

Wavelet-Driven GAN with Attention Mechanism for Improved Raindrop Removal

Muthukumar Balamurugan , Shivarama K. Holla , and Varun P. Gopi , *Senior Member, IEEE*

Abstract—The adhesion of raindrops to camera lenses causes severe distortions and image aberrations, posing a major problem for computer vision systems used in autonomous driving, surveillance, and environmental monitoring. This paper presents a powerful multi-scale restoration architecture, D2W-GAN, that combines the Dual-Tree Complex Wavelet Transform (DT-CWT) with a U-shaped network, an attention mechanism, and Generative Adversarial Networks (GANs) to remove raindrops. Current techniques tend to struggle with removing artifacts introduced by raindrops and maintaining overall image fidelity. The proposed method minimally distorts raindrop distortions by using DT-CWT, which simultaneously provides fine high-frequency and large-scale low-frequency components, thereby improving the separation of the distortion. The addition of attention mechanisms enables the network to focus more on the regions of the raindrops, thereby improving restoration accuracy. Refinements to the adversarial learning framework yielded more realistic and consistent results. The proposed method achieves balanced performance across standard evaluation metrics and superior SSIM scores on the RainDrop dataset, indicating better structural preservation and perceptual quality. The results show that wavelet-based feature extraction, feature attention, and adaptive optimization have been successfully integrated to achieve effective image restoration under unstable and demanding conditions.

Link to graphical and video abstracts, and to code:
<https://latam.ieee9.org/index.php/transactions/article/view/10735>

Index Terms—Single Image Raindrop Removal, Complex Wavelet Transform (CWT), GAN, Attention.

I. INTRODUCTION

RAINY conditions heavily degrade visual data as raindrops stick to camera lenses, which is a significant problem for computer vision systems. Raindrops degrade the important details of a scene, compromise image quality, and adversely affect object detection, recognition, tracking, and scene interpretation. Many standard image-restoration methods have been proposed, but removing all distortions caused by raindrops has proven difficult, especially when they vary in

size, shape, and intensity. Denoising diffusion models are powerful generative methods capable of generating high-quality images from noisy or damaged inputs [20], [32]. Generative models based on scores have demonstrated high efficiency in image restoration as conditioned models when confronted with degraded images (e.g., images with raindrops) [21]. A number of diffusion-based frameworks have been specialized in restoration tasks following iterative generative processes [3], [4]. Stable diffusion is a typical diffusion algorithm that runs the reverse process, starting with Gaussian noise, but it is sensitive to the quality of the degraded result and requires a large number of sampling steps. This frequently results in inefficiencies, particularly when trying to recover fine high-frequency information while maintaining global low-frequency structures. Image-to-image translation processes based on deep learning can be broadly divided into generative models, such as Generative Adversarial Networks (GANs), and hierarchical models, such as Convolutional Neural Networks (CNNs). CNN-based models continuously decompose features with convolutional and pooling layers to allow the model to learn both local textures and more global contextual features [15], [17]. An example is the context-and-detail interaction networks, which use a coarse-to-fine restoration scheme that entails both rain-suppression and detail-enhancement modules to refine the result iteratively [11]. Conversely, GAN-based models include a generator and a discriminator that assess the validity of generated images as realistic, which helps the generator to produce visually plausible results.

Although these designs are successful, traditional CNN designs can be limited in dealing with the complexity of raindrop removal. They can have very high computational requirements or inadequately handle contextual information. Standard CNNs are limited in that their depth and receptive field are relatively shallow, and finer details of variations in rain patterns are relatively hard to model. To overcome such challenges, recent methods have more detailed architectures that extract multi-scale features, employ attention mechanisms, and use perceptual loss functions to improve restoration performance [23], [24], [27]. However, recognizing and rebuilding raindrop-affected regions are difficult, as they have irregular, unclear forms, are occluded, and often lack background information. This challenge is exacerbated when raindrops are large and sparsely distributed. To address this issue, discriminative networks were introduced to GAN-based approaches to assess the realism of restored images. The proposed framework restores raindrop-degraded images by integrating frequency decomposition based on the Dual-Tree Complex Wavelet Transform (DT-CWT), attention-based feature refinement, and adversarial

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos Thomaz (*Corresponding author: Varun P. Gopi*).

M. Balamurugan is with the Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tamil Nadu, India, and with Continental Autonomous Mobility, Pvt. Ltd., Advanced Driver Assistance Systems Division, Bengaluru, Karnataka, India (e-mail: muthukumar.balamurugan@continental-corporation.com).

S. K. Holla is with Continental Autonomous Mobility, Pvt. Ltd., Advanced Driver Assistance Systems Division, Bengaluru, Karnataka, India (e-mail: shivarama.2.holla.k@continental-corporation.com).

Varun P. Gopi is with the Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tamil Nadu, India (e-mail: varun@nitt.edu).

learning. The main contributions of this work include the following:

- A framework is introduced that integrates DT-CWT and an adversarial-based approach to single-image raindrop removal, which utilizes both spatial and frequency-domain information.
- A multi-scale coarse-to-fine raindrop removal approach that operates at successive levels of DT-CWT decomposition was incorporated.
- A mask-guided localization approach was embedded in the framework during the training step to focus the removal process in the raindrop-contaminated areas of the images. This approach has no impact on inference time.

