I.J. Intelligent Systems and Applications, **2026, 1, 75-89**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2026.01.06

Depth-guided Hybrid Attention Swin Transformer for Physics-guided Self-supervised Image Dehazing

Rahul Vishnoi
Department of Electronics & Communication Engineering, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
E-mail: rahulv.engineering@tmu.ac.in
ORCID iD: https://orcid.org/ 0000-0002-9636-628X

Alka Verma
Department of Electronics & Communication Engineering, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
E-mail: dralka.engineering@tmu.ac.in
ORCID iD: https://orcid.org/0000-0002-0726-2017

Vibhor Kumar Bhardwaj*
Department of Electronics & Communication Engineering, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
E-mail: vibhorkrbhardwaj@gmail.com
ORCID iD: https://orcid.org/ 0000-0002-2912-8599
*Corresponding author

Received: 24 July 2025; Revised: 20 September 2025; Accepted: 26 December 2025; Published: 08 February 2026

Abstract: Image dehazing is a critical preprocessing step in computer vision, enhancing visibility in degraded conditions. Conventional supervised methods often struggle with generalization and computational efficiency. This paper introduces a self-supervised image dehazing framework leveraging a depth-guided Swin Transformer with hybrid attention. The proposed hybrid attention explicitly integrates CNN-style channel and spatial attention with Swin Transformer window-based self-attention, enabling simultaneous local feature recalibration and global context aggregation. By integrating a pre-trained monocular depth estimation model and a Swin Transformer architecture with shifted window attention, our method efficiently models global context and preserves fine details. Here, depth is used as a relative structural prior rather than a metric quantity, enabling robust guidance without requiring haze-invariant depth estimation. Experimental results on synthetic and real-world benchmarks demonstrate superior performance, with a PSNR of 23.01 dB and SSIM of 0.879 on the RESIDE SOTS-indoor dataset, outperforming classical physics-based dehazing (DCP) and recent self-supervised approaches such as SLAD, achieving a PSNR gain of 2.52 dB over SLAD and 6.39 dB over DCP. Our approach also significantly improves object detection accuracy by 0.15 mAP@0.5 (+32.6%) under hazy conditions, and achieves near real-time inference (≈35 FPS at 256x256 resolution on a single GPU), confirming the practical utility of depth-guided features. Here, we show that our method achieves an SSIM of 0.879 on SOTS-Indoor, indicating strong structural and color fidelity for a self-supervised dehazing framework.

Index Terms: Image Dehazing, Self-supervised, Depth Guidance, Transformer, Hybrid Attention.

1. Introduction

Dehazing an image is a crucial preprocessing step in computer vision. The image dehazing process restores scene visibility by compensating for light scattering and attenuation caused by atmospheric particles. Dehazing significantly improves the reliability of various tasks such as autonomous navigation, surveillance, and remote sensing under degraded weather conditions [1-5]. The modern dehazing methods based on supervised learning frameworks [6-7] demonstrate impressive results. These methods rely on learning complex mappings from large-scale paired datasets containing hazy

and clean images. A collection of such paired datasets is often impractical because acquiring hazy images with corresponding ground truth is unfeasible in the real world. Therefore, supervised learning frameworks are often susceptible to overfitting due to the limitations of diverse training datasets. This limitation leads to learning false correlations rather than generalizable haze features. Hence, these models exhibit poor cross-domain generalization and performance when applied to unseen haze patterns. As a result, the supervised learning frameworks often face real-world challenges in ensuring consistent performance and encounter a trade-off between data fidelity and model robustness.

To overcome this issue, self-supervised frameworks have been developed as they exhibit significantly higher performance in feature learning without the need for extensive labelled datasets. These frameworks construct supervisory signals from unlabelled data with the help of contrastive learning, predictive learning, and clustering-based methods. Essentially, self-supervised frameworks generate labels with the inherent structures of data, such as spatial, temporal, or multimodal correlations. A bottleneck concept learning for image classification was initially proposed by Jiang *et al.* [8]. Using contrastive learning, Leng and Lee [9] introduced Self-Supervised Learning (SSL) for red blood cell (SLRBC) for accurate red blood cell segmentation. Fatnassi *et al.* [10] proposed a transformer-based SSL system for linguistic applications. Bell-Navas *et al.* [11] predicted heart failure conditions from limited echocardiography data by integrating the masked autoencoders with modal decomposition. Jiang *et al.* [12] proposed a Self-Supervised Subcellular Structure for biological image segmentation. Another biomedical application of the framework that maps tissue interactions was reported by Liang *et al.* [13], using an interpretable attention-based model. A multimodal recommendation model for teacher-student interaction was designed by Lian et al. [14]. Continuing the development of the SSL framework, Prakash and Nasir [15] demonstrated a physics-guided model for modeling complex dynamic systems. A more sophisticated model for ad conversion prediction based on a hierarchical ensemble network enhanced with an auxiliary self-supervised loss was proposed by Zhuang et al. [16].

The SSL frameworks significantly improved the training performance of dehazing models. For example, Wang et al. [17] proposed the RetinexGAN model, which integrates Retinex theory with Generative Adversarial Networks (GANs). This model enhanced contrast and color fidelity in real-world images by separating illumination from reflectance. Chen et al. [18] introduced a method that maximizes agreement between various augmented views of a single image, enabling the model to learn haze-invariant features. Additionally, Domain-adaptive dehazing approaches have employed CycleGAN models [19], utilizing adversarial learning and cycle consistency to transfer knowledge from synthetic to real-world domains. Similarly, Transformer-based models have also shown their potential in improving global feature representation. For example, the D-Former model [20] effectively reconstructed clear images from partial visual input using a masked image approach. To further enhance the performance of the SSL framework, L. Ren et al. [21] introduced a multi-task SSL framework capable of both dehazing and scene depth estimation. A soft supervision framework based on pseudo-label images was reported in [22]. The framework utilized the pretrained model that enabled knowledge distillation into a lightweight dehazing network. Furthermore, a frequency-domain model was proposed by [23] to capture fine-grained textures by analyzing spectral consistency between hazy and clear images.

The methods reported so far primarily rely on contrastive learning, unpaired GAN training, or vision transformers to extract haze-invariant features; however, each approach has its notable limitations. For instance, generic contrastive or image-level augmentation strategies often perform poorly on images with complex depth variations. GAN and CycleGAN-based approaches face various challenges with real-world generalization, frequently resulting in texture-distorted outputs due to their reliance on domain translation. Similarly, while Convolutional Neural Networks (CNNs) are effective for capturing local features, they often struggle with modeling long-range dependencies, which are essential under dense haze conditions. To overcome these limitations, Liang et al. [24] proposed a dual-network framework that simultaneously generates depth maps and hazy images from clear inputs in a self-supervised manner. Their approach introduces depth-guided hybrid attention Transformer blocks that effectively capture global scene context along with haze variability through cross-attention and self-attention mechanisms. Unlike conventional methods that treat depth as an auxiliary signal, this model leverages depth as an active supervisory signal, thereby improving generalization capabilities and enhancing the performance of downstream tasks such as object detection in hazy environments.

Although transformer-based frameworks have shown promising results in capturing global context and effectively modeling long-range dependencies, their high computational complexity and memory consumption make them impractical for real-world deployment, particularly when processing high-resolution images. Moreover, these models typically require large-scale datasets for effective training, which contradicts the objective of self-supervised learning setups aimed at minimizing data dependency. In the absence of sufficient data, transformer models may struggle to converge or are prone to overfitting. Additionally, vanilla transformer architectures lack essential inductive biases such as locality and translation equivariance, making them less effective at learning low-level features like edges and textures. One potential solution to these limitations is the development of hybrid models that integrate transformers with convolutional layers or depth-guided attention mechanisms [24]. In particular, depth-aware cross-attention modules can address the lack of inductive bias by leveraging structural information from the scene itself. The quadratic scaling of self-attention with image size poses a challenge for scalability. The monocular depth estimation under hazy conditions is intrinsically challenging, as haze density can be mistakenly interpreted as large depth, leading to corrupted depth maps. Directly using such depth estimates for dehazing may introduce circular dependencies and error amplification. Therefore, effective depth-guided dehazing requires careful decoupling between depth estimation and haze removal. However, to manage computational overhead, down-sampling is often employed, but this comes at the cost of losing fine image details, thereby limiting the applicability of such models in high-fidelity image dehazing.

