C^{2}AIR: Consolidated Compact Aerial Image Haze Removal

Ashutosh Kulkarni[1], Shruti S. Phutke[2], Santosh Kumar Vipparthi[1], and Subrahmanyam Murala[3]
[1]CVPR Lab, Indian Institute of Technology Ropar, India
[2]Institute for Integrated and Intelligent Systems, Griffith Univeristy, Australia
[3]CVPR Lab, School of Computer Science and Statistics, Trinity College Dublin, Ireland
ashutosh.20eez0008@iitrpr.ac.in

Abstract

Aerial image haze removal deals with improving the visibility and quality of images captured from aerial platforms, such as drones and satellites. Aerial images are commonly used in various applications such as environmental monitoring, and disaster response. These applications usually require cleaner data for accurate functioning. However, atmospheric conditions such as haze or fog can significantly degrade the quality of these images, reducing their contrast, color saturation, and sharpness, making it difficult to extract meaningful information from them. Existing methods rely on computationally heavy and haze density (light, moderate, dense) specific architectures for aerial image dehazing. In light of these limitations, we propose a novel lightweight and consolidated approach for aerial image dehazing. In this approach, we propose Density Aware Query Modulated Block for learning weather degradations in input features and guiding the restoration process. Further, we propose Cross Collaborative Feed-Forward Block for learning to restore varying sizes of the structures in the input images. Finally, we propose Gated Adaptive Feature Fusion block to achieve inter-scale and intra-feature attentive fusion, effective for aerial image restoration. Extensive analysis on benchmark aerial image dehazing datasets and real-world images, along with detailed ablation studies validate the effectiveness of the proposed approach. Further, we have analysed our method for other restoration task such as underwater image enhancement to experiment its wide applicability. The code is available at https://github.com/AshutoshKulkarni4998/C2AIR.

1. Introduction

Aerial imagery is an important tool for various applications such as surveillance [31, 32], urban planning, disaster management, and military reconnaissance. However, the images captured from an aerial platform can often be degraded due to the presence of atmospheric haze with various densities, which can reduce the visual quality and make it difficult to discern important details. Therefore, the development of an efficient and effective aerial image dehazing model is crucial for improving the performance of dependent application significantly.

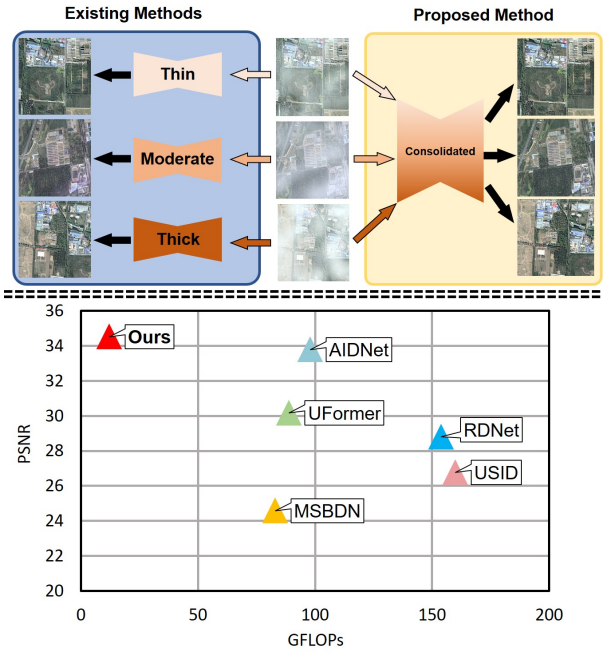

Figure 1. *First Row:* For different densities, the existing methods require separately trained models vs proposed method which consists of a consolidated model. *Second Row:* Graphical illustration of the computational complexity and performance of the proposed method with existing state-of-the-art methods in terms of GFLOPs (on image size 256 × 256) vs PSNR (on RICE dataset). The proposed consolidated method achieves better performance with lesser computational complexity.

For aerial image dehazing, existing handcrafted approaches utilized virtual point clouds [28], dark-channel prior [10, 28], frequency correlation [44], *etc.* However, such hand crafted methods perform better on specific set

of images and do not generalize well on real-world scenarios. The recent advancements in deep learning architectures boosted the research in the direction of image restoration [16, 17]. Specifically, using the adaptive capabilities of convolutional neural networks (CNNs), the researchers have proposed improved approaches for aerial image dehazing. These approaches include generative adversarial networks (GANs) [30], unsupervised learning [29], *etc*. Huang *et al*. [12] proposed a synthetic aperture radar (SAR) prior based approach for aerial image dehazing, which has a limitation of prior dependency. For eliminating the data prior dependency, [15] proposed a transformer based approach for aerial image dehazing, but imposes higher computation burden and lower runtime. All the above-mentioned learning based methods utilize separate training checkpoints for different densities of haze.

In brief, the existing aerial image dehazing approaches have certain limitations: **1) Single-Domain Applicability:** Existing methods utilize separately trained models for dehazing images with different densities. **2) Computational Complexity:** The existing approaches possess high computational complexity, making them inapplicable in practical scenarios. Therefore, there is a need of dehazing model that has less computational complexity, making it suitable for real-world applications, and generalizes well to different types of aerial images, ensuring better performance and accuracy.

