LFD-Net: Lightweight Feature-Interaction Dehazing Network for Real-Time Remote Sensing Tasks

Yizhu Jin, Jiaxing Chen, Feng Tian, and Kun Hu, *Member, IEEE*

***Abstract*—Currently, remote sensing equipments are evolving toward intelligence and integration, incorporating edge computing techniques to enable real-time responses. One of the key challenges in enhancing downstream decision-making capabilities is the preprocessing step of image dehazing. Existing dehazing methods usually suffer from steep computational costs with densely connected residual modules, as well as difficulties in maintaining visual quality. To tackle these problems, we designed a lightweight atmosphere scattering model based network structure to extract, fuse, and weight multiscale features. Our proposed LFD-Net demonstrates strong interpretability by exploiting the gated fusion module and attention mechanism to realize feature interactions between multilevel representations. The experimental results of LFD-Net on SOTS dataset reach an average frequency per second of 54.41, approximately eight times faster than seven most popular methods with equivalent metrics. After image dehazing by LFD-Net, the performance of object detection is significantly improved. The mean average precision when IoU = 0.5 (mAP@0.5) based on YOLOv5 is improved by 4.73% on DAIR-V2X dataset, which verified the practicability and adaptability of LFD-Net for real-time vision tasks.**

***Index Terms*—Interpretablity, model compression, real-time application, single image dehazing.**

I. Introduction

REMOTE sensing refers to the process of collecting information or data about an object, area, or phenomenon from a distance, typically using sensors mounted on aircraft, satellites, or other platforms [1]. Nowadays, the construction of space-air-ground integrated remote sensing land observation networks is of great importance for various industrial applications [2], [3]. Based on platforms, such as satellites, aircraft, drones, vehicles, and ground monitoring devices, one can obtain comprehensive information about ground objects. However, real-time accurate information extraction in complex and highly dynamic conditions, such as traffic regulation, crime tracking, and disaster relief, remains particularly difficult [4], [5], [6]. A key preprocessing step to improve the image quality is to remove the negative effects of prevailing haze, and it would be a good option to deploy dehazing algorithms on remote sensing terminal platforms, which could significantly reduce data transmission costs and achieve faster response. Therefore, it is necessary to propose a lightweight dehazing algorithm to remove the constraints of limited power and computing resources on edge devices, and optimize the dehazing efficiency while ensuring accuracy and reliability.

Dehazing methods for remote sensing images are mainly of three types: prior knowledge-based methods, physical model-based methods, and deep learning-based methods. Most of the earliest dehazing methods are based on prior knowledge. For instance, dark channel prior (DCP) makes an approximation that haze-effected pixels have at least one relatively low intensity value among RGB channels [7]; a semiphysical guided-filter-based approach is adopted to refine the coarse haze thickness map to restore textural information [8]; depth estimation and image segmentation are incorporated with DCP to generate the final transmittance [9]. These prior knowledge based methods are typically subject to empirical or statistical regularities, leading to limited application scenarios.

In addition, ASM has been extensively introduced in physics model based dehazing methods. It is physically grounded for an unrestricted access to various image scenes through the estimation of global atmosphere light and transmission map. For instance, an end-to-end DehazeNet combines dark channel, maximum contract, color attenuation as well as hue disparity prior to compute the transmission map and assigns a default value to atmosphere light [10]; a Haze Density Prediction Network is designed for a more accurate approximation of atmosphere light to better fit for nighttime occasions [11]; a multidecoder framework is presented to handle multiple bad weather restoration, with rain veiling effect embedded into the conventional ASM [12], and a differential guided layer is embedded with the backbone and substituted to the physical scattering equation [13]. Approaches based on ASM are usually more lightweight, but they may produce unnatural color tones due to inaccurate estimation of atmospheric light.

Compared with traditional dehazing methods, deep learning-based methods gradually become the research hotspot due to

Manuscript received 5 May 2023; revised 29 June 2023 and 1 August 2023; accepted 21 August 2023. Date of publication 6 September 2023; date of current version 13 October 2023. This work was supported in part by the Institute of Artificial Intelligence, in part by the Youth talent support program of Beihang University under Grant YWF-22-L-1277, in part by the National Innovation Center of Intelligent and Connected Vehicles, and in part by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China under Grant KLSMNR-202208. *(Yizhu Jin and Jiaxing Chen are co-first authors.) (Corresponding author: Kun Hu.)*

Kun Hu is with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China (e-mail: kunhu@buaa.edu.cn).

Yizhu Jin is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: 19374316@buaa.edu.cn).

Jiaxing Chen and Feng Tian are with the National Innovation Center of Intelligent and Connected Vehicles, Beijing 102607, China (e-mail: chenjiaxing@china-icv.cn; FengTianReal95@gmail.com).

Our codes are available at https://github.com/RacerK/LFD-Net.

Digital Object Identifier 10.1109/JSTARS.2023.3312515

their stronger modeling and generalization capabilities. Dehazing methods based on convolutional neural networks (CNNs) are extensively adopted, which will be discussed in detail in Section II.

Our proposed Lightweight Feature-interaction Dehazing Network (LFD-Net) utilizes convolutional layers of different kernel sizes as a sequence to extract multilevel features. The feature interaction process is addressed coherently by taking in, redistributing, and reassigning weights to the extracted features. Each component of our network performs its own function, but also interacts efficiently and effectively as a whole. Moreover, we utilize multiple metrics for evaluation, which are highly relevant and sensitive to remote sensing tasks. Overall, our main contributions are threefold as follows.

1) Our method employs ASM to jointly approximate the atmospheric light and transmission map to enhance image restoration capability and inference efficiency. It incorporates the convolutional operations into more specialized modules while maintaining the conciseness.
2) Our proposed method is designed to provide interpretability by assigning distinct tasks to each module, as demonstrated by the results of our visualization and ablation experiments. The feature-interaction process relies heavily on elementwise multiplication, which has been shown to enhance the performance of pure convolutional operations.
3) Our proposed method has been extensively validated across various scenarios of space-air-ground remote sensing land observation tasks to demonstrate its stability, practicability, and generalization capabilities. It can effectively address common challenges such as halo effect, gridding artifact, and color inconsistency, and achieves an excellent tradeoff between accuracy and efficiency, which considerably improves the performance of object detection.

II. Related Work

The increasing prevalence of intelligent remote sensing devices that support real-time responses, as opposed to relying on data transmission to servers, has highlighted the importance of studying lightweight dehazing methods, which are crucial for context-aware and fast-response remote sensing systems. However, there exists a tradeoff between the efficiency and accuracy of lightweight dehazing approaches. Some approaches employ knowledge distillation [14], [15] or pruning techniques [16], which may sacrifice accuracy for efficiency. In contrast, other methods directly construct lightweight networks to address this issue. For instance, AOD-Net [17] serves as a baseline for other lightweight dehazing models by concatenating multilevel features using different patterns. FAOD-Net [18] and GAOD-Net [19] utilized depthwise and pointwise convolutions to reduce parameters and aggregate context information in a pyramid pooling module. FAMED-Net [20] employed cascaded and densely connected pointwise convolutional and pooling layers at multiple scales. LD-Net [21] tackles the semantic gap by concatenating convolutional layers and incorporates a Color Visible Restoration module to enhance color consistency.

Nevertheless, achieving a balance between high performance on specific datasets and generalization to diverse practical applications remains a central challenge. The design and evaluation of dehazing methods should consider this tradeoff comprehensively. While current methods may exhibit promising results under certain conditions, their lack of efficiency and generalization capabilities limit their suitability for real world and real-time applications.

Our proposed LFD-Net considers dehazing as an image reconstruction task with an emphasis on feature extraction and feature utilization processes, as discussed in Sections II-A and II-B. In contrast to stacking deep residual modules in these procedures, we employ the gated fusion and attention mechanism only once, which improves both efficiency and interpretability. Moreover, it is important to design comprehensive evaluation metrics for dehazing methods, as described in Section II-C.

A. Feature Extraction

One of the key challenges in image reconstruction is the extraction of multilevel or multiscale features, which can be facilitated by using a symmetric encoder-decoder structure. The U-Net architecture, originally designed for effective extraction of context information at different scales or levels [22], has been widely used as a backbone in various reconstruction tasks. In [23], the Strengthen-Operate-Subtract boosting strategy is incorporated into the decoder, and a dense feature fusion module utilizing a back-projection feedback scheme is leveraged to compensate the missing spatial information from high-resolution features. In [24], the U-Net architecture is modified to incorporate discrete wavelet transform and inverse discrete wavelet transform in place of conventional downsampling and upsampling. In [25], hybrid convolution is applied in the U-Net encoder, which combines standard convolution with dilated convolution, to expand the receptive field and extract image features in more detail.

As opposed to a fixed backbone like U-Net, some methods utilize more flexible structures with multiple paths to diversify color information or perform various tasks. For instance, in [26], image dehazing and depth estimation are addressed simultaneously in a framework with four decoders sharing information from the same encoder. In [27], a multicolor space encoder that incorporates RGB, LAB, and HSV is applied to extract representative features in separate paths. In [28], quadruple color-cue is integrated into a multilook architecture with multiweighted training loss for autonomous vehicular application. These color spaces are often designed manually, which work well for specific applications, but may lack adaptability and generalization for others.

B. Feature Utilization

Another major challenge in image reconstruction tasks is the efficient utilization of extracted features, which has prompted the exploration of various feature fusion strategies and attention mechanisms. For instance, in [29], a novel attention-based multiscale estimation module is implemented in the backbone on a grid network to alleviate the bottleneck issue encountered in conventional multiscale approaches. In [30], a block structure integrated with channelwise attention (CA), pixelwise attention (PA) is stacked to form a group structure, which is progressively triple-stacked and concatenated to feed into another CA-PA attention mechanism for feature fusion. In [31], a multilevel fusion module is presented to integrate low-level and high-level representations. In addition, a residual mixed-convolution attention module is developed to guide the network to focus on significant features during the learning process. In [32], the feature fusion method progressively aggregates the features of hazy image and generated reference image to remove the useless features.

