Research on Defogging Algorithm Based on DRSformer

Peizhou Huang[1]

[1] Software and Systems Engineering, Lappeenranta-Lahti University of Technology, Lappeenranta 53850, Finland
Peizhou.Huang@student.lut.fi

Abstract. Recently, advancements in remote sensing technology have led to the collection of a large volume of remote sensing images. Nevertheless, these images are always influenced by atmospheric phenomena such as haze, which reduce their clarity and quality. Traditional dehazing methods and convolutional neural network (CNN)-based techniques show certain limitations when dealing with the complex and uneven pattern of haze in remote sensing images. The article repurposes DRSformer, a pioneering transformer network created for rain removal tasks, to tackle the problem of image dehazing in this research. DRSformer utilizes Sparse Transformer Blocks (STB) and a Mixture of Experts Feature Compensator (MEFC) to effectively address the challenges posed by nonuniform haze scenarios. Based on experimental results, DRSformer achieved good performance. It surpasses current appeached, achieving superior PSNR and SSIM values under various haze conditions. Furthermore, qualitative assessments indicate that DRSformer significantly improves visual clarity and detail preservation. Looking ahead, the adaptability of DRSformer can be further explored to enhance its performance under other atmospheric conditions and expand its applicability to a wider range of remote sensing image processing tasks.

Keywords: Image Dehazing, Transformer Networks, Atmospheric Correction.

1 Introduction

Advances in remote sensing (RS) technology in recent history have produced a large number of RS images, which are essential for land cover classification and ocean monitoring, among other uses [1]. However, undesirable weather circumstances like haze and fog frequently degrade the quality of these photographs, resulting in low contrast, blurring, and other image degradation [1].

The quality of satellite photos is improved and the genuine landscape taken in hazy conditions is shown when atmospheric haze is removed [2]. Enhancement approaches or prior-based procedures are typically employed in conventional dehazing techniques. Examples are dark channel prior (DCP) [3] and haze-optimized transformation [4], both of which have been applied extensively. While these techniques are usually simple and fast to implement, they can have drawbacks such as artifact introduction and inefficiency in complicated scenarios with non-uniform haze distribution.

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_51

Deep learning-based techniques have shown a great deal of promise recently for dehazing natural photographs. Convolutional neural networks (CNNs) are used in methods like DehazeNet [5], MSFNet [6], and GridDehazeNet [7] to estimate the parameters of atmospheric scattering models or to directly produce clear images. These techniques, however, are less effective for RS photos with more varied and nonuniform haze distributions because they are primarily designed for natural photographs and assume a uniform haze distribution.

Researchers have explored deep learning frameworks specifically designed for dehazing RS images to address these images' unique challenges. Developed methods include the First-Coarse-Then-Fine Network (FCTF-Net) [8], the Dual-Step Cascaded Residual Dense Network (DCRD-Net) [9], and the Dense Attentive Dehazing Network (DADN) [10]. These techniques restore haze-free images through the use of advanced network architectures and have achieved success, though they still face challenges with the highly variable haze distribution in RS images.

Drawing inspiration from the effectiveness of Vision Transformers in handling distant dependencies and representing global information [11], DRSformer introduces a novel Transformer-based approach for RS image dehazing and deraining. In view of the good application of the original model DRSformer in removing rain from images, the article extended this model and experimentally verified the rationality of the conjecture on the state1k data set.

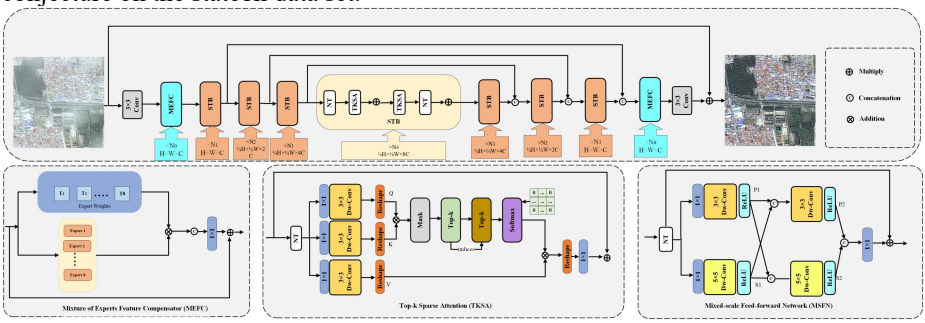

Fig. 1. DRSformer framework and its components. (Picture credit: Original)

