Unsupervised Contrastive Masking for Visual Haze Classification

Jingyu Li
Shandong University
jingyu_lee@mail.sdu.edu.cn

Haokai Ma
Shandong University
mahaokai@mail.sdu.edu.cn

Xiangxian Li
Shandong University
xiangxian_lee@mail.sdu.edu.cn

Zhuang Qi
Shantou University
97qizhuang@gmail.com

Lei Meng*
Shandong University
lmeng@sdu.edu.cn

Xiangxu Meng
Shandong University
mxx@sdu.edu.cn

ABSTRACT

Haze classification has gained much attention recently as a cost-effective solution for air quality monitoring. Different from conventional image classification tasks, it requires the classifier to capture the haze patterns of different severity degrees. Existing efforts typically focus on the extraction of effective haze features, such as the dark channel and deep features. However, it is observed that the light-haze images are often mis-classified due to the presence of diverse background scenes. To address this issue, this paper presents an unsupervised contrastive masking (UCM) algorithm to segment the haze regions without any supervision, and develops a dual-channel model-agnostic framework, termed magnifier neural network (MagNet), to effectively use the segmented haze regions to enhance the learning of haze features by conventional deep learning models. Specifically, MagNet employs the haze regions to provide the pixel- and feature-level visual information via three strategies, including Input Augmentation, Network Constraint, and Feature Enhancement, which work as a soft-attention regularizer to alleviates the trade-off between capturing the global scene information and the local information in the haze regions. Experiments were conducted on two datasets in terms of performance comparison, parameter estimation, ablation studies, and case studies, and the results verified that UCM can accurately and rapidly segment the haze regions, and the proposed three strategies of MagNet consistently improve the performance of the state-of-the-art deep learning backbones.

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**.

KEYWORDS

Haze classification, Unsupervised masking, Contrastive map, Magnifier network

*Corresponding author

ICMR '22, June 27–30, 2022, Newark, NJ, USA.

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531370

ACM Reference Format:
Jingyu Li, Haokai Ma, Xiangxian Li, Zhuang Qi, Lei Meng, and Xiangxu Meng. 2022. Unsupervised Contrastive Masking for Visual Haze Classification. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531370

1 INTRODUCTION

Haze classification [5, 25, 29] is an emerging AI-powered application in the computer version domain, which has been widely used in air quality estimation [11, 19] and visibility estimation in autonomous driving [7, 8]. In contrast to conventional image classification tasks, the classification of haze images aims to estimate the severity levels of the visual haze, instead of distinguishing between different objects. Existing studies usually look into extracting effective haze features [2, 16, 22, 28]. However, it has been observed that even the state-of-the-art convolutional neural networks (CNNs) [16, 27, 30] encounter issues in distinguishing the haze classes in light haze settings, caused by the diversity of background scenes. Therefore, new solutions are needed to alleviate the influence of background objects when learning the visual features of haze.

Existing methods for haze classification lie in three main categories, including threshold-based models, handcrafted-feature-based classifiers, and deep learning algorithms. The threshold-based models develop mathematical equations to compute the values of haze degrees directly from haze images, such as color [3, 9, 18] and dark channel prior [7, 14]. This line of research heavily depends on observation data and the computation may be cumbersome, leading to an unstable performance to classifying data from different domains. The handcrafted-feature-based classifiers [24, 29, 31] usually focus on extracting effective visual features and use the classifiers such as SVM to classify the haze classes. Such data-driven approach alleviates introducing manual biases, but it introduces the need of feature engineering with high computational cost for feature extraction [2, 22, 28]. Finally, the deep learning algorithms use CNNs [32, 33] with network ensembles [22, 27], multi-branch training [17, 28], and pre-training [2, 5], to classify haze images in an end-to-end manner, which typically achieve much better performance than the threshold-based and the handcrafted-feature-based methods. However, it has been observed that these models usually perform worse on the classification of light haze images, due to the fact that the background scene objects work as spurious causal features in classification [16, 27, 30] and the lack of integration of corresponding features [6, 12, 20].

To address the aforementioned problems, this paper presents an unsupervised contrastive masking (UCM) algorithm to segment the

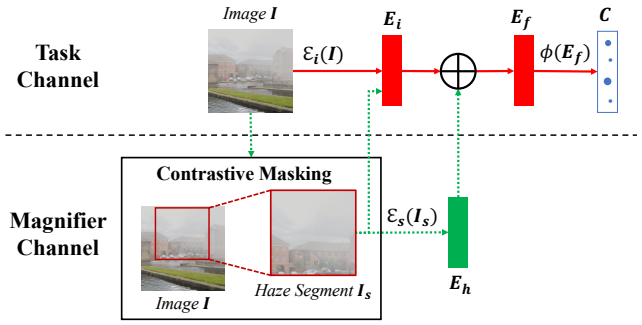


Figure 1: Illustration to UCM-MagNet for haze classification. In addition to the task channel $\mathbf{I} \mapsto \mathbf{E}_i$, UCM-MagNet extracts visual features from the haze portion segmented by the UCM module $\mathbf{I} \mapsto \mathbf{I}_s \mapsto \mathbf{E}_h$. The final feature $\mathbf{E}_f$ is obtained by a fusion of the features from both channels for haze classification $\mathbf{E}_i \oplus \mathbf{E}_h = \mathbf{E}_f \mapsto C$.

haze regions in a cost-effective manner, and develops a dual-channel model-agnostic framework, termed magnifier neural network (MagNet), to integrate the segmented haze regions into conventional CNNs to improve their classification performance. Figure 1 shows the entire framework, termed UCM-MagNet. Specifically, UCM considers the knowledge prior of haze colors and employs a contrastive masking method to distinguish the haze regions from the background scenes; MagNet is a dual-channel network that uses the extracted haze regions as a magnifier to enhance the learning of haze features in terms of the pixel- and feature-level visual information, achieved by input augmentation, network constraint, and feature enhancement. The Input Augmentation module serves as a magnifier that complement the input image with the magnified haze regions, which makes the model to better capture the features of the haze information; the Network Constraint module works as a soft-attention regularizer [13, 26] to the feature encoder in the task channel by aligning its intermediate visual features to the haze features produced by the encoder in the magnifier channel, and this guides the encoder in the task channel to pay more attention to the haze regions; and the Feature Enhancement module alleviates the trade-off between capturing the global and local information in the haze images by fusing the features learned from both the task and magnifier channels.