Moreover, the self-attention mechanism proposed in transformer has also been practiced in dehazing methods. For instance, a transformer-based channel attention module and a spatial attention module are combined to form an attention module that enhances channel and spatial features [33]. Long-range dependencies of image information can be effectively extracted through transformer blocks in image dehazing [34]. Recently, it has been revealed in [35] that self-attention mechanism inherently functions as a two-order feature interaction. In our method, gated convolution has been developed as an alternative method to achieve an competitive results to self-attention, while reducing the computational cost.

C. Quality Evaluation

Existing methods usually focus on high performance quantified by metrics in terms of peak-signal-to-noise-ratio (PSNR) and structure similarity index (SSIM). More specifically, PSNR measures the ratio between the maximum possible value of a pixel and the power of corrupting noise that affects the restoration fidelity. Instead of directly estimating absolute error, SSIM reveals interdependencies within pixels by luminance masking and contrast masking between spatially close image pairs. Besides, CIE2000 Delta E formula (CIEDE2000) and Spatial-Spectral Entropy-based Quality (SSEQ) are also introduced in our comparison metrics, because color and texture are significant for object recognition and terrain classification of remote sensing applications. CIEDE2000 is used to quantify the visual difference between two colors. It takes into account the chromaticity and luminance of the colors being compared, as well as the surrounding colors and the viewing conditions [36]. SSEQ is calculated by separating the image into its spatial and spectral components, calculating and combining the entropy of each component [37]. Halo effect in many remote sensing images can lead to significant degradation over large areas compared to high spatial resolution close-range images. In the comparison experiments, we calculate the average CIEDE2000 of each pixel in image pairs and the average absolute value of relative error on SSEQ (i.e., ΔSSEQ).

III. Proposed Method

A. Preliminaries

ASM is employed in our method to overcome the difficulty of raw pixel prediction from reconstructed images via light model. It is physics based, more suitable for real-world scenarios, and less prone to overfitting during training. The conventional ASM can be reformulated to jointly estimate the global atmosphere light A and the transmission map t, resulting in a reduction of parameters [17]

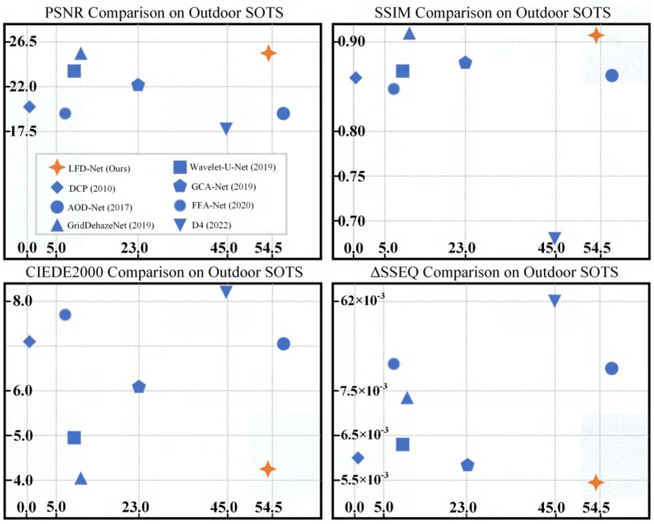


Fig. 1. Comparison metrics on outdoor SOTS, in terms of PSNR, SSIM, CIEDE2000, ΔSSEQ (↑), and FPS (→).

$$I(\theta) = J(\theta) \times t(\theta) + A(1 - t(\theta)) \tag{1}$$

where A is treated as a constant, $t \in (0, 1]$ denotes the pixelwise transmittance of light, θ represents the pixel coordinate of an $H \times W$ image of height H and width W, with I and J being the hazy input and haze-free output, respectively. Therefore, the haze-free approximation $J(\theta)$ can be written as

$$J(\theta) = \frac{I(\theta) - A}{t(\theta)} + A. \tag{2}$$

To encapsulate these two factors (i.e., A and $t(\theta)$) into one variable, the formula of the reformulated ASM is as follows:

$$J(\theta) = K(\theta) \times I(\theta) - K(\theta) + b \tag{3}$$

where $K(\theta)$ represents the new incorporated variable, which can be derived as

$$K(\theta) = \frac{\frac{1}{t(\theta)} \times (I(\theta) - A) + (A - b)}{I(\theta) - 1}. \tag{4}$$

To be specific, K is the intermediate evaluation parameter of the network. The ultimate goal is to generate a separate K value for each input channel, typically in terms of RGB. That is, K in size $3 \times H \times W$ is substituted into (3) at the end of the network, with a most commonly used default value $b = 1$.

B. Network Design

The proposed LFD-Net distinguishes itself from both heavyweight and lightweight frameworks with its concise and effective approach to feature extraction and interaction, as shown in Fig. 1 and Fig. 2. To optimize the lightweight structure design of the LFD-Net, the gated fusion module and attention mechanism are used only once, instead of being incorporated as parts of more

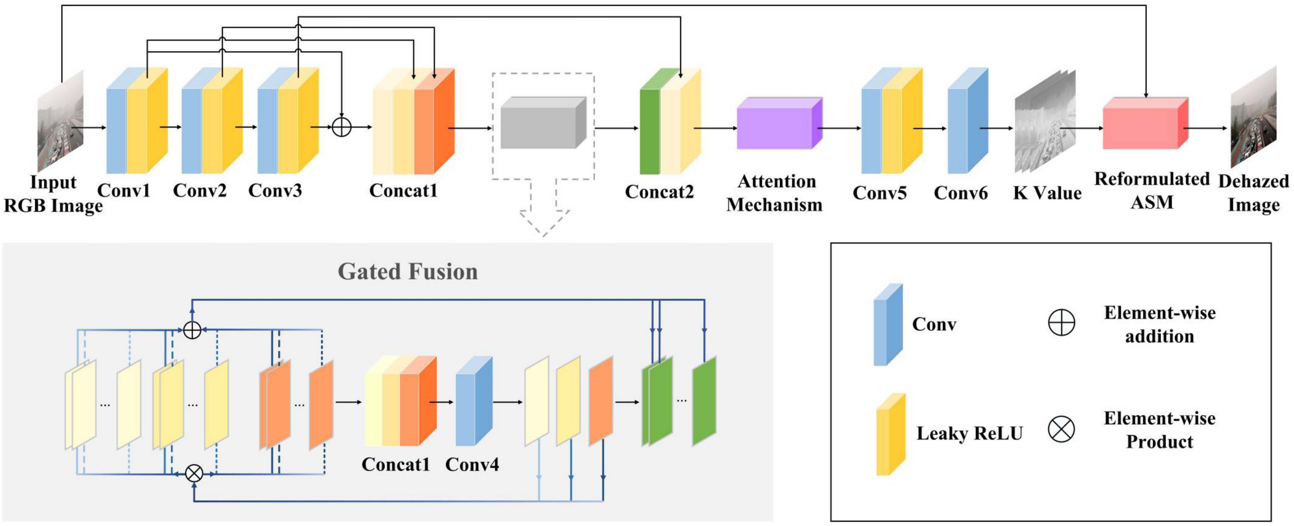


Fig. 2. Architecture of the lightweight feature-interaction dehazing network. The reformulated ASM generates an explicit output by substituting the evaluated K value. The network primarily consists of convolutional layers and concatenation layers, with the use of elementwise product in the gated fusion module and attention mechanism.

complex blocks. This approach significantly improves efficiency while maintaining strong performance in dehazing tasks.

In CNNs, convolution kernels of varying sizes are used to extract features at different levels of abstraction. Specifically, smaller size kernels are effective at capturing local features, while larger size kernels are more suited for capturing features with larger receptive fields, which are considered as more global features. The most commonly used kernel size is 3 × 3. However, stacking convolutional layers with this typical kernel size are not efficient enough in lightweight models. The concatenation layers are utilized to combine the low-level and high-level features, which compensates the loss of information from the initial layers as the network proceeds deeper. Therefore, the formation of convolutional and concatenation layers is crucial and needs to be designed flexibly to meet specific needs. Different from existing methods, we further simplify the formation of convolutional layers during feature extraction. Based on this, we also introduce feature interaction strategies including the gated fusion module and attention mechanism.

To be specific, in feature extraction, a sequence of convolutional layers with ascending kernel sizes is implemented, ranging from $3 \times 3, 5 \times 5$, to 7×7, namely *Conv 1*, *Conv 2*, and *Conv 3*. A residual connection between *Conv 1* and *Conv 3* is utilized to refine feature representations between low-level and high-level features.

A concatenation layer, namely *Concat 1*, is applied to combine the multilevel features from the extraction process. These features are then fed into the gated fusion module for spatial interactions, which includes a convolutional operation, namely *Conv 4*. The output features are passed to the second concatenation layer, namely *Concat 2*, which progressively integrates the features extracted in *Conv 3* layer. This is because higher level information is always more global, and thus being distributed to lower levels in the gated fusion module while performing feature interactions. This information is also indispensable for image restoration, especially for the following attention mechanism, which makes it necessary to involve the *Concat 2* layer. The attention mechanism adaptively learns channelwise and pixelwise weights to enhance conducive features. After that, all features are fed into the high-resolution stage, which consists of two convolutional layers, namely *Conv 5* and *Conv 6*, respectively. The details of the proposed method are illustrated in Table I.

TABLE I
DETAILS OF THE LFD-NET ARCHITECTURE

	Kernel Size	Stride	Padding	Channel (In / Out)
Conv 1	3	1	1	3 / 32
Conv 2	5	1	2	32 / 32
Conv 3	7	1	2	32 / 32
Concat 1	*Conv 1*, *Conv 2*, *Conv 1* + *Conv 3*			
Conv 4	3	1	1	96 / 3
Concat 2	*Conv 3*, the output of Gated Fusion module			
Conv 5	3	1	1	64 / 16
Conv 6	1	1	9	16 / 3

C. Gated Fusion Module

Our proposed LFD-Net replaces densely connected residual blocks with effective feature-interaction-based strategies. The gated fusion module aims to perform two-order interactions among multilevel features. This idea is demonstrated in transformer-based architecture through two successive pixelwise products (i.e., K, V) [38]. While transformers are effective, the computational cost is huge when dealing with low-level preprocessing tasks. Transformer-ensembled CNNs usually expand the flexibility of convolutional operations through adding dynamic weights to improve the modeling power of convolution [35], [39], [40]. Similar techniques have been practiced in image dehazing methods [41], [42], but are still in need of further exploration and interpretation.

