Aerial Image Dehazing with Attentive Deformable Transformers

Ashutosh Kulkarni and Subrahmanyam Murala
CVPR Lab, Indian Institute of Technology Ropar, INDIA
{ashutosh.20eez0008, subbumurala}@iitrpr.ac.in

Abstract

Aerial imagery is widely utilized in visual data dependent applications such as military surveillance, earthquake assessment, etc. For these applications, minute texture in the aerial image are essential as any disturbance can cause inaccurate prediction. However, atmospheric haze severely reduces the visibility of the scene to be analysed, and hence takes a toll on accuracy of higher level applications. Existing methods either utilize additional prior while training, or produce sub-optimal outputs on different densities of haze degradation, due to absence of local and global dependencies in the extracted features. Therefore, it is essential to have a texture preserving algorithm for aerial image dehazing. In light of this, we propose a work that introduces a novel deformable multi-head attention with spatially attentive offset extraction based solution for aerial image dehazing. Here, the deformable multi-head attention is introduced to reconstruct fine level texture in the restored image. We also introduce spatially attentive offset extractor in the deformable convolution for focusing on relevant contextual information. Further, edge boosting skip connections are proposed for effectively passing edge features from shallow layers to deeper layers of the network. Thorough experimentation on synthetic as well as real-world data, along with extensive ablation study, demonstrate that the proposed method outperforms the prevailing works on aerial image dehazing. The code is provided at `https://github.com/AshutoshKulkarni4998/AIDTransformer`.

1. Introduction

Steep development in aerial imaging technology in the recent past has lead to ameliorated quality of aerial images which can be applied to many fields, *e.g.* building extraction [9], earthquake damage assessment [3], and image decomposition [44]. Significant performance of these applications mainly depends on the clean aerial data. Taking into account the fact that aerial images are captured from a long distance, they are susceptible to low visibility, color shifts, and blurriness as a result of changes in atmospheric conditions and presence of the clouds or fog. Due to diminished visibility in haze or cloud affected scene, monitoring and assessment of the situations such as disaster management becomes a challenging task. Hence, there is a dire need for a robust aerial image visibility enhancement method.

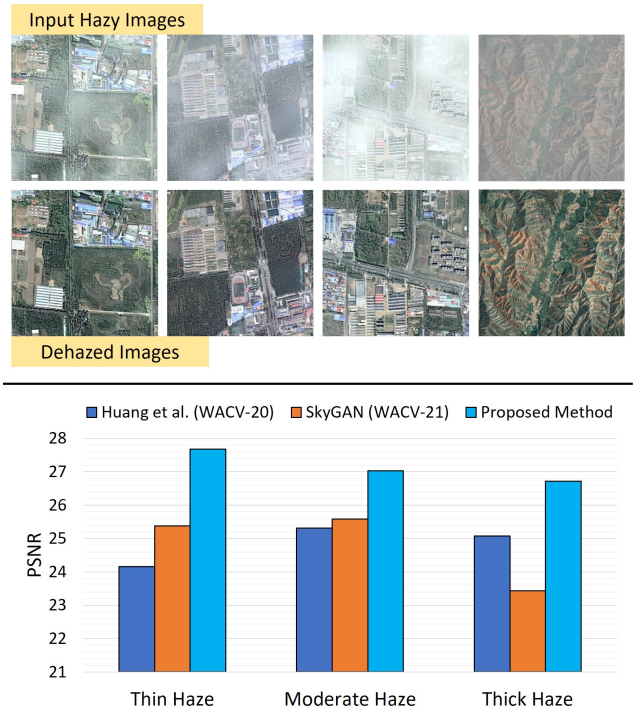

Figure 1: Sample results for aerial image dehazing (*top row*) and performance comparison (*bottom row*) of the proposed method with existing state-of-the-art methods (Huang *et al.*[17] and SkyGAN [28]) on Sate1K dataset. As seen from the results, the proposed method is effective in both quantitative and qualitative evaluations.

Working towards this, existing method [48] attempted usage of a correction technique assisted by finding the correlation between the low and high-frequency color bands. Further, Liu *et al.* [24] used virtual cloud point method for haze removal. Long *et al.* [27] utilized dark channel prior (DCP) proposed by He *et al.* [16] for removing haze from natural scene images.

The emergence of deep learning [19, 20, 29, 31, 32, 33, 34] further promoted the research towards aerial image restoration. With the generalization capability of convolutional neural networks (CNNs), authors have proposed methods which include conditional generative adversarial networks (cGAN) [17], unsupervised learning [28], channel refinement [15], *etc.* for aerial image dehazing. The rapid growth of transformers due to their ability to capture global dependencies in an image has resulted in different architectures [43, 47] for image restoration tasks such as deblurring, denoising, deraining, *etc.* Yet, the transformers have not been explored to deal with haze degraded aerial images.