To circumvent these limitations, we propose a novel lightweight and consolidated approach for aerial image dehazing, having only 1.7M trainable parameters (*half* of the existing method [5]), and requiring only single checkpoint model for restoring images with different densities of haze. In contrast to existing method [12], the proposed method does not require any additional data as prior information. Next, the existing methods do not have provision for learning density-related features, whereas the proposed network adopts learning of such features through the proposed Density Aware Query Modulation block. This block allows the proposed network to remove various haze densities present in aerial images in a consolidated manner. It is noteworthy that the proposed method *requires only single trained checkpoint*. Further, state-of-the-art aerial image dehazing approach AIDNet [15] directly passes inter-stream information, making it prone to passing irrelevant degraded features. Unlike AIDNet [15], we give provision to the proposed network for fusing the inter-stream features in an adaptive manner, and propagate most relevant features with the proposed Adaptive Gated Feature Fusion block. Further, in contrast to previous approaches which do not consider multi-receptive learning in their basic blocks, we propose a Cross Collaborative Feed-Forward blocks to focus on multi-receptive information, essential for restoring images. The main contributions of this work are enlisted as:

- First consolidated approach for aerial hazy image restoration, which is a single trained model to restore images with various densities of haze, and has less computational complexity than existing methods (refer Figure 1 for overview).
- A Density Aware Query Modulated (DAQM) Block is proposed for adapting to various densities of haze present in aerial images (Sec. 3.1).
- A Cross Collaborative Feed-Forward (CCF) Block equipped with multi-kernel feature extractors is proposed to capture varying sizes of structures in the aerial images (Sec. 3.2).
- An Adaptive Gated Feature Fusion (AGFF) Block is proposed for fusing and propagating relevant information within the proposed network (Sec. 3.3).

Substantial experiments on synthetic and real world images, along with detailed ablation studies, verify the effectiveness of the proposed method for hazy aerial image restoration. We further evaluate the applicability of the proposed method by experimenting on other image restoration task, such as underwater image enhancement.

2. Literature Review

Initial efforts were focused on haze removal from outdoor images using hand-crafted priors, as reported in several studies such as [1, 7, 10, 36, 37, 47]. He *et al.* [10] introduced DCP which is a baseline prior relevant to haze for obtaining coarse-level depth information to de-haze the image. However, it exhibited a halo effect near complex edge structures and failed to produce satisfactory results in the sky regions. Salazar-Colores *et al.* [34] improved upon DCP by combining it with mathematical morphology operations like erosion and dilation to efficiently compute transmission maps. In recent years, researchers [2, 5, 18, 27, 33, 35, 40, 41, 45] have developed convolutional neural networks (CNNs) for transmission map estimation followed by atmospheric scattering models for achieving haze removal. Cai *et al.* [2] proposed a deep network that estimates the transmission map and uses an atmospheric scattering model for haze removal from the image. Dong *et al.* [5] proposed a multiscale boosted decoder based on dense connections. Zhao *et al.* [45] proposed a two-stage framework with weakly-supervised learning and unpaired adversarial learning. Jia *et al.* [13] utilized a network by leveraging meta attention, and Liu *et al.* [26] proposed a feature extraction-based method for integrating all characteristic information for haze removal in a multi-branch manner. Chen *et al.* [4] proposed the adaptation of network trained on synthetic data to enhance performance on real-world data. Li *et al.* [20] proposed a compact multi-scale

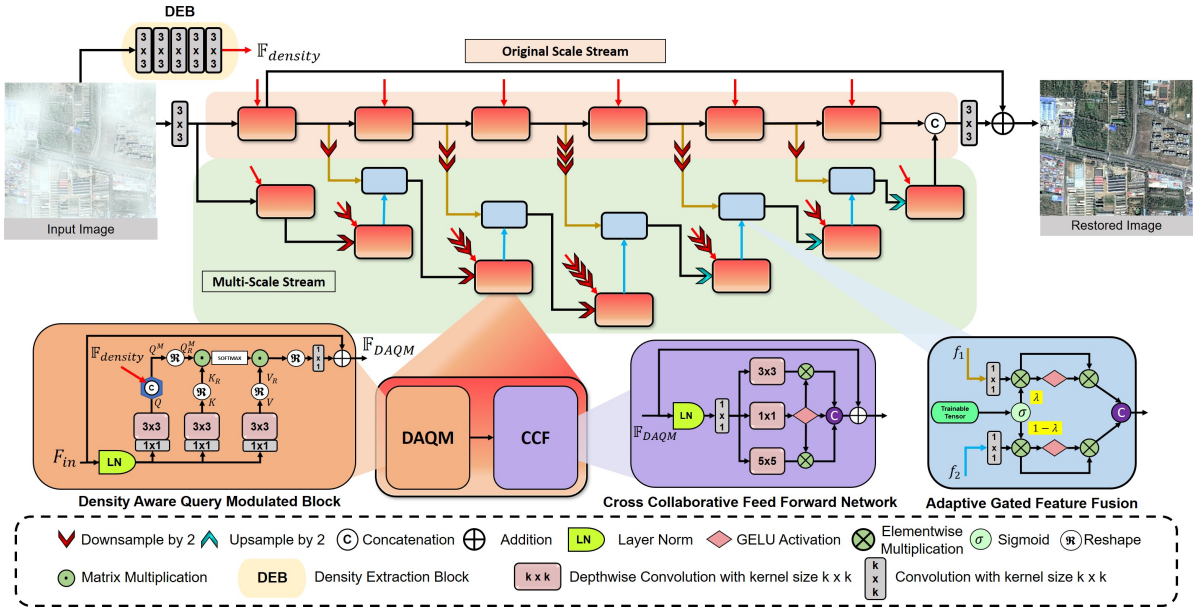

Figure 2. Representation of the proposed network for aerial image haze removal.

feature attention and multi-frequency representation learning network trained in unsupervised manner.