Experiments were conducted on the Hazel-level [14] and the Haze-wild datasets, where Hazel-level dataset contains 3024 images in 9 classes, and Haze-wild dataset contains 100,000 images in 10 classes. Notebly, the Haze-wild dataset is created by us considering the shortage of publicly-accessible large-scale benchmarking datasets in the literature, and it will be release to the community. The performance of UCM-MagNet is evaluated with performance comparison, parameter estimation, ablation studies, and the case studies. The results show that UCM-MagNet may achieve consistent performance gains as compared with the CNN backbones.

To summarize, this paper includes three main contributions:

(1) An unsupervised haze segmentation algorithm, termed UCM, is proposed, which does not require any supervision and can accurately segment the haze regions in a fast speed and with mild parameter settings.
(2) A model-agnostic framework, termed MagNet, is proposed to effectively integrate UCM into conventional CNNs and can effectively use the haze segments for performance improvement.
(3) A large-scale dataset has been created for haze classification, named Haze-Wild. As compared with existing datasets, Haze-wild contains a wide range of outdoor scenes and employs the state-of-the-art depth estimation technique to create the haze. It will be release to the research community.

2 RELATED WORKS

Existing studies on haze classification can be categorized into the three types according to their different methods in haze feature extraction and classification, as illustrated below.

2.1 Threshold-based Methods

Threshold-based methods need to model the haze functions by theory or observation and then substitute statistical or image processing information to determine the class according to thresholds [1]. The information includes the lowest/highest pixel value of the original RGB image [9, 18], dark channel priors [7], depth map [14], transmitted image [3], etc., through logarithm, division, and pooling calculation parameters for operations such as transformation [14]. These methods are limited by the construction of specific functions and have poor scalability.

2.2 Classifiers with Handcrafted Features

Manually extracting features from images through feature engineering and training the model with machine learning methods. Such methods usually manually extract color histograms [24, 29], color model parameters [31], regions of interest (ROI) and power spectral slopes [16] from raw RGB images, depth maps, dark channel maps [30] as features, then adapting a regression model such as a support vector regression (SVR) to predict the degree of haze [16] and find the corresponding class from the index, or directly classify images by multiple or cascaded support vector machine (SVM) [30]. These methods avoid the construction of specific functions, but the selection of features limits the precision of model classification.

2.3 Deep Learning Methods

Deep learning is an end-to-end approach, inputting images and the model would learn to extract features and output classification results autonomously. Since it is hard to train models from complete haze images, related works improve the model training framework: model ensemble methods [22, 27] achieve feature enhancement by training multiple basic models and adapting a meta-learner to learn to fusion basic models' output; the multi-branch method [17, 28] proposes multiple classifiers in different training branches to get better predictions. There are also ways of pre-training [2, 5], multi-task training [32, 33] and feature fusion [15, 21]. The end-to-end method does not require feature engineering, however, most of the current works lack constraints on the background problem of haze classification.

In summary, end-to-end methods have made a progress currently, but still require a lot of computing time and computation for model ensembling or pre-training, and there is lacking methods to handle the noise from the background of haze images.

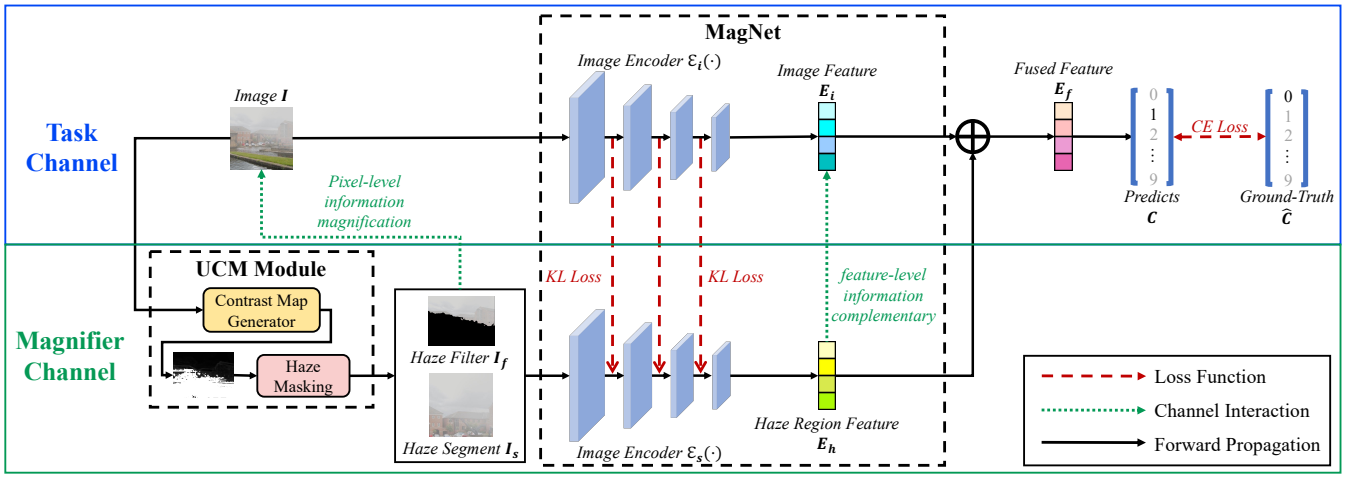


Figure 2: The illustration of UCM-MagNet unsupervised segments the haze portion of the original image to assist in haze classification. In forward propagation, Image Encoder $\varepsilon_s(\cdot)$ extracts the visual features from the image I and the haze region $\mathbf{I}_s$ segmented by the UCM module for haze classification. In addition to classification loss $\mathcal{L}_{cls}$, KL-divergence$\mathcal{L}_{sim}$ is also introduced to enhance the constraints on the feature extraction process for image I. With the pixel-level information magnification and the feature-level information complementary, the magnifier channel can enable the task channel to focus more on the haze partition and strengthen its haze classification performance at multiple levels and perspectives

3 PROBLEM FORMULATION

This paper investigates the use of haze segments to enhance visual haze classification. As shown in Figure 2, compared with the conventional setting, Magnifier Channel is added to obtain the haze region feature. Given a dataset including haze images $\mathcal{I} = \{\mathbf{I}_i \,|i = 1, ..., N\}$ and corresponding labels of J classes $\tilde{C} = \{c_j \,|j = 1, ..., J\}$. Two image encoders $\varepsilon_s(\cdot)$ and $\varepsilon_i(\cdot)$ are trained to learn the haze region feature and the original image feature, denoted as $\mathbf{E}_h$ and $\mathbf{E}_i$, respectively. And the fused feature $\mathbf{E}_f$ is then mapped to prediction label. Finally, the Cross Entropy loss function about prediction label C and ground-truth $\tilde{C}$ is minimized to modify image encoder, i.e., minimizing $CE(C, \tilde{C})$. Meanwhile, we try to make the encoder of the task channel learn from the encoder of the magnifier channel by minimizing the KL-divergence $KL(\mathbf{E}_i \,||\, \mathbf{E}_{fi})$.

