, my research path would be more arduous.

Finally, I thank my family for their continued and unwavering support. My parents have always encouraged my pursuits and I would not have completed this journey without them. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The escalating concern over environmental challenges has spurred a growing interest in harnessing machine learning and computer vision techniques to represent scenes in environmental applications. Accurate and efficient scene representations play a pivotal role in addressing environmental issues, including air pollution, fire detection, and remote sensing analysis. This dissertation delves into the field of scene representations in machine learning and computer vision, with a specific focus on image-based approaches for environmental applications.

For vision-based air pollution applications, air quality can be estimated by observing haze effects in images; hence, digital cameras can be used to quantify pollutants across large areas. We propose to use vision-based air pollution algorithms to predict the level of air pollution within the environment. The prevalence of images suggests that images can be used to estimate high spatial resolution air pollutant concentrations. However, there are many challenges to develop a portable, inexpensive, and accurate method for pollutant analysis, such as image quality variability, sufficient data for training, and hardware and software optimizations to meet constraints.

I address those challenges by designing image-based air pollution prediction methods for sensing and forecasting, developing benchmark datasets to test and validate vision-based pollution estimation algorithms, and determining how sensing accuracy depends on point sensor density and use of cameras. My efforts can be divided into three categories: (1) We design an image-based multi-pollutant estimation algorithm that is capable of modeling atmospheric absorption in addition to scattering, spatial variation, and color dependence of pollution; (2) We use different spatial densities of sensors and vision-based algorithms to estimate air pollution concentrations and analyze hazy images; (3) We construct an image-based air quality forecasting model that fuses a history of $PM_{2.5}$ measurements with colocated images (at the same spot); and (4) We develop an image-based air quality prediction model specifically tailored to the nighttime case.

All the techniques are evaluated and validated using real-world data. Experimental results show that our techniques can reduce sensing error significantly. For example, our multi-pollutant estimation technique reduces single-pollutant estimation RMSE (root mean square error) by 22% compared to previous existing vision-based techniques; for the im-

ages in our benchmarking dataset, using images decreases MAE (mean absolute error) by 8.4% on average; therefore, adding a camera to collect images helps more than adding more sensors. Finally, experiments on Shanghai data show that our forecasting model improves $PM_{2.5}$ prediction accuracy by 15.8% in RMSE and 10.9% in MAE compared to previous forecasting methods.

Furthermore, two innovative deep learning models were introduced to address segmentation tasks in different environmental domains. The first model focused on fire segmentation in images, incorporating a multi-scale aggregation module and a context-oriented module to achieve accurate and rapid fire detection by extracting discriminative features from various receptive fields and capturing both local and global context information. The proposed fire segmentation network outperformed previous methods with a significant 2.7% improvement in Intersection over Union (IoU). The second model targeted remote sensing segmentation in aerial images, enhancing feature representation in the spatial and frequency domains through a Frequency Weighted Module and a Spatial Weighting Module, respectively. Additionally, a Multi-Domain Fusion Module was employed to combine features from different domains, leading to state-of-the-art performance on remote sensing datasets with a mean F1-score accuracy improvement of 1.9%.

CHAPTER 1

Introduction

The increasing concern over environmental issues has led to a growing interest in leveraging machine learning and computer vision techniques for scene representations in the context of environmental applications. Accurate and efficient scene representations play a crucial role in understanding and addressing environmental challenges, such as air pollution, fire detection, and remote sensing analysis. This dissertation aims to explore and advance the field of scene representations in machine learning and computer vision, focusing on image-based approaches for environmental applications.

A large part of this dissertation focuses on image-based air pollution applications. In particular, image-based air pollution prediction estimates air quality from images through visibility and haze level analysis. Existing research in visibility physics demonstrates significant correlation between visibility and $PM_{2.5}$ levels [4–6]. Scene visibility is affected by atmospheric particles due to the scattering and absorption of light and by extension, the level of air pollution at the time [7]. When light travels through the atmosphere, it encounters atmospheric particles and gases that affect its path. Light attenuation is caused by both scattering and absorption. Scattering occurs when particles cause the light to change its direction of travel. Absorption occurs when part of the light disappears due to the transfer of energy to the particles.

In addition to air quality sensors, image sensors can be used to estimate PM concentrations, potentially at higher spatial resolutions [7, 8]. In cities, using images for $PM_{2.5}$ estimation will typically increase field estimation accuracy by 14.3% compared to only using point sensors such as particle sensors [9]. The increasing popularization of smartphones and webcams makes such applications possible. Particle counters are more expensive than webcams and generally require more maintenance. In contrast, consumer-grade cameras can gather pollution data over wide fields of view. Moreover, image-based methods are drift-resistant, which reduces maintenance costs. While some work I did relied on a fixed, known camera location, some related ideas would also work with smartphone images.

Figure 1.1: Varying air quality in images in Shanghai: clear (left), medium haze (middle), and heavy haze (right) [1].

Adding image sensors with high spatial resolution can increase field estimation accuracy and spatial resolution. Visibility, in the context of air quality and the environment, refers to the degree to which objects and landmarks can be seen and identified at a given location and time [10]. It is a crucial parameter used to assess atmospheric clarity and the presence of air pollutants or particles that may obscure vision. Visibility is high on low air pollution days, but visibility is low during high pollution because light is scattered away from the camera by PM. This is exhibited in Figure 1.1. Images have the ability to capture complex relationships between air quality and the factors influencing it. Additionally, images can be useful in various applications in air pollution, such as monitoring and forecasting of visible pollutants including PM and NO_x.

Past research has devised an atmospheric model describing an image influenced by haze as follows [11]:

$$\boldsymbol{I}(x) = \boldsymbol{J}(x)t(x) + \boldsymbol{A}(1 - t(x)), \tag{1.1}$$

where x is the pixel location, $\boldsymbol{I}$ is the observed image, $\boldsymbol{J}$ is the scene radiance (image without any haze), $\boldsymbol{A}$ is the atmospheric light (also known as the airlight), and t is the transmission function. Additionally, transmission can be expressed as follows:

$$t(x) = e^{-\beta d(x)}, \tag{1.2}$$

where t is the transmission, β is the scattering coefficient of the atmosphere, and d is the depth from the camera to the pixel object.

The model relies on several assumptions about atmospheric haze: 1) light attenuation is influenced by only scattering; 2) the properties of light attenuation are color-independent; and 3) the density of haze is uniform in the atmosphere [11]. Past research has used the transmission from the haze model as a feature to estimate air quality from images, which have been shown to be effective [12]. Transmission can be used to describe the attenuation of scene radiance [11]. To calculate the transmission, the dark channel prior was proposed

and widely used, which relies on some pixels in an image having zero or very low intensity for at least one color channel [13].

Before image-based air pollution sensing can be used in real-world applications, there are still many challenges to overcome. It requires complex signal processing and machine learning algorithms. However, it holds the promise of a portable, inexpensive, and accurate method for pollutant analysis in urban and industrial areas. This dissertation will demonstrate that commodity cameras and smartphones can be widely used as an accurate tool for real-time monitoring of air pollution levels.

This dissertation will describe frameworks for image-based air pollution prediction and explore features that will potentially affect the performance of models. I will validate the concept of image-based air pollution prediction through real-world experiments. We have demonstrated that using atmospheric modeling, data fusion techniques, and image processing algorithms can enhance image-based estimation and forecasting algorithms, making them more practical in real-world applications.

Moreover, we build upon vision-based air quality estimation to develop effective fire detection systems using deep learning principles. By adaptively incorporating information from multiple levels of the CNN and enhancing the receptive field network, we can utilize consumer cameras for real-time fire detection and monitoring. This advancement aids in early fire detection, prompt response measures, and minimizing the impact of wildfires on the environment and human lives. Additionally, we apply expertise from vision-based air quality algorithms to remote sensing segmentation tasks, enabling the processing of satellite or aerial images for environmental monitoring, land management, urban planning, and disaster response efforts. This segmentation analysis empowers decision-makers with critical information for informed actions.

1.1 Single-Point Image-Based Estimation

I developed an image-based pollution estimation technique that uses wavelength-dependent scattering and absorption properties to enable analysis of multi-pollutant systems and improve estimation accuracy. In addition, I use the position-dependent properties of light attenuation within images to improve prediction accuracy by accounting for nonuniform pollution distribution. Given training and testing images and training ground-truth pollution concentrations, the objective is to minimize mean squared error of pollution concentration estimates over time.

It is important to estimate pollutant concentrations from images using accurate models of pollutants in images. Previous atmospheric imaging models consider the scattering

properties of haze, but not the space-and color-dependent properties of scattering and absorption. The model from 1.1 approximately represents haze in realistic conditions, but relaxing the model's assumptions results in a 34.6% improvement in $PM_{2.5}$ level prediction.

I am the first to use images to estimate pollutant concentrations in systems with multiple pollutants. I achieve this by considering the differences in scattering and absorption spectra between different pollutants. My system improves the accuracy of $PM_{2.5}$, PM_{10}, and NO_2 estimation by 29.9% for single-scene images in Shanghai compared to existing image-based techniques.

1.2 Novel Dataset for Image-Based Estimation

Ground truth pollution data are typically obtained from the nearest of several sparsely deployed monitoring stations. Lack of high-resolution ground truth pollution measurements makes it difficult to evaluate vision-based pollution estimation techniques, as pollution can vary rapidly with position [14]. Moreover, vision-based pollution estimation techniques generally assume homogeneous distributions of particles and gases within images, implying constant light attenuation [14]. In reality, pollution concentration changes rapidly in space. Therefore, accurate evaluation of vision-based estimation algorithms requires high-resolution datasets [15].

We (including collaborators in Hangzhou) develop a novel dataset with the goal of validating vision-based pollution estimation algorithms. The dataset contains images and corresponding "ground truth" pollution concentration values in images. In particular, pollution concentration varies over time and in the images. The novel contribution is a densely distributed, low-cost $PM_{2.5}$ and PM_{10} dataset with high temporal and spatial resolution enabling scientific discovery and the evaluation of air pollutant estimation algorithms. The dataset also includes images that capture the distributed area of the sensors. This dataset, the first of its kind, is significant because air quality estimation substantially improves using both data from existing sensors and image-based analysis of air pollution.

My main contribution to this dataset work was performing extensive analysis on how the estimation of $PM_{2.5}$ depends on point sensor density and absence/presence of cameras. My main findings show that using the images in our benchmarking dataset decreases MAE by 8.4% on average; hence adding a camera to collect images helps more than adding more sensors. Additionally, I find that pollutant concentrations are spatially correlated; spatial variation of $PM_{2.5}$ is high; and temperature and humidity had limited correlation with PM concentration in our dataset.

1.3 Nighttime Pollutant Estimation

Collaborators in Hangzhou and I introduce a novel vision-based technique for nighttime $PM_{2.5}$ concentration estimation. Specifically, we first derive a glow map using the image brightness and transmission. Then, we design a deep convolutional neural network algorithm to estimate the $PM_{2.5}$ concentration quantitatively. Finally, we evaluate our methods using real-world data and images. Experimental results demonstrate that our proposed method can achieve an improvement in accuracy by 29.3% compared to that of the daytime method. To the best of our knowledge, this is the first work to measure nighttime $PM_{2.5}$ concentration using a vision-based method.

1.4 Image-Based Air Quality Forecasting

The problem of air quality forecasting is important but also challenging because air quality is affected by a diverse set of complex factors. I construct the first image-based air quality forecasting model. It fuses a history of $PM_{2.5}$ measurements with colocated images. Past research showed that images have the ability to inform on air quality in a region over time. I construct a multi-level attention-based recurrent network that uses images and $PM_{2.5}$ data to represent variation over space and time. Experiments on Shanghai data show that the forecasting model improves $PM_{2.5}$ prediction accuracy by 15.8% in RMSE and 10.9% in MAE compared to previous forecasting methods. In addition, I evaluate the impact of each model component via ablation studies.

1.5 Fire Segmentation

Accurate and rapid detection of fire is useful for environmental protection and public safety. The problem of fire segmentation in images is difficult because images contain fire with different kinds of shapes, sizes, illumination, and backgrounds. I construct a novel fire segmentation model, which utilizes multi-scale aggregation as well as global scene information through a context-oriented module. For example, the multi-scale aggregation module reconstructs the segmentation using features from multiple receptive fields. Also, the context-oriented module obtains local and global context information to expand the receptive field and extract more discriminative features. Using our fire segmentation network improves accuracy by 2.7% in IoU compared to previous methods.

1.6 Remote Sensing Segmentation

For the problem of remote sensing segmentation, we introduced a new deep learning model for segmenting aerial images. Our model improves feature representation in both the spatial and frequency domains, preserving important details and textures to enhance feature learning across different frequency scales. We incorporated a Frequency Weighted Module and a Spatial Weighting Module to capture contextual information in the frequency and spatial domains, respectively. Additionally, we developed a Multi-Domain Fusion Module to combine features from different domains, providing valuable complementary information.

Our proposed model outperformed previous methods on various remote sensing datasets, achieving state-of-the-art performance. It improved the mean F1-score accuracy by 1.9% compared to existing techniques. We conducted ablation studies to validate the effectiveness of each component of our model. Our approach shows promise for enhancing a range of remote sensing applications, such as vegetation classification, urban structure detection, and crop monitoring.

1.7 Dissertation Organization

This dissertation is organized as follows.

1. Chapter II describes our image-based method that estimates multiple pollution concentrations from images using scattering and absorption properties. Our wavelength-sensitive, absorption and spatial variation aware multi-pollutant vison-based estimation technique improves accuracy by 29.9% in Shanghai.

2. Chapter III discusses a publicly released novel dataset appropriate for evaluating vision-based pollution estimation algorithms. We determined how accuracy depends on point sensor density and absence/presence of cameras. For the images in our benchmarking dataset, using images decreases MAE by 8.4% on average; hence adding a camera to collect images helps more than adding more sensors.

3. Chapter IV introduces a novel vision-based approach for estimating nighttime $PM_{2.5}$ concentration. Our method involves deriving a glow map based on image brightness and transmission, followed by the design of a deep convolutional neural network algorithm for quantitative estimation.

4. Chapter V explains an image-based forecasting model of future $PM_{2.5}$ concentrations by using a multi-level attention-based recurrent network. Experiments on Shanghai

data show that our forecasting model improves $PM_{2.5}$ prediction accuracy by 15.8% in RMSE and 10.9% in MAE compared to previous forecasting methods.

5. Chapter VI details an image-based fire segmentation model, which utilizes multi-scale aggregation as well as global scene information through a context-oriented module. Using the proposed fire segmentation network improves accuracy by 2.7% in IoU compared to previous methods.

6. Chapter VII introduces a novel deep learning model for aerial image segmentation, enhancing feature representation in both spatial and frequency domains. The proposed model achieved state-of-the-art performance on remote sensing datasets, improving accuracy by 1.9% in the mean F1-score compared to previous methods.

7. Chapter VIII concludes the dissertation.

CHAPTER 2

Estimation of Multiple Atmospheric Pollutants through Image Analysis

2.1 Introduction

Major air pollutants such as $PM_{2.5}$, PM_{10}, and NO_2 degrade the air quality in many areas of the world. The degradation of air quality in visual scenes is caused by two phenomena of light, which are scattering and absorption. The level of scattering and absorption in the atmosphere varies with respect to individual pollutants and their concentrations such as $PM_{2.5}$, PM_{10}, and NO_2. This holds for some important but not all pollutant types. In polluted air, the intensity and color of light is affected by its interactions with the atmosphere.

Prior research used scattering principles of haze to estimate air quality from images [14, 16]. However, representing atmospheric haze in realistic scenarios is complex and has always been a difficult problem. Nonetheless, it is very important for rendering realistic outdoor scenes and predicting air quality.

Haze is a kind of aerosol which consists of small particles suspended in the atmosphere [11]. Haze has various sources, including combustion material and volcanic ashes. Air molecules are smaller than haze particles, but fog and cloud droplets are bigger than haze particles [11]. In particular, haze affects the visibility of objects when present and its color is usually either gray or bluish.

Computer vision systems typically assume that image scenes are not occluded due to weather conditions such as haze and fog. In clear scenes, objects are immersed in a transparent medium and light rays are reflected by objects and travel without attenuation. However, computer vision systems need to also deal with haze scenarios. Moreover, it is important to model haze more accurately, and this has implications on multi-pollutant air quality estimation.

This chapter presents a method that estimates air quality from images using properties derived from the principles of visibility physics. Our work makes the following main

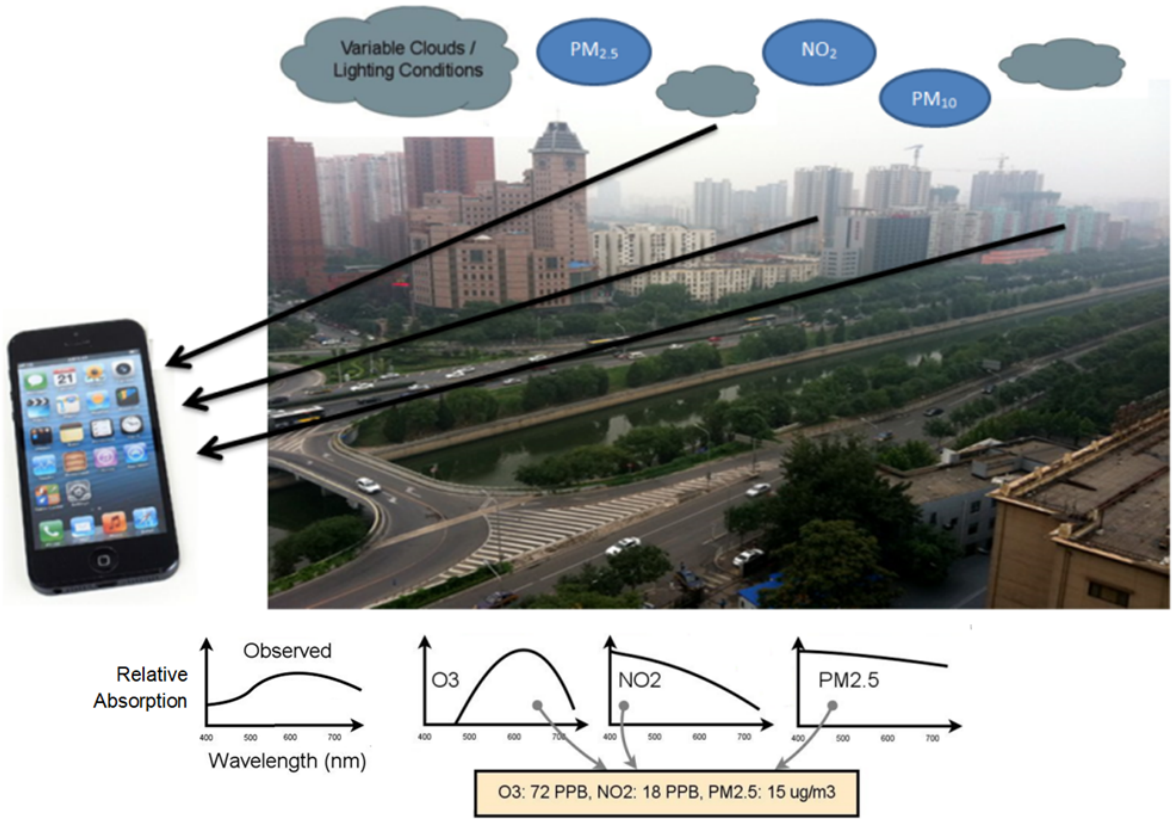


Figure 2.1: I consider the differences in scattering and absorption spectra between different pollutants in RGB color space. As an example, for components of $PM_{2.5}$ and PM_{10} that are smaller than the wavelength of light, relative scattering is inversely proportional to wavelength. This enables the estimation of concentrations of multiple simultaneous pollutants.

contributions.

1. This work is the first to develop a system that estimates the concentrations of multiple pollutants from images, namely $PM_{2.5}$, PM_{10}, and NO_2.

2. We formulate and solve the multi-pollutant estimation problem by using the position- and color-dependent properties of pollutant-specific scattering and absorption. I achieve multi-pollutant estimation from images by considering the differences in scattering and absorption spectra across different pollutants.

3. Our system improves the accuracy of $PM_{2.5}$, PM_{10}, and NO_2 estimation by 29.9% for a Shanghai dataset compared to previous techniques for estimating air quality via images.

The rest of this chapter is organized as follows. Section 2.2 discusses background material regarding visibility physics principles. Section 4.4 describes the multi-pollutant esti-

mation system using novel haze modeling techniques. Section 2.5 describes the evaluation results, and Section 4.7 concludes the chapter.

2.2 Background

Previous research on estimating air quality from images and on dehazing images uses an atmospheric model to describe an image influenced by haze [17, 18], as described in the introduction. We summarize the effect of both light attenuation and airlight on atmospheric scattering, as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)) \text{ and} \tag{2.1}$$

$$t(x) = e^{-\beta d(x)}. \tag{2.2}$$

In clear weather, images appear vivid because objects reflect the energy from the illumination source, and little illumination is lost in the photo. For days affected by air pollution, there are two mechanisms that influence images: the direct attenuation and the airlight. First, the direct attenuation causes the intensity of the pixels in the image to decrease in a multiplicative manner. This is because the $J(x)t(x)$ part of the equation reflects the direct attenuation.

Additionally, the atmospheric light (also known as the airlight) is caused by light scattering from particles and gases in the atmosphere. The term $A(1 - t(x))$ in the haze model equation represents the effect of the airlight, and the effect of the airlight on the light intensities is additive. It shifts the color of the scene radiance towards the aggregate color of the particles and gases in the atmosphere. The effect of the airlight increases as more particles and gases in the atmosphere increase.

The haze model makes multiple assumptions about atmospheric pollution. Existing work used the haze model for single image dehazing using a dark channel defined as the darkest pixels within the localized patches [19]. A color attenuation prior was used to model a hazy scene in HSV (hue, saturation, and value) color space [20]. Finally, another dehazing method based on the CNN (convolutional neural network) jointly estimates the transmission map and the atmospheric light from the haze model [21].

Furthermore, prior research used the haze model to estimating air quality from images [14, 16]. Liu et al. used support vector regression (SVR) to estimate $PM_{2.5}$ concentration based on image features from the haze model [14], and a deep network extract the level of haze in images through the haze model [16]. However, the existing haze model makes assumptions that do not hold, and I provide a more accurate atmospheric model.

2.2.1 Visibility Physics

Light attenuation is influenced by both scattering and absorption, but past work ignores absorption [22,23]. For instance, elemental carbon is a major contributor of absorption and accounts for 20-30% of total light absorption [24]. Hence, there should be two coefficients for β, β_s for scattering and β_a for absorption. Light attenuation also varies by location; the two coefficients should be β_s^x for scattering and β_a^x for absorption where x represents the location. Air pollution demonstrates spatial variation, as pollutant concentrations are expected to vary within field of view. We use the most updated coefficients in the model which is more accurate, and I model spatial variation of pollution to improve accuracy.

Furthermore, past work on image-based estimation assumes pollution is gray. However, the properties of light attenuation are wavelength-dependent. For example, the absorption coefficient of aerosols, especially black carbon (BC), vary depending on wavelength [25]. This assumption makes the simultaneous estimation of multiple pollutant concentrations impossible when they are mixed together. A full wavelength spectrum would be ideal; RGB suffices. We can estimate concentrations of multiple pollutants simultaneously by considering wavelength-dependent optical properties of pollutants. In particular, β_s^x, β_a^x, and A (airlight) are RGB vectors.

The processes that contribute to visibility in the atmosphere are wavelength-dependent. Two kinds of light scattering take place: Mie and Rayleigh scattering. Mie scattering occurs when the size of atmospheric particles is at least the wavelength of light; its effect on visibility is wavelength-independent. Rayleigh scattering occurs when the particles are much smaller than the wavelength of light (i.e., nitrogen and oxygen) and is wavelength-dependent.

Visibility is also decreased by absorption [22, 23]; elemental carbon (black carbon) and organic carbon (brown carbon) are two main causes of absorption in $PM_{2.5}$ and PM_{10}. Specifically, elemental carbon accounts for 20-30% of total light absorption [24] [26], and organic carbon is also a major contributor [27]. Their relative absorptions are inversely proportional to wavelength [28] [29] [30] [31]. Also, NO_2 absorbs blue light heavily.

The atmosphere model used in prior work also assumes that the attenuation coefficient, which is related to pollution concentration, is constant for an entire image. In realistic conditions, the density of particles and gases changes as a function of position and altitude, leading to a non-uniform light attenuation coefficient [32] [33]. We explicitly consider this effect. In summary, the properties of light attenuation are color-dependent, the light attenuation coefficient is influenced by scattering and absorption, and the atmosphere is non-homogeneous.

Past research assumed that the only cause of reduced visibility in images is scattering.

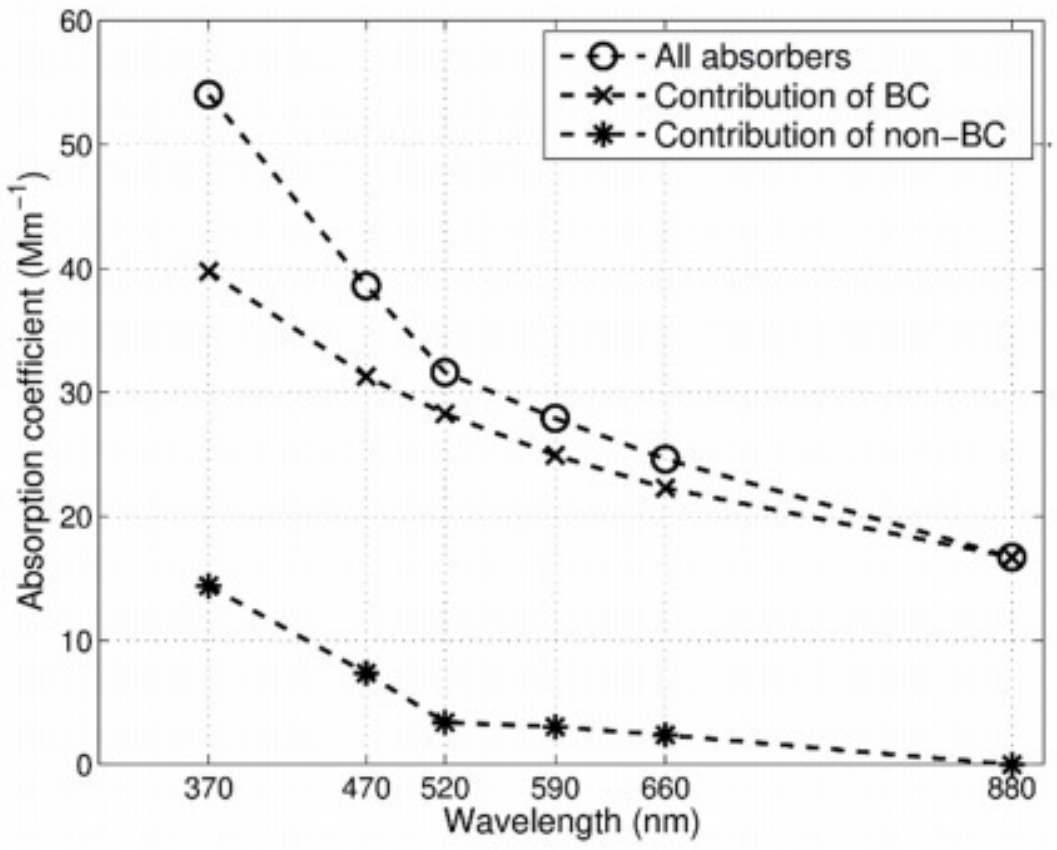


Figure 2.2: The absorption coefficient of aerosols, especially black carbon (BC), vary depending on wavelength [2].

Even though light extinction is mainly caused by scattering, absorption also reduces visibility. We extend the haze model from Eq. (1) and (2) to account for both scattering and absorption, as follows:

$$I_c(x) = J_c(x)t_{1_c}(x) + A_c(1 - t_{2_c}(x)), \tag{2.3}$$

$$t_{1_c}(x) = e^{-(\beta_{s_c}+\beta_{a_c})d(x)}, \text{ and} \tag{2.4}$$

$$t_{2_c}(x) = e^{-\beta_{s_c}d(x)}. \tag{2.5}$$

In Eq. (3), we incorporate t_1, the transmission, as a function of both scattering and absorption, and t_2, the transmission, as a function of only scattering. In Eq. (4) and Eq. (5), β_s corresponds to the scattering coefficient of the atmosphere and β_a corresponds to the absorption coefficient. Additionally, every variable in Eq. (3) to (5) has a subscript c to demonstrate color-dependent light interactions. For this model, two mechanisms influence images: the direct attenuation and the airlight. The direct attenuation is caused by both

scattering and absorption and corresponds to the first term $J_c(x)t_{1_c}(x)$. The second term, airlight, is estimated from the image. Although $A_c(1 - t_{2_c}(x))$ explicitly considers only scattering, the measured airlight color implicitly considers absorption.

2.3 Related Work

In this section, we review related work on image dehazing and hazy image datasets.

2.3.1 Image Dehazing Algorithms

Image dehazing has attracted a lot of attention, and many image dehazing algorithms have been developed in the last decade. Some haze removal techniques use the formulation in Eq. 1 and 2 to estimate the transmission map and airlight. Some early dehazing methods require the depth information either as an input or from 3D models [34, 35]. Other approaches estimate the haze using multiple images of the same scene with different polarization properties [36, 37].

Afterward, the accuracy of image dehazing techniques increased by using assumptions or priors to estimate the haze level. He et al. use a dark channel, defined as the darkest pixels within the localized patches, to estimate the transmission and airlight from the haze imaging model [19]. They assume that the values of the dark channel in clear images are close to zero, but the prior does not work when objects in the original scene have colors similar to the airlight.

Various prior-based methods work very well when the underlying assumption is satisfied, but may fail otherwise [38, 39]. Zhu et al. use a color attenuation prior to model the scene depth through saturation and value [20]. However, Zhu et al. assume that the level of scattering is constant across the entire image. Berman et al. assume that clear images contain only a finite number of color clusters, and each color cluster corresponds to a haze-line in RGB space as a function of distance [40]. The method may not work where the airlight is significantly brighter than the original scene radiance.

In the last few years, dehazing techniques utilize deep learning to estimate the amount of haze. Cai et al. [41] and Ren et al. [42] use a convolution neural network (CNN) to estimate the transmission map and airlight separately. However, estimating transmission and airlight separately will introduce errors for each independent estimation. Moreover, combining transmission and airlight to construct the dehazed image amplify those errors further.

While some deep learning methods estimate the transmission and airlight separately,

other methods use an end-to-end model to directly estimate the clear image without having to estimate other parameters. Li et al. constructed a CNN that jointly estimates transmission and airlight to form the clear image [21]. Finally, dehazing methods incorporate an end-to-end Generative Adversarial Network (GAN) framework to achieve increased accuracy [43–45].

2.3.2 Existing Image Datasets and Benchmarking

Hazy image datasets are used to evaluate the effectiveness of dehazing and other computer vision algorithms. FRIDA (Foggy Road Image Database) is used to evaluate automatic driving systems in hazy environments [46]. Sakaridis et al. [47] applied their fog synthesis method on the Cityscapes dataset [48] to construct the Foggy Cityscapes dataset.

Past research has also constructed datasets with synthetic hazy images using Eq. 1 and 2. They used the indoor NYU2 Depth Database [49] and the Middlebury stereo database [50], which contain clear images and the corresponding depth meta-data. Hazy images are also synthesized from outdoor images obtained from websites, where the depth map is estimated from past work in monocular depth estimation. However, Eq. 1 and 2 neglects various haze properties such as absorption, color-dependent haze, and region-dependent haze.

Recently, a realistic haze dataset O-HAZE was introduced, which contains 45 pairs of realistic hazy and haze-free ground-truth images for the same scene [51]. The haze originated from a haze machine. O-HAZE, however, does not contain enough pairs for training and validating machine learning approaches. For example, AOD-Net [21] uses over 27,000 images for training and almost 4,000 images for evaluation. Additionally, it is very likely that the spatial distribution of haze in O-HAZE is not realistic since the haze was distributed manually. The ground truth haze concentration distribution was not given. On the contrary, our method considers absorption, color-dependent haze, and region-dependent haze.

Prior work uses hazy datasets to benchmark image dehazing methods. Li et al. [52] and Ancuti et al. [53] evaluate dehazing algorithms using a synthetic hazy image dataset from Eq. 1 and 2. On the other hand, Ancuti et al. [53] evaluate dehazing algorithms using O-HAZE [51]. The haze in the O-HAZE dataset may not reflect the spatial variation of haze in natural scenes influenced by weather or pollution because all hazy images were produced using the same haze machine and chemical process.

2.4 Methodology

Our technique consists of two main steps: obtaining the transmissivities of scattering and absorption for all three color channels based on Eq. (3) to (5), and obtaining predicted concentrations for $PM_{2.5}$, PM_{10}, and NO_2 based on the transmissivities from the prior step.

2.4.1 Obtaining Transmissivities

We used the atmospheric model described in Eq. (3) to (5) to obtain β_{s_c} and β_{a_c} for all c. $I(x)$ is the input image. Using the webcam image dataset and ground truth pollutants, we obtain $J(x)$ by collecting the images with the lowest PM_{10} concentrations and taking the mean of their color intensities so $J(x)$ contains as little air pollution as possible (typically about 5% of the maximum). The depth map is obtained by running a convolutional neural network by Li et al. [54] on $J(x)$. The airlight is estimated using the technique in Berman et al. [55]. Afterward, the only unknown variables in Eq. (3) are β_{s_c} and β_{a_c}.

Algorithm 1 Gradient Descent

Input: $I(x)$, $J(x)$, $d(x)$, A, height, width
Output: β_s, β_a
while $\mid \beta_a - \beta'_a \mid > \gamma$ or $\mid \beta_s - \beta'_s \mid > \sigma$ **do**
- $\beta'_a = \beta_a$, $\beta'_s = \beta_s$
- $\hat{I}(x) = J(x)e^{-(\beta_s+\beta_a)d(x)} + A(1 - e^{-\beta_s d(x)})$
- $\epsilon(x) = (\hat{I}(x) - I(x))/$ (height $\times$ width)
- $C(x) = \frac{1}{2} \times (\hat{I}(x) - I(x))^2$
- $\frac{dC(x)}{d\beta_s} = \epsilon(x)d(x)e^{-\beta_s d(x)}(A - J(x)e^{-\beta_a d(x)})$
- $\frac{dC(x)}{d\beta_a} = -\epsilon(x)d(x)J(x)e^{-(\beta_s+\beta_a)d(x)}$
- $\beta_s = \beta_s - \alpha \times \sum_x \frac{dC(x)}{d\beta_s}$
- $\beta_a = \beta_a - \alpha \times \sum_x \frac{dC(x)}{d\beta_a}$

end

We use gradient descent to find β_s and β_a by minimizing the cost function $C(x) = \frac{1}{2}(\hat{I}(x) - I(x))^2$, where $\hat{I}(x)$ is the predicted image calculated in Eq. (3) and $I(x)$ is the actual image. The gradient descent algorithm is shown in Algorithm 1. Gradient descent is computationally efficient and produces a stable error gradient and a stable convergence. To improve convergence, the algorithm keeps track of the last ten calculated β_{s_c} and β_{a_c}. If the past beta values are stable, the step size (i.e., learning rate or α) decreases in the last two expression in Algorithm 1.

In realistic conditions, the distributions of particles and gases are not uniform and change as a function of position and altitude. Light attenuation in a single image varies

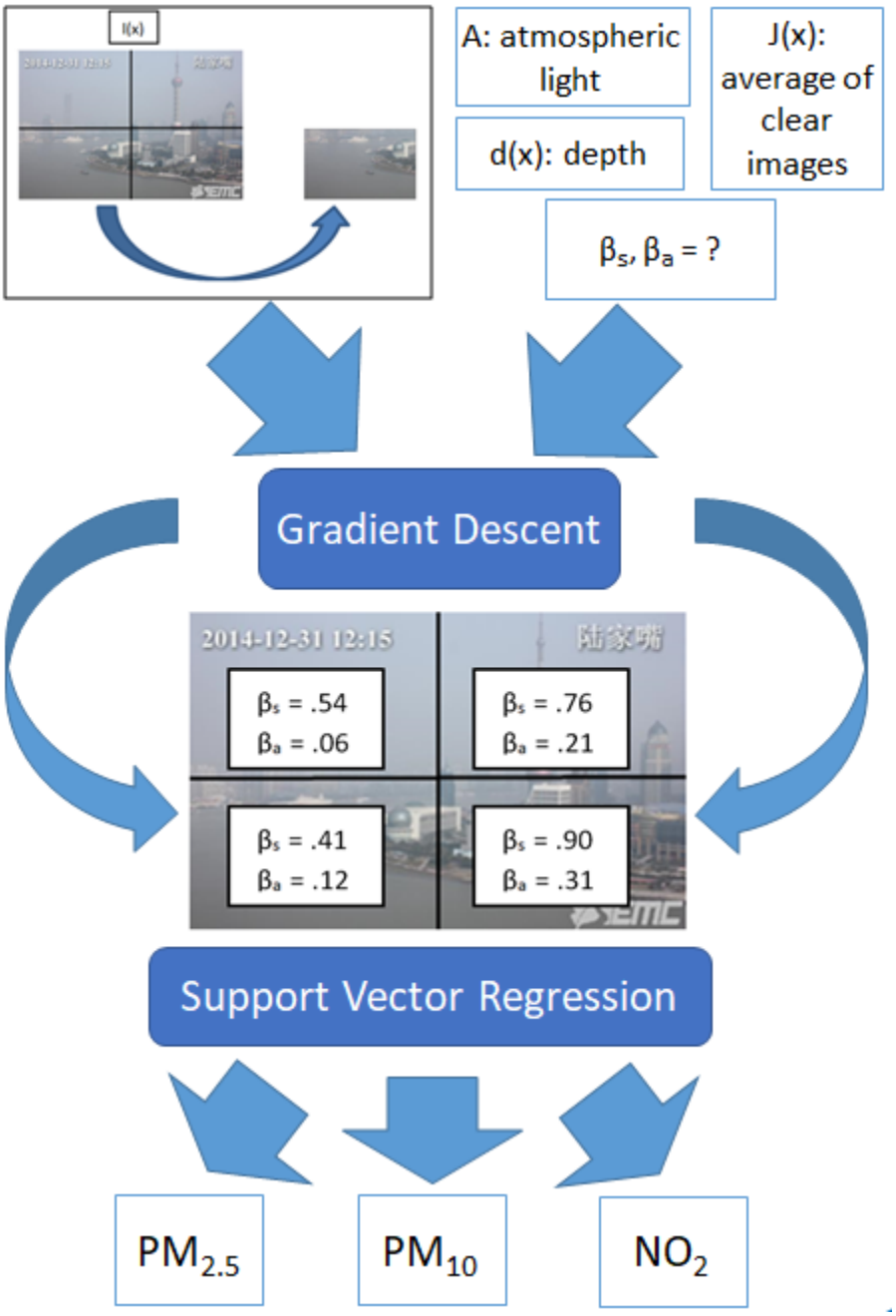


Figure 2.3: The methodology involves two main steps. We first obtain all the light attenuation coefficients β_{sc} and β_{ac} of scattering and absorption. We then determine their relationships with pollutant concentrations using support vector regression. $PM_{2.5}$, PM_{10}, and NO_2 have different color-dependent properties for scattering and absorption. Hence, we can predict all those pollutants from a single image.