Various techniques have been developed, especially for haze removal from aerial images. For example, Zhang *et al.* [44] proposed a correction technique that correlates multi-level color bands. Liu *et al.* [24] utilized virtual cloud points, while Long *et al.* [28] utilized DCP proposed by He *et al.* [10] for removing haze from real-world images. Guo *et al.* [9] employed residual learning strategies and channel attention modules for fast network convergence and effective channel correlation. Pan [30], proposed a model with local-to-global spatial attention for cloud and haze removal. Grohnfeldt *et al.* [8] used a cGAN with the fusion of SAR prior and multi-spectral image data for cloud removal. Huang *et al.* [12] used RGB and SAR prior information for aerial image de-hazing with dilated convolution based GAN. Mehta *et al.* [29] proposed SkyGAN, a GAN framework incorporating hyper-spectral images (HSI) guidance for aerial image de-hazing. However, these methods lack in capturing the long range dependencies. This issue is mitigated with advancements of transformer architectures.

Transformers have gained popularity over CNNs due to their ability to capture long-range dependencies. They have been used in computer vision through Vision Transformers (ViT) [6], which employ flattened patches of images while training. Image processing transformers have also been used for low-level vision tasks, as demonstrated by [3], who showed how pre-training on large datasets can improve performance. UFormer [39] uses a U-Net like structure with transformers for image restoration tasks. [15] proposed a transformer-based network for aerial image dehazing, which extracts transformer embeddings in a spatially attentive manner. However, this approach requires separately trained models and higher computational cost. Further, these explained methods do not consider various densities of haze present in the aerial images as they do not contain feature processing blocks required for generalization, which is addressed in the proposed method.

3. Proposed Aerial Image Dehazing Method

The main aim of the proposed method is to restore the visibility in aerial images containing different densities of haze in a consolidated manner (single checkpoint model) while maintaining low computational complexity. To achieve this, we propose: **(a)** Density Aware Query Modulation Blocks (DAQM) to deal with the degradation-relevant feature extraction, **(b)** Cross Collaborative Feed-forward Blocks (CCF) to learn multi-receptive features, and **(c)** Adaptive Gated Feature Fusion Block (AGFB) to collect and adaptively fuse the features from both the streams in network. In the following sections, we provide a detailed explanation of the proposed blocks. Proposed network architecture is illustrated in Figure 2.

Overall Pipeline: The input image is passed through Density Extraction Block (DEB), whose output ($\mathbb{F}_{density}$, denoted by red arrows in Figure 2) is provided to each DAQM block. Then, the input image is passed through initial convolution layer which is then further passed through two-stream interconnected network. The first stream processes the input in the original scale, and the second stream extracts features in a multi-scale manner. The streams are fused with the proposed Adaptive Gated Feature Fusion blocks. Finally, outputs of both the streams are concatenated and passed through a final convolution block, after

which it is added with the input to get a haze removed image. *The architecture configurations are provided in the supplementary material.*

3.1. Density Aware Query Modulated Block

Aerial images are often degraded by haze in varying densities, ranging from thin to moderate and thick. To effectively restore such images, it is necessary for the deep learning network to adapt the varying density of haze present in the images. Inspired by application of transformers in natural language processing (NLP) [6, 38, 39, 42, 46], where from the embedding of a word (query) from the input text, the importance of every other word embeddings (keys) in the same text (importance with respect to query) is measured and an updated embedding is created by merging their information. Inspired from such update strategy, we correspond the queries as density extracted information, where the output features can be obtained with relevance to the extracted density features. To achieve this, we propose the Density Aware Query Modulated blocks (DAQM) detailed as follows.

Let F_{in} be the input features to the DAQM. The DAQM block involves initial extraction of query (Q), key (K) and value (V) from the layer normalized tensor (F_{in}^{LN}) by applying 1×1 convolutions followed by 3×3 depth-wise convolutions for encoding non-local and channel-wise spatial context. The density features $\mathbb{F}_{density}$ are extracted from the input hazy image using a Density Extraction Block (DEB) (see Figure 2). DEB contains a series of 3×3 convolution layers. This is intended to gradually increase the receptive field and learn diverse features from the input image. The earlier convolutional layers learn edges and textures affected with different densities of haze whilst the later (deeper) layers learn the global density features, due to gradual increase in receptive field. Precisely, it captures patterns in the input image that correlates with varying levels of haze density. The extracted density features are then fed into the Density Aware Query Modulation Block, where they are fused with the extracted queries through concatenation to get Q^M. Later, the modulated query (Q^M), key (K) and value (V) projections are reshaped (denoted as Q_R^M, K_R, V_R) to maintain low computational complexity in order of feature channels instead of spatial dimensions [42]. We further obtain the output of DAQM block ($\mathbb{F}_{DAQM}$) as:

$$\mathbb{F}_{DAQM}(F_{in}, \mathbb{F}_{density}) = F_{in} + C_1(\Re(V_R \odot \delta(Q_R^M \odot K_R))) \tag{1}$$

where, C_N is convolution with kernel size $N \times N$, $\Re(\cdot)$ is the reshaping operation which reshapes the tensor back to shape of F_{in}. $\odot$ is matrix multiplication and $\delta(\cdot)$ is the softmax operation.

The fusion of density features into the queries in the DAQM block allows the network to achieve a restoration process that adapts to the density of haze present in the image, promoting its robustness and generalization across various types of hazy images. The effectiveness of the proposed DAQM is provided in Sec. 5.