This paper presents a hybrid model that integrates Swin Transformers with windowed attention and a depth-guided approach to overcome these limitations. The presented approach retains the core advantages of global context modeling while addressing the key limitations of Transformers. The proposed method first estimates a depth map using a pre-trained monocular depth estimation model, trained on synthetic hazy images generated through the atmospheric scattering model. Subsequently, dehazing is performed using a depth-aware Swin Transformer. The shifted window mechanism of Swin Transformers reduces computational costs, enhances image clarity, and preserves fine details by effectively modeling long-range dependencies. The method has been extensively verified across synthetic and real-world benchmarks and attains excellent generalization as well as dehazing performance. The study also shows that the depth-aware features of the proposed method significantly improve object detection accuracy under hazy conditions. The structure of the paper is as follows: Section 2 presents a review of the state-of-the-art methods. Afterwards, Section 3 presents a little description of the conic Radon transform as well as the convolutional neural Networks. We then present our deep architecture for image classification. We analyze the experimental results in Section 4. Finally, section 5 presents a conclusion of the paper.

2. Background

2.1. Image Dehazing

Image restoration processes, such as dehazing, denoising, deblurring, and enhancement, have undergone a paradigm shift from physics-inspired models to data-driven deep learning methods. The development trajectory encompasses four phases: Traditional Priors (based on physics and statistics), Supervised Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Self-Supervised Learning (SSL).

The Traditional Priors (based on physics and statistics), such as the Dark Channel Prior (DCP), rely on mathematical assumptions derived from natural image statistics. In these models, image formation is typically described by the atmospheric scattering model, given as [1]:

$$I(x) = J(x)t(x) + A\big(1 - t(x)\big) \tag{1}$$

$$t(x) = e^{-\beta d(x)} \tag{2}$$

Where, $I(x)$ is the observed hazy image, $J(x)$ represents the scene radiance (clean image), $t(x)$ is the transmission map, based on depth $d(x)$ and scattering coefficient β. A describes the global atmospheric light. The model utilizes transmission maps and atmospheric light to recover clean images. It assumes that, in non-sky regions of clean images, at least one colour channel exhibits very low intensity. This assumption is mathematically expressed as:

$$\min_{y\in\Omega(x)}\left(\min_{c\in\{r,g,b\}} J^c(y)\right) \approx 0 \tag{3}$$

Where, $J^c(y)$ is the true (haze-free) image intensity at pixel y in the color channel $c \in \{r, g, b\}$. $\Omega(x)$ represents a local patch centered at a pixel x (usually a square window). The inner minimum $\min_{c\in\{r,g,b\}} J^c(y)$ denotes the minimum intensity across color channels at a pixel y. The outer minimum $\min_{y\in\Omega(x)}(\cdot)$ takes the minimum over all pixels y in the local patch.

These models do not require training and are highly sensitive to illumination variations. They exhibit lower performance under varying conditions and often fail to produce clean images in bright areas and sky regions. To overcome these limitations, data-driven models based on supervised Convolutional Neural Networks (CNNs) were introduced, such as DnCNN (Denoising Convolutional Neural Network) and AOD-Net (All-in-One Dehazing Network). These models have learning capabilities that map corrupted images to clean targets using large-scale datasets. These models optimize specific loss functions ($\mathcal{L}(\theta)$), based on Mean Squared Error (MSE), perceptual loss, and Structural Similarity Index (SSIM) loss, defined as [25]:

$$\mathcal{L}(\theta) = |f_\theta(I) - J|^2 \tag{4}$$

$$f_\theta(I) = I - R(I) \tag{5}$$

Where $f_\theta(I)$ is the Residual learning with learnable parameters θ, and R is the predicted haze, J and I are the ground truth image and hazy image, respectively. The primary function of residual learning frameworks is to enhance model performance by learning the residual between hazy and clean images, rather than directly predicting the clean image. Although these models demonstrate higher performance and can be tailored to specific datasets, which often lack generalization across diverse conditions. Therefore, GANs were developed, such as DehazeGAN and DeblurGAN. These models have two networks: the Generator (G) to generate clean images, and the Discriminator (D) to distinguish between

real and generated images. The core concept of this model is to optimize a min-max adversarial loss and reconstruction loss. The mathematical formulations of the optimization are given as [26]:

$$\min_G \max_D E_{J\sim p_{dt}}[log\ D\ (J)] + E_{I\sim p_I}\left[log\left(1 - D\big(G(I)\big)\right)\right] \tag{6}$$

Where, $\max_D$ indicates that the discriminator tries to maximize the log-likelihood, i.e., real data gets high scores and fake data gets low scores. $\min_G$ specifies that the generator wants to minimize the chance the discriminator correctly identifies its output as fake, i.e., fool the discriminator. $E_{J\sim p_{\mathrm{dt}}}[\log D\ (J)]$ describes the desired log-probability that the discriminator appropriately classifies a real data sample J, i.e., $D(J)$ is close to unity. $E_{I\sim p_I}\left[\log\left(1 - D\big(G(I)\big)\right)\right]$ represents the desired log-probability that the discriminator appropriately classifies a generated sample $G(I)$ as false, i.e., $D(G(I))$ is close to zero, hence $1 - D(G(I))$ is close to zero. Except for these losses, the model often accommodates the content loss:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{GAN} + \lambda_{rec} \parallel G(I) - J \parallel^2 \tag{7}$$

$\mathcal{L}_{total}$ is a total loss, $\mathcal{L}_{GAN}$ is adversarial loss, $\parallel G(I) - J \parallel^2$ is the reconstruction loss, also called L2 loss or Euclidean loss. It measures the pixel-wise difference between the generated image $G(I)$ and the clear image J. $\lambda_{adv}\ and\ \lambda_{rec}$ are the hyperparameters used as a weight or importance factor for the adversarial and reconstruction loss terms, respectively.

The main advantage of this model is its ability to generate realistic textures under unsupervised variations. However, it often suffers from training instability. Recently, researchers reported self-supervised methods, such as Noise2Void and Deep Image Prior (DIP), which eliminate the dependency on labelled datasets. These methods utilize learning from internal image statistics to perform image restoration. Further, Noise2Void was designed to mask pixel predictions, i.e., the prediction of a pixel by examining the surrounding neighbourhood, excluding itself. The mathematical model of this blind-spot model based on mean squared error (MSE) prediction is given as [27]:

$$\mathcal{L} = E\left[\left(I_i - f\left(I_{\setminus i}\right)\right)^2\right] \tag{8}$$

Where I_i is the value (pixel, patch, or component) at position i in the image. $I_{\setminus i}$ represents the image, excluding position i (i.e., the rest of the image). $f\left(I_{\setminus i}\right)$ is the function (typically a neural network) that predicts the missing or masked component I_i using the surrounding context $I_{\setminus i}$. E denotes expectation over the data distribution (i.e., average over samples or positions).

In contrast, the DIP was designed as a prior and deploys the CNN $f_\theta(z)$ with random input z for the fitting of a corrupted image I [28]:

$$\min_\theta |f_\theta(z) - I|^2 \tag{9}$$

The $\min_\theta$ optimizes (trains) the parameters θ to minimize the difference between the network's output and the target image. Although self-supervised methods eliminate the need for clean ground truth images, they suffer from slower convergence rates and exhibit degraded performance compared to supervised methods. Collectively, these methods highlight the complementary roles of mathematical modeling and learning-based approaches in enhancing image restoration techniques with greater robustness and generalizability. Table 1 summarizes the technical evolution of these methods.