Table 2.1: Variance of Light Attenuation within the Grid

	β_{s_b}	β_{s_g}	β_{s_r}
2×2	0.17	0.16	0.15
4×4	0.24	0.28	0.30
6×6	0.29	0.31	0.33
8×8	0.33	0.35	0.36
10×10	0.34	0.36	0.37

significantly at different places; light attenuation often varies with altitude. This observation implies that multiple pollution concentrations may occur in the same image. Hence, we split an image up into an $n \times n$ grid and obtain β_{s_c} and β_{a_c} for each n^2 grid element using gradient descent to improve prediction accuracy. This step may help compensate for errors in the depth map and scene radiance and noise in the input image. In Table 1, we calculate the variance of the light attenuation of all grid elements in the image, which increases as n increases. It would be useful to track light attenuation at multiple parts of an image since the variance across grid elements shows that multiple visibility levels exist in a single image.

2.4.2 Estimation of Pollutant Concentrations

After all coefficients β_{s_c} and β_{a_c} are extracted from each image, we determine their relationships with pollutant concentrations using support vector regression (SVR). We use an SVR with a radial basis kernel function because it has the ability to map the coefficients β_{s_c} and β_{a_c} to pollutant concentration through high-dimensional space. In particular, the feature space of the RBF kernel has an infinite number of dimensions. Since $PM_{2.5}$, PM_{10}, and NO_2 have different color-dependent properties for scattering and absorption, it is possible to predict all those pollutants from a single image. Given a dataset $\{(x_1, y_1), ..., (x, y)\}$, we find a regression function $f(x) = w\phi(x) + b$ that solves the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||_2^2 + C\sqrt{\sum_{n=1}^{m}(\xi_i^+)^2} + C\sqrt{\sum_{n=1}^{m}(\xi_i^{-2})} \\ \text{subject to} \quad & y_i - f(x_i) \leq \epsilon + \xi_i^+, \\ & y_i - f(x_i) \geq -(\epsilon + \xi_i^-), \text{ and} \\ & \xi_i^+ \geq 0,\ \xi_i^- \geq 0,\ i = 1, \ldots, m. \end{aligned}$$

This formulation uses l_2 regularization in the case of non-linearly separable datasets and outliers, where C value is the penalty parameter.

2.5 Results and Discussion

This section describes data collection, experimental evaluation, and findings.

2.5.1 Data Collection and Experimental Evaluation

The data consist of single-scene images taken in Shanghai and their ground truth pollutant concentrations. The Shanghai dataset consists of 1,890 images taken from May to December in 2014 at various times and were captured at the Oriental Pearl Tower [1]. We use the $PM_{2.5}$, PM_{10}, and NO_2 data provided by sensor stations within the cities as ground truth, provided by the Ministry of Environmental Protection of China. The units for $PM_{2.5}$ and PM_{10} are $\mu g/m^3$ and for NO_2 is parts per billion (ppb).

The SVR model uses $6n^2$ features from β values evaluated with two-fold cross validation. The C value for Shanghai is 200. The two evaluation metrics used are the R^2 coefficient of determination and the root mean squared error (RMSE) between the estimated and ground truth pollutant concentration.

2.5.2 Effect of Absorption and Color Properties

We evaluate the impact of absorption and color-dependent light extinction. We use the 8×8 grid size for Shanghai. When absorption is neglected, β_{s_c} is still determined using gradient descent for all c in Eq. 2.1 and 2.2. We also consider neglecting color-dependent properties and find β_s and β_a using gradient descent on a grayscale version of the problem. As shown in Fig. 2.4, considering each property generally improves results.

2.5.3 Effect of Grid Resolution

We evaluate the effect of using n^2 grid elements since light attenuation varies across an image. For the Shanghai dataset, shown in Fig. 4, the RMSE keeps decreasing as the grid size increases to 10×10. Obtaining β values for an increasing number of grid elements initially rapidly increases accuracy and then levels off. A simple approach to selecting grid resolution would be to use 10×10 for all pollutants, which always enabled accuracy near that of the optimal resolution and is computationally tractable.

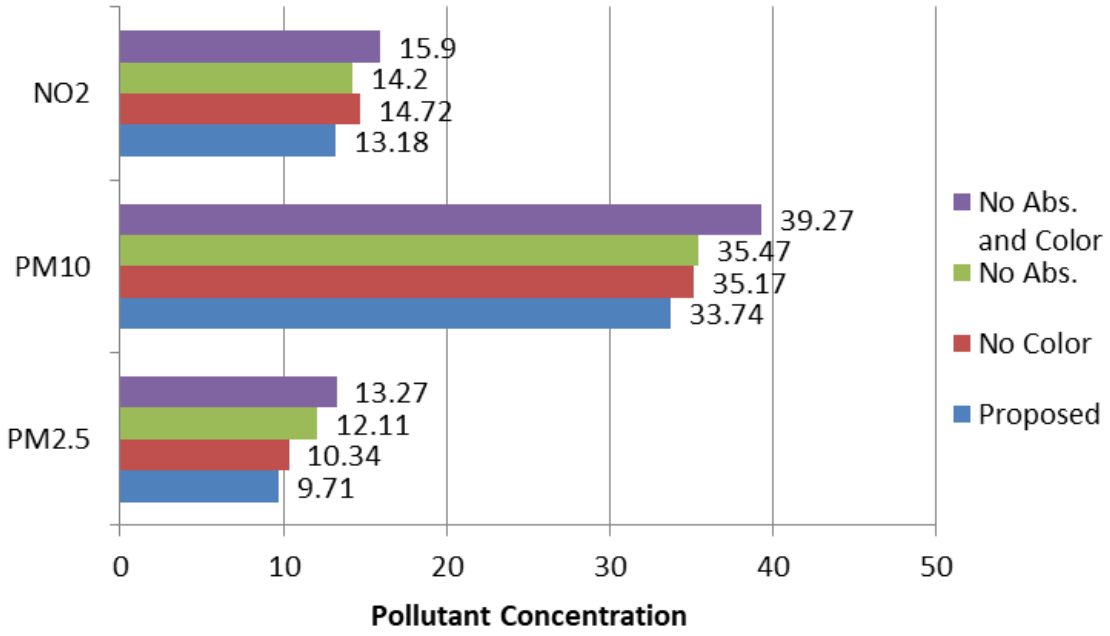


Figure 2.4: RMSE for absorption and color (Shanghai). Wavelength-dependent scattering and absorption properties can enable analysis of multi-pollutant systems and improve estimation accuracy. Considering each property generally improves results.

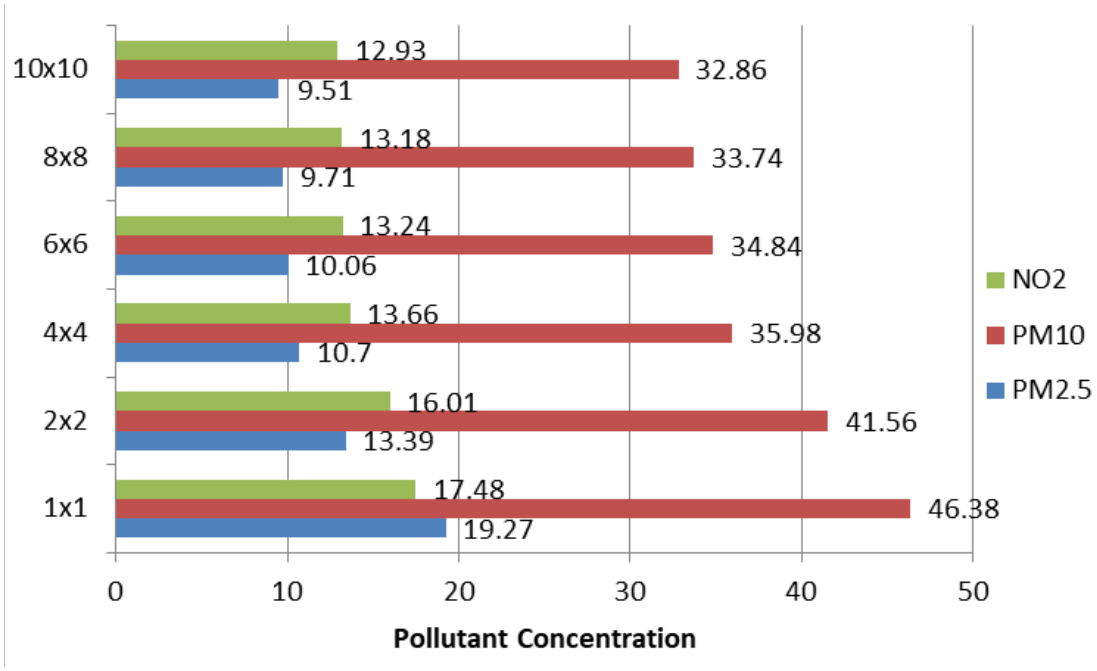


Figure 2.5: RMSE for various grid resolutions (Shanghai). We evaluate the effect of using n^2 grid elements since light attenuation varies across an image. For the Shanghai dataset, shown in Fig. 4, the RMSE keeps decreasing as the grid size increases to 10×10.

Table 2.2: Comparison of Results with Other Research

Shanghai		$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	NO_2 (ppb)
Proposed w.o. Weather	RMSE	8.68	28.00	11.67
	r^2	0.917	0.779	0.750
Proposed w. Weather	RMSE	8.32	27.18	11.52
	r^2	0.924	0.794	0.757
Liu et al.	RMSE	13.65	35.46	15.83
	r^2	0.76	0.640	0.548
Li et al.	RMSE	25.66	52.94	19.41
	r^2	0.260	0.208	0.302
Improvement	%	39.05	23.35	27.23

2.5.4 Discussion

We compare the performance of the proposed approach with the best known existing techniques for estimating air quality via images in Table 2.2 use feature selection (FS) of coefficients to increase accuracy, eliminating those features that increase the root mean square error. We also evaluate our technique with additional weather features, incorporating humidity, temperature, pressure, and wind speed. For the proposed approach, the largest grid size 10×10 is used. It is possible to obtain an even higher accuracy using a more optimal grid size as shown in Figure 1. The current approach outperforms Liu et al. [56] and Li et al. [57] for all three pollutants.

Various factors influenced the prediction accuracy. The distance between the air quality sensors and image sensors is greater than 25 kilometers. All three pollutants may have high spatial and temporal variation so the large distance might introduce error in the ground truth data (i.e., it is possible that the reported error is higher than the actual error). In addition, the distribution of pollutant concentrations were skewed towards lower values. This could result in underscoring insignificant features and missing significant features. Also, the current state-of-the-art depth map estimation may contain errors in certain areas in the image. Any errors in depth estimates may introduce errors in β_{s_c} and β_{a_c} estimation because in Eq. (3) the depth is an exponent of e.

2.6 Conclusion

Vision-based air pollution estimation is an emerging research area with the advantages of low cost, wide coverage, and high spatial resolution. Estimating air pollution from

images is valuable since high spatial resolution data are needed for exposure estimation and pollutants such as PM and ozone have concentrations that can vary at small spatial scales. We have shown that using position- and color-dependent features of both scattering and absorption are useful for estimating multiple pollutants in images, namely $PM_{2.5}$, PM_{10}, and NO_2. Our future work consists of developing a technique to accurately predict pollutant concentrations using images from web crawling and crowdsensing and estimating position-dependent concentration variation within images.

CHAPTER 3

HVAQ: A High-Resolution Vision-Based Air Quality Dataset

3.1 Introduction

This chapter describes a dataset containing high spatial (one sensor every 2.5 km^2) and temporal (one second interval) resolution particle counter based pollution measurements with corresponding images, in addition to auxiliary information including GPS locations, humidity, and temperature. These properties are significant: this is the first publicly available dataset capable of being used to train and evaluate vision-based pollution estimation and forecasting techniques at high spatial resolutions. To the best of our knowledge, there have been no publicly available datasets enabling evaluation in this context.

The PM concentrations are correlated with source distributions. For example, PM has heterogeneous sources [58], for example, automobiles, manufacturing, and building construction. In addition, numerous factors, including wind, humidity, and geography [59, 60], are related to PM distributions. Increasing sensor density or adding image sensors supporting high spatial resolution captures can increase the field estimation accuracy and resolution of pollutant concentrations.

Ground truth pollution data are typically obtained from the nearest of several sparsely deployed monitoring stations. Lack of high-resolution ground truth pollution measurements makes it difficult to evaluate vision-based pollution estimation techniques, as pollution can vary rapidly with position [12]. Moreover, vision-based pollution estimation techniques generally assume homogeneous distributions of particles and gases within images, implying constant light attenuation [12]. In reality, pollution concentration changes rapidly in space. Therefore, accurate evaluation of vision-based estimation algorithms requires high-resolution datasets [15].

There exist datasets containing high-resolution [15, 61] and wide coverage [62, 63] air pollution data. However, none of them contain corresponding synchronized images. I made

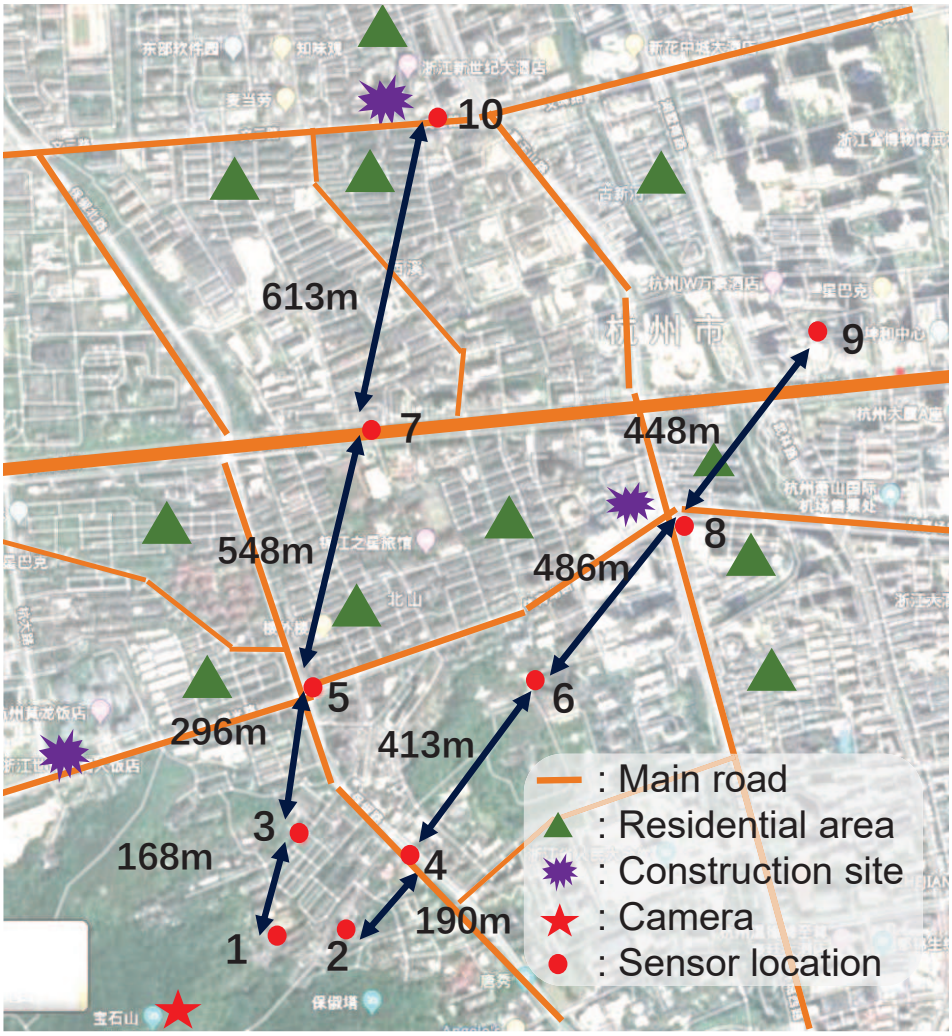


Figure 3.1: The distribution of sensors and pollution sources. Sensor locations are numbered in ascending order according to the distance from the camera.

several observations from the dataset, e.g., the rate of spatial variation in pollution concentration and the evaluation of several existing vision-based PM concentration prediction algorithms. To evaluate the improvement brought by increasing sensor spatial resolution and using images, I use heterogeneous information for concentration estimation, i.e., images and particle counter-based PM concentrations, which were not considered in the previous vision-based algorithms. For the images in HVAQ, using images decreases MAE by 8.4% on average; hence adding a camera to collect images helps more than adding more particle counting sensors.

3.2 Contributions

I participated in a collaborative effort to develop a novel air quality dataset with the goal of validating vision-based pollution estimation algorithms. I will describe some contributions of other team members to provide context for my contributions. Other contributors led work on collecting data for the air quality dataset, with my help and advice. The dataset contains images and corresponding "ground truth" pollution concentration values in images. In particular, there is varying pollution concentration over time and in the images. The novel contribution is a densely distributed, low-cost $PM_{2.5}$ and PM_{10} database with high temporal and spatial resolution enabling scientific discovery and the evaluation of air pollutant estimation algorithms. The dataset also includes images that capture the distributed area of the sensors. This dataset, the first of its kind, is significant because we have shown that air quality estimation accuracy substantially improves by using both data from existing sensors and image-based analysis of air pollution.

I led the work on the analysis of the data with respect to vision-based pollution estimation algorithms. My main contribution is performing extensive analysis on how the estimation of PM2.5 depends on point sensor density and absence/presence of cameras. I made several observations from the dataset, e.g., the rate of spatial variation in pollution concentration increases with distance. To evaluate the improvement brought by increasing sensor spatial resolution and using images, I used heterogeneous information for concentration estimation, i.e., images and particle counter-based PM concentrations, which were not considered in the previous vision-based algorithms. For HVAQ, I found that using images decreases MAE by 8.4% on average; hence adding a camera to collect images helps more than adding more particle counting sensors. Additionally, I find that the pollutant distribution is spatially correlated; spatial variation of PM2.5 is high; and temperature and humidity had limited correlation with PM concentration in our dataset.

3.3 Related Work

Most existing air quality monitoring systems [64] have low temporal and spatial resolutions. For example, Janssens-Maenhout et al. [65] provide a harmonized gridded air pollution emission dataset. This dataset includes multiple pollutants on a global scale with $0.1^\circ \times 0.1^\circ$ spatial resolution (latitude and longitude). De et al. [66] collect an air quality dataset containing 9,358 instances of hourly averaged responses from metal oxide chemical sensors. Devices are deployed in a highly polluted area in Italy at high spatial resolution. Li et al. [67] provide a dataset of mobile air quality measurements in Zurich. They use sensor

Location	P1	P2	P3
Longitude	120.153173°	120.15488°	120.153894°
Latitude	30.269884°	30.268726°	30.27096°
P4	P5	P6	P7
120.156252°	120.153905°	120.15936°	120.155162°
30.270242°	30.27358°	30.273139°	30.278369°
P8	P9	P10	Photo location
120.161912°	120.164792°	120.156541°	120.153955°
30.276465°	30.279437°	30.283932°	30.267191°

Table 3.1: GPS Locations of the Sensors and Images in HVAQ

boxes installed on mobile trams. Static installations are also deployed close to high-quality reference stations for calibration. Their sensors move around the city, recording every $5\,\mathrm{s}$. Apte et al. [15] use two Google street view vehicles equipped with data acquisition platforms to collect the air pollution data of a $30\,\mathrm{km}^2$ city area. Their data contains nitrogen oxides and black carbon. Unlike our dataset, it does not contain PM concentrations, images, and weather conditions. Wei et al. [62] describe the ChinaHighPM10 dataset, which integrates multiple data sources and contains hourly PM10 data in China with $1\,\mathrm{km}$ resolution. Generally, these datasets typically do not include images for the evaluation of vision-based pollution estimation. On the other hand, HVAQ is unique in providing high spatial and temporal resolution pollution measurements with corresponding images. In addition, our dataset makes it possible to evaluate vision-based air pollution algorithms in a high-resolution scenario since ground-truth data previously were of low spatial resolution.

Prior work uses hazy datasets to benchmark image dehazing methods. Li et al. [52] and Ancuti et al. [53] evaluate dehazing algorithms using a synthetic hazy image dataset. Recently, a realistic haze dataset O-HAZE was introduced, which contains 45 pairs of realistic hazy and haze-free ground-truth images for the same scene [51]. The haze originated from a haze machine. O-HAZE, however, does not contain enough pairs for training and validating machine learning approaches. In addition, the haze in the O-HAZE dataset may not reflect the spatial variation of haze in natural scenes influenced by weather or pollution because all hazy images were produced using the same haze machine and chemical process. Moreover, there is a lack of ground-truth distribution for O-HAZE.

Past research also constructed synthetic datasets for other image restoration applications. Li et al. presents a benchmark dataset of both synthetic and real-world rainy images to evaluate existing image deraining algorithms [68]. Lai et al. presents a benchmark dataset of both synthetic and real-world blurred images to evaluate existing image deblurring algorithms [69]. Finally, multiple underwater image datasets have been proposed for the evaluation of underwater image enhancement algorithms [70, 71].

3.4 Sensor Deployment

This section describes the deployment of our sensors and cameras in order to collect both high-resolution ground-truth data and corresponding images.

3.4.1 Sensor Calibration

Our sensing platform is equipped with humidity and temperature sensors. The precision is 3% relative humidity (RH) and $\pm 0.3\,^{\circ}\mathrm{C}$, respectively. The PM sensor can detect particles with $0.3\,\mu\mathrm{m}$ to $10\,\mu\mathrm{m}$ diameters according to the scattered light intensity in a specific direction [72].

We calibrated our particle counting sensors based on the measurements from the air quality monitoring site of Hangzhou Meteorological Bureau in Hemu Primary School using co-location [73]. Our device is co-located with an air monitoring station, which is equipped with a high-precision sensor. The device collects data for 58 hours continuously. The station data are considered ground truth. We use the least-squares method to fit a quadratic function to the ground truth data. According to the measured concentration, we fit the data into a two-stage piecewise linear function. The fitted function is as follows.

$$y = \begin{cases} 1.61x + 16.01 & \text{for } 0 \leqslant x \leqslant 30 \text{ and} \\ 0.13x + 29.48 & \text{for } x > 30, \end{cases} \tag{3.1}$$

where x is the original value and y is the calibrated one. The calibration data are gathered during a 2 day period and the result is shown in Figure 3.2. Calibration reduces root mean squared error from $32.74\,\mu\mathrm{g\,m^{-3}}$ to $3.88\,\mu\mathrm{g\,m^{-3}}$, while the variance for all the data over all locations is $9.33\,\mu\mathrm{g\,m^{-3}}$.

3.4.2 Deployment Details

Our dataset[1] contains both high-resolution ground truth data and wide-view images. We deploy our sensors in the urban area of Hangzhou, a city frequently affected by high PM2.5 concentrations [74]. Existing research [75] shows that in the main urban area of Hangzhou, the sources of PM2.5 are biomass burning/construction dust (41.6%), vehicle exhaust/metallurgical (metals' production and purification) dust (29.3%), unknown source (11.2%), oil combustion (9.8%), and soil (8.0%). As shown in Figure 3.1, the main pollution sources are marked on the map and the sensors are located on two straight lines from

[1] Available on https://github.com/implicitDeclaration/HVAQ-dataset/tree/master.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0	113	168	281	464	697	1012	1150	1730	1625
P2		0	211	190	509	603	1040	1089	1537	1660
P3			0	245	296	575	844	1000	1450	1457
P4				0	388	413	871	899	1347	1510
P5					0	464	548	819	1250	1161
P6						0	625	486	934	1170
P7							0	656	918	613
P8								0	448	953
P9									0	877
P10										0

Table 3.2: Pairwise Sensors Distance (meters)

	Camera	Quadcopter	Sensors	Battery	Platform
Price ($)	200-460	1500	3-15	12	37
Model	Quadcopter default camera, iPad, Onplus 7 phone camera	Dji Phantom 4 Pro	PM2.5, PM10, humidity, tempera-ture	4000 mAh	Raspi 3B+

Table 3.3: Used equipments.

the observation point. The GPS locations and distances between each pair of sensors are listed in Table 3.3 and Table 3.4.2. The sensors are sampled every second and an image is captured every $20\,\mathrm{minutes}$. Since the temporal variation of PM concentration is slower than the acquisition rate of the sensors, and the image acquisition rate is also limited by the flight time of quadcopter, images are captured less frequently.

To derive a wide-view image covering all sensor locations, we mount a camera on a quadcopter. We use two approaches to take photos. First, we use a quadcopter equipped with a 4864×3648 camera at $90\,\mathrm{m}$ altitude. Moreover, since the quadcopter has limited carrying and battery capacities, we take pictures from a fixed location on the mountain top with a resolution of 2592×1936 using an iPad at a resolution of 4000×3000 using a phone. The parameters of our cameras are listed in Table 3.4.2. The quadcopter camera uses the Sony Exmor R CMOS sensor. The other two cameras types are the Apple iSight and Sony IMX586. 106 images and over 300 thousand PM samples were gathered in three days.

The cost of vision-based approach (about $30) is much less than the total cost of the sensors (about $670). Figure 3.4.2 lists our equipment. Thus, predicting air pollution concentrations using images is more convenient and less expensive than particle counters, but requires sophisticated image processing algorithms.

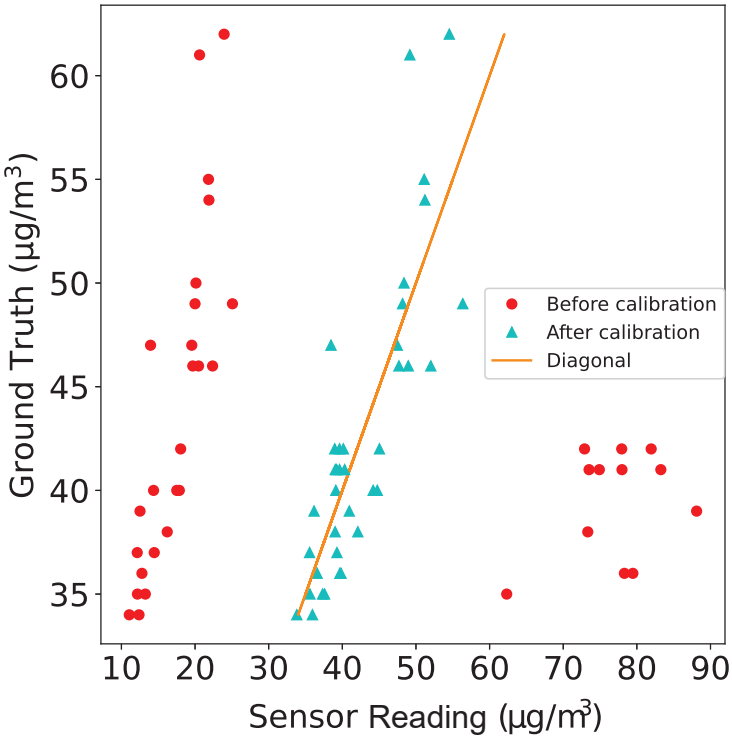


Figure 3.2: Pollution concentration calibration.

Camera	Pixel	Sensor	Aperture
quadcopter	20,000,000	Sony Exmor R	f/2.8-f/11
pad	8,000,000	iSight	f/2.4
phone	48,000,000	Sony IMX586	f/1.6

Table 3.4: Camera Parameter.

Figure 3.3: The quadcopter (Dji Phantom 4 Pro) used in our deployment.

3.5 Dataset Analysis

We investigate four questions pertaining to our dataset. First, it is important to investigate the relationship between environmental conditions and measurement accuracy. We also need to study the benefits of increasing the spatial resolution of a pollution sensor network. We then examine the relationship between PM2.5 concentration and distance Finally, it is important to understand how the sensor density and the use of images relates to the PM2.5 estimation accuracy.

Q1) What is the impact of environmental conditions on measurement accuracy? (Answer is A1 in the later section.)

Q2) What are the spatial variation characteristics of PM2.5 concentration? (Answer is A2 in the later section.)

Q3) How does the correlation of pollution concentrations at two different locations depend on their separation? (Answer is A3 in the later section.)

Q4) How much do additional point sensors and vision-based methods improve estimation accuracy? (Answer is A4 in the later section.)

3.5.1 Temperature and Humidity Correlations with Pollution Concentrations

Environmental factors such as weather conditions can affect sensor readings. For the deployments on Oct. 19 and Nov. 10, we calculate the R^2 correlation coefficients for PM2.5 and several environmental factors.

A1: Environmental conditions have limited impact. The results in Table 3.5 shows that the correlations between PM2.5 and weather factors are insignificant. Wu et al. [76] report that temperature is not significantly correlated to PM concentrations in Hangzhou and PM2.5 concentration is only significantly elevated when RH is higher than 60%. Moreover, the correlation between environmental factors and PM concentration also depends on the region and season, e.g., Zhu et al. [77] report that PM concentration and RH show seasonal correlation. However, since temperature and humidity do not directly affect the measurement process of particle counters, we do not include temperature and humidity in our calibration functions.

3.5.2 Correlations of PM Readings

Our measurements demonstrate that it typically takes more than 10 minutes for concentration to change by $10\,\mu\mathrm{g\,m^{-3}}$. Our data sampling period is one second and is considered

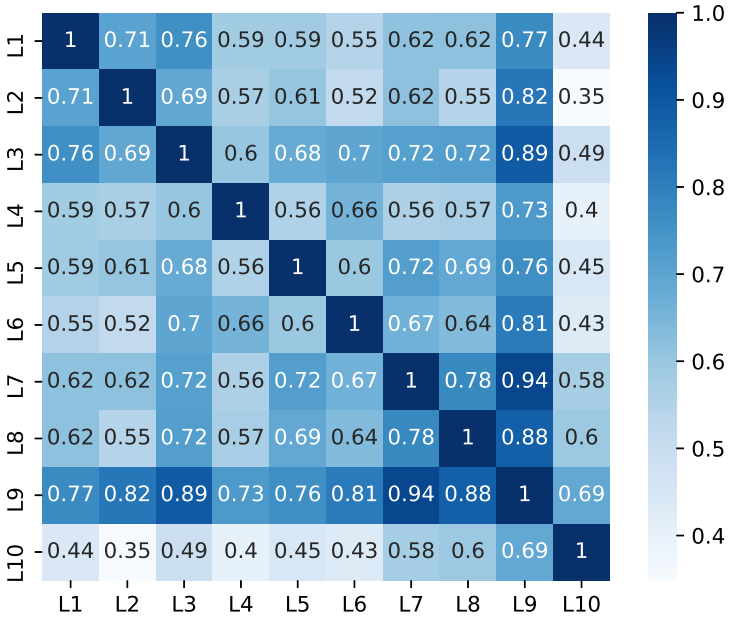


Figure 3.4: The confusion matrix for PM2.5 correlation. The number on the axis represents the number of the corresponding location.

Table 3.5: PM2.5 Correlation with PM10 and Environmental Factors on Oct. 19 and Nov. 10

	PM2.5	Temperature	Humidity
PM2.5	1.0	0.298	0.306
Temperature	0.298	1.0	0.808
Humidity	0.306	0.808	1.0

Data in Different Days	Standard Deviation	Data Range	Average Value	Date
PM2.5 ($\mu g\,m^{-3}$)	2.03	4.56	13.93	Jul. 24
PM2.5 ($\mu g\,m^{-3}$)	2.22	5.11	25.03	Jul. 06
PM2.5 ($\mu g\,m^{-3}$)	6.68	22.21	56.90	
Temperature (°C)	5.64	16.02	26.40	Oct. 19
RH (%)	12.36	37.63	45.13	
PM2.5 ($\mu g\,m^{-3}$)	4.77	16.51	48.76	
Temperature (°C)	4.53	14.21	24.79	Nov. 10
RH (%)	10.20	34.71	40.16	

Table 3.6: Statistics for PM and Environmental Data

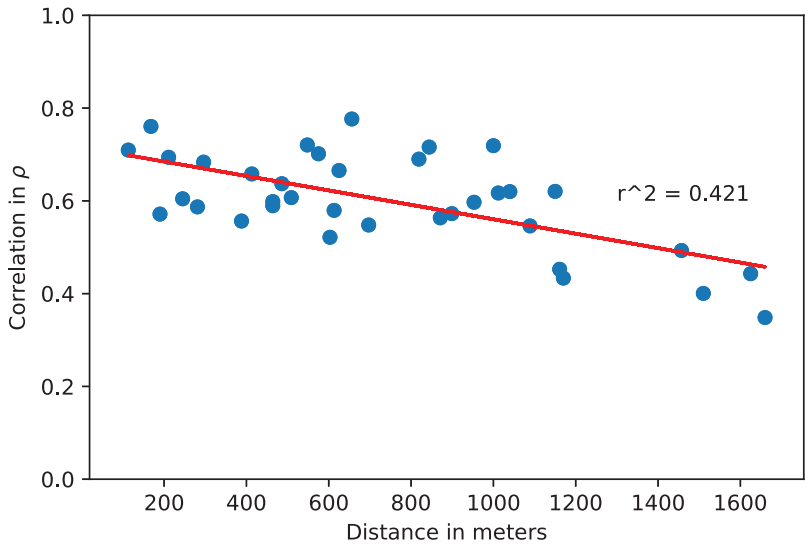


Figure 3.5: Pairwise correlation between sensors for Oct. 19 and Nov. 10 as functions of their distances.

adequate given the relatively slow concentration change. We quantify the spatial variation of pollution by calculating the standard deviations over all sensors.

A2: Spatial variation of PM2.5 is high. As shown in Section 3.5.2 and Figure 3.5, the 2 locations have different concentration and variation trends. Moreover, the large differences indicate that multiple pollution levels coexist in a single image.

A3: Pollution concentrations are spatially correlated. We quantify the correlations for further analysis. Figure 3.5 shows sensor correlations as a function of pairwise distance. The correlation is measured by the Spearman correlation coefficient:

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}, \tag{3.2}$$

where d_i represents the position difference of the paired variables after the two variables are sorted separately and N is the total number of samples. The slope of the fitting line is $0.155\,\mathrm{km}^{-1}$ and the coefficient of determination (R^2) is 0.421, which implies that closer sensors enable more accurate estimates. Figure 3.4 shows the sensor correlations during the Nov. 10 and Oct. 19 deployments. We expect the correlations to decrease with increasing inter-sensor distances. The deployment results confirm our hypothesis.

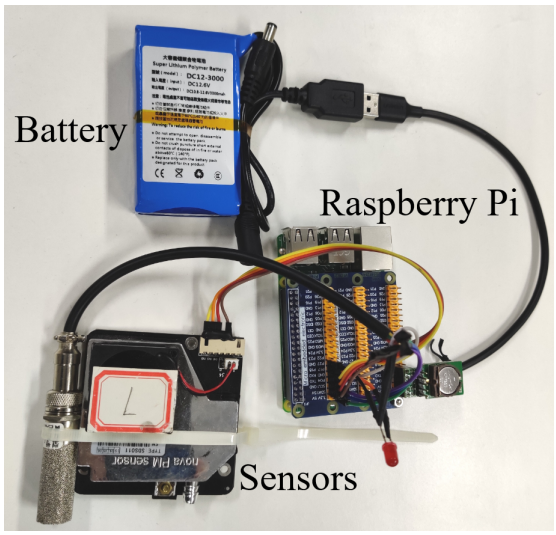


Figure 3.6: The sensing platform consisting of battery, Raspberry Pi, and sensors.

3.6 Experimental Results

In this section, we evaluate the impacts of changing sensor density and using vision-based techniques on estimation accuracy. Specifically, we estimate PM2.5 concentration based on vision-based analysis using transmission information and standard deviation of gray-scale pixel values and compare the performance of several state-of-art estimation algorithms.

3.6.1 Experimental Setup

We design a portable sensing platform to collect, process, and transmit data, as shown in Figure 3.6. The system battery life is 3.5 hours. The following algorithms are used to estimate pollutant concentration from the measured data.

1. **Gradient boosting regression (GBR):** This method combines a group of weak learners with low complexity and low training cost. It reduces the problem of over-fitting and modifies the weights at each training round to produce a strong learner. Gradient boosting modifies its models based on the gradient descent direction of the loss functions of the previously established models.

The algorithms are implemented using the Python package *sklearn*. The RFR parameter *n_estimators* is set to 100 and *criterion* is set to mean squared error. We set the GBR parameter *learning_rate* to 0.1 and *n_estimators* to 100, using least squares regression. We set the SVR parameter *c* to 1.0 and *epsilon* to 0.1, using a radial basis function kernel. The parameters are chosen to maximize the training performance. We use the following

equation to combine sensor readings and image properties:

$$S(x,t) = G(s_1(t), s_2(t), ..., s_{10}(t), t_{dcp}(t), \beta_{sd}(t)), \tag{3.3}$$

where x is the location index, $S(x,t)$ is the concentration estimation at time t, $s_1, s_2, ..., s_{10}$ are the available 10 sensors, $t_{dcp}(t)$ is the transmission information at time t, $\beta_{sd}(t)$ is the standard deviation of gray-scale image, and G is the estimation algorithm, which refers to RFR, GBR, or SVR. We use $t_{dcp}(t)$ for low-altitude data and $\beta_{sd}(t)$ for high-altitude data. We later describe how $t_{dcp}(t)$ and $\beta_{sd}(t)$ are obtained for each image in the dataset in section 5.B.