Huang *et al.* [17] utilized synthetic aperture radar (SAR) prior for training the network. In contrast to this, the proposed method only utilizes RGB haze and haze free aerial images while training the proposed network, hence avoiding additional requirement of data priors. The existing methods for aerial [17, 28] and outdoor [10, 22, 49] image dehazing do not consider geometric adaptability for feature extraction which leads to improper restoration of crucial structures in the image. In contrast to this, the proposed method achieves geometric adaptability with the proposed novel space aware deformable convolution block. Further, traditional deformable convolutions [8] may extract irrelevant features because the offsets may extend beyond relevant regions. To avoid this, we introduce a novel spatially attentive offset extractor which leads to spatially relevant feature extraction. Edge boosting skip connections are proposed (instead of simple skip connections) to preserve edge information in the restored image. *Conceptual differences between the proposed method and existing methods are discussed in the supplementary material*. The main contributions of this work are summarized as:

- We introduce an end-to-end trainable attentive transformer network for aerial image de-hazing. In that, we leverage space aware deformable convolution based multi-head self attention to preserve crucial texture while de-hazing an image.
- We propose spatially attentive offset extractor for extracting relevant spatial information from the features.
- We propose edge boosting skip connections for effectively passing edge features from shallow layers to deeper layers of the network.

Extensive experiments on various synthetic datasets and real-world images demonstrate that the proposed method outperforms the existing state-of-the-art methods on all the evaluation metrics. Sample results of the proposed method are provided in Figure 1.

2. Related Works

2.1. Prevailing Methods for Image De-hazing

Initial attempts were directed towards image de-hazing using hand-crafted priors [1, 12, 16, 40, 41, 51]. He *et al.* [16] proposed a baseline haze relevant prior to get the image coarse-level depth information for de-hazing. However, it fails in sky regions and exhibits halo effect near complicated edge structures. Salazar-Colores *et al.* [36] combined DCP with mathematical morphology operations, *e.g.*, erosion and dilation, to compute transmission maps efficiently. Researchers [4, 35, 37, 45, 46, 10, 49, 21, 26] have been developing CNNs to calculate the transmission map of a scene followed by atmospheric scattering models to reconstruct the haze-free image for the past decade. Cai *et al.* [4] have proposed deep network to estimate the transmission map followed by atmospheric scattering model to recover the haze-free image. A boosted decoder was proposed by Dong *et al.* [10], where, only reconstruction error calculated using ground truth is used as supervision for gradually obtaining the haze-free image. Zhao *et al.* [49] proposed a weakly supervised two stage framework which utilizes unpaired adversarial learning. Recently, Jia *et al.*[18] proposed a meta-attention based network for restoration of hazy images. Liu *et al.* [25] proposed a multi-branch feature extraction based method for integrating all characteristic information and reconstructing the haze-free image. Chen *et al.* [7] proposed generalization of a pre-trained network on synthetic data to adapt on real-world images. Li *et al.* [22] proposed an unsupervised learning based approach with compact multi-scale feature attention and multi-frequency representations. These methods have been mainly applied on hazy images taken at surface level.

Particularly for aerial image de-hazing, various methods have been proposed. Zhang *et al.* [48] found correlation between low and high color bands using correction technique. Further, Liu *et al.* [24] used virtual cloud point method for haze removal. Long *et al.* [27] utilized DCP proposed by He *et al.* [16] for removing haze from natural scene images. Guo *et al.* [15] availed residual learning strategies for fast convergence of the network along channel attention modules to achieve strong channel correlation. In [30], a model which focuses on the cloudy area with local-to-global spatial attention is proposed for cloud and haze removal. With the fusion of SAR information and multi-spectral image data, Grohnfeldt *et al.* [13] proposed a cGAN for cloud removal. Huang *et al.* [17] utilized information from both RGB and SAR prior and proposed a dilated convolution based generative adversarial network. Mehta *et al.* [28] put forward a GAN framework named SkyGAN which incorporates hyper-spectral images (HSI) guidance in an image-to-image translation network for aerial image de-hazing. Although the existing de-hazing approaches have shown some