Our proposed method also takes advantage of pixelwise multiplication by directly implementing it to successive feature levels, the concatenation layer *Concat 1* that combines the sequence of convolutional layers *Conv 1*, *Conv 2*, and *Conv 3*. For illustration, these features are denoted as $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$, respectively. In addition, these three convolutional operations are denoted as $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$, and the ith feature map of the output layer is denoted as $\mathcal{G}_i$. The process of gated fusion module can be expressed mathematically as follows:

$$\begin{aligned}\mathcal{G}_i &= \sum_{k=1}^{3} \mathcal{C}_k(\mathcal{F}) \otimes \mathcal{F}_{k,i} \\ &= \sum_{k=1}^{3} \mathcal{C}_k(\mathcal{F}_{k,i} \oplus \sum_{j \neq i} \mathcal{F}_{k,j}) \otimes \mathcal{F}_{k,i} \\ &= \sum_{k=1}^{3} \mathcal{C}_k(\mathcal{F}_{k,i}) \otimes \mathcal{F}_{k,i} + \sum_{j \neq i} \mathcal{C}_k(\mathcal{F}_{k,j}) \otimes \mathcal{F}_{k,j} \end{aligned} \tag{5}$$

where $\mathcal{F}_{k,i}$ is the original ith feature map of the kth group. As shown in (5), the input of gated fusion module consists of three levels. The number of output feature maps reduces the input by one-third, equal to the number of feature maps in each level of the input. The gated fusion module enhances the features within a feature map with neighboring pixels and introduces interactions by dynamically assigning weights to other feature maps through pixelwise multiplication. This reinforces the ability of convolution to retain and utilize multilevel features in an intensive and expansive manner.

D. Attention Mechanism

According to (5), the gated fusion module adaptively enhances and interacts with multilevel features. However, in cases where the haze is unevenly distributed, as often occurs in aerial imaging, accurately assessing the extent and density of the haze region remains challenging. This can result in the presence of fancy shades or dark spots. Attention mechanisms, which have been designed to focus on distinctive parts when processing large amounts of information [43], can be utilized to address this issue in image dehazing. Specifically, CA selects the feature levels for features associated with the haze region, while pixelwise attention refines the selected haze region. In [30], attention mechanism [44] is integrated into a block structure and stacked in feature extraction process. While in our proposed method, the attention mechanism is utilized only once as a single module to finalize feature weights before the high-resolution stage, leaving a large space for weight adjustment.

The adopted attention mechanism is composed of channelwise attention (CA) and pixelwise attention (PA), as depicted in Fig. 3, serving as a compensation to the gated fusion module. All of the convolution operations used in the attention mechanism have a kernel size of 1 × 1, similar to a multilayer perceptron architecture, with global average pooling and channelwise mixing [45]. In this mechanism, elementwise product is also used in place of absolute convolutional operations to increase the flexibility and reduce computational complexity.

In detail, CA first assigns weights to each channel by a global average pooling. The average pooling value of the c-th feature map, namely $\mathcal{M}_c$, can be formulated as follows:

$$\mathcal{M}_c = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} \mathcal{M}_{c,i,j}. \tag{6}$$

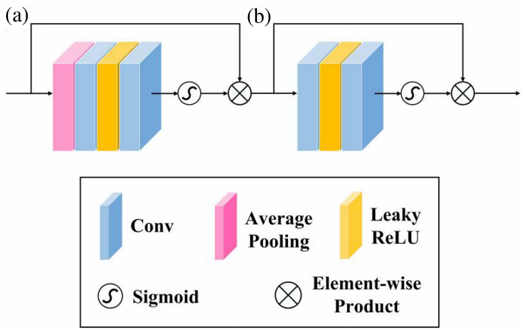


Fig. 3. Structure of attention mechanism. (a), (b) Stand for CA and PA separately.

Then, two successive convolutional layers with activation layers are utilized as linear transformation to obtain a 1-D weight vector that elementwisely multiplies the cth feature map as follows:

$$\mathcal{M}_c^* = \sigma(\mathcal{C}_2^* \delta(\mathcal{C}_1^*(\mathcal{M}_c))) \otimes \mathcal{M}_c \tag{7}$$

where $\mathcal{C}_1^*$ and $\mathcal{C}_2^*$ are the two convolutional layers, respectively, with $\delta(\cdot)$ and $\sigma(\cdot)$ being the corresponding activation function.

Similarly, PA transforms the output feature maps of CA $\mathcal{M}^*$ on a pixel scale with the output namely $\mathcal{M}^\circ$ derived as follows:

$$\mathcal{M}^\circ = \sigma(\mathcal{C}_2^\circ \sigma(\mathcal{C}_1^\circ(\mathcal{M}^*))) \otimes \mathcal{M}^* \tag{8}$$

where C_1° and C_2° are the two convolutional operations, respectively, with $\sigma(\cdot)$ being the shared activation function.

Unlike the gated fusion module, which reduces the number of channels by one-third, the attention mechanism maintains an equal number of input and output channels. This suggests that the attention mechanism is able to effectively preserve the feature representation through channelwise interaction, leading to fine-tuning of pixelwise features with relatively low computational cost. In comparison to the approach presented in [21], which utilizes 1 × 1 convolutional layers at the beginning and end of the network, our method incorporates fully connected layers into the attention mechanism with elementwise product to further enhance the power of convolutional operations.

E. Loss Function

While a combination of L1 loss, L2 loss, SSIM, or perceptual loss as loss functions has been shown to achieve good performance in previous works [46], [47], [48], [49], our experiments on LFD-Net indicate that the most widely used L2 loss, is the most suitable loss function for LFD-Net. The L2 loss is defined as follows:

$$\mathcal{L} = \frac{1}{H \times W} \sum_{s=1}^{W} \sum_{t=1}^{H} (I_{s,t} - J_{s,t})^2 \tag{9}$$

where I is the input hazy image and J is the haze-free output. The intermediate value being approximated is K, which is not a direct output and thus introducing a natural discrepancy with the output from VGG, rendering it impractical to utilize perceptual loss. Furthermore, the small number of parameters in the

TABLE II
AVERAGE COMPARISON OF METRICS ON SOTS FOR 492 JPG IMAGES

	PSNR↑	SSIM↑	CIEDE↓	ΔSSEQ ↓	FPS↑
DCP	19.81	0.8622	7.07	0.0061	0.17
AOD-Net	19.45	0.8593	7.12	0.0080	**56.85**
GridDehazeNet	25.07	**0.9108**	**4.02**	0.0074	10.04
Wavelet-U-Net	23.73	0.8661	4.93	0.0064	8.28
GCA-Net	22.13	0.8766	6.19	0.0058	23.01
FFA-Net	19.36	0.8472	7.70	0.0131	6.91
D4	18.09	0.6668	8.36	0.0624	44.79
LFD-Net	**25.12**	0.9087	4.24	**0.0054**	54.41

TABLE III
AVERAGE COMPARISON OF METRICS ON O-HAZE FOR 45 JPG IMAGES

	PSNR↑	SSIM↑	CIEDE↓	ΔSSEQ ↓	FPS↑
DCP	15.79	0.6379	16.19	0.0040	0.04
AOD-Net	15.06	0.5412	17.53	0.0146	13.93
GridDehazeNet	16.68	0.6361	13.54	0.0055	1.09
Wavelet-U-Net	15.87	0.5058	14.93	**0.0041**	8.66
GCA-Net	17.24	0.6523	13.81	0.0077	4.92
FFA-Net	14.62	0.5881	14.72	0.0067	1.52
D4	11.51	0.2564	18.58	0.1612	2.70
LFD-Net	**17.67**	**0.6532**	**11.80**	**0.0041**	**14.65**

TABLE IV
AVERAGE COMPARISON OF METRICS ON RICE1 FOR 500 PNG IMAGES

	PSNR↑	SSIM↑	CIEDE↓	ΔSSEQ ↓	FPS↑
DCP	18.30	0.8209	14.80	0.0032	0.16
AOD-Net	14.80	0.6578	16.73	0.0508	**58.62**
GridDehazeNet	19.14	0.8351	11.44	0.0021	5.08
GCA-Net	18.35	0.7237	15.25	0.0075	23.61
FFA-Net	19.92	0.8117	10.39	0.0029	7.06
MSBDN	19.77	0.8477	10.65	0.0022	21.38
D4	19.29	0.8258	12.21	0.0202	15.59
LFD-Net	**30.88**	**0.9420**	**3.32**	**0.0008**	45.51

proposed method minimizes the risk of overfitting, so regularization terms (i.e., L1 Loss) may even be counterproductive.

IV. EXPERIMENTS

A. Dataset

To validate the dehazing effect of LFD-Net for space-air-ground integrated remote sensing land observation, we first conduct training (i.e., outdoor training set **OTS**) and validation (i.e., synthetic objective testing set **SOTS**, hybrid subjective testing set **HSTS**) experiments of ground-based observation from the REalistic Single Image DEhazing dataset **RESIDE** [50]. To validate the generalization ability of LFD-Net, we also use the real hazy and haze-free outdoor images dataset **O-HAZE** [51]. We fine-tine the pretrained weights from ground-based observation data by using the aerial image dataset AID [52] for satellite (i.e., space-based) and drone (i.e., air-based) and test on the Remote sensing Image Cloud rEmoving dataset **RICE** [53]. However, we lack a dataset to test the performance of the downstream perception task under hazy conditions. To solve this problem, we synthesize hazy images on **DAIR-V2X** [54] and **VisDrone2019** [55], and evaluate the performance of object detection tasks using hazy and dehazed images for comparison.

B. Experiment Results

We faithfully reproduce seven methods for various outdoor scenarios, including DCP [7], AOD-Net [17], GridDehazeNet [29], Wavelet-U-Net [24], GCA-Net [42], FFA-Net [30], and D4 [56]. All the experiments are conducted on a PC with an R9-5900HX CPU (E5-1650) and an NVIDIA RTX-3080 GPU. Quantitative comparison results on the outdoor SOTS and O-HAZE datasets can be found in Tables II and III, respectively. The visual comparison results from the outdoor SOTS and O-HAZE datasets are shown in Figs. 4 and 5. Furthermore, we also perform experiments using real-world hazy images with no reference both from HSTS and randomly selected images from the Internet, as depicted in Figs. 6 and 7.