We use data from Jul. 24, Oct. 19, and Nov. 10 for all time stamps and use the mean absolute error (MAE) as the evaluation criteria.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{3.4}$$

where y_i is the actual PM2.5 concentration, and $\hat{y}_i$ is the predicted PM2.5 concentration.

There are two classes of images in our dataset. Those in the first class were captured from the ground, on top of a mountain ($78\,\mathrm{m}$). The second class contains images from a quadcopter flying above the same mountain ($78\,\mathrm{m}$ + $90\,\mathrm{m}$). We tried to take the images from the quadcopter and mountain at the same angle and keep the images as similar to each other as possible, but there is some (unavoidable) variation in camera orientation.

We divide our data into "low-altitude" and "high-altitude" subsets according to the image class and evaluate our algorithms separately on the two classes. We analyze 19 images from the high-altitude dataset and 26 images from the low-altitude dataset. Furthermore, each dataset is divided as follows: 75% of the data and images are randomly selected as a training set and the rest are the testing set. We run the prediction model 50 times per random split of the training and testing datasets.

3.6.2 Image Enhanced Concentration Estimation

Images can be used to estimate PM concentrations in large areas because images can extract the haziness information. We predict PM2.5 concentrations at locations without sensors. Since PM2.5 attenuates light, we estimate PM2.5 concentration in part through the light attenuation coefficient β from the haze model. We use the following image features to estimate PM2.5: the dark channel $t_{dcp}(x)$ and the standard deviation β_{sd}. $t_{dcp}(x)$ is determined using Equation 3.13 and β_{sd} is determined using Equation 3.21. Note that certain

kinds of weather conditions like rain and snow might be misinterpreted as pollution. Our PM2.5 estimation algorithm mainly works in daytime sunny and cloudy conditions, which are common in Hangzhou.

The atmospheric model describing an image influenced by haze follows [11]:

$$\boldsymbol{I}(x) = \boldsymbol{J}(x)t(x) + \boldsymbol{A}(1 - t(x)), \tag{3.5}$$

where x is the pixel location, $\boldsymbol{I}$ is the observed image, $\boldsymbol{J}$ is the scene radiance (image without any haze), $\boldsymbol{A}$ is the atmospheric light, t is the transmission function.

Images with higher air pollution tend to look hazier due to lower transmission and contrast. Hence, image features that correlate with haze level enable pollutant concentration estimation.

3.6.2.1 Dark Channel Prior

The dark channel prior, which has been widely used for haze removal, can be used to estimate the transmission of each image pixel. The dark channel prior method is based on the observation that in most haze-free patches, at least one color channel has some pixels with very low intensities. The dark channel is defined as the minimum of all pixel colors in a local patch and can be calculated using the following equation [78]:

$$J_{dark}(x) = \min_{c \in r,g,b} \left(\min_{y \in \Omega_r(x)} J^c(y) \right), \tag{3.6}$$

where J^c is an RGB channel of $\boldsymbol{J}$ and $\Omega_r(x)$ is a local patch centered at x with the size of 15×15. Assume the atmospheric light $\boldsymbol{A}$ is given and the transmission in a local patch $\Omega_r(x)$ is constant, taking the minimum operation in the local patch on Equation 3.5, we have

$$\min_{y \in \Omega_r(x)} (I^c(y)) = \tilde{t}(x) \min_{y \in \Omega_r(x)} (J^c(y)) + (1 - \tilde{t}(x))A^c, \tag{3.7}$$

where $\tilde{t}(x)$ is the patch's transmission. The minimum operation is performed on three color channels independently, it is equivalent to

$$\min_{y \in \Omega_r(x)} \left(\frac{I^c(y)}{A^c} \right) = \tilde{t}(x) \min_{y \in \Omega_r(x)} \left(\frac{J^c(y)}{A^c} \right) + (1 - \tilde{t}(x)). \tag{3.8}$$

By taking the minimum of three color channels, we have

$$\min_c \left(\min_{y \in \Omega_r(x)} \left(\frac{I^c(y)}{A^c} \right) \right) = \tilde{t}(x) \min_c \left(\min_{y \in \Omega_r(x)} \left(\frac{J^c(y)}{A^c} \right) \right) + (1 - \tilde{t}(x)). \tag{3.9}$$

According to the definition of dark channel prior, the dark channel J_{dark} of the haze-free radiance $\boldsymbol{J}$ tends to be zero

$$J_{dark}(x) = \min_c \left(\min_{y \in \Omega_r(x)} J^c(y) \right) = 0. \tag{3.10}$$

Because A^c is always positive, this lead to

$$\min_c \left(\min_{y \in \Omega_r(x)} \frac{J^c(y)}{A^c} \right) = 0. \tag{3.11}$$

Substituting Equation 3.11 into Equation 3.9, we can estimate the transmission as follows.

$$\tilde{t}(x) = 1 - \min_c \left(\min_{y \in \Omega_r(x)} \frac{I^c(y)}{A^c} \right). \tag{3.12}$$

In practice, the atmosphere always contains some haze, which provides depth information. We can optionally keep a small amount of haze by introducing a constant parameter ω $(0 < \omega < 1)$ into Equation 3.12

$$\tilde{t}(x) = 1 - \omega \min_c \left(\min_{y \in \Omega_r(x)} \frac{I^c(y)}{A^c} \right). \tag{3.13}$$

We fix the value of ω to 0.95 because this approximates the sparse haze present even on relatively clear days. The atmospheric light A is estimated through this procedure: we pick the top 0.1% brightest pixels in the dark channel and the input image $\boldsymbol{I}$ to calculate the atmospheric light. For each hazy image in our dataset, we take the average of the pixel-level transmissions estimated using Equation 3.13. The resulting average is used for low-altitude data.

3.6.2.2 Standard Deviation

In order to calculate the intensity standard deviation for each image, we convert the RGB image to a gray-scale image, then calculate the standard deviation of all the pixel intensities. The standard deviation is closely related to haze density [79]. The scattering coefficient is a measure of haze density. The higher the scattering coefficient, the higher the haze density.

The transmission function is

$$t(x) = e^{\beta d(x)}, \tag{3.14}$$

where β is the scattering coefficient and d is the depth. Substituting Equation 3.14 into Equation 3.5, we have

$$\boldsymbol{I}_g(x) = \boldsymbol{J}(x)e^{\beta d(x)} + \boldsymbol{A}(1 - e^{\beta d(x)}). \tag{3.15}$$

The variance of a gray-scale image is

$$\begin{aligned} \sigma_{I_g}^2 &= \frac{1}{N}\sum_{i=1}^{N}(I_g(i) - \frac{1}{N}\sum_{j=1}^{N} I_g(j))^2 \\ &= e^{2\beta d(x)}\frac{1}{N}\sum_{i=1}^{N}(J(i) - \frac{1}{N}\sum_{j=1}^{N} J(j))^2, \end{aligned} \tag{3.16}$$

where I_g is the gray-scale image and N is the number of pixels in the image. When $\beta = 0$, we have

$$\sigma_0^2 = \frac{1}{N}\sum_{i=1}^{N}(J(i) - \frac{1}{N}\sum_{j=1}^{N} J(j))^2. \tag{3.17}$$

Combining Equation 3.17 and Equation 3.16 yields

$$\sigma_{I_g} = e^{-2\beta}\sigma_0^2. \tag{3.18}$$

After taking the logarithm of both sides, the scattering coefficient can be expressed as

$$\beta = \ln \sigma_0 - \ln \sigma_{\mathrm{I_g}}. \tag{3.19}$$

Since σ is changed at each image, Equation 3.19 can be expressed using the first-order Taylor Series approximation

$$\beta = 1 + \ln \sigma_0 - \sigma_{\mathrm{I_g}}. \tag{3.20}$$

When $\beta = 0$, the variance of the scene radiance approximates 1. Thus we have

$$\beta = 1 - \ln \sigma_{\mathrm{I_g}}. \tag{3.21}$$

Therefore we can estimate the concentration using the standard deviation of the gray-scale image. The standard deviation is used for high-altitude data.

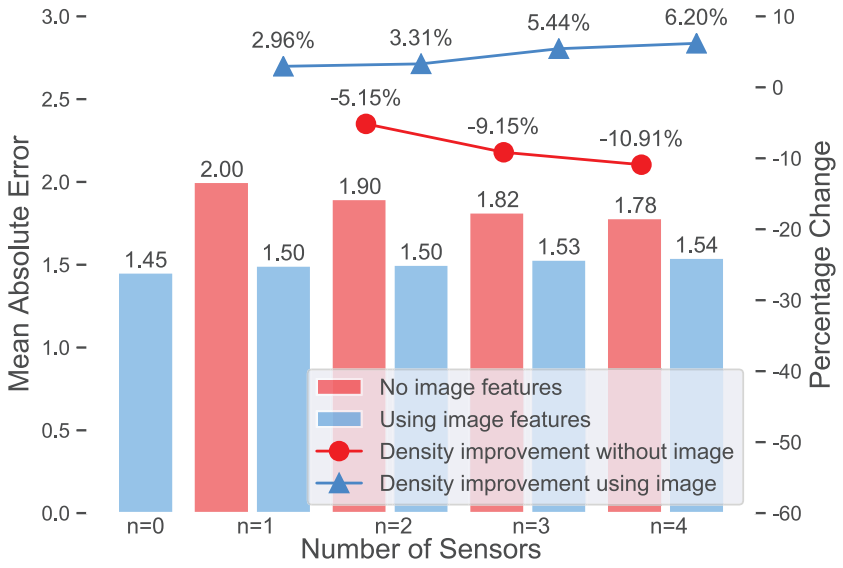


Figure 3.7: The relationship between mean average error and sensor density for Gradient Boosting Regression on high-altitude data.

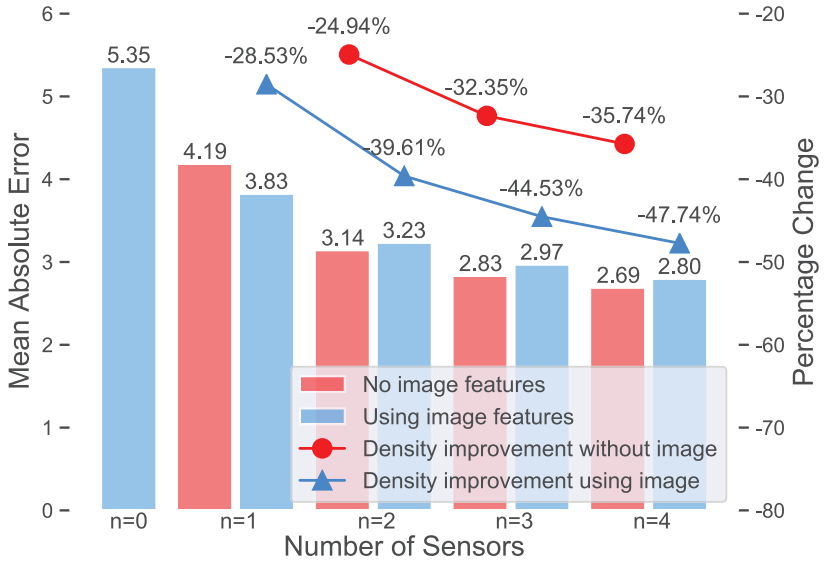


Figure 3.8: The relationship between mean average error and sensor density for Gradient Boosting Regression on low-altitude data.

3.6.3 Concentration Estimation Results

For each available sensor deployment location, we speculatively remove one or more sensor's data and use estimation techniques with access to the remaining sensors to infer concentration(s), thereby allowing comparison with ground truth measurements. We investigate the impact of the number of sensors on the estimation accuracy by using all possible combinations of speculatively removed sensors and averaging the results. We also consider

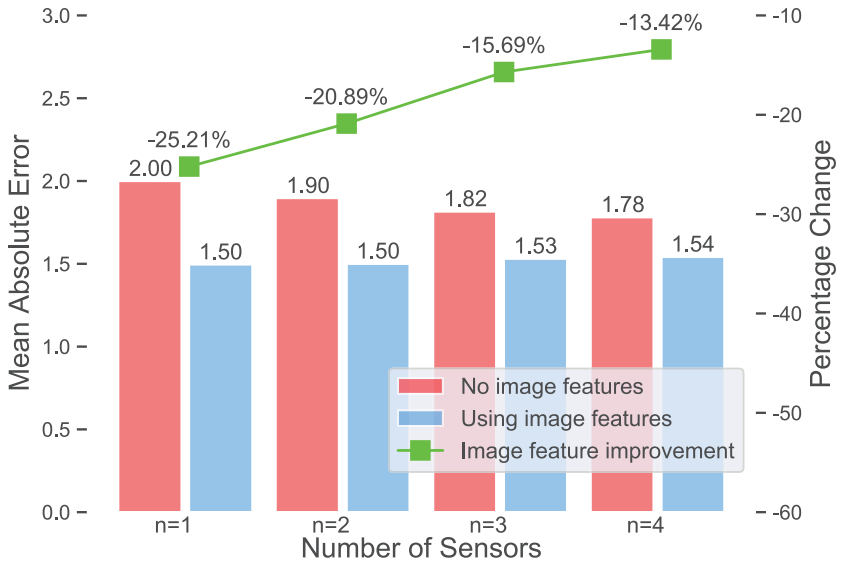


Figure 3.9: The relationship between mean average error and using images for Gradient Boosting Regression on high-altitude data.

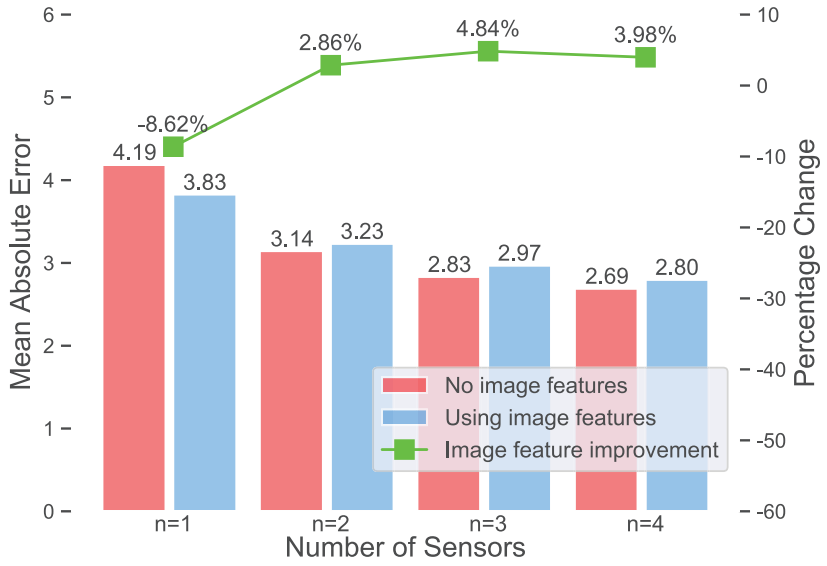


Figure 3.10: The relationship between mean average error and using images for Gradient Boosting Regression on low-altitude data.

the impact of using image data on estimation accuracy.

We plotted the results from the best-performing algorithm: GBR. The density improvements in Figures 3.7 and 3.8 are compared with the $n = 0$ (no images) case. If image features are used, the improvements are compared with the $n = 1$ case. Note that the lower MAE is, the better the estimation accuracy. Thus a negative change indicates improvement. The bars in Figures 3.7–3.10 represent the MAE resulting from using different numbers of

sensors. The lower bars indicate higher accuracies. The lines in Figures 3.7–3.10 indicate percentage change to MAE in different cases (e.g., using images or more sensors). The lower the percentage change, the larger the improvement, and positive percentage changes imply (undesirable) increases in MAE.

As sensor density increases, the estimation MAE decreases. This implies that accuracy improves with sensor density, even at densities much higher than those of modern stationary sensor deployments. Moreover, as shown in Figure 3.5, sensor correlations decrease with increasing distance. This is the reason estimation accuracy improves with increasing sensor density. Using the four nearest sensors instead of one sensor improves estimation accuracy by 23.3% on average without using images, and 20.75% when using images. The fact that increasing sensor density improves accuracy less when images are available does not imply that images are unhelpful. In contrast, it implies that using images allows higher accuracy when few sensors are available, leaving less potential for improvement if sensor density is later increased.

A4 Vision-based techniques significantly improve estimation accuracy. As shown in Figures 3.9 and 3.10, PM2.5 concentration prediction accuracy improves when images are used. In the case of $n = 0$, we average all the available concentrations. The MAE is $20.821\,\mu\mathrm{g\,m^{-3}}$ for high-altitude data and $6.929\,\mu\mathrm{g\,m^{-3}}$ for low-altitude data. The high-altitude images enable higher accuracy because some sensor locations are occluded in the low-altitude images. When images are used, MAE drops to $1.45\,\mu\mathrm{g\,m^{-3}}$ and $5.35\,\mu\mathrm{g\,m^{-3}}$ respectively. For the case where $n = 1$, when we use the PM2.5 concentrations of the nearest available sensor for estimation image data improves prediction accuracy by 16.9% on average. The benefits of using images are greatest when the fewest particle counters are used.

To determine whether the improvement is systemic and statistically significant, we use the Kolmogorov-Smirnov test. We compare the data for the MAE without using images that using images for each number of sensors ($n = 1, ..., 4$) to determine whether the two data sets have the same distribution. The test is non-parametric and requires no knowledge about the distribution of data.

We determine that there is a statistically significant difference between the distribution of the MAE when using images and the distribution without using images, i.e., the difference between the two distributions is not due only to chance or noise but due to a genuine improvement in accuracy when using the image data. If the p-value is below 0.10, we can reject the null hypothesis and conclude that using images does improve our results. As shown in Table 3.7 and Table 3.8, on low-altitude data, the p-values of GBR in all cases are less than 0.10, and on high-altitude data, GBR's p-values are less than 0.10. As a result, we

Table 3.7: P-Values of GBR on High-Altitude Data

GBR			
n=1	n=2	n=3	n=4
< 0.001	< 0.001	< 0.001	0.064

Table 3.8: P-Values of GBR on Low-Altitude Data

GBR			
n=1	n=2	n=3	n=4
0.068	< 0.001	< 0.001	< 0.001

are confident that using images improves the estimates.

Certain fixed-location images show slightly negative results for the GBR method. Since the fixed-location images are taken from a low altitude, some of the sensor locations are blocked by buildings. For the quadcopter images taken at a higher altitude, all estimation techniques improve accuracy. In general, images decrease MAE by 8.44% on average, when $n \leqslant 1$, adding a camera to collect images helps more than adding more sensors.

Increasing sensor density and using images change the relative accuracy of the estimation algorithms. When we use only one sensor and no images, RFR has lower MAE than GBR on low-altitude data. When the sensor density is increased ($n = 4$) and images are used, GBR outperforms RFR. This result demonstrates that it is important to evaluate estimation algorithms using appropriate sensor densities and access to image data. A sparsely deployed and less accurate sensor network can lead to false conclusions about pollution concentrations, and about which pollution concentration estimation algorithms are most accurate.

To summarize, higher sensor densities and image data both improve estimation accuracy, and adding image data has a similar effect to increasing particle counter density by 0.61 sensors km^{-2}. Of the three estimation techniques evaluated, GBR had the highest accuracy with MAE=$1.45\,\mu\mathrm{g\,m}^{-3}$. In particular, our developed method is unlikely to work as well on night-time images and images with adverse weather conditions such as rain and snow. The main limitation of our dataset is that it does not contain nighttime images.

3.7 Conclusion

This project has presented a PM dataset with high spatial and temporal resolution. In contrast with existing datasets, it contains images covering the locations of stationary point sensors, making it suitable for evaluating and validating vision-based pollution estimation algorithms. Through our analysis, we find that (1) the estimation accuracy can be improved significantly using vision-based techniques; (2) the spatial pollutant distribution is

spatially correlated; (3) spatial variation of PM2.5 is high; and (4) temperature and humidity had limited impact on PM concentration in our dataset. We also evaluate our data using state-of-the-art prediction methods. Accuracy correlates with density with a coefficient of $0.2875\,\mu\mathrm{g\,m^{-3}}$ MAE per sensor and vision-based estimation improves accuracy by $0.1813\,\mu\mathrm{g\,m^{-3}}$ MAE, on average.

CHAPTER 4

Nighttime Vision-based $PM_{2.5}$ Estimation

4.1 Introduction

Air quality (AQ) has attracted widespread concern in recent years. $PM_{2.5}$, or ambient fine particulate matter, is a major pollutant consisted of various particles with diameter less than $2.5\,\mu\mathrm{m}$. It is a class I carcinogen certified by the world health organization (WHO) and hence, a great threat to human health [80]. Long-term exposure to high concentrations of $PM_{2.5}$ can damage the cardiovascular and respiratory systems, leading to respiratory disease, heart disease, stroke, and many other health problems. In that case, high $PM_{2.5}$ concentration can significantly increase mortality rates and reduce life expectancy [81–83]. Therefore, large-scale $PM_{2.5}$ monitoring system is important [84, 85].

Traditionally, people rely upon the ground-based observation laboratories to measure $PM_{2.5}$ concentration. Although this method is accurate, it also suffers from the problems of limited coverage and long analysis time. Ground-based AQ sensor networks have similar limitations [84, 86]. On the other hand, as the widespread use of surveillance devices and smartphones, the number of cameras increases rapidly. The scattering and absorption of light caused by atmospheric pollutant particles can have a negative impact on the quality of captured images. In other words, images taken in heavily polluted environments differ significantly from normal images in many attributes, such as brightness, contrast, and saturation. There are several accurate vision-based research works on daytime AQ estimation by leveraging image features such as spatial contrast, dark channel, and variations in sky-earth colors [87, 88].

However, estimating $PM_{2.5}$ concentration during nighttime is still challenging where existing daytime vision-based techniques are not applicable. In low-light conditions, visibility is significantly reduced, causing difficulty in capturing scene details and increased noise levels. Moreover, differences in brightness, saturation, hue, sharpness, and other factors between daytime and nighttime images make it hard to use the daytime algorithm

directly [12, 89, 90]. Meanwhile, rapidly increasing nighttime human activities imply severe $PM_{2.5}$ exposure, which can not be accurately and efficiently monitored using current vision-based techniques. Therefore, a nighttime large-scale $PM_{2.5}$ monitoring technique is useful and urgently needed.

To address the nighttime vision-based $PM_{2.5}$ monitoring problem, we propose to utilize halation [91], which represents the glow around lights, to estimate the $PM_{2.5}$ concentration at night. Glow widely exists in night images, e.g., light can be scattered multiple times before reaching the observer, resulting in a glow around the source. Thus, the glow is typically visible and noise-resistant at night, even if the illumination is low. The floating particle concentrations are correlated to the scattering effect, making it possible to estimate $PM_{2.5}$ concentrations through the glow effect. Other features, such as dark channel and contrast, depend on scene visualization and are susceptible to noise, atmospheric light (moonlight, cloud), and varying artificial light sources. Compared to dark channel based techniques, our glow map based method outperform by 29.3% in our experimental evaluation.

Specifically, we first extract features related to luminance and its attenuation from the source image. These features are combined to create a glow feature map. Then we employ a convolutional neural network (CNN) to match the glow map with ground-truth $PM_{2.5}$ concentration. Finally, we create a real-world dataset to validate our approach.

We summarize our contributions as follows.

1. We propose a nighttime $PM_{2.5}$ concentration estimation method, which first extracts light source brightness, glow intensity attenuation, and glow features from a single image, then predicts the concentration by a CNN. To the best of our knowledge, this is the first vision-based method targeted at night scenarios.

2. We collected a dataset containing 11,753 multi-location images in night scenes with corresponding environment parameters, including $PM_{2.5}$, PM_{10}, temperature, and humidity. This dataset is of high temporal resolution and can contribute to the studies about air quality measurement.

Experimental results show that our method achieves accurate $PM_{2.5}$ estimation, with an error of $2.58\,\mu g/m^3$.

The rest of this chapter is organized as follows. Section 4.3 presents related work. Section 4.4 describes details the proposed method. Section 4.5 describes the data collection and analysis process. Section 7.5 presents the experimental results. Section 4.7 discusses the potential limitations and concludes this chapter.

4.2 Contributions

I participated in a collaborative effort to develop a novel air quality dataset with the goal of developing a novel vision-based pollution estimation algorithm at night. I will describe some of the contributions of the other team members to provide context for my contributions. Other contributors led work on collecting data for the nighttime image-based air quality dataset, with my help and advice. The dataset contains nighttime images and corresponding ground truth pollution concentration values in the images. In particular, an air quality sensing platform is placed in the center of the region to record ground truth pollutant readings for $PM_{2.5}$ concentrations. I provided significant advice on the construction of the vision-based pollution estimation algorithm for night-time applications.

4.3 Related Work

The related work can be generalized into four categories, including $PM_{2.5}$ monitoring, vision-based air quality estimation, remote sensing based $PM_{2.5}$ estimation, and image dehazing.

4.3.1 $PM_{2.5}$ Monitoring

Air quality sensors can accurately measure $PM_{2.5}$ concentration. However, their deployment can be time-consuming, which limit the coverage [92]. Moreover, $PM_{2.5}$ concentration varies dramatically over time and space [93]. It is reported that the concentration can vary up to $10\,\mu g/m^3$ within a 10-minute interval [9]. Therefore, long-term, large-scale, and real-time monitoring is impractical for this type of sensors.

Existing researches estimate $PM_{2.5}$ concentrations using current readings and historical data. For example, Krishan et al. consider the spatial diffusion and long-term dependence of pollutant concentration and develop an air quality prediction model based on long short-term memory (LSTM) [94]. Ma et al. propose a bidirectional LSTM (BLSTM) model for air quality prediction [95]. Guo et al. propose an unsupervised $PM_{2.5}$ estimation method using a time distributed convolutional gated recurrent unit (TCGRU) and k-nearest neighbor inverse distance weighted (KIDW) interpolation to monitor areas without air monitoring stations [96]. Zhang et al. propose a CNN-LSTM hybrid network to model the spatio-temporal correlations between haze images and $PM_{2.5}$ concentrations [97]. It uses multi-level attention to forecast $PM_{2.5}$ concentration. Those methods suffers from the spatial and temporal scarsity problems.

4.3.2 Vision-based air quality estimation

Vision-based estimation methods have much higher spatial and temporal resolution. There are various sources of images available which can be utilized for air quality measurement, including quadcopters, social media, and Google Street View [98] etc.

Zhang et al. propose a method which utilizes scattering and absorption features for the concurrent estimation of multiple pollutants [7]. Su et al. propose an end-to-end CNN to estimate multiple atmospheric environmental parameters [99]. Wang et al. develop a dual-channel air quality measurement method based on videos [88]. Yang et al. design ImgSensingNet, a vision-guided aerial-ground sensing system which consists of unmanned aerial vehicles (UAVs) and a ground sensor network. It combines vision-based air quality monitoring and the ground sensing network to improve accuracy [87]. However, none of those methods can be used directly for nighttime monitoring.

4.3.3 Remote sensing based $PM_{2.5}$ estimation

Recent works on nighttime $PM_{2.5}$ concentration estimation are based on remote sensing techniques. However, due to the absence of sunlight, it is impractical to estimate aerosols from visible bands. Weng et al. propose to use thermal channel data and aerosol absorbing properties to estimate $PM_{2.5}$ [100]. Wang et al. analyze the relationship between nighttime light radiance, meteorological elements, and topographic elements. Then they use multiple linear regression and support random forest methods to develop seasonal and annual $PM_{2.5}$ concentration estimation models [101]. However, satellite remote sensing techniques are limited by cost, accuracy, and speed etc.

4.3.4 Image dehazing

Image dehazing and vision-based AQ estimation techniques are closely related since they both process the haze effect in images. Image dehazing aims to remove the haze effect and enhance image quality, while AQ estimation quantifies the degree of the haze effect.

He et al. proposes to use the dark channel prior (DCP) to estimate the transmission and atmospheric light in the haze images [78]. However, although the dark channel is widely used for daytime air quality estimation [87, 102], it may not be as effective for nighttime images due to low color intensities. Li et al. propose to separate the glow layer from nighttime images based on a smoothness prior [103]. However, this method is prone to noise and color shift problems in the resulting glow maps. Therefore, we develop our nighttime $PM_{2.5}$ estimation technique based on glow map.

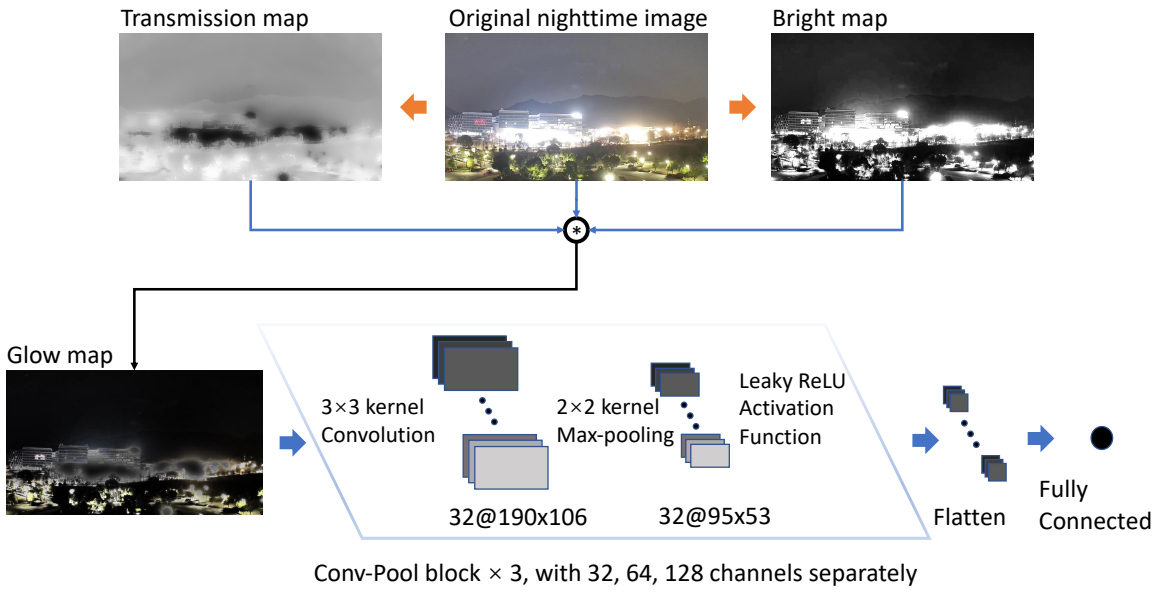


Figure 4.1: The architecture of the proposed method.

4.4 Methodology

In this section, we first present our nighttime haze imaging model. Then we describe how to extract haze-related features and the corresponding model to estimate $PM_{2.5}$ concentration. The flow of our system is shown in Figure 4.1.

4.4.1 Nighttime Haze Imaging Model

The nighttime haze imaging model is an extension of the optical daytime model [104], where the observed intensity at pixel x is modeled as a linear combination of the direct attenuation $\mathcal{D}(x)$ and the global atmospheric light or air light $\mathcal{A}(x)$ as following equation.

$$\begin{aligned} \mathcal{I}(x) &= \mathcal{D}(x) + \mathcal{A}(x) \\ &= \mathcal{J}(x)t(x) + A[1 - t(x)], \end{aligned} \tag{4.1}$$

where $\mathcal{I}(x)$ is the observed intensity at pixel x, $\mathcal{J}(x)$ is the scene radiance assuming no scattering particles, and A is the global atmospheric light constant. $t(x)$ is the transmission that indicates the portion of scenes reaching the camera. It is defined as in the following equation.

$$t(x) = e^{-\beta d(x)}, \tag{4.2}$$

where β is the atmosphere scattering coefficient and d is the scene depth. $\mathcal{J}(x)t(x)$ represents the direct transmission. $\mathcal{A}(x)$ is the air light indicating the particle veil induced by

the scattering, which varies with location.

In the nighttime haze scenario, besides the global source, i.e., moonlight, the air light is also intensified by other active light sources, which is very common in night scenes. Thus, it generates the glow region, which is defined as the regions close to light sources. The observed glow effect G can be modeled as the convolution of light source with an atmospheric point spread function (APSF) expressed by Legendre polynomial [91].

$$\mathcal{I}(x) = \mathcal{D}(x) + \mathcal{A}(x) + G \text{ and} \tag{4.3}$$

$$G = A_L(x) * APSF, \tag{4.4}$$

where A_L is the active light source and its intensity is convolved with APSF to derive the G. There are three factors involved with the glow generation, including active light sources (illumination and shape of light sources), properties of the scattering medium, and the scene depth or scattering distance. We discuss features related to these factors in the following sections.

4.4.2 Feature maps

In this subsection, we first introduce two different feature maps: the bright map and the transmission map. Then we describe how to combine them to generate the glow map.

4.4.2.1 Bright map

Various artificial light sources causes uneven distribution of nighttime atmospheric light. In that case, the color and intensity of glow regions can vary significantly depending on the distance from adjacent light sources. Thus, assumption of globally constant atmospheric light for the daytime scenarios is not applicable at night.

In an RGB image, pixels with high values across all three channels indicate the proximity to light sources. For the colored light sources, one or two channels may have significantly higher pixel values. The channel difference, which is defined as the gap between the maximum and minimum values of each color channel at every pixel position, is used to generate the bright map. By merging the maximum value and channel difference, we can derive the bright map as shown in the following equation.

$$B(x) = \min\left((2 * \max_{c\in\{r,g,b\}} \mathcal{I}^c(x) - \min_{c\in\{r,g,b\}} \mathcal{I}^c(x))^\gamma, 1\right), \tag{4.5}$$

where x is the pixel index, $\mathcal{I}^c(x)$ is the pixel value at x position on channel c, and γ is a rectified factor to modify the distribution of bright maps.

4.4.2.2 Transmission map

Transmission is defined as the proportion of reflected light that is not scattered and is capable of reaching the camera. The level of scattering in reflected light is affected by the distance between the camera and the scene as well as the density of haze. The correlation between the transmission $\tilde{t}(x)$ and scene depth $d(x)$ can be calculated using the following equation [78].

$$\tilde{t}(x) = 1 - \omega \min_{x' \in \Omega(x)} \left(\min_{c} \frac{\mathcal{I}^c(x')}{A^c} \right) \text{ and} \tag{4.6}$$

$$\beta = \frac{1}{d(x)} \ln \frac{1}{\tilde{t}(x)}, \tag{4.7}$$

where Ω is a local patch centered at x, ω is haze-retention constant (fixed at 0.95), A^c is global atmospheric light, and x' is the location index inside the patch. When $d(x)$ is fixed, the scattering coefficient β, i.e., haze level, is inversely correlated with the $\tilde{t}(x)$. We use guided image filtering [105] to reduce the halos and block artifacts introduced by the patch Ω; thus the transmission map can capture the edges of objects.

At night, the glow effect is negligible in regions distant from the light source, as expressed in Equation 4.1. However, in regions close to the light source, non-uniform illumination significantly enhances atmospheric light A, glow G, and restrains transmission $t(x)$, making scene radiance irrelevant, i.e., $\frac{t(x)}{A} \to 0$. By combining Equation 4.1 and Equation 4.3, we have

$$\frac{\mathcal{I}(x)}{A} = \frac{t(x)}{A}\mathcal{J}(x) + 1 - t(x) + \frac{G}{A}. \tag{4.8}$$

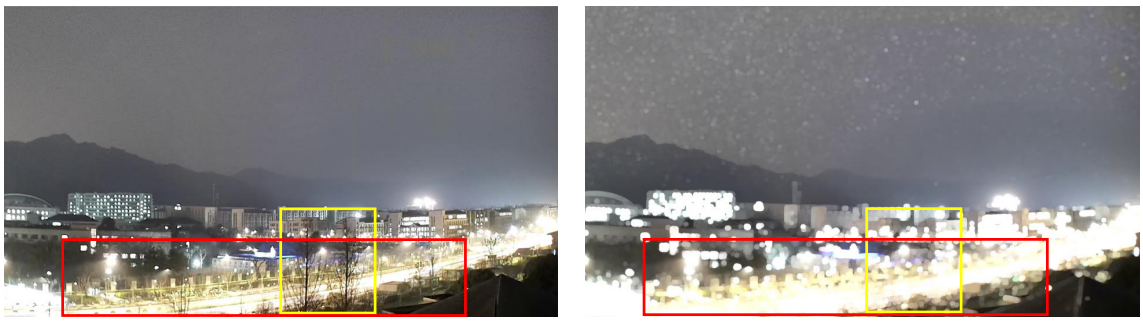
It can be further approximated as

$$\begin{aligned} \mathcal{I}(x) &\approx G + A(1 - t(x)) \\ &\approx G + a, \end{aligned} \tag{4.9}$$

where a is an air light related constant. Note that for an image patch, the scene depth and atmospheric light are assumed to be stable, which means A and $t(x)$ can both be approximated to constants.

4.4.2.3 Glow map

The glow effect originates from nighttime artificial light sources. In dense haze conditions, the brightness of the glow diminishes gradually as the distance from the light source increases, while the sharp edges of backgrounds become more discernible. Since the arti-



(a) before dilate operation (b) after dilate operation

Figure 4.2: Visualization of dilate function.

ficial light sources influence both the glow map and bright map, their attenuation patterns remain consistent. To extract the glow map, we utilize the bright map to identify regions of light sources, effectively reducing the impact of global air light and scene radiance.

It is also observed that the background visibility increases with distance from light sources in the transmission map since the transmission is essentially related to the proportion of scene radiance being perceived. They can be combined to reflect the glow intensity, i.e. the degree of radial scattering, and produce the glow map as follows.

$$G^c = \text{dilate}(x) \cdot t(x) \cdot B(x) \text{ and} \tag{4.10}$$

$$\text{dilate}(x) = \max_{x' \in \Omega(x)} \mathcal{I}^c(x'), \tag{4.11}$$

where $c \in \{r, g, b\}$ is the color channel, '$\cdot$' is pixel-wise multiplication, and $\Omega(x)$ is a circular structuring element centered on x with diameter of 15 pixels. We use the dilate function to enhance the glow effect since most glow regions contain foregrounds and backgrounds, such as tree branches. Figure 4.2 shows that the dilate function can successfully eliminate the foreground tree branch (yellow rectangle) and the background urban area (red rectangle).