3.2. Cross Collaborative Feed-Forward Block

Aerial images contain a diverse range of structures, including buildings, roads, vegetation, and water bodies, each with distinct shapes and sizes. One of the challenges in image restoration tasks is to learn such intricate and varying structures. Traditional approaches often use a fixed-size kernel to capture the variations in size, which limits their effectiveness in handling complex structures. To address this challenge, we have introduced the Cross Collaborative Feed-Forward Block (CCF) in the proposed network architecture. The term "cross-collaborative" refers to the intergating mechanism between features extracted using convolutions with different kernel sizes. The output features from CCF can be equated mathematically as:

$$\mathbb{F}_{CCF}(F_{in}) = F_{in} + \langle \wp\left(C_3^d(\phi), C_1^d(\phi)\right), \wp\left(C_5^d(\phi), C_1^d(\phi)\right) \rangle \tag{2}$$

$$\wp(x, y) = x * \zeta(y) \tag{3}$$

$$\phi = C_1\left(\mathbb{F}_{DAQM}^{LN}\right) \tag{4}$$

where, $\langle a, b \rangle$ represents concatenation of a, b, $\wp(\cdot)$ is the gating operation, $\zeta(\cdot)$ is GELU activation [11], and $\mathbb{F}_{DAQM}^{LN}$ represents layer norm of $\mathbb{F}_{DAQM}$. Incorporating cross-kernel collaborative learning enables the ability to effectively attend to diverse receptive fields within input features, leading to enhanced perceptual quality in the resultant images. The effectiveness of the proposed CCF block is analysed in Sec. 5.

3.3. Adaptive Gated Feature Fusion

We observe that relevant feature fusion and propagation are important prerequisites for generalized aerial image haze removal. The existing method outlined in [15] employed a feature fusion strategy that involved the direct merging of features extracted from distinct scales of the input image. Such approach can be susceptible to the propagation of irrelevant and degraded information. To avoid this, we have introduced Adaptive Gated Feature Fusion Blocks (AGFF) to merge and disseminate features from both original scale stream and multi-scale stream. AGFF aims to offer a two-tiered attention mechanism that enhances selectivity during feature propagation. AGFF provides i) scale-level feature attention and ii) intra-feature attention. To provide relevant importance to features from either original or multi-scale transformer stream, we utilize mixup strategy [43] (originally proposed for data augmentation) weighing the features. We make the weighing parameter trainable using a trainable tensor. This parameter is then multiplied with the respective stream features for deciding their importance. Following this, we provide gated attention to the

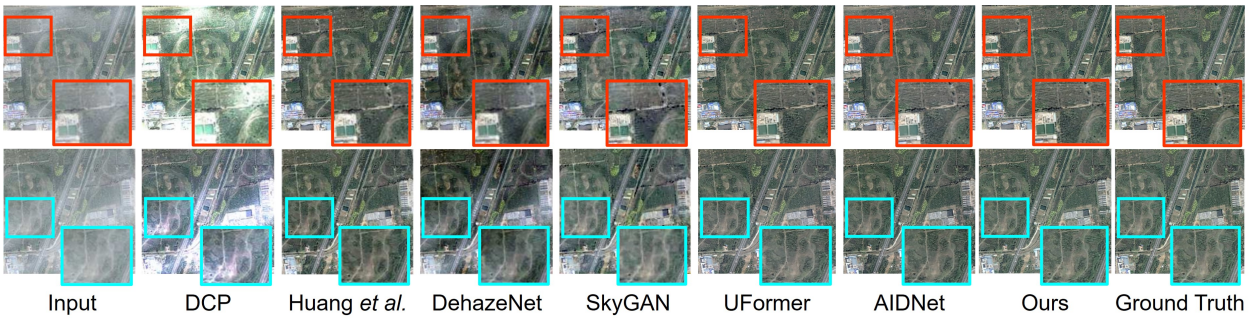

Figure 3. Qualitative results on Sate-1K dataset. The compared methods are: DCP [10], Huang *et al*. [12], DehazeNet [2], SkyGAN [29], UFormer [39], AIDNet [15] and the proposed method (Ours).

Table 1. Quantitative results comparison of the proposed method with existing methods on Sate-1K dataset having density splits: Thin, Moderate and Thick. Here, TS: Task Specific, trained on separate datasets, C: Consolidated, trained in a consolidated manner. **Red** represents best and Blue represents second best performance values.

Methods	Thin Haze		Moderate Haze		Thick Haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP [10]	13.15	0.7246	9.78	0.5735	10.25	0.5850
SAR-Opt-cGAN [8]	20.19	0.8419	21.66	0.7941	19.65	0.7573
DehazeNet [2]	19.75	0.8950	18.12	0.8552	14.33	0.7064
Huang *et al.* [12]	24.16	0.9061	25.31	0.9264	25.07	0.8640
SkyGAN [29]	25.38	0.9248	25.58	0.9035	23.43	0.8925
UFormer [39]	25.79	0.9270	26.11	0.9308	25.15	0.9017
AIDNet [15]	27.68	0.9511	27.03	0.9472	26.72	0.9290
Proposed Method - TS	**31.85**	**0.9750**	**30.53**	**0.9714**	**29.31**	**0.9589**
Proposed Method - C	30.32	0.9723	29.65	0.9677	27.97	0.9413

weighted features to provide intra-feature importance to relevant regions in the features. AGFF can be equated mathematically as:

$$\mathbb{F}_{AGFF} = \langle f_1' * \zeta(f_1'), f_2' * \zeta(f_2') \rangle \tag{5}$$

$$f_1' = \lambda * f_1, f_2' = (1 - \lambda) * f_2 \tag{6}$$

where, f_1 and f_2 are the inputs to AGFF block, where, the f_1 features are obtained from the original scale stream and f_2 are obtained from multi-scale stream. $\lambda = \sigma(\theta)$ is the adaptive weighing parameter obtained with a trainable tensor θ activated with Sigmoid activation $\sigma(\cdot)$.

All the proposed modules contribute towards providing a consolidated solution for aerial hazy image restoration. The effectiveness of each proposed module is elaborated in Sec. 5.