In the remote sensing domain, to the best of our knowledge, pretrained models for dehazing methods are not publicly available. However, we also reproduce seven SOTA methods using default outdoor weights, including AOD-Net [17], GridDehazeNet [29], GCA-Net [42], FFA-Net [30], MSBDN [23], D4 [56], and DehazeFormer [57]. As expected, the performance of AOD-Net is limited due to its small number of parameters, while the other methods show similar performance before fine-tuning. In this article, our pretrained model is open to the public for further comparison.

To demonstrate the effectiveness and efficiency of our proposed method, we present a comprehensive comparison using various metrics including PSNR, SSIM, CIEDE2000, ΔSSEQ, and FPS. The comparison results are summarized in Tables II, III, and IV. In addition, we provide a comparison of model sizes in Table V.

Observations reveal that many networks suffer from inconsistencies within color blocks or misrepresenting original information, as reflected in terms of CIEDE2000 and ΔSSEQ. For instance, lightweight methods such as AOD-Net [17] and D4 [56] produce relatively dark visual quality, resulting in a significant loss of texture information and making it difficult to distinguish objects for downstream tasks. DCP [7], a traditional dehazing method, exhibits relatively high dehazing capacity; however, it is susceptible to severe color shift as it heavily relies on prior assumptions about color distributions. While GCA-Net [42] encounters color shift occasionally in the synthetic SOTS dataset, it performs well in realistic scenarios like O-HAZE, which has thick and irregular haze. However, its halo effect and color imbalance are magnified in RICE1, which makes it less adaptive to generalized scenarios, as shown in Fig. 8. FFA-Net [30] performs well on specific datasets but distinctly lacks dehazing capability on RICE1, where there are a variety of landforms and terrains, rendering it not generalizable enough for shifted domains.

From the experiment, we can observe that incorporating attention mechanisms may prevent the image from being uniformly dehazed without region discrepancy (i.e., FFA-Net) compared to networks with absolute convolutional and concatenation layers (i.e., AOD-Net, LD-Net). However, a stack of sophisticated modules incorporating attention mechanisms may also confuse the model when selecting regions of interest, leading

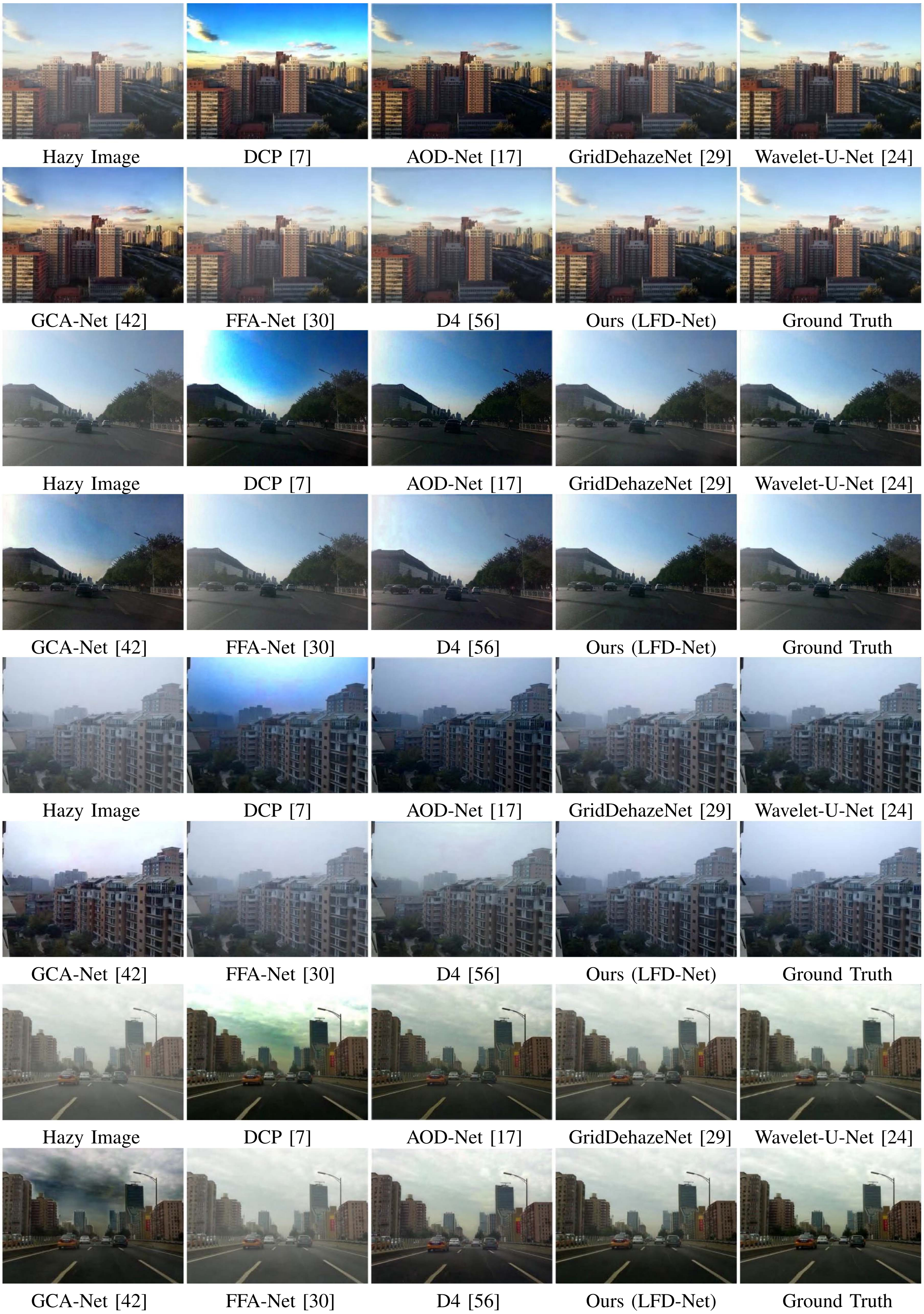


Fig. 4. Visual comparison on outdoor SOTS. We compare our methods with DCP [7], AOD-Net [17], GridDehazeNet [29], Wavelet-U-Net [24], GCA-Net [42], FFA-Net [30], and D4 [56]. Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

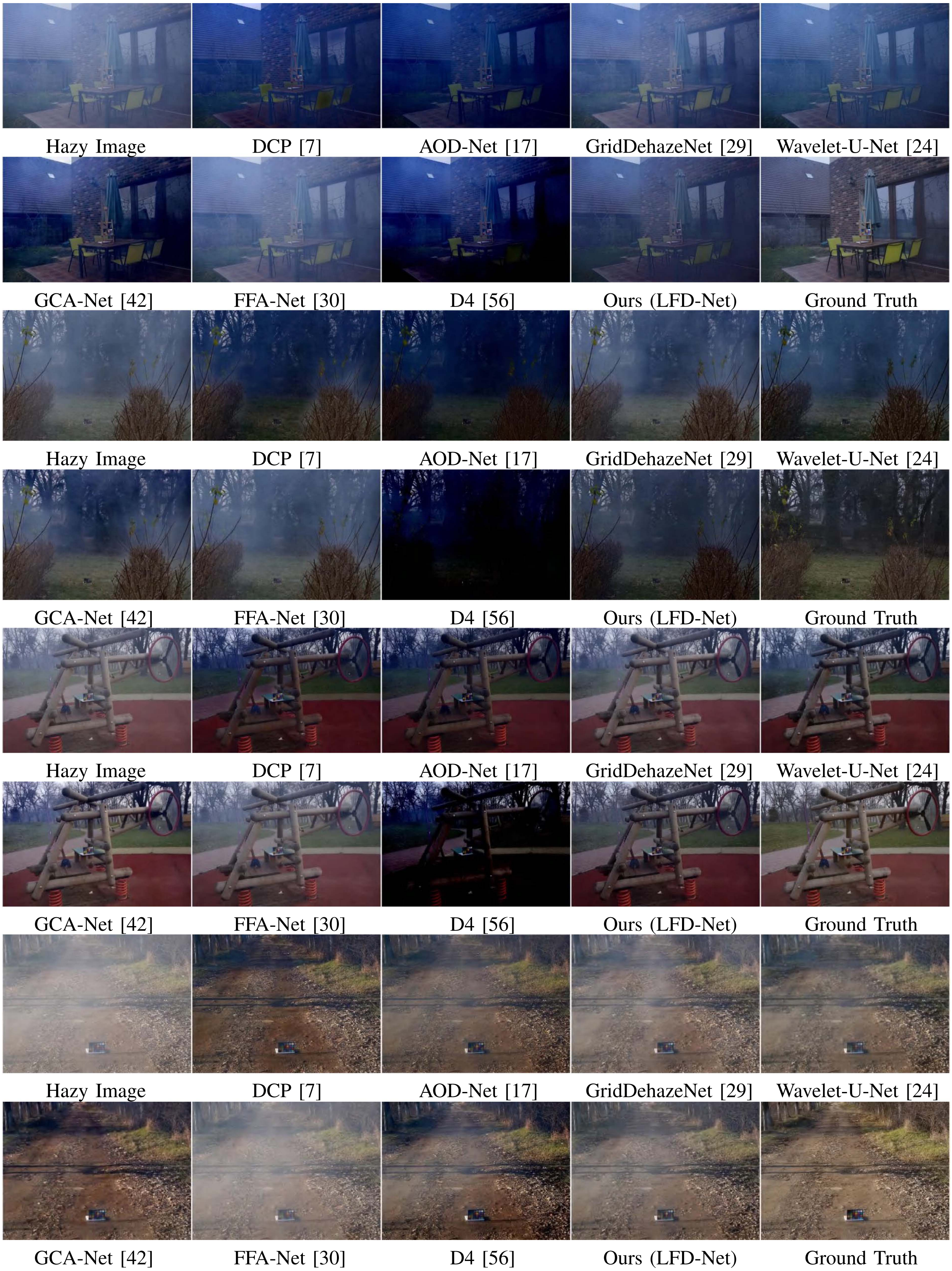


Fig. 5. Visual comparison results on O-HAZE. We compare our methods with DCP [7], AOD-Net [17], GridDehazeNet [29], Wavelet-U-Net [24], GCA-Net [42], FFA-Net [30], and D4 [56]. AOD-Net and D4 produce relatively dark in visual quality. GCA-Net performs well on irregular haze but suffers from inconsistency in color blocks. Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