As shown in Figure 4.3, the raw image contains the atmospheric light, light sources, and glow simultaneously. The bright map eliminates the air light in the sky and reserves the artificial light sources and their adjacent region. The transmission map locates the light source and restrains the air light effect of artificial light sources. The glow map demonstrates that only the regions with sufficient light intensity and smooth variation are identified as glow features. Figure 4.4 shows the pixel values at the red arrow locations in Figure 4.3. The red line in Figure 4.4 (a) is the glow intensity, which is in accordance with the visual change of glow. Figure 4.4 (b) shows that the variation of glow is consistent with the raw image.

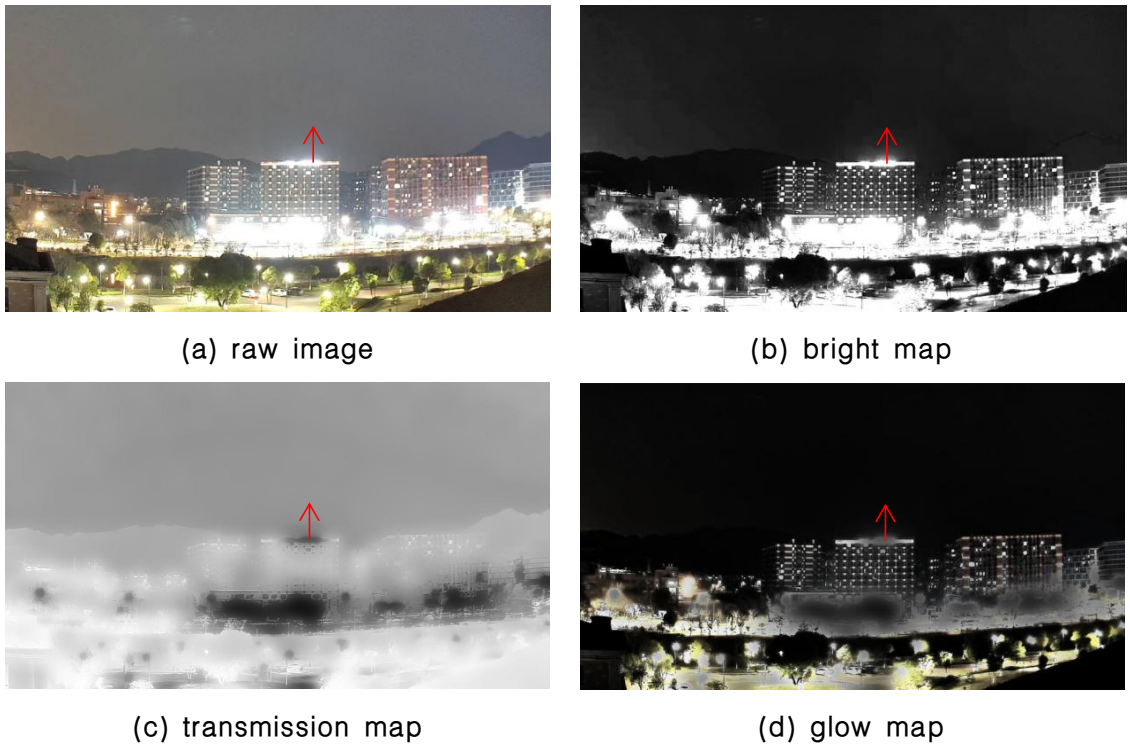


Figure 4.3: Visualization of feature maps.

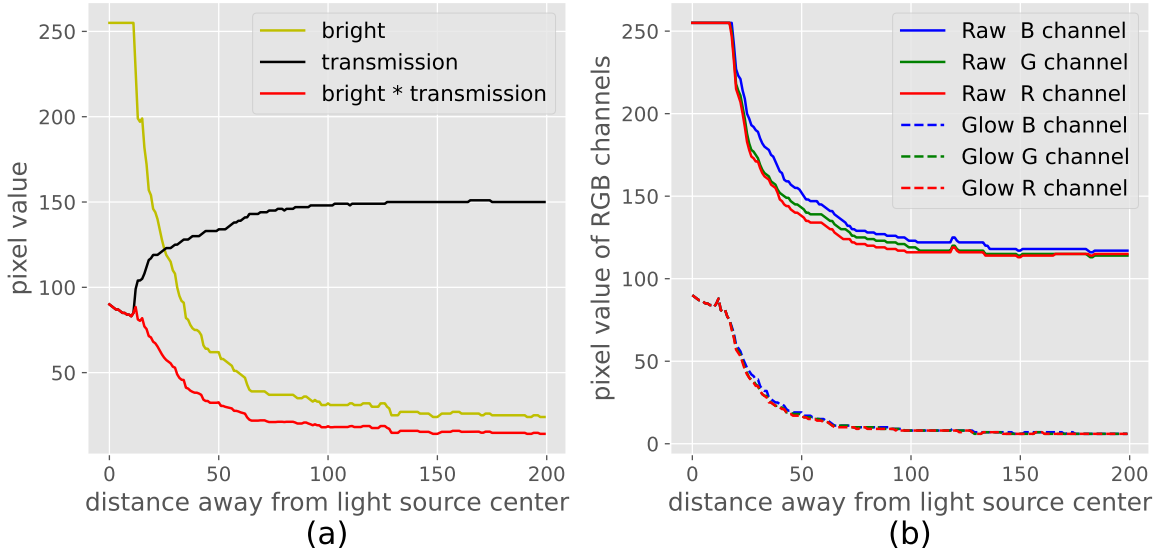


Figure 4.4: Visualization of pixel values along red arrow line in four images. The x-axis is the number of pixels away from the arrowhead along the vertical direction and the y-axis is the corresponding pixel value.

The difference between the raw image and the glow map is the constant a introduced in Equation 4.9.

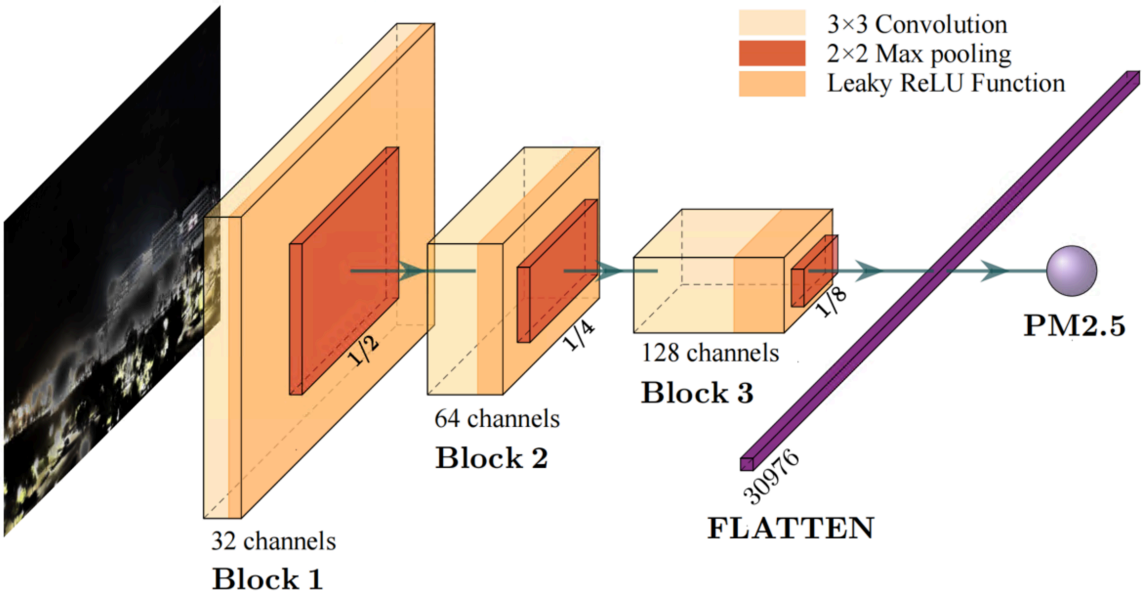


Figure 4.5: Architecture of CNN.

4.4.3 Mapping algorithm

CNNs have been widely used for image processing and classification. In this work, we employ the CNN technique to map the relationship between $PM_{2.5}$ and the glow map. As shown in Figure 4.5, the extracted glow map is resized to $108 \times 192 \times 3$ to reduce the input dimension and also computation cost. For the CNN, we set convolutional kernel size to be 3×3. In the subsequent pooling layer, 2×2 max pooling is used. We also use the leaky ReLU function as the activation function to avoid the dying ReLU problem [106]. This Conv-Pool block is repeated three times consecutively. Finally, the flattened vector is input to a fully connected layer to generate the concentration of $PM_{2.5}$.

4.5 Data processing

To validate our algorithm, we first build the corresponding dataset. Therefore, in this section, we detail the setting of our dataset, the deployment of our sensor network, and the data analysis methods.

4.5.1 Overview

The dataset consists of 11753 nighttime images and the corresponding ground truth sensor data collected from Nov. 2022 to Mar. 2023 in the west Hangzhou city. The longitude and latitude are $120°02'49.631''$E, $30°13'56.022''$N, respectively. Particulate matter is the main

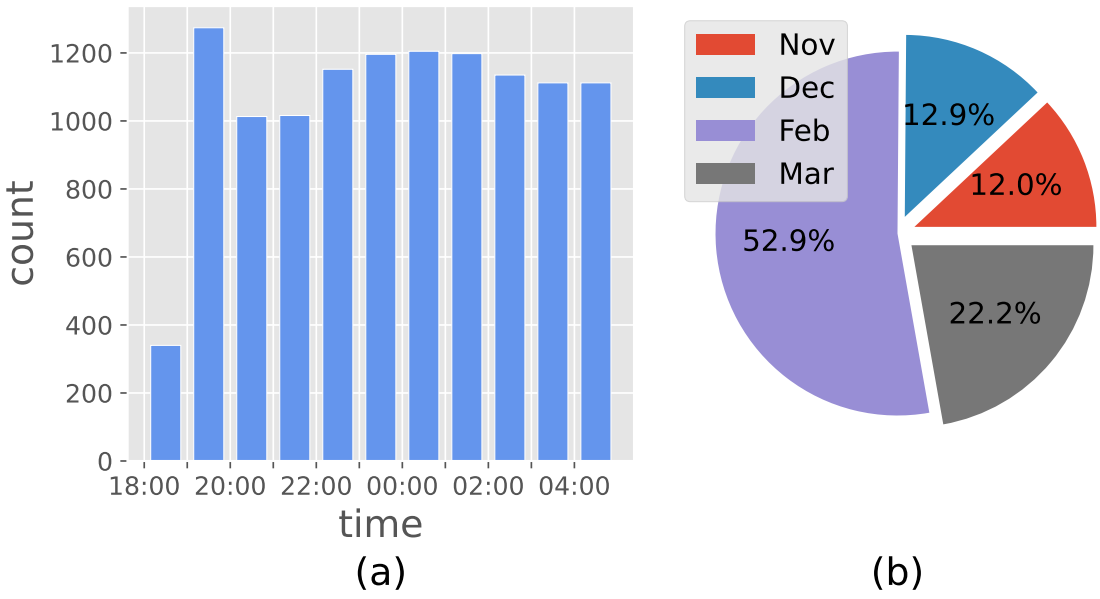


Figure 4.6: The time distribution of our dataset.

local atmospheric pollutant at night [107]. Missing data on certain dates are mainly due to weather conditions. The dataset is used to validate our algorithms. It is also made public for other researches in the field.

The image dataset contains frames captured every 5 minutes, with a 1080×1920 resolution. Due to the uneven temporal occurrence distribution of high $PM_{2.5}$ pollution in Hangzhou, we increase the sampling rate to one per minute for certain period of time. For example, the rate is increased from 06:20 pm to 08:00 pm on Feb. 19, during which the $PM_{2.5}$ concentration is between $58\,\mu g/m^3$ and $119\,\mu g/m^3$.

The overall distribution of image capture time is shown in Figure 4.6. The images are taken between approximately 06:20 P.M. and 05:00 A.M. when the sun is set and the illumination is low. The experiment lasts for 5 months (from November to March next year). Figure 4.7 shows the $PM_{2.5}$ concentration distribution during the same period.

The majority of images were captured in February, with the remaining collected in November, December, and March. it exhibits a short-tailed pattern, indicating that only a few readings exceed $80\,\mu g/m^3$. To address this imbalance, we use resampling technique as described in Section 4.6.1.

4.5.2 Data collection

To collect the ground truth $PM_{2.5}$ readings and the corresponding images, we deploy sensor and camera networks in downtown Hangzhou, with many residential communities and

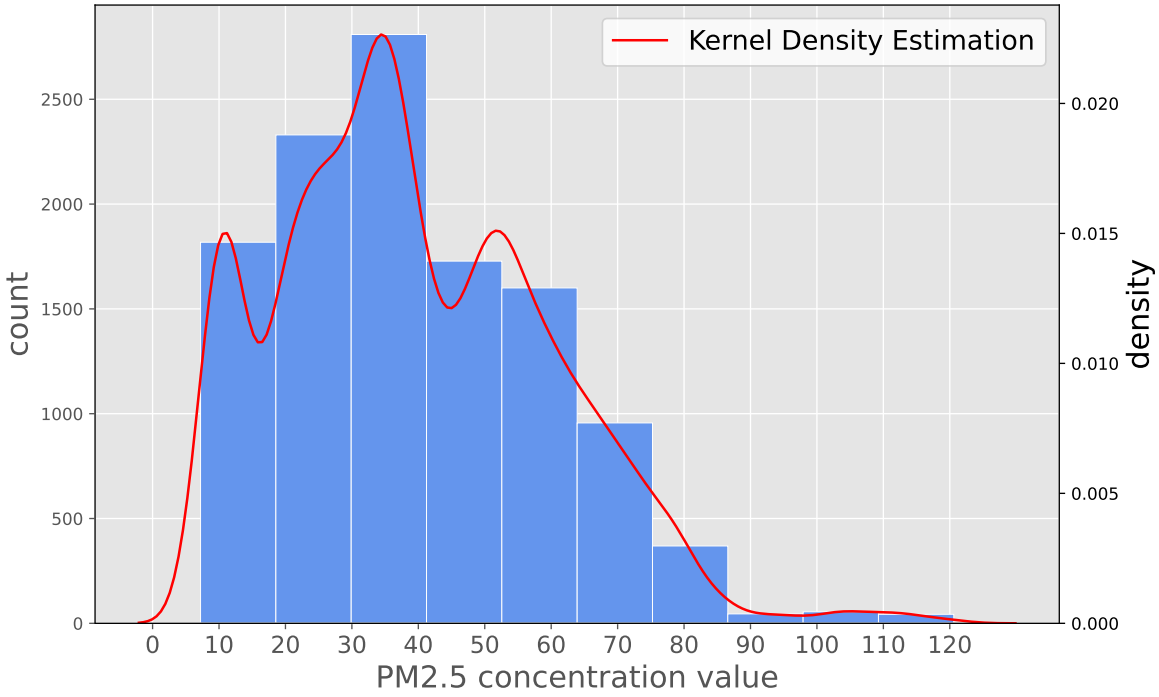


Figure 4.7: The $PM_{2.5}$ distribution of our dataset.

schools. The cameras are placed along a main city road, as demonstrated in Figure 4.8. Meanwhile, an air quality sensing platform is placed in the center of the region to record ground truth pollutant readings.

We develop a portable sensing platform to collect and store data, including a Nova PM sensor module, temperature and humidity sensors, and an Arduino UNO controller (see Figure 4.9). This platform is powered using lithium batteries and records data once per second. Details regarding the parameters of our sensors is in Table 4.1 and Table 4.2.

We employ five XiaoMi intelligent cameras to capture videos of the monitored area (see Figure 4.9). Further details regarding the camera parameters can be found in Table 4.2. The key camera parameters are len and aperture sizes. Increasing those sizes allows effective monitoring of large areas in low-light conditions. Each camera covers an approximate area of $2\,\mathrm{km}^2$, while a network of air quality sensors can only cover about $500\,\mathrm{m}^2$ [86, 108]. It is important to note that we utilize off-the-shelf standard camera modules, making our technique highly suitable for real-world applications. These cameras are strategically positioned within buildings situated amidst schools and residential communities, spanning a distance of approximately 4.8 kilometers along the main road. Detailed information about the specific camera locations can be found in Table 4.1 and Figure 4.8. We adjust the camera depression angles to capture comprehensive scenes including both the ground and the sky.

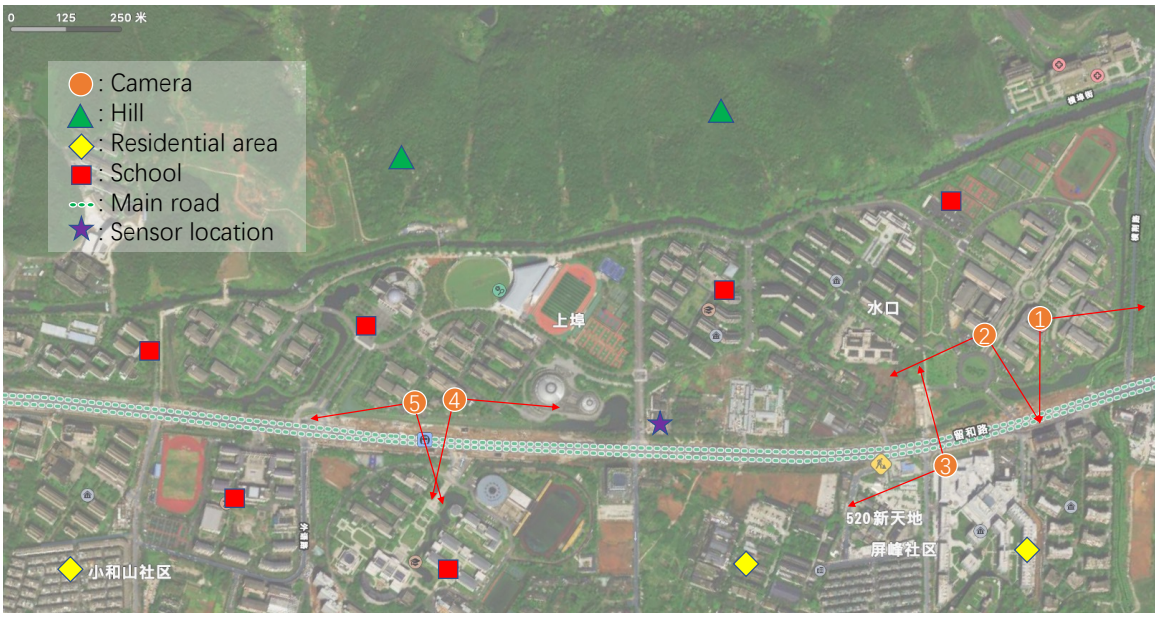


Figure 4.8: Sensor and cameras locations. The cameras directions are marked by red arrows.

Table 4.1: Deployment specifications for environmental sensors and cameras

Device	**GPS Location**	**Height (m)**	**Depression Angle(°)**	**Acquisition interval**
Camera1	120°03′16.488″E 30°14′15.946″N	18	20	5 minutes
Camera2	120°03′12.236″E 30°14′12.787″N	18	10	5 minutes
Camera3	120°03′21.372″E 30°14′8.083″N	33	5	5 minutes
Camera4	120°02′38.086″E 30°13′47.709″N	18	20	5 minutes
Camera5	120°02′34.189″E 30°13′45.393″N	15	5	5 minutes
Sensors	120°02′53.592″E 30°13′54.323″N	18	-	1 second

Table 4.2: Parameters of PM Device and Cameras

Nova PM sensor :	
Sensor range	[$PM_{2.5}$] 0.0 - $999.9\,\mu g/m^3$
	[PM_{10}] 0.0 - $1999.9\,\mu g/m^3$
Operating temperature	-10-50°C
Maximum operating humidity	70%
The response time	1 second
Serial port data output frequency	1 Hz
Minimum resolution particle size	$0.3\,\mu m$
The relative error	Max. ±15% and ±10$\mu g/m^3$
	(Note: $25\,^{\circ}C$,50%RH)
Standard certification	CE/FCC/RoHS
Camera :	
Lens	FOV: 110°
	Aperture: f/1.4
	Shooting Range: 0.6 m to ∞
Video Resolution	1080×1920, MP4 (H.265/HEVC)
Operating temperature	-10 - $45\,^{\circ}C$

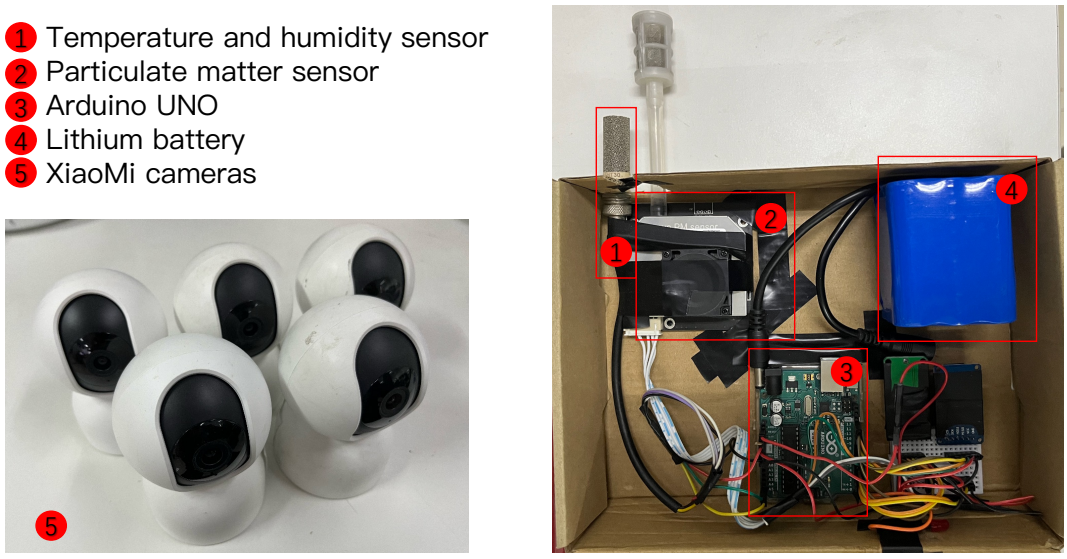


Figure 4.9: Sensing platform and cameras used in our deployment.

Table 4.3: Correlation with environmental factors

	$PM_{2.5}$	PM_{10}	Temperature	Humidity
$PM_{2.5}$	1.0	0.758	0.325	0.349
PM_{10}	0.758	1.0	0.253	0.302
Temperature	0.325	0.253	1.0	0.787
Humidity	0.349	0.302	0.787	1.0

4.5.3 Nighttime observations

In theory, the formation, propagation, and dissipation of $PM_{2.5}$ are affected by climates and weather conditions. For instance, low temperatures during the night can cause pollutants to stay closer to the ground. However, the influence of those environmental factors on PM concentration is limited. We calculate the R^2 correlation coefficients for the two environmental factors and discover that the correlations are weak, as demonstrated in Table 4.3.

Air pollution also shows significant temporal variation throughout the day [109]. Especially, the decreased human activities during the night can lead to reduced pollution level. The drop in temperature during the night can also cause atmospheric pollutants to be trapped and accumulated.

To assess the differences in the rate of change of $PM_{2.5}$ between day and night, we use both absolute and relative rates of variation. Since a portion of our $PM_{2.5}$ data does not cover a full 24-hour period due to various reasons, we select 13 days which contains 24-hour data and calculate their average. Assume the $PM_{2.5}$ data for day t is p_t. The absolute

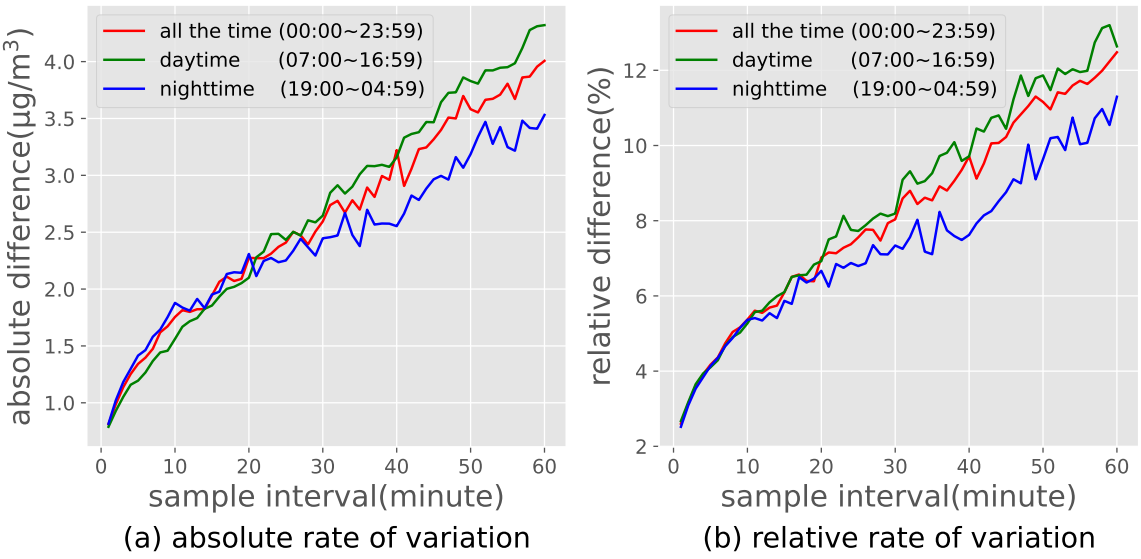


Figure 4.10: Diurnal difference in $PM_{2.5}$ variation rate.

rate of variation v_a and the relative rate of variation v_r are calculated as follows.

$$v_a = \frac{1}{N-1}\sum_{t=1}^{N-1}|p_{t+1} - p_t| \text{ and} \tag{4.12}$$

$$v_r = \frac{1}{N-1}\sum_{t=1}^{N-1}|p_{t+1} - p_t|/p_t, \tag{4.13}$$

where $N = \frac{A}{s}$ represents the length of the sequence, A is the number of data points covering 24 hours, s is the sampling interval, and p_t denotes the $PM_{2.5}$ concentration at the t-th sampling point.

The results are presented in Figure 4.10, which clearly demonstrates that the level of $PM_{2.5}$ exhibits significantly reduced variation during nighttime compared to daytime. This finding suggests that the $PM_{2.5}$ models developed for daytime estimation may not be suitable for accurate estimation during nighttime.

4.5.4 Glow effect

We first design an algorithm, as shown in Algorithm 2, to determine the attenuation rate of light intensity near the light source, which is closely related to the $PM_{2.5}$ concentration. Specifically, we put a 101×101 image patch $P(\boldsymbol{x})$ in a Cartesian coordinate system and $\boldsymbol{x} = (x_1, x_2)$ is the pixel coordinate. The algorithm identifies the light source in $P(\boldsymbol{x})$ and returns a binary mask $M(\boldsymbol{x})$ that highlights the pixels identified as light source and its

Algorithm 2 Calculation of Light Source

Data: 101×101 image patch $P(\boldsymbol{x})$
Result: light source binary map $M(\boldsymbol{x})$ and central coordinate $\boldsymbol{c}$

Init 1: $\boldsymbol{c} \leftarrow$ the mean coordinate of the pixels with [255,255,255] value in the $P(\boldsymbol{x})$ Init 2: $M = O_{101\times101}$, $M(\boldsymbol{c}) = 1$ Step 1: Use Euclidean distance to pixel $\boldsymbol{c}$ to resort pixels in $P(\boldsymbol{x})$ as $P_{resort}(\boldsymbol{x})$ **for** p in $P_{resort}(\boldsymbol{x})$ **do**

    Step 2.1:
    if p == [255,255,255] and the Hamiltonian distance between its coordinate and the nearest light source pixel in M is 1 **then**
        $M[coordinate(p)] = 1$; // identified as light source
    end
    Step 2.2:
    if Less than 20 of the last 200 pixels processed are identified as light sources **then**
        break
    end

end

Step 3: $\boldsymbol{c} \leftarrow$ the mean coordinate of the pixels identified as the light source in M return $M, \boldsymbol{c}$

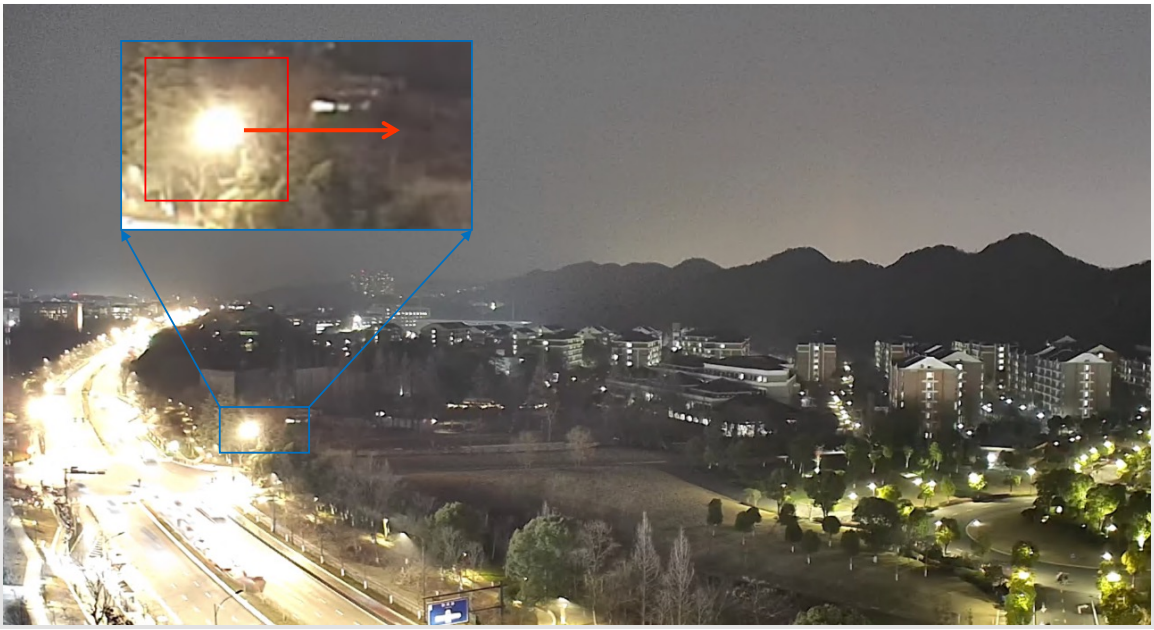


Figure 4.11: Light source of location 3 (see Figure 4.8). The red arrow represents the transmission direction for attenuation analysis.

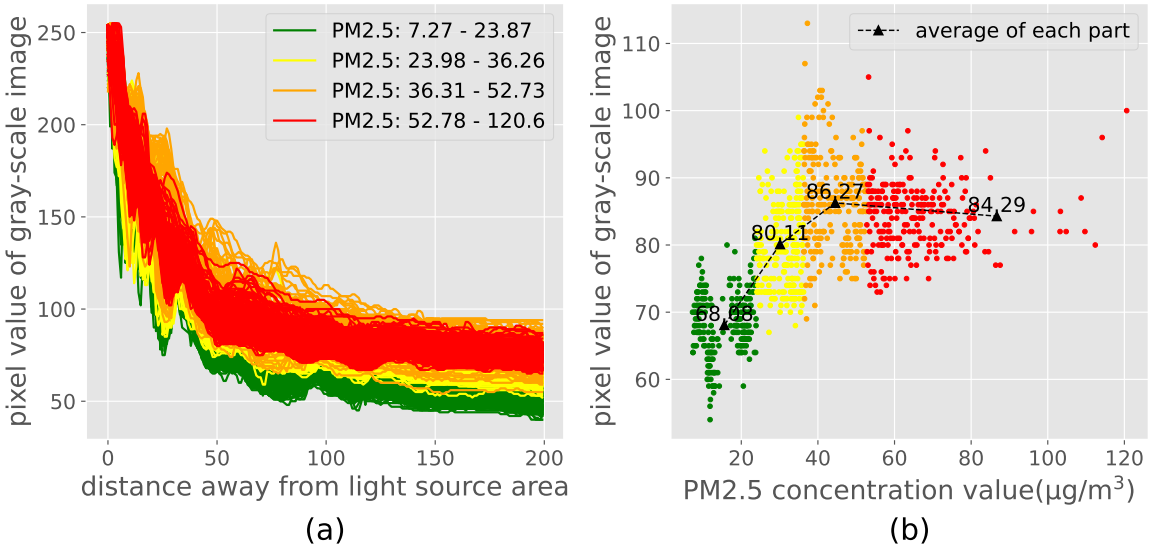


Figure 4.12: Attenuation of the glow region: (a) The x-axis is the number of pixels away from the light source along the horizontal direction and the y-axis is the corresponding pixel value; (b) The x-axis is $PM_{2.5}$oncentration value of each line and the y-axis is pixel value at 100-pixels away from light source area.

central coordinate $\boldsymbol{c}$.

We use the light source located in Figure 4.11 to evaluate the correlation between the glow effect and $PM_{2.5}$ concentration. 1000 images are selected randomly and divided into four sets based on their concentration range. The region with a gray-scale pixel value of 255 is considered the light source. Then we designate the region 200 pixels away from the edge of the light source as the glow region. We plot the pixel values as a function of distance away from the edge of the light source along the red arrow. As shown in Figure 4.12 (a). Each line represents an image, and its color indicates the corresponding $PM_{2.5}$ range. There is an obvious layering by color, indicating that the pixel values decreased more rapidly with a decrease in $PM_{2.5}$ concentration, thus confirming the correlation between the glow effect and $PM_{2.5}$ concentration. Specifically, take pixels at 100 pixels away from the light source (along the horizontal direction) as an example, Figure 4.12 (b) shows that the average pixel value increasing rate is 0.45 pixel units every $\mu g m^{-3}$ when the concentration is lower than $50\,\mu g/m^3$. Note that the red part does not meet this trend, possibly due to the exposure adjustment of intelligent cameras given the increased brightness due to increased PM. Moreover, the most significant attenuation in pixel values occurred within a distance of 100 pixels. This guides our selection of hyperparameters in later experiments.

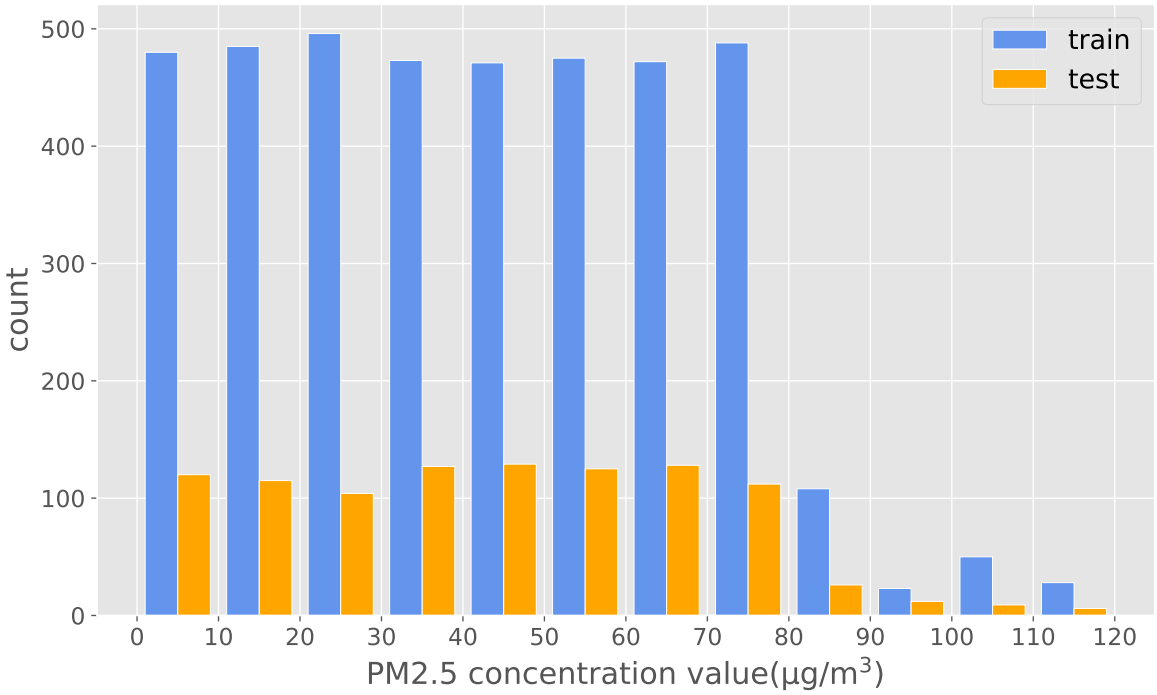


Figure 4.13: The data distribution of our dataset. There are 534 samples from highly polluted environments ($PM_{2.5}$ $>80\,\mu g/m^3$).

4.6 Experimental results

In this section, we introduce the experimental setup and the corresponding experimental design and results.

4.6.1 Experimental setup

We randomly select 5334 images from our dataset and divide them by a ratio of 4:1 for training and testing sets. Since high concentration samples are scarce ($PM_{2.5}$ $>80\,\mu g/m^3$), we replicate high-concentration samples in the training set. Figure 4.13 shows the sample distribution of the training set and testing set. The algorithm is developed using PyTorch (version 1.11.0, CUDA 10.2), Adam optimizer with a batch size of 128 [110], and mean squared error loss function. We employ multi-step learning rate decay strategy, where the learning rate is initialized to $5\,e^{-3}$ and stepped down to $1\,e^{-5}$ in 400 epochs.

The model is trained on a server equipped with a Tesla V100 GPU and Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz. The mean absolute error (MAE) and mean relative error (MRE) are used as the evaluation metrics. Their calculations are shown in the following

Figure 4.14: A example of images at different times from the evening of March 4 to the early morning of March 5.

equations.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \text{ and} \tag{4.14}$$

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{4.15}$$

where N is the total number of samples, $\hat{y}_i$ is the predicted value of i-th sample, and y_i is the ground truth value.

