

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 towards 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 pre-processing 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 (ASM) 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 multi-level representations. The experimental results of LFD-Net on SOTS dataset reach an average Frequency Per Second (FPS) of 54.41, approximately 8 times faster than seven most popular SOTA 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. Our codes are available at <https://github.com/RacerK/LFD-Net>.

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

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]–[6]. A key pre-processing step to improve image quality is to

Manuscript received May 3, 2023; revised Jun 23, 2023. This work was supported in part by the Institute of Artificial Intelligence, the Youth talent support program of Beihang University under Grant KLSMNR-202208, 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 YWF-22-L-1277. (*Corresponding author: Kun Hu*)

Kun Hu is with the Institute of Artificial Intelligence, Beihang, China. Yizhu Jin is with the School of Automation Science and Electrical Engineering, Beihang, China (e-mail: kunhu@buaa.edu.cn, 19374316@buaa.edu.cn).

Jiaxing Chen and Feng Tian are with the National Innovation Center of Intelligent and Connected Vehicles, Beijing, China (e-mail: chenjiaxing@chinaicv.cn, fengtianreal95@gmail.com).

(Yizhu Jin and Jiaxing Chen contributed as co-first authors.)

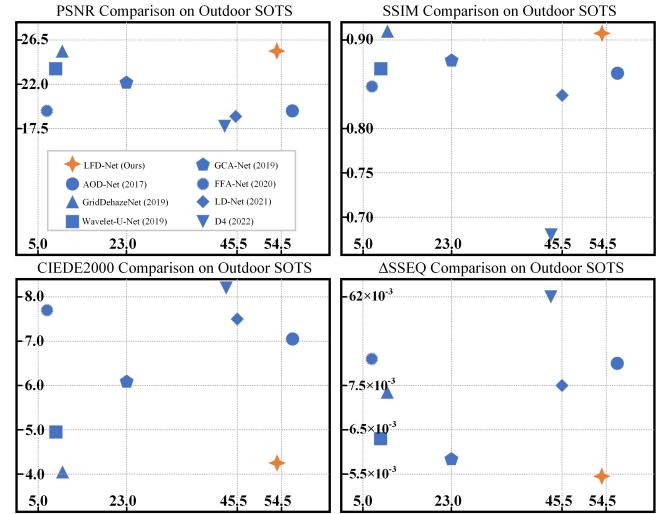


Fig. 1. Comparison Metrics on Outdoor SOTS, in terms of PSNR, SSIM, CIEDE2000, Δ SSEQ (\uparrow) and FPS (\rightarrow).

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 fast 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 makes an approximation that haze-effected pixels have at least one relatively low intensity value among RGB channels [2]; a semi-physical guided-filter based approach is adopted to refine the coarse haze thickness map to restore textural information [7]; depth estimation and image segmentation are incorporated with Dark Channel Prior to generate the final transmittance [8]. 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 contrast, color attenuation as well as hue disparity

prior to compute the transmission map and assigns a default value to atmosphere light [9]; a Haze Density Prediction Network is designed for a more accurate approximation of atmosphere light to better fit for nighttime occasions [10]; a multi-decoder framework is presented to handle multiple bad weather restoration, with rain veiling effect embedded into the conventional ASM [11], and a differential guided layer is embedded with the backbone and substituted to the physical scattering equation [11]. 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 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 multi-level 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:

- Our method employs ASM to jointly approximate atmospheric light and transmission maps to enhance image restoration capability and inference efficiency. It incorporates the convolutional operations into more specialized modules while maintaining the conciseness.
- 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 element-wise multiplication, which has been shown to enhance the performance of pure convolutional operations.
- Our proposed method has been extensively validated across various scenarios of space-air-ground remote sensing land observation tasks to demonstrate its stability, practicality, and generalization capabilities. It can effectively address common challenges such as halo effects, gridding artifacts, and color inconsistency, and achieves an excellent trade-off 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 trade-off between the efficiency and accuracy of lightweight dehazing approaches.

Some approaches employ knowledge distillation [12], [13] or pruning techniques [14], which may sacrifice accuracy for efficiency. In contrast, other methods directly construct lightweight networks to address this issue. For instance, AOD-Net [15] serves as a baseline for other lightweight dehazing models by concatenating multi-level features using different patterns. FAOD-Net [16] and GAOD-Net [17] utilize depth-wise and point-wise convolutions to reduce parameters and aggregate context information in a pyramid pooling module. FAMED-Net [18] employs cascaded and densely connected point-wise convolutional and pooling layers at multiple scales. LD-Net [19] 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 trade-off 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 Section 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 multi-level or multi-scale 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 [20], has been widely used as a backbone in various reconstruction tasks. In [21], 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 [22], the U-Net architecture is modified to incorporate discrete wavelet transform and inverse discrete wavelet transform in place of conventional down-sampling and up-sampling. In [23], 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 with 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 [24], image dehazing and depth estimation are addressed simultaneously in a framework with four decoders sharing information from the same encoder. In [25], a multi-color space encoder that incorporates RGB, LAB, and HSV is

applied to extract representative features in separate paths. In [26], quadruple color-cue is integrated into a multi-look architecture with multi-weighted training loss for autonomous vehicular application. These color spaces are often designed manually, which may work well for specific applications, but lack adaptability and generalization.

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 [27], a novel attention-based multi-scale estimation module is implemented in the backbone on a grid network to alleviate the bottleneck issue encountered in conventional multi-scale approaches. In [28], a block structure integrated with Channel-wise Attention (CA), Pixel-wise 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 [29], a multi-level 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 [30], 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 [31]. Long-range dependencies of image information can be effectively extracted through Transformer blocks in image dehazing [32]. Recently, it has been revealed in [33] that self-attention mechanism inherently functions as a two-order feature interaction. Based on this, gated convolution has been developed as an alternative method to achieve 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 inter-dependencies 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 significance 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 [34]. SSEQ is calculated by separating the

image into its spatial and spectral components, calculating and combining the entropy of each component [35]. Halo effects 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.* Δ SSEQ).