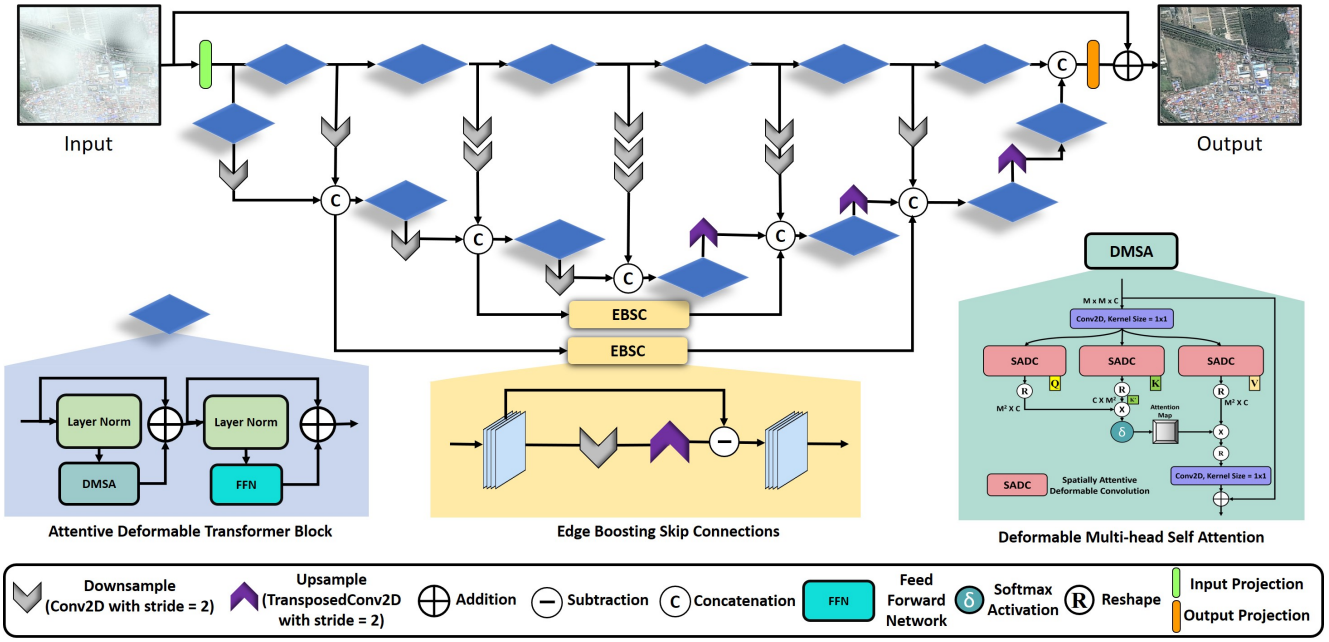

Figure 2: Architectural illustration of the proposed network for aerial image de-hazing. *The input is cropped into patches and passed through input projection which contains a convolution layer. These features are passed through series of proposed attentive deformable transformer blocks. Features from the upper part of the architecture are downscaled and passed to the lower part of the architecture. After processing through all the layers, output residual is obtained from output projection layer involving a 3 channel convolution filter and finally, output image is obtained with addition of the residual with the input.*

significant improvements, they lack in consideration of simultaneous local and global dependencies required for robust de-hazing, which we address in the proposed work.

2.2. Transformers for Low Level Computer Vision

Due to the proven superiority of the Transformers over CNNs in capturing long-range dependencies, they have been vastly used for several applications. Initial work for computer vision using transformers was done in the form of Vision Transformers (ViT) [11] for visual recognition. It uses flattened patches of images for training the Transformer network. The transformers are also utilized for low-level vision applications. Using the image processing transformer, [6] illustrated how pre-training on large datasets can lead to improved performance for low-level applications. Uformer [43] utilized a U-Net like structure using transformers for image restoration problems *i.e.,* deraining, deblurring and denoising. To the best of our knowledge, the proposed approach is the first transformer based approach specifically designed for aerial image de-hazing.

3. Proposed Method

This section reveals several contributing elements involved in the proposed network for aerial image de-hazing.

3.1. Overview

As aerial image de-hazing is a texture susceptive task, it is essential to have processing modules which preserve textural information. Existing methods for aerial image de-hazing [17, 28, 30] produce sub-optimal outputs due to lack of local and global dependencies in the extracted feature maps. Furthermore, the way of conveying features from encoder layers to decoder layers determines the robustness of the output image. Taking this into account, we propose deformable attention based transformer for aerial image de-hazing. We induce spatially attentive offset extraction in the deformable attentive transformer block to extract relevant spatial features crucial for effective de-hazing. Further, we provide edge boosting skip connections in the network to propagate significant edge features. All of these core components are explained in the proceeding subsections.

3.2. Attentive Deformable Transformer Block

Compared to traditional CNNs, transformers have proven their superiority on various tasks such as restoration [43], segmentation [50], object detection [39], *etc.* This performance is derived due to the capability of transformers to capture long range dependencies with the help of multi-head self attention. Furthermore, deformable convolutions have shown superior performance because of their ability to adapt to the geometric variations of the objects. Therefore,