II. RELATED WORKS

Studies on image enhancement in poor-weather scenarios have made substantial progress by adopting deep learning methods, particularly GANs. Initial methods incorporated attention mechanisms into GANs to discriminate and reject raindrops in images [15]. A technique for detecting raindrops was proposed using two attentions, in which artificial raindrop images were compared with candidate areas in the input images [17]. Although it can work under limited conditions on raindrop geometry, the method shows minimal generalizability when droplet sizes and shapes vary. There have also been investigations in the field of stereo vision where multi-camera arrangements are used to identify precipitation by measuring differences between images and the distance from the lens surface. The regions identified as raindrops are then filled using texture data from adjacent regions. However, these methods make assumptions that fail when two droplets overlap the same background area.

A cycle-consistent network architecture was proposed to model images as a combination of three components: a raindrop layer, a transparency mask, and a clean background [24]. The model is a significant improvement, as it uses high-level contextual cues to enhance low-level features through feedback and iteration, achieving high performance on benchmark datasets. In the same manner, a two-stage GAN method performs preliminary coarse removal of raindrops before a finer phase of improving visual quality [27]. Although effective, these techniques are not universal as the physical properties of rain droplets on lenses differ significantly from those of other weather phenomena, such as haze, fog, or rain streaks. CNN methods have also yielded encouraging results, especially when working with images of small, sparsely distributed droplets. However, when large or dense raindrops appear, the performance is impaired, and the methods tend to produce blurred or incompletely restored images [5], [23]. This task has been solved by using general-purpose frameworks such as Pix2Pix [6], which uses conditional GANs to examine image-to-image translation. These models are not specifically designed to remove raindrops, but they can still achieve promising results. Neural networks such as CycleGAN also enable unpaired image translation, although they struggle with transparent rain droplets that have

odd shapes and with separating them from visually similar background areas. Alternative approaches to traditional GAN methods have also been explored. Real-time imaging systems based on unmanned aerial vehicles can be used for tasks such as transmission line inspection and difference learning to improve the performance of GANs and remove raindrops [29]. The U-Net architecture has also become popular because of its high performance in pixel-based prediction tasks, and its applicability for detecting and removing raindrops. Many approaches have been proposed to address the frequency-domain representation with adversarial learning, but they have had unacceptable performance. The Mutual Channel Prior-guided Dual-domain Interaction Network (MDINet) uses spatial and frequency-domain interactions and an adaptive modulation scheme to handle various degradation patterns [16]. Table I shows a comparison of different methods based on frequency domain usage, attention mechanisms, model type, and auxiliary localization capability. Many current methods cannot fully capture the nature of real-world raindrop artifacts, and they are often not explicitly modeled due to factors such as droplet shapes and orientations, and the effects of air veils. Therefore, this paper presents an application-based framework that represents features in the wavelet domain and employs an attention-guided GAN, achieving balanced results across common benchmark datasets. Transformer-based archi-

TABLE I
COMPARISON OF DIFFERENT METHODS

Method	Frequency Domain	Attention	Diffusion/GAN	Auxiliary Localization
Resfusion	×	×	Diffusion	×
WaveDM	✓	×	Diffusion	×
MDINet	✓	✓	CNN	×
D2W-GAN	✓	✓	GAN	✓

tectures have enabled significant advances in image restoration with excellent long-range performance. Restormer is a model that uses efficient self-attention for high-resolution image restoration [30]. PromptIR improves the model's adaptability to different types of degradation via prompts [14], and DiffIR combines diffusion models with restoration methods to ensure high perceptual quality [26]. The methods show good performance but generally have high computational overhead and are not ideal for online or resource-limited systems. Unlike transformer-based approaches, the proposed D2W-GAN combines a frequency-domain representation with adversarial learning. The method highlights multi-scale feature decomposition and structural consistency, which are effective for raindrop-specific distortions, whereas transformer models excel at capturing global dependencies. The framework builds on several well-established deep learning components. The DT-CWT provides improved directional selectivity and shift invariance for multi-scale feature representation [7], [19]. The U-Net architecture, widely used for image restoration and segmentation tasks, enables hierarchical feature extraction through its encoder-decoder structure [18]. Attention mechanisms have been extensively adopted to improve feature representation by focusing on relevant regions [30]. Additionally, GANs have demonstrated strong capabilities for producing

realistic image reconstructions in image-to-image translation tasks [27].

III. METHODOLOGY

The proposed D2W-GAN is based on the common architecture of a U-Net generator, a PatchGAN discriminator, and an auxiliary mask generator, as illustrated in Fig. 1. The generator is based on the DT-CWT [Editor3.1] and is embedded in an encoder-decoder framework that enables joint spatial-frequency processing. The encoder consists of convolutional blocks with 3×3 kernels, followed by downsampling and increasingly higher feature channels. The decoder uses transposed convolutions for upsampling, with skip connections to preserve spatial details. The decoder uses attention modules to improve the focus on rain-affected areas. The discriminator is based on the PatchGAN architecture and consists of multiple convolutional layers with LeakyReLU activations that distinguish genuine and generated image patches. In addition, a lightweight mask generator is used to generate soft raindrop localization maps, based on automatically generated ground-truth masks during training. The whole model is trained adversarially to guarantee reconstruction fidelity and perceptual quality with the Adam optimizer. Table II provides a summary of the architectural configuration of the proposed D2W-GAN framework.