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 [15]:

$$I(\theta) = J(\theta) \times t(\theta) + A(1 - t(\theta)), \quad (1)$$

where A is treated as a constant, $t \in (0, 1]$ denotes the pixel-wise 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:

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

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

$$J(\theta) = K(\theta) \times I(\theta) - K(\theta) + b, \quad (3)$$

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

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

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 equation (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 heavy-weight and lightweight frameworks with its concise and effective approach to feature extraction and interaction, as shown in Fig. 1. To optimize the lightweight structure design of the LFD-Net, the Gated Fusion module and attention mechanism are used only once, instead of incorporating them as parts of more 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,

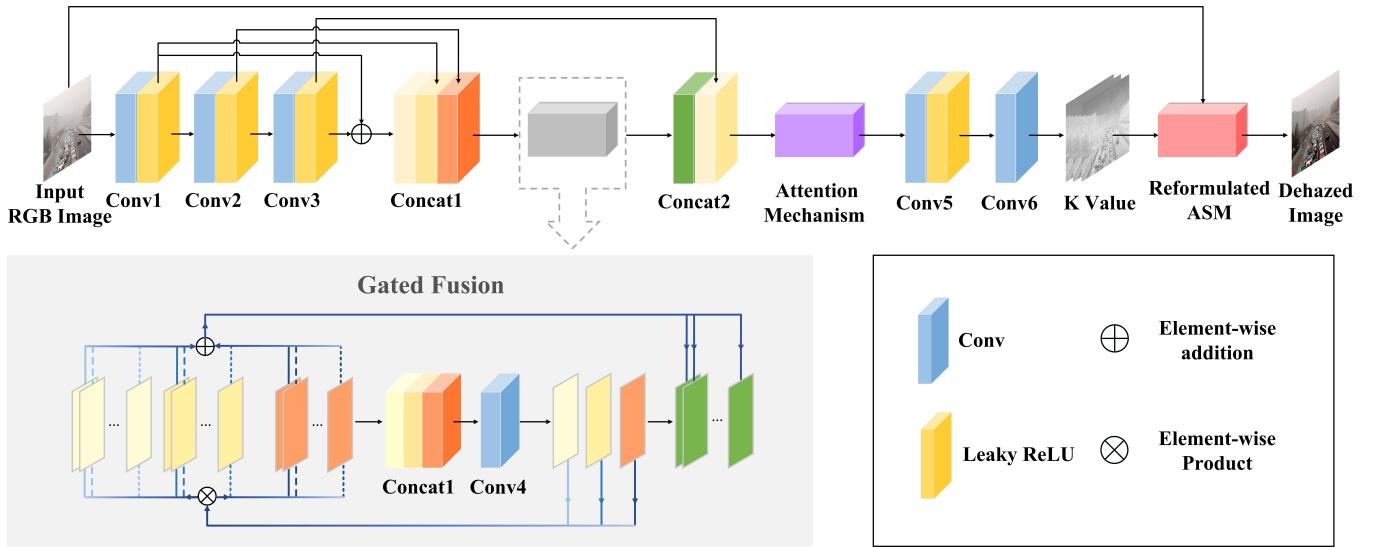


Fig. 2. Architecture of the Lightweight Feature-interaction Dehazing Network. The reformulated ASM generates an explicit output by substituting the evaluated K value. The network primarily consists of convolutional layers and concatenation layers, with the use of element-wise product in the Gated Fusion module and attention mechanism.

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 is 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 3×3 , 5×5 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 multi-level 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 channel-wise and pixel-wise 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)
<i>Conv 1</i>	3	1	1	3 / 32
<i>Conv 2</i>	5	1	2	32 / 32
<i>Conv 3</i>	7	1	2	32 / 32
<i>Concat 1</i>				<i>Conv 1</i> , <i>Conv 2</i> , <i>Conv 1 + Conv 3</i>
<i>Conv 4</i>	3	1	1	96 / 3
<i>Concat 2</i>				<i>Conv 3</i> , The Output of Gated Fusion module
<i>Conv 5</i>	3	1	1	64 / 16
<i>Conv 6</i>	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 multi-level features. This idea is demonstrated in Transformer-based architecture through two successive pixel-wise product (i.e. K , V) [36]. 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 [33], [37], [38]. Similar techniques have been practiced in image dehazing methods [39], [40] but are still in need of further exploration and interpretation.

Our proposed method also takes advantage of pixel-wise 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 i -th feature map of the

output layer is denoted as \mathcal{G}_i . The process of Gated Fusion module can be expressed mathematically as follows:

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

where $\mathcal{F}_{k,i}$ is the original i -th feature map of the k -th group. As shown in equation (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 pixel-wise multiplication. This reinforces the ability of convolution to retain and utilize multi-level features in an intensive and expansive manner.

D. Attention Mechanism

According to equation (5), the Gated Fusion module adaptively enhances and interacts with multi-level features. However, in instances where the haze is not uniformly distributed, which is common in aerial imaging, it is still challenging to accurately assess the extent and thickness of the haze region, which might lead to the presence of fancy shades or black spots. Attention mechanisms, which have been designed to focus on distinctive parts when processing large amounts of information [41], can be utilized to address this issue in image dehazing. Specifically, channel-wise attention selects the feature levels for features associated with the haze region, while pixel-wise attention refines the selected haze region. In [28], attention mechanism [42] 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.

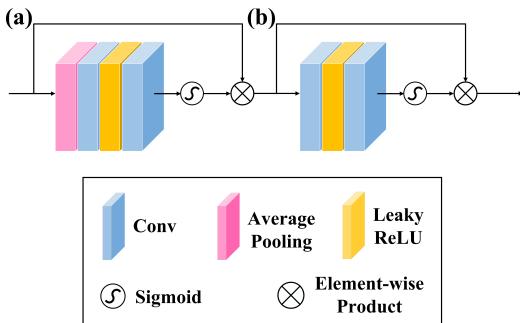


Fig. 3. The structure of Attention Mechanism. (a), (b) stand for channel-wise attention (CA) and pixel-wise attention (PA) separately.

The adopted attention mechanism is composed of Channel-wise Attention (CA) and Pixel-wise 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 Multi-Layer Perceptron architecture, with global average pooling and channel-wise mixing [43]. In this mechanism, element-wise 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:

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

Then, two successive convolutional layers with activation layers are utilized as linear transformation to obtain a one-dimensional weight vector that element-wisely multiplies the c -th feature map as follows:

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

where \mathcal{C}_1^* and \mathcal{C}_2^* are the two convolutional layers respectively, with $\delta(\cdot)$ and $\sigma(\cdot)$ be 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}° derived as follows:

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

where \mathcal{C}_1° and \mathcal{C}_2° are the two convolutional operations respectively with $\sigma(\cdot)$ be 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 channel-wise interaction, leading to fine-tuning of pixel-wise features with relatively low computational cost. In comparison to the approach presented in [19], 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 element-wise 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 [44]–[47], 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:

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

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 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** [48]. To validate the generalization ability of LFD-Net, we also use the real hazy and haze-free outdoor images dataset **O-HAZE** [49]. We fine-tune the pretrained weights from ground-based observation data by using the Aerial Image dataset AID [50] for satellite (*i.e.* space-based) and drone (*i.e.* air-based) and test on the Remote sensing Image Cloud rEmoving dataset **RICE** [51]. 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** [52] and **VisDrone2019** [53], and evaluate the performance of object detection tasks using hazy and dehazed images for comparison.