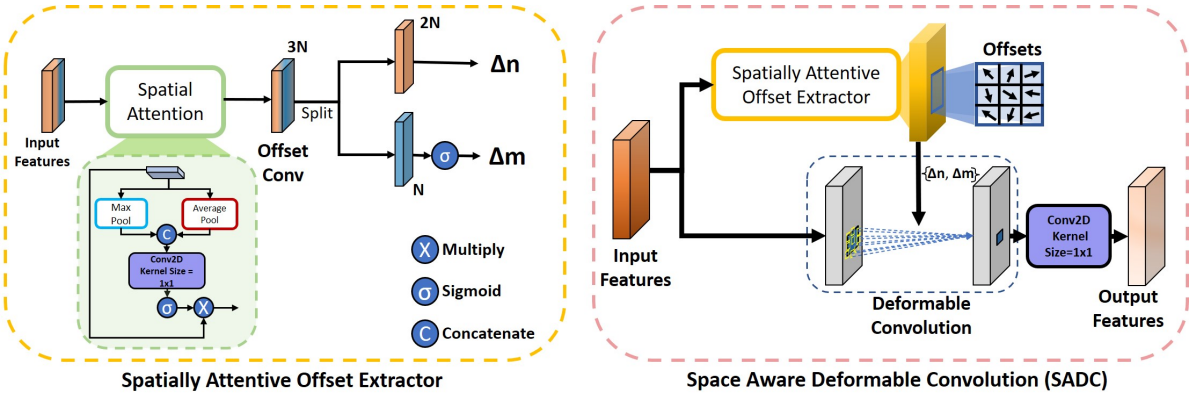

Figure 3: Detailed illustration of the proposed spatially attentive offset extractor and space aware deformable convolution.

in contradiction to prevailing transformer methods [42, 43], we use space aware deformable convolution as a feature extractor for queries (**Q**), keys (**K**) and values (**V**). Collectively, we call the space aware deformable convolution based multi-head self attention as *deformable multi-head self attention* (DMSA) which can be represented as:

$$DMSA = X_{in} + Conv_{1\times1}\left(\delta\left(\mathbf{QK^T}\right)\cdot\mathbf{V}\right) \quad (1)$$

Here, $\delta(\cdot)$ represents Softmax activation layer. The queries, keys and values are obtained as:

$$\mathbf{Q, K, V} = SADC_{3\times3}(Conv_{1\times1}(T_{Norm})) \quad (2)$$

where, $SADC_{3\times3}$ is space aware deformable convolution (*explained in Sec.3.3*) with kernel size 3×3, $Conv_{1\times1}$ is convolution with size 1×1 and T_{Norm} are input features after layer normalization. Further, the process flow of feed forward network (FFN) is: `Conv1 -> Reshape -> DepthwiseConv`$_{3\times3}$`-> Flatten ->Conv2-> Add(Conv1,Conv2)`. Here, `Conv1` and `Conv2` are convolution filters with kernel size 1×1. With this implementation, we provide the network with enough spatially rich contextual information due to double provision of spatial attention in the offset extraction for deformable convolution and long range capturing ability of multi-head self attention. In summary, the attentive deformable transformer block can capture both, long range (global) dependencies through multi-head self attention and feature dependent local contextual information through space aware deformable convolutions.

3.3. Spatially Attentive Offset Extractor and Space Aware Deformable Convolution

The deformable convolutions alone have proven to adapt with the geometric variations in the input features. However, the offsets may extend beyond relevant regions. This may cause irrelevant feature propagation [52] resulting in partially restored image. To avoid this, we introduce spatially attentive offset generation module called as spatially attentive offset extractor. In that, the offsets and modulation values are extracted from the same offset convolution, with spatially attentive features as the input (*please see Figure 3*). This makes the deformable convolutions focus on relevant image regions by providing texture-relevant offsets. We collectively call such setting as space aware deformable convolution (SADC). The SADC can be explained mathematically as:

$$SADC\left(X_n\right) = \sum_{i=1}^{N} DefConv_{3\times3}\left(X_{n+n_i+\Delta n_i}\right)\cdot\Delta m_{n_i} \quad (3)$$

where, N is sampling location in a 3×3 convolution grid. $DefConv_{3\times3}(\cdot)$ is deformable convolution with kernel size 3×3, n is location in the feature, Δn are the offsets extracted from spatially attentive offset extractor and Δm are the modulator scalars extracted from the spatially attentive offset extractor block and $n_i \in \{(-1,-1),(-1,0)\ldots(1,1)\}$. Such practice of spatially attentive offset extraction leads to adept texture attentive queries, keys and values are passed to the multi-head self-attention and avoids the problem of irrelevant feature propagation faced in prevailing method [8].

3.4. Edge Boosting Skip Connections

Aerial images usually contain edge sensitive regions such as roads, buildings, *etc.* and it is a crucial task to reconstruct these texture after restoring the image. Also, skip connections are well-known for their capability to avoid the problem of vanishing gradients. Previous methods [17, 28, 43] provide simple skip connections from shallow to deeper layers in the network. This may neglect the edge information present in the initial layers of the network. To avoid this, we propose edge boosting skip connections (EBSC) in the network. This is achieved by extracting high frequency edge information through learnable layers. The output of EBSC can be obtained as:

$$O_{EBSC}(X_{in}) = X_{in} - TrConv_{s=2}(Conv_{s=2}(X_{in})) \quad (4)$$

Figure 4: Qualitative results comparison of the proposed method with existing state-of-the-art methods DCP [16], Huang *et al.* [17], DehazeNet [4], SkyGAN [28] and UFormer [43] on Sate1K dataset.