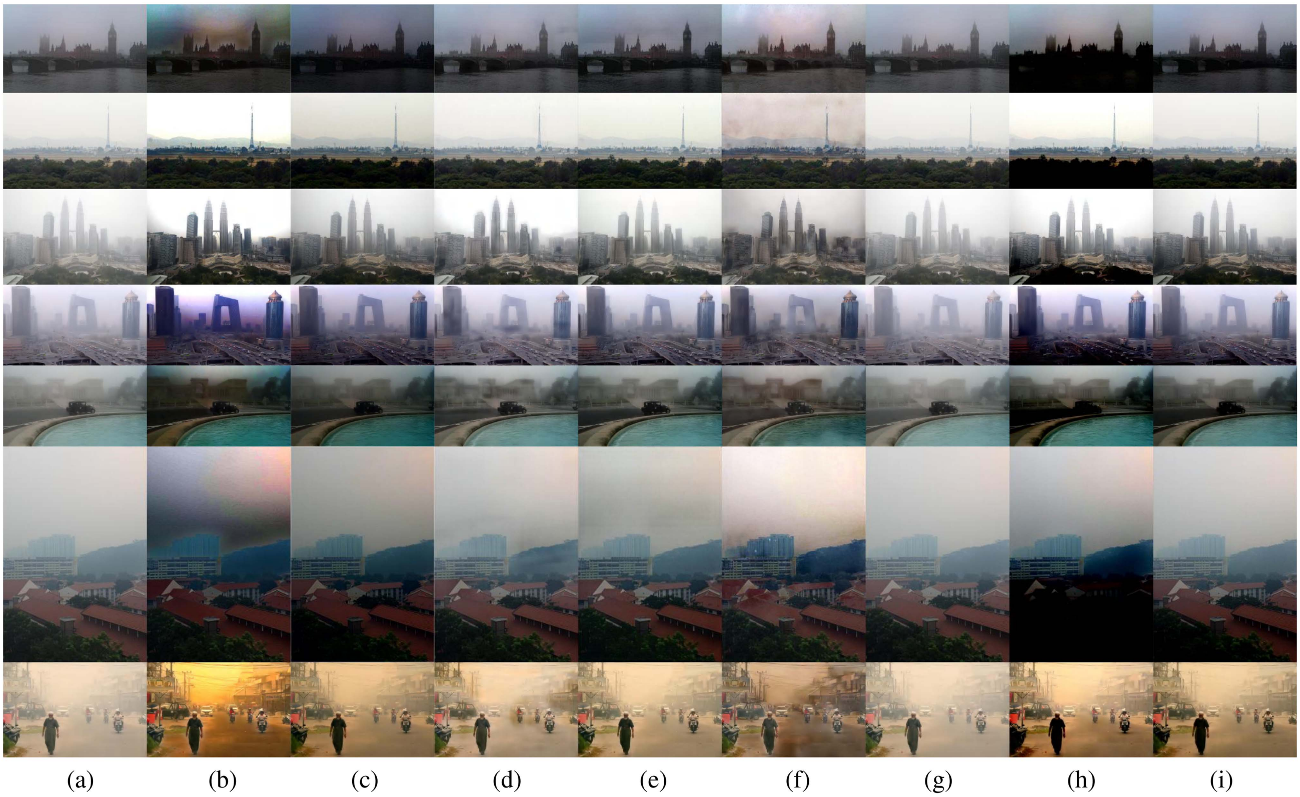


Fig. 6. Visual Comparison Results on Real-world HSTS. (a) Hazy image. (b) DCP [7]. (c) AOD-Net [17]. (d) GridDehazeNet [29]. (e) Wavelet-U-Net [24]. (f) GCA-Net [42]. (g) FFA-Net [30]. (h) D4 [56]. (i) Ours (LFD-Net). Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

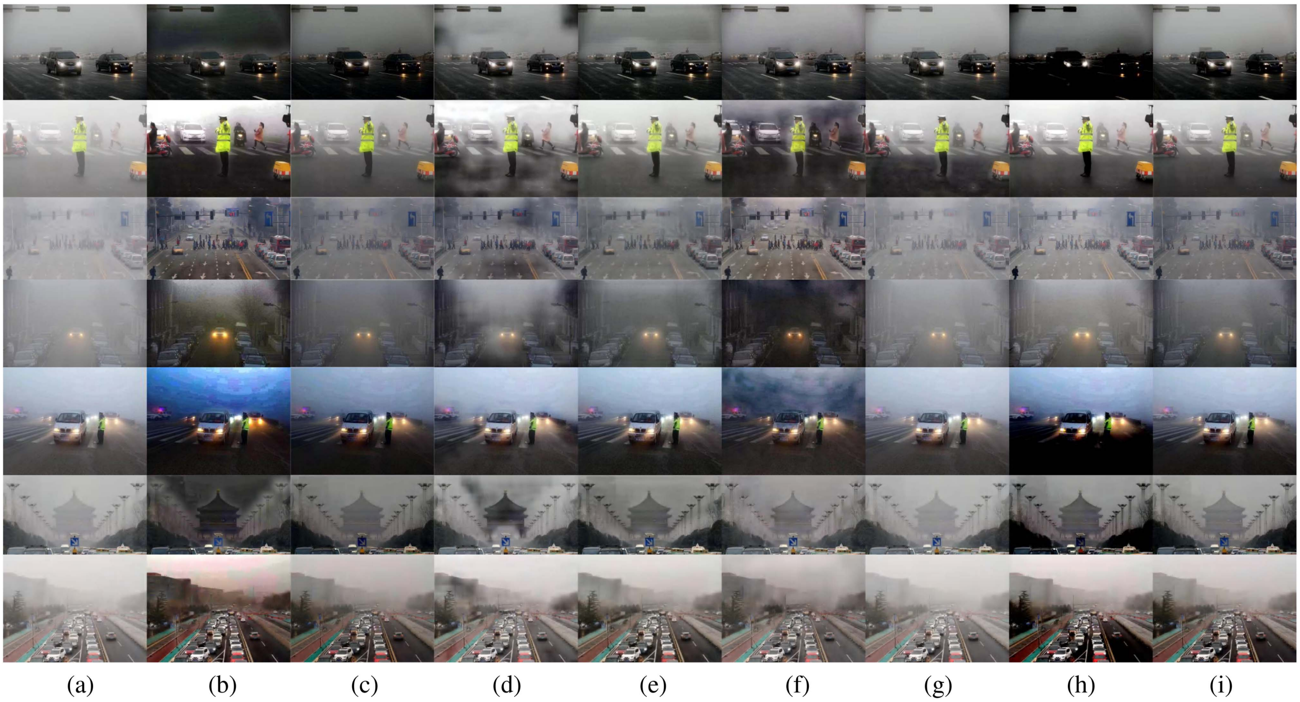
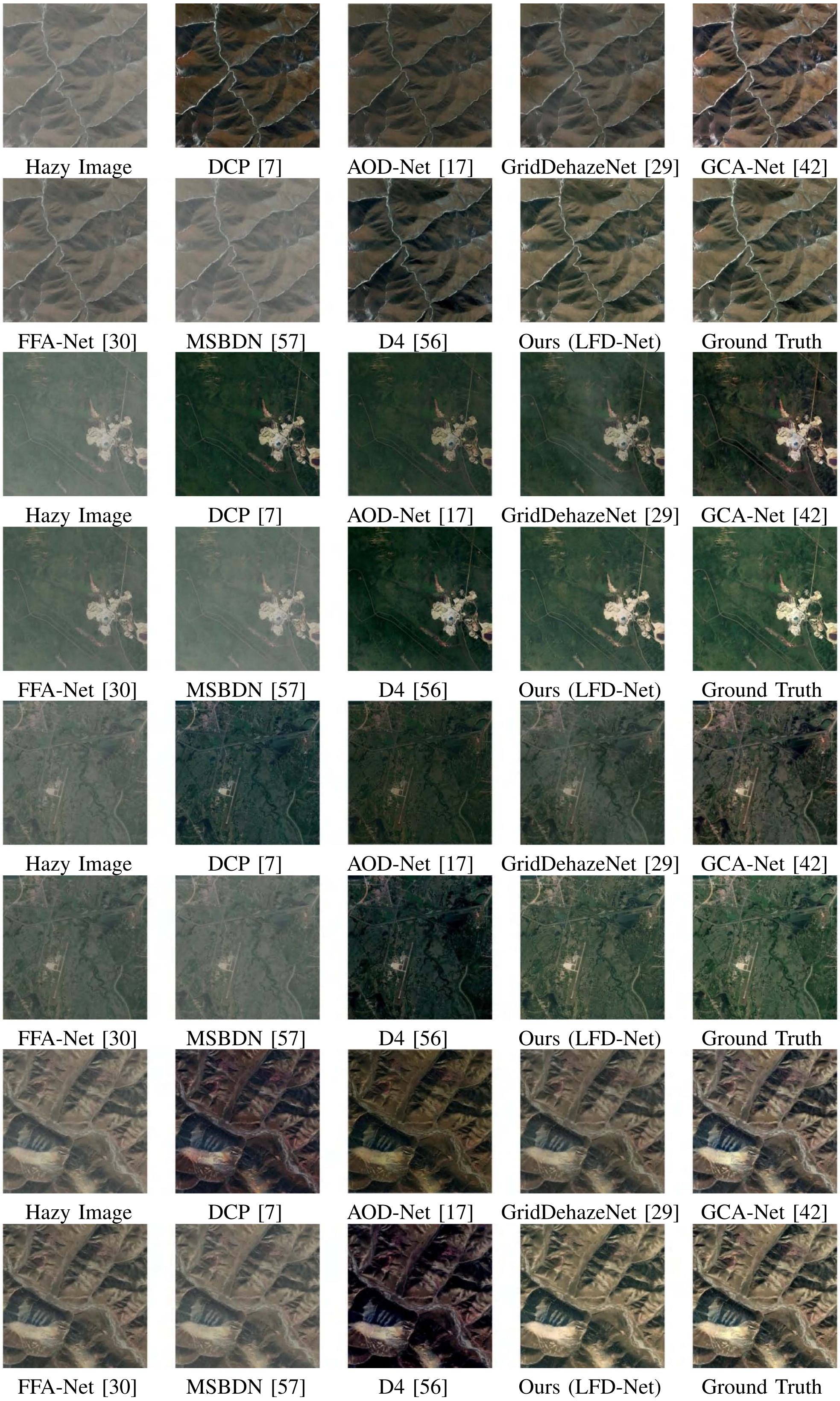


Fig. 7. Visual comparison results on randomly selected real-world images. (a) Hazy image. (b) DCP [7]. (c) AOD-Net [17]. (d) GridDehazeNet [29]. (e) Wavelet-U-Net [24]. (f) GCA-Net [42]. (g) FFA-Net [30]. (h) D4 [56]. (i) Ours (LFD-Net). Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.