B. Experiment Results

We faithfully reproduce 7 State Of The Art (SOTA) methods for various outdoor scenarios, including AOD-Net [15], Grid-DehazeNet [27], Wavelet-U-Net [22], GCA-Net [40], FFA-Net [28], LD-Net [19] and D4 [54]. 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 Fig. 4 and Fig. 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 Fig. 6 and Fig. 7.

In the remote sensing domain, to the best of our knowledge, pre-trained models for dehazing methods are not publicly available. However, we also reproduce 7 SOTA methods using default outdoor weights, including AOD-Net [15], GridDehazeNet [27], GCA-Net [40], FFA-Net [28], MSBDN [55], D4 [54] and DehazeFormer [56]. 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 paper, 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,

Δ SSEQ, and FPS. The comparison results are summarized in Tables II, III, and IV. Additionally, we provide a comparison of model sizes in Table V.

TABLE II. Average Comparison of Metrics on SOTS for 492 JPG Images

	PSNR↑	SSIM↑	CIEDE↓	Δ SSEQ ↓	FPS↑
AOD-Net	19.45	0.8593	7.12	0.0080	56.85
GridDehazeNet	<u>25.07</u>	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	<u>0.0058</u>	23.01
FFA-Net	19.36	0.8472	7.70	0.0131	6.91
LD-Net	18.62	0.8380	7.52	0.0075	45.44
D4	18.09	0.6668	8.36	0.0624	44.79
LFD-Net	25.12	<u>0.9087</u>	<u>4.24</u>	0.0054	<u>54.41</u>

TABLE III. Average Comparison of Metrics on O-HAZE for 45 JPG Images

	PSNR↑	SSIM↑	CIEDE↓	Δ SSEQ ↓	FPS↑
AOD-Net	15.06	0.5412	17.53	0.0146	<u>13.93</u>
GridDehazeNet	16.68	0.6361	<u>13.54</u>	0.0055	1.09
Wavelet-U-Net	15.87	0.5058	14.93	0.0041	8.66
GCA-Net	<u>17.24</u>	<u>0.6523</u>	13.81	0.0077	4.92
FFA-Net	14.62	0.5881	14.72	0.0067	1.52
LD-Net	14.72	0.5583	16.74	0.0057	10.62
D4	11.51	0.2564	18.58	0.1612	2.70
LFD-Net	17.67	0.6532	<u>11.80</u>	0.0041	14.65

TABLE IV. Average Comparison of Metrics on RICE1 for 500 PNG Images

	PSNR↑	SSIM↑	CIEDE↓	Δ SSEQ ↓	FPS↑
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	<u>10.39</u>	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
DehazeFormer	<u>19.95</u>	<u>0.8615</u>	10.82	<u>0.0018</u>	4.29
LFD-Net	30.88	0.9420	<u>3.32</u>	0.0008	<u>45.51</u>

Observations reveal that many networks suffer from inconsistencies within color blocks or misrepresenting original information, as reflected in terms of CIEDE2000 and Δ SSEQ. For instance, lightweight methods such as AOD-Net [15], LD-Net [19], and D4 [54] produce relatively dark visual quality, resulting in a significant loss of texture information and making it difficult to distinguish objects for downstream tasks. While GCA-Net [40] encounters severe 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 effects and color imbalances are magnified in RICE1, which makes it less adaptive to generalized scenarios. FFA-Net [28] 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 to insufficient attention paid to each hazy region and overfitting on specific datasets with limited data diversity,

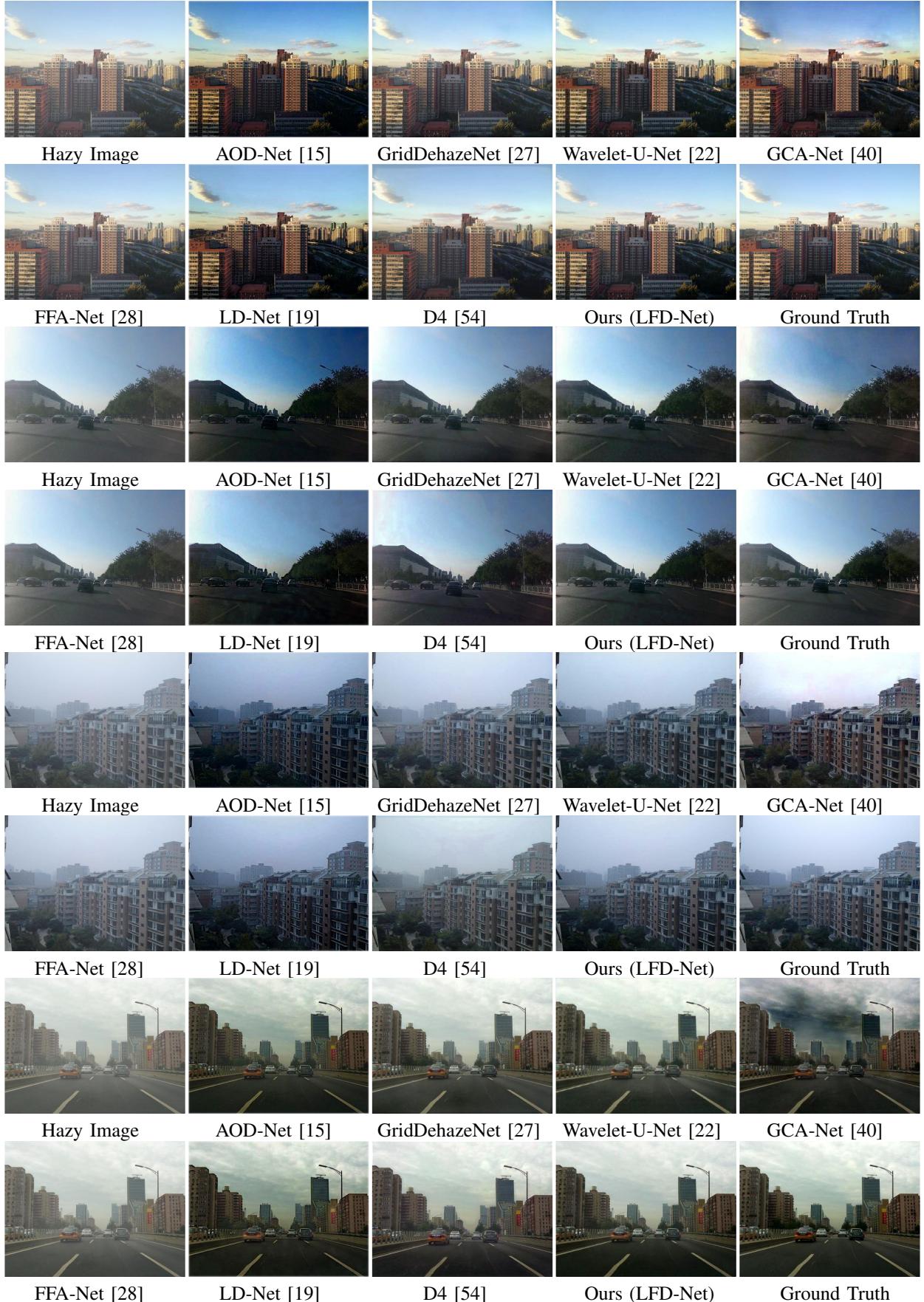


Fig. 4. Visual Comparison on Outdoor SOTS. We compare our methods with AOD-Net [15], GridDehazeNet [27], Wavelet-U-Net [22], GCA-Net [40], FFA-Net [28], LD-Net [19] and D4 [54]. Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.