2 Methodology

This section outlines the adaptation of the DRSformer architecture, initially designed for image deraining, to address the challenges of image defogging. DRSformer have adopted the architecture as described by Chen et al [12], without modifications, due to its proven effectiveness in managing complex image distortions akin to those encountered in foggy scenarios. DRSformer uses a hierarchical encoder-decoder framework to process foggy images. The core components retained from the original architecture include: Sparse Transformer Block (STB), Mixture of Experts Feature Compensator (MEFC), Loss Function. The specific operations of this architecture (e.g., feed-forward networks and multi-scale processing) are an integral part of the STB and are described in the following sections.

2.1 Overall Framework

According to the Fig. 2, DRSformer employs a layered encoder-decoder architecture specifically tailored for processing fog-affected images. The model takes a foggy image input, $I_{\text{fog}} \in R^{H\times W\times 3}$. In this model, H denote height and Wdenote the width of the image resolution. Using overlapping patch embeddings applied through 3×3 convolutions, the model prepares the input data for feature extraction. In its core, four Sparse Transformer Blocks (STBs) are used to extract spatially variant fog features and manage multi-scale fog representations. The encoders and decoders at each stage are meticulously designed to address specific spatial resolutions and channel dimensions, ensuring precise image detail recovery. Additionally, skip connections are incorporated into the STBs to maintain feature information continuity during various training stages and stabilize the training process. At the initial and final stages of training, DRSformer integrates N_0 Mixture of Experts Feature Compensators (MEFC) for supplementary feature refinement, which ultimately results in high-quality clear output. This hybrid approach allows DRSformer to leverage both adaptive content and the inherent characteristics of foggy images, effectively distinguishing between unwanted fog and the underlying clear background. Experimental results show significant quality improvement due to these design choices. Consequently, the final reconstructed result is obtained using the following formula:

$$I_{\text{defog}} = F(I_{\text{fog}}) + I_{\text{fog}} \#(1)$$

where $F(\cdot)$ is the entire network. The article train the network by minimizing the following lossfunction:

$$L = |I_{\text{defog}} - I_{\text{gt}}|_1 \#(2)$$

In the formula, I_{gt} represents the real image, and $// \cdot //_1$ represents the L1 norm.

2.2 Sparse Transformer Block (STB)

The DRSformer employ the STB, which is specifically de- signed for enhanced feature extraction in image denoising tasks. STB leverages the sparsity emerging in neural networks to optimize the efficiency and effectiveness of feature extraction. Specifically, given the input features from the previous block X_{l-1}, the coding process is defined as follows

$$X_l^{'} = X_{l-1} + \text{TKSA}(\text{LN}(X_{l-1})), \#(3)$$

$$X_l = X_l^{'} + \text{MSFN}\left(\text{LN}(X_l^{'})\right), \#(4)$$

Here, LN denotes layer normalization. This technique enhances the stability of the network and aids in its convergence by standardizing the inputs. The variable $X_l^{'}$

corresponds to the output from the top-k sparse attention (TKSA) module. This module is designed to efficiently manage sparse data by concentrating on the top-k elements. Meanwhile, X_l is derived from the mixed-scale feed-forward network (MSFN). The MSFN enhances feature representation by integrating information across different scales.

Top-k Sparse Attention (TKSA). The Reformer model innovates by enhancing the traditional self-attention mechanism used in Transformers through the introduction of TKSA. In the TKSA mechanism, the initial step involves encoding the channel-wise context using (1×1) convolutions, which help in managing the spatial dimensions and preparing the data for further processing. This is followed by (3×3) depth-wise convolutions, which are crucial for capturing more detailed and localized features within the data.