As shown in Ficture 2, there are three main processes:

- **Haze filter:** UCM uses three small modules to obtain contrast map $\mathbf{I}_c$, denoising map $\mathbf{I}_{de}$ and haze filter(segment) map $\mathbf{I}_f(\mathbf{I}_s)$ in order, i.e. $\mathbf{I} \mapsto \mathbf{I}_c \mapsto \mathbf{I}_{de} \mapsto \mathbf{I}_f(\mathbf{I}_s)$.
- **Feature extraction:** MagNet uses two encoder to extract features independently in dual channels: $\mathbf{I} \mapsto \mathbf{E}_i$ in task channel and $\mathbf{I} \mapsto \mathbf{I}_f(\mathbf{I}_s) \mapsto \mathbf{E}_h$ in magnifier channel.
- **Haze image label prediction:** The final prediction on haze classes C is obtained by a fused feature obtained from both the task and magnifier channels, i.e., $\mathbf{E}_i \oplus \mathbf{E}_h = \mathbf{E}_f \mapsto C$.

Conventional haze detection algorithms typically need to design structure $\mathcal{N}$, and use the dataset $\mathcal{T}$ with haze region boundary B_T for training, i.e. $B_T = \mathcal{N}(\mathcal{T})$, which is a challenge in itself. This motivates this study to propose the UCM module, which adaptively extracts haze region for any dataset without a training process. Subsequently, the dual-channel design of MagNet makes feature fusion possible to make the model pay more attention to haze areas. The fusion feature not only reduces the interference of irrelevant background, but also makes up for the lack of feature information of haze region.

4 TECHNIQUE

4.1 Framework Overview

As shown in Figure 2, UCM-MagNet framework has two channels: Task Channel and Magnifier Channel. These channels can be divided into two main modules: UCM module and MagNet module. UCM module generates haze region image which will be used as the input data for Magnifier channel in MagNet. MagNet module can use a variety of methods to classify with enlarged image features and original features.

4.2 UCM Module

As shown in the Figure 3, the UCM algorithm contains three main sub-modules: Contrast Module, Denoised Module, and Mask Module. UCM can enlarge the haze region of image I and get the haze region image I_h which contains the haze filter map I_f the haze segment map I_s, both of which can be used as the haze region.

4.2.1 Contrast Map Module. This module uses the information from the input image **I** to obtain a preliminary comparison map containing the haze region $\mathbf{I}_c$. There are two main procedures to use these information in Contrast module.

- **Calculate the Dark/Bright Channel Images.** This procedure is used to count the pixel information in the image **I**. Image **I** is processed by channel filter $\mathbf{F}_b$ and $\mathbf{F}_d$ to get the corresponding channel map $\mathbf{I}_b$, $\mathbf{I}_d$. The bright channel filter $\mathbf{F}_b$ and the dark channel filter $\mathbf{F}_d$ use the equations below:

$$\mathbf{F}_b(x) = \max_{y \in \Omega(x)} \left(\max_{c \in \{R,G,B\}} \mathbf{I}^c(y) \right), \tag{1}$$

$$\mathbf{F}_d(x) = \min_{y \in \Omega(x)} \left(\min_{c \in \{R,G,B\}} \mathbf{I}^c(y) \right), \tag{2}$$

where x is one of the pixel in this image, $\mathbf{I}^c(\cdot)$ denotes the c channel of image **I**, $\Omega(x)$ means the neighboring pixels of x.

- **Calculate the Contrast Map.** This procedure uses the statistical information in the previous step to obtain the comparison map

$\mathbf{I}_c$. We subtract $\mathbf{I}_b$ and $\mathbf{I}_d$ to get a preliminary comparison map $\mathbf{v}_c$. After that we get the comparison map $\mathbf{I}_c$ by Eqs. 3 and 4.

$$I_{\mathrm{c}}(x) = \begin{cases} 0 & \text{if } v_{\mathrm{c}} < v_{\mathrm{threshold}} \\ 255 & \text{otherwise} \end{cases}, \tag{3}$$

$$I_{\mathrm{c}}(x) = \begin{cases} 255, & \text{if } x > v_{\mathrm{threshold}} \text{ or } \mathrm{img}_{\mathrm{mean}} < \mathrm{gray} \\ x, & \text{otherwise} \end{cases}, \tag{4}$$

where $v_{threshold} = \frac{1}{2}\left(mean(v_c) + median(v_c)\right)$, $v_{\mathrm{c}} = I_{\mathrm{b}}(x) - I_d(x)$, $\mathrm{img}_{\mathrm{mean}} = \frac{1}{3}\sum_{c\in\{R,G,B\}}(I^c)$ and $\mathrm{gray} = 0.3I^R + 0.33I^G + 0.45I^B$.

4.2.2 Denoised Map Module. The input to this module is the comparison map $\mathbf{I}_c$.In this module $\mathbf{I}_c$ will be filtered out of scattered haze areas. We set a filter $\mathbf{F}_{de}$ to select the most frequent pixel value in a range as the pixel value of each pixel point in this range. So we can get the denoised map $\mathbf{I}_{de}$ by using Eqs. 5.

$$I_{\mathrm{de}}(x) = \forall_{y\in\Omega(x)}\mathbf{F}_{de}(y), \tag{5}$$

where x is a pixel. $\mathbf{F}_{de}$ denotes the method of denoising as we described.

4.2.3 Haze Mask Module. This module use the position information of the denoised map $\mathbf{I}_{de}$ to get the haze filter map $\mathbf{I}_f$ and haze segment map $\mathbf{I}_s$. There are three main procedures in Mask module.