TABLE II

DETAILED ARCHITECTURE OF THE PROPOSED D2W-GAN

Stage	Operation	Kernel	Channels	Output Size
Input	RGB Image	–	3	$H \times W \times 3$
DT-CWT	L3-decomposition	–	12	$H/2 \times W/2$
Encoder-1	Conv+BN+ReLU	3×3	64	$H/2 \times W/2$
Encoder-2	Conv+BN+ReLU	3×3	128	$H/4 \times W/4$
Encoder-3	Conv+BN+ReLU	3×3	256	$H/8 \times W/8$
Attention	Channel+Spatial	–	256	$H/8 \times W/8$
Bottleneck	Conv Blocks	3×3	512	$H/8 \times W/8$
Decoder-1	UpConv	3×3	256	$H/4 \times W/4$
Decoder-2	UpConv	3×3	128	$H/2 \times W/2$
Decoder-3	UpConv	3×3	64	$H \times W$
Output	Restoration Head	1×1	3	$H \times W \times 3$

A. Generator and Discriminator Integration

1) *Generator*: A GAN consists of a generator and a discriminator. The generator is based on DT-CWT and a U-Net model, which is effective for multi-scale image processing [29]. DT-CWT has superior directional selectivity and approximate shift invariance compared to standard wavelet decompositions, making it suitable for complex raindrop patterns. Integrating DT-CWT within U-Net enables processing across multiple frequency bands and raindrop removal at different scales while preserving fine image details. The encoder-decoder structure of U-Net supports feature extraction at varying spatial resolutions [6]. Attention modules further improve the generator by focusing on raindrop-affected regions based on size and transparency, which improves background restoration. Dynamic attention allows adaptability across diverse input images. Skip connections preserve spatial information by transferring encoder features directly to the decoder and ensure contextual consistency. The combination of

attention mechanisms and skip connections allows simultaneous focus on global and local features, producing high-quality, artifact-free images suitable for real-world applications.

2) *Pix2Pix Discriminator*: The discriminator distinguishes between real clean images and generated raindrop-free images. Pix2Pix is a conditional GAN that was designed for image-to-image translation and evaluates whether the generator output closely resembles the ground truth. The discriminator guides the generator toward realistic outputs by penalizing artifacts while preserving texture, color, and contextual information. This training scheme ensures effective raindrop removal without compromising intrinsic image structures.

B. U-Net with Attention and DT-CWT

The generator employs a modified U-Net architecture for raindrop removal. The encoder reduces spatial resolution while learning hierarchical features, and the decoder restores the image resolution using skip connections to preserve fine details. DT-CWT was applied at the input stage to improve frequency and orientation awareness.

1) *DT-CWT Layer*: DT-CWT decomposes the input image into low- and high-frequency components at multiple scales, capturing detailed features essential for restoration tasks [12], [13]. The DT-CWT operator transforms the input image I as:

$$x' = \mathcal{W}(I) \quad (1)$$

s is the scale and o represents the directional sub-bands. Unlike conventional DWT, DT-CWT provides six directional sub-bands per scale ($\pm 15^\circ$, $\pm 45^\circ$, $\pm 75^\circ$) with improved shift invariance. Reconstruction is achieved using inverse DT-CWT. Additional implementation details are provided in the supplementary material.

2) *Encoder Block*: The encoder progressively downsamples the input image while increasing feature channels:

$$x_1, r_1 = \text{Encoder}(x') \quad (2)$$

$$x_2, r_2 = \text{Encoder}(x_1) \quad (3)$$

$$x_3, r_3 = \text{Encoder}(x_2) \quad (4)$$

where r_1, r_2 , and r_3 are skip connections.

3) *Regularization Block*: The deepest feature representation is refined as:

$$x_4 = \text{Regularization}(x_3) \quad (5)$$

4) *Decoder Block*: The decoder restores spatial resolution using skip connections:

$$x_5 = \text{Decoder}(x_4, r_3) \quad (6)$$

$$x_6 = \text{Decoder}(x_5, r_2) \quad (7)$$

$$x_7 = \text{Decoder}(x_6, r_1) \quad (8)$$

A multi-output generator produces intermediate outputs at reduced scales along with a final high-resolution output. Intermediate supervision is applied using multiscale reconstruction loss with fixed weighting coefficients ($\lambda = 1.0$), which emphasizes the final output. This coarse-to-fine strategy enables effective removal of raindrops with varying sizes and densities.