Additionally, the mechanism computes the similarities between all reshaped queries and keys. During this process, elements that exhibit lower attention weights are masked within the transposed attention matrix M By doing so, the TKSA mechanism ensures that only the most relevant and critical components are retained, effectively minimizing the interference caused by irrelevant information. This adaptive top-k selection strategy enables the most important features to be noticed in the model, thus improving the overall performance. The standard self-attention mechanism, which TKSA replaces, is mathematically described as follows:

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\lambda}\right)\mathrm{V}\#(5)$$

In this context, **(Q)** represents the matrix of queries, **(K)** denotes the matrix of keys, and (**V**) stands for the matrix of values. The parameter $\boldsymbol{\lambda}$, which is optional, is defined as $(\boldsymbol{\lambda} = \sqrt{\mathbf{d}})$. Each of the k new sets of queries **(Q)**, keys **(K)**, and values **(V)** undergoes multi-head attention. This process results in channel dimensional outputs of $(\mathbf{d} = \mathbf{C/k})$.

The resulting outputs are combined and then transformed through a linear projection, yielding the final output of the attention mechanism

To provide an alternative to the standard self-attention mechanism, the TKSA mechanism is introduced. This new mechanism operates as follows:

$$\mathrm{SparseAtt}(Q,K,V) = \mathrm{softmax}\left(\frac{\tau_k(QK^T)}{\lambda}\right)\mathrm{V},\#(6)$$

where the learnable top-k selection operator $\tau_k(\cdot)$ is used:

$$\tau_k(QK^T)_{ij} = \begin{cases} S_{ij} S_{ij} \geq t_i \\ 0\ otherwise \end{cases} \#(7)$$

In this context, (t_i) refers to the (k)-th largest value found in the (j)-th row of (QK^T). This method of adaptive selection is effective in enabling the attention mechanism to transition smoothly from a dense configuration to a sparse one, thereby enhancing the model's efficiency and focus.

Mixed-scale Feed-Forward Network (MSFN). To effectively capture the features of image degradations such as fog or rain streaks across various scales, the DRS-Former introduces multiple scale-wise separable convolution paths within the MSFN.

Next, the DRS-Former employs feature transformation through two parallel branches. One branch utilizes (3×3) depth-wise convolutions, while the other employs (5×5) depth-wise convolutions. This dual-path approach enhances the extraction of multi-scale local information, allowing the model to effectively handle features at different scales.

The detailed process for feature fusion in the MSFN is described as follows: $\widehat{X}_l = f_{i_1}\left(LN(X_{l-1})\right)\#(8)$

$$X_l^{P_1} = \sigma\left(f_{3\times3}^{dwc}\left(\widehat{X}_l\right)\right)\#(9)$$

$$X_l^{P_2} = \sigma\left(f_{3\times3}^{dwc}\left(X_l^{P_1}, X_l^{S_1}\right)\right)\#(10)$$

$$X_l^{S_2} = \sigma\left(f_{5\times5}^{dwc}\left(X_l^{S_1}, X_l^{P_1}\right)\right)\#(11)$$

$$X_l = f_{1\times1}^{c}\left(X_l^{P_2}, X_l^{S_2}\right) + X_{l-1}\#(11)$$

In this context, the ReLU activation function is denoted by $\sigma(\cdot)$, introducing a non-linear component to the model. The expression $f_{c1\times1}$ signifies a (1×1) convolution, which is utilized to enhance the channel dimension and facilitate subsequent processing. Depth-wise convolutions with kernel sizes of (3×3) and (5×5) are represented by $f_{3\times3}^{dw}$ and $f_{5\times5}^{dw}$dw , respectively. These depth-wise convolutions are crucial for capturing local features at multiple scales. The channel-wise concatenation operation, symbolized by $|\cdot|$, merges outputs from different convolutional paths into a unified feature representation.

2.3 Mixture of Experts Feature Compensator (MEFC)

The DRSformer incorporates a novel component called the MEFC which enhances performance through a combination of sparsity and feature compensation techniques. This approach builds upon the foundational design of effective CNN models, widely referenced in existing literature. In the MEFC module, DRSformer employs a series of parallel layers, each comprising multiple sparse CNN operations, referred to as "Experts." These experts include a variety of specialized layers: an average pooling layer with a 3×3 receptive field, separable convolution layers with kernel sizes of (1×1), (3×3), (5×5), and (7×7), along with dilated convolution layers of identical kernel sizes to capture diverse spatial features.