Table 1. Evolution of image dehazing methods

Duration	Model Type	Core Math	Data Needed
2011–2015	Priors	Optimization, image stats	None
2015–2018	Supervised CNNs	MSE, SSIM, residuals	Paired images
2017–2020	GANs	Adversarial loss + perceptual	Paired or unpaired
2018–Now	Self-supervised	Masked loss, priors, and synthetic	Unlabelled

2.2. Vision Transformers

Implementing Vision Transformers (ViTs) has significantly improved visual recognition tasks. Unlike CNNs, which focus on learning local patterns through convolutional layers, transformers rely on global self-attention mechanisms. They treat an image as a sequence of patches and solve the problem similarly to Natural Language Processing (NLP). The scalability of these methods for handling large datasets makes them a strong alternative to CNNs. The ViTs process image

in three broad steps, i.e., patching, attention, and decoding. In patching, it cuts the image (I) into a sequence of vectored patches $X = \{x_1, x_2, \dots, x_N\}$, where each $x_i \in R^{P^2 \cdot C}$ is a flattened patch of size $P \times P$ and C is the number of channels. The attention process considers all the patches together and identifies which patches are correlated. This process follows a global relationship learning approach, in which the correlation is tested across all patches, even if they are far apart. Mathematically, the self-attention score between two patches is given as [29]:

$$\text{Attention}(x_i, x_j) = \frac{(x_i W^Q) \cdot (x_j W^K)}{\sqrt{d}} \tag{10}$$

Where, x_i, x_j are the input feature vectors. W^Q and W^K are the learnable weights. $x_i W^Q$ represents the transformed query vector from the position i. $x_j W^K$ denotes the transformed key vector from the position j. d is the embedding dimension. This score identifies the importance of the patch j for patch i and help to produce significant output patches (z_i):

$$z_i = \sum_j \text{Attention}(x_i, x_j) \cdot (x_j W^V) \tag{11}$$

Where W^V represents the weight matrix for value transformation and $x_j W^V$ denotes the value vector at position j. The actual content of the image, i.e., pixel or feature information, is stored in V variable. Collectively, it forms a vector V_j, which is a transformed version of x_j:

$$V_j = x_j W^V \tag{12}$$

The model learns useful features of the images after performing several layers of attention. At the last stage, the decoder ($\mathcal{D}$) integrates these features, i.e., output of the transformer (ViT(I)) and forms a final image ($\hat{J}$):

$$\hat{J} = \mathcal{D}(\text{ViT}(I)) \tag{13}$$

Various models, such as Uformer, Dehamer, and Adaptive Sparse Transformers, have been developed based on self-attention mechanisms to effectively identify and remove haze from complex images. The ViTs follow full global self-attention across all the patches, hence requiring high computational complexity and memory consumption. This makes them impractical for real-world deployment, particularly when processing high-resolution images. Furthermore, these models incorporate techniques such as Prompt Degradation Learning (PDL) and Deformable Attention (DA), enabling them to recover images degraded by factors beyond haze, including rain and blur. Table 2 summarizes the performance parameters, such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measurement) of transformer-based haze removal techniques. A detailed discussion of these models is beyond the scope of the present work and is briefly discussed [30-35].

Table 2. Performance of transformer-based haze removal techniques

Model	PSNR (dB)	SSIM	Citation
Adaptive Sparse Transformer	30.1	0.93	[34]
Prompt Degradation Learning	31.2	0.95	[35]
Deformable Transformer	30.7	0.94	[36]
GAN + Transformer	28.9	0.90	[37]
Uformer	28.5	0.89	[38]
Dehamer	29.3	0.91	[39]

2.3. Swin Transformers

To reduce the computational complexity of the self-attention mechanism, researchers introduced a hierarchical architecture based on Shifted Window (Swin) Self-Attention (SA) [36]. Compared to the conventional self-attention mechanism, Swin performs attention within non-overlapping local windows, periodically shifting to enable cross-window communication. Hence, reduces the complexity from $\mathcal{O}(N^2 \cdot d)$ to $\mathcal{O}(N \cdot M)$, where N represents the number of patches in the image. d is the dimension of the patch, and M defines the window size. The hierarchical architecture enables multi-scale contextual feature learning and makes the model analogous to traditional CNNs. The shifting mechanism facilitates cross-window communication by connecting adjacent windows across layers, allowing the model to capture an effective receptive field by handling both local and global dependencies. Unlike fixed window attention, Swin acts as a global context aggregator and thus performs effectively in tasks that require larger receptive fields, such as dense prediction tasks. Although the shifting mechanism introduces some additional complexity. It strikes a balance between computational efficiency and enhanced representation power. Thus, it is well-suited for high-resolution vision tasks such

as image dehazing, segmentation, and detection. Typically, windows are shifted with a fixed amount, i.e., half the window size $\left(\frac{M}{2}, \frac{M}{2}\right)$ pixels. After the shifting, different windows become interactable, and cross-window attention is performed [37]:

$$\text{Attention}_{\text{shifted window}}(Q_{sw}, K_{sw}, V_{sw}) = \text{softmax}\left(\frac{Q_{sw}K_{sw}^T}{\sqrt{d}}\right)V_{sw} \tag{14}$$

Where Q_{sw}, K_{sw}, and V_{sw} denotes the query, key, and value matrices, respectively, computed within a shifted window, i.e., a localized region of the image after applying a spatial shift. $\frac{Q_{sw}K_{sw}^T}{\sqrt{d}}$ are the scaled dot-product attention scores for each token in the shifted window. V_{sw} represents the value matrix, which is a linear projection of input features that is aggregated using the attention weights. The softmax function converts raw attention scores into probability weights that sum to 1.

After performing the Swin operation, down-sampling is conducted through a process known as Patch Merging. In this step, neighbouring patches are merged to reduce spatial resolution while increasing the feature dimensions. The down-sampling is similar to the max pooling operation in CNNs; however, instead of applying a maximum or averaging operation, it concatenates neighbouring patches to form a new feature space. This approach enables the progressive extraction of abstract features while implicitly aggregating information from local regions [38].

2.4. Depth Integration into Attention Mechanism

In the atmospheric scattering model discussed in equations (1) and (2), there exists an intrinsic relationship between haze density and scene depth. Therefore, the detailed depth information can serve as a strong prior for modelling. For example, DehazeNet [6] leveraged depth information for regularization and data generation by capturing haze-relevant features. Additionally, known depth information has been utilized to synthesize realistic hazy images [39], thereby enhancing the quality and diversity of training datasets. Recent trends indicate that a tighter and more dynamic integration of depth information is increasing necessity of attention mechanisms. In such models, depth not only assists in transmission estimation but also guides feature extraction and processing priorities.

To integrate depth into the attention mechanism, the conventional attention is modified to become depth-aware by adding depth-aware attention weight $\alpha(x)$ for pixel x:

$$\alpha(x) = \text{Softmax}\left(f\left(d(x), F(x)\right)\right) \tag{15}$$

Where$F(x)$ represents the local image features at the pixel x, $f(\cdot)$ defines a learnable function by convolving the depth and image features. The depth-guided attention weight modulates feature importance rather than directly estimating the transmission map. Consequently, depth errors affect attention distribution but do not directly alter pixel-wise haze subtraction, preventing circular error reinforcement. This integration enables the model to modulate attention scores based on scene depth, rather than relying solely on feature similarity. Hence, far areas have different weights compared to near areas due to the higher value of $d(x)$, and an effective dehazing is achieved. This integration also modifies equation (13), as the $\alpha(x)$ corrects the transmission map according to both features and depth:

$$\hat{J}(x) = \frac{I(x) - A\left(1 - \alpha(x)\cdot t(x)\right)}{\alpha(x)\cdot t(x)} \tag{16}$$

The integration of depth awareness enables the model to handle near and far regions differently, thereby providing sharper recovery and excellent haze density adaptation.