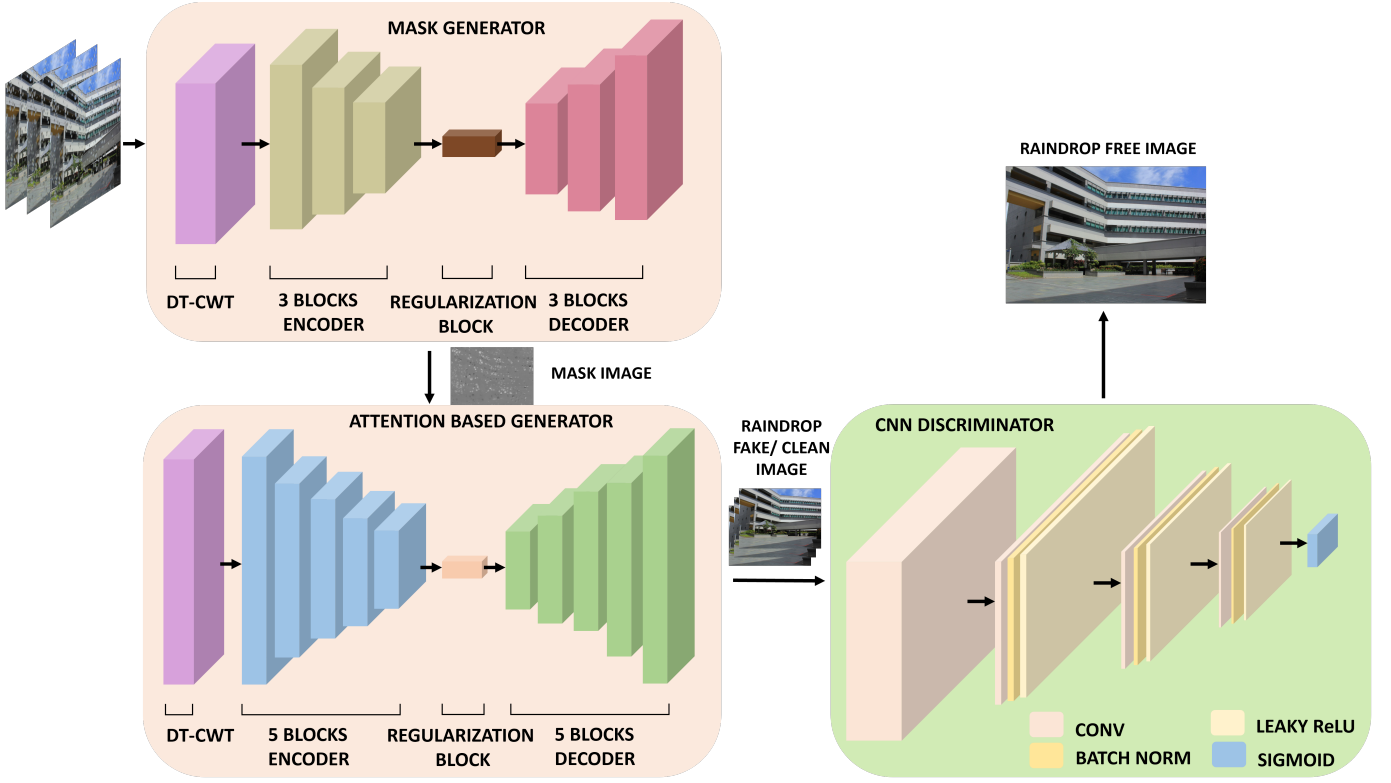


Fig. 1. Overall process of the proposed D2W-GAN using DT-CWT-based Attention GAN.

5) *Final Output*: The final output is obtained via:

$$C_{Out} = Conv(x_7) \quad (9)$$

Attention mechanisms, including spatial and channel attention, guide the network toward raindrop-contaminated regions while preserving unaffected areas. Channel attention prioritizes feature channels relevant to raindrop patterns, thereby improving restoration quality.

6) *Mask Generator for Raindrop Localization*: An auxiliary mask generator was added to help target areas affected by rain during training. The mask generator was designed as a lightweight encoder-decoder, with three convolutional layers in the encoder and three upsampling layers in the decoder. All convolutional layers are followed by ReLU activations, and the output layer uses a sigmoid activation to produce a one-channel soft mask. The ground-truth masks are automatically generated based on the differences between the image of the raindrop and the clean target image, and then automatically thresholded to produce binary mask representations. To guide the network to accurately localize regions of raindrops, the mask generator was trained with a squared error loss function and was used only during training, which means that it incurs no additional inference cost.

C. Loss Functions

The total training objective combines multiple loss terms:

$$L_D = \frac{1}{2} [\mathbb{E}_{y \sim p_{data}} \log D(y) + \mathbb{E}_{\hat{y} \sim p_{gen}} \log(1 - D(\hat{y}))] \quad (10)$$

$$L_G = \mathbb{E}_{\hat{y} \sim p_{gen}} [\log D(\hat{y})] \quad (11)$$

$$L_{multiscale} = \frac{1}{N} \sum_{i=1}^N MSE(y_{pred,i}, y_{true}) \quad (12)$$

$$L_{SSIM} = 1 - SSIM(\hat{y}, y) \quad (13)$$

$$L_{VGG} = \|\phi(\hat{y}) - \phi(y)\|_2^2 \quad (14)$$

$$L_{Mask} = MSE(m, m_{true}) \quad (15)$$

The total loss is

$$L_{Total} = L_G + L_{VGG} + L_{SSIM} + L_{multiscale} \quad (16)$$

Although the mask loss L_{Mask} was defined separately, it was used solely to train the auxiliary mask generator to provide spatial guidance. It was not included in the total loss function used to optimize the main generator network. This design ensures that the mask generator assists in learning without affecting the primary optimization objective.

D. Experimental Setup

Experiments were conducted on a system with an Intel Xeon E5-2667 v3 CPU, 32 GB of RAM, and four NVIDIA Tesla K80 GPUs. Training was performed using the Adam optimizer with a learning rate of 2×10^{-4} , $\beta_1 = 0.5$, and $\beta_2 = 0.999$.