4.6.2 Relationship between glow and traffic

As depicted in Figure 4.14, we observe that the overall color of certain images gradually transitions from yellow to white over time. This phenomenon can be attributed to the properties of aerosol light absorption. Aerosols, particularly black carbon (BC), which comprises elemental carbon (EC) and organic carbon (OC), play a dominant role in absorbing light in the atmosphere [111]. EC is present in both traffic emissions and $PM_{2.5}$ in urban areas and can significantly influence the visual perception, especially during nighttime. Therefore, a decrease in the number of vehicles on the road during specific hours (07:00 P.M. - 05:00 A.M.) can reduce EC concentraion [112] and the corresponding level of light absorption.

Figure 4.15 shows the brightness of the light source area in different months. The x-axis

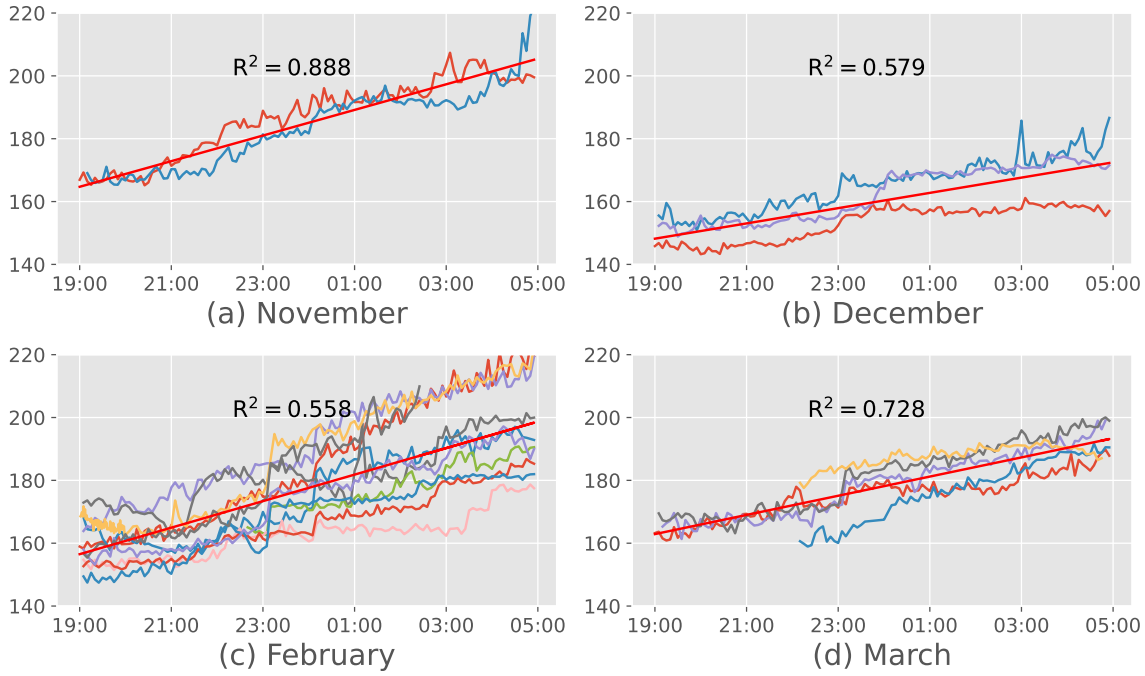


Figure 4.15: The variation of brightness in the light source area.

is night timeline and the y-axis is the average pixel values in the light source area. Each line represents the brightness variation of one night. We define the light source area as the 101×101 block centered on the light source and the brightness as the average pixel values in the area. We observe that brightness increases over time as expected.

4.6.3 Evaluation of glow map based model

We compare our technique with state-of-art techniques with the same CNN architecture and different input types. The comparing methods include glow map [103], dark channel, and transmission [78]. As shown in Table 4.4, the MAE and MRE of our method are better than other methods that take in other inputs, indicating the glow map can accurately reflect the density of haze. Compared with the second-best method, our technique is improved by 21.3% on MRE.

The effectiveness of the transmission map-based method is diminished during nighttime [113], as it tends to mistakenly identify artificial light sources as light sources at infinity. This issue arises due to the uneven distribution of brightness caused by low illumination and artificial light sources. Consequently, many regions in the transmission map exhibit pixel values that are either close to 0 (black regions) or 255 (white regions). This disrupts the overall relationship between transmission and scene depth, resulting in poor transmission map performance.

Table 4.4: Comparison of different features using our model

Method	MAE ($\mu g/m^3$)	MRE
Our Glow Map	**2.58**	**8.18%**
Raw Image	3.69	10.92%
Dark Channel [78]	3.65	10.82%
Li's Glow Map [103]	4.18	14.86%
Bright Equation 4.5	3.28	9.92%
Transmission [103]	5.23	14.91%
Bright+Transmission+Glow	4.11	12.29%

On the other hand, the dark channel-based method seeks the minimal value of RGB channels within a patch to generate a coarse output. However, this approach neglects a significant amount of valuable image information. Li's glow map approach addresses the color shift problem by imposing a global-based RGB channel constraint and extracting a smooth layer using a spectrum-based algorithm. However, when dealing with large-sized images containing multi-color lights, unexpected color shifts can occur, and object edges may appear in the glow map.

Compared with Li's glow map method, which relies on statistical prior knowledge of images, our glow map estimates the proportion of radially scattered light in a raw image based on physical imaging model. Experimental results demonstrates its performance in$PM_{2.5}$ estimation.

We also compare our CNN based technique with traditional machine learning models, using the same glow map input. We split the glow map into a 9×16 grid and calculate the average for each region. The resultant feature vector is used as input. As shown in Table 4.6.3, our technique is compared with six baselines. Both MAE and MRE of our technique have the best performance, indicating our model can better fit the relationship between the glow map and corresponding $PM_{2.5}$ concentration. Compared with other methods, our model achieves at least 8% improvement for MRE and MAE.

4.6.4 Receptive field

For classic deep learning models such as VGG16 [114], the receptive field of a neuron increases with its depth. However, in our case, attenuation of the light intensity is within 100 pixels. Therefore, we envision that the receptive field of the neuron in the final layer may contain an area of approximately 100×100 pixels in the input image.

To test this hypothesis, we train models with varying input sizes and evaluate each

Table 4.5: Comparison of different models using our glow map

Model	MAE ($\mu g/m^3$)	MRE
Ours	**2.58**	**8.18%**
Bayesian Ridge	6.03	20.95%
Linear Regression	6.12	21.60%
Elastic Net	8.72	29.64%
Support Vector Regression	14.49	62.71%
Gradient Boosting Regression	5.55	18.14%
Random Forest	2.83	8.87%

Table 4.6: Comparison with different input size

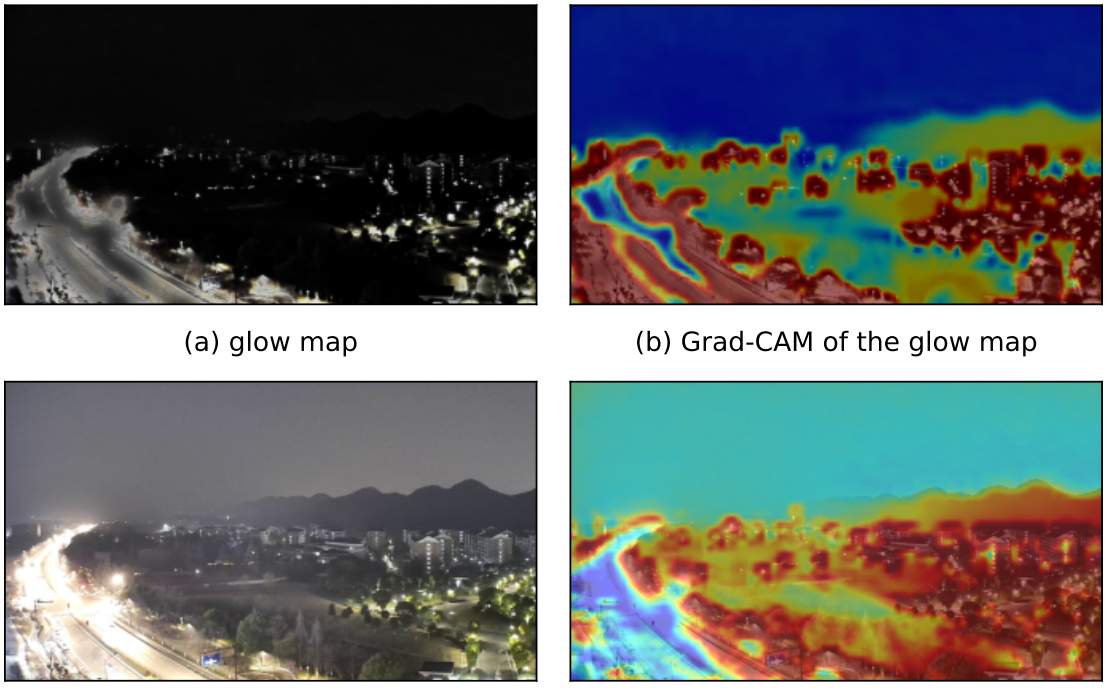
Input size	Receptive field	MAE	MRE	Flops	Parameters
(252,448)	(37, 35)	3.88	12.16%	1.09G	293,697
(216,384)	(43, 41)	3.58	12.08%	789.13M	240,449
(180,320)	(54, 50)	3.51	10.98%	539.76M	190,529
(144,256)	(67, 64)	2.95	8.97%	337.61M	154,689
(108,192)	**(98, 87)**	**2.58**	**8.18%**	**182.7M**	**124,225**
(81,144)	(135, 120)	3.40	10.37%	96.28M	109,633
(54,96)	(216, 192)	3.22	9.87%	39.64M	99,649
(27,48)	(1080, 480)	4.80	15.00%	6.79M	93,761

case, as shown in Table 4.6. The results indicate that the best receptive field is with an input size of $(108, 192)$ and a corresponding receptive field of $(98, 87)$, which is close to our hypothesis. Increasing the input resolution beyond that point leads to a decrease in the receptive field and an increase in MAE and MRE. Moreover, decreasing the input resolution causes a loss of image details and also a deterioration in network accuracy.

4.6.5 Sky region impact

The sky is crucial in estimating daytime $PM_{2.5}$ concentration in outdoor environments [88]. Typically, the daytime airlight is assumed to be the intensity of the sky region at an infinite distance. However, the nighttime sky region is not applicable due to low illumination and cloud.

To investigate the impact of the sky region, we partition the image into two regions: the sky region and the non-sky region, respectively. As shown in Table 4.6.5, training the model using only the sky region results in a significantly larger MRE compared with the non-sky one. Removing the nighttime sky region have little impact on the overall model



(a) glow map (b) Grad-CAM of the glow map

(c) raw image (d) Grad-CAM of the raw image

Figure 4.16: Grad-CAM of the first convolution layer.

Table 4.7: Evaluation of different region

	sky region	non-sky region	sky + non-sky
MAE($\mu g/m^3$)	13.21	3.04	**2.58**
MRE	54.55%	9.46%	**8.18%**

accuracy.

We then use gradient-weighted class activation mapping (Grad-CAM) [115] to visualize the network attention on input images. Grad-CAM generates a heatmap of each pixel's contribution to the final prediction by utilizing back propagation gradients. Since the resolution of the Grad-CAM heatmap is consistent with the target hidden layer, we select the first convolution layer to highlight the fine-grained details of the image. As shown in Figure 4.16, the network focuses its attention on the glow of light sources. In comparison to the raw image, the glow map pays less attention to the non-light source region, airlight, and background while focuses more on the glow itself.

4.7 Conclusion and discussions

In this work, we propose a nighttime $PM_{2.5}$ concentration estimation technique based on light source glow map extraction method. We also build a corresponding dataset including various related parameters such as $PM_{2.5}$ concentration, humidity, and temperature etc. The experimental results demonstrate that our glow map-based approach produces a 38.3% improvement in the accuracy for MAE (i.e., from 4.18 to 2.58) compared to the state of art.

During the deployment, due to the relatively low pollution level in Hangzhou (compared to the rest of China), most $PM_{2.5}$ values are under $80\,\mu g/m^3$. Moreover, rainy weather also prevented the deployment. Our method mainly works at night and requires artificial light sources dominating the scene. Therefore, it may not work at dawn or dusk when sunlight dominates the ambient light.

Future work includes incorporating other environmental parameters, such as wind and rain, to improve the accuracy of $PM_{2.5}$ concentration estimation. Moreover, collecting a more diverse and larger-scale nighttime image dataset can further improve the accuracy of our method.

CHAPTER 5

Image-Based Air Quality Forecasting through Multi-Level Attention

5.1 Introduction

Past work developed data-driven models for time-series forecasting of air quality. For example, researchers designed a dual-stage attention model for time series prediction [116]. Also, deep neural networks are used to combine multiple sources of data such as weather and geo-context data for $PM_{2.5}$ forecasting [117, 118]. However, researchers have yet to consider image-based air quality forecasting.

Images have the ability to capture the level of atmospheric scattering and absorption due to airborne particles [119, 120]. In particular, images can be valuable for air pollution forecasting since cameras and webcams are less expensive and easier to maintain than most commonly used air quality sensors. Also, cameras capture a large amount of data over large spatial regions, whereas air pollution data are now commonly collected by sparsely distributed single-point monitoring stations. By augmenting $PM_{2.5}$ data with images, we can better estimate the air quality at a particular location and use contextual information to achieve better forecasts.

Past research includes numerous image-based haze detection techniques [13, 121], but they do not capture the complex spatio-temporal correlations of haze in images over time. Our objective is to forecast $PM_{2.5}$ concentrations by fusing $PM_{2.5}$ concentrations with colocated images, which requires spatio-temporal analysis of air quality in the images. In this chapter, we jointly use a convolutional neural network (CNN) and a long short-term memory (LSTM) to model the level of haze in the images over time.

It is necessary to learn intricate relationships between the images and the $PM_{2.5}$ data. Inspired by the success of attention networks in low-level computer vision [122], our method incorporates spatial attention, which learns the image regions to focus on, and feature attention, which learns the importance of each feature extracted from the image. Spatial

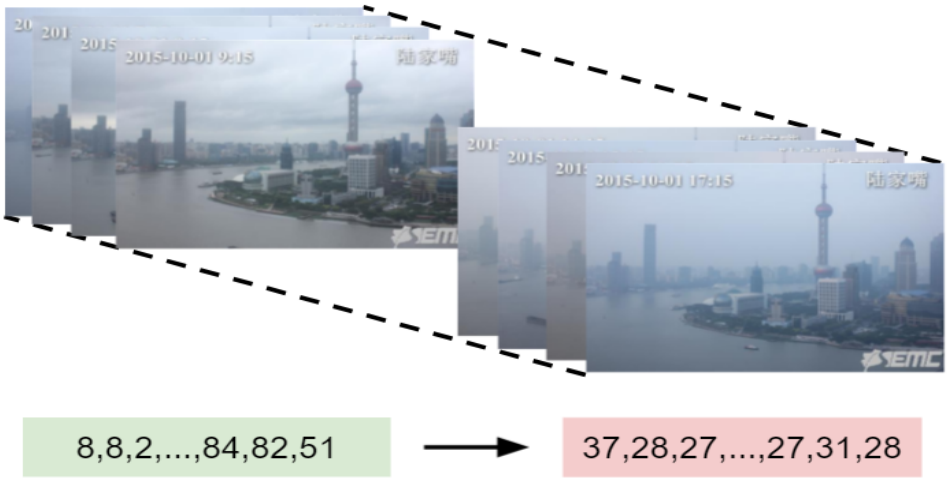


Figure 5.1: Image-based air quality forecasting uses a sequence of images and past $PM_{2.5}$ concentrations (left, green box) to forecast future $PM_{2.5}$ concentrations (right, red box).

attention selects the regions based on their similarity with $PM_{2.5}$ latent features. We hypothesize that spatial attention can improve predictions by identifying image regions with $PM_{2.5}$ concentrations that are better correlated with the ground-truth sensor location.

We evaluate our model on Shanghai data containing hourly $PM_{2.5}$ measurements and webcam images, and experimentally compare our forecasting model's accuracy with that of previous forecasting models. The main contributions are:

1. defining and solving the image-based air pollution forecasting problem,
2. developing a forecasting model capturing the level of haze in images over time with a combined CNN and RNN, which is novel in this context, and
3. incorporating multi-level attention to learn intricate relationships between images and the $PM_{2.5}$ data.

5.2 Background

Air pollution exhibits complex spatio-temporal correlations, stemming from the intricate interplay between various factors, including emission sources, atmospheric conditions, and meteorological dynamics. Understanding and capturing these correlations are useful for accurate air pollution forecasting and effective mitigation strategies. Let us delve into the nature of these correlations in more detail.

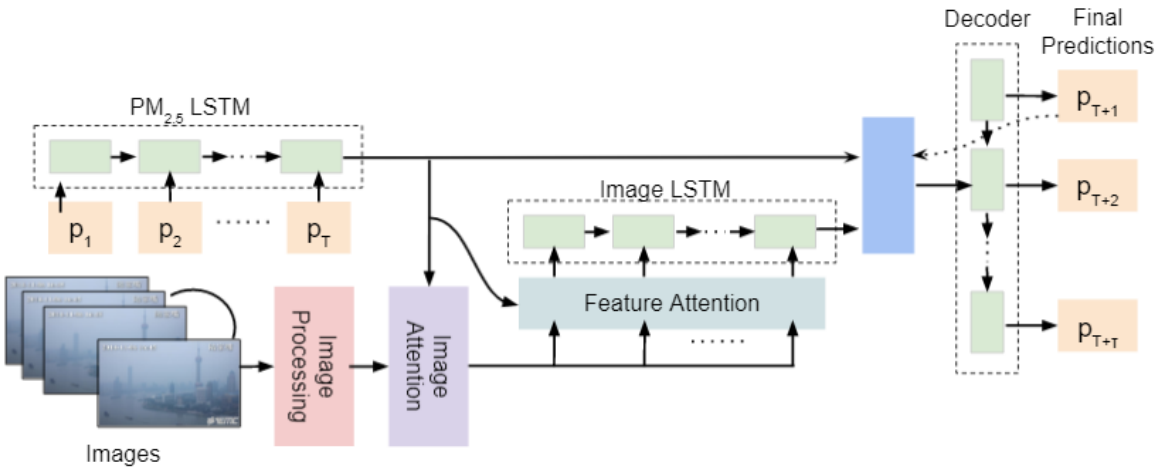


Figure 5.2: Image-based air quality forecasting model overview.

Spatio-temporal correlations refer to the relationships between air pollution levels across different locations and over varying time scales. In the spatial domain, pollutants disperse and interact with the surrounding environment, leading to spatial patterns and gradients in pollution concentrations. Localized emission sources, such as industrial complexes or urban traffic, contribute to localized pollution hotspots, characterized by elevated pollutant levels in close proximity to the sources. Furthermore, the terrain, land use patterns, and meteorological conditions influence the spatial distribution of pollutants, causing variations in pollution concentrations across different regions. Understanding and modeling these spatial relationships are important for identifying pollution sources, assessing exposure risks, and implementing targeted pollution control measures.

In the temporal domain, air pollution exhibits intricate dynamics due to diurnal, seasonal, and long-term variations. Diurnal patterns arise from human activities and natural processes that follow daily cycles. For example, traffic congestion during morning and evening rush hours can lead to peak pollution levels in urban areas, while nighttime cooling and reduced emissions often result in lower pollution levels. Seasonal variations are driven by changes in weather patterns, vegetation growth, and human activities. For instance, in many regions, air pollution is more pronounced during winter due to increased energy consumption and unfavorable meteorological conditions for pollutant dispersion. Long-term trends capture gradual changes in air pollution levels over extended periods, reflecting the effects of evolving emission regulations, urban development, and environmental policies.

The spatio-temporal relationships of air pollution also stem from the transport and dispersion of pollutants through the atmosphere. Pollutants emitted in one location can be transported by wind over long distances, affecting air quality in remote areas. Atmospheric

conditions such as temperature inversions, wind speed, and stability influence the vertical mixing and horizontal transport of pollutants, leading to the formation of pollution plumes and the spread of pollutants across different regions. These transport processes introduce temporal lags and dependencies, as pollutant concentrations at a particular location can be influenced by emissions occurring hours or even days earlier in other locations.

Moreover, the spatio-temporal correlations of air pollution interact with meteorological dynamics. Weather conditions, including temperature, humidity, precipitation, and wind patterns, play a significant role in pollutant dispersion, chemical reactions, and pollutant transformation. For example, high temperatures and sunlight can enhance the photochemical reactions that lead to the formation of secondary pollutants, such as ozone. Wind patterns determine the direction and speed of pollutant transport, affecting the spatial distribution of pollution. Additionally, atmospheric stability and mixing height influence the vertical mixing of pollutants, impacting their concentration profiles across different altitudes.

Capturing these complex spatio-temporal correlations requires sophisticated modeling approaches that consider the interplay of emission sources, atmospheric conditions, meteorological dynamics, and pollutant transport. Machine learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in leveraging the spatial and temporal information encoded in air pollution data and environmental variables to capture and predict these correlations. By understanding and modeling the spatio-temporal dynamics of air pollution, we can gain valuable insights into pollution sources, assess exposure risks, develop effective pollution control strategies, and ultimately work towards improving air quality and safeguarding public health.

5.2.1 Challenges

Modeling spatio-temporal correlations in air pollution presents several challenges:

Data Availability and Quality: Acquiring high-quality and comprehensive data on air pollution levels across different locations and time intervals can be challenging. Spatial coverage of monitoring stations may be limited, leading to gaps in data representation. Moreover, temporal resolution might not capture the fine-scale variations of pollution levels, hindering the accurate modeling of spatio-temporal correlations.

Complex Interactions: Air pollution is influenced by a wide range of factors, including meteorological conditions, emission sources, topography, and human activities. The interactions between these factors are complex and non-linear, making it challenging to capture their combined effect accurately. Modeling spatio-temporal correlations requires

sophisticated techniques that can handle the multi-dimensional and dynamic nature of the data.

Scale Discrepancies: Air pollution patterns can exhibit significant variations at different spatial and temporal scales. Modeling spatio-temporal correlations necessitates addressing scale discrepancies and finding an appropriate balance between capturing fine-grained local variations and understanding broader regional or global patterns. Failure to account for scale dependencies can lead to biased predictions and inaccurate assessments of pollution levels.

Uncertainty and Noise: Air pollution data are inherently noisy and subject to various sources of uncertainty. Measurement errors, sensor biases, and missing data can introduce uncertainties in the collected data. Additionally, the complex and dynamic nature of air pollution makes it challenging to separate the signal (correlated pollution patterns) from the noise (random variations). Modeling techniques must be robust enough to handle uncertainties and noise to obtain reliable spatio-temporal correlations.

These challenges can be addressed by advanced modeling techniques, such as machine learning algorithms, deep learning architectures, and spatial statistical models. These approaches can effectively capture the spatio-temporal correlations in air pollution by incorporating multiple data sources, accounting for scale dependencies, and handling the complexities of the underlying processes. By accurately modeling spatio-temporal correlations, researchers and policymakers can make informed decisions and take proactive measures to mitigate air pollution and safeguard public health.

5.3 Problem Formulation

This section describes the mathematical notation used throughout the chapter and the problem formulation for air pollution forecasting. We forecast $PM_{2.5}$ concentrations measured by a monitoring station every hour. Images are captured hourly near the monitoring station. Assuming a time window of length T, we are given as input $PM_{2.5}$ data $\mathbf{P_i} = \{p_t\}_{t=1}^{T} \in R^T$. The corresponding images are specified as $\mathbf{I} = \{i_t\}_{t=1}^{T} \in R^{T \times C \times H \times W}$, where $i_t \in R^{C \times H \times W}$, C is the number of channels, H is the height, and W is the width of the image. In this chapter, $C = 3$, $H = 32$, and $W = 32$. We aim to predict the $PM_{2.5}$ concentrations over the next τ hours where the ground-truth is represented by $\mathbf{P_f} = \{p_{T+t}\}_{t=1}^{\tau} \in R^{\tau}$.

The goal is to predict $PM_{2.5}$ concentrations over the next τ hours. We formulate the problem as $\hat{P}_{\mathrm{f}} = M(\mathbf{P_i}, \mathbf{I})$, where M denotes the forecasting model, and the predictions are denoted as $\hat{P}_{\mathrm{f}} = \{\hat{p}_{T+t}\}_{t=1}^{\tau} \in R^{\tau}$.

5.4 Benefits of Using Images

The fusion of images and $PM_{2.5}$ data in air quality forecasting offers several notable benefits, advancing our understanding of air pollution dynamics and enabling more accurate and comprehensive predictions. This fusion leverages the complementary nature of visual information captured in images and the quantitative measurements provided by $PM_{2.5}$ sensors, providing a holistic perspective on air quality that surpasses the capabilities of traditional approaches.

Firstly, integrating images into air quality forecasting models allows for enhanced spatial analysis. By incorporating visual data, we gain access to a wealth of spatial information that can help identify localized patterns and variations in air pollution. Images captured by cameras and webcams provide a fine-grained view of the environment, enabling the detection of specific sources of pollution, such as industrial emissions or vehicular exhaust. This spatial awareness empowers decision-makers to implement targeted pollution control measures and optimize resource allocation based on the specific areas and sources contributing to poor air quality.

Secondly, the fusion of images and $PM_{2.5}$ data enables a more comprehensive understanding of air pollution by incorporating contextual information. Images not only capture the presence of particulate matter but also provide valuable insights into meteorological conditions, such as cloud cover, humidity, and wind patterns. These contextual factors play a crucial role in air pollution formation and dispersion, influencing the spatial and temporal distribution of $PM_{2.5}$ concentrations. By considering these contextual cues alongside the $PM_{2.5}$ measurements, forecasting models gain a more holistic view of the complex interplay between meteorology and air pollution, leading to more accurate and nuanced predictions.

Moreover, fusing images and $PM_{2.5}$ data allows for improved estimation and interpolation of air quality at unmonitored locations. Traditional monitoring stations are sparsely distributed, limiting their ability to capture the heterogeneity of air pollution within a city or region. However, images provide a vast amount of visual data collected from various locations, offering a more extensive coverage of the environment. By integrating this image-based information with $PM_{2.5}$ data, forecasting models can better estimate air quality at unmonitored locations by leveraging spatial correlations and patterns observed in the images. This capability is particularly valuable in densely populated urban areas with limited monitoring resources.

In summary, the fusion of images and $PM_{2.5}$ data in air quality forecasting has significant benefits. It enhances spatial analysis, enabling targeted pollution control measures and resource allocation. It provides a more comprehensive understanding of air pollution

dynamics by incorporating contextual information captured in images. Additionally, it improves estimation and interpolation of air quality at unmonitored locations, addressing the limitations of traditional monitoring networks. By harnessing the synergistic power of visual data and quantitative measurements, this fusion approach pushes the boundaries of air quality forecasting, empowering decision-makers with more accurate and actionable information to combat the adverse effects of air pollution on human health and the environment.

5.5 Image-Based Forecasting Model

We describe a novel multi-level attention LSTM network designed for air pollution forecasting. Unlike previous haze detectors [13, 123], our model can represent changes in haze over both space and time. Our proposed model integrates a CNN and an LSTM; the CNN extracts the haze from each image and the LSTM predicts $PM_{2.5}$ concentrations over time. We also incorporate multi-attention to learn intricate relationships between images and the $PM_{2.5}$ data.

Figure 7.1 shows the architecture of the proposed model, resembling the encoder-decoder framework for time-series forecasting [124]. We develop three LSTM sequences: one encoding the previous $PM_{2.5}$ time-series data, another encoding the sequence of images, and another forecasting future $PM_{2.5}$ concentrations. We feed the past $PM_{2.5}$ data into an LSTM encoder to obtain its latent representation, and the image processing module learns to identify hazy regions from the images. Next, the image attention module weights each image region using the $PM_{2.5}$ hidden representation. The feature attention module then embeds each image and weights each image feature, and the image features are fed into another LSTM encoder. Finally, the LSTM decoder forecasts future $PM_{2.5}$ concentrations from the outputs of the two encoders.

5.5.1 Data Representation

This part will discuss the representation of the set of $PM_{2.5}$ data and images.

5.5.1.1 $PM_{2.5}$ Data Representation

The encoder of the $PM_{2.5}$ data is comprised of a sequence of LSTMs of length T. The $PM_{2.5}$ concentration p_t at time t is fed as an input to the encoder as $h_t = f_e(h_{t-1}, p_t)$, where f_e represents an LSTM unit and h_t represents the t-th hidden state. We obtain hidden states for each time step $H = \{h_1, ..., h_T\}$, where $h_t \in R^n$ is the t-th hidden state and n is the

Layer	# of Filters	Filter size	Activation
Conv	16	3 x 3	-
RDB 1-3	16	3 x 3	-
Conv	32	3 x 3	ReLU
Pool	32	H/2 x W/2	-
Conv	64	3 x 3	ReLU
Pool	64	H/4 x W/4	-
Conv	128	3 x 3	ReLU
Pool	128	H/8 x W/8	-

Table 5.1: The architecture of the image processing module. Padding makes image sizes consistent. For the pooling layers, the output of the filter size is denoted.

Layer	Input size	Output size	Activation
FC	$128 \times \frac{H}{8} \times \frac{W}{8}$	128	ReLU
FC	128	128	ReLU
FC	128	output size	ReLU

Table 5.2: The fully connected (FC) layers of the image embedding layers.

size of each hidden state. The output of the encoder is the hidden representation h_T of the entire $PM_{2.5}$ sequence.

5.5.1.2 Image Representation

Since images can represent the level of air quality, we learn to identify hazy regions from the images. For this purpose, we adapt the Residual Dense Block (RDB) [125], which has been used for single-image dehazing [123]. We develop the image sequence processing module outlined in Table 5.1. From the input i_t, the module begins with a conv layer and proceeds with three RDB blocks[1], and finally three Conv-Pool layers. The output i'_t consists of feature maps representing the level of haze for each region. Its output dimensions are $128 \times \frac{H}{8} \times \frac{W}{8}$, or $128 \times 4 \times 4$.

5.5.2 Image Attention Module

While the RDB has the ability to capture haze in an image via dense connections [125], it treats every pixel equally although images may contain uneven haze. It is important to weight image regions according to their relationship with the $PM_{2.5}$ data from the particle counters to accommodate different data modalities. We need signals from the $PM_{2.5}$ data when encoding each image. Hence, the image attention module preserves spatial information by selecting image regions enabling the most accurate $PM_{2.5}$ prediction. The experimental results show extracting salient features from images via attention improves accuracy.

For each time t, we calculate the attention weight for each 4×4 region using the the latent representation of the $PM_{2.5}$ data h_T. We compute the dot product between $W_i i'_t(x, y)$ and $W_h h_T$, where (x, y) is the location of the region, $i'_t(x, y) \in R^{128}$ is the 128-dimensional representation of the region at (x, y), and $h_T \in R^n$. The parameters to learn are $W_i \in R^{128\times128}$ and $W_h \in R^{128\times n}$. The attention weight $s_t(x, y)$ denotes the importance of the (x, y) region at time t and represents the similarity between the region and h_T:

$$s_t(x, y) = [W_i i'_t(x, y)]^T W_h h_T. \tag{5.1}$$

The attention weights are then normalized by the softmax over all regions. Finally, we

[1]For RDB, the depth rate (number of input features) is 16, the number of dense layers is 4, and the growth rate is 16. More details about RDB are in Zhang et al. 2018 [122].

multiply the attention weight matrix by i'_t to obtain the output i''_t.

$$\alpha_t(x,y) = \frac{\exp[s_t(x,y)]}{\sum_{x=1}^{4}\sum_{y=1}^{4}\exp[s_t(x,y)]}, \text{ and} \tag{5.2}$$

$$i''_t = \alpha_t i'_t. \tag{5.3}$$

Air pollution modeling using images introduces unique challenges due to the spatial nature of air quality patterns. Images capture detailed spatial information, but not all regions within an image may contribute equally to the prediction of pollutant concentrations. Some regions may contain more relevant features or exhibit stronger correlations with pollutant levels. By incorporating an image attention mechanism, the model can selectively focus on informative regions, effectively filtering out irrelevant or noisy information. This attention mechanism enables the model to assign different weights to image regions based on their importance, enhancing the model's predictive accuracy by emphasizing the most relevant regions while suppressing the influence of less informative areas.

Also, image attention plays a crucial role in integrating image data with PM2.5 data, which is obtained from particle counters. The model needs to leverage information from both data sources to capture a more complete picture of air pollution. Image attention allows the model to selectively attend to image regions based on their similarity to the latent representation of PM2.5 data. This integration enables the fusion of multiple data modalities, leveraging the complementary information provided by images (spatial patterns) and PM2.5 measurements (concentration levels). By incorporating PM2.5 data through image attention, the model can effectively capture the complex spatio-temporal correlations between air pollution patterns and the corresponding pollutant concentrations.

5.5.3 Feature Attention Module

We flatten the output i''_t to one dimension and feed it to the image embedding layers for each time t as described in Table 5.2, where the dimensions become $i'' \in R^{m\times T}$. The size of the output layer m is a hyper-parameter selected during training. Afterward, the image feature attention module represents the relationship between each image feature and the latent features h_T of $PM_{2.5}$. It adaptively selects the image features most relevant to h_T when predicting the future time series.

For time t, we calculate the attention weight of each image feature j via h_T. We compute the dot product between $W'_i i''(j)$ and $W'_h h_T$, where $i''(j) \in R^T$, and $h_T \in R^n$. The

parameters to learn are $W_i' \in R^{n \times T}$ and $W_h' \in R^{n \times n}$.

$$s(j) = [W_i' i''(j)]^T W_h' h_T. \tag{5.4}$$

The attention weight $s(j)$ represents the importance of j-th feature. The weights are normalized by the softmax over all m features.

$$\alpha(j) = \frac{\exp[s(j)]}{\sum_{k=1}^{m} \exp[s(k)]}. \tag{5.5}$$

The attention weights denote the importance of the individual features. Once the attention weights are computed, the input vector for time t is as follows:

$$\tilde{x}_t^{img} = [a(1)i_t''(1), a(2)i_t''(2), ..., a(m)i_t''(m)]^T. \tag{5.6}$$

Feature attention is a crucial component chosen to address the complex spatio-temporal correlations between air pollution and various features or variables that influence it. Air pollution is influenced by a multitude of factors, such as meteorological conditions, traffic patterns, land use characteristics, and emissions from specific sources. These factors exhibit varying degrees of relevance and impact on pollutant concentrations at different times. By incorporating feature attention, the model can dynamically assign importance weights to different features, allowing it to focus on the most influential factors and adaptively adjust the attention weights as the conditions change.

5.5.4 Model Architecture

The image features $\tilde{x}^{img}$ are fed into an LSTM encoder for images, comprised of a sequence of LSTMs of length T. The image features $\tilde{x}_t^{img}$ at time t are fed as an input to the encoder as $h_t^{img} = f_e^{img}(h_{t-1}^{img}, \tilde{x}_t^{img})$, where f_e^{img} represents an LSTM unit for the image and $h_t^{img} \in R^n$ represents the t-th hidden state of size n for the image. The output is the hidden representation h_T^{img} of the entire image sequence.

In the decoder with length τ, we concatenate the hidden representation of the image sequence h_T^{img} and the $PM_{2.5}$ data h_T. $h_0^d = [h_T^{img}; h_T] \in R^{2n}$ is then initialized as the first hidden state of the decoder. The previous output of the LSTM becomes the input of the next LSTM p_t' to update the decoder hidden state.

$$h_t^d = f_d(h_{t-1}^d, p_t'), \tag{5.7}$$

where f_d is an decoder LSTM unit. Afterward, we can estimate y_t:

$$y_t = W_y^T h_t^d + b_y. \tag{5.8}$$

The learned parameters are $W_y \in R^{2n}$, $b_y \in R$, which determines the prediction y_t.

5.6 Experimental Results

We evaluate our proposed model on air quality data and images from Shanghai. We first introduce the dataset and the experimental protocol. Next, we evaluate our proposed forecasting method and compare it with other methods. Afterward, we investigate the effect of each individual component of our proposed forecasting model.

5.6.1 Dataset and Implementation Details

We use air quality data from Shanghai from July 1st, 2014 to December 31, 2014 from the U.S. Consulates in Shanghai. The data contain hourly $PM_{2.5}$ measurements in $\mu g/m^3$. We also use webcam images taken by the Shanghai Environmental Monitoring Center near the air quality measurement station [126, 127]. The images were taken at the Oriental Pearl Tower. Our dataset includes images in the same data range approximately every hour from 8:00 am to 10:00 pm. We resized the images to $3 \times 32 \times 32$ (C $\times$ H $\times$ W). There are 2,296 chronologically ordered images.

The sequence length of the encoder is $T = 6$ (the window size) and the decoder time-step is $\tau = 6$. During the training phase, we conduct grid search to determine hyperparameter values. We set the learning rate to 0.005 and the batch size to 4, and apply early stopping for model training. The hidden size of each LSTM unit is 32, and the output size of the FC unit for the image processing module is 16 units.

We divide the dataset using an 8:1:1 ratio for training, validation, and testing data, which do not overlap. We use Adam to optimize parameters during training and use mean squared error (MSE) as the loss function. We evaluate our model's root mean squared error (RMSE) and mean absolute error (MAE). We also use gradient clipping with a parameter of 0.1. All experiments are run on a machine with an NVIDIA GeForce 940MX GPU.

5.6.2 Model Comparison

We compare our model with the existing pollutant forecasting methods listed below. We present the best performance of each method under different parameter settings.

Method	RMSE	MAE
HA	54.84	44.43
SVR	43.97	29.52
GBR	38.75	24.57
LSTM	27.81	18.69
Seq2seq	27.99	17.78
Only image processing module	25.59	16.90
Proposed approach with only image attention	24.78	16.32
Proposed approach	23.57	15.84

Table 5.3: Comparisons with previous forecasting methods in Shanghai (in $\mu g/m^3$) for six-hour forecasts.

- **Historical Average (HA)**: predict $PM_{2.5}$ concentrations using the mean of previous $PM_{2.5}$ concentrations.
- **Support Vector Regression (SVR)**: a supervised regression model that can map lower dimension data into a higher dimension space.
- **Gradient Boosting Regression (GBR)**: a supervised regression model using an ensemble of decision trees.
- **Long Short-Term Memory (LSTM)**: a recurrent network that models long-term temporal relationships.
- **Seq2seq**: an architecture that incorporates an LSTM to encode the input sequences and another LSTM to forecast time-series values.