DRSformer utilizes self-attention mechanisms as a dynamic selector for the experts. This design allows the model to adaptively prioritize different representations based on the characteristics of the incoming data.

Starting from an input feature map $X_{l-1} \in R^{H\times W\times C}$, the DRSformer initially computes a C-dimensional channel descriptor $z_c \in R^C$ by applying a channel-wise average.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{l-1}(i,j), \#(12)$$

where $X_{l-1}(i,j)$ denotes the feature at position (y, x) within the feature map. Each expert's influence is then modulated by learnable weight matrices $W_1 \in R^{T \times C}$ and $W_2 \in R^{O \times T}$, where T is the dimension of these matrices.

The output for the l-th MEFC layer is computed as follows:

$$T_l = W_2 \sigma(W_1 z_c), \#(13)$$

$$X_l = f_{c1\times1}\left(\sum_{i=1}^{o} f_{exp}(X_l, T_l)\right) + X_{l-1}, \#(14)$$

In this context, f_{xp} denotes the expert operations, while o signifies the count of experts.The term $f_{c1\times1}$ refers to a 1×1 convolution, while $\sigma(\cdot)$ denotes the ReLU activation function. The symbol $[\cdot]$ is used for channel-wise concatenation. This design intricately incorporates the MEFC into the main STBs, enabling it to adaptively manage the diverse appearances of image degradations such as rain streaks and fog.

3 Results

3.1 Datasets and Metrics

The article uses 1,200 synthetic image pairs from the SateHaze1k dataset [13] in this work. Based on the degree of fog, this dataset is separated into three categories: thin, moderate, and thick. There are 400 image pairs in each category for training. A selection of 120 image pairs, proportionately dispersed over the three fog types, are chosen for testing. As with the previously stated datasets, the article measure the efficacy of dehazing approach in this study using assessment metrics such as the Structural Similarity Index Measure (SSIM) and the Peak Signal-to-Noise Ratio (PSNR) [14, 15]. These measurements aid in validating the level of image restoration quality that the article's method provides under hazy settings.

3.2 Experimental Setup

Datasets. As mentioned above, the article experiments with defogging using the SateHaze1k dataset. The diverse levels of fog in this dataset offer a thorough assessment of the model's method's performance in various scenarios.

Comparing Methods. The article evaluate DRSformer model against a number of cutting edge dehazing techniques, including as transformer-based models like M2SCN [16] and SkyGAN[17, 18], CNN-based techniques like FCFT-Net [8, 16] and SAR-Opt-cGAN[13, 19], and conventional techniques like DCP [3, 16]. The information used in these models comes from earlier assessments conducted by Chen [13] and Huang [13], Zhou [16]. To ensure a fair comparison, the article retrain recent models in the event that no pretrained versions are available. The article refer to published results in the literature for other methodologies.

Evaluation Metrics. As the article previously indicated, the article uses PSNR and SSIM measures to evaluate the dehazing algorithm's performance on the SateHaze1k dataset. By comparing the restored image's similarity to the original, unaltered image, PSNR assesses the restoration's correctness. On the other hand, SSIM looks at the structural similarity between the reference and restored images, providing information about how well the important structural details are preserved.

Instructional Specifics. In the DRS model, $\{4, 4, 6, 6, 8\}$ is defined as $\{N_0, N_1, N_2, N_3, N_4,\}$. The number of attention heads for each of the four STBs at each level is set to $\{1, 2, 4, 8\}$. The initial count (C) has 48 channels with an expansion ratio of 2. For the weight matrix, **T** is set to **32**, and the number of experts in the MEFC is **O = 8**. Due to the simplicity of the fog patterns in SateHaze1k, MEFC is not utilized for training. Additionally, the channel expansion factor r in MSFN for the sparseness values $\{1/2, 4/5\}$ in the STB is 2.66.

3.3 Results and Comparison

Table 1. Quantitative PSNR and SSIM comparison based on SateHaze1k.