Methods	Publication	Thin Haze		Moderate Haze		Thick Haze	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Original	-	12.77	0.7241	12.58	0.7399	8.58	0.4215
DCP [16]	TPAMI-10	13.15	0.7246	9.78	0.5735	10.25	0.5850
SAR-Opt-cGAN [13]	IGARSS-18	20.19	0.8419	21.66	0.7941	19.65	0.7573
DehazeNet [4]	TIP-18	19.75	0.8950	18.12	0.8552	14.33	0.7064
Huang *et al.* [17]	WACV-20	24.16	0.9061	25.31	0.9264	25.07	0.8640
SkyGAN [28]	WACV-21	25.38	0.9248	25.58	0.9035	23.43	0.8925
UFormer [43]	CVPR-22	25.79	0.9270	26.11	0.9308	25.15	0.9017
Proposed Method	-	**27.68**	**0.9511**	**27.03**	**0.9472**	**26.72**	**0.9290**

Table 1: Quantitative results comparison of the proposed method with existing methods on Sate1K dataset for non-uniform satellite image haze removal.

here, X_{in} represents input to EBSC, $TrConv_s$ and $Conv_s$ represents transposed convolution and convolution respectively having stride = s. These outputs of the EBSC are provided to deeper layers of the proposed network. Adapting such edge boosting skip connections overcomes the limitation of normal skip connections. The effectiveness of all of the explained blocks is scrutinized in ablation study (Section 4.5).

3.5. Loss Functions

Our network is trained in an end-to-end fashion using L1 loss ($\mathbb{L}_1$) between the output and ground truth. To maintain edge consistencies, we have used edge loss ($\mathbb{L}_{Edge}$). Furthermore, we have used a perceptual loss ($\mathbb{L}_P$) that measures the deviation between the features of the predicted output and the ground truth. We extract features from intermediate layers (*viz.* 3^{rd}, 8^{th} and 15^{th}) of a pretrained VGG-16 [38] model for loss calculation. We provide separate weights (λ_{loss}) to the individual loss functions to control their contribution in the overall loss function ($\mathbb{L}_{Total}$) which is formulated as:

$$\mathbb{L}_{Total} = \lambda_{L1}\mathbb{L}_1 + \lambda_{Edge}\mathbb{L}_{Edge} + \lambda_P\mathbb{L}_P \qquad (5)$$

We set the weights as $\lambda_{L1} = 1$, $\lambda_{Edge} = 5$ and $\lambda_P = 10$ which are set empirically. *The detailed equations of loss functions are given in supplementary material.*

4. Experimental Discussion

In this section, we discuss the datasets, training details, comparative analysis and ablation study of the proposed network.

4.1. Datasets

• **Sate1K Dataset [17]**: This dataset contains pairs of clean and degraded aerial images with different (thin, moderate and thick) densities of haze. With data augmentation, we have used a total of 640 pairs (hazy and clean) for training and 45 pairs for testing in each haze density. Degradations present in the images of this dataset mimic non-uniform haze present in the aerial images.

• **RICE Dataset [23]:** This dataset consists of haze degraded aerial images covering different types of earth surfaces *i.e.,* urban scenes, ocean, desert, mountains, *etc*. With data augmentation, we have used 800 pairs of training images and 100 pairs of testing images. The images in this

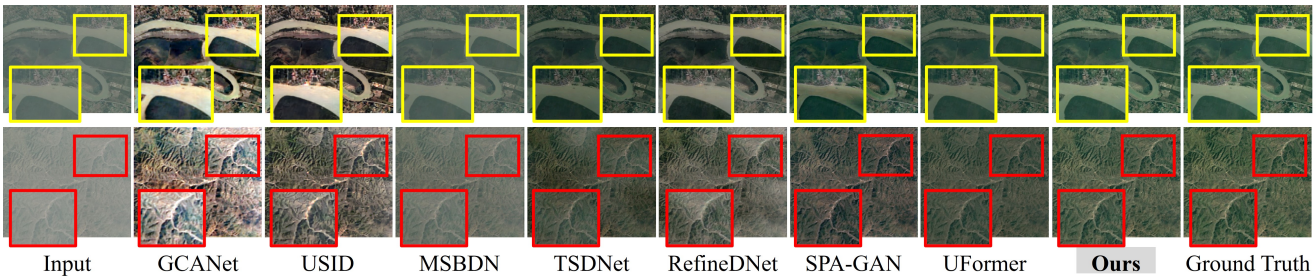

Figure 5: Qualitative results comparison of the proposed method with existing state-of-the-art methods GCANet [5], USID [22], MSBDN [10], TSDNet [25], RDNet [49], SPA-GAN [30] and UFormer [43] on RICE dataset.