Table 7.1 compares several air quality forecasting methods. We average the results of three runs. Experiments on Shanghai data show that our forecasting model improves accuracy by 15.8% in RMSE and 10.9% in MAE.

We also evaluate the impact of each model component via ablation studies (see Table 7.1). Notably, using images improves accuracy by 8.6% relative to Seq2seq when we add an encoder that extracts image features through the image processing module. Furthermore, adding the image attention module improves accuracy because it selects image regions by computing a dot product with the $PM_{2.5}$ latent features. This module further improves accuracy by 11.5% relative to Seq2seq. Finally, adding the feature attention module improves accuracy by 15.8% by weighting the extracted image features through a dot product with the $PM_{2.5}$ latent features.

Figure 5.3: The left figure is the original image with a 4×4 grid. The image attention module emphasizes certain regions of the image. In the right figure, the regions showing the original scene have attention, and the white regions do not have attention.

We hypothesize that the image attention module improves accuracy because it can identify image regions with $PM_{2.5}$ concentrations that are best correlated with the ground-truth sensor location. As shown in Figure 5.3, the image attention module emphasizes certain regions of the image. Since those regions tend to be clustered around the same area, we believe that the attention module can evaluate the correlations in $PM_{2.5}$ concentrations of different regions with sensor location.

5.7 Conclusion

The problem of air quality forecasting is important but also challenging because air quality is affected by a diverse set of complex factors. This chapter describes the first image-based air quality forecasting model. It fuses a history of $PM_{2.5}$ measurements with colocated images. We construct an image- and attention-based LSTM architecture to forecast $PM_{2.5}$ concentration, which uses multi-level attention to represent the spatio-temporal relationship of visual haze with measured $PM_{2.5}$ concentration over time. Experiments on Shanghai data show that our model improves $PM_{2.5}$ RMSE prediction accuracy by 15.8% and MAE by 10.9% compared to previous forecasting methods.

5.8 Future Work

Integration of Additional Data Sources: Expand the scope of data integration by incorporating other relevant data sources, such as meteorological data, traffic data, or satellite

imagery. By combining a wider range of data modalities, it is possible to capture more comprehensive and diverse factors that influence air quality. Investigate how the inclusion of these additional data sources improves the accuracy and robustness of the forecasting model.

Transfer Learning and Generalization: Explore the applicability and generalization of the developed model to different geographical regions or cities. Investigate transfer learning techniques that leverage pre-trained models on one region's data and fine-tune them for another region with limited data. Assess the model's ability to adapt to varying environmental conditions and evaluate its performance in different urban settings.

Real-time Implementation and Deployment: Develop strategies for real-time implementation and deployment of the image-based air quality forecasting model. Consider the computational requirements and scalability of the model for processing large volumes of image and air quality data in real-time. Explore techniques for efficient deployment on edge devices or cloud-based platforms, enabling widespread access to accurate air quality forecasts.

User Interface and Visualization: Design user-friendly interfaces and visualization tools that present the air quality forecasts in an easily understandable manner. Develop interactive visualizations that allow users to explore the spatial and temporal variations in air pollution, aiding decision-making processes for individuals, policymakers, and city planners.

By exploring these future research directions, we can advance the field of image-based air quality forecasting, enhance the accuracy and applicability of the models, and contribute to effective air pollution management and public health initiatives.

5.9 Broader Applications

The broader applications of image-based air quality forecasting through multi-level attention follow:

1. Air Quality Management: Accurate air quality forecasting can assist environmental agencies and policymakers in making informed decisions regarding pollution control measures, urban planning, and public health interventions. By providing timely and accurate predictions, it enables proactive actions to mitigate the adverse effects of air pollution on human health and the environment.

2. Healthcare Systems: Poor air quality has a significant impact on public health, contributing to respiratory diseases, cardiovascular problems, and other health issues.

By forecasting air quality at a local level, healthcare systems can prepare for potential increases in patient visits and allocate resources accordingly. The information can also be used to issue health advisories, enabling individuals to take necessary precautions and reduce exposure to harmful pollutants.

3. Smart Cities and IoT: Image-based air quality forecasting can be integrated into smart city initiatives and Internet of Things (IoT) frameworks. By leveraging existing camera networks or deploying new cameras in strategic locations, cities can collect real-time visual data and combine it with air quality measurements. This integrated approach enhances situational awareness, supports traffic management systems, and facilitates the development of intelligent urban environments.

4. Environmental Monitoring: Traditional air quality monitoring stations are limited in their coverage and spatial resolution. Image-based forecasting extends the monitoring capabilities by utilizing cameras and webcams that capture a larger spatial region. This approach enables the estimation of air quality in areas where sensor networks are sparse or nonexistent, providing a more comprehensive understanding of pollution patterns and aiding in the identification of pollution sources.

5. Data-Driven Decision Making: The proposed model combines multiple data sources, including images and $PM_{2.5}$ concentrations, to improve air quality forecasting accuracy. This data-driven approach can be extended to incorporate other relevant data, such as meteorological data, traffic data, and land-use information. By integrating diverse datasets, decision-makers can gain insights into the complex interactions between various factors and make informed decisions to improve air quality and urban sustainability.

6. Research and Development: The research on image-based air quality forecasting contributes to the broader field of environmental science and data analytics. It opens up avenues for further exploration and innovation in the fusion of different data modalities, the development of advanced deep learning models, and the application of attention mechanisms in environmental forecasting. This work can inspire researchers to explore similar approaches for other environmental variables and improve the understanding and prediction of complex environmental phenomena.

Overall, the broader applications of image-based air quality forecasting extend beyond the field of air pollution management and have implications for public health, urban planning, smart city initiatives, environmental monitoring, and data-driven decision making.

CHAPTER 6

A Context-Oriented Multi-Scale Neural Network for Fire Segmentation

6.1 Introduction

Accurate and rapid detection of fire is useful for environmental protection and public safety. It is also essential for perception systems in robotics and autonomous vehicle systems, especially those used in fire fighting. They must quickly react to unexpected situations and potentially catastrophic events.

Past research has focused on extracting information about wildfires using unmanned aerial vehicles (UAVs) [128]. Early and accurate fire detection is possible through the combination of computer vision and UAVs. UAVs are small, inexpensive, have the ability to navigate many areas, and often have hardware capable of automated data analysis. Therefore, UAVs can reliably monitor large areas such as woodlands and forests and determine the location and severity of fires [128, 129].

Since wildfires spread quickly and can be difficult to control, detection and suppression speeds are crucial. Wildfires cause millions of dollars of damage and kill thousands of people per year. While traditional fire detection technologies such as smoke sensors are inexpensive, they only detect nearby fire sources. Hence, there is an increasing interest in long-range, image-based fire detection.

The earliest image-based fire detection techniques use hand-crafted features from color, shape, and texture to detect fire regions [130,131]. With deep learning algorithms achieving remarkable progress in many fields [132, 133], they were also applied to fire detection recently [134, 135].

In image segmentation, deep learning methods have better performance than earlier methods using predetermined features, such as U-Net [136] and PSP-Net [137]. Hossain et al. detect forest fires with a neural network using color space local binary patterns of both

Figure 6.1: Images contain fire with different kinds of shapes, sizes, and illumination. The left column contains the original image and the right column contains the ground truth segmentation map. It is important to recognize flames that are present and also minimize false alarms.

flame and smoke signatures [135]. Choi et al. assign pixel-level labels of fire in images via a CNN residual network [138]. A recent study performed fire segmentation using a squeezed fire binary segmentation network with depthwise separable convolutions [139].

Despite the progress of fire detection methods, the accuracy of existing models decreases for many difficult scenarios. For example, small or occluded flames are difficult to identify. Also, complex backgrounds make it difficult to distinguish the fire from its surroundings and objects with similar color. Finally, the highly variable sizes, shapes, and colors of flames exacerbate the problem of fire segmentation.

Determining scene context, which refers to relationships among distant pixels, reduces false positives and false negatives. To handle small flame sizes (e.g., less than 5% of the image) as well as differentiate between the flame and background, it is necessary to enlarge the receptive field in order to effectively determine relationships among distant pixels. Also, to handle multiple scales of flames, multi-scale aggregation selectively combines useful information from different network layers. However, existing fire detection methods do not take into account these two important factors.

In this chapter, we propose a Context-Oriented Multi-Scale CNN. It does multi-scale aggregation, which outputs the segmentation map from multi-scale features and adaptively refines the features from different receptive fields. We also introduce a novel Context-

Oriented Module (COM) for our fire detection network. It extracts discriminant feature representations by building associations among features with global context, which uses relationships of all pixels in the feature map. In the COM, the input is fed into multiple branches with convolutions, average pooling, and global pooling. Then, the COM integrates the features from all branches.

High-resolution CNNs model relationships among nearby locations in the image, but their inductive biases make modeling long-range relationships difficult. Low-resolution, downsampled CNNs model long-distance relationships effectively, but disallow consideration of short-distance relationships due to downsampling. Our approach considers relationships at multiple length scales, and the additional cost of doing this is low because the downsampled analysis paths need only consider a small fraction of the data in the high-resolution path.

The main contributions are (1) a novel fire segmentation model, which utilizes global scene information and multi-scale aggregation, (2) a context-oriented module, which obtains local and global context information to expand the receptive field and extract more discriminative features, and (3) a multi-scale aggregation module, which reconstructs the segmentation using features from multiple receptive fields. Using our fire segmentation network improves accuracy by 2.7% in IoU compared with previous methods.

The remainder of this chapter is organized as follows. Section 2 details prior work in related fields. Section 3 describes our proposed model for fire segmentation. Section 4 demonstrates the performance of the proposed model compared to prior models and additional experimental analysis. Finally, Section 5 concludes the chapter.

6.2 Problem Importance

Wildfires, or unplanned wildland fires, cause tens of billions of dollars of damage and kill thousands of people per year in the U.S. [140]. They are becoming increasingly harmful over time, with a 90% increase in yearly damage from 2009 to 2018 [141–143]. Wildfires are commonly started in remote areas [144] by human accident or lightning strikes. In minutes, fires can develop from small, easily controllable blazes to out-of-control infernos. Even after detection, there are commonly substantial delays in delivering fire suppression equipment.

If wildfires could be identified and suppressed early enough, control would be inexpensive, with low risk to human life and property [145]. While natural fires are required to burn in many cases to keep the level of combustible material on the forest floor to a minimum, controlling when and where these burns occur can reduce the damage that they cause. Fur-

Figure 6.2: Scripps Ranch, California wildfire under 30 miles to the NIWC Pacific facility [3]. Rapid detection and response to fires worldwide has become an increasing concern for defense services to protecting coastal forests, harbors, ships, assets, and waterways.

thermore, the latency of information available to forestry wildfire services needs reduce from hours to minutes if to support early responses that keep wildfires under control.

My goal is to research and develop aerial fire observation and suppression technologies that monitor high-risk regions prone to fire with sophisticated imaging systems and machine learning algorithms. Over the past decade, there have been enabling advances to optical imaging and stabilization, machine learning computer vision algorithms for smoke plume and pollutant detection, and to multi-rotor drones capable of carrying payloads over long distances. With a small investment in research, they can be translated into life- and property-saving technologies.

6.3 Related Work

This section discusses related work in semantic segmentation and fire detection.

6.3.1 Semantic Segmentation

CNNs have achieved state-of-the-art performance in many computer vision fields. For instance, fully convolutional networks are used in image semantic segmentation and perform end-to-end classification of all pixels [146]. However, the receptive field is not large enough for feature representation of all the pixels in the image.

In order to differentiate between objects of different scales and illumination, it is necessary to enhance the discriminative ability of feature representations. One way to improve the performance of FCNs is multi-scale feature aggregation. PSPNet [137] uses spatial pyramid pooling to combine multi-scale information. The Deeplab model uses atrous spatial pyramid pooling (ASPP) with different dilation rates to capture contextual information [147].

In addition, attention mechanisms are applied for pixel-level recognition in order to enhance discriminative features. Zhao et al. introduce a pointwise spatial attention network that encodes relative position information in pixel space [148]. EncNet proposes an encoding layer on top of the network to capture global context [149]. Fu et al. include a self-attention module to model long-range dependencies [150].

Some methods incorporated attention mechanisms to learn feature weights and emphasize important features. OCNet learns feature weights according to object context [151]. Also, CCNet obtains contextual information based on all pixels in the criss-cross path [152]. Furthermore, the Dual Relation-Aware Attention Network [153] uses a self-attention mechanism that utilizes different pooling kernels to emphasize certain spatial areas. It also represents associations between channel dimensions to generate channel weights.

AttaNet [154] highlights certain pixels through a strip operation as well as a cross-level aggregation strategy. BiSeNetV2 [155] incorporates a detail path to preserve the spatial information and a semantic path to process feature maps with a large receptive field. Finally, ConvNeXt [156] constructs a revolutionary convolutional architecture containing inverted bottlenecks, larger kernel sizes, and other significant architectural differences.

6.3.2 Fire Detection

Prior image-based fire detection algorithms use the color and features of the fire [157, 158]. The most straight-forward fire detection methods are color-based [159]. They analyze images in the RGB, HSI, or YCbCr color spaces to obtain possible fire regions based on color thresholds [130, 131]. Other past work improves the accuracy of detection by considering additional features as shapes and optical flow [160, 161].

Deep learning algorithms perform automatic extraction of features and can greatly outperform conventional fire detection methods in detection accuracy. For example, Muhammad et al. [134] compared their CNN-based method with other hand-crafted fire detection methods and outperforms them in terms of accuracy by 0.88% and false positives by 11.6%. Yin et al. constructed a deep normalization and convolutional neural network attaining smoke detection rates at least 96.4% [162]. Another CNN-based method called

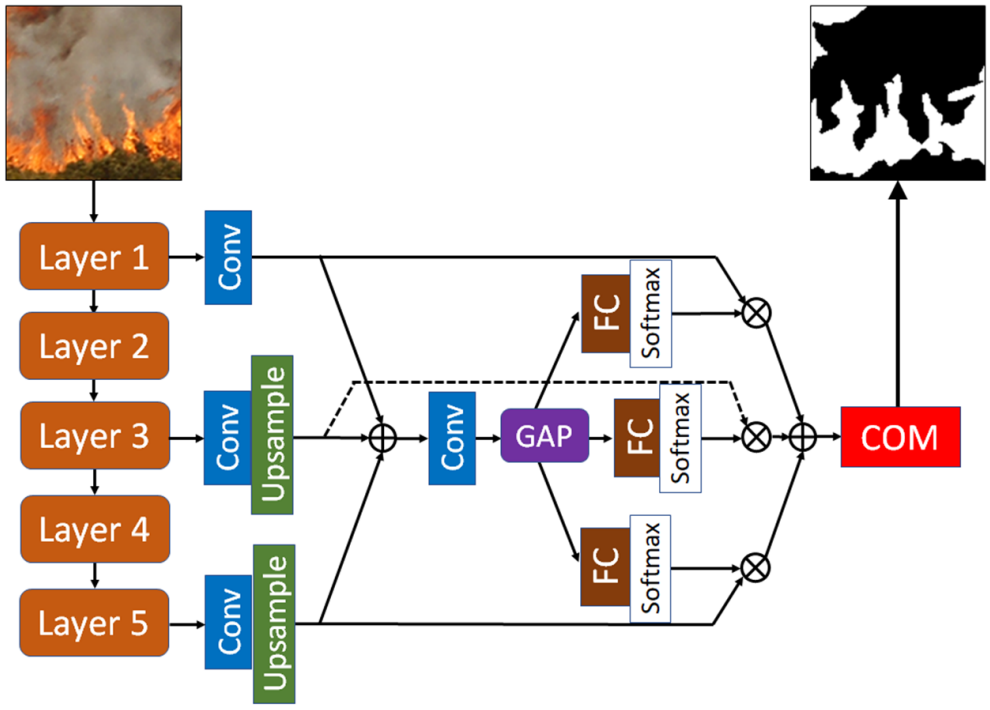


Figure 6.3: We propose a Context-Oriented Multi-Scale Network for fire segmentation with both a Multi-Scale Aggregation (MSA) layer and Context-Oriented Module (COM). MSA considers relationships at multiple layers in the network and performs adaptive feature refinement. COM is explained in the next figure.

the DCNN incorporates a deep dual-channel neural network for smoke detection and has a detection rate of 99.5% on average [163].

Hossain et al. detect forest fires with a neural network using color and multi-color space local binary patterns of both flame and smoke signatures [135]. Saponara et al. implemented a fully real-time CNN for fire detection using the YOLOv2 framework on a NVIDIA Jetson Nano [164]. Muhammad et al. described a framework based on the AlexNet architecture for fire detection and obtain an accuracy of 94.39% and a false positive rate of 9.07% [134, 165].

6.4 Methodology

This section provides an overview of the proposed model and describes each of its key components in detail.

6.4.1 Overview

Figure 7.1 shows the architecture of the proposed model. Initially, we use a five-layer ResNet-50 backbone to extract its features, denoted as $f_i(i = 1, 2, ..., 5)$. The backbone maps the input scene to feature representations, but it cannot capture both the local and global information of the scene well.

In order to exploit the multi-scale structure of the flames and deal with different flame sizes, we incorporate a multi-scale aggregation module. We perform adaptive feature refinement at multiple network levels in order to consider relationships at multiple length scales. The implications of this involve enhancing the intra-class and inter-class recognition.

Since contextual information can be used to improve the performance of CNNs, we expand the size of the receptive field by incorporating global contextual information via our Context-Oriented Module (COM). In scenes with diverse backgrounds and varied shapes, the COM can adaptively aggregate global contextual information, which refers to the relationships of all pixels in the feature map, improving feature representation for fire segmentation.

6.4.2 Multi-Scale Aggregation

We incorporate multi-scale aggregation (MSA) to capture different scales of flames more accurately. We incorporate a gating mechanism to adjust the level of information from different layers. It adaptively passes important semantic information at multiple layers in order to improve accuracy. Hence, the network focuses on more informative contextual features.

The structure of this module is shown in Figure 7.1. From the backbone layer, each of f_1, f_3, and f_5 form separate branches, go through a conv layer, and are each upsampled to the dimension of f_1. The outputs are f_1', f_3', and f_5', respectively, containing features at different scales.

We select multiple layers where each layer is downsampled by a different amount. Earlier layers have more spatial information and later layers have more semantic information about the image. Next, we combine all outputs using an element-wise sum as $F = f_1' + f_3' + f_5'$. Afterward, we apply global average pooling (GAP) across the spatial dimension of $F \in R^{W \times H \times C}$ to compute channel-wise statistics $s \in R^{1 \times 1 \times C}$.

Later, we feed s into three independent fully connected layers, FC_1, FC_3, and FC_5, and apply softmax to the outputs to obtain w_3, w_4, and w_5. We then perform channel-wise multiplication for $f_1' \cdot w_1$, $f_3' \cdot w_3$, and $f_5' \cdot w_5$ and then fuse them via element-wise summation.

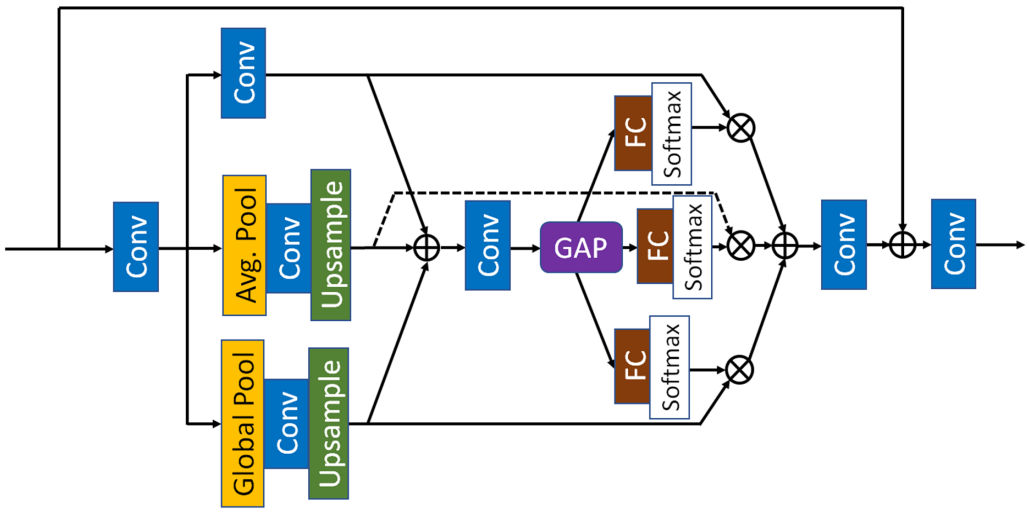


Figure 6.4: We propose a Context-Oriented Module (COM) that extracts discriminant feature representations by building associations among features with average and global pooling.

This is described as follows:

$$\begin{aligned} s &= \text{GlobalPooling}(F), \\ w_1, w_3, w_5 &= \text{softmax}([FC_1(s), FC_3(s), FC_5(s)]), \text{ and} \\ V &= C(F_1 \cdot w_1 + F_3 \cdot w_3 + F_5 \cdot w_5). \end{aligned} \tag{6.1}$$

6.4.3 Context-Oriented Module

We adopt the Context-Oriented Module (COM) to expand the receptive field to capture richer features. The network initially obtains feature representations by stacking conv layers, but it cannot capture both local and global information simultaneously. Incorporating more contextual information via local and global pooling can improve fire segmentation accuracy.

Past work has shown that global context information improves various computer vision tasks [137, 166]. We obtain more discriminative feature representations for better scene understanding by building associations with features through global context. Feature aggregation allows the network to focus on more informative contextual features.

The detailed structure of the Context-Oriented Module is shown in Figure 6.4. The output of the MSA layer V is fed into an input conv layer to output V'. Next, V' is fed to three branches: one branch contains a conv layer, another branch contains an average pooling layer, followed by a conv layer and upsampling block, and the other branch contains a

global pooling layer, followed by a conv layer and upsampling block.

The outputs are F_c, F_l, and F_g, representing local and global features, respectively. All three branches contain features with different receptive fields. Then, we combine both local and global features using an element-wise sum as: $F = F_c + F_l + F_g$. This is described as follows:

$$\begin{aligned} V' &= C(V), \\ F_c &= C(V'), \\ F_l &= U(C(P(V'))), \\ F_g &= U(C(G(V'))), \text{ and} \\ F &= F_c + F_l + F_g, \end{aligned} \tag{6.2}$$

where C, P, G, and U represent convolution, average pooling, global pooling, and upsampling, respectively. We then apply global average pooling (GAP) across the spatial dimension of $F \in R^{W \times H \times C}$ to compute channel-wise statistics $s \in R^{1 \times 1 \times C}$.

Later, we feed s into three independent fully connected layers, FC_c, FC_l, and FC_g, and apply softmax to the outputs to obtain w_c, w_l, and w_g. We then perform channel-wise multiplication for $F_c \cdot w_c$, $F_l \cdot w_l$, and $F_g \cdot w_g$ and fuse them via element-wise summation. The output F' selectively incorporates local and global attention based on their content and characteristics. These operations are described as follows, which is similar to those of the MSA:

$$\begin{aligned} s &= \text{GlobalPooling}(F), \\ w_c, w_l, w_g &= \text{softmax}([FC_c(s), FC_l(s), FC_g(s)]), \\ F' &= C(F_c \cdot w_c + F_l \cdot w_l + F_g \cdot w_g), \text{ and} \\ F' &= C(F' + V). \end{aligned} \tag{6.3}$$

6.4.4 Difference Between the Two Modules

The two modules serve different purposes. The multi-scale aggregation (MSA) module uses features from low-level and high-level features to capture spatial details better. Low-level and high-level features are complementary, where low-level features are rich in spatial details but lack semantic information, and vice-versa for high-level features. To bridge the gap between high-level and low-level features, MSA adaptively combines both features with a novel design. MSA improves accuracy by better capturing spatial details from low-level features. Low-level features contain information from a lower receptive field, so MSA does not expand the receptive field but improves the spatial reasoning of the network

through local details.

In contrast, the context-oriented module (COM) further expands the receptive field from the output of the backbone network to additional length scales by average pooling and global pooling. In particular, the COM further improves the network's ability to extract semantic information.

6.4.5 Loss Function

The binary cross-entropy loss ($\mathcal{L}_{BCE}$) is used to calculate the loss of each pixel in the predicted segmentation map compared to the ground-truth map. This is formulated as

$$\mathcal{L}_{BCE} = \sum_{i=1}^{M} \sum_{j=1}^{N} -[y_{ij} \log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij})], \tag{6.4}$$

where M is the number of images in the dataset, N is the number of pixels in the image in a flattened array, and p_{ij} and y_{ij} are the values of the jth pixel in the predicted segmentation map and the ground truth map of the ith image, respectively.

6.5 Experimental Results

We first introduce the dataset and the experimental protocol. Next, we evaluate our proposed method on images containing wildfires and compare it with other methods. We then investigate the effect of each individual component of our model.

6.5.1 Dataset and Implementation Details

We use a benchmark dataset of wildfires, consisting of 595 images of varying size [167]. The dataset includes annotation of all fire pixels and each is resized from a larger size down to 512×512. We then augment the dataset by applying random cropping five times for each image to size 224×224 to end up with 2,975 images in total.

The training dataset contains 2,000 images, while the testing dataset contains 975 images. During the training phase, we set the learning rate to 2e-4, the batch size to 2, and the number of epochs to 40 for model training. Also, we set the momentum parameter to 0.9 and use Adam to optimize the parameters during training. All experiments are run on a machine with an NVIDIA GeForce 940MX GPU.

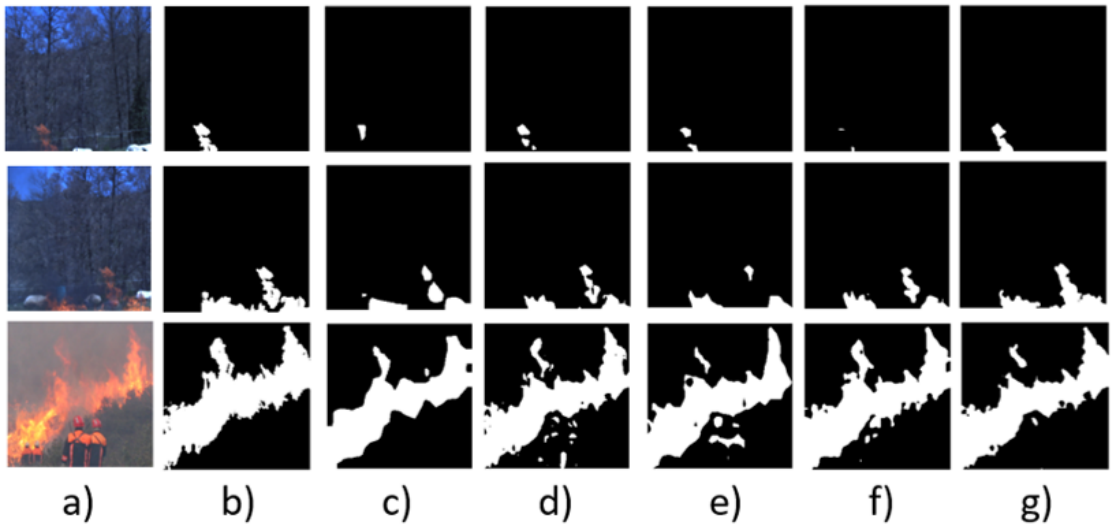


Figure 6.5: Visual results of our method and four previous segmentation methods. Our model is effective at segmenting flames of various sizes and distinguishing flames from complex backgrounds. a) Input image, b) Ground-truth image, c) DeepLabv3, d) DRAN, e) AttaNet, f) BiSeNetV2, g) Proposed method. Our model is capable of accurately segmenting wildfires in complex scenes.

6.5.2 Model Comparison

We compare our model with past fire segmentation methods, shown in Table 7.1. For a fair comparison, we calculate each method's accuracy with the same parameters. The list of previous methods are U-Net [136], PSP-Net [137], DeepLabv3 [168], CPD [169], RAS [170], DRAN [153], AttaNet [154], BiSeNetV2 [155], and ConvNeXt [156]. Although more recent work is on fire detection instead of fire segmentation[1], we include more recent methods for generic segmentation to compensate for the lack of recent fire segmentation methods.

Experiments on the benchmark dataset show that our model improves accuracy by 2.7% compared to RefineNet. We report all segmentation results in terms of mean Intersection over Union (mIoU) and Dice error, which are widely used to evaluate the overall performance of semantic segmentation algorithms. The mIoU metric reflects the degree of the overlap between the predicted segmentation and the corresponding ground truth versus their union.

[1]Fire segmentation determines which pixels in the image contain fire, whereas fire detection decides whether or not any pixels in the image contain fire.

Table 6.1: Results of fire segmentation with other methods. The results of the best existing fire segmentation method and the proposed method are bolded.

Methods	**IoU**	**Dice**
U-Net (2015)	0.705	0.792
PSP-Net (2017)	0.653	0.757
DeepLabv3 (2017)	0.755	0.834
CPD (2019)	0.681	0.779
RAS (2020)	0.686	0.780
DRAN (2020)	0.751	0.829
AttaNet (2021)	0.747	0.827
BiSeNetV2 (2021)	**0.781**	**0.852**
ConvNeXt (2022)	0.632	0.741
Ours w/o MSA	0.675	0.771
Ours w/o COM	0.789	0.858
Ours	**0.808**	**0.873**

6.5.2.1 Ablation Analysis

We also conducted an ablation analysis to evaluate the effectiveness of each module. First, we remove the Context-Oriented Module (COM) and only keep the Multi-Scale Aggregation (MSA) module in order to examine the effectiveness of the COM. From Table 7.1, we observe that our model with the COM outperforms our model by 1.9% without the COM. Hence, the COM improves accuracy by expanding the receptive field in order to consider relationships of longer length scales in the feature map.

We then remove the MSA module and retain the COM. From Table 7.1, we observe that our model with the MSA module outperforms our model by 13.3% without it.

6.5.2.2 Visualization of Results

Figure 6.5 shows the qualitative comparison of our proposed method and past fire segmentation methods. We select some representative examples from the dataset. It can be seen that our method is capable of accurately segmenting flames in challenging scenes and performs significantly better than other models.

In the first row, previous methods were not able to discern the small flame in the image. Some methods in the second row confused the background with the fire. Furthermore, some existing models confused the flames with the background which has similar appearance with the fire. In contrast, our method can accurately infer the flame region in each case.

This is mainly because Multi-Scale Aggregation (MSA) can handle flames with differ-

ent scales via adaptive feature refinement at multiple levels of the CNN. Also, the Context-Oriented Module can help discriminate the flames from the background in complex scenes.

6.6 Conclusion

This chapter describes a Context-Oriented Multi-Scale CNN for fire detection in images. The proposed approach leverages global scene information, multi-scale aggregation, and a context-oriented module to improve fire segmentation accuracy. Our method is able to handle challenging scenarios such as small or occluded flames, complex backgrounds, and highly variable sizes, shapes, and colors of flames. Compared to existing methods, our approach improves accuracy by 2.7% in IoU.

The Context-Oriented Module extracts discriminant feature representations by building associations among features with global context, and the Multi-Scale Aggregation module outputs the segmentation map from multi-scale features and adaptively refines the features from different receptive fields. The combination of these modules effectively determines relationships among distant pixels, reduces false positives and false negatives, and enhances the discriminative ability of feature representations. Our approach considers relationships at multiple length scales, which allows for modeling both short-distance and long-distance relationships. The proposed method has important practical applications in environmental protection, public safety maintenance, and robotics, particularly in autonomous systems designed for fire fighting.

6.7 Future Work

For future work, one may extend the output segmentation map for flames to additional parts. For instance, I can construct a 2D probability-of-fire map that can assess regions to scan with greater resolution and occurrence. It can provide the likelihood of fire occurring based on the land use patterns, information from weather services, images from satellites, and historical data.

Moreover, I can construct a 2D risk-of-damage map to aid in assessing whether to deploy suppression measurements or not. It can use simulation information based on weather data and imagery to predict the direction and speed of a fire if it occurred at each location on the map. It can correlate these data with dwelling and property data to estimate the range, likelihood, and cost of damages.

Another area for future research is the integration of multiple data sources for fire segmentation. Other data sources such as aerial imagery, LiDAR, and ground-based sensors

could be used in combination with satellite data to improve the accuracy and resolution of fire segmentation models.

One important consideration for future research is the need to balance accuracy with computational efficiency. While deep learning techniques have shown promise in improving accuracy, they can also be computationally expensive, which can be a limiting factor for real-time applications. Developing more efficient algorithms that can run in real-time on resource-constrained devices could greatly expand the potential applications for fire segmentation technology.

Another important area for future research is the development of automated systems for fire segmentation and suppression. While current fire segmentation algorithms are capable of detecting fires, they typically require human intervention to initiate suppression efforts. Developing automated systems that can detect and suppress fires in real-time could greatly improve the speed and effectiveness of fire suppression efforts.

Finally, there is a need for further research on the impact of climate change on wildfire behavior and the efficacy of fire segmentation and suppression technologies. As wildfires become more frequent and intense due to climate change, it is critical to understand how these changes will affect the effectiveness of fire segmentation and suppression technologies. This research could inform the development of more effective and adaptive fire segmentation algorithms that can adapt to changing wildfire behavior.

CHAPTER 7

Spatial-Frequency Network for the Segmentation of Remote Sensing Images

7.1 Introduction

Remote sensing technologies have enabled the collection of a large number of optical satellite images, and the spatial resolution of remote sensing images has increased up to the degree of centimeters. Satellite-based remote sensing images are used for various applications, including classification of vegetation, urban structures, or crop type. As a result, it is important to have the ability to accurately detect land usage in pre-processing of optical satellite images.

Identifying land use patterns from satellite imagery is an important problem, where each pixel is precisely classified in the output. Earlier techniques adopted hand-crafted features using support vector machines and other models, but convolutional neural networks (CNNs) have recently made major breakthroughs in many subfields of computer vision with their vastly improved accuracy; hence, past work used CNNs for land use segmenation as well due to their ability to generalize.

Ding et al. incorporate patch attention to enhance the feature extraction of contextual information and leveraging multi-layer fusion [171]. Also, Yu et al. use multiscale feature extraction via the pyramid pooling module for semantic segmentation on aerial images [172]. Another model developed boundary losses in order to improve the edge extraction in satellie images [173]. Furthermore, Marmanis et al. propose using a class-boundary detection network to improve accuracy [174].

Satellite image-based land use detection is a challenging problem for the following reasons: remote sensing images are high-resolution with many diverse objects. Additionally, images contain many different kinds of terrain and different lighting which varies over time. In particular, understanding scene context is important to process high-resolution satellite images by extracting the relationships of each pixel with surrounding pixels. This

is essential for discerning fine-grained spatial areas and modeling the relationship between different semantic classes.

Past computer vision research, e.g., on object detection, has found that texture information encoded by CNNs is very useful for accurate localization. Other research found that frequency-domain information can denote texture, noise, and low-level information in images [175, 176]. Edges correspond to higher frequencies and the inner surfaces correspond to lower frequencies. Past work learns identical parameters for all frequency components, whereas learning different parameters for different frequency levels can enhance feature representation.

This kind of segmentation problem requires learning more expressive feature representations for intricate scene understanding at the pixel and frequency domains. Learning features at various frequencies, especially high-frequency features, can help with reducing confusion between different semantic classes. This chapter describes a spatial-frequency CNN for aerial segmentation, which extracts the relationships of each pixel with surrounding pixels in the spatial and frequency domains.

We introduce a Frequency Weighted Module to regularize the network based on the frequency-based features to refine the segmentation details. Also, we develop a Spatial Weighting Module that encodes which spatial areas of the input the network should focus on. Finally, we develop a Multi-Domain Fusion Module to aggregate features from different domains, which can provide important complementary information from different domains. Using our spatial-frequency segmentation network improves accuracy by 1.9% in IoU compared to previous methods. In addition, we evaluate the impact of each model component via ablation studies.

The main contributions follow.

- We describe a novel deep learning model for aerial image segmentation that enhances feature representation in both the spatial and frequency domains. This technique preserves essential details and textures in order to improve the learning of features at multiple frequency scales.
- We design a Frequency Weighted Module to encode contextual information based on the frequency domain via a Fourier transform.
- In order to enhance contextual information in the spatial domain, we employ a Spatial Weighting Module to effectively determine which relationships among distant pixels are important through a multi-scale pooling layer.

The remainder of this chapter is organized as follows. Section 2 reviews previous research in related fields. Section 3 describes our spatial-frequency-based segmentation

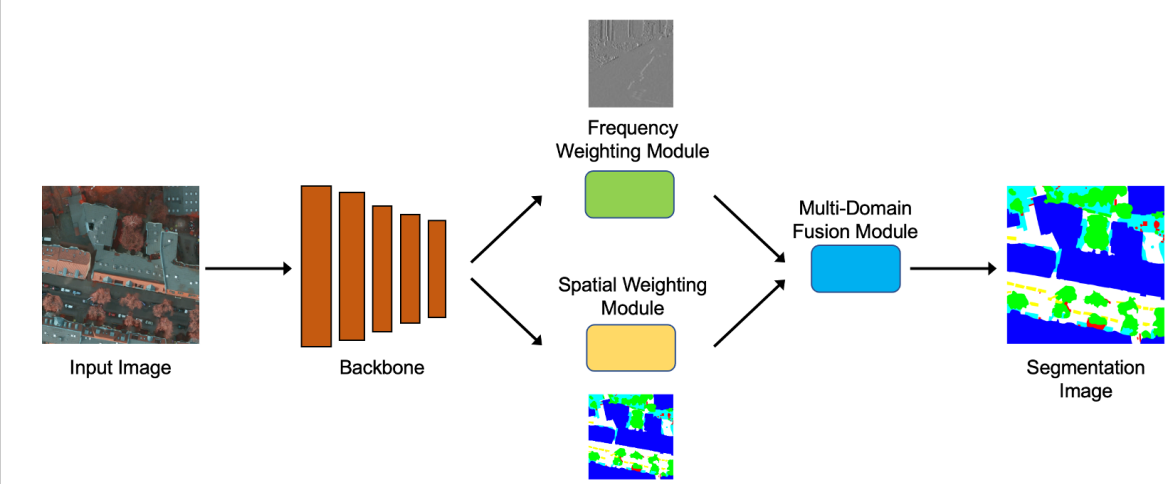


Figure 7.1: We propose a novel model designed for segmentation of satellite images that enhances feature representation in both the spatial and frequency domains. This model preserves essential details and textures in order to improve the learning of features at multiple frequencies. Finally, we develop a Multi-Domain Fusion Module to aggregate features from different domains, which can provide important complementary information.

model for remote sensing images. Section 4 demonstrates the performance of the proposed model compared to previous forecasting models and additional experimental analysis. Finally, Section 5 concludes the chapter.