- **Calculate the Mask matrix.** The matrix Mask stores the position information of the haze region in the denoised map **D**. We set a filter $\mathbf{F}_m$ to get this information. Filter $\mathbf{F}_m$ is defined as

$$\mathbf{F}_m(x) = \max_{y\in\Omega(x)} \mathbf{I}_{de}(y), \tag{6}$$

  where x is a pixel. Now we can get the Mask matrix by the following Eqs. 7.

$$\mathbf{Mask}(x) = \begin{cases} 1 & \text{if } \mathbf{F}_m(x) < 255 \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

- **Get the haze filter map.** In the previous step we obtained the matrix **Mask**. We can use it to generate the haze filter map $\mathbf{I}_f$ by the following Eqs. 8.

$$I_{\mathrm{f}}(x) = \mathbf{I}(x) \times \mathbf{Mask}(x). \tag{8}$$

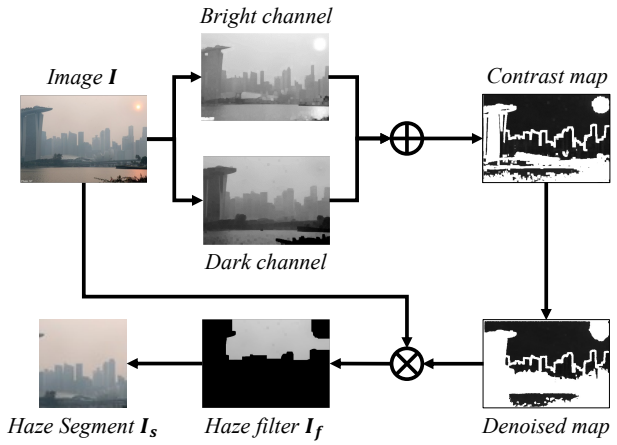
- **Get the haze segment map.** In the filter map $\mathbf{I}_f$, if a pixel is in the haze region, its value must be 0. We can get the bound values of the haze region **Bound** in the filter map $\mathbf{I}_f$ by the Eqs. 9.

$$Bound = [\min_{\mathbf{I}(x,y)>0}(x), \max_{\mathbf{I}(x,y)>0}(x), \min_{\mathbf{I}(x,y)>0}(y), \max_{\mathbf{I}(x,y)>0}(y)], \tag{9}$$

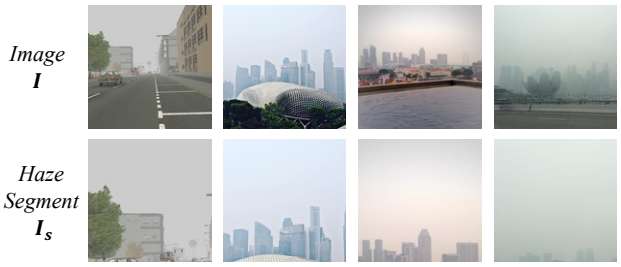
  where x, y is the position of a pixle. The haze segment map $\mathbf{I}_s$ can be obtaioned by using **Bound** to crop the Image **I**.

4.3 MagNet Module

Too much or too little background information will have an impact on the haze classification results. To solve this issue, the MagNet module receives inputs from both channels and processes them using different methods. As shown in the Figure 4, there are three approaches that are used to fuse the haze region image I_h extracted by UCM module and original image I to achieve feature-level information complementary, including Input Augmentation, Network Constraint, and Feature Enhancement.



(a)



(b)

Figure 3: (a) Illustration of the UCM module, which segments haze region $\mathbf{I}_s$ from original image I based on the contrastive information between the bright and dark channels. (b) Some examples of the haze regions $\mathbf{I}_s$ segmented from the image I by the UCM module.

4.3.1 Input Augmentation. Using a multi-channel hybrid map containing the haze region image extracted by UCM module and the original image to achieve Input Augmentation, as shown in Figure 4(a). A CNN network is then implemented to extract the feature $\mathbf{E}_a$, it consists of the original image feature and the haze region feature, and is used for classification.

To obtain the augmented image $\mathbf{I}_a$, the haze region image $\mathbf{I}_h$ and original image **I** are concatenated in input set. And this operation can be implemented by

$$\mathbf{I}_a = concat(\mathbf{I}, \mathbf{I}_h). \tag{10}$$

4.3.2 Network Constraint. The Network Constraint method defines a new loss function $\mathcal{L}_{sim}$ to guide the update of the image encoder, the process is illustrated in Figure 4(b). It is used to extract as many features about the haze region as possible from the original image. In the haze filter image, the non-haze areas are shown as black, so the CNN network can only extract features from the haze regions. In this way, the features extracted from the original image are as close as possible to the haze region features to achieve the goal of the Network Constraint method. Notably, using the filtered image as haze region image instead of the haze segment image in this section, since the location of the haze region in the filter map is the same as the original image.

In this method, the input data contains an original image $\mathbf{I}$ and a haze filter map $\mathbf{I}_f$. Using the same image encoder in both channels to extract features for $\mathbf{I}$ and $\mathbf{I}_f$, denoted as $\mathbf{E}_i$ and $\mathbf{E}_{fi}$, respectively. Finally, the loss $\mathcal{L}_{sim}$ can be calculated by the Eqs. 11.

$$\mathcal{L}_{sim} = KL(\mathbf{E}_i \mid\mid \mathbf{E}_{fi}). \tag{11}$$

4.3.3 Feature Enhancement. The aim of the Feature Enhancement method is to complement the features of the original map and haze region to perform a more accurate classification. Feature extraction using two channels can retain the feature information of the original image, and supplement it with the feature information of the haze area at the feature-level. Both channels use the same CNN network for feature extraction, so that each channel can learn the feature information of the other channel and pay more attention to the feature information that the channel can provide.

In this method, both the original image and the two haze region maps extracted by UCM can be fused for feature-level information complementation. As shown in the Figure 4(c), MagNet uses two channels to extract features from $\mathbf{I}$ and $\mathbf{I}_h$ to obtain features $\mathbf{E}_i$, $\mathbf{E}_h$ respectively. We use the following Eqs. 12 to $\mathbf{E}_f$.

$$\mathbf{E}_f = concat(\mathbf{E}_i, \mathbf{E}_h). \tag{12}$$

4.4 Training Strategies

UCM-MagNet consists of three main processes, including (a) segmenting the haze portion of the image, i.e., $\mathbf{I} \mapsto \mathbf{I}_f \mapsto \mathbf{I}_s$, (b) extracting visual features independently in dual channels, i.e., $\mathbf{I} \mapsto \mathbf{E}_i$ in the task channel and $\mathbf{I} \mapsto \mathbf{I}_f(\mathbf{I}_s) \mapsto \mathbf{E}_h$ in the magnifier channel, (c) fusing visual features from two channels for haze classification, i.e., $\mathbf{E}_i \oplus \mathbf{E}_h = \mathbf{E}_f \mapsto C$. These three processes are constantly interactive and interdependent. Ideally, they are in a state of dynamic equilibrium, and it is non-trivial to train them simultaneously. Therefore, we use two strategies to guarantee a smooth training effect:

- **Segmenting the haze region $\mathbf{I}_s$ from the original image I:** That is, before training MagNet, we first train the UCM module to segment the haze region $\mathbf{I}_s$ from the original image $\mathbf{I}$. This module's two critical parameters can significantly affect its segmentation results, namely the gray threshold $gray$ and the filter ratio F_m. The effects of the different values are described in detail in Section 5.4.
- **Extracting and fusing visual features for haze classification:** MagNet is optimized with two groups of loss terms, including losses for haze classification, i.e. $\mathcal{L}_{cls}$ and losses for interactive information between task channel and magnifier channel, i.e. $\mathcal{L}_{sim}$. We proposed two loss tactics used for the above three fusion methods described in Section 4.3:
  - For the Input Augmentation method and the Feature Enhancement method, MagNet is optimized by minimizing loss $\mathcal{L}_{cls}$.

$$\mathcal{L}_{cls} = CE\left(S, S^{'}\right) \tag{13}$$

  - For the Network Constrained method, the $\mathcal{L}_{cls}$ and $\mathcal{L}_{sim}$ are combined by a certain coefficient m to get loss $\mathcal{L}_{com}$, we train MagNet model by minimizing this fused loss, i.e.,

$$\mathcal{L}_{com} = m\mathcal{L}_{sim} + \mathcal{L}_{cls} \tag{14}$$

5 EXPERIMENTS

5.1 Datasets

Experiments were conducted on two datasets for visual haze classification. One, called Hazel-level. Notably, there is few published

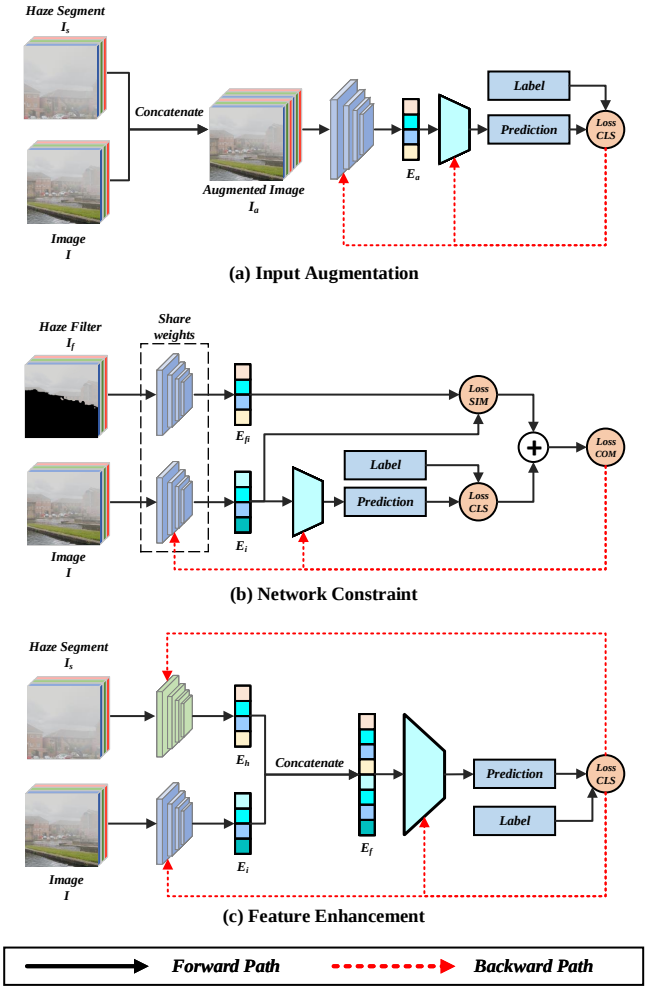


Figure 4: Schematic diagrams of three network designs for applying the haze segment and haze filter output by the UCM module. Loss CLS, Loss SIM, and Loss COM refer to Classification Loss, Similarity Loss, and Combined Loss respectively.

dataset. To better evaluate the generalization capability of UCM-MagNet, a new dataset, named Haze-wild, was created. The details about two datasets are as follows:

- **The Image Hazel-level Dataset (Hazel-level Dataset)**: The image Hazel-level dataset [14] contains 3024 synthetic images with 9 classes. These images are based on the algorithms and dataset provided by the FRIDA dataset.
- **Haze-wild Dataset**: We use source datasets that contain 5000 sunny and 5000 cloudy images to generate 10 levels of fogging images with the original images as level 0 through a monocular depth estimation model [4].

5.2 Model details

We implemented the proposed UCM-MagNet and the algorithms in comparison in-house with python 3.6.5, with the parameter settings following the original papers. The details are reported as follows:

- Threshold based methods: There are three algorithms using threshold based feature including Filter-Based Fog Detection

(a) Hazel-level Dataset

(b) Haze-wild Dataset

Figure 5: Examples of Hazel-level and Haze-wild dataset, and the level of haze increases from left to right; (a) Hazel-level Dataset includes 9 classes of images and (b) Haze-wild includes 10 classes of images.

Table 1: Performance comparison of haze classification algorithms on Hazel-level and Haze-wild datasets.

Dataset	Threshold-based Methods			Handcrafted Methods	Deep Learning Methods				
	S&RCDA [1]	FBFDA [9]	HBFDA [18]	SVM [31]	CNN [27]	LeNet5 [10]	ResNet18 [8]	MagNet(LeNet5)	MagNet(ResNet18)
Hazel-level	0.1529	0.2063	0.2431	0.4532	0.7436	0.8745	0.8998	0.8976	**0.9108**
Haze-wild	0.1018	0.2121	0.1779	0.2601	0.5181	0.7692	0.8412	0.7921	**0.8524**

[9], Saturation & RGB-correlation Detection [1] and HSV-Based Fog Detection [18]. These methods use pixel value information to calculate the haze concentration by the formulas.

- Classifiers with handcrafted features: The SVM [31] is based cascaded SVMs and uses four handcrafted features for classification.
- Deep learning methods: The CNN [27] contains nine convolutional layers, two pooling layers, and two dropout layers. It achieves the classification of air pollution levels through a ReLU-based activation function. The LeNet5 [10] contains three convolutional layers, two subsampling layers and two fullconnection layers. The ResNet18 [8] classifies haze images by residual network.