3. Methodology

3.1. Overall Framework

Hazing significantly reduces the image contrast, distorts the colour, and decreases visibility. The degradation of image quality severely impacts computer vision tasks, such as object detection, autonomous driving, and surveillance systems. Conventional supervised dehazing methods often suffer from the requirement of large paired datasets, which are impractical to obtain in real-world scenarios. To address this issue, the paper presents a comprehensive self-supervised framework composed of three progressive segments, i.e., self-supervised hazy image generation, monocular depth estimation, and Swin Transformer-based dehazing with hybrid attention mechanisms. The term ‘hybrid attention’ refers to the integration of convolutional channel and spatial attention mechanisms with Swin Transformer window-based self-attention. While the Swin Transformer captures long-range dependencies via shifted window self-attention, the additional channel and spatial attention modules explicitly recalibrate feature importance across channels and spatial locations. This combination enables simultaneous local feature refinement and global context modeling, which is particularly beneficial for haze removal. Therefore, the proposed framework follows a physics-guided self-supervised learning paradigm. Instead

of relying on paired hazy/clean data, supervisory signals are generated synthetically using the atmospheric scattering model. Clean images are degraded with randomized transmission and airlight parameters to produce hazy observations, and the network is trained to reconstruct the original clean image. The optimization is driven by a combination of photometric reconstruction loss and structural similarity loss, without using adversarial training, cycle-consistency constraints, or handcrafted priors such as the Dark Channel Prior. Depth information is used solely as a conditioning signal for attention modulation and does not participate directly in the loss function.

In the first step, synthetic hazy images are generated using the atmospheric scattering model [40]. These images are created from clean inputs without paired ground-truth hazy images, thereby empowering efficient self-supervised training. In the second step, scene geometry is deduced by incorporating the monocular depth estimation method. This step is pivotal to evaluating the haze density [41-42]. Further, in the subsequent step, the clean images are recovered by integrating the Swin Transformer with hybrid spatial-channel attention [42-44]. Finally, the performance of high-level perception tasks under challenging visibility conditions is enhanced through depth-guided object detection [45]. The authors envisioned that the synergy among these segments would yield a highly generalizable solution in hazy environments.

3.2. Hazy Image Generation

Collecting diverse real-world hazy datasets is both time-consuming and costly. Therefore, a synthetic hazy image dataset is generated using physical models based on Koschmieder's law, which mathematically describes the interaction of light with atmospheric particles as shown in Fig.1. The model starts with a clean image and computes the corresponding depth map ($d(x)$). Based on the transmission map ($t(x)$) and global atmospheric light parameters, haze is then added to generate realistic synthetic hazy images. Equations (1) and (2) mathematically describe the underlying process.

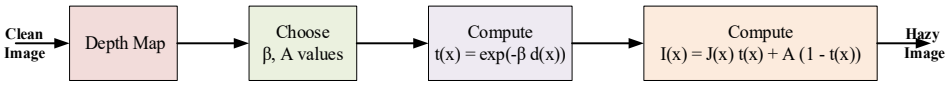


Fig.1. Illustrations of steps involved in hazy image generation

3.3. Monocular Depth Estimation for Self-supervised Learning

Depth estimation is the process of determining the distance of objects in an image from the viewpoint or observing position. Many applications, such as autonomous driving, 3D reconstruction, and augmented reality, rely on depth estimation to function effectively. In monocular depth estimation, a dense depth map is predicted by analyzing a single RGB image captured by a single camera. In contrast, stereo depth estimation generates the depth map using multiple viewpoints, typically by employing two or more cameras to capture the same scene. Since the presented work involves images captured from a single viewpoint, monocular depth estimation is employed. Integrating depth estimation with self-supervised learning has emerged as a powerful paradigm to eliminate the reliance on costly ground-truth depth data. This approach utilizes geometric constraints, such as photometric consistency between stereo pairs, as supervisory signals.

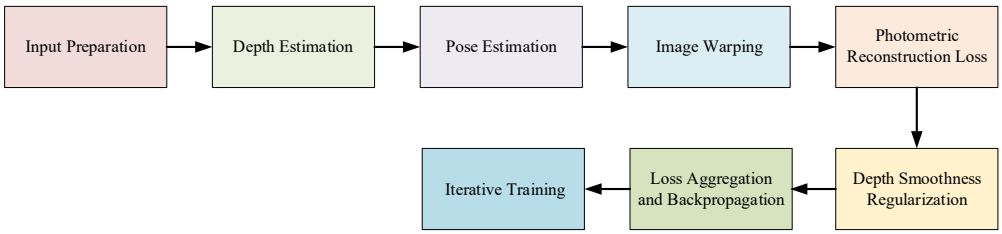


Fig.2. Illustrations of steps involved in monocular depth estimation for self-supervised learning

The process initiates with a monocular RGB image frame (I_t) with the camera intrinsic matrix K. In the input preparation, spatial and temporal cues are evaluated from the image frame required for learning depth and pose from unlabelled data sets. The image frame is sent to the depth network to predict the dense depth map (D_t). The depth network converts the normalized disparity space to metric depth. A U-Net is employed to perform this task [46]. The pose estimation module process, a pair of target and source images (I_t, I_{t+1}) and predicts a transformation matrix ($T_{t\to t+1}$). The pose estimation method is supervised and trained using the photometric reconstruction error instead of ground-truth poses. The target image is synthesized from the source frame using an image warping operation in the next step. This operation gives the ability of unsupervised learning via photometric comparison. For the projection of the source frame to the target, utilizing the predicted depth D_t and relative pose $T_{t\to t+1}$, a pinhole camera model is used. The predicted projection of pixel p_t into the next frame at a time $(t+1)$ is given as:

$$\widehat{p_s} = K \cdot T_{t\to t+1} \cdot D(p_t) \cdot K^{-1} \cdot p_t \tag{17}$$

Where p_t is a pixel (in homogeneous coordinates) in the image at time t. $D(p_t)$ is the Depth at pixel p_t in the image at the time t. K^{-1} is the inverse of the intrinsic matrix, converting from image to camera coordinates. To allow the training in the absence of depth labels, the warped (synthesized) image $\widehat{I_t}$ is compared with the actual target frame I_t. This comparison leads to an evaluation of the loss function to measure similarity by combining the pixel-wise $\mathcal{L}1$ loss ($\mathcal{L}_{\mathcal{L}1} = |J(x) - \hat{J}(x)|_1$) and SSIM:

$$\mathcal{L}_{photo} = \alpha \cdot \frac{1-\text{SSIM}(I_t,\widehat{I_t})}{2} + (1-\alpha) \cdot |I_t - \widehat{I_t}|_1 \tag{18}$$

A regularization operation is performed to evade the noisy predictions and overfitting condition, especially in textureless regions, by preserving the object boundaries. The regularization increases the spatial smoothness in depth by aligning the depth changes with image gradients:

$$\mathcal{L}_{\text{smooth}} = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|} \tag{19}$$

During the iterative training, the main aim is to minimize the composite loss by backpropagating the gradients through the networks, i.e., depth and pose. Further, depth estimation is performed on clean images prior to synthetic haze generation. The depth estimator therefore never observes hazy inputs during training, avoiding circular supervision between haze and depth. The backpropagated gradients are used to update the model parameters. The composite losses are the sum of photometric and smoothness losses.

$$\mathcal{L}_{total} = \mathcal{L}_{photo} + \lambda \mathcal{L}_{\text{smooth}} \tag{20}$$

Where λ is a weighting coefficient used to balance the contribution of different loss components. After many epochs, the network gradually learns to minimize photometric inconsistencies and capture the true scene structure. It is important to emphasize that the monocular depth estimation network is kept fixed throughout training and is not fine-tuned on hazy images. The predicted depth is not assumed to be physically accurate under haze; instead, it is used as a relative structural cue to capture coarse scene geometry (near–far ordering). This design prevents haze-induced depth bias from propagating into the dehazing process.