Fig. 5. Visual Comparison Results on O-HAZE. We compare our methods with AOD-Net [15], GridDehazeNet [27], Wavelet-U-Net [22], GCA-Net [40], FFA-Net [28], LD-Net [19] and D4 [54]. AOD-Net and LD-Net 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.

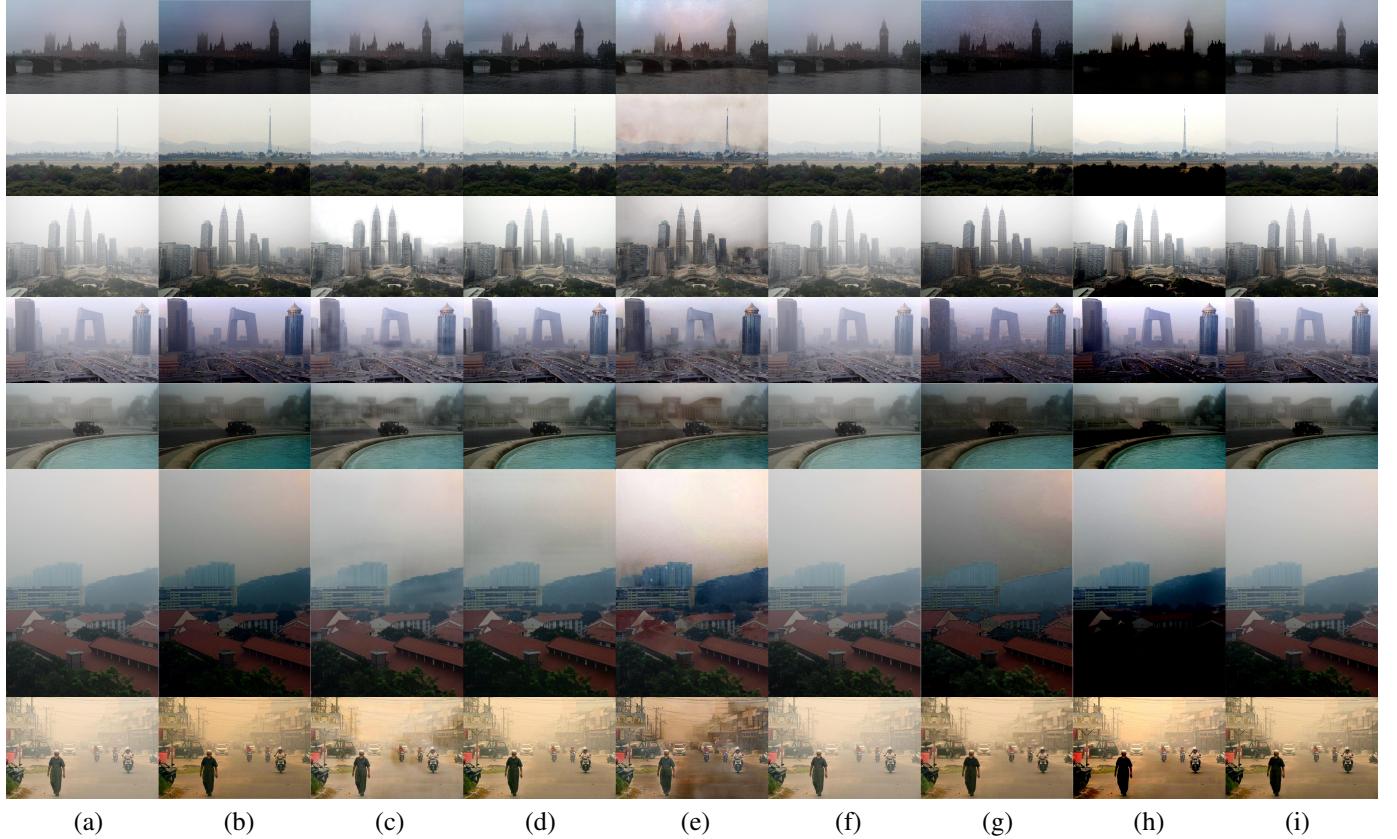


Fig. 6. Visual Comparison Results on Real-world HSTS. (a) Hazy image, (b) AOD-Net [15], (c) GridDehazeNet [27], (d) Wavelet-U-Net [22], (e) GCA-Net [40], (f) FFA-Net [28], (g) LD-Net [19], (h) D4 [54] and (i) Ours (LFD-Net). Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