Methods	Publication	PSNR	SSIM
GCANet [5]	WACV-19	22.13	0.7917
MSBDN [10]	CVPR-20	24.58	0.8341
RDNet [49]	TIP-21	28.81	0.9193
USID [22]	TMM-22	26.77	0.8733
TSDNet [25]	TII-22	29.07	0.9274
SPA-GAN [30]	-	30.23	0.9540
UFormer [43]	CVPR-22	30.17	0.9531
Proposed Method	-	**33.79**	**0.9703**

Table 2: Quantitative analysis of the proposed method with existing state-of-the-art methods on RICE dataset for uniform haze removal from aerial images.

Methods	# Par (M)	FLOPs ($\times 10^{11}$)	Run-time (sec/image)
USID [22]	3.70	1.60	0.15
MSBDN [10]	31.35	0.83	0.12
RDNet [49]	65.13	1.54	0.20
UFormer [43]	50.88	0.89	0.16
Ours	20.32	0.98	0.13

Table 3: Comparative analysis of the proposed method with existing state-of-the-art methods in terms of number of parameters, FLOPs and run-time (*on image with size 256 × 256*).

dataset have haze degradation which mimics uniform haze present in aerial images.

4.2. Training Details

Traditional transformer used key-query dot-product which results in quadratic growth of with the spatial resolution of the input. To avoid this problem, we use non-overlapped patches of inputs with resolution of M×M = 256×256 for processing through the multi-head self-attention [43]. While training, the ADAM optimizer is used having initial learning rate of 2×10^{-4} varying with cosine annealing strategy. The proposed network is implemented using Pytorch library, and trained on NVIDIA-DGX station with 2.2 GHz processor, Intel Xeon E5-2698, NVIDIA Tesla V100 16 GB GPU for an average of 120 epochs ($\sim$ 22 GPU hours).

4.3. Quantitative Analysis

We evaluate performance of the proposed method quantitatively in terms of average PSNR and SSIM with several existing methods for aerial image de-hazing. Quantitative results on Sate1K dataset are provided in Table 1. The quantitative results on RICE dataset are provided in Table 2. For fair comparison, the quantitative and qualitative values in the manuscript are provided after re-training the existing methods on RICE and Sate1K datasets. Apparently, the proposed method pushes the quantitative scores by a noticeable amount for both datasets. Furthermore, comparison based on computational complexity is provided in Table 3 in terms of number of trainable parameters, floating point operations (FLOPs) and run-time. Although being moderately complex, the proposed method outperforms the existing methods qualitatively and quantitatively.

4.4. Qualitative Analysis

The results of the proposed method are compared in a qualitative manner with existing state-of-the-art methods to scrutinize its improved perceptual quality. The qualitative results on Sate1K dataset are compared in Figure 4 and results on RICE dataset are compared in Figure 5. In the research of image restoration task such as de-hazing, we know that there is significant gap between synthetic and real data. Hence, we evaluate the proposed method on real-world aerial image de-hazing and display the results in Figure 6. As seen from the highlighted regions in the respective comparison figures, the proposed method is able to dehaze the aerial images while preserving more textural content, color balance and perceptual quality. *More qualitative results are provided in the supplementary material.*

4.5. Ablation Study

In this section, we discuss the contribution of each block and loss function in the proposed method. The analysis is carried out on the Sate1K-Moderate Haze dataset.

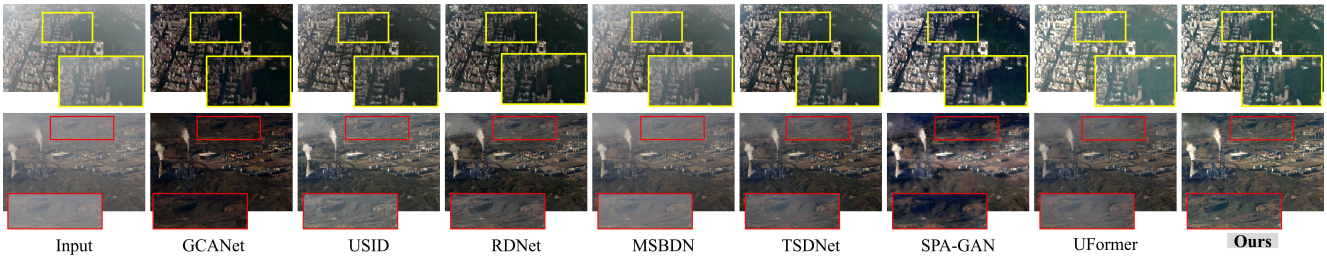

Figure 6: Qualitative results of the proposed method in comparison with existing methods GCANet [5], USID [22], RDNet [49], MSBDN [10], TSDNet [25], SPA-GAN [30] and UFormer [43] on real-world hazy aerial images.