The batch size was 4, and training was performed for 1700 epochs. The dataset used was the RainDrop dataset, which contains 861 training pairs and 58 test pairs (test set A) with an average resolution of 720×480 . Data augmentation included random cropping and flipping. The training alternated between discriminator, mask generator, and generator updates.

IV. RESULTS AND DISCUSSION

A. Evaluation of the RainDrop Dataset

An analysis of alternative approaches to removing raindrops is presented in Table III using the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Natural Image Quality Evaluator (NIQE) as metrics. NIQE is a no-reference perceptual image-quality metric that assesses the naturalness of an image using statistical analyses of unaltered natural scenes. NIQE does not depend on a reference image, unlike PSNR and SSIM. As such, it provides additional insight into the perceptual realism of restored outputs. A lower NIQE score reflects an improvement in perceptual quality. Among the assessed methods, Resfusion had good overall performance, with the highest PSNR (32.61), the second-highest SSIM (0.938), and the lowest NIQE (2.224), suggesting high perceptual quality and naturalness of the restored images. However, the proposed D2W-GAN achieved the highest SSIM (0.949), indicating the best preservation of structural information, while both PSNR (32.14) and NIQE (2.230) were also competitive. The proposed architecture has about 133 million parameters, which is much larger than those of Resfusion (7.7 million) and RainDropDiff128 (109.7 million). This complexity offsets the added computational cost while achieving better structural similarity and perceptual quality, especially in fine detail. The success of the approach based on DT-CWT is explained by its ability to perform multi-scale and multi-resolution analysis, which allows for the separation of raindrop artifacts, preserves high-frequency details and low-frequency information, and helps to preserve the image's context. The method achieves superior structural

preservation compared to lightweight models such as Pix2Pix and AttentiveGAN, as evidenced by higher SSIM scores and lower NIQE values.

Although approaches such as RainDropDiff128 and WeatherDiff64 [32] achieve high PSNR and SSIM values, they yield slightly less realistic outputs, as evidenced by the high NIQE values of Resfusion [20] and the proposed approach. The multi-scale definition provided by DT-CWT is crucial for effectively isolating raindrop distortions from the background and preserving key image characteristics. Although the proposed method does not outperform all competing methods across all metrics, it offers a balanced trade-off between reconstruction accuracy and perceptual realism. This behavior suggests that the model emphasizes structural consistency and visual coherence, particularly in regions affected by raindrops. Fig. 2 and Fig. 3 show a visual comparison of the Resfusion model and the proposed D2W-GAN framework. This figure contains the input image containing raindrops, the ground truth image, and the outputs of the two methods. The proposed method was better at removing raindrop artifacts while retaining background elements, and it minimized distortions left by other methods. Incorporating DT-CWT into the U-Net framework and using attention-based feature refinement techniques improved the retention of intricate image details in both raindrop-present and raindrop-absent areas of the scene. The image-wise PSNR, SSIM, and NIQE values are reported. The relatively small standard deviations indicate uniform performance across the different test images. The other methods produce outputs at the image level, but the information is not publicly available, and a direct paired test of the statistical significance of those outputs cannot be performed.

B. Evaluation of the RainDS Dataset

Table IV compares the various raindrop-removal techniques on the RainDS dataset, which consists of 248 image pairs of real raindrop images and clean backgrounds. 98 pairs were allocated to the test, and all the measures listed in the table were assessed to provide objective measurements in realistic imaging conditions. On the RainDS dataset, the proposed method achieved a PSNR of 29.01 and NIQE of 6.41, which are comparable to those of the best method, Resfusion. Although Resfusion achieves a higher PSNR and SSIM, the proposed method shows expected restoration performance while generating visually realistic outputs. The results show that the proposed framework generalizes to real-world raindrop images reasonably well, but more work is needed to consistently surpass state-of-the-art diffusion-based methods for all evaluation metrics. The proposed framework is just slightly worse than Refusion. To some degree, SSIM was lower than that of certain other methodologies, but it largely measures structural proximity and may fail to reflect improvements in perception in real-world raindrop-removal scenarios, such as high occlusion, complex lighting conditions, and a lack of background information. NIQE is a perceptual quality metric and was lowest for the proposed method, indicating that the visuals appear more natural. Overall, the results show that when applied to real-world raindrop-degraded images, the proposed

TABLE III

QUANTITATIVE COMPARISON ON THE RAINDROP DATASET. VALUES ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION (SD). THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULT IS UNDERLINED. THE SYMBOL " $\uparrow$ " INDICATES THAT HIGHER VALUES ARE BETTER, WHEREAS " $\downarrow$ " INDICATES THAT LOWER VALUES ARE BETTER. A DASH (" $-$ ") DENOTES THAT THE CORRESPONDING RESULT WAS NOT REPORTED