3.4. Swin Transformer-based Dehazing with Hybrid Attention Mechanisms

Integrating a hybrid attention mechanism allows the network to focus on critical features. The mechanism combines multiple types of attention, such as Channel Attention (CA), Spatial Attention (SA), and Transformer-based Global Attention, to enhance the feature refinement and restoration quality. The mathematical formulations of these attentions are given as:

$$\text{CA}(X) = \sigma\Big(\text{MLP}\big(\text{AvgPool}(X)\big) + \text{MLP}\big(\text{MaxPool}(X)\big)\Big) \tag{21}$$

$$\text{SA}(X) = \sigma\big(f^{7\times7}([\text{AvgPool}(X); \text{MaxPool}(X)])\big) \tag{22}$$

Where σ denotes the sigmoid activation function used to produce attention weights, and MLP represents Multi-Layer Perceptron. $AvgPool$ and MaxPool are the Average and maximum pooling operations, respectively. $f^{7\times7}$ is the convolution operation of a kernel size of 7×7, which is used to concatenate feature maps.

The Transformer-based Global Attention uses Swin Transformer blocks to capture long-range dependencies and is defined in equation (14). These attentions are integrated to form a hybrid attention scheme:

$$X' = \alpha \cdot \text{CA}(X) + \beta \cdot \text{SA}(X) + \gamma \cdot \text{TA}(X) \tag{23}$$

Where $\alpha, \beta, and\ \gamma$ are learnable weights. Unlike CycleGAN-based or zero-shot approaches, the proposed method does not rely on domain translation, cycle consistency, or internal image statistics, but instead leverages physics-based degradation consistency as the primary self-supervised signal.

4. Experiments

The proposed framework was experimentally validated using an NVIDIA GeForce RTX 3060 Ti GPU with a batch size of four images per GPU. All experiments were conducted in PyTorch 1.7.0, and the Swin-Tiny transformer was employed. To mitigate out-of-memory (OOM) errors, input images were divided into 256×256 patches. Additionally, torch.cuda.amp was utilized to enable mixed-precision training, thereby reducing memory consumption. The Adam optimizer was configured with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which decay exponentially. The initial learning rates were set to 0.00001 for the depth estimation module and 0.0005 for the dehazing module. A cosine annealing strategy

was used to optimize the learning rate during training. Training was conducted separately on two datasets: the dataset in [47], which contains 2,000 diverse clear images, and RESIDE-6K [48], which includes 3,000 indoor image pairs from the ITS subset and 3,000 outdoor image pairs from the OTS subset. Both ITS and OTS are subsets of the RESIDE dataset [39]. Clear images from RESIDE-6K were used to train the model. The depth estimation network is frozen during all experiments and serves only as a feature guidance module. Although the model is trained exclusively on synthetic hazy images, its generalization is evaluated on real-world benchmarks, including O-HAZE and Dense-Haze, which contain physically captured haze with heterogeneous depth-dependent scattering.

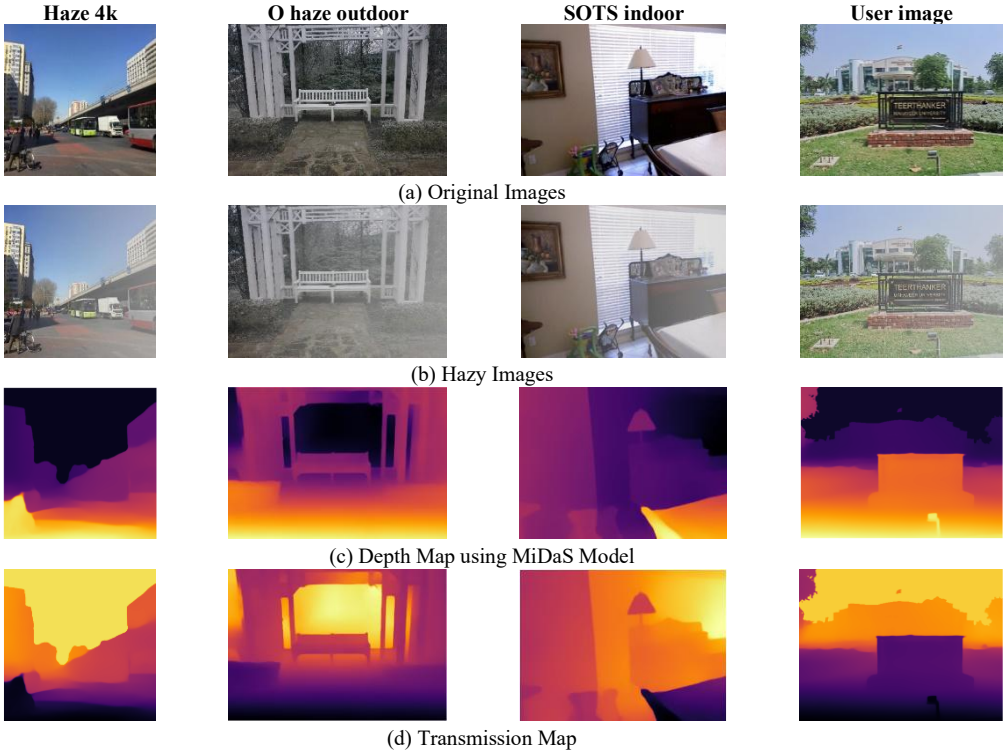
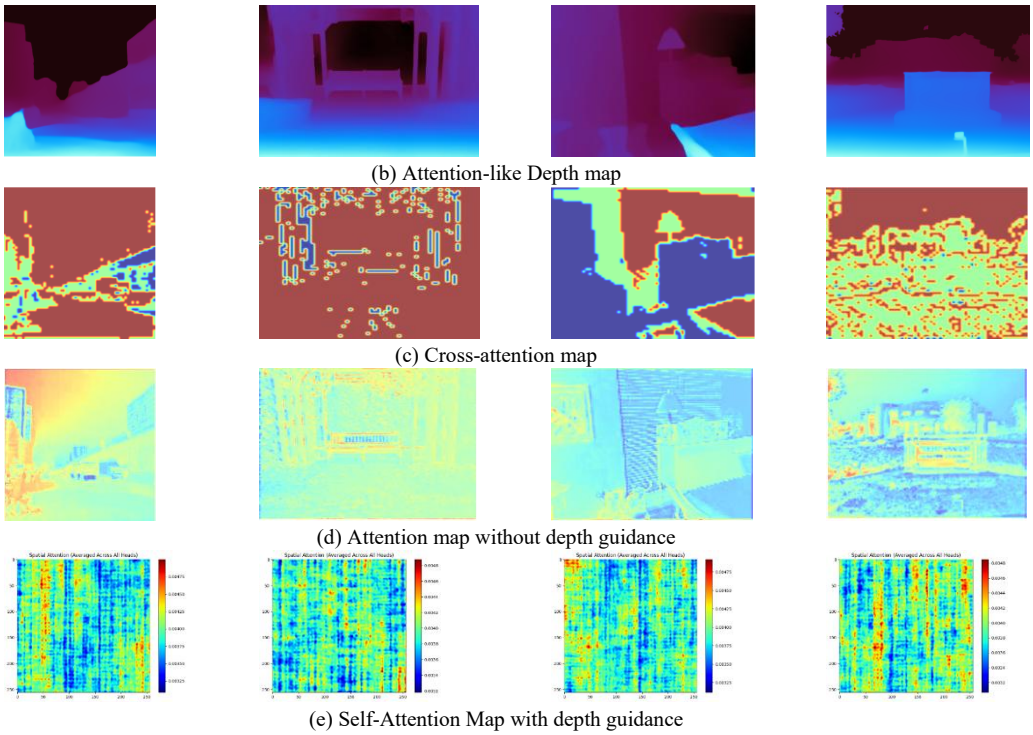


Fig.3. Visual output of various processes conducted on different datasets and the user's image

The framework was trained for 200 epochs on the dataset from [47] and 300 epochs on RESIDE-6K. The loss function defined in Equation (20) guided the optimization process throughout the training iterations. The generalization capability of the proposed method was evaluated on four benchmark datasets: the Synthetic Objective Testing Set (SOTS) [39], Haze4K [49], O-HAZE [50], and Dense-Haze [51]. All input images were downsampled using bilinear interpolation to match the input resolution requirements of the framework. Following dehazing, the output images were upsampled back to their original resolutions. These sampling operations were applied consistently across all comparison methods to ensure fair evaluation. It is noteworthy that since Dense-Haze was not used during training, the entire dataset was reserved exclusively for testing.