For proposed UCM and MagNet, the model details are as follows:

(1) Regarding the parameters used in the UCM, we always set the size of filter $\mathbf{F}_b$ and $\mathbf{F}_d$ 1/40 of the input image size, the size of filter $\mathbf{F}_{de}$ 1/40 of the input image size and the size of filter $\mathbf{F}_{mask}$ 1/70 of the input image size. For the setting of the threshold, we recommend using our setting in Eqs. 4.
(2) As for MagNet, parameters are set differently for different datasets. When using the Hazel-level dataset, batch size is usually set between [8,16], learning rate is usually chosen between [5e-5, 1e-4, 5e-4, 1e-3], and optimizer is chosen Adam. When using the Haze-wild dataset, batch size is usually set as 128, learning rate is usually chosen between [5e-4, 1e-3, 5e-3]. The decay is set by 0.1 or 0.5 for every N epoches. When using the Network Constraint method, the ratio m of $\mathcal{L}_{com}$ is suggested to be a small positive value, such as 0.001. We use ResNet18 [8] and LeNet5 [10] as the base models of the MagNet. The original image and the haze region map shrunk to the size (64×64). The corresponding changes in different methods are done on these two networks. All models were implemented in CUDA 11.3 environment with Pytorch 1.7.1.

5.3 Performance Comparison

This section presents a performance comparison between MagNet and existing haze classification methods, including three threshold-based features methods: Saturation & RGB-correlation Detection (S&RCD) [1], Filter-Based Fog Detection (FBFD) [9], HSV-Based Fog Detection (HBFD) [18], the handcrafted features mathod SVM [31], and the Deep Learning features method CNN [27], LeNet5 [10] and ResNet18 [8]. For both algorithms, we fine-tune their hyper-parameters to obtain their best performance in the experiments. We can observe the followings as shown in the table 1:

- Among various methods, the precision on Haze-wild dataset is lower than that of Hazel-level, which shows that the complexity of the background of haze images and the number of categories affect the precision of haze classification.
- Threshold-based method performs poorly on both datasets, which may be caused by the loss of information for threshold calculating. They are only valid for data that conform to assumptions, so the overall progress is still limited.
- The handcrafted feature extraction method achieves a 46.21% performance improvement over calculating haze values by equations in both datasets. But the quality of feature engineering limits the performance of machine learning, and the fitting process is difficult to intervene.
- Among the end-to-end feature extraction methods, the CNN method achieves 64.07% precision improvement than the handcrafted feature extraction method on the Hazel-level dataset and achieves 99.19% precision improvement on the Haze-wild dataset with a simpler background, which shows the advantages of data fitting.
- On both datasets, the proposed MagNet achieves significant precision improvement than existing methods using different backbones(20.71% with LeNet5 and 25.87% with ResNet18), demonstrating the model-agnostic of our method and verifies the effectiveness of our proposed UCM module and MagNet.

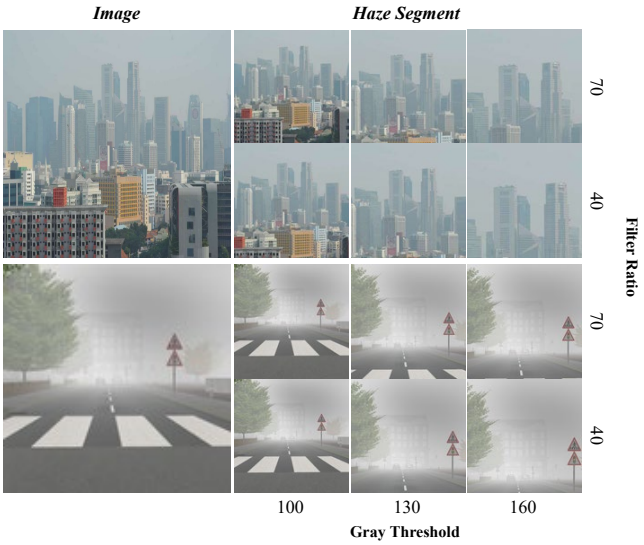


Figure 6: Haze region maps output by UCM under different parameter conditions. Gray Threshold is the threshold of calculating the haze region in gray map, and Filter Ratio is the ratio of the size of haze image to the filter.

Table 2: Time cost comparison of UCM for filtering images of different size.

Size	600*300	780*520	1080*809	1920*1441
Pixels	180K	400K	870K	2760K
Time	0.23s	0.42s	1.09s	4.12s

In summary, threshold-based and handcrafted feature extracting methods perform limited results since the loss of information and the mismatching of assumption and data. The proposed MagNet achieves better precision on both Hazel-level and Haze-wild datasets compared to other methods. Compared with the performance of the other models, MagNet achieves good performance by complementing the amplified haze features with the original image features. This result also demonstrates the role of haze region maps in haze classification.

5.4 In-depth Analysis of UCM

UCM module has two main parameters: the Gray Threshold and the Filter Ratio, and both of which affect the segmented size of the detected haze area.

As shown in the Figure 6, as the gray threshold increases, the percentage of haze in our extracted images increases. As the Filter Ratio decreases, the haze region we extract will carry less background information. Note that the threshold value is set with respect to image size, where an over-large value may lead to an overlook to haze regions and a too-small one may affect the performance of background detection. However, it is observed that a desired result can be achieved with the parameter values in mild ranges.

As shown in the table 2, the UCM algorithm takes very little time to process an image and tends to grow linearly with the increase of pixel values. Notably, the processing time can be further accelerated by a batch-manner GPU processing.

5.5 Ablation Study

In this section, we investigate the effect of the extracted haze region images on the classification results under different methods.

Table 3: Classification performance of different combinations of components on Hazel-level dataset and Haze-wild dataset. F: using haze filter map; S: using haze segment map; IA:using method Input Augmentation; NC:using method Network Constraint; FE:using method Feature Enhancement