Method	Parameters	RainDrop Dataset		
		PSNR $\uparrow$	SSIM $\uparrow$	NIQE $\downarrow$
Pix2Pix [6]	–	28.02 $\pm$ 1.88	0.855 $\pm$ 0.65	2.499
AttentiveGAN [15]	6.2M	31.59 $\pm$ 1.98	0.917 $\pm$ 0.29	2.442
DuRN [10]	10.2M	31.24/–	0.926/–	–
RaindropAttn [17]	–	31.44/–	0.926/–	–
All-in-One [8]	–	31.12/–	0.927/–	–
IDT [28]	16.4M	31.87/–	0.931/–	–
WeatherDiff64 [32]	82.9M	30.71 $\pm$ 1.64	0.931 $\pm$ 0.04	–
RainDropDiff128 [32]	109.7M	<u>32.43</u> $\pm$ 1.53	0.933 $\pm$ 0.02	–
Resfusion [20]	7.7M	32.61	0.938 $\pm$ 0.02	2.22
Proposed Method	133M	32.14 $\pm$ 1.14	0.949 $\pm$ 0.01	2.23

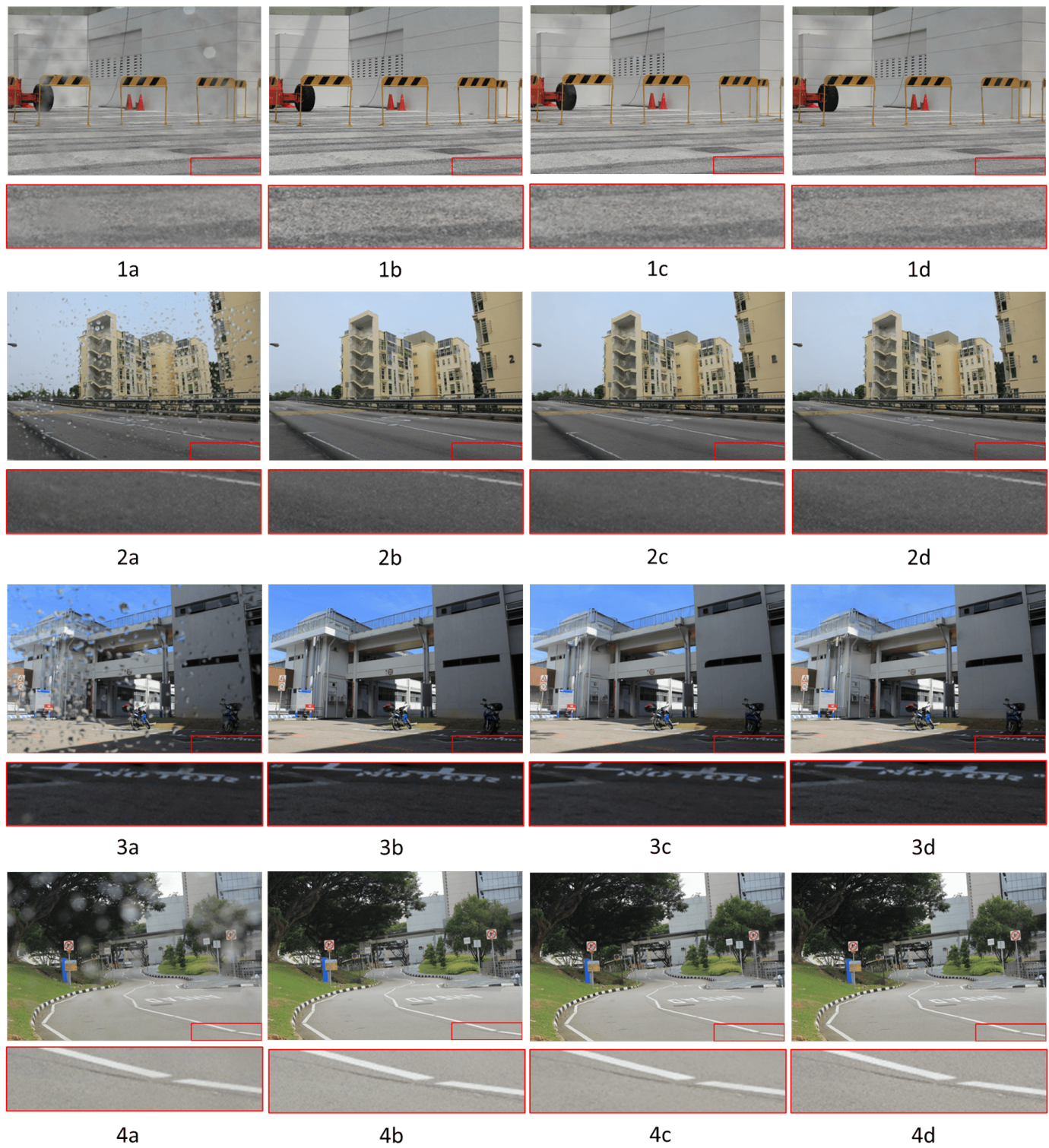


Fig. 2. Qualitative reconstruction comparisons on the RainDrop dataset for test images 1–4. From left to right: (a) Raindrop images, (b) Ground truth, (c) ResFusion, and (d) D2W-GAN.

method achieves an acceptable compromise between restoration fidelity and perception-based image quality. RainDS is a highly dense, highly irregular raindrop dataset, and its density allows for accurate structural recovery. The proposed model emphasizes perceptual naturalness and artifact removal over pixel-wise structural similarity, leading to a low SSIM score.

This shows that the model has some limitations and that under very harsh degradation conditions, fine structural consistency can be slightly affected.