Model	Thin		Moderate		Thick	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Original	12.77	0.7316	12.59	0.7408	8.58	0.4215
DCP [3, 16]	13.15	0.7246	9.783	0.5735	10.25	0.5850
SAR-Opt-cGAN[13, 19]	20.19	0.8419	21.66	0.7941	19.65	0.7573
FCFT-Net [8, 16]	23.59	0.9127	22.88	0.9272	20.03	0.8156
M2SCN [16]	25.21	0.9175	26.11	0.9416	21.33	0.8289
SkyGAN[13, 18]	25.38	0.9248	25.58	0.9035	23.43	0.8925
DRSformer	25.93	0.9310	27.12	0.9477	23.33	0.8672

The mean PSNR and SSIM values for the test procedure in thin, medium, and thick haze conditions are displayed in Table 1. The table illustrates that the SSIM is 0.7246 and the PSNR of DCP [3, 16] is 13.15 dB in misty conditions. These findings highlight DCP's shortcomings, particularly when dealing with thicker haze.

In hazy conditions, SkyGAN [13, 18] demonstrates significantly better performance, achieving 25.38 dB in PSNR and 0.9248 in SSIM, highlighting the effectiveness of advanced deep learning models. SAR-Opt-cGAN [13, 19], although achieving a respectable SSIM of 0.8419 and a PSNR of 20.19 dB, does not perform as well as other models. FCFT-Net [8, 16] and M2SCN [16] both achieve high PSNR

and SSIM values under hazy conditions. Specifically, M2SCN reaches an SSIM of 0.9175, while FCFT-Net records a PSNR of 23.59 dB, with both models achieving an SSIM of 0.9127. These results indicate their strong capability to mitigate haze.

However, the DRS-Former model, which is the focus of this study, outperforms all other models. In thin haze conditions, DRS-Former achieves a PSNR of 25.93 dB and an SSIM of 0.9310. For medium haze conditions, it records a PSNR of 27.12 dB and an SSIM of 0.9477. Even in thick haze conditions, DRS-Former maintains superior performance with a PSNR of 23.33 dB and an SSIM of 0.8672. These results clearly demonstrate that DRS-Former sets a new state-of-the-art in haze removal, proving its robustness and effectiveness across various haze intensities.

However, the DRS-Former model, which is the focus of this study, outperforms all other models. In thin haze conditions, DRS-Former achieves 25.93 dB in PSNR and 0.9310 in SSIM. For medium haze conditions, it records 27.12 dB in PSNR and 0.9477 in SSIM. Even in thick haze conditions, DRS-Former maintains superior performance with 23.33 dB in PSNR and 0.8672 in SSIM. These results clearly demonstrate that DRS-Former remains effective on the SateHaze1k dataset, extending its applicability in the field of image restoration.

3.4 Discussion

The article’s trials' outcomes show how well the DRS-Former performs in dehazing RS images in a range of haze situations. DRSformer robustness and effectiveness are demonstrated by the strong PSNR and SSIM values observed in the thin, moderate, and thick haze categories. In particular, this performance was much improved by the MEFC and Sparse STB, which improved feature extraction and processed multi-scale haze patterns.

The model performs noticeably better, especially in heavy haze, than both sophisticated deep learning models like SkyGAN and more conventional techniques like DCP. This is explained by the DRSformer's capacity to efficiently manage non-uniform haze dispersion by utilizing the advantages of Transformer-based systems. By utilizing MSFN and TKSA, the DRSformer can effectively handle different haze patterns and concentrate on pertinent features.

The article’s study does have several limitations, though. Even while the SateHaze1k dataset is extensive, its synthetic nature might not adequately represent the complexities of actual atmospheric conditions. Subsequent investigations may concentrate on verifying the model with authentic datasets and investigating the incorporation of other contextual data to augment dehazing efficacy. Further insights into the practical usefulness of the DRSformer could be gained by investigating its computing efficiency and scalability in bigger and more diverse datasets.

4 Conclusion

The article introduced the DRSformer in this work, which is a Transformer-based network designed to dehaze images obtained from remote sensing (RS). Using a MEFC and novel Sparse Transformer Blocks (STB), the DRSformer manages the

non-uniform haze distribution that is frequently present in RS pictures. The article's approach ensures high-quality dehazing outcomes by efficiently capturing pertinent characteristics and processing multi-scale haze patterns.