Hazy Image DCP [7] AOD-Net [17] GridDehazeNet [29] GCA-Net [42]

FFA-Net [30] MSBDN [57] D4 [56] Ours (LFD-Net) Ground Truth

Hazy Image DCP [7] AOD-Net [17] GridDehazeNet [29] GCA-Net [42]

FFA-Net [30] MSBDN [57] D4 [56] Ours (LFD-Net) Ground Truth

Hazy Image DCP [7] AOD-Net [17] GridDehazeNet [29] GCA-Net [42]

FFA-Net [30] MSBDN [57] D4 [56] Ours (LFD-Net) Ground Truth

Hazy Image DCP [7] AOD-Net [17] GridDehazeNet [29] GCA-Net [42]

FFA-Net [30] MSBDN [57] D4 [56] Ours (LFD-Net) Ground Truth

Fig. 8. Visual comparison results on O-HAZE. We compare our methods with DCP [7], AOD-Net [17], GridDehazeNet [29], Wavelet-U-Net [24], GCA-Net [42], FFA-Net [30], and D4 [56]. AOD-Net and D4 produce relatively dark in visual quality. GCA-Net performs well on irregular haze but suffers from inconsistency in color blocks. Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

TABLE V
COMPARISON OF THE PARAMETERS OF MODELS

	AOD-Net	GridDehazeNet	Wavelet-U-Net	GCA-Net	FFA-Net	D4	MSBDN	LFD-Net
Parameters(×M)	0.002	0.956	11.3	0.702	4.46	10.7	28.7	0.086

TABLE VI
ABLATION EXPERIMENT OF LFD-NET ON OUTDOOR SOTS DATASET

	Concat 2	Gated Fusion Module	Attention Mechanism	PSNR↑	SSIM↑	CIEDE↓	ΔSSEQ ↓
Case 1	✗			24.35	0.9062	5.09	0.0055
Case 2		✗		23.33	0.8910	5.56	0.0066
Case 3			✗	21.62	0.8642	6.65	0.0116
Case 4		✗	✗	23.17	0.8878	5.54	0.0076
Default				**25.12**	**0.9087**	**4.24**	**0.0054**

to insufficient attention paid to each hazy region and overfitting on specific datasets with limited data diversity, rendering these approaches not flexible enough for real-world vision tasks.

Nevertheless, attention mechanisms are well adapted to U-Net or U-Net ensembling structures, where multiscale features are addressed symmetrically. Wavelet-U-Net [24] and GridDehazeNet [29] have excellent performance, but they may come at the cost of inference time, 6.6× and 5.4× longer compared to our proposed method, respectively. Wavelet-U-Net transforms the image into the wavelet space using discrete wavelet transformation, which adds to the computational cost to some extent. GridDehazeNet also utilizes attention mechanisms but as a bridge of multiscale features, which ensembles the design of U-Net [22]. It has three rows and six columns, with each row corresponding to a different scale, constructing a grid network, which may compromise the inference speed.

However, their performance on the HSTS dataset from Fig. 6 and randomly selected hazy images from Fig. 7 demonstrates that they may also suffer occasional degradation when dealing with remote objects that are occluded, as well as objects located in areas uniformly covered with thick haze but with limited prior semantic information. While the images randomly selected for our study in Fig. 7 may not be representative of specific datasets, they are still valuable for consideration as they reflect scenarios that can occur in real-world practices. Although accurately verifying the generalization of algorithms is challenging, our approach has demonstrated effectiveness even when encountering severe domain shifts, as evidenced by our experiments. Our proposed method does not adopt the U-Net structure for efficiency, nor does it leverage stacked attention mechanisms, which saves the computational cost to a large extent, exhibits adaptability to diverse scenarios, and possesses a noteworthy level of generalization.

C. Ablation Study

The experimental results confirm that our proposed LFD-Net is effective and efficient for real-time applications. Since it has a different principle than other methods, we perform a series of ablation studies to ensure that each component of the network is indispensable. The detailed experimental conditions and corresponding metrics tested on outdoor SOTS are listed in Table VI.

Inspired by [17] and [21], we add a second concatenation layer (i.e., *Concat 2*), to our method. In Case 1, we omit *Concat 2* and observe a slight loss of detailed texture information due to the reduced high-level information.

In Cases 2, 3, and 4, we investigate the importance of the gated fusion module and attention mechanism in our model. These cases demonstrate that these two subnetworks work together to facilitate feature interaction. Specifically, the removal of the attention mechanism leads to the occasional appearance of black spots on the images, which significantly degrades the overall performance. In comparison with other lightweight methods, our method partially addresses this issue. In addition, the gated fusion module is a crucial component in enhancing the dehazing capability, serving as a bridge between the multilevel feature extraction process ending at the first concatenation layer *Concat 1*, and the attention mechanism begining at the second concatenation layer *Concat 2*.

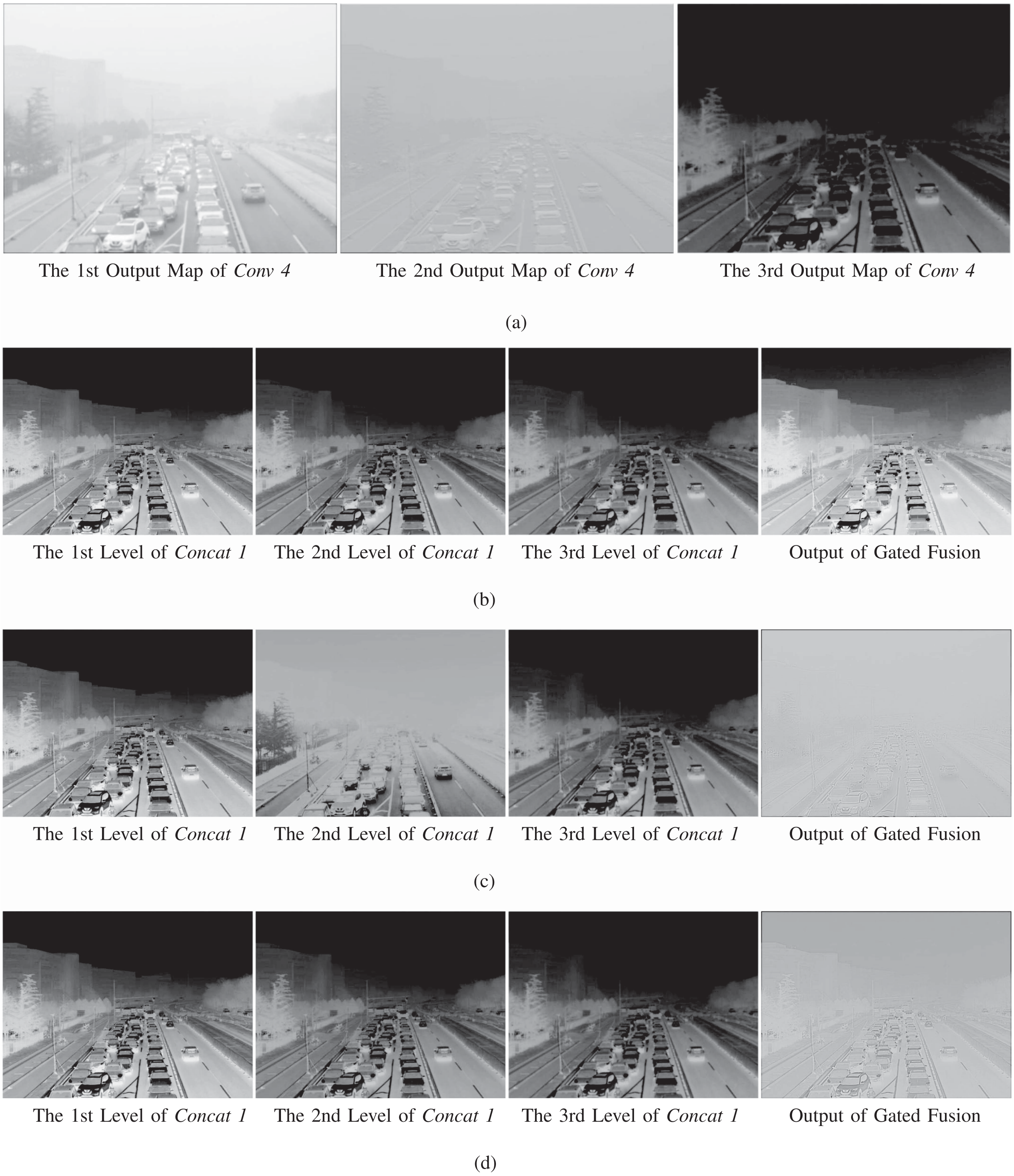
When both the attention mechanism and the gated fusion module are involved, the detailed information in the images is further refined, making it more authentic and faithful to the original information. This structure helps to preserve and interact with multilevel information to improve the overall image quality.

D. Visualization Results

We have visualized the intermediate feature maps before and after the Gated Fusion module, as depicted in Fig. 9. As shown in (a), the incorporated convolutional layer combines features of three levels from *Conv 1*, *Conv 2* and *Conv 1* + *Conv 3* to generate three distinctive feature maps. They are distinguished from each other by their focus on close or distant objects and the lightness or contrast of the pixels.

In Fig. 9(b)–(d), we demonstrate the changes in specific feature maps after the gated fusion module. Fig. 9(b) shows that the contrast of the image is enhanced with the hierarchical information, resulting in distant objects becoming more distinct. Fig. 9(c) and (d) shows more abstract feature representations, which are significantly shifted compared to the input features. Specifically, Fig. 9(c) emphasizes the outline of substances, while Fig. 9(d) highlights the blocks within substances.