4. Experimental Discussion

4.1. Datasets

i) Sate-1K Dataset [12]: The dataset comprises of pairs of aerial images, namely clean and degraded images, with varying levels of haze density - thin, moderate, and thick. Through data augmentation techniques such as random flipping, 640×3 = 1920 image pairs (degraded and ground-truth) are used for training purpose, containing light, medium, and dense haze, and 45 image pairs are dedicated for testing in each level of haze density.

ii) RICE Dataset [22]: The dataset includes aerial images degraded by haze that cover various types of earth surfaces such as urban scenes, oceans, deserts, mountains, and more. With the help of data augmentation techniques, we have utilized 800 pairs of images for training and 100 pairs for testing purposes.

For consolidated training, we have merged these two datasets for combined learning of the degradations. Hence, a total 2720 pairs of training images are utilized. And the testing is done on testing sets of respective datasets (45 per density split for Sate-1K dataset, and 100 for RICE dataset).

4.2. Training Details

The input images are resized to 256× 256 for training of the proposed network. We have utilized $\mathbb{L}_1$ and perceptual loss [14] (calculated between the restored output and ground truth) in a weighted manner for training of the proposed network. *Detailed equations of these loss functions are provided in the supplementary material.* During training, we utilized the ADAM optimizer with an initial learning rate of 1×10^{-3}, and vary it using cosine annealing strategy. The proposed network is implemented using the PyTorch library and trained on NVIDIA-DGX sta-

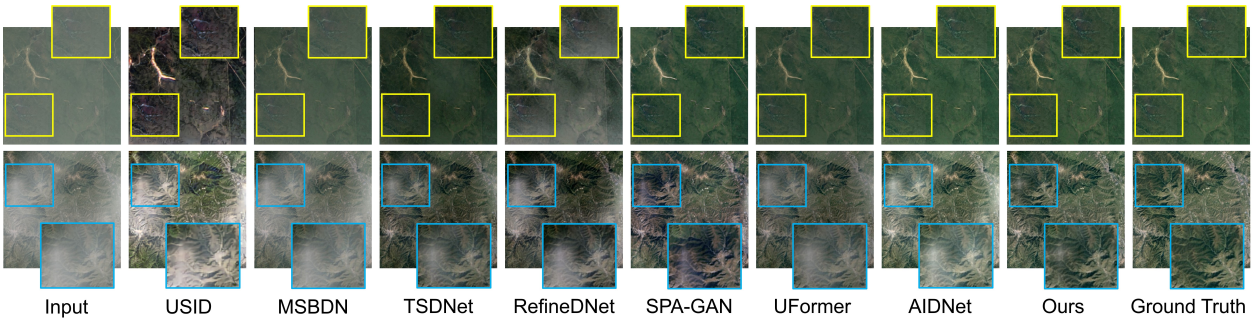

Figure 4. Qualitative results comparison on RICE dataset. The compared methods are: USID [20], MSBDN [5], TSDNet [26], RefineDNet [45], SPA-GAN [30], UFormer [39], AIDNet [15] and the proposed method (Ours).

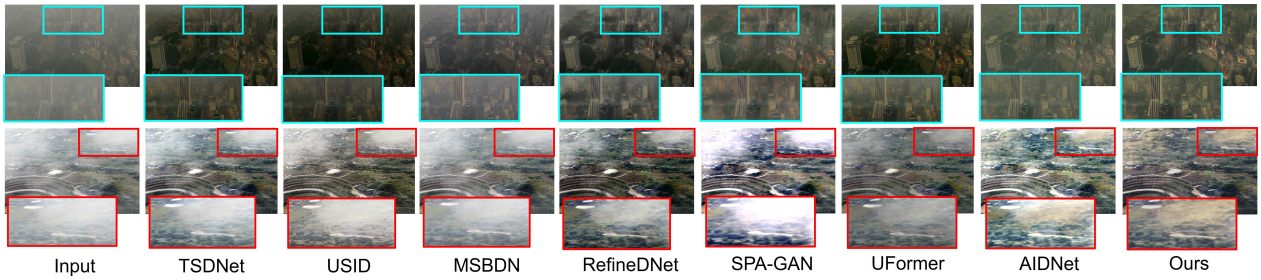

Figure 5. Qualitative results comparison of the proposed method (Ours) with existing state-of-the-art methods (USID [20], MSBDN [5], TSDNet [26], RefineDNet [45], SPA-GAN [30], UFormer [39], AIDNet [15]) on realworld aerial images.

Table 2. Quantitative results comparison of the proposed method with existing methods on RICE dataset.

Methods	PSNR	SSIM
MSBDN [5]	24.58	0.8341
RDNet [45]	28.81	0.9193
USID [20]	26.77	0.8733
TSDNet [26]	29.07	0.9274
SPA-GAN [30]	30.23	0.9540
UFormer [39]	30.17	0.9531
AIDNet [15]	33.79	0.9703
Proposed Method - TS	**36.33**	**0.9881**
Proposed Method - C	35.21	0.9815

Table 3. Computational complexity analysis in terms of number of parameters, GFLOPs and run-time (*with image size 256 × 256*).

Methods	# Par (M)	GFLOPs	Run-time (sec/image)
USID [20]	3.70	160	0.15
MSBDN [5]	31.35	83	0.12
RDNet [45]	65.13	154	0.20
UFormer [39]	50.88	89	0.16
AIDNet [15]	20.32	98	0.13
Proposed Method	**1.49**	**12.1**	**0.07**

tion equipped with an Intel Xeon E5-2698 processor and NVIDIA Tesla V100 16 GB GPU for 400K iterations, taking approximately 25 GPU hours.