The article have shown through extensive testing on the SateHaze1k dataset that the DRSformer works much better than both deep learning-based and classical dehazing approaches. Across various hazy conditions, DRSformer exhibited PSNR and SSIM values that highlight its exceptional performance. The DRSformer is a viable option for enhancing the usefulness of RS pictures in real-world vision applications because of its capacity to adjust to various haze levels and improve image quality.

In the future, DRSformer can be further improved to handle a wider range of complex atmospheric conditions and investigate its potential applications in further image restoration. The article's goal is to promote RS technology and its applications to Earth observation by further developing DRSformer capabilities.

References

1. Zheng, X., Sun, H., Lu, X., Xie, W.: Rotation-Invariant Attention Network for Hyperspectral image classification. IEEE Transactions on Image Processing 31, 4251–4265 (2022).
2. Bi, G., Si, G., Zhao, Y., Qi, B., Lv, H.: Haze removal for a single remote sensing image using Low-Rank and sparse prior. IEEE Transactions on Geoscience and Remote Sensing 60, 1–13 (2022).
3. Zhang, Y., Guindon, B., Cihlar, J.: An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. Remote Sensing of Environment 82, 173–187 (2002).
4. He, N.K., Sun, N.J., Tang, N.X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 2341–2353 (2011).
5. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: DeHazeNet: an End-to-End system for single image Haze removal. IEEE Transactions on Image Processing 25, 5187–5198 (2016).
6. Zhu, X., Li, S., Gan, Y., Zhang, Y., Sun, B.: Multi-Stream fusion network with generalized smooth L1 loss for single image dehazing. IEEE Transactions on Image Processing 30, 7620–7635 (2021).
7. Liu, X., Ma, Y., Shi, Z., Chen, J.: GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. IEEE/CVF International Conference on Computer Vision (2019).
8. Gu, Z., Zhan, Z., Yuan, Q., Yan, L.: Single remote sensing image dehazing using a Prior-Based dense attentive network. Remote Sensing 11, 3008 (2019).
9. Li, Y., Chen, X.: A Coarse-to-Fine Two-Stage attentive network for haze removal of remote sensing images. IEEE Geoscience and Remote Sensing Letters 18, 1751–1755 (2021).
10. Huang, Y., Chen, X.: Single Remote Sensing Image Dehazing Using a Dual-Step Cascaded Residual Dense Network. IEEE International Conference on Image Processing (ICIP) (2021).
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. arXiv (Cornell University) 30, 5998–6008 (2017).

12. Chen, X., Li, H., Li, M., Pan, J.: Learning A Sparse Transformer Network for Effective Image Deraining. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).
13. Huang, B., Li, Z., Yang, C., Sun, F., Song, Y.: Single Satellite Optical Imagery Dehazing using SAR Image Prior Based on conditional Generative Adversarial Networks. IEEE Winter Conference on Applications of Computer Vision (WACV) (2020).
14. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. Electronics Letters 44, 800 (2008).
15. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612 (2004).
16. Li, S., Zhou, Y., Xiang, W.: M2SCN: Multi-Model Self-Correcting Network for satellite Remote Sensing Single-Image Dehazing. IEEE Geoscience and Remote Sensing Letters 20, 1–5 (2023).
17. Chen, X., Li, Y., Dai, L., Kong, C.: Hybrid High-Resolution Learning for Single Remote sensing satellite Image dehazing. IEEE Geoscience and Remote Sensing Letters. 19, 1–5 (2022). https://doi.org/10.1109/lgrs.2021.3072917.
18. Mehta, A., Sinha, H., Mandal, M., Narang, P.: Domain-Aware Unsupervised Hyperspectral Reconstruction for Aerial Image Dehazing. IEEE/CVF Winter Conference on Applications of Computer Vision (2021).
19. Grohnfeldt, C., Schmitt, M., Zhu, X.: A Conditional Generative Adversarial Network to Fuse Sar and Multispectral Optical Data For Cloud Removal From Sentinel-2 Images. IEEE International Geoscience and Remote Sensing Symposium (2018).