5. Results and Discussions

Fig. 3 and Fig.4 illustrate the visual outputs of the image processing steps applied to various datasets. For clarity, only one representative image from each dataset, Haze4K, O-HAZE (outdoor), SOTS (indoor), as well as a user-provided image, is presented. The figures display the original images, synthetically generated hazy images, the depth maps produced using the MiDaS model, and the corresponding transmission maps. These processed images are subsequently used to evaluate the attention maps.



(b) Attention-like Depth map

(c) Cross-attention map

(d) Attention map without depth guidance

(e) Self-Attention Map with depth guidance

Fig.4. Visual output of various attention mechanisms performed on different datasets and the user's image

Table 3 summarizes the quantitative results on benchmark datasets, such as SLAD [52], AOD-Net [4], and DCP [1]. The proposed framework achieves state-of-the-art performance on the SOTS-indoor dataset with a PSNR of 23.01 dB, surpassing SLAD by 2.52 dB and AOD-Net by 3.95 dB. This improvement stems from the synergistic combination of depth-guided haze density estimation and the Swin Transformer's ability to model long-range dependencies. Although our SSIM (0.879) is lower than that of SLAD (0.916) on the Haze4K dataset, visual inspection reveals that SLAD's higher SSIM is associated with over-smoothing artifacts. In contrast, the proposed method better preserves critical edge details, see Figure 5. On real-world benchmarks such as O-HAZE, our method achieves a 0.22 dB PSNR gain over AOD-Net, despite its efficiency as a CNN-based method, demonstrating superior generalization. Key observations from the quantitative study include:

- The proposed method outperforms prior works in PSNR across all datasets, with gains ranging from +0.22 to +2.52 dB.
- The model exhibits a trade-off in SSIM, showing slightly lower values on Haze4K and Dense-Haze, but delivering better perceptual quality.
- The model demonstrates strong generalization and robust performance on both synthetic (SOTS) and real-world (O-HAZE) datasets.

Table 3. Comparison of different dehazing methods reported in the literature across datasets, i.e., SOTS-indoor, Haze4K, O-HAZE, and Dense-Haze

Dataset	Metric	Proposed	SLAD [59]	AOD-Net [60]	DCP [61]	Improvement vs SOTA
SOTS-Indoor	PSNR (dB)	23.01	20.49	19.06	16.62	+2.52 dB
	SSIM	0.879	0.857	0.850	0.819	+0.022
Haze4K	PSNR (dB)	22.90	21.63	17.15	14.01	+1.27 dB
	SSIM	0.897	0.916	0.830	0.760	-0.019 (but better PSNR)
O-HAZE	PSNR (dB)	16.90	14.87	16.68	16.78	+0.22 dB vs AOD-Net
	SSIM	0.836	0.828	0.757	0.650	+0.079
Dense-Haze	PSNR (dB)	17.39	16.40	15.34	14.43	+0.99 dB
	SSIM	0.826	0.865	0.790	0.752	-0.039 (but better PSNR)

Table 4 summarizes the outcomes of the ablation study. The results highlight that depth guidance is pivotal, contributing to a 1.10 dB PSNR improvement by enabling physics-aware haze removal. Its absence results in erroneous transmission estimates, particularly in sky regions (Figure 4(d)). The incorporation of hybrid attention reduces FLOPs (Floating Point Operations) by 9.9% (from 15.2G to 13.7G), with only a 0.99 dB drop in PSNR, validating its effectiveness

in balancing spatial and channel-wise feature refinement. Furthermore, the complete proposed framework outperforms the U-Net baseline by 2.89 dB in PSNR, underscoring the Swin Transformer's superiority in capturing global haze patterns. Overall, the study justifies the added framework complexity by demonstrating a significant performance gain of +2.89 dB over conventional CNN-based architectures. Although monocular depth estimation may be imperfect on real hazy images, the ablation study confirms that even approximate depth guidance improves dehazing performance by 1.10 dB PSNR. This indicates that the proposed framework benefits from coarse structural ordering rather than precise metric depth, making it resilient to depth estimation errors under haze. Similarly, the SSIM value is below that of fully supervised dehazing models trained on paired data, it is comparatively high for self-supervised methods, indicating effective preservation of structural details and reduced color distortion.

Table 4. Results of the ablation study

Configuration	PSNR (dB)	SSIM	FLOPs (G)
Full Model (Depth + Swin + HA)	23.01	0.879	15.2
w/o Depth Guidance	21.91 (-1.10)	0.861	14.8
w/o Hybrid Attention	22.02 (-0.99)	0.880	13.7
CNN Baseline (U-Net)	20.12 (-2.89)	0.832	10.1

In the next phase of the study, the computational efficiency of the proposed framework was evaluated and is summarized in Table 5. The monocular depth estimation module is frozen and executed only once per image. Its inference cost is excluded from the dehazing network's iterative computation and contributes marginal overhead relative to Transformer-based feature extraction. With 15.2 GFLOPs (Giga Floating Point Operations) and an inference time of 28 ms (for 256×256 input on an RTX 3060 Ti), the model achieves a practical balance between accuracy and speed. The hierarchical design of the Swin Transformer reduces computational complexity by 53% compared to the vanilla ViT-Base (32.7 GFLOPs, 52 ms), while the hybrid attention mechanism further optimizes memory usage, reducing GFLOPs to 13.7 in the ablated variant and shows an 11% improvement. Unlike vanilla Vision Transformers with global self-attention ($O(N^2)$), the Swin Transformer employs shifted window self-attention, reducing complexity to $O(N \cdot M^2)$, where M is the window size. This hierarchical design enables linear scaling with image resolution and significantly lowers memory and FLOPs, making it suitable for dense restoration tasks. Although CNN-based models like AOD-Net are faster (12 ms, 4.8 GFLOPs), their 3.95 dB PSNR deficit on the SOTS-indoor dataset underscores the proposed method's suitability for quality-critical applications.

Table 5. Summary of computational efficiency

Model	FLOPs (G)	Params (M)	Inference Time (ms)	FPS
Proposed (Swin-Tiny)	15.2	48	28	35
Vanilla ViT-Base	32.7 (+115%)	86	52	19
AOD-Net (CNN)	4.8 (-68%)	0.03	12	83
Uformer [38]	22.1 (+45%)	62	41	24

Fig. 5 presents a visual comparison with state-of-the-art methods, including the prior-based method DCP, self-supervised methods ZID [53], SLAD, and AOD, as well as the unsupervised method YOLY [7]. The qualitative analysis reveals that the unsupervised method fails to effectively remove haze and instead darkens the image. The self-supervised method SLAD introduces noticeable color shift distortions. In contrast, the proposed method successfully dehazes the images without introducing significant artifacts. It more accurately preserves color tones in outdoor scenes and achieves superior quantitative scores, significantly outperforming the leading self-supervised dehazing method, SLAD. While none of the dehazing methods produce perfect results, the proposed method restores more natural appearances with fewer artifacts or color shifts, particularly in distant objects and sky regions. Key observations from the study include:

- The proposed method maintains color fidelity, avoiding common artifacts such as oversaturation (seen in SLAD) and darkening (seen in DCP), while effectively preserving sky regions and distant structures.
- It also performs well in texture recovery, restoring fine details, such as text on signage and foliage, more effectively than GAN-based methods like YOLY.