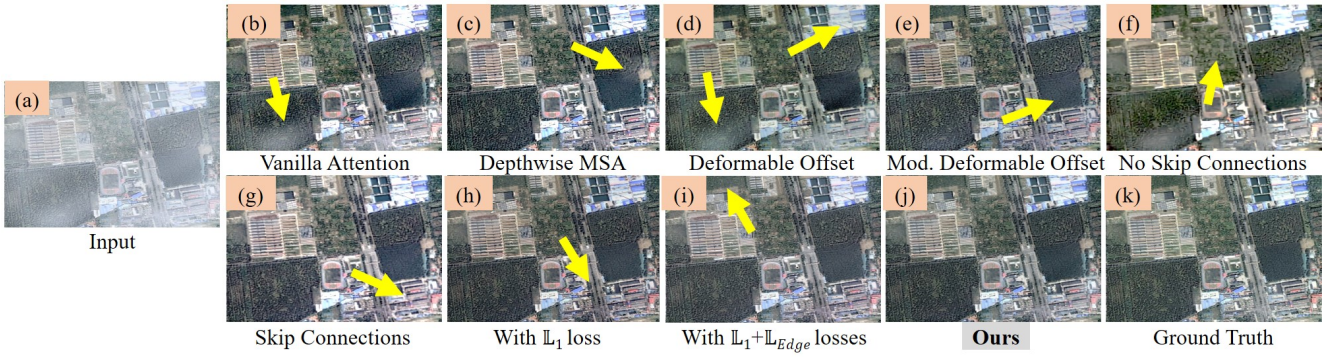

Figure 7: Qualitative results of different network settings mentioned in ablation study. As observed, the proposed method produces plausible outputs as compared to other network settings.

Attention Type	PSNR	SSIM
Vanilla Attention	24.13	0.8957
Depthwise Convolution MSA	25.77	0.9160
Attentive Deformable MSA	**27.03**	**0.9472**

Table 4: Quantitative analysis of the performance of various types of attention modules settings.

Offset Type	PSNR	SSIM
Deformable Offset	25.95	0.9176
Modulated Deformable Offset	26.69	0.9217
Spatially Attentive Deformable Offset	**27.03**	**0.9472**

Table 5: Quantitative analysis of the performance of various offset settings.

Setting	PSNR	SSIM
No Skip Connections	23.89	0.8821
Skip Connections	26.46	0.9258
Edge Boosting Skip Connections	**27.03**	**0.9472**

Table 6: Quantitative analysis of the various types of skip connections.

- **Analysis of the proposed Attentive Deformable Transformer Block:** *The proposed attentive deformable transformer block focuses more on adapting geometrical variations in the input features and hence, preserves more texture in the dehazed image*. This can be scrutinized with

Loss Setting	PSNR	SSIM
$\mathbb{L}_1$	24.90	0.9211
$\mathbb{L}_1 + \mathbb{L}_{Edge}$	25.88	0.9325
$\mathbb{L}_1 + \mathbb{L}_{Edge} + \mathbb{L}_P$	**27.03**	**0.9472**

Table 7: Quantitative analysis of the performance of various loss settings.

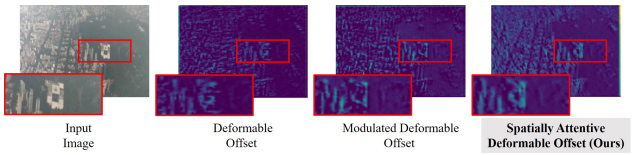

Figure 8: Feature map visualization with different offset extraction methods.

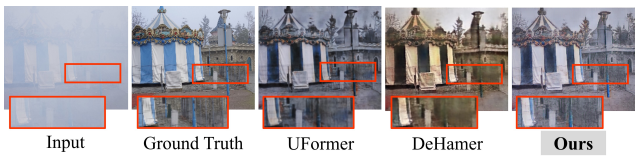

Figure 9: Results of the proposed network on the tasks of outdoor de-hazing compared with UFormer [43] and DeHamer [14] trained on N-Haze dataset [2].

comparative analysis with traditional transformer blocks and the proposed attentive deformable transformer block

Methods	UFormer [43]	DeHamer [14]	Ours
PSNR	19.73	20.66	**21.08**
SSIM	0.6751	0.6844	**0.7013**

Table 8: Quantitative results comparison of the proposed method with existing methods on N-Haze [2] dataset.

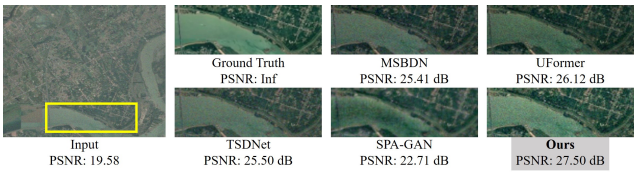

Figure 10: Impact of noise on the performance of the proposed method.

provided quantitatively in Table 4 and visually in Figure 7 *(b), (c) and (j)*. The analysis shows that the proposed transformer block works more effectively than existing transformer blocks.