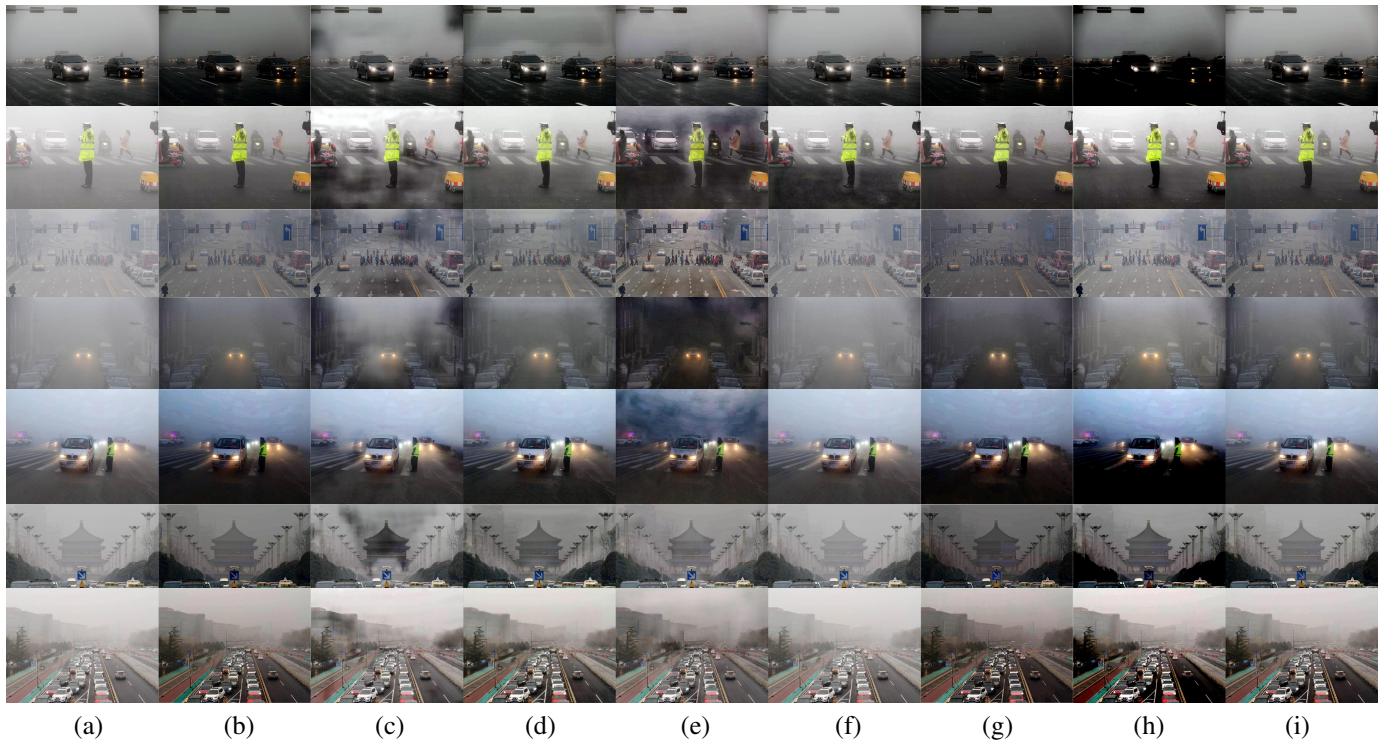


Fig. 7. Visual Comparison Results on Randomly Selected Real-world Images. (a) Hazy image, (b) AOD-Net [15], (c) GridDehazeNet [27], (d) Wavelet-U-Net [22], (e) GCA-Net [40], (f) FFA-Net [28], (g) LD-Net [19], (h) D4 [54] and (i) Ours (LFD-Net). Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

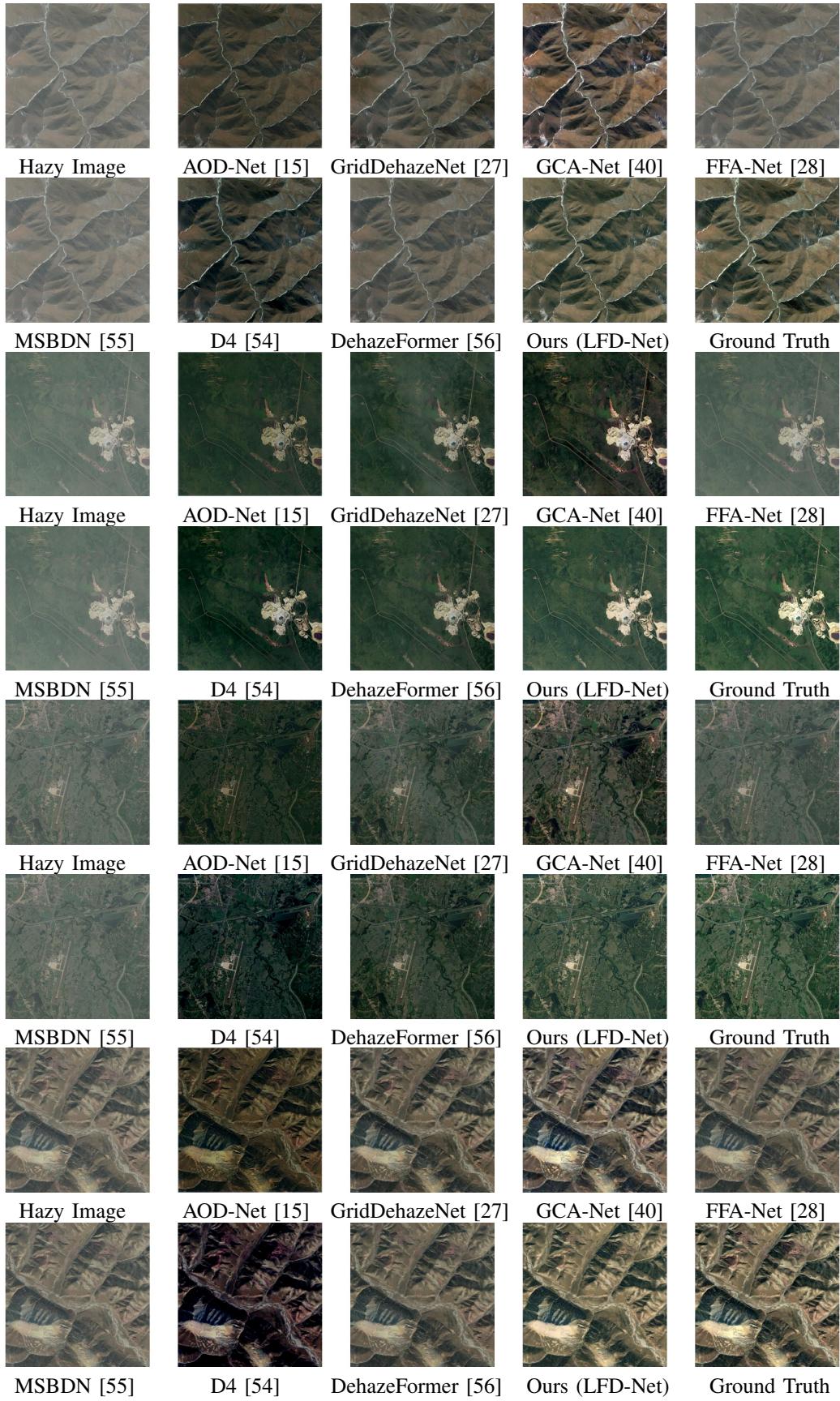


Fig. 8. Visual Comparison Results on O-HAZE. We compare our methods with AOD-Net [15], GridDehazeNet [27], Wavelet-U-Net [22], GCA-Net [40], FFA-Net [28], LD-Net [19] and D4 [54]. AOD-Net and LD-Net produce relatively dark in visual quality. GCA-Net performs well on irregular haze but suffers from inconsistency in color blocks. Our proposed method exhibits adaptability to diverse scenarios and possesses a noteworthy level of generalization.