4.3. Quantitative Analysis

In this section, we evaluate performance of the proposed method in quantitative manner. The performance of the methods is evaluated in terms of peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). We compare the quantitative results obtained by our method with existing state-of-the-art techniques on the Sate-1K dataset in Table 1, and on the RICE dataset in Table 2. For fair comparison, the quantitative and qualitative results are provided after re-training the existing methods on RICE and Sate1K datasets. As seen from the results, the proposed network performs better than the existing state-of-the-art methods on both the datasets *viz.* having different densities such as light, medium, and dense. It is noteworthy that the proposed method is distinct from existing methods which typically undergo separate training procedures and have different checkpoints for each dataset and their respective splits. The proposed approach, on the other hand, is designed to restore aerial images in a consolidated manner, signifying that it requires only one checkpoint to effectively restore images in density-invariant manner, ensuring its applicability in practical scenarios.

4.4. Qualitative Analysis

In this section, we evaluate the performance of proposed method in terms of visual results. Figure 3 compares the qualitative results obtained on the Sate1K dataset, while Figure 4 presents the results obtained on the RICE dataset. Further, we evaluate the proposed method on real-world hazy images and display the results in Figure 5. As seen from the results, the proposed method stands out to be more effective in restoring images with better color and detail

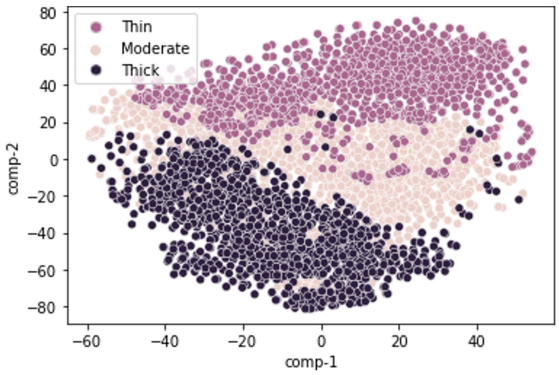

Figure 6. t-SNE visualization of the features extracted from DEB after providing inputs with various haze densities.

preservation. *More qualitative results are provided in the supplementary material.*

4.5. Computational Complexity Analysis

As a preprocessing step, it is crucial for any image restoration module to have a low computational complexity and fast inference speed. Table 3 presents a comparison of the proposed method with prevailing methods in terms of computational complexity. From the values presented in the table, the proposed method has approximately half the number of trainable parameters compared to the prevailing methods. Moreover, the proposed method requires only about one-seventh of the GFLOPs (floating-point operations per second). Further, the proposed network has a comparatively lower run-time requirement, only about half of the runtime of existing approach [5]. The computational complexity of normal convolutional layers is in the order of $\mathcal{O}(C^2K^2HW)$, where, C is the number of channels, K is the kernel size of the convolution filter, H, W are height and width of the feature map. The methods USID [20], MSBDN [5], and RDNet [45] contain convolution filters with number of channels as much as upto 256. Whereas the proposed Adaptive Gated Feature Fusion (AGFF) blocks, and inter-sharing of multi-stream features, allows the network to learn diverse features, and minimizes the need of expanding the channels throughout the network (which are upto 128 in the proposed network (*please see architecture details section in the supplementary material*). As for the transformer based methods, the computational complexity of the key-query dot-product in UFormer [39] and AIDNet [15] grows quadratically with the window-size (M), *i.e.,* $\mathcal{O}(M^4)$, whereas, inspired from Restormer [42], the computational complexity of the proposed method grows quadratically with the number of channels, *i.e.,* $\mathcal{O}(C^2)$, where, $C < M$. These configurations result in the lesser computational complexity and inference time of the proposed method than existing methods. From this, it is verified that the proposed method can achieve better results in efficient (faster) manner.

Table 4. Analysis on effectiveness of Query Modulation (QM) in terms of PSNR and SSIM on Sate1K - **Thin / Moderate/ Thick** dataset.

Setting	PSNR	SSIM
w/o QM	28.09/ 27.34/ 24.76	0.9483/ 0.9451/ 0.9083
with QM (Additive)	28.85/ 28.08/ 26.22	0.9634/ 0.9502/ 0.9295
with QM (Ours)	**30.32/ 29.65/ 27.97**	**0.9723/ 0.9677/ 0.9413**

Table 5. Evaluation of various feed-forward block settings.

Feed-Forward Block Setting	PSNR	SSIM
Includes FPN [23]	26.59	0.9421
Includes only $\wp(C_1^d, C_1^d)$	25.87	0.9311
Includes only $\wp(C_3^d, C_3^d)$	27.21	0.9516
Includes only $\wp(C_5^d, C_5^d)$	26.73	0.9467
Includes only $\wp(C_3^d, C_1^d)$	28.26	0.9599
Includes only $\wp(C_5^d, C_1^d)$	27.91	0.9562
Ours (Includes $\langle \wp(C_3^d, C_1^d), \wp(C_5^d, C_1^d) \rangle$)	**29.65**	**0.9677**

Table 6. Evaluation of different feature fusion approaches.

Feature Fusion Approach	PSNR	SSIM
No Fusion	25.09	0.9287
Concatenation	26.55	0.9431
Gated Feature Fusion (*without* λ)	26.95	0.9489
Adaptive Feature Fusion (*without gating*)	28.15	0.9513
Adaptive Gated Feature Fusion	**29.65**	**0.9677**

Table 7. Evaluation of influence of loss functions.

Training Losses	PSNR	SSIM
Only $\mathbb{L}_1$ Loss	28.01	0.9511
Only Perceptual Loss	27.86	0.9485
$\mathbb{L}_1$ + Perceptual Loss	**29.65**	**0.9677**

5. Ablation Study

In this section, we analyse the influence of every key element and design choice in the formulation of the proposed method. All the experimental settings follow the same training procedures explained in Sec. 4.