The gated fusion module reallocates the distributed feature representations of the multilevel layers through feature-interaction strategies. The feature extraction process is compressed into three successive convolutional layers, for which



The 1st Output Map of *Conv 4* The 2nd Output Map of *Conv 4* The 3rd Output Map of *Conv 4*

(a)

The 1st Level of *Concat 1* The 2nd Level of *Concat 1* The 3rd Level of *Concat 1* Output of Gated Fusion

(b)

The 1st Level of *Concat 1* The 2nd Level of *Concat 1* The 3rd Level of *Concat 1* Output of Gated Fusion

(c)

The 1st Level of *Concat 1* The 2nd Level of *Concat 1* The 3rd Level of *Concat 1* Output of Gated Fusion

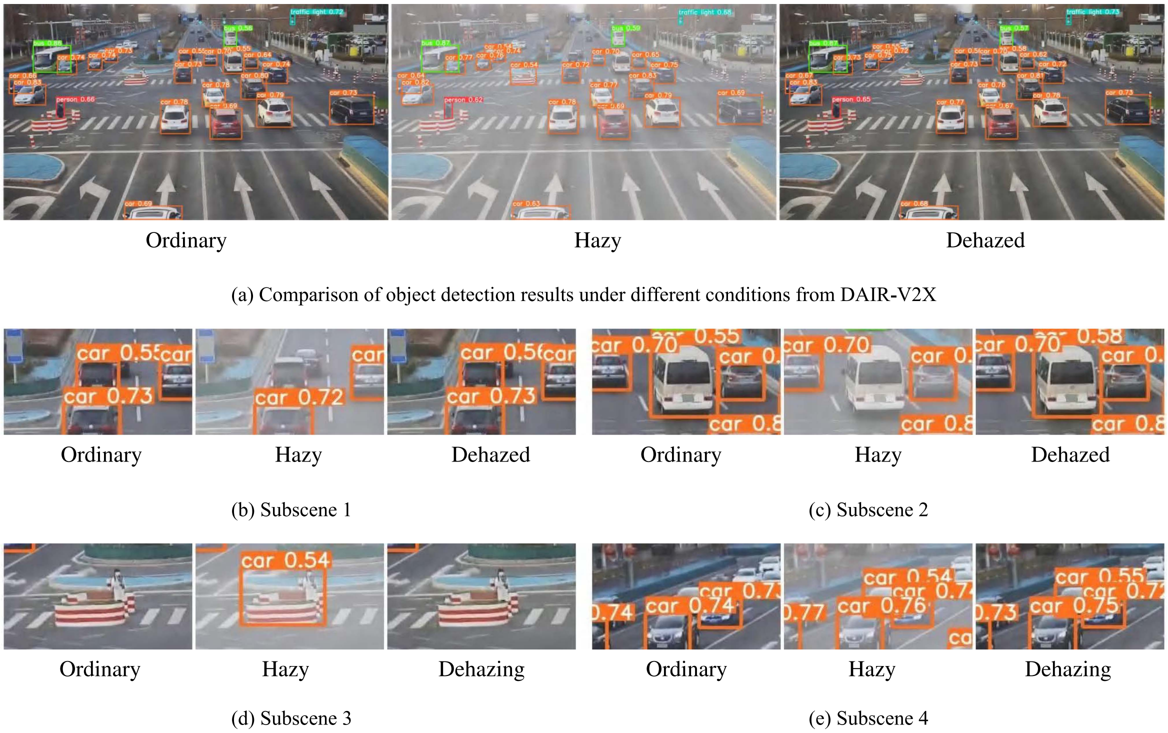
(d)

Fig. 9. Visualization results of the changes in layers before and after the gated fusion module. (a) represents the three output feature maps of the convolutional operation *Conv 4* incorporated into Gated Fusion module, (b)-(d) stand for the changes in the 15th, 30th, and 31st feature map of the layers, respectively. (b) shows that the contrast of the image is enhanced, resulting in distant objects becoming more distinct. (c) and (d) show more abstract feature representations, which are significantly shifted compared to the input. Specifically, (c) emphasizes the outline of substances, while (d) highlights the blocks within substances. (a) The output of convolutional operation in gated fusion module. (b) The 15th Feature Map of Each Level of *Concat 1* and the Output of Gated Fusion Module. (c) The 30th Feature Map of Each Level of *Concat 1* and the Output of Gated Fusion Module. (d) The 31st Feature Map of Each Level of *Concat 1* and the Output of Gated Fusion Module.

we compensate by intralevel enhancement and interlevel combination.

E. Application for Object Detection Task

As a severe weather condition, haze can significantly reduce the effectiveness of remote sensing land observation system. For instance, in autonomous navigation applications, object detection can be significantly impacted by hazy environments, resulting in degraded image quality and potentially jeopardizing the safety of the system. Therefore, preprocessing procedure for image enhancement before performing those tasks is of great importance. As far as we know, there is no dataset with built-in synthetic hazed images for object detection. In our experiment,



Ordinary Hazy Dehazed

(a) Comparison of object detection results under different conditions from DAIR-V2X

Ordinary Hazy Dehazed Ordinary Hazy Dehazed

(b) Subscene 1

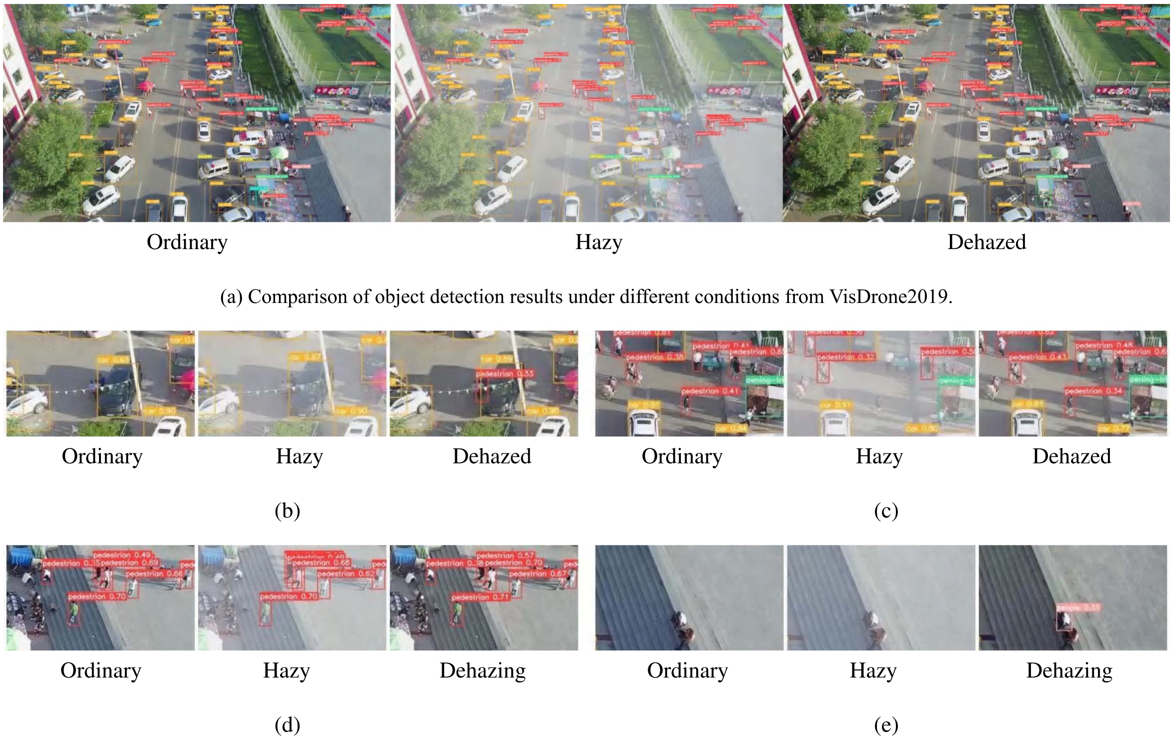
(c) Subscene 2

Ordinary Hazy Dehazing Ordinary Hazy Dehazing

(d) Subscene 3

(e) Subscene 4

Fig. 10. Reference object detection results. (a) Comparison of object detection results under ordinary, simulated hazy, and dehazed conditions. (b)–(e) Detailed subscenes of detection results under different conditions. (b) and (c) Demonstrate an improvement in the detection rate, detecting an additional car instance in the dehazed condition compared to the hazy condition. (d) Corrects the error of mistaking a roadblock for a car in the hazy condition. (e) Shows the detection of another car compared to the ground-truth clear image.



Ordinary Hazy Dehazed

(a) Comparison of object detection results under different conditions from VisDrone2019.

Ordinary Hazy Dehazed Ordinary Hazy Dehazed

(b)

(c)

Ordinary Hazy Dehazing Ordinary Hazy Dehazing

(d)

(e)

Fig. 11. Reference remote sensing object detection results. (a) Comparison of remote sensing object detection results under ordinary, simulated hazy, and dehazed conditions. (b)–(e) Detailed subscenes of detection results under different conditions, in which the detection rate for pedestrians is enhanced to a large extent. In particular, (b) and (e) highlight instances of pedestrians that are not visible in the ordinary conditions but are detected after dehazing, similar to the results from DAIR-V2X.

we randomly select 100 images from each dataset (DAIR-V2X and VisDrone2019) and produce their synthetic hazy versions. We use the default outdoor pretrained weight for the former and the fine-tuned remote sensing pretrained weight for the latter. Both object detection processes are based on YOLOv5. Our experiment results show that the mean average precision when IoU = 0.5 (mAP@0.5) of the dehazed condition improves by 4.73% compared to the hazy condition in DAIR-V2X, while by 0.81% in VisDrone2019.

Furthermore, overall detection result of a particular scene is shown in Fig. 10(a), while Fig. 10(b)–(e) illustrates the most representative perspectives of the dehazing effect. In Fig. 10(b) and (c), it can be seen that dehazing improves the detection rate, as an additional car instance is detected in the dehazed condition compared to the hazy condition. In Fig. 10(d), the roadblock is mistakenly identified as a car in the hazy condition, but the dehazing method is able to remove this error. In Fig. 10(e), another car instance is shown before and after dehazing the synthetic hazed image.