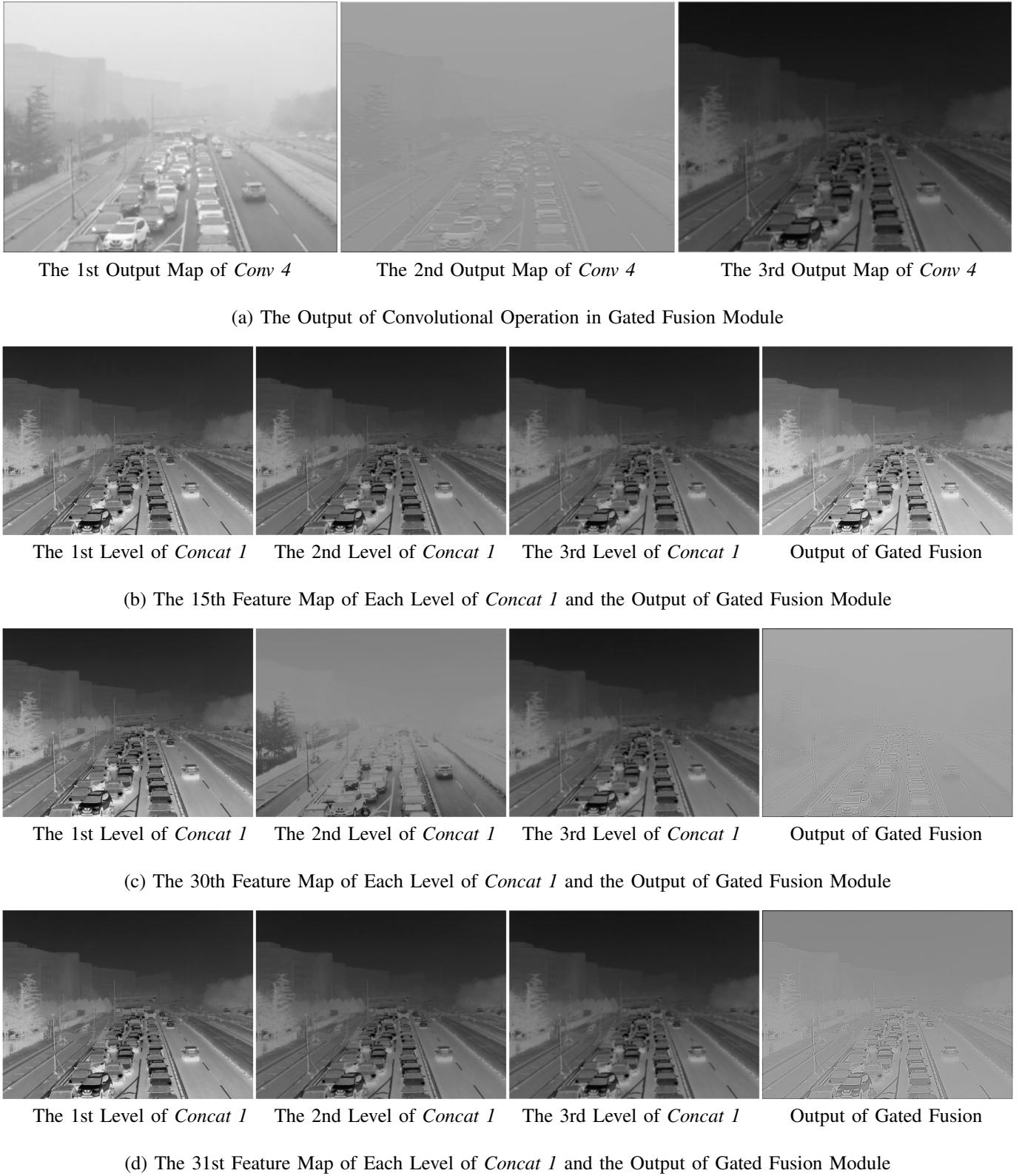


Fig. 9. Visualization Results of the changes in layers before and after the Gated Fusion Module: (a) represents the three output feature maps of the convolutional operation *Conv 4* incorporated into Gated Fusion module, (b)-(d) stand for the changes in the 15th, 30th and 31st feature map of the layers 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.

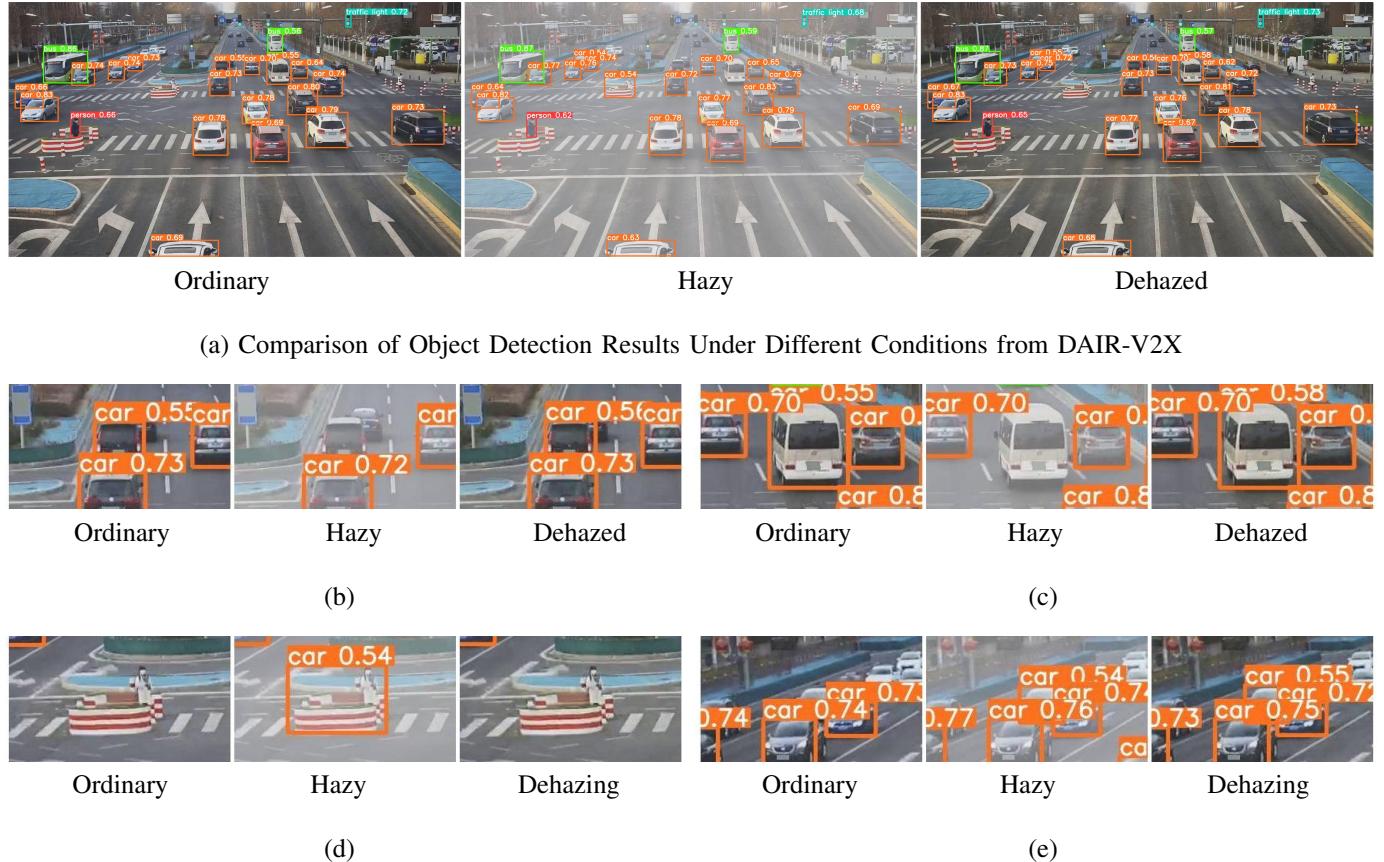


Fig. 10. Reference Object Detection Results: (a) comparison of object detection results under ordinary, simulated hazy, and dehazed conditions, (b)-(e) detailed sub-scenes of detection results under different conditions. (b) and (c) demonstrate an improvement in the detection rate, detecting an additional 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.