7.2 Problem Importance

High-resolution satellite image segmentation has a wide range of applications in various fields, including agriculture, forestry, urban planning, and environmental monitoring. Here are some specific examples of how high-resolution satellite image segmentation can be used in each of these fields.

In agriculture, high-resolution satellite image segmentation can be used to monitor crop health and yield estimation. By segmenting satellite images into different crop types, farmers can better understand which crops are thriving and which ones need more attention. This can help farmers make more informed decisions about how to allocate resources such as water and fertilizers, which ultimately leads to higher crop yields and lower costs.

In forestry, high-resolution satellite image segmentation can be used to identify different types of trees and vegetation cover. This information is crucial for forest management and conservation efforts, such as identifying areas that require reforestation, monitoring deforestation and illegal logging, and identifying forest fires and their spread. Accurate

segmentation can also help with predicting forest growth patterns and estimating carbon sequestration potential.

In urban planning, high-resolution satellite image segmentation can be used to identify different types of buildings, roads, and other infrastructure. This information is useful for analyzing urban sprawl, monitoring urban expansion, and identifying areas that require urban renewal. By understanding the distribution of different land use types, urban planners can make more informed decisions about zoning and development policies, as well as optimize transportation networks and public services.

In environmental monitoring, high-resolution satellite image segmentation can be used to track changes in land use and natural resources such as water bodies, wetlands, and wildlife habitats. By comparing satellite images over time, researchers can identify changes in land use patterns, detect environmental degradation such as soil erosion and desertification, and assess the impact of climate change on ecosystems. This information is crucial for developing sustainable development policies and mitigating the negative impact of human activities on the environment.

Overall, high-resolution satellite image segmentation is a powerful tool for understanding the earth's surface and how it is changing over time. By accurately segmenting satellite images into different land use types and other features, we can better understand the distribution of resources, monitor changes in the environment, and make more informed decisions about how to manage our planet's natural resources.

7.3 Related Work

This section discusses related work in semantic segmentation of natural images and semantic segmentation of remote sensing images.

7.3.1 Semantic Segmentation

Convolutional neural networks (CNNs) have achieved state-of-the-art performance on many computer vision problems, such as object detection and image generation. CNNs perform extraction of features through convolutional and pooling layers. For instance, fully convolutional networks (FCNs) are used in image semantic segmentation and perform end-to-end classification of all pixels [146]. However, the receptive field is not large enough for feature representation of all the pixels in the image.

To differentiate between objects of different scales and illumination, it is necessary to enhance the discriminative ability of feature representations. One way to improve the per-

formance of FCNs is multi-scale feature aggregation. PSPNet [137] uses spatial pyramid pooling to combine multi-scale information. The Deeplab model uses atrous spatial pyramid pooling (ASPP) with different dilation rates to capture contextual information [147].

In addition, attention mechanisms are applied for pixel-level recognition in order to enhance discriminative features. Existing work in computer vision use convolutional networks to model context in images in order to improve accuracy in various vision tasks such as object recognition. Zhao et al. introduce a pointwise spatial attention network that encodes relative position information in pixel space [148]. EncNet proposes an encoding layer on top of the network to capture global context [149]. Fu et al. include a self-attention module to model long-range dependencies [150].

In addition, FCNs embed semantic information into high-level feature maps by downsampling the input image. While the features capture fine details in the image, they lose information about the precise location of each pixel. The U-Net model adds skip connections between the feature maps of the encoder and decoder to fuse low-level and high-level features [136]. Zhang et al. incorporate semantic information into low-level features and spatial inforamtion into high-level features [177]. Moreover, Yu et al. create the BiSeNet model with a spatial path to preserve the spatial information and a context path to process feature maps with a large receptive field [178].

Some methods incorporated attention mechanisms to learn feature weights and emphasize important features. For example, the DRANet adaptively weights important features for both the spatial and channel dimensions [179]. OCNet learns weights for features according to their object context [151]. Also, CCNet obtains contextual information based on all pixels in the criss-cross path [152].

7.3.2 Semantic Segmentation of Remote Sensing Images

The earliest algorithms for segmentation of remote sensing images use pre-determined features [180, 181]. Later, models adopted commonly used machine learning algorithms including random forests and support vector machines [182, 183]. Past work used automatic fuzzy clustering for remote sensing image classification [184].

With deep learning algorithms achieving remarkable progress in many fields [132, 133], they were also applied to satellite image segmentation recently [172, 173]. Past work used CNNs for land use segmenation for their improved accuracy as well due to their ability to generalize. For example, a CNN model includes boundary losses in order to improve the edge extraction in satellite images [173]. Marmanis et al. propose a class-boundary detection network to improve accuracy [174].

Some models applied multi-scale feature extraction to learn contextual information at different scales. Yu et al. used the pyramid pooling module for semantic segmentation on aerial images [172]. Furthermore, a model employs two stages for effective multi-scale processing of remote sensing images [171]. DDCM-Net combines both low-level and high-level features and obtains features using dense dilated layers [185].

Furthermore, attention mechanisms are applied in order to increase the discriminative power of feature representations. Ding et al. incorporate patch attention to enhance feature extraction of contextual information and leverage multi-layer fusion [171]. MANet uses multiple attention layers through the kernel and channel [186].

While CNN-based segmentation algorithms for satellite images achieved improved performance, segmentation can be difficult because images capture different kinds of terrain with different scales, occlusion, and illumination levels. A filter in a convolutional network only covers a finite region of the image and represents a local receptive field. It is essential that each pixel in the output has access to a large receptive field from the input image so that contextual information is taken into account.

7.4 Methodology

This section first provides an overview of our segmentation model for remote sensing images and then describe each of its key components in detail.

7.4.1 Overview

We describe a novel model designed for remote sensing image segmentation (see Figure 7.1). Specifically, we adopt a ResNet-50 as our backbone network to extract multi-level features from the input image, i.e., f_i $(i = 1, 2, ..., 5)$. We next extract more discriminative features from both the spatial and frequency domain. We define features obtained from spatial information compared to features obtained from frequency information as different domains.

In order to enhance contextual information in the spatial domain, we employ a Spatial Weighting Module to effectively determine relationships among distant pixels. This is useful to discern fine-grained spatial areas, especially confusing areas and boundaries. Following that, we apply a Frequency Weighted Module to encode contextual information based on the frequency domain via a Fourier transform. Since remote sensing images contain significant information on the texture and outline as well as noise, we selectively combine useful information from different frequency bands.

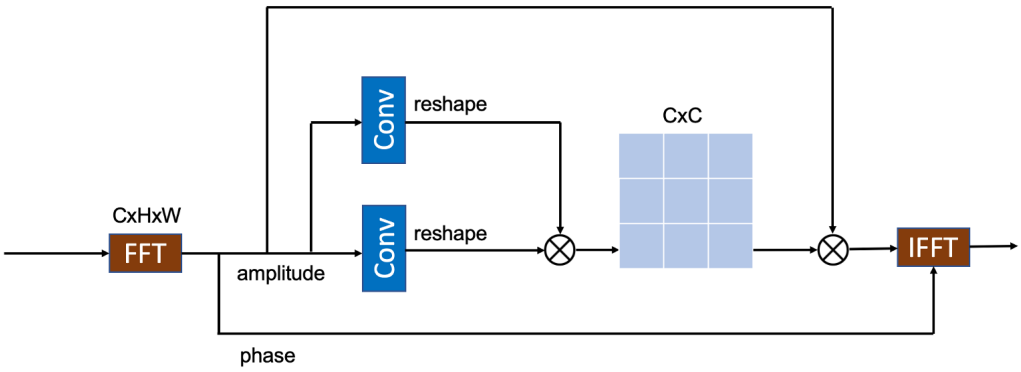


Figure 7.2: Frequency Weighting Module.

Finally, we develop a Multi-Domain Fusion Module to aggregate features from different domains, which can provide important complementary information. This module has the capability to learn shared representations for two different feature representations through a cross-fusion technique. Overall, the proposed model has the ability to retain both local and global features through the expansion of the receptive field.

7.4.2 Frequency Weighting Module

Semantic segmentation of satellite images involves handling the problem of intra-class and inter-class variations. It is difficult to discriminate between many objects in the scene area in remote sensing images and it is affected by both the image's texture and the context. To alleviate this problem, we propose a Frequency Weighting Module (FWM) to enhance important information in the extracted features based on frequency level.

Remote sensing images have large spatial size and contain significant fine-grained information. While remote sensing images contain plentiful contextual information, it is important to evaluate semantic information at different frequencies to distinguish among object classes. Features of higher frequencies tend to provide important texture information, and features of lower frequencies tend to provide important shape information.

As a result, we adjust the extracted features in the frequency domain in order to perform dynamic frequency modulation, whereas past work treated each feature in any frequency level equally. This module can help facilitate the information flow and learning complementary representations of features. Moreover, this mechanism can help suppress any noisy feature representations.

We use the Fourier transform $\mathcal{F}$ to convert the features from the spatial domain to the frequency domain, and the inverse Fourier transform $\mathcal{F}^{-1}$ to convert the features from the

frequency domain to the spatial domain. Both transforms are implemented through the FFT algorithm. The Fourier transform applies an image $x \in R^{C\times H\times W}$, and the equation is described below:

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{i=0}^{N-1} \sum_{j=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)}. \tag{7.1}$$

The Fourier transform outputs both amplitude and phase components, and the Fourier transform and inverse are computed independently on each channel of feature maps. The equations of both the amplitude component and phase component are denoted below:

$$\begin{aligned} \mathcal{A}(x)(u,v) &= \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \text{ and} \\ \mathcal{P}(x)(u,v) &= \arctan \frac{I(x)(u,v)}{R(x)(u,v)}. \end{aligned} \tag{7.2}$$

where $R(x)(u,v)$ is the real component of $\mathcal{F}(x)$ and $I(x)(u,v)$ is the imaginary component of $\mathcal{F}(x)$. In particular, the amplitude component tends to contain low-level statistics of the original image [175, 176].

We now describe the architecture of the Frequency Weighting Module (FWM) shown in Figure 7.2. We apply the Fourier transform to the output of the backbone network, and we feed the amplitude component $F \in R^{C\times H\times W}$ into the FWM. We reshape F to two dimensions $R^{C\times(H\times W)}$ and obtain the weights $W \in R^{C\times C}$ by doing a matrix multiplication of F with F', and then applying the softmax operation:

$$w_{ji} = \frac{\exp(F_i \cdot F_j)}{\sum_{i=1}^{C} \exp(F_i \cdot F_j)}. \tag{7.3}$$

Afterward, the transpose of the weighting map W is multiplied by the amplitude feature map F. Then we reshape their result to R^{CHW} to obtain the amplitude-based weighted features. Then we multiply the result by a parameter β and perform an element-wise sum operation with F to obtain the amplitude-based weighted features:

$$F_j = \beta \sum_{i=1}^{C} (w_{ji} F_i) + F_j. \tag{7.4}$$

Finally, we take the modified amplitude component and the phrase component through an inverse Fourier transform to obtain the spatial feature maps.

Our module has many capabilities. It aggregates contextual information from many frequency levels to model channel relationships. It focuses on more discriminative and in-

formative features by exploiting the inter-dependencies among frequency-based features. Finally, the correlations between object classes can be modeled by the Frequency Weighting Module through non-local context.

7.4.3 Spatial Weighting Module

Past work in segmentation used convolutional layers that only operate with a local receptive field, which is unable to capture contextual information outside of the local region. It is important to utilize relationships of all pixels in the feature map in order to obtain features that is able to discern fine-grained regions in the image.

Images typically exhibit different attributes at different length scales. To enhance spatial details, we introduce a multi-scale pooling layer that uses average pooling operations with different bin sizes in order to capture contextual information. In our pooling layer, we use bin sizes of 1×1, 2×2, 3×3, and 6×6, and then upsample the pooled feature maps to the original size. After that, we concatenate the feature maps.

In the Spatial Weighting Module, we feed the output of the backbone network into a 3×3 convolutional layer to obtain F. We then feed F into a multi-scale pooling layer described earlier to obtain F'. We reshape F to two dimensions $R^{(H\times W)\times C}$ and also F' to $R^{C\times(H\times W)}$. We obtain the weights $W \in R^{(H\times W)\times(H\times W)}$ by doing a matrix multiplication of F with its transpose, and then applying the softmax operation:

$$w_{ji} = \frac{\exp(F_i \cdot F'_j)}{\sum_{i=1}^{C} \exp(F_i \cdot F'_j)}. \tag{7.5}$$

Afterward, the transpose of the weighting map W is multiplied by F'. Then we reshape their result to R^{CHW}, and we multiply the result by a parameter λ and perform an element-wise sum operation with F to obtain the position-based weighted features:

$$F_j = \lambda \sum_{i=1}^{H\times W} (w_{ji} F'_i) + F_j. \tag{7.6}$$

7.4.4 Multi-Domain Fusion Module

As shown in Figure 7.3, we propose a Multi-Domain Fusion Module to fuse cross-domain features. This block improves accuracy because it learns the complex correlations from features of different domains. While other methods have directly concatenated different feature vectors from different domains into one long vector, this does not fully extract the complementary information from spatial and frequency features.

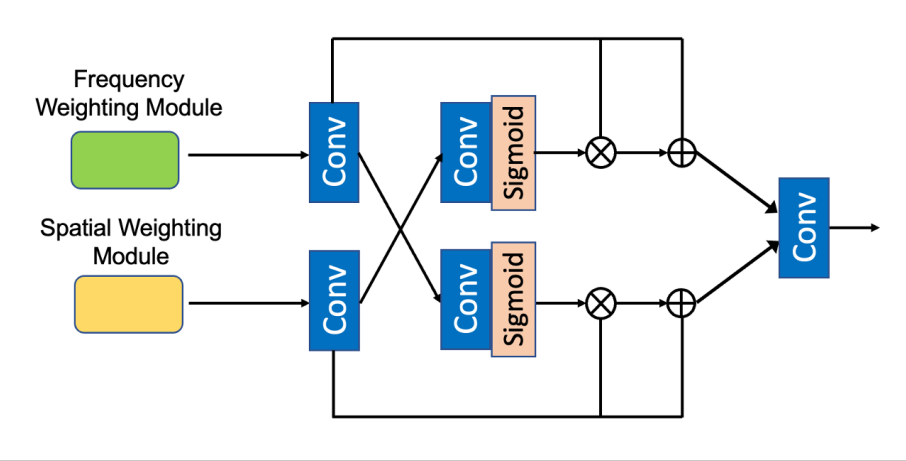


Figure 7.3: Multi-Domain Fusion Module.

We initially perform enhancement of features in both the spatial domain x_s and in the frequency domain x_f by boosting features in one domain through a normalized weighted map in the other domain. Initially, we feed the two kinds of features into a 3×3 convolutional layer in order to embed both features into the same feature space. Next, we feed both features into a 3×3 convolutional layer and then a sigmoid activation layer. Hence, we have normalized feature maps for both the spatial and frequency domains, w_s and w_f, respectively.

At this point, we weight the feature map of the spatial domain x_s by using the normalized feature map from the frequency domain w_f, and vice-versa. This is used to represent the correlations between the two feature domains. We also add a residual connection in order to retain the original information of each domain. The output x'_f is the cross-enhanced feature representation from w_s, and the output x'_s is the cross-enhanced feature representation from w_f.

$$\begin{aligned} x'_f &= x_f + x_f \times w_s \\ x'_s &= x_s + x_s \times w_f \end{aligned} \tag{7.7}$$

Afterward, the module integrates the features by concatenating and then feeding them into a 3×3 convolutional layer. Finally, we obtain the output which combines information from multiple domains.

The information in x_s and x_f are complimentary, so the multi-domain fusion module exploits the relationship between the different features. The normalized feature maps can be regarded as feature-level attention maps to adaptively weight the feature representations of another domain. This leads to more discriminative features and improves segmentation

Methods	Impervious surface F1	Building F1	Low vegetation F1	Tree F1	Car F1	Overall F1
SegNet [187]	0.551	0.537	0.368	0.308	0.684	0.490
U-Net [136]	0.488	0.518	0.438	0.500	0.702	0.529
RefineNet [188]	0.578	0.587	0.469	0.502	0.746	0.576
LANet [171]	0.641	0.665	0.450	0.511	0.736	0.600
BiSeNetV2 [155]	0.627	0.673	0.458	0.435	0.790	0.597
MACUNet [189]	0.565	0.555	0.445	0.517	0.755	0.567
MA-Net [186]	0.626	0.678	0.479	0.531	0.720	0.607
Proposed	0.599	0.699	0.526	0.548	0.761	0.626

Table 7.1: Results of aerial image segmentation with other segmentation methods.

accuracy for remote sensing images.

7.5 Experimental Results

This section describes the data, experimental evaluation, and discussion.

7.5.1 Dataset and Implementation Details

We evaluate our segmentation model for remote sensing images using the Potsdam dataset [190], which is publicly available. It comprises of 38 true orthophotos (TOPs) of size 6000×6000, consisting of satellite views of a historic city. The ground-truth contains six semantic categories: buildings, trees, cars, low-vegetation, impervious surfaces, and background/clutter. We select 24 RGB images for training and the remaining 14 images for testing. For both the training and testing datasets, we augment the dataset by applying random cropping 30 times for each image to size 224×224 to end up with 1,180 images in total.

During the training phase, we set the learning rate to 5×10^{-4}, the batch size to 8, and the number of epochs to 100 for model training. Also, we set the momentum parameter to 0.9 and use Adam to optimize the parameters during training.

7.5.2 Evaluation Metrics

In order to compare our method with past work, we use the F1 score to evaluate semantic segmentation models of satellite images. The F1 score for a certain class is defined as the weighted average of precision and recall, and is useful when semantic classes are more imbalanced.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{7.8}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \text{ and} \tag{7.9}$$

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{7.10}$$

where

- TP stands for true-positive,
- TN stands for true-negative,
- FP stands for false-positive, and
- FN represents false-negative.

7.5.3 Model Comparison

We compare our model with past segmentation methods for aerial images in the Potsdam dataset, shown in Table 7.1. For a fair comparison, we calculate each method's accuracy with the same parameters and the cross-entropy loss function. Also, we use ResNet-50 pretrained on ImageNet as the backbone network for all previous methods.

Using our spatial-frequency segmentation network improves mean F1 score by 1.9% compared to previous methods. We assess a variety of methods including those containing multi-scale fusion and attention mechanisms. We include SegNet and U-Net as an early method baseline. More recent aerial image-based segmentation methods comprise of LANet [171], MACUNet [189], and MA-Net [186].

MA-Net is the most recent aerial image-based segmentation method and uses attention mechanisms based on the kernel operation and channel dimension. It generally performs well compared to previous baseline segmentation models. However, our spatial-frequency segmentation network further improves accuracy by 1.9% in mean F1-score over MA-Net because our model has the ability to discern fine-grained spatial regions and discriminate between object classes.

7.5.4 Ablation Study

In order to assess the capabilities of the proposed modules, we conduct ablation experiments using different settings. Table 7.1 shows the results of the ablation experiments on

Methods	mean F1
Ours w/o FWM + Fusion	0.581
Ours w/o SWM + Fusion	0.611
Ours w/o Fusion	0.618
Ours	0.626

Table 7.2: Evaluation of the accuracy of each component of our proposed segmentation method.

the Potsdam data set.

First, we remove the Multi-Domain Fusion Module from the network in order to examine its effectiveness. Instead of the fusion module, we sum the outputs of the SWM and FWM and feed the output through a 3×3 convolutional layer. We observe that our model with the fusion module outperforms our model without it. Next, to verify that the SWM module can further improve the accuracy, we conduct analysis without the SWM module and only kept the FWM. From Table 7.1, we observe that our model with the SWM module outperforms our model without it.

Moreover, we remove the FWM from the network and only keep the SWM module in order to examine the effectiveness of the FWM. We observe that our model with the FWM outperforms our model without it. This reflects that the Frequency Weighting Module is necessary for improving the accuracy by using frequency levels capture richer features and discriminate between object classes.

7.5.5 Visualization of Results

Figure 7.4 shows the qualitative comparison of our proposed method versus past aerial image segmentation methods. The aerial images shown are representative examples from the Potsdam dataset. It can be seen that our method is capable of accurately segmenting challenging areas of the satellite image (i.e., discrimination of fragmented segments, distinguishing between different semantic classes) and performs significantly better than other models in F1-score.

Earlier methods contain multiple fragmented segments, while the proposed method concatenate the segments into one for the most part. Additionally, the our method demonstrates more accurate segmentation of each semantic class. This is due to our spatial weighting module which is capable of combining local contextual information.

Furthermore, our model is adept in discerning between semantic classes, while other methods confused different object classes with one another. While the segmentation output of MANet is more accurate than past methods, it still lacks the ability to discern between

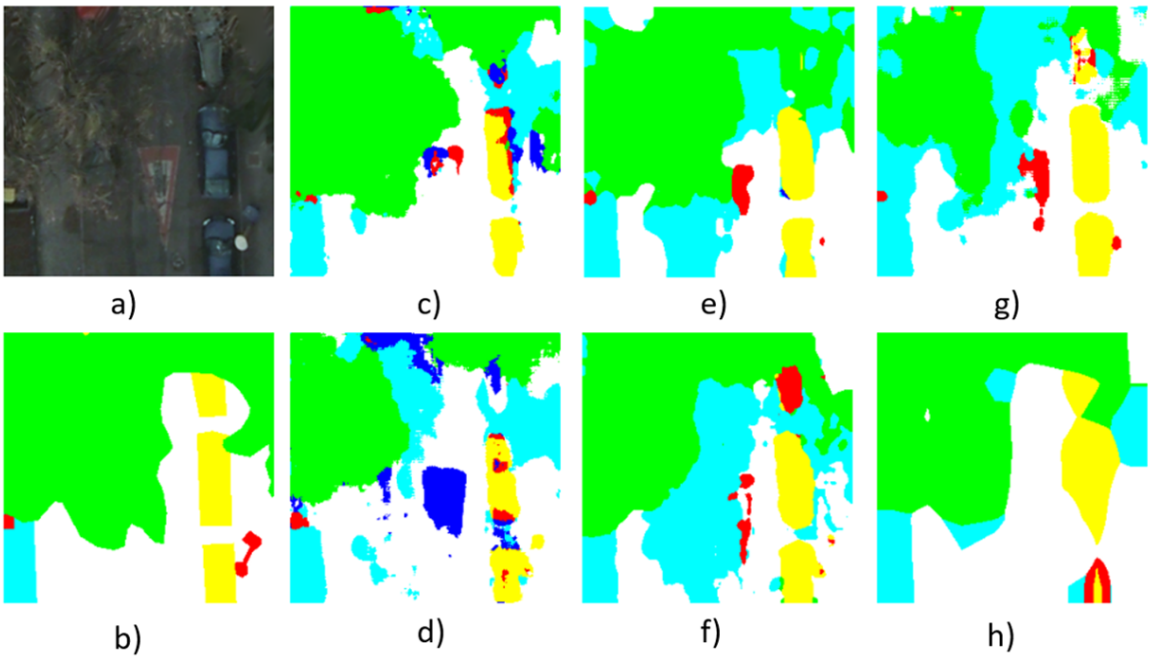


Figure 7.4: Visualization of segmentation results between our method and other segmentation methods on the Potsdam test set. (a) Input image. (b) Ground-truth segmentation map. c) U-Net with F1-score of 0.556. d) MACUNet with F1-score of 0.477. e) BiSeNetv2 with F1-score of 0.533. f) LANet with F1-score of 0.694. g) MANet with F1-score of 0.708. h) Proposed method with F1-score of 0.752. (white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter).

semantic classes. In particular, the frequency weighting module has the ability to improve the discrimination of each semantic class.

Hence, our method improves both the discrimination of semantic categories and the preservation of spatial details.

7.6 Conclusion

In this chapter, we proposed a novel deep learning model for aerial image segmentation that enhances feature representation in both the spatial and frequency domains. Our approach preserves essential details and textures in order to improve the learning of features at multiple frequency scales. Specifically, we introduced a Frequency Weighted Module and a Spatial Weighting Module to encode contextual information based on the frequency and spatial domains, respectively. Moreover, we developed a Multi-Domain Fusion Module to aggregate features from different domains, which can provide important complementary information.

Our proposed model achieved state-of-the-art performance on several remote sensing datasets, improving accuracy by 1.9% in the mean F1-score compared to previous methods. We also performed ablation studies to demonstrate the effectiveness of each model component. Our approach has the potential to improve many remote sensing applications, including vegetation classification, urban structure detection, and crop monitoring.

Overall, our work contributes to the field of remote sensing and computer vision by addressing the challenges of high-resolution satellite image segmentation and improving the accuracy of land use detection. Future work can explore extensions of our model to other types of remote sensing data, such as LiDAR and SAR, and investigate additional techniques for improving feature extraction and fusion in both the spatial and frequency domains.

7.7 Future Work

While the proposed spatial-frequency segmentation network shows significant improvements over previous methods, there are still many avenues for future research in this area. In particular, future work could focus on the following areas.

Firstly, while our proposed network is effective in segmenting land use in high-resolution satellite images, there may be potential to extend this approach to other types of remote sensing data. For example, this approach could be applied to multispectral or hyperspectral data, which contain information about the reflectance of different wavelengths of light. This would require modifications to the network architecture to handle the additional dimensions of the input data.

Secondly, our current approach focuses on pixel-wise classification of land use categories. However, there may be potential to incorporate additional information into the segmentation process, such as elevation data or spatial context. This could be achieved through the use of additional input channels or through modifications to the network architecture.

Thirdly, the proposed network relies on a Fourier transform to extract features in the frequency domain. However, there may be alternative methods for extracting frequency-based features that could be explored. For example, recent work has shown promising results using wavelet transforms for feature extraction in image segmentation tasks.

Lastly, the proposed network has been evaluated on several benchmark datasets, but there may be potential to apply this approach to real-world applications. For example, this approach could be applied to monitor land use changes over time or to detect and track specific features of interest, such as buildings or vegetation. This would require additional

evaluation on real-world datasets and potentially modifications to the network architecture to handle the additional complexity of real-world data.

CHAPTER 8

Conclusion

The increasing concern regarding environmental challenges has ignited a heightened interest in utilizing machine learning and computer vision techniques to portray scenes in environmental applications. The accurate and effective representation of scenes holds paramount significance in tackling environmental problems such as air pollution, fire detection, and remote sensing analysis. This dissertation delves deep into the realm of scene representations in machine learning and computer vision, concentrating particularly on image-based methods tailored for environmental applications.

Initially, this dissertation is dedicated to the development and evaluation of vision-based air quality estimation and prediction algorithms. The goal is to accurately estimate high spatial resolution air pollutant concentrations. It is possible to estimate pollution concentrations by observing the impact on light attenuation using commodity consumer cameras. The field is closer to developing a portable, inexpensive, and accurate image-based pollutant sensing method in urban and industrial areas.

Overall, I envision a world in which it is possible to accurately estimate human exposure to pollution. My research is currently on that path with the immediate goal of high resolution estimation of air quality. This will enable public officials to help track and identify potential problems in air pollution. Officials can act on situations much quicker and treat the public with the necessary care.

After achieving high-resolution estimation of air quality using vision-based algorithms, my research has expanded into other related areas in the environmental domain such as fire detection and remote sensing segmentation, leveraging the knowledge and techniques developed in the field of vision-based air quality estimation.

Building upon the progress made in vision-based air quality estimation, we developed effective fire detection systems using principles in deep learning. By adaptively incorporating information from multiple levels of the CNN and enhancing the receptive field network, we can utilize commodity consumer cameras to detect and monitor fires in real-time. This

advancement would contribute to early fire detection, enabling prompt response measures, and assisting firefighters in tackling wildfires and minimizing their impact on the environment and human lives.

Additionally, the expertise gained in vision-based air quality algorithms can be applied to remote sensing segmentation tasks. Remote sensing imagery, such as satellite or aerial images, can be processed using computer vision techniques to segment and classify various land cover types, vegetation, water bodies, and human-made structures. This segmentation analysis facilitates environmental monitoring, land management, urban planning, and disaster response efforts, empowering decision-makers with critical information for informed actions.

8.1 Contributions

My contributions can be summarized as follows.

1. Designing a wavelength-sensitive, absorption and spatial variation aware multi-pollutant vision-based estimation technique. It improves accuracy by 22% compared to previous image-based pollution estimation methods.

2. Contributing to the first publicly released dataset appropriate for evaluating vision-based pollution estimation algorithms. The dataset is a densely distributed, low-cost PM2.5 and PM10 database with high temporal and spatial resolution along with images taken at the location of the sensors.

3. Determining how accuracy depends on point sensor density and the presence or absence of cameras. I show that the prediction of air pollution concentrations at various locations is enhanced when images are used at various sensor densities.

4. Contributing to novel vision-based approach for estimating nighttime $PM_{2.5}$ concentration. The method involves deriving a glow map based on image brightness and transmission, followed by the design of a deep convolutional neural network algorithm for quantitative estimation.

5. Developing an image-based $PM_{2.5}$ forecasting model that capture the level of haze in images over time. The model incorporates a multi-level attention to learn intricate relationships between images and the $PM_{2.5}$ data. This direction accomplishes multi-sensor air pollution prediction.

6. Designing a Context-Oriented Multi-Scale Network for fire segmentation. This network adaptively integrates local and global context and uses multi-scale aggregation in order to give more precise segmentation results. Our method improves IoU accuracy by 2.7% compared to past work.

7. Producing a remote sensing segmentation model that enhances feature representation in both the spatial and frequency domains. As remote sensing images have high spatial resolution, this technique preserves essential details and textures in order to improve the learning of features at various frequencies, especially high-frequency features.

8.2 Future Work

My research on vision-based air quality estimation can be expanded into several related topics, which are described in the following sections: 3D air pollution estimation, quantifying human exposure to air pollution and health effects through high-resolution static sensors, and vision-based air quality estimation by learning from synthetic hazy images.

8.2.1 3D Air Pollution Estimation

The first direction is low-cost, visual, high-resolution 3D pollution field estimation. This involves combining overlapping images taken at different locations to reconstruct an accurate prediction of air pollution in 3D space. Such an algorithm would allow us to estimate pollutant concentrations with high spatial resolution, enabling analysis of air pollution flows within a city.

Since there are few, sparsely distributed monitoring stations, it is essential to estimate fine-grained air quality at arbitrary locations. In particular, monitoring stations are expensive and require maintenance; on the other hand, fixed webcams are less expensive and can cover large spatial areas. Moreover, there are many image-based haze detection algorithms [19, 20, 123, 191], but they do not have the ability to measure fine-grained pollutant concentrations. I aim to use techniques from haze estimation as well as 3D reconstruction (i.e., Colmap) to produce a 3D haze estimation algorithm.

The plan for the algorithm design is as follows: given multiple camera views with a common scene, i.e., a region of a city, we define a 3-dimensional Cartesian space filled with unit voxels. Each unit voxel represents a cubic volume of constant size in the real scene. For each camera view, we will utilize the depth of the scene and generate a projection of camera view in our voxel space. As shown in Figure 8.1, each camera view contains

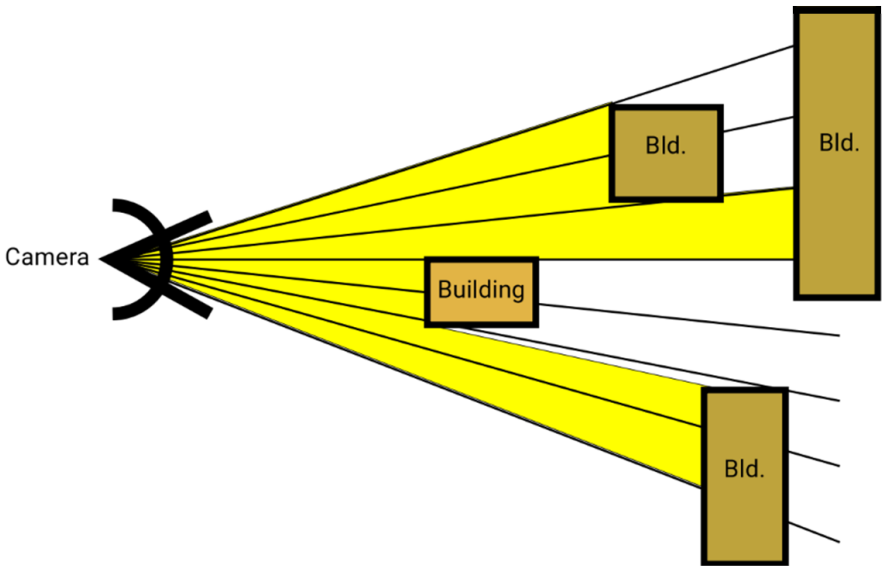


Figure 8.1: This figure represents the camera view of multiple prisms with different depths, each ending when the prism hits a building. The camera view relies on the physical structure of scene.

multiple prisms with different depths, each ending when the prism hits a building. Following existing single image air quality estimation models, we define air pollution as the transmittance per unit length of a true object color as light rays travel through a distance before hitting the camera sensor.

We will develop and validate an algorithm that translates from 2D views without knowledge of heterogeneity in pixel-associated prisms. For each camera view, we calculate the projection of the pollution and have voxels seen by multiple cameras. We define a constraint satisfaction problem in which all the incident voxels from a camera to an object in finite distance must have transmittance that sums up to the total transmittance implied by the color shift on a particular pixel. The color shift is explain from the equation (described in the first chapter) $I(x) = J(x)t(x) + A(1 - t(x))$. $I(x)$ is the observed hazy image, $J(x)$ is the haze-free scene radiance to be recovered. Additionally, there are two critical parameters: A denotes the global atmospheric light, and $t(x)$ is the transmission. Inspired by cone-beam reconstruction in medical applications, we develop a pollutant reconstruction technique from 2D camera views into a 3D volume based on the concepts of transmittance and image reconstruction.

To validate our approach, we will generate a synthetic hazy image of a scene containing several buildings via extruded rectangles and two different colors of haze, each with substantially varying concentrations and hot spots at different locations in the image, as shown in Figure 8.2. We will generate randomly distributed haze densities in 3D space, where

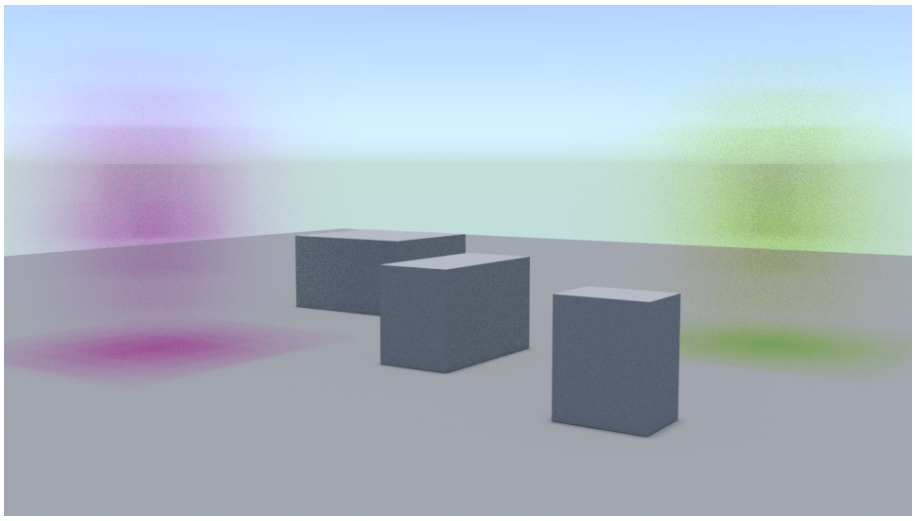


Figure 8.2: This figure represents the concentrations of multiple pollutants in three-dimensional space by using different colors.

the haze densities are inversely correlated with each other with respect to distance. Finally, we will construct the synthetic haze images to correspond to real world haze, where haze concentrations are inversely correlated with distance.

8.2.2 Quantifying Human Exposure to Air Pollution and Health Effects through High-Resolution Static Sensors

For air pollution exposure analysis, I will perform human exposure analysis utilizing an air quality dataset with fine-grained spatial resolution. I will perform personal exposure analysis to PM2.5 indirectly by utilizing a high spatial resolution air quality dataset in Hangzhou (from HVAQ [9]), and I will represent human motion through Hangzhou taxi traces. The main question I will examine is how does the density of sensors used in a city affects personal exposure analysis? I will show that relying on low spatial resolution pollution concentration data results in an increase in the error of per-person exposure estimates compared to high spatial resolution data.

It is necessary to accurately quantify individual human exposure in order to understand human exposure to PM2.5 in the future. Existing research in human exposure modeling evaluates the impact of air quality (e.g., PM2.5 levels) on human health. Initially, personal exposure to pollutants was estimated over large and coarse scales (e.g., on the order of tens of km) [192]. Coarse-grained personal exposure estimates may lead to large errors in epidemiological studies [193]. For instance, Paolella et al. concluded that model-estimated PM2.5 exposure is lower with coarser grids than with finer grids: the estimated mean expo-

sure increases by 27% in the United States when the grid cell edge length is decreased from 69 km to 5.9 km [194]. High resolution data are essential for human exposure assessments because pollutant levels can vary at small spatial and temporal scales in a difficult-to-predict way [9].