1) Effectiveness of the proposed query modulation: For this, we train the network with different settings of the density guided queries and report the corresponding results in Table 4. As seen from the results, incorporation of query modulation provides favorable gain of 2.31 dB PSNR. This can be justified with the feature discriminative ability of the DEB. Figure 6 illustrates the learned variations of haze densities by DEB. The clusters in the Figure 6 show the ability of DEB for learning the haze density specific features effectively. Further, the qualitative results provided in Figure 7 obtained with and without inclusion of query modulation proves effectiveness of DAQM in learning various densities of haze. **2) Influence of variations in the Feed-Forward block**: Upon inspection of Table 5 which provides performance of various settings in Feed-Forward blocks, and Figure 7 which provides the result

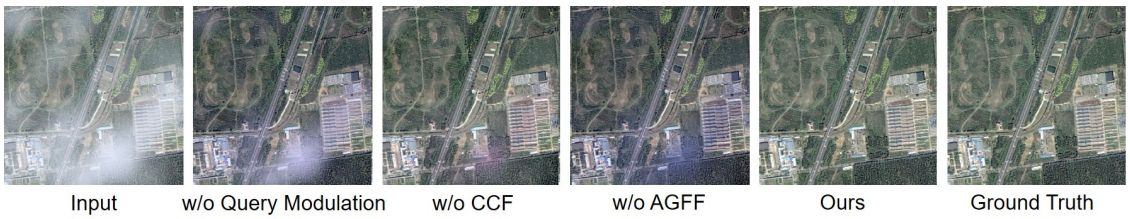

Figure 7. Qualitative evaluation of the influence of the proposed block. "w/o" refers to exclusion of a particular block from the network.

Table 8. Quantitative analysis on UIEB dataset for underwater image enhancement.

Methods	PSNR	SSIM
CLUIE [21]	20.37	0.89
TACL [25]	22.30	0.88
Proposed Method	**24.53**	**0.92**

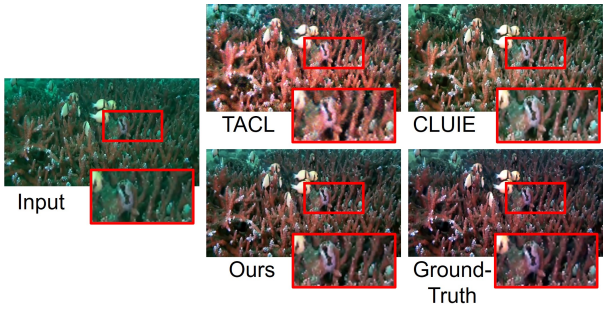

Figure 8. Qualitative results comparison on UIEB dataset.

when CCF is not used in the proposed network, it is verified that utilizing different kernel sizes in a collaborative manner (proposed CCF block) provide better performance than other (non-collaborative) settings. **3) Performance evaluation of other fusion mechanisms vs AGFF**: The results of this study are reported quantitatively in Table 6. Further, the visual results obtained with and without incorporation of AGFF are provided in Figure 7. As seen from the results, the proposed AGFF is able to produce better perceptual quality. *The architectural diagrams of the networks implemented for the ablation studies are displayed in the supplementary material.* **4) Loss Functions**: We notice that the performance of the proposed network enhances when trained with combined loss functions ($\mathbb{L}_1$ + Perceptual Loss) than using each of these loss functions separately. Table 7 provides the quantitative analysis for the same.

6. Applicability

Until now, we discussed and analysed the proposed method on aerial image dehazing. In this section, we verify the applicability of the proposed network for another widely utilized restoration task *i.e.* underwater image enhancement. We train the proposed transformer network on UIEB dataset [19] containing 800 pairs (further augmented to 2400 via random flipping) for training and 90 images for testing. The quantitative results in comparison with existing

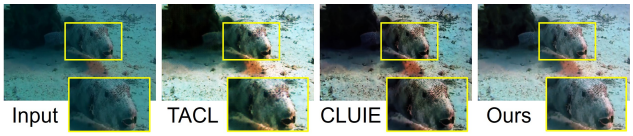

Figure 9. Qualitative results comparison for real-world underwater image enhancement.

state-of-the-art methods for underwater image enhancement are provided in Table 8. The qualitative results comparison on UIEB dataset is provided in Figure 8. We further compare the qualitative results on challenging real-world underwater images in Figure 9. As seen from the results analysis, the proposed method shows superior performance both qualitatively and quantitatively, hence proving the potential applicability of the proposed method for other image restoration tasks. *More qualitative results are provided in the supplementary material.*

7. Conclusion

In this paper, we proposed a lightweight and consolidated approach for aerial image dehazing, which has advantages of generalizability (single trained model) and lower computational complexity. To achieve this, a density aware query modulated block is proposed for learning restoration of aerial images with various haze densities in a consolidated manner. Further, a cross collaborative feed-forward network is proposed for extracting structures with varying sizes within an image using depthwise convolutions with varying kernel sizes. Lastly, an adaptive gated feature fusion block is introduced for providing dual (scale-level and intra-feature) attention while fusing and propagating the features in the network. Substantial experiments on synthetic as well as real-world images, along with extensive ablation studies demonstrated the effectiveness of proposed method for consolidated aerial hazy image restoration. We further analysed the proposed approach for the task of underwater image enhancement, which shows applicability of the proposed approach for other image restoration task.

Acknowledgement

Ashutosh Kulkarni is supported by TCS Research Scholar Program (Cycle 17), and Subrahmanyam Murala is supported by DST-SERB, India, under Grant CRG/2022/006876. The authors would also like to thank all the members of CVPR Lab for their support on this work.