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 multi-scale features are addressed symmetrically. Wavelet-U-Net [22] and GridDehazeNet [27] have excellent performance, but they may come at the cost of inference time, ***6.6 \times* and *5.4 \times* 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 multi-scale features, which ensembles the design of U-Net [20]. 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 [15] and [19], 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 or attention mechanism in our model. These cases demonstrate that these two sub-networks

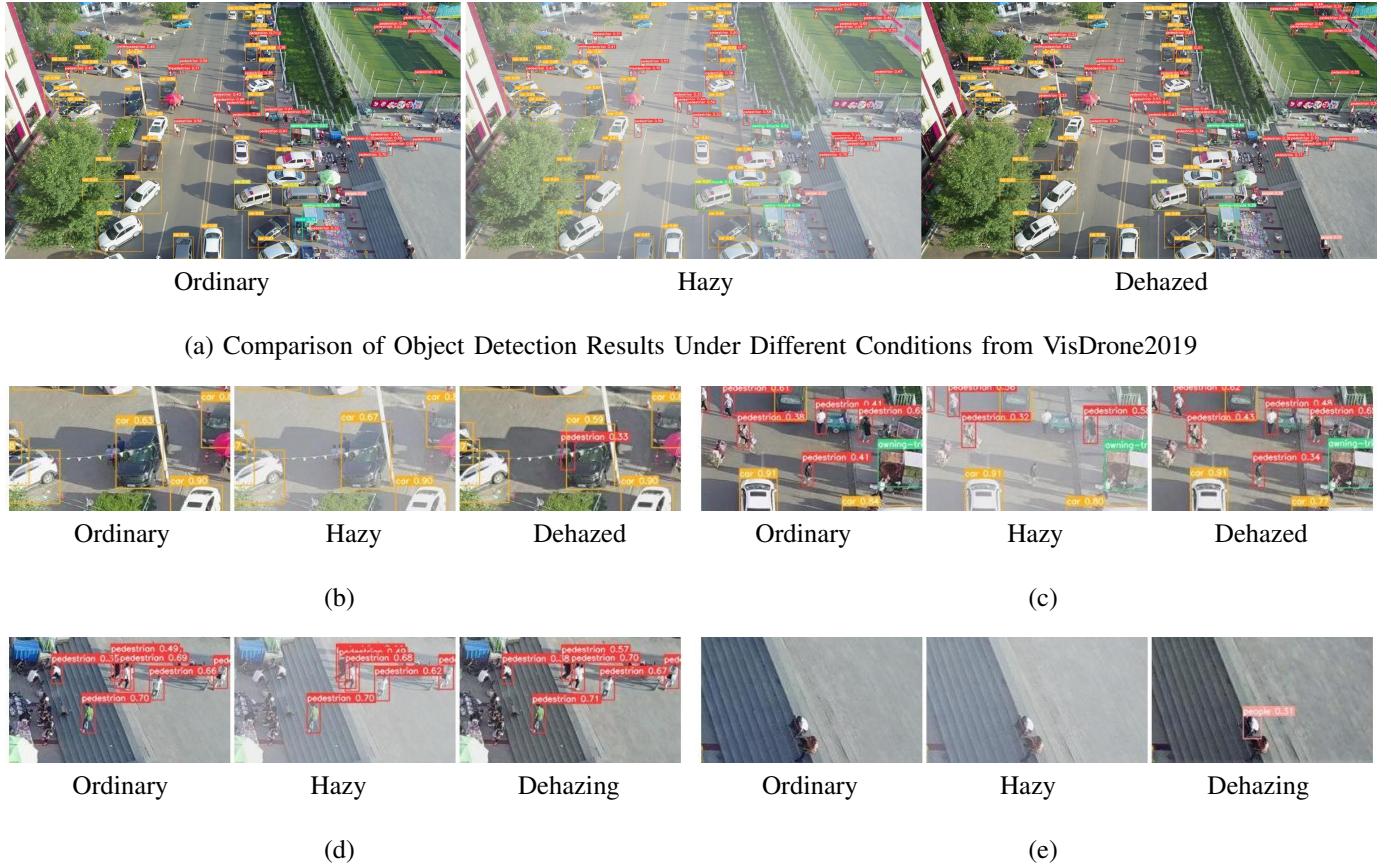


Fig. 11. Reference Remote Sensing Object Detection Results: (a) comparison of remote sensing object detection results under ordinary, simulated hazy, and dehazed conditions, (b)-(e) detailed sub-scenes of detection results [under different conditions](#), in which the detection rate for pedestrians is 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.

TABLE V. Comparison of the Parameters of Models

	AOD-Net	GridDehazeNet	Wavelet-U-Net	GCA-Net	FFA-Net	LD-Net	D4	MSBDN	DehazeFormer	LFD-Net
Params($\times M$)	0.002	0.956	11.3	0.702	4.46	0.030	10.7	28.7	2.51	0.086

TABLE VI. Ablation Experiment of LFD-Net on Outdoor SOTS Dataset

	Concat 2	Gated Fusion Module	Attention Mechanism	PSNR↑	SSIM↑	CIEDE↓	$\Delta SSEQ \downarrow$
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

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. Additionally, the Gated Fusion module is a crucial component in enhancing the dehazing capability, serving as a bridge between the multi-level feature extraction process, which ends at the first concatenation layer *Concat 1*, and the attention mechanism, which begins 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 multi-level 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 Fig. 9 (d) show 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 re-allocates the distributed feature representations of the multi-level layers through feature-interaction strategies. The feature extraction process is compressed into three successive convolutional layers, for which we compensate by intra-level enhancement and inter-level combination.