Fig. 3. Qualitative reconstruction comparisons on the RainDrop dataset for test images 1–4. From left to right: (a) Raindrop images, (b) Ground truth, (c) ResFusion, and (d) D2W-GAN.

C. Impact of Raindrop Removal on Object Detection

From the perspective of downstream object-detection tasks, distortions caused by raindrops can also significantly degrade detection accuracy. Notably, there is no significant increase in the cost of computing D2W-GAN. The findings show that D2W-GAN not only improves the quality of visual images but

also provides practical benefits for autonomous driving and surveillance systems. These advantages are evident in Fig. 4. The detection performance of the rain removal was evaluated using the YOLOv7 and YOLOv7-D6 models. The detectors were not fine-tuned on the raindrop dataset and were pre-trained on the COCO dataset to make a fair assessment of



Fig. 4. Visual impact of D2W-GAN-based raindrop removal on object detection accuracy for test images 1–4. (a) Raindrop-degraded images with YOLOv7-d6 detections. (b) D2W-GAN-enhanced images with YOLOv7-d6 detections.

the restoration effects. We evaluated the method on the test subset of the RainDrop dataset (58 image pairs). Common object detection metrics were used, such as precision, recall, F1-score, and mean average precision (mAP). An Intersection over Union (IoU) threshold of 0.5 was applied, and the metric was mAP@0.5. Common classes of objects in the dataset were

used in the evaluation, including vehicles and walking pedestrians. This configuration enables the evaluation of the benefit of raindrop removal on the downstream detection performance in realistic conditions. For YOLOv7 and YOLOv7-D6, the precision metric improved by +2.4% and +11.0%, the F1 score improved by +8.0% +13.0%, respectively. Furthermore,

TABLE IV

QUANTITATIVE COMPARISON ON THE RAINDS DATASET.

VALUES ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION (SD). THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULT IS UNDERLINED. THE SYMBOL " $\uparrow$ " INDICATES THAT HIGHER VALUES ARE BETTER, WHEREAS " $\downarrow$ " INDICATES THAT LOWER VALUES ARE BETTER. A DASH ("–") DENOTES THAT THE CORRESPONDING RESULT WAS NOT REPORTED

Method	PSNR $\uparrow$	SSIM $\uparrow$	NIQE $\downarrow$
UA_GAN [1]	19.72	0.60	–
CycleGAN [31]	20.38	0.69	–
AGGAN [22]	21.54	0.75	–
IDT [28]	25.67	0.71	–
UDR-S2Former [2]	26.71 $\pm$ 1.42	0.71 $\pm$ 0.45	–
Resfusion [20]	29.42 $\pm$ 0.81	0.67 $\pm$ 0.135	6.48
Proposed Method	<u>29.01</u> $\pm$ 0.55	0.56 $\pm$ 0.130	6.41

the mAP increased almost twofold with the raindrop-removal approach for YOLOv7-D6, as shown in Table V.

TABLE V

IMPACT OF D2W-GAN-BASED RAINDROP REMOVAL ON OBJECT DETECTION PERFORMANCE

Metric	YOLOv7	YOLOv7-D6	Δ Improvement
Precision (%)	44.2 $\rightarrow$ 54.5	51.4 $\rightarrow$ 64.9	+10.3 / +13.5
Recall (%)	74.1 $\rightarrow$ 76.5	78.0 $\rightarrow$ 89.0	+2.4 / +11.0
F1 Score (%)	55.6 $\rightarrow$ 63.6	62.0 $\rightarrow$ 75.0	+8.0 / +13.0
mAP@0.5 (%)	0.1672 $\rightarrow$ 0.2789	0.2443 $\rightarrow$ 0.5078	+0.1117 / +0.2635
Avg. Inference Time (ms)	760.59 $\rightarrow$ 704.15	1115.32 $\rightarrow$ 1065.60	56.44 / 49.72

D. Computational Analysis

The computational efficiency of the proposed D2W-GAN was evaluated by measuring the number of parameters, floating-point operations (FLOPs), inference time, and memory consumption on a typical workstation with an NVIDIA RTX 3060 (12 GB VRAM), an Intel i7 CPU, and 16 GB of RAM. The proposed model contains approximately 133 million parameters and requires 479.22 GFLOPs per 720×480 image. On the specified hardware, the proposed model achieved an average inference time of 0.2766 seconds per image ($\approx$ 3.6 FPS) with peak GPU memory consumption of 2.3 GB. Compared to existing lightweight approaches, the method is computationally intensive and is better suited for applications where restoration quality is prioritized over strict real-time constraints. To achieve practical deployment, compression based on model pruning, pruning based on knowledge distillation, mixed-precision inference, or post-training quantization [9] could be used to reduce memory cost and latency without further compromising restoration quality. These techniques would make it easier to use embedded GPU platforms, such as the NVIDIA Jetson series, in medium-latency applications.