In Fig. 11(a), we show the overall remote sensing object detection results from the perspective of a drone in a particular scene. In the images captured by the drone, the types of land cover are more complex and the objects are smaller when compared to the driving perspective from DAIR-V2X. Fig. 11(b)–(e) illustrates the difficulties object detection methods encounter when detecting smaller pedestrian instances, especially in hazy conditions. However, dehazing methods can partially address this issue and enhance the detection rate of small objects like pedestrians, as shown in Fig. 11(c) and (d). In Fig. 10(b) and (e), two additional pedestrian instances are detected after dehazing compared to the original conditions, similar to that in Fig. 10(e). Experimental results show that haze can have unpredictable effects on normal conditions, and our method can provide a better solution compared to the ground truth in representing high-level semantic information to some extent.

V. Conclusion

In this article, we propose a novel end-to-end model called LFD-Net for remote sensing image dehazing. As a preprocessing for downstream vision tasks, it not only ensures the effectiveness and efficiency required for real-time applications, but also outperforms SOTA methods in terms of region-balance and color-fidelity. By designing this framework, we demonstrate the potential of CNN-based networks by performing two-order spatial interaction. Specifically, we show that the capabilities of deep neural networks can be enhanced not only by adding more complex modules to be deeper, but also by effectively combining individual and natural feature extraction, fusion, and attention with feature interaction strategies, particularly in the field of image superresolution. The experiments on various scenarios also shows that performance of a model is not always proportional to the number of parameters, and less parameters to some extent may help mitigate overfitting, which might be conducive for future network design.

References

[1] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*. Hoboken, NJ, USA: Wiley, 2014.

[2] A. S. Li, V. Chirayath, M. Segal-Rozenhaimer, J. L. Torres-Perez, and J. v. d. Bergh, "NASA NeMO-Net's convolutional neural network: Mapping marine habitats with spectrally heterogeneous remote sensing imagery," *IEEE J. Sel. Topics Appl. earth Observ. Remote Sens.*, vol. 13, pp. 5115–5133, 2020.

[3] P. Tam, S. Math, C. Nam, and S. Kim, "Adaptive resource optimized edge federated learning in real-time image sensing classifications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10929–10940, 2021.

[4] Y. Zheng, J. Su, S. Zhang, M. Tao, and L. Wang, "Dehaze-AGGAN: Unpaired remote sensing image dehazing using enhanced attention-guide generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[5] Y. Han, M. Yin, P. Duan, and P. Ghamisi, "Edge-preserving filtering-based dehazing for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[6] A. Makarau, R. Richter, R. Müller, and P. Reinartz, "Haze detection and removal in remotely sensed multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5895–5905, Sep. 2014.

[7] 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, Dec. 2011.

[8] F. Liu, Y. Lv, B. Li, S. Gao, and Y. Qin, "A semiphysical approach of haze removal for landsat image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7410–7421, 2021.

[9] B. Xie, J. Yang, J. Shen, and Z. Lv, "Image defogging method combining light field depth estimation and dark channel," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng.*, 2021, pp. 745–749.

[10] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[11] Y. Liao, Z. Su, X. Liang, and B. Qiu, "HDP-Net: Haze density prediction network for nighttime dehazing," in *Proc. Pacific Rim Conf. Multimedia*. 2018, pp. 469–480.

[12] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3175–3185.

[13] L. Jiao, C. Hu, L. Huo, and P. Tang, "Guided-Pix2Pix: End-to-end inference and refinement network for image dehazing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3052–3069, 2021.

[14] M. Hong, Y. Xie, C. Li, and Y. Qu, "Distilling image dehazing with heterogeneous task imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3462–3471.

[15] J. S. Mitheran, A. Suresh, J. Nisha, and V. P. Gopi, "Rich feature distillation with feature affinity module for efficient image dehazing," *Optik*, vol. 267, 2022, Art. no. 169656.

[16] L. Liu, Z. Meng, K. Wang, J. Zhang, and Z. Wang, "Aerial image dehazing network compression: Towards micro-UAV'S real-time online inferencing," in *Proc. Int. Conf. Image, Vis. Intell. Syst.*, 2022, pp. 442–453.

[17] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4770–4778.

[18] W. Qian, C. Zhou, and D. Zhang, "FAOD-Net: A fast AOD-Net for dehazing single image," *Math. Problems Eng.*, vol. 2020, 2020, Art. no. 4945214.

[19] S. Chen, J. Cheng, and Z. Huang, "GADO-Net: An improved AOD-Net single image dehazing algorithm," in *Proc. 3rd Int. Academic Exchange Conf. Sci. Technol. Innov.*, 2021, pp. 640–646.

[20] J. Zhang and D. Tao, "FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2020.

[21] H. Ullah et al., "Light-DehazeNet: A novel lightweight CNN architecture for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 8968–8982, 2021.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[23] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2157–2167.

[24] H.-H. Yang and Y. Fu, "Wavelet U-Net and the chromatic adaptation transform for single image dehazing," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2736–2740.

[25] T. Feng, C. Wang, X. Chen, H. Fan, K. Zeng, and Z. Li, “URNet: A U-Net based residual network for image dehazing,” *Appl. Soft Comput.*, vol. 102, 2021, Art. no. 106884.

[26] B.-U. Lee, K. Lee, J. Oh, and I. S. Kweon, “CNN-based simultaneous dehazing and depth estimation,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 9722–9728.

[27] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, “Underwater image enhancement via medium transmission-guided multi-color space embedding,” *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, 2021.

[28] A. Mehra, M. Mandal, P. Narang, and V. Chamola, “ReViewNet: A fast and resource optimized network for enabling safe autonomous driving in hazy weather conditions,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4256–4266, Jul. 2021.

[29] X. Liu, Y. Ma, Z. Shi, and J. Chen, “GridDehazeNet: Attention-based multi-scale network for image dehazing,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7314–7323.

[30] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, “FFA-Net: Feature fusion attention network for single image dehazing,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 11908–11915.

[31] X. Zhang, T. Wang, W. Luo, and P. Huang, “Multi-level fusion and attention-guided CNN for image dehazing,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4162–4173, Nov. 2021.

[32] H. Bai, J. Pan, X. Xiang, and J. Tang, “Self-guided image dehazing using progressive feature fusion,” *IEEE Trans. Image Process.*, vol. 31, pp. 1217–1229, 2022.

[33] G. Gao, J. Cao, C. Bao, Q. Hao, A. Ma, and G. Li, “A novel transformer-based attention network for image dehazing,” *Sensors*, vol. 22, no. 9, 2022, Art. no. 3428.

[34] Y. Yang, H. Zhang, X. Wu, and X. Liang, “MSTFDN: Multi-scale transformer fusion dehazing network,” *Appl. Intell.*, vol. 53, pp. 5951–5962, 2023.

[35] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S. N. Lim, and J. Lu, “Hornet: Efficient high-order spatial interactions with recursive gated convolutions,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 10353–10366.

[36] M. R. Luo, G. Cui, and B. Rigg, “The development of the CIE 2000 colour-difference formula: CIEDE2000,” *Color Res. Appl.*, vol. 26, no. 5, pp. 340–350, 2001.

[37] L. Liu, B. Liu, H. Huang, and A. C. Bovik, “No-reference image quality assessment based on spatial and spectral entropies,” *Signal Process.: Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.

[38] A. Vaswani et al., “Attention is all you need,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[40] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, “Dynamic convolution: Attention over convolution kernels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11030–11039.

[41] W. Ren et al., “Gated fusion network for single image dehazing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3253–3261.

[42] D. Chen et al., “Gated context aggregation network for image dehazing and deraining,” in *Proc. IEEE winter Conf. Appl. Comput. Vis.*, 2019, pp. 1375–1383.

[43] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[45] I. O. Tolstikhin et al., “MLP-Mixer: An All-MLP Architecture for Vision,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 24261–24272.

[46] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[47] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[48] M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, “SROBB: Targeted perceptual loss for single image super-resolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2710–2719.

[49] Y. Guo, J. Chen, X. Ren, A. Wang, and W. Wang, “Joint raindrop and haze removal from a single image,” *IEEE Trans. Image Process.*, vol. 29, pp. 9508–9519, 2020.

[50] B. Li et al., “Benchmarking single-image dehazing and beyond,” *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 754–762.

[52] G.-S. Xia et al., “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[53] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, “A remote sensing image dataset for cloud removal,” 2019, *arXiv:1901.00600*.

[54] H. Yu et al., “DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21361–21370.

[55] D. Du et al., “VisDrone-DET2019: The vision meets drone object detection in image challenge results,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.

[56] Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, and D. Tao, “Self-augmented unpaired image dehazing via density and depth decomposition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2037–2046.

[57] Y. Song, Z. He, H. Qian, and X. Du, “Vision transformers for single image dehazing,” *IEEE Trans. Image Process.*, vol. 32, pp. 1927–1941, 2023.

Yizhu Jin is currently working toward bachelor’s degree in engineering with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China.

Her research interests include the intersection between valuable application and general methods with high interpretability in computer vision and artificial intelligence.

Jiaxing Chen received the M.S. degree in electrical and computer engineering from the University of Illinois, Chicago, IL, USA, in 2021.

He worked as an Algorithm Engineer with National Innovation Center of Intelligent and Connected Vehicles in 2022. He is currently a Researcher with Tsinghua University, Beijing, China. His research interests include computer vision with deep learning, machine learning, multimodal fusion, and trajectory prediction.

Feng Tian received the B.S. degree in engineering from Zhejiang College, Tongji University, Zhejiang, China, in 2018. She is currently working toward the postgraduate degree in transportation engineering with the College of Traffic and Logistics Engineering, Xinjiang Agricultural University, Xinjiang, China.

Her research interests include machine learning, data mining, intelligent transportation, and target detection.

Kun Hu (Member, IEEE) received the B.Sc. degree in remote sensing science and technology and the M.Sc. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2010, 2012, and 2016, respectively.

He was an Assistant Professor at the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, from 2012 to 2017, as a Postdoctoral at Ohio State University, Ohio, USA, from 2017 to 2019, and an Associate Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences, from 2019 to 2022. He is currently an Associate Professor with the Institute of Artificial Intelligence, Beihang University, Beijing. His research interests focus on the accurate processing and intelligent application of multisource remote sensing data, such as camera calibration, image production, information fusion, target detection, clarification, 3-D reconstruction, and quality evaluation.