E. Application for Object Detection Task

As a severe weather condition, haze can significantly reduce the effectiveness of remote sensing land obsevation system. For instance, object detection results in automatic driving applications suffer from hazy environments and may be put into risky situations due to the degradation of image quality. Therefore, pre-processing 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 fog images for object detection. In our experiment, 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) illustrate 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) illustrate 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 paper, we propose a novel end-to-end model called LFD-Net for remote sensing image dehazing. As a pre-processing 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 super-resolution. The experiments on various scenarios also shows that performance of a model is not always propotional 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*. John Wiley & Sons, 2014.
- [2] A. S. Li, V. Chirayath, M. Segal-Rozenhaimer, J. L. Torres-Perez, and J. van den Bergh, “Nasa nemo-net’s convolutional neural network: mapping marine habitats with spectrally heterogeneous remote sensing imagery,” *IEEE journal of selected topics in applied earth observations and remote sensing*, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 929–10 940, 2021.
- [4] Y. Zheng, J. Su, S. Zhang, M. Tao, and L. Wang, “Dehaze-agan: Unpaired remote sensing image dehazing using enhanced attention-guide generative adversarial networks,” *IEEE Transactions on Geoscience and Remote Sensing*, 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 Geoscience and Remote Sensing Letters*, 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 Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, pp. 5895–5905, 2014.
- [7] F. Liu, Y. Lv, B. Li, S. Gao, and Y. Qin, “A semophysical approach of haze removal for landsat image,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7410–7421, 2021.
- [8] B. Xie, J. Yang, J. Shen, and Z. Lv, “Image defogging method combining light field depth estimation and dark channel,” in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 2021, pp. 745–749.
- [9] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “Dehazenet: An end-to-end system for single image haze removal,” *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.
- [10] Y. Liao, Z. Su, X. Liang, and B. Qiu, “Hdp-net: Haze density prediction network for nighttime dehazing,” in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 469–480.
- [11] R. Li, R. T. Tan, and L.-F. Cheong, “All in one bad weather removal using architectural search,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3175–3185.
- [12] M. Hong, Y. Xie, C. Li, and Y. Qu, “Distilling image dehazing with heterogeneous task imitation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3462–3471.
- [13] A. Suresh, J. Nisha, V. P. Gopi *et al.*, “Rich feature distillation with feature affinity module for efficient image dehazing,” *Optik*, vol. 267, p. 169656, 2022.

- [14] 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 *International Conference on Image, Vision and Intelligent Systems*. Springer, 2022, pp. 442–453.
- [15] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4770–4778.
- [16] W. Qian, C. Zhou, and D. Zhang, "Faod-net: A fast aod-net for dehazing single image," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–11, 2020.
- [17] S. Chen, J. Cheng, and Z. Huang, "Gado-net: An improved aod-net single image dehazing algorithm," in *2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, 2021, pp. 640–646.
- [18] J. Zhang and D. Tao, "Famed-net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE*, no. 1, 2020.
- [19] H. Ullah, K. Muhammad, M. Irfan, S. Anwar, M. Sajjad, A. S. Imran, and V. H. C. de Albuquerque, "Light-dehazenet: a novel lightweight cnn architecture for single image dehazing," *IEEE Transactions on Image Processing*, vol. 30, pp. 8968–8982, 2021.
- [20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [21] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2157–2167.
- [22] H.-H. Yang and Y. Fu, "Wavelet u-net and the chromatic adaptation transform for single image dehazing," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2736–2740.
- [23] T. Feng, C. Wang, X. Chen, H. Fan, K. Zeng, and Z. Li, "Urnet: A u-net based residual network for image dehazing," *Applied Soft Computing*, vol. 102, p. 106884, 2021.
- [24] B.-U. Lee, K. Lee, J. Oh, and I. S. Kweon, "Cnn-based simultaneous dehazing and depth estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9722–9728.
- [25] 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 Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021.
- [26] 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 Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4256–4266, 2020.
- [27] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7314–7323.
- [28] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 908–11 915.
- [29] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided cnn for image dehazing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4162–4173, 2020.
- [30] H. Bai, J. Pan, X. Xiang, and J. Tang, "Self-guided image dehazing using progressive feature fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 1217–1229, 2022.
- [31] 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, p. 3428, 2022.
- [32] Y. Yang, H. Zhang, X. Wu, and X. Liang, "Mstfdn: Multi-scale transformer fusion dehazing network," *Applied Intelligence*, pp. 1–12, 2022.
- [33] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-N. Lim, and J. Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," *arXiv preprint arXiv:2207.14284*, 2022.
- [34] M. R. Luo, G. Cui, and B. Rigg, "The development of the cie 2000 colour-difference formula: Ciede2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350, 2001.
- [35] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal processing: Image communication*, vol. 29, no. 8, pp. 856–863, 2014.
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [38] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 030–11 039.
- [39] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3253–3261.
- [40] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1375–1383.
- [41] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [43] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24 261–24 272.
- [44] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.
- [45] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.
- [46] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2710–2719.
- [47] Y. Guo, J. Chen, X. Ren, A. Wang, and W. Wang, "Joint raindrop and haze removal from a single image," *IEEE Transactions on Image Processing*, vol. 29, pp. 9508–9519, 2020.
- [48] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.
- [49] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 754–762.
- [50] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [51] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," *arXiv preprint arXiv:1901.00600*, 2019.
- [52] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.
- [53] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.
- [54] Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, and D. Tao, "Self-augmented unpaired image dehazing via density and depth decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2037–2046.
- [55] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2157–2167.
- [56] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *arXiv preprint arXiv:2204.03883*, 2022.



Yizhu Jin is currently a final year undergraduate student in the School of Automation Science and Electrical Engineering from Beihang University, Beijing, China. Jin's research interests lie in the intersection between valuable application and general methods with high interpretability in Computer Vision and Artificial Intelligence.



Jiaxing Chen received M.S. degree in Electrical and Computer Engineering from University of Illinois, Chicago, USA, in 2021 and worked as an Algorithm engineer in National Innovation Center of Intelligent and Connected Vehicles in 2022. He is currently a researcher with Tsinghua University, Beijing, China. Chen's research interests include Computer Vision with Deep Learning, Machine Learning, Multimodal Fusion, and Trajectory Prediction.



Feng Tian received her B.S. degree in engineering in 2018. She is currently studying in the College of Traffic and Logistics Engineering, Xinjiang Agricultural University, Xinjiang, China, pursuing a Postgraduate degree in transportation engineering. Tian's research interests include Machine Learning, Data Mining, Intelligent Transportation, and Target Detection.



Kun Hu received the B.Sc. degree in Remote Sensing Science and Technology from Wuhan University, Wuhan, China, in 2010, the M.Sc. and Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University in 2012 and 2016, respectively. He worked as an assistant professor with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China from 2012 to 2017, as a post-doctor with the Ohio State University, Ohio, USA, from 2017 to 2019, and as an associate professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China from 2019 to 2022.

He is currently an associate professor with the Institute of Artificial Intelligence, Beihang University, Beijing, China. His research interests focus on the accurate processing and intelligent application of multi-source remote sensing data, such as camera calibration, image production, information fusion, target detection, clarification, 3D reconstruction and quality evaluation.