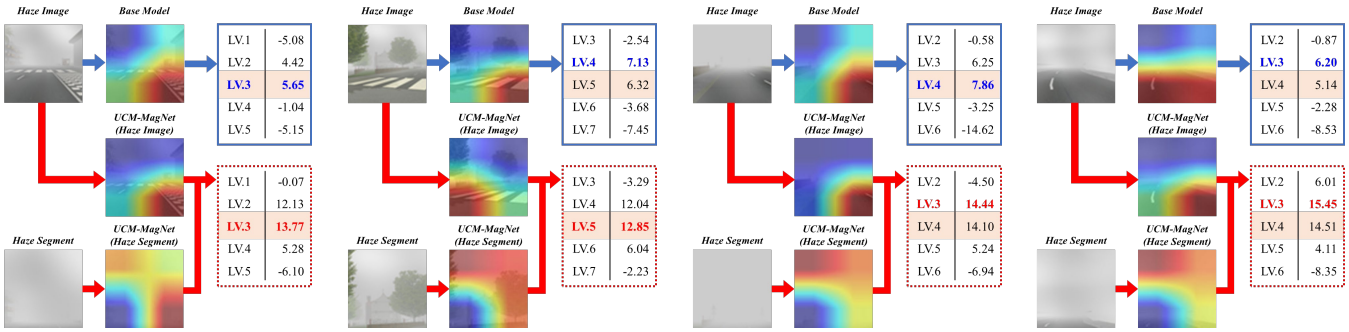
Model	Hazel-level		Haze-wild	
	LeNet5	RseNet18	LeNet5	RseNet18
Base	0.8745	0.8998	0.7692	0.8412
Base+IA+S	0.7913	**0.9405**	0.6897	0.8170
Base+IA+F	0.7997	0.8712	0.6439	0.7725
Base+NC+F	0.8624	0.8943	0.7046	**0.8510**
Base+FE+F	0.8161	0.8932	0.6552	0.8223
Base+FE+S	**0.8976**	**0.9108**	**0.7921**	**0.8524**

1. **Evaluation on Base Models:** As observed in the row "Base" of Table 3, both of the models achieve comparably good performance on the two datasets. Overall the performance of ResNet18 is better than LeNet5 on both datasets.

2. **Evaluation on Input Augmentation:** As shown in the row "Base+IA+S" of Table 3, using the segment map as augmented contents of ResNet18 leads to an improvement in performance on Hazel-level dataset and a drop on Haze-wild dataset. Using the segment map as the augmented contents of LeNet5 leads to an obvious drop on both datasets. In the row "Base+IA+F" of Table 3, using the filter map as the augmented contents of both networks leads to a drop on the two datasets. The good performance of using the segment map on Hazel-level dataset demonstrates the haze region map we extracted from UCM can improve the precision of classification. But the drop in performance on both datasets shows that Input Augmentation is not a general method for all CNN models. This method is limited to the choice of network and dataset.

3. **Evaluation on Network Constraint:** As shown in the row "Base+NC+F" of Table 3, on Hazel-level dataset, using this method leads to slightly lower performance. On Haze-wild dataset, using this method on ResNet18 has a better performance in the precision of classification than the base model. But it leads to a drop when using this method on LeNet5. The performance on Haze-wild dataset and ResNet18 demonstrates that the idea of having the features extracted by the network cover the whole haze region as much as possible is feasible.

4. **Evaluation on Feature Enhancement:** As shown in the row "Base+FE+F" of Table 3, using the filter map as the complement to the feature of LeNet5 leads to a drop in performance on both of Hazel-level dataset and Haze-wild dataset. It also leads to slightly lower performance on Hazel-level dataste when using the filter map as the complement to the feature of ResNet18, while it gets a better performance on Haze-wild dataset. In the row "Base+FE+S" of Table 3, using the segment map as the complement to the feature of both two networks leads to an improvement in performance on the two datasets. The result shows that this method using the haze segment map is a general method on the CNN models. We will provide an in-depth analysis in the following section.



(a) (b) (c) (d)

Figure 7: Visualization of feature attentions and model predictions of Base model (the blue branch) and UCM-MagNet (the red branch). We use GradCAM [23] to generate heatmaps and show outputs near the ground-truth, the ground-truth class is emphasised by background color, while the predicted class of the Base model and UCM-MagNet are marked in bolded bule and bolded red respectively.

5.6 Case Study

In this section, we will further analyze how UCM-MagNet improves haze classification in the view of feature attentions and outputs of model. We randomly selected four examples from the Hazel-level dataset and analyzed the different outputs by Base model (ResNet-18) and UCM-MagNet (ResNet-18).

When the foreground is regular and distinguishable, both methods give correct prediction as shown in Figure 7(a). The features learned by the Base model focus on foreground objects and lack attention to the haze region. For UCM-MagNet, the task channel is more focused than Base model since another channel would deal with the haze region; the magnifier channel extracts important features from the haze segment because the model has learned the whole picture, thus the prediction of ground-truth is more prominent. This case suggests that there is a cooperative mechanism in the UCM-MagNet channels to learn image features at different levels. As shown in Figure 7(b), when the foreground objects are more complex (like trees and buildings mixed), the Base model hardly pays attention to the region where the haze gathers and is disturbed by the distinguishable foreground information, so outputs a wrong prediction. Due to the dual-channel collaboration, haze image channel of UCM-MagNet pays attention to the complex foreground and magnifier channel provides the haze area as a reference at the same time, for UCM-MagNet to learn the features of the complex details and make a correct classification.

In some situations like Figure 7(c) shows, where the dense haze occupies most of the image. In this case, due to the lack of comparable objects, both of the two models need to focus on the boundary of foreground and dense haze region, and the Base model predicts correctly while the UCM-MagNet gives the wrong prediction. Since UCM-MagNet gets smaller attention in task channel, the boundary information it gets in the haze image is limited, but due to the haze information provided by magnifier channel, the output at the ground-truth class and the predicted class are actually very close. Another case of this situation is shown in Figure 7(d). At this time, both models predict incorrectly. The Base model heavily focuses on the foreground, and the output bias is more serious than UCM-MagNet; while UCM-MagNet uses dual-channel to first reduce excessive attention to the foreground, and secondly, for the magnifier channel, it strengthens the attention to the boundary of the foreground and the dense haze region, and the prediction is still relatively close to the ground-truth.

From the above analysis, we found there is a cooperative mechanism in the dual-channel structure of UCM-MagNet. The Cooperation during the training process provides model multi-level information from images. When UCM-MagNet learns the overall features of the haze image, it reduces excessive attention to the foreground, and the magnifier channel filters and concentrates important regions for classification, which enables UCM-MagNet to deal with more complex haze images.

6 CONCLUSION

This paper presents a dual-channel and model-agnostic framework (UCM-MagNet) for fast and robust haze classification. Conventional methods use the original image features, which pay too much attention to the background and cannot capture the feature of haze in the image. UCM-MagNet uses an unsupervised contrastive masking algorithm (UCM) to obtain the haze region image and performs pixel-level magnification. A dual channel structure is designed to realize feature fusion, which not only solves the problem of insufficient feature information of the magnifier channel but also reduces the interference of irrelevant background in the task channel. Experimental results show that our method can extract the haze area quickly and accurately and using haze segments makes MagNet outperform existing methods in visual haze classification.