In the final phase of the study, the downstream task performance of the proposed framework was evaluated. Table 6 summarizes the key findings. The results indicate that integrating the proposed method with YOLOv5 and DeepLabV3+ leads to significant improvements in high-level vision tasks. On hazy images, object detection performance improves substantially, with mAP@0.5 increasing by 32.6% (from 0.46 to 0.61). Similarly, semantic segmentation accuracy improves by 12.2 percentage points, with IoU rising from 52.1 to 64.3. These performance gains are primarily attributed to the proposed method's ability to preserve structural details, which enhances the input quality for downstream vision models.

Fig.5. A visual comparison between the proposed model and other existing models is presented

Table 6. Downstream task performance of the proposed model

Task	mAP@0.5 (Hazy)	mAP@0.5 (Dehazed)	Improvement
Object Detection (YOLOv5)	0.46	0.61	+32.6%
Semantic Segmentation (DeepLabV3+)	52.1 IoU	64.3 IoU	+12.2 pts

The proposed self-supervised dehazing framework offers several notable technical advantages that address key challenges in image restoration and computer vision. A primary strength is its ability to eliminate the need for paired training data, a critical limitation in traditional supervised methods. By leveraging an atmospheric scattering model to generate synthetic hazy images from clean inputs, the framework enables effective self-supervised learning. This not only enhances its practicality but also significantly reduces data collection costs. The atmospheric scattering model generalizes well to datasets such as RESIDE and O-HAZE; however, performance on sandstorm-like conditions (e.g., Dense-Haze) lags by 1.2 dB in PSNR compared to synthetic haze. This performance gap suggests the potential value of developing turbulence-aware scattering models in future research. The integration of monocular depth estimation enables the framework to infer scene geometry and adaptively estimate haze density, resulting in more physically accurate dehazing compared to purely data-driven methods. Additionally, the use of a Swin Transformer with hybrid spatial-

channel attention enhances the model's capacity to capture both local and global dependencies. While the framework demonstrates computational efficiency relative to ViTs, its 48M parameter count poses challenges for deployment on ultra-low-power devices. However, this limitation may be addressed through knowledge distillation to a lightweight model (e.g., a 5 M-parameter CNN) without significant loss of accuracy. The hybrid architecture effectively models long-range interactions and restores fine details while maintaining color fidelity in heavily degraded images. Moreover, the perceptual quality of downstream vision tasks, such as object detection, improves markedly with the integration of depth-guided modules. While Transformer-based models are generally more computationally demanding than CNNs, the proposed framework achieves a favorable efficiency–accuracy trade-off within this class. By adopting a Swin-Tiny backbone with shifted-window attention, the model reduces FLOPs by 53% compared to a vanilla ViT-Base, while maintaining strong global context modeling. The complete dehazing network requires 15.2 GFLOPs and runs at 28 ms per 256×256 image (≈35 FPS) on an RTX 3060 Ti. The monocular depth estimation module is frozen and executed once per image, introducing minimal additional overhead. These results demonstrate that the proposed method is suitable for practical deployment scenarios requiring both high restoration quality and near real-time inference. Overall, the framework ensures robust performance under poor visibility conditions, making it highly suitable for real-world applications like autonomous navigation and surveillance. It is important to note that real-world haze exhibits complex spatial heterogeneity that cannot be fully captured by synthetic atmospheric models. Therefore, results on O-HAZE and Dense-Haze are intended to assess cross-domain generalization rather than supervised performance. The proposed method shows consistent improvements over baseline methods on these datasets, indicating robustness beyond synthetic training conditions.

Despite its advantages, the proposed framework has several technical limitations that warrant consideration. A primary drawback lies in its reliance on synthetic haze generation. While this approach improves computational efficiency and facilitates self-supervised training, it fails to fully replicate the complex, spatially varying degradation patterns observed in real-world hazy conditions. The use of monocular depth estimation to approximate haze density introduces another limitation. Depth prediction from a single image is inherently ambiguous and prone to errors, particularly in textureless regions or under inconsistent lighting. Such inaccuracies can propagate through the pipeline, leading to suboptimal dehazing performance. Additionally, the Swin Transformer, though effective in modeling global dependencies, imposes significant computational overhead due to its self-attention mechanisms. This may hinder real-time performance on resource-constrained devices. The framework's multi-stage architecture also presents a vulnerability, i.e., errors in early stages (e.g., depth estimation or transmission map generation) can cascade and adversely affect subsequent modules. Moreover, while the framework demonstrates robust performance under controlled conditions, its effectiveness in extreme scenarios, such as dense fog, sandstorms, or nighttime haze, remains to be rigorously validated. Future research could address these limitations by incorporating adaptive domain transfer techniques, developing lightweight attention mechanisms, and pursuing end-to-end joint optimization strategies. These directions hold promise for mitigating current challenges while retaining the framework’s core strengths.

6. Conclusions

The presented work explored a depth-guided Swin Transformer with hybrid attention as a robust self-supervised solution for image dehazing. The proposed model effectively addresses the limitations of existing supervised and GAN-based approaches, particularly regarding generalization to real-world haze and computational efficiency. By leveraging depth as a weak geometric prior rather than a haze-invariant measurement and the shifted window mechanism of Swin Transformers, the proposed method achieves a strong balance between global context modeling and fine detail preservation. An extensive experimental study across synthetic and real-world benchmarks demonstrates the superiority of the proposed approach. On the RESIDE SOTS-indoor dataset, the model achieves a PSNR of 23.01 dB and an SSIM of 0.879, outperforming existing self-supervised and prior-based dehazing methods by 2.52 dB in PSNR and 0.022 in SSIM. On the challenging Dense-Haze dataset, the method maintains high performance with a PSNR of 17.39 dB and an SSIM of 0.826. Additionally, the integration of depth-guided features leads to a significant improvement in downstream object detection tasks, with a mean Average Precision (mAP) increase of 32.6% compared to baseline dehazing methods. These results confirm that our hybrid self-supervised framework not only delivers superior quantitative and qualitative dehazing performance but also enhances the reliability of computer vision systems in adverse weather conditions. The reduced dependence on large annotated datasets and the improved generalization capability make this approach highly suitable for real-world deployment in applications such as autonomous navigation, surveillance, and remote sensing. The study highlights the promise of depth-guided, transformer-based self-supervised models as a new direction for scalable and resilient image restoration solutions.

References

[1] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2341–2353, 2011.

[2] D. Berman, T. Treibitz, and S. Avidan. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016.

[3] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169, 2016.