• **Analysis on influence of the proposed Spatially Attentive Offset Extractor:** *The proposed spatially attentive offset extractor provides offsets which influence the deformable convolution to focus on relevant contextual information.* To evaluate this, we have experimented with different types of offsets for extracting queries, keys and values to be passed in multi-head self attention block. The quantitative result comparison of this experiment is provided in Table 5 and qualitative comparison is provided in Figure 7 *(d), (e) and (j)*. Further, feature visualization of various offset extraction schemes are provided in Figure 8. As seen from the results, the performance of the proposed deformable convolution with spatially attentive offset extraction is better than other types of offset extraction schemes.

• **Analysis on the capability of the Edge Boosting Skip Connections in comparison with traditional skip connections:** *Edge boosting skip connections help in edge enhancement by passing edge features from shallow layers to the deeper layers of the network.* This can be justified by comparison of the performance with different types of feature passing modalities *(i.e. no skip connections and normal skip connections).* The quantitative outcomes of the experimental comparison are provided in Table 6 and Figure 7 *(f), (g) and (j)*. Upon analysis of the values, it is verified that passing of features with edge boosting skip connections perform better than other modalities.

• **Analysis on different loss functions and their impact for training of the proposed network:** Loss functions are used to minimize the discrepancies between the dehazed output and expected ground truth while optimizing the network. We have used a combination of loss functions as stated in Sec. 3.5. We study the effect of these loss functions on the training of the proposed network and provide the results in Table 7 and Figure 7 *(h), (i) and (j)*. From the reported values, it is verified that the combination of various losses performs better. *The detailed information about each network configuration used in ablation study is provided in supplementary material.*

5. Applicability of the Proposed Network

In this paper, we have analysed the proposed network mainly for aerial image de-hazing. For exploring the possible applicability of the proposed network, we carry out experiment on outdoor non-uniform haze removal. The results of the proposed and existing methods after training on N-Haze dataset [2] (*for outdoor non-uniform haze removal*) are provided in Figure 9 and quantitative results are provided in Table 8. As seen from the results, the proposed method can be adopted for various image restoration tasks like general as well aerial image de-hazing.

6. Limitation

Aerial images are susceptible to introduction of noise occasionally. As seen from the results given in Figure 10, the proposed as well as existing methods are unable to remove the noise completely, however, the proposed method performs better visually and quantitatively. The performance may be improved upon training the network on images containing combined hazy and noisy degradations, which can be considered in the future work.

7. Conclusion

In this paper, we proposed a novel attentive deformable transformer network for aerial image de-hazing. In this network, a deformable convolution based multi-head self attention is utilized for preserving crucial textural content in an image. We introduce spatially attentive offset extractor for extracting relevant spatial information from the input features. Along with these, edge boosting skip connections are utilized for passing enhanced edge information from shallow layers to deeper layers in the network. Through comparative quantitative and qualitative analysis on various datasets, we evaluated the superior performance of the proposed network. Extensive ablation study demonstrated the contribution and influence of each block in the introduced network. We also discussed the application of the proposed method for general and aerial image de-hazing.

Acknowledgement

This work was supported by the DST-SERB, India, under Grant ECR/2018/001538.

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] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 444–445, 2020.

[3] Dominik Brunner, Guido Lemoine, and Lorenzo Bruzzone. Earthquake damage assessment of buildings using vhr optical and sar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2403–2420, 2010.

[4] 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.

[5] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019.

[6] 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.

[7] 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.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[9] Mauro Dalla Mura, Jón Atli Benediktsson, Björn Waske, and Lorenzo Bruzzone. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3747–3762, 2010.

[10] 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.

[11] 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.

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

[13] 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.

[14] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5812–5820, 2022.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Ashutosh Kulkarni and Subrahmanyam Murala. Wipernet: A lightweight multi-weather restoration network for enhanced surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[20] Ashutosh Kulkarni, Prashant W. Patil, and Subrahmanyam Murala. Progressive subtractive recurrent lightweight network for video deraining. *IEEE Signal Processing Letters*, 29:229–233, 2022.

[21] 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.

[22] 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.

[23] 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.

[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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Adaptive feature consolidation network for burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1279–1286, 2022.

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

[31] 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.

[32] Prashant W Patil, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Dual-frame spatio-temporal feature modulation for video enhancement. *Pattern Recognition*, page 108822, 2022.

[33] Prashant W. Patil and Subrahmanyam Murala. Msfgnet: A novel compact end-to-end deep network for moving object detection. *IEEE Transactions on Intelligent Transportation Systems*, 20(11):4066–4077, 2019.

[34] Shruti S. Phutke and Subrahmanyam Murala. Fasnet: Feature aggregation and sharing network for image inpainting. *IEEE Signal Processing Letters*, 29:1664–1668, 2022.

[35] 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.

[36] 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.

[37] 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.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Chen Xu, Min Li, and Xiaoli Sun. An edge-preserving variational method for image decomposition. *Chinese Journal of Electronics*, 22(1):109–113, 2013.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.