Our proposed solution achieves impressive results. The future work includes the investigation of the fine-grained haze features for improved predictive performance and the incorporation of multiple data sources besides photos to gain awareness of both visual and invisible pollution. We want to expand the current version of UCM to detect air quality in night photos and more challenging weather, such as rain. In addition, we hope to expand the combination of different methods in MagNet to classify better.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China (Grant no. 2021YFC3300203) and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

REFERENCES

[1] Salma Alami, Abdelhak Ezzine, and Fouad Elhassouni. 2016. Local fog detection based on saturation and RGB-correlation. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. IEEE, 1–5.

[2] Avijoy Chakma, Ben Vizena, Tingting Cao, Jerry Lin, and Jing Zhang. 2017. Image-based air quality analysis using deep convolutional neural network. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3949–3952.

[3] Yan Chen, Jiangtao Wang, Shuai Li, and Weiwei Wang. 2019. Multi-feature based Foggy Image Classification. In *IOP Conference Series: Earth and Environmental Science*, Vol. 234. IOP Publishing, 012089.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. 2019. Digging into Self-Supervised Monocular Depth Prediction. (October 2019).

[5] Lu Guo, Jing Song, Xin-rui Li, He Huang, Jing-jing Du, Yong-chao He, and Cheng-zhuang Wang. 2019. Haze image classification method based on alexnet network transfer model. In *Journal of Physics: Conference Series*, Vol. 1176. IOP Publishing, 032011.

[6] Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan. 2020. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition. *IEEE Transactions on Multimedia* 23 (2020), 1785–1798.

[7] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* 33, 12 (2010), 2341–2353.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Kyeongmin Jeong, Kwangyeon Choi, Donghwan Kim, and Byung Cheol Song. 2018. Fast fog detection for de-fogging of road driving images. *IEICE TRANSACTIONS on Information and Systems* 101, 2 (2018), 473–480.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[11] Runya Li, Xiangnan Liu, and Xuqing Li. 2015. Estimation of the PM2. 5 pollution levels in Beijing based on nighttime light data from the defense meteorological satellite program-operational linescan system. *Atmosphere* 6, 5 (2015), 607–622.

[12] Xiangxian Li, Haokai Ma, Lei Meng, and Xiangxu Meng. 2021. Comparative Study of Adversarial Training Methods for Long-tailed Classification. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*. 1–7.

[13] Xiang Li, Lei Wu, Chen Xu, Lei Meng, and Xiangxu Meng. 2021. DSE-NET: Artistic Font Image Synthesis Via Disentangled Style Encoding. In *IEEE International Conference on Multimedia Expo (ICME'22), accepted, 2022*. 1–7.

[14] Yuncheng Li, Jifei Huang, and Jiebo Luo. 2015. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*. 1–5.

[15] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat Seng Chua. 2020. Multi-source Domain Adaptation for Visual Sentiment Classification. (2020).

[16] Chenbin Liu, Francis Tsow, Yi Zou, and Nongjian Tao. 2016. Particle pollution estimation based on image analysis. *PloS one* 11, 2 (2016), e0145955.

[17] Jian Ma, Kun Li, Yahong Han, and Jingyu Yang. 2018. Image-based air pollution estimation using hybrid convolutional neural network. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 471–476.

[18] Jun Mao, Uthai Phommasak, Shinya Watanabe, and Hiroyuki Shioya. 2014. Detecting foggy images and estimating the haze degree factor. *Journal of Computer Science & Systems Biology* 7, 6 (2014), 226–228.

[19] Shike Mei, Han Li, Jing Fan, Xiaojin Zhu, and Charles R Dyer. 2014. Inferring air pollution by sniffing social media. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE, 534–539.

[20] Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. 2019. Learning using privileged information for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*. 557–565.

[21] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3460–3468.

[22] Nabin Rijal, Ravi Teja Gutta, Tingting Cao, Jerry Lin, Qirong Bo, and Jing Zhang. 2018. Ensemble of deep neural networks for estimating particulate matter from images. In *2018 IEEE 3rd international conference on image, Vision and Computing (ICIVC)*. IEEE, 733–738.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[24] Haoqian Wang, Xin Yuan, Xingzheng Wang, Yongbing Zhang, and Qionghai Dai. 2014. Real-time air quality estimation based on color image processing. In *2014 IEEE Visual Communications and Image Processing Conference*. IEEE, 326–329.

[25] Xiaoyu Wang, Lei Zhang, Qirong Bo, Jun Feng, Jingzhao Hu, Yuxin Kang, and Jing Zhang. 2020. Feature Enhancement And Fusion For Image-Based Particle Matter Estimation With F-MSE Loss. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 768–772.

[26] Lei Wu, Xi Chen, Lei Meng, and Xiangxu Meng. 2020. Multitask adversarial learning for Chinese font style transfer. In *2020 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[27] Chao Zhang, Junchi Yan, Changsheng Li, Xiaoguang Rui, Liang Liu, and Rongfang Bie. 2016. On estimating air pollution from photos using convolutional neural network. In *Proceedings of the 24th ACM international conference on Multimedia*. 297–301.

[28] Chao Zhang, Junchi Yan, Changsheng Li, Hao Wu, and Rongfang Bie. 2018. End-to-end learning for image-based air quality level estimation. *Machine Vision and Applications* 29, 4 (2018), 601–615.

[29] Yuanyuan Zhang, Guangmin Sun, Qian Ren, and Dequn Zhao. 2013. Foggy images classification based on features extraction and SVM. In *Proceeding of 2013 International Conference on Software Engineering and Computer Science*. 142–14.

[30] Zheng Zhang, Huadong Ma, Huiyuan Fu, Liang Liu, and Cheng Zhang. 2016. Outdoor air quality level inference via surveillance cameras. *Mobile Information Systems* 2016 (2016).

[31] Zheng Zhang, Huadong Ma, Huiyuan Fu, and Xinpeng Wang. 2015. Outdoor air quality inference from single image. In *International Conference on Multimedia Modeling*. Springer, 13–25.

[32] Xiangwei Zhao, Jiaojiao Jiang, Kang Feng, Bo Wu, Jishan Luan, and Min Ji. 2021. The Method of Classifying Fog Level of Outdoor Video Images Based on Convolutional Neural Networks. *Journal of the Indian Society of Remote Sensing* 49, 9 (2021), 2261–2271.

[33] Xing Zhao, Ting Zhang, Wenxin Chen, and Wei Wu. 2020. Image Dehazing Based on Haze Degree Classification. In *2020 Chinese Automation Congress (CAC)*. IEEE, 4186–4191.