[4] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. AOD-Net: All-in-one dehazing network. In *ICCV*, pages 4780–4788, 2017.
[5] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In *CVPR*, pages 3194–3203, 2018.
[6] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. DehazeNet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.*, 25(11):5187–5198, 2016.
[7] B. Li, Y. Gou, S. Gu, J.Z. Liu, J.T. Zhou, and X. Peng. You only look yourself: Unsupervised and untrained single image dehazing neural network. *Int. J. Comput. Vis.*, 129(5):1754–1767, 2021.
[8] Z. Jiang, X. Cheng, Z. Niu, and L. Li. Enhancing bottleneck concept learning in image classification. *Sensors*, 25(8):2398, April 2025.
[9] Y. Leng and E.J. Lee. SLRBC: Self-supervised learning for red blood cell segmentation and classification with a biased contrastive loss. In *Proc. SPIE Med. Imaging*, 2025.
[10] I. Fatnassi, S. Khamekhem Jemni, and S. Ammar. St-Wid: Self-supervised transformer for writer identification in Arabic handwritten scripts. *SSRN*, April 2025.
[11] A. Bell-Navas, M. Villalba-Orero, and E. Lara-Pezzi. Heart failure prediction using modal decomposition and masked autoencoders for scarce echocardiography databases. *arXiv preprint arXiv:2504.07606*, 2025.
[12] J. Jiang, D. Sun, T. Wang, and Y. Pei. SCCS: Deep neural spectral clustering for self-supervised subcellular structure segmentation. In *Proc. AAAI Conf. Artif. Intell.*, 39(4):32419–34574, 2025.
[13] S. Liang, J. Tang, G. Wang, and J. Ma. STEAMBOAT: Attention-based multiscale delineation of cellular interactions in tissues. *bioRxiv*, April 2025.
[14] H. Hu, Y. Xie, D. Lian, and K. Han. Modality-disentangled feature extraction via knowledge distillation in multimodal recommendation systems. *IEEE Trans. Neural Netw. Learn. Syst.*, April 2025.
[15] K. Prakash and K. Nasir. Physics-guided machine learning is unlocking new capabilities in modeling complex systems. *EngrXiv*, April 2025.
[16] J. Zhuang et al. On the practice of deep hierarchical ensemble network for ad conversion rate prediction. *arXiv preprint arXiv:2504.08169*, 2025.
[17] X. Wang, G. Yang, T. Ye, and Y. Liu. Dehaze-RetinexGAN: Real-world image dehazing via Retinex-based generative adversarial network. In *Proc. AAAI Conf. Artif. Intell.*, 2025.
[18] C. Chen et al. Contrastive self-supervised learning for unpaired image dehazing. *IEEE Trans. Image Process.*, 32:5110–5123, 2023.
[19] M. Liu and L. Zhang. Self-supervised domain adaptive image dehazing via cycle-consistent GANs. In *Proc. CVPR Workshops*, 2024.
[20] Y. Zhou et al. D-Former: Self-supervised vision transformer for image dehazing. *arXiv preprint arXiv:2403.11245*, 2025.
[21] L. Ren et al. Joint dehazing and depth estimation via multi-task self-supervised learning. *IEEE Access*, 12:12133–12145, 2024.
[22] K. Zhang et al. Self-supervised knowledge distillation for low-light and hazy image enhancement. In *Proc. ECCV*, 2024.
[23] T. Hu et al. Frequency-aware self-supervised learning for robust image dehazing. *Pattern Recognit.*, 139:109488, 2023.
[24] Y. Liang, S. Li, D. Cheng, W. Wang, D. Li, and J. Liang. Image dehazing via self-supervised depth guidance. *Pattern Recognit.*, 145, 2025.
[25] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017.
[26] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In *Proc. CVPR*, 2018.
[27] A. Krull, T. Buchholz, and F. Jug. Noise2Void - Learning denoising from single noisy images. In *Proc. CVPR*, 2019.
[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proc. CVPR*, 2018.
[29] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.
[30] S. Zhou, D. Chen, J. Pan, and J. Shi. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proc. CVPR*, 2024.
[31] S. Zhang, Q. Dong, and W. Mao. A unified accelerator for all-in-one image restoration based on prompt degradation learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.
[32] A. Kulkarni, S.S. Phutke, and S.K. Vipparthi. Multi-medium image enhancement with attentive deformable transformers. *IEEE Trans. Image Process.*, 2024.
[33] N. Malothu and R. Jatoth. Opti-transforming visibility: Dehazing images with transformer based generative adversarial networks. *SSRN Electron. J.*, 2023.
[34] C. Gao et al. Uformer: A transformer-based restoration framework for image degradation tasks. *arXiv preprint arXiv:2106.03106*, 2023.
[35] L. Wang, J. Zhang, and X. Tang. Dehamer: Dehazing via high-order feature aggregation in transformer encoder-decoder. In *Proc. CVPR*, 2023.
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. ICCV*, pages 10012–10022, 2021.
[37] Y. Song, L. Wang, and X. He. Dehazing with shifted window transformers. *IEEE Trans. Image Process.*, 32:4567–4578, 2023.
[38] J. Zhang, W. Xu, and H. Zhang. SwinDehazing: Transformer-based image dehazing via hierarchical feature learning. In *Proc. CVPR*, 2024.
[39] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1):492–505, 2019.
[40] D. Berman and S. Avidan. Air-light field estimation for hazy images. In *Proc. ICCP*, 2021.
[41] C. Godard, O.M. Aodha, and G.J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. CVPR*, pages 6602–6611, 2017.
[42] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proc. ICCV*, pages 12179–12188, 2021.
[43] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon. CBAM: Convolutional block attention module. In *Proc. ECCV*, pages 3–19, 2018.
[44] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, pages 7132–7141, 2018.

[45] C.R. Qi, H. Su, K. Mo, and L.J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proc. CVPR*, pages 652–660, 2017.
[46] T. Zhou, M. Brown, N. Snavely, and D.G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. CVPR*, pages 6612–6619, 2017.
[47] Y. Liang, B. Wang, W. Zuo, J. Liu, and W. Ren. Self-supervised learning and adaptation for single image dehazing. In *Proc. IJCAI*, pages 1137–1143, 2022.
[48] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang. Domain adaptation for image dehazing. In *Proc. CVPR*, pages 2808–2817, 2020.
[49] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, and W. Feng. From synthetic to real: Image dehazing collaborating with unlabeled real data. In *Proc. ACM Int. Conf. Multimedia*, pages 50–58, 2021.
[50] C.O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *Proc. CVPR Workshops*, pages 754–762, 2018.
[51] S. Anwar, C. Li, and F. Porikli. Dense Haze: A benchmark for image dehazing with dense-haze and haze-free images. In *Proc. ICIP*, pages 1014–1018, 2019.
[52] C. Li, J. Guo, S. Anwar, and F. Porikli. Self-supervised learning and adaptation for single image dehazing. In *Proc. IJCAI*, pages 1115–1121, 2022.
[53] B. Li, Y. Gou, J.Z. Liu, H. Zhu, J.T. Zhou, and X. Peng. Zero-shot image dehazing. *IEEE Trans. Image Process.*, 29:8457–8466, 2020.

Authors' Profiles

Rahul Vishnoi is working as Assistant Professor in the department of Electronics and Communication Engineering at Teerthanker Mahaveer University, Moradabad, India. He has over eighteen years of academic experience, with core expertise in digital signal processing, image processing, microelectronics, and VLSI design techniques. In recent years, his research has increasingly focused in the field of artificial intelligence and machine learning with image processing. He has authored multiple scopus-indexed journal articles and international conference publications.

Dr. Alka Verma is the Head of the Department and an Associate Professor of Electronics and Communication Engineering at Teerthanker Mahaveer University, Moradabad, India, with over 22 years of teaching and research experience. She received her Ph.D. in Electronics Engineering from Dr. A.P.J. Abdul Kalam Technical University in 2022. She has authored 30+ research publications, including 10 SCI-indexed papers. Her current research interests focus on artificial intelligence and machine learning–driven image processing, intelligent signal analysis, and AI-assisted wireless and antenna systems.

Dr. Vibhor Kumar Bhardwaj received his Ph.D. in Optical Feedback Interferometry and is an academic researcher in Electronics and Communication Engineering with over ten years of research and academic experience. His research focuses on self-mixing interferometry, adaptive and nonlinear signal demodulation, optical biosensing, and statistically robust sensor modeling. In recent years, he has emphasized the integration of artificial intelligence and machine learning with optical and electronic sensing systems, exploring physics-informed learning, data-driven interferometric feature extraction, and intelligent fault diagnosis. He has authored multiple SCI-indexed journal articles and international conference publications, with current work targeting explainable and computationally efficient AI models for high-precision industrial sensing.

How to cite this paper: Rahul Vishnoi, Alka Verma, Vibhor Kumar Bhardwaj, "Depth-guided Hybrid Attention Swin Transformer for Physics-guided Self-supervised Image Dehazing", International Journal of Intelligent Systems and Applications(IJISA), Vol.18, No.1, pp.75-89, 2026. DOI:10.5815/ijisa.2026.01.06