References

[1] Codruta O Ancuti, Cosmin Ancuti, Chris Hermans, and Philippe Bekaert. A fast semi-inverse approach to detect and remove the haze from a single image. In *Asian Conference on Computer Vision*, pages 501–514. Springer, 2010. 2

[2] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 2, 5

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[4] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7180–7189, 2021. 2

[5] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2157–2167, 2020. 2, 6, 7

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[7] Raanan Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):1–9, 2008. 2

[8] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729. IEEE, 2018. 3, 5

[9] Jianhua Guo, Jingyu Yang, Huanjing Yue, Hai Tan, Chunping Hou, and Kun Li. Rsdehazenet: Dehazing network with channel refinement for multispectral remote sensing images. *IEEE Transactions on geoscience and remote sensing*, 59(3):2535–2549, 2020. 3

[10] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1, 2, 3, 5

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[12] Binghui Huang, Li Zhi, Chao Yang, Fuchun Sun, and Yixu Song. Single satellite optical imagery dehazing using sar image prior based on conditional generative adversarial networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1806–1813, 2020. 2, 3, 5

[13] Tongyao Jia, Jiafeng Li, Li Zhuo, and Guoqiang Li. Effective meta-attention dehazing networks for vision-based outdoor industrial systems. *IEEE Transactions on Industrial Informatics*, 18(3):1511–1520, 2022. 2

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[15] Ashutosh Kulkarni and Subrahmanyam Murala. Aerial image dehazing with attentive deformable transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6305–6314, 2023. 2, 3, 4, 5, 6, 7

[16] Ashutosh Kulkarni, Prashant W Patil, and Subrahmanyam Murala. Progressive subtractive recurrent lightweight network for video deraining. *IEEE Signal Processing Letters*, 29:229–233, 2021. 2

[17] Ashutosh Kulkarni, Prashant W Patil, Subrahmanyam Murala, and Sunil Gupta. Unified multi-weather visibility restoration. *IEEE Transactions on Multimedia*, 2022. 2

[18] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017. 2

[19] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2019. 8

[20] Jiafeng Li, Yaopeng Li, Li Zhuo, Lingyan Kuang, and Tianjian Yu. Usid-net: Unsupervised single image dehazing network via disentangled representations. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 2, 6, 7

[21] Kunqian Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. Beyond single reference for training: underwater image enhancement via comparative learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 8

[22] Daoyu Lin, Guangluan Xu, Xiaoke Wang, Yang Wang, Xian Sun, and Kun Fu. A remote sensing image dataset for cloud removal. *arXiv preprint arXiv:1901.00600*, 2019. 5

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[24] Changbing Liu, Jianbo Hu, Yu Lin, Shihong Wu, and Wei Huang. Haze detection, perfection and removal for high spatial resolution satellite imagery. *International Journal of Remote Sensing*, 32(23):8685–8697, 2011. 3

[25] Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31:4922–4936, 2022. 8

[26] Ryan Wen Liu, Yu Guo, Yuxu Lu, Kwok Tai Chui, and Brij B. Gupta. Deep network-enabled haze visibility enhancement for visual iot-driven intelligent transportation systems. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2022. 2, 6

[27] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7314–7323, 2019. 2

[28] Jiao Long, Zhenwei Shi, Wei Tang, and Changshui Zhang. Single remote sensing image dehazing. *IEEE Geoscience and Remote Sensing Letters*, 11(1):59–63, 2013. 1, 3

[29] Aditya Mehta, Harsh Sinha, Murari Mandal, and Pratik Narang. Domain-aware unsupervised hyperspectral reconstruction for aerial image dehazing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 413–422, 2021. 2, 3, 5

[30] Heng Pan. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. *arXiv preprint arXiv:2009.13015*, 2020. 2, 3, 6

[31] Prashant Patil, Jasdeep Singh, Praful Hambarde, Ashutosh Kulkarni, Sachin Chaudhary, and Subrahmanyam Murala. Robust unseen video understanding for various surveillance environments. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2022. 1

[32] Prashant W Patil, Akshay Dudhane, Ashutosh Kulkarni, Subrahmanyam Murala, Anil Balaji Gonde, and Sunil Gupta. An unified recurrent video object segmentation framework for various surveillance environments. *IEEE Transactions on Image Processing*, 30:7889–7902, 2021. 1

[33] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pages 154–169. Springer, 2016. 2

[34] Sebastian Salazar-Colores, Eduardo Cabal-Yepez, Juan M Ramos-Arreguin, Guillermo Botella, Luis M Ledesma-Carrillo, and Sergio Ledesma. A fast image dehazing algorithm using morphological reconstruction. *IEEE Transactions on Image Processing*, 28(5):2357–2366, 2018. 2

[35] Sanchayan Santra, Ranjan Mondal, and Bhabatosh Chanda. Learning a patch quality comparator for single image dehazing. *IEEE Transactions on Image Processing*, 27(9):4598–4607, 2018. 2

[36] Robby T Tan. Visibility in bad weather from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[37] Ketan Tang, Jianchao Yang, and Jue Wang. Investigating haze-relevant features in a learning framework for image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2995–3000, 2014. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[39] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, June 2022. 3, 4, 5, 6, 7

[40] Dong Yang and Jian Sun. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In *Proceedings of the european conference on computer vision (ECCV)*, pages 702–717, 2018. 2

[41] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, June 2022. 4, 7

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4

[44] Ying Zhang and Bert Guindon. Quantitative assessment of a haze suppression methodology for satellite imagery: Effect on land cover classification performance. *IEEE Transactions on Geoscience and Remote Sensing*, 41(5):1082–1089, 2003. 1, 3

[45] Shiyu Zhao, Lin Zhang, Ying Shen, and Yicong Zhou. Refinednet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing*, 30:3391–3404, 2021. 2, 6, 7

[46] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 4

[47] Qingsong Zhu, Jiaming Mai, and Ling Shao. Single image dehazing using color attenuation prior. In *BMVC*. Citeseer, 2014. 2