As high resolution pollutant data may not always be available, air quality modeling is critical for accurate individual exposure estimates [195–197]. Existing air pollution research found that PM2.5 concentrations varied on the scale of hundreds of meters [9], where two different particle counters less than 1 km apart have differing pollutant concentrations. Even though recent personal exposure studies used finer resolution air quality data as fine as 1 km^2 for assessments, those methods still have certain limitations. In the human exposure analysis from Tan et al., the relative errors at 9 km^2 (3 km resolution) ranged from 26% to 245%, while the relative errors at 1 km^2 (1 km resolution) ranged from –25% to 59%. However, Tan et al. claimed that a grid resolution of 1 km^2 cannot consistently capture the low-level variations in highly industralized areas because the relative errors at 1 km^2 can still be improved [198].

Existing human exposure studies are still not adequate enough without being conducted at a more granular level and should be conducted at scales less than 1 km^2. For accurate human exposure assessment, we need to determine the pollutant concentrations for each point in space and time a person occupies. Newly developed high-resolution air quality datasets can enable us to monitor personal exposure to air pollutants more directly. For example, Chen et al. released an air quality dataset with a density of 10 sensors in an area less than 1 km^2 and a sampling period of one second called HVAQ [9]. However, fewer studies have been conducted investigating the relationships between the density of air pollution measurements and human exposure calculations.

Furthermore, measurements that are acceptably accurate for estimating average human exposure can be very inaccurate when estimating a non-linear response for health effects via pollutant concentration. There is an association between exposure to PM2.5 and an increased risk of rheumatoid arthritis, connective tissue diseases, and inflammatory bowel diseases [199]. Specifically, an additional 7% risk of having autoimmune disease was linked to an increase of 10 $\mu g/m^3$ in PM10 concentration [199]. Also, exposure to PM10 above 30 $\mu g/m^3$ and PM2.5 above 20 $\mu g/m^3$ was associated with a 12% and 13% higher risk of autoimmune disease, respectively [199].

In this work, we will examine the effect of the spatial resolution of air pollution data on human exposure estimates in Hangzhou. We will leverage fine-grained air quality data from multiple stationary sensors in Hangzhou via HVAQ [9], and we will use taxi traces to represent the motion patterns of people. Existing human exposure models are based on

pollutant data with resolution as low as 1 km^2, but HVAQ has a density of 10 sensors in an area less than 1 km^2; we will calculate individual human exposure and perform analysis from the highest resolution data yet.

8.2.3 Vision-based air quality estimation by learning from synthetic hazy images

Another plan is to develop a vision-based air quality estimation method to predict the PM2.5 concentration in various locations and utilize synthetic hazy images generated using realistic models to improve the accuracy of PM2.5 estimation. The main problem that we are trying to solve is the lack of large quantities of training images labeled with PM2.5 concentrations. Through model-based regularization, we assume that the PM2.5 concentration of synthetic haze images are strongly correlated with their scattering coefficient corresponding to realistic situations.

PM2.5 contributes to degraded air quality in many areas of the world and is associated with millions of premature deaths annually. Air quality sensing generally has the goal of helping scientists understand the complex process of air pollution formation and propagation. However, most existing approaches for air pollution monitoring and human exposure estimation are spatially sparse. Hence, we will construct a vision-based air quality estimation method that can predict the PM2.5 concentration in various locations within a city.

Deep neural networks have achieved extremely high accuracy for a multitude of computer vision applications, such as image classification, object detection, and semantic segmentation. Their performance and accuracy is attributable through supervised learning, which requires a labeled dataset, and that training deep networks on larger datasets produces better performance. Nonetheless, it can be difficult to obtain a dataset with a sufficiently large number of labeled images. For instance, labeling data often requires human labor; images taken throughout the day need to be associated with air pollution concentrations. In some cases, there is no way to measure the air pollution concentration at the moment the image was taken.

A powerful approach for training deep models is through synthetic images. This mitigates the requirement for labeled data by providing a means of utilizing synthetic images. Since synthetic data can often be obtained with minimal human labor, any performance boost often comes with low cost. For hazy images, the atmospheric scattering model has been used to model the level of image in an image. The classical description for the generation of haze in images is (also described in the first chapter): $I(x) = J(x)t(x)+A(1-t(x))$. $I(x)$ is the observed hazy image, $J(x)$ is the haze-free scene radiance to be recovered. Ad-

ditionally, there are two critical parameters: A denotes the global atmospheric light, and $t(x)$ is the transmission.

We will develop a method of estimating air quality from images and use synthetic hazy images to improve the accuracy of air pollution estimation from realistic hazy images. For the synthetic hazy images, we will incorporate regularization by assuming that two images with the same scattering coefficient, β, have the same PM2.5 concentration. We also will assume that an image with a greater scattering coefficient, β, than that for another image will have a greater PM2.5 concentration. As a result, we expect our vision-based PM2.5 prediction system will produce accurate estimates of PM2.5 concentrations. Our experimental results so far provide a strong case for the benefits of applying vision-based methods for convenient and accurate estimation of air quality. Furthermore, our method could help increase public awareness of the relationship between their behavior and exposure to polluted air and provide valuable information to scientific researchers, government officials, and health professionals.

BIBLIOGRAPHY

[1] N. Jacobs, N. Roman, and R. Pless, "Consistent Temporal Variations in Many Outdoor Scenes," March 2007.

[2] A. Leskinen, A. Ruuskanen, P. Kolmonen, Y. Zhao, D. Fang, Q. Wang, C. Gu, J. Jokiniemi, M.-R. Hirvonen, K. Lehtinen, S. Romakkaniemi, and M. Komppula, "The contribution of black carbon and non-bc absorbers on aerosol absorption coefficient in nanjing, china," Aerosol and Air Quality Research, vol. 20, pp. 590–605, 2020.

[3] "Cedar fire's lessons, 10 years later," Oct. 2013. [Online]. Available: https://www.sandiegouniontribune.com/sdut-wildfire-cedar-anniversary-fire-2013oct24-htmlstory.html

[4] J. Wang, Y. Zhang, M. Shao, and X. Liu, "The quantitative relationship between visibility and mass concentration of pm 2.5 in beijing," 2006.

[5] X. Fu, X. Wang, Q. Hu, G. Li, X. Ding, Y. Zhang, Q. He, T. Liu, Z. Zhang, Q. Yu, R. Shen, and X. Bi, "Changes in visibility with pm2.5 composition and relative humidity at a background site in the pearl river delta region." Journal of environmental sciences, vol. 40, pp. 10–9, 2016.

[6] X. Zhang, X. Ding, D. Talifu, X. Wang, A. Abulizi, M. Maihemuti, and S. Rekefu, "Humidity and pm2.5 composition determine atmospheric light extinction in the arid region of northwest china." Journal of environmental sciences, vol. 100, pp. 279–286, 2021.

[7] T. Zhang and R. P. Dick, "Estimation of multiple atmospheric pollutants through image analysis," in 2019 IEEE Int. Conf. on Image Processing (ICIP). IEEE, 2019, pp. 2060–2064.

[8] C. Zhang, J. Yan, C. Li, X. Rui, L. Liu, and R. Bie, "On estimating air pollution from photos using convolutional neural network," in Proc. Int. Conf. on Multimedia, 2016, pp. 297–301.

[9] Z. Chen, T. Zhang, Z. Chen, Y. Xiang, Q. Xuan, and R. P. Dick, "Hvaq: A high-resolution vision-based air quality dataset," IEEE Trans. Instrum. Meas., vol. 70, pp. 1–10, 2021.

[10] William C. Malm, "Introduction to Visibility," May 1999.

[11] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," Int. J. of Computer Vision, vol. 48, no. 3, pp. 233–254, 2002.

[12] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle pollution estimation based on image analysis," PloS One, vol. 11, no. 2, p. e0145955, 2016.

[13] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2341–2353, December 2011.

[14] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle Pollution Estimation Based on Image Analysis," PloS one, vol. 11, 2016.

[15] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-resolution air pollution mapping with Google street view cars: exploiting big data," Environmental Science & Technology, vol. 51, no. 12, pp. 6999–7008, 2017.

[16] Z. Pan, H. Yu, C. Miao, and C. Leung, "Crowdsensing air quality with camera-enabled mobile devices," 2017, p. 4728–4733.

[17] S. G. Narasimhan and S. K. Nayar, "Vision and the Atmosphere," Int. J. of Computer Vision, vol. 48, pp. 233–254, July 2002.

[18] ——, "Chromatic framework for vision in bad weather," vol. 1, February 2000, pp. 598 – 605 vol.1.

[19] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2341–2353, December 2011.

[20] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3522–3533, Nov 2015.

[21] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 4780–4788.

[22] N. P. Hyslop, "Impaired visibility: the air pollution people see," Atmospheric Environment, vol. 43, pp. 182–195, 2009.

[23] J. Wang, W. Nie, Y. Cheng, Y. Shen, X. Chi, J. Wang, X. Huang, Y. Xie, P. Sun, Z. Xu, X. Qi, H. Su, and A. Ding, "Light absorption of brown carbon in eastern china based on 3-year multi-wavelength aerosol optical property observations and an improved absorption ångström exponent segregation method," Atmospheric Chemistry and Physics, 2018.

[24] N. Hyslop, "Impaired visibility: the air pollution people see," Atmospheric Environment, vol. 43, pp. 182–195, January 2009.

[25] A. Leskinen, A. Ruuskanen, P. Kolmonen, Y. Zhao, D. Fang, Q. Wang, C. Gu, J. Jokiniemi, M.-R. Hirvonen, K. Lehtinen, S. Romakkaniemi, and M. Komppula, "The contribution of black carbon and non-bc absorbers on aerosol absorption coefficient in nanjing, china," Aerosol and Air Quality Research, vol. 20, pp. 590–605, 2020.

[26] J. Watson, J. Chow, L. Pritchett, L. Richards, D. Dietrich et al., Comparison of three measures of visibility extinction in Denver, Colorado, January 1989.

[27] J. Wang, W. Nie, Y. Cheng, Y. Shen, X. Chi et al., "Light absorption of brown carbon in eastern China based on 3-year multi-wavelength aerosol optical property observations at the SORPES station and an improved Absorption Angstrom exponent segregation method," Atmospheric Chemistry and Physics Discussions, pp. 1–31, January 2018.

[28] A. Hansen, H. Rosen, and T. Novakov, "The aethalometer — An instrument for the real-time measurement of optical absorption by aerosol particles," Science of The Total Environment, vol. 36, pp. 191–196, August 1983.

[29] L. Ran, Z. Deng, P. Wang, and X. Xia, "Black carbon and wavelength-dependent aerosol absorption in the north china plain based on two-year aethalometer measurements," Atmospheric Environment, vol. 142, July 2016.

[30] M. Xie, M. Hays, and A. Holder, "Light-absorbing organic carbon from prescribed and laboratory biomass burning and gasoline vehicle emissions," Scientific Reports, vol. 7, December 2017.

[31] V. Bernardoni, R. Pileci, L. Caponi, and D. Massabo, "The multi-wavelength absorption analyzer (mwaa) model as a tool for source and component apportionment based on aerosol absorption properties: Application to samples collected in different environments," Atmosphere, vol. 8, p. 218, November 2017.

[32] Z. Shi, J. Long, W. Tang, and C. Zhang, "Single image dehazing in inhomogeneous atmosphere," Optik - International J. for Light and Electron Optics, vol. 125, p. 3868–3875, August 2014.

[33] A. J. Preetham, P. Shirley, and B. Smits, "A practical analytic model for daylight," in Proc. Conf. on Computer Graphics and Interactive Techniques, ser. SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 91–100. [Online]. Available: http://dx.doi.org/10.1145/311535.311545

[34] S. Narasimhan and S. Nayar, "Interactive (de)weathering of an image using physical models," IEEE Workshop on Color and Photometric Methods in Computer Vision, vol. 10, 12 2015.

[35] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, "Deep photo: Model-based photograph enhancement and viewing," ACM Trans. Graph., vol. 27, no. 5, Dec. 2008. [Online]. Available: https://doi.org/10.1145/1409060.1409069

[36] D. B. Chenault and J. L. Pezzaniti, "Polarization imaging through scattering media," in Polarization Analysis, Measurement, and Remote Sensing III, D. B. Chenault, M. J. Duggin, W. G. Egan, D. H. Goldstein, W. G. Egan, and M. J. Duggin, Eds., vol. 4133, International Society for Optics and Photonics. SPIE, 2000, pp. 124 – 133. [Online]. Available: https://doi.org/10.1117/12.406619

[37] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, Dec 2001, pp. I–I.

[38] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in IEEE International Conference on Computer Vision, Dec 2013, pp. 617–624.

[39] Y. Cho, J. Jeong, and A. Kim, "Model assisted multi-band fusion for single image enhancement and applications to robot vision," IEEE Robotics and Automation Letters (RA-L) (with IROS), vol. 3, no. 4, pp. 2822–2829, 2018.

[40] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 1674–1682.

[41] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," Trans. Img. Proc., vol. 25, no. 11, pp. 5187–5198, Nov. 2016. [Online]. Available: https://doi.org/10.1109/TIP.2016.2598681

[42] W. Ren, S. Liu, H. Zhang, J. shan Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in ECCV, 2016.

[43] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2018, pp. 8202–8211.

[44] R. Malav, A. Kim, S. R. Sahoo, and G. Pandey, "Dhsgan: An end to end dehazing network for fog and smoke," in Computer Vision – ACCV 2018, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer International Publishing, 2019, pp. 593–608.

[45] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," CoRR, vol. abs/1803.08396, 2018. [Online]. Available: http://arxiv.org/abs/1803.08396

[46] J. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," IEEE Intelligent Transportation Systems Magazine, vol. 4, no. 2, pp. 6–20, Summer 2012.

[47] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," International Journal of Computer Vision, 08 2017.

[48] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[49] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in ECCV, 2012.

[50] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., vol. 1, June 2003, pp. I–I.

[51] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop, ser. NTIRE CVPR'18, 2018.

[52] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 492–505, Jan 2019.

[53] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-hazy: A dataset to evaluate quantitatively dehazing algorithms," in 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 2226–2230.

[54] Z. Li and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," in Computer Vision and Pattern Recognition, 2018.

[55] D. Berman, T. Treibitz, and S. Avidan, "Air-light estimation using haze-lines," Proc. Int. Conf. on Computational Photography, pp. 1–9, 2017.

[56] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle Pollution Estimation Based on Image Analysis," PloS one, vol. 11, 2016.

[57] Y. Li, J. Huang, and J. Luo, "Using user generated online photos to estimate and monitor air pollution in major cities," in Proc. Int. Conf. on Internet Multimedia Computing and Service.

[58] Y. Zhang, J. Cai, S. Wang, K. He, and M. Zheng, "Review of receptor-based source apportionment research of fine particulate matter and its challenges in China," Science of the Total Environment, vol. 586, pp. 917–929, 2017.

[59] G. Lin, J. Fu, D. Jiang, W. Hu, D. Dong, Y. Huang, and M. Zhao, "Spatio-temporal variation of PM2. 5 concentrations and their relationship with geographic and socioeconomic factors in China," Int. J. of Environmental Research and Public Health, vol. 11, no. 1, pp. 173–186, 2014.

[60] X. Zhao, W. Zhou, L. Han, and D. Locke, "Spatiotemporal variation in PM2. 5 concentrations and their relationship with socioeconomic factors in China's major cities," Environment Int., vol. 133, p. 105145, 2019.

[61] Y. Guan, M. C. Johnson, M. Katzfuss, E. Mannshardt, K. P. Messier, B. J. Reich, and J. J. Song, "Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google street view vehicles," J. of the American Statistical Association, vol. 115, no. 531, pp. 1111–1124, 2020.

[62] J. Wei, Z. Li, W. Xue, L. Sun, T. Fan, L. Liu, T. Su, and M. Cribb, "The ChinaHighPM10 dataset: generation, validation, and spatiotemporal variations from 2015 to 2019 across China," Environment Int., vol. 146, p. 106290, 2021.

[63] J. Khan, K. Kakosimos, O. Raaschou-Nielsen, J. Brandt, S. S. Jensen, T. Ellermann, and M. Ketzel, "Development and performance evaluation of new AirGIS–a GIS based air pollution and human exposure modelling system," Atmospheric environment, vol. 198, pp. 102–121, 2019.

[64] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu, and T. Chi, "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," Science of the Total Environment, vol. 654, pp. 1091–1099, 2019.

[65] G. Janssens-Maenhout, F. Dentener, J. Van Aardenne, S. Monni, V. Pagliari, L. Orlandini, Z. Klimont, J.-i. Kurokawa, H. Akimoto, T. Ohara et al., "EDGAR-HTAP: a harmonized gridded air pollution emission dataset based on national inventories," European Commission Publications Office, Ispra (Italy). JRC68434, EUR report, vol. 25, pp. 299–2012, 2012.

[66] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," Sensors and Actuators B: Chemical, vol. 129, no. 2, pp. 750–757, 2008.

[67] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, "Sensing the air we breathe: the OpenSense Zurich dataset," in AAAI Conf. on Artificial Intelligence, 2012.

[68] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," CoRR, vol. abs/1903.08558, 2019. [Online]. Available: http://arxiv.org/abs/1903.08558

[69] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in IEEE Conferene on Computer Vision and Pattern Recognition, 2016.

[70] T. Ueda, K. Yamada, and Y. Tanaka, "Underwater image synthesis from rgb-d images and its application to deep underwater image restoration," in 2019 IEEE International Conference on Image Processing (ICIP), Sep. 2019, pp. 2115–2119.

[71] C. Li, J. Guo, R. Cong, Y. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5664–5677, Dec 2016.

[72] H. Grimm and D. J. Eatough, "Aerosol measurement: the use of optical light scattering for the determination of particulate size distribution, and particulate mass, including the semi-volatile fraction," J. of the Air & Waste Management Association, vol. 59, no. 1, pp. 101–107, 2009.

[73] M. Kamionka, P. Breuil, and C. Pijolat, "Calibration of a multivariate gas sensing device for atmospheric pollution measurement," Sensors and Actuators B: Chemical, vol. 118, no. 1-2, pp. 323–327, 2006.

[74] G. Liu, J. Li, D. Wu, and H. Xu, "Chemical composition and source apportionment of the ambient PM2. 5 in Hangzhou, China," Particuology, vol. 18, pp. 135–143, 2015.

[75] Q. Jin, R. Ren, and L. Gong, "Research on the elemental characterization and source apportionment of PM2.5 in main urban area of Hangzhou," Chinese J. of Health Laboratory Technology, vol. 27, no. 22, pp. 3200–3205, 2017.

[76] J. Wu, C. Xu, Q. Wang, and W. Cheng, "Potential sources and formations of the PM2.5 pollution in urban Hangzhou," Atmosphere, vol. 7, no. 8, p. 100, 2016.

[77] Q. Zhu, Y. Liu, W. Xu, and M. Huang, "Analysis on the pollution characteristics and influence factors of pm2. 5 in guangzhou," Environ. Monit. China, vol. 29, no. 2, p. 15, 2013.

[78] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2341–2353, 2010.

[79] D. Park, D. K. Han, C. Jeon, and H. Ko, "Fast single image de-hazing using characteristics of RGB channel of foggy image," IEICE Trans. on Information and Systems, vol. 96, no. 8, pp. 1793–1799, 2013.

[80] J. Park, E. H. Park, J. J. Schauer, S.-M. Yi, and J. Heo, "Reactive oxygen species (ros) activity of ambient fine particles (pm2.5) measured in seoul, korea," Environment International, vol. 117, pp. 276–283, 2018.

[81] M. Rohrer, A. Flahault, and M. Stoffel, "Peaks of fine particulate matter may modulate the spreading and virulence of covid-19," Earth Systems and Environment, vol. 4, no. 4, pp. 789–796, 2020.

[82] S. M. Ali, F. Malik, M. S. Anjum, G. F. Siddiqui, M. N. Anwar, S. S. Lam, A.-S. Nizami, and M. F. Khokhar, "Exploring the linkage between pm2.5 levels and covid-19 spread and its implications for socio-economic circles," Environmental Research, vol. 193, p. 110421, 2021.

[83] R. K. Chakrabarty, P. Beeler, P. Liu, S. Goswami, R. D. Harvey, S. Pervez, A. van Donkelaar, and R. V. Martin, "Ambient pm2.5 exposure and rapid spread of covid-19 in the united states," Science of the Total Environment, vol. 760, p. 143391, 2021.

[84] B. Feenstra, V. Papapostolou, S. Hasheminassab, H. Zhang, B. Der Boghossian, D. Cocker, and A. Polidori, "Performance evaluation of twelve low-cost pm2.5 sensors at an ambient air monitoring site," Atmospheric Environment, vol. 216, p. 116946, 2019.

[85] X. Querol, A. Alastuey, S. Rodriguez, F. Plana, E. Mantilla, and C. R. Ruiz, "Monitoring of pm10 and pm2.5 around primary particulate anthropogenic emission sources," Atmospheric Environment, vol. 35, no. 5, pp. 845–858, 2001.

[86] Y. Lu, G. Giuliano, and R. Habre, "Estimating hourly pm2.5 concentrations at the neighborhood scale using a low-cost air sensor network: A los angeles case study," Environmental Research, vol. 195, p. 110653, 2021.

[87] Y. Yang, Z. Hu, K. Bian, and L. Song, "Imgsensingnet: Uav vision guided aerial-ground air quality sensing system," in IEEE INFOCOM 2019-IEEE Conf. on Computer Communications. IEEE, 2019, pp. 1207–1215.

[88] Z. Wang, S. Yue, and C. Song, "Video-based air quality measurement with dual-channel 3-d convolutional network," IEEE Internet Things J., vol. 8, no. 18, pp. 14 372–14 384, 2021.

[89] K. Gu, J. Qiao, and X. Li, "Highly efficient picture-based prediction of pm2.5 concentration," IEEE Trans. Ind. Electron., vol. 66, no. 4, pp. 3176–3184, 2019.

[90] J. Qiao, Z. He, and S. Du, "Prediction of pm 2.5 concentration based on weighted bagging and image contrast-sensitive features," Stochastic Environmental Research and Risk Assessment, vol. 34, no. 3-4, pp. 561–573, 2020.

[91] S. G. Narasimhan and S. K. Nayar, "Shedding light on the weather," in 2003 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1. IEEE, 2003, pp. I–I.

[92] M. L. Zamora, J. Rice, and K. Koehler, "One year evaluation of three low-cost pm2.5 monitors," Atmospheric Environment, vol. 235, p. 117615, 2020.

[93] S. Zhai, Y. Zhang, J. Huang, X. Li, W. Wang, T. Zhang, F. Yin, and Y. Ma, "Exploring the detailed spatiotemporal characteristics of pm2.5: Generating a full-coverage and hourly pm2.5 dataset in the sichuan basin, china," Chemosphere, vol. 310, p. 136786, 2023.

[94] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, "Air quality modelling using long short-term memory (lstm) over nct-delhi, india," Air Quality, Atmosphere & Health, vol. 12, pp. 899–908, 2019.

[95] J. Ma, Y. Ding, V. J. Gan, C. Lin, and Z. Wan, "Spatiotemporal prediction of pm2.5 concentrations at different time granularities using idw-blstm," IEEE Access, vol. 7, pp. 107 897–107 907, 2019.

[96] C. Guo, G. Liu, L. Lyu, and C.-H. Chen, "An unsupervised pm2.5 estimation method with different spatio-temporal resolutions based on kidw-tcgru," IEEE Access, vol. 8, pp. 190 263–190 276, 2020.

[97] T. Zhang and R. P. Dick, "Image-based air quality forecasting through multi-level attention," in 2022 IEEE Int. Conf. on Image Processing (ICIP), 2022, pp. 686–690.

[98] M.-S. Dao, K. Zettsu, and U. K. Rage, "Image-2-aqi: Aware of the surrounding air qualification by a few images," in Advances and Trends in Artificial Intelligence. From Theory to Practice. Springer, 2021, pp. 335–346.

[99] P. Su, Y. Liu, S. Tarkoma, A. Rebeiro-Hargrave, T. Petäjä, M. Kulmala, and P. Pellikka, "Retrieval of multiple atmospheric environmental parameters from images with deep learning," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[100] F. Weng, H. Huang, and X. Han, "Monitoring pm2.5 distributions over china from geostationary satellite observations," in IGARSS 2020 - 2020 IEEE Int. Geoscience and Remote Sensing Symposium, 2020, pp. 5581–5583.

[101] M. Wang, Y. Wang, F. Teng, S. Li, Y. Lin, and H. Cai, "Estimation and analysis of pm2.5 concentrations with npp-viirs nighttime light images: A case study in the chang-zhu-tan urban agglomeration of china," Int. J. Environ. Res. Public Health, vol. 19, no. 7, p. 4306, 2022.

[102] C.-H. Hsieh, K.-Y. Chen, M.-Y. Jiang, J.-J. Liaw, and J. Shin, "Estimation of pm 2.5 concentration based on support vector regression with improved dark channel prior and high frequency information in images," IEEE Access, vol. 10, pp. 48 486–48 498, 2022.

[103] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in Proc. IEEE Int. Conf. on Computer Vision, 2015, pp. 226–234.

[104] H. Koschmieder, Theorie der horizontalen Sichtweite. Keim & Nemnich, 1925.

[105] K. He, J. Sun, and X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1397–1409, 2012.

[106] A. K. Dubey and V. Jain, "Comparative study of convolution neural network's relu and leaky-relu activation functions," in Applications of Computing, Automation and Wireless Systems in Electrical Engineering. Springer, 2019, pp. 873–880.

[107] H. P. Government, "Hangzhou ecological environment status bulletin 2021," 2023. [Online]. Available: https://www.hangzhou.gov.cn/art/2022/6/2/art_1228974625_59058626.html

[108] E. M. Considine, C. E. Reid, M. R. Ogletree, and T. Dye, "Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network," Environmental Pollution, vol. 268, p. 115833, 2021.

[109] P.-Y. Wong, H.-J. Su, S.-C. C. Lung, and C.-D. Wu, "An ensemble mixed spatial model in estimating long-term and diurnal variations of pm2.5 in taiwan," Science of The Total Environment, p. 161336, 2023.

[110] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[111] H. Moosmüller, R. Chakrabarty, and W. Arnott, "Aerosol light absorption and its measurement: A review," Journal of Quantitative Spectroscopy and Radiative Transfer, vol. 110, no. 11, pp. 844–878, 2009.

[112] L. Xu, J. Zhang, X. Sun, S. Xu, M. Shan, Q. Yuan, L. Liu, Z. Du, D. Liu, D. Xu et al., "Variation in concentration and sources of black carbon in a megacity of china during the covid-19 pandemic," Geophysical research letters, vol. 47, no. 23, p. e2020GL090444, 2020.

[113] X. Jiang, H. Yao, S. Zhang, X. Lu, and W. Zeng, "Night video enhancement using improved dark channel prior," in 2013 IEEE Int. Conf. on Image Processing, 2013, pp. 553–557.

[114] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[115] 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. on Computer Vision, 2017, pp. 618–626.

[116] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017.

[117] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, p. 965–973.

[118] Z. Luo, J. Huang, K. Hu, X. Li, and P. Zhang, "Accuair: Winning solution to air quality prediction for kdd cup 2018," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, p. 1842–1850.

[119] K. Gu, J. Qiao, and X. Li, "Highly efficient picture-based prediction of PM2. 5 concentration," IEEE Transactions on Industrial Electronics, vol. 66, no. 4, pp. 3176–3184, 2018.

[120] T. Zhang and R. P. Dick, "Estimation of multiple atmospheric pollutants through image analysis," in Proceedings of the International Conference on Image Processing, 2019, pp. 2060–2064.

[121] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3522–3533, November 2015.

[122] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European Conference on Computer Vision, September 2018.

[123] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in Proceedings of the IEEE International Conference on Computer Vision, 2019.

[124] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112. [Online]. Available: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[125] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[126] N. Jacobs, N. Roman, and R. Pless, "Consistent Temporal Variations in Many Outdoor Scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, March 2007.

[127] N. Jacobs, W. Burgin, N. Fridrich, A. Abrams, K. Miskell, B. H. Braswell, A. D. Richardson, and R. Pless, "The global network of outdoor webcams: Properties and applications," in ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 2009, pp. 111–120.

[128] E. Hinkley and T. Zajkowski, "Usda forest service-nasa: Unmanned aerial systems demonstrations - pushing the leading edge in fire mapping," Geocarto International, vol. 26, pp. 103–111, 04 2011.

[129] V. Sherstjuk, M. Zharikova, and I. Sokol, "Forest fire monitoring system based on uav team, remote sensing, and image processing," in IEEE International Conference on Data Stream Mining Processing (DSMP), 08 2018, pp. 590–594.

[130] C.-H. Chen, P.-H. Wu, and Y.-C. Chiou, "An early fire-detection method based on image processing," Proceedings of the International Conference on Image Processing, vol. 3, pp. 1707–1710, 2004.

[131] T. Çelik, H. Özkaramanli, and H. Demirel, "Fire and smoke detection without sensors: Image processing based approach," Proceedings of the European Signal Processing Conference, pp. 1794–1798, 2007.

[132] X. Zhang, H. Xiong, W. gang Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1134–1142, 2016.

[133] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[134] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," Neurocomputing, vol. 288, pp. 30–42, 2018.

[135] F. A. Hossain, Y. M. Zhang, and M. A. Tonima, "Forest fire flame and smoke detection from uav-captured images using fire-specific color features and multi-color space local binary pattern," Journal of Unmanned Vehicle Systems, vol. 8, no. 4, pp. 285–309, 2020.

[136] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in MICCAI, 2015.

[137] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 6230–6239, 2017.

[138] H.-S. Choi, M. Jeon, K. Song, and M. joo Kang, "Semantic fire segmentation model based on convolutional neural network for outdoor image," Fire Technology, pp. 1–15, 2021.

[139] K. Song, H.-S. Choi, and M. joo Kang, "Squeezed fire binary segmentation model using convolutional neural network for outdoor images on embedded device," Machine Vision and Applications, vol. 32, 2021.

[140] N. I. F. Center, "Wildland fire statistics," https://www.nifc.gov/fireInfo/fireInfo_statistics.html, accessed 25 May 2022.

[141] A. L. Westerling, "Increasing western us forest wildfire activity: sensitivity to changes in the timing of spring," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 371, 2016.

[142] M. Burke, A. Driscoll, S. Heft-Neal, J. Xue, J. A. Burney, and M. Wara, "The changing risk and burden of wildfire in the united states," Proceedings of the National Academy of Sciences of the United States of America, vol. 118, 2021.

[143] J. T. Abatzoglou and A. P. Williams, "Impact of anthropogenic climate change on wildfire across western us forests," Proceedings of the National Academy of Sciences, vol. 113, pp. 11 770 – 11 775, 2016.

[144] A. D. Syphard and J. E. Keeley, "Location, timing and extent of wildfire vary by cause of ignition," International Journal of Wildland Fire, vol. 24, no. 1, pp. 37–47, 2015.

[145] ABC 10 News San Diego. [Online]. Available: https://www.10news.com/news/here-s-what-the-wildfires-devastation-in-california-looks-like#id2

[146] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

[147] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 834–848, 2018.

[148] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision, 2018.

[149] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7151–7160, 2018.

[150] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3141–3149, 2019.

[151] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," ArXiv, vol. abs/1809.00916, 2018.

[152] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, H. Shi, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," 2019 IEEE International Conference on Computer Vision, pp. 603–612, 2019.

[153] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, pp. 2547–2560, 2020.

[154] Q. Song, K. Mei, and R. Huang, "Attanet: Attention-augmented network for fast and accurate scene parsing," in AAAI Conference on Artificial Intelligence, 2021.

[155] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," International Journal of Computer Vision, vol. 129, pp. 3051–3068, 2021.

[156] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022 IEEE Conference on Computer Vision and Pattern Recognition, pp. 11 966–11 976, 2022.

[157] T. Wittkopp, C. Hecker, and D. Opitz, "Cargo fire monitoring system (CFMS) for the visualisation of fire events in aircraft cargo holds." in International Conference on Automatic Fire Detection "AUBE '01", 12th. Proceedings. National Ir 2001-03-25 2001. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=916859

[158] C. Lai, Y. Jie-Ci, and Y. Chen, "A real time video processing based surveillance system for early fire and flood detection," in 2007 IEEE Instrumentation and Measurement Technology Conference IMTC 2007, 06 2007, pp. 1 – 6.

[159] D. Han and B. Lee, "Development of early tunnel fire detection algorithm using the image processing," in Advances in Visual Computing, G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 39–48.

[160] X. Han, J. Jin, M. Wang, W. Jiang, L. Gao, and L. Xiao, "Video fire detection based on gaussian mixture model and multi-color features," Signal, Image and Video Processing, vol. 11, pp. 1419–1425, 2017.

[161] K. Trambitckii, K. Anding, V. Musalimov, and G. Linß, "Colour based fire detection method with temporal intensity variation filtration," in 2014 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security, 2015.

[162] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," IEEE Access, vol. 5, pp. 18 429–18 438, 2017. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2747399

[163] K. Gu, Z. Xia, J. Qiao, and W. Lin, "Deep dual-channel neural network for image-based smoke detection," IEEE Transactions on Multimedia, vol. 22, no. 2, pp. 311–323, 2019.

[164] S. Saponara, A. Elhanashi, and A. Gagliardi, "Real-time video fire/smoke detection based on CNN in antifire surveillance systems," Journal of Real-Time Image Processing, vol. 18, no. 3, pp. 889–900, 2021.

[165] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 7, pp. 1419–1434, 2018.

[166] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7151–7160, 2018.

[167] T. Toulouse, L. Rossi, A. Campana, T. Çelik, and M. A. Akhloufi, "Computer vision for wildfire research: An evolving image dataset for processing and analysis," Fire Safety Journal, vol. 92, pp. 188–194, 2017.

[168] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," ArXiv, 2017.

[169] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3902–3911, 2019.

[170] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. R. Fu, "Reverse attention-based residual network for salient object detection," IEEE Transactions on Image Processing, vol. 29, pp. 3763–3776, 2020.

[171] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, pp. 5367–5376, 2020.

[172] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, pp. 3252–3261, 2018.

[173] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "Ern: Edge loss reinforced semantic segmentation network for remote sensing images," Remote Sensing, vol. 10, p. 1339, 2018.

[174] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," ArXiv, vol. abs/1612.01337, 2016.

[175] B. C. Hansen and R. F. Hess, "Structural sparseness and spatial phase alignment in natural scenes." Journal of the Optical Society of America, vol. 24 7, pp. 1873–85, 2007.

[176] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," 2021 IEEE Conference on Computer Vision and Pattern Recognition, pp. 14 378–14 387, 2021.

[177] Z. Zhang, X. Zhang, C. Peng, D. Cheng, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," ArXiv, vol. abs/1804.03821, 2018.

[178] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," ArXiv, vol. abs/1808.00897, 2018.

[179] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, pp. 2547–2560, 2020.

[180] A. Mohan, G. Sapiro, and E. Bosch, "Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images," IEEE Geoscience and Remote Sensing Letters, vol. 4, pp. 206–210, 2007.

[181] R. Qin and W. Fang, "A hierarchical building detection method for very high resolution remotely sensed images combined with dsm using graph cut optimization," Photogrammetric Engineering and Remote Sensing, vol. 80, pp. 873–883, 2014.

[182] A. Puissant, S. Rougier, and A. J. Stumpf, "Object-oriented mapping of urban trees using random forest classifiers," International Journal of Applied Earth Observation and Geoinformation, vol. 26, pp. 235–245, 2014.

[183] M. Volpi, D. Tuia, F. Bovolo, M. F. Kanevski, and L. Bruzzone, "Supervised change detection in vhr images using contextual information and support vector machines," International Journal of Applied Earth Observation and Geoinformation, vol. 20, pp. 77–85, 2013.

[184] U. Maulik and I. Saha, "Automatic fuzzy clustering using modified differential evolution for image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, pp. 3503–3510, 2010.

[185] Q. Liu, M. C. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, pp. 6309–6320, 2020.

[186] R. Li, S. Zheng, C. Duan, and J. Su, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.

[187] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 2481–2495, 2015.

[188] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5168–5177, 2017.

[189] R. Li, J. Su, C. Duan, and S. Zheng, "Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[190] ISPRS. (2022) 2d semantic labeling contest-potsdam. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[191] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," Trans. Img. Proc., vol. 25, no. 11, pp. 5187–5198, November 2016. [Online]. Available: https://doi.org/10.1109/TIP.2016.2598681

[192] S. C. Anenberg, L. W. Horowitz, D. Q. Tong, and J. J. West, "An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling," Environmental Health Perspectives, vol. 118, pp. 1189 – 1195, 2010.

[193] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-resolution air pollution mapping with google street view cars: Exploiting big data." Environmental science & technology, vol. 51 12, pp. 6999–7008, 2017.

[194] D. A. Paolella, C. W. Tessum, P. J. Adams, J. S. Apte, S. E. Chambliss, J. D. Hill, N. Z. Muller, and J. D. Marshall, "Effect of model spatial resolution on estimates of fine particulate matter exposure and exposure disparities in the united states," Environmental Science & Technology Letters, 2018.

[195] H. ye Tao, J. Xing, H. Zhou, J. E. Pleim, L. Ran, X. Chang, S. Wang, F. Chen, H. Zheng, and J. Li, "Impacts of improved modeling resolution on the simulation of meteorology, air quality, and human exposure to pm2.5, o3 in beijing, china," Journal of Cleaner Production, vol. 243, p. 118574, 2020.

[196] S. Yarza, L. Hassan, A. Shtein, D. Lesser, L. Novack, I. Katra, I. Kloog, and V. Novack, "Novel approaches to air pollution exposure and clinical outcomes assessment in environmental health studies," 2020.

[197] L. Gerharz, O. Klemm, A. V. Broich, and E. J. Pebesma, "Spatio-temporal modelling of individual exposure to air pollution and its uncertainty," Atmospheric Environment, vol. 64, pp. 56–65, 2013.

[198] J. Tan, Y. Zhang, W. Q. Ma, Q. Yu, J. M. Wang, and L. Chen, "Impact of spatial resolution on air quality simulation: A case study in a highly industrialized area in shanghai, china," Atmospheric Pollution Research, vol. 6, pp. 322–333, 2015.

[199] G. Adami, M. Pontalti, G. Cattani, M. Rossini, O. Viapiana, G. Orsolini, C. Benini, E. Bertoldo, E. Fracassi, D. Gatti, and A. Fassio, "Association between long-term exposure to air pollution and immune-mediated diseases: a population-based cohort study," RMD Open, vol. 8, 2022.