E. Ablation Study

The proposed method was evaluated through ablation studies (Table VI and Table VII) with the RainDrop dataset [25] to assess the contributions of different modules. In Experiments

1–3, the effects of attention and the discriminator module can be observed. Experiment 1 (DT-CWT + attention) showed moderate performance (PSNR 29.25, SSIM 0.785), while Experiment 2 showed higher performance (PSNR 30.60, SSIM 0.871) as a result of introducing the discriminator. Experiment 3 illustrates that the discriminator can be used on its own to increase performance (PSNR 30.14, SSIM 0.829). Experiments 4–9 examine the influence of various wavelet configurations and DT-CWT decomposition levels, with focus on the discriminator. These variations led to gradual improvements, but the ultimate combination of DT-CWT Level 3, attention, and the discriminator in the proposed method yielded the best results (PSNR 32.14, SSIM 0.949). The analysis shows that each module alone is insufficient: attention and DT-CWT can only enhance image quality, but removing the discriminator results in the largest decrease in PSNR and SSIM. The results of the ablation study indicate that DT-CWT Level 3 performs best, offering a good trade-off between retaining local details and broader contextual information. At lower decomposition levels, the network captures only limited frequency-scale information, which may be insufficient to differentiate raindrop artifacts from background texture. The third decomposition level, however, provides richer low-frequency context and tuned high-frequency subbands. This helps the model better represent raindrops with different sizes, shapes, and levels of transparency. This multi-scale representation helps to preserve edges and fine structures while reducing distortion induced by droplets. Discriminator and attention modules work best in tandem. The attention module helps the generator focus on relevant channels and areas affected by raindrops, while limiting modifications to the unaltered background. The discriminator provides adversarial guidance so that the reconstructed images resemble clean natural images. Attention, in conjunction with the discriminator, provides adequate spatial selectivity and visual realism while also helping to minimize blurred reconstruction artifacts. This explains the results achieved by the configuration that combines DT-CWT Level 3, attention, and a discriminator, compared to the other configurations in the ablation study.

F. Limitations

Although the proposed method has shown considerable structural preservation on the RainDrop dataset, extremely dense or overlapping raindrops can obscure significant portions of the scene and occasionally cause mild over-smoothing or texture inconsistencies. Very large or irregular droplets and scenes with strong reflections, low light, or backlighting may leave faint artifacts, slight color deviations, or partial blurring. Thin structures and fine high-frequency details, such as wires or text, can be softened when partially occluded. Training the model on a particular RGB dataset can also lead to lower performance in other domains, and the model's relatively large number of parameters can be a limitation for deployment in low-resource settings. These limitations can be eliminated by incorporating low-light enhancement features, applying adaptive exposure normalization algorithms, and training models on a larger ensemble of raindrop and lighting conditions.

TABLE VI
ABLATION STUDY OF THE PROPOSED D2W-GAN ON THE RAINDROP DATASET

Exp.	DWT (db1)	DWT (db3)	DT-CWT L1	DT-CWT L2	DT-CWT L3	Attention	Discriminator	PSNR $\uparrow$	SSIM $\uparrow$
1	×	×	×	×	×	✓	×	29.25	0.785
2	×	×	×	×	×	✓	✓	30.60	0.873
3	×	×	×	×	×	×	✓	30.14	0.829
4	×	✓	×	×	×	✓	✓	30.19	0.896
5	×	✓	×	×	×	×	✓	30.04	0.881
6	×	×	✓	×	×	✓	✓	30.92	0.893
7	×	×	×	✓	×	✓	✓	31.04	0.914
8	✓	×	×	×	×	✓	✓	28.04	0.830
9	✓	×	×	×	×	×	✓	26.55	0.825
D2W-GAN	×	×	×	×	✓	✓	✓	32.14	0.949

TABLE VII
LOSS-BASED ABLATION STUDY ON THE RAINDROP DATASET

L_G	L_{VGG}	L_{SSIM}	L_{Multi}	PSNR / SSIM
✓	✓	×	×	29.10 / 0.894
✓	×	✓	×	31.11 / 0.928
✓	×	×	✓	29.23 / 0.884
×	✓	✓	×	29.78 / 0.906
✓	✓	✓	✓	32.14 / 0.949

While the proposed method has shown considerable structural preservation on the RainDrop dataset, it does not have a superior edge in terms of PSNR or SSIM on several other benchmark datasets. Efforts to improve performance on other datasets would mainly focus on developing more advanced, less computationally intensive, frequency-adaptive diffusion and transformer architectures, as well as on improving overall reconstruction performance while preserving the structural advantages of the proposed framework. Another drawback is that many methods do not publicly provide output data for comparative analysis. Therefore, many of the methods could not undergo statistical significance testing due to insufficient data. In the future, we will evaluate state-of-the-art methods within a common experimental setup to perform more rigorous paired statistical tests. D2W-GAN offers better structural preservation, and lightweight diffusion models (such as Resfusion) better manage the trade-off between increased model size and improvement in PSNR. Future work will involve designing compact wavelet-based architectures that maintain restoration quality while reducing computational cost.

V. CONCLUSION AND FUTURE WORK

This paper proposes a wavelet-driven D2W-GAN framework with attention mechanisms to remove raindrops. The comparison of alternative raindrop removal methods suggests that both Resfusion and the proposed task-specific D2W-GAN method provide an acceptable trade-off between image quality and perceptual realism. Experiments demonstrated the balanced performance of the proposed method when compared to the current approaches. It does not consistently outperform all baselines in terms of PSNR and NIQE, but it shows strong SSIM scores, reflecting its good structural preservation. By incorporating attention-guided adversarial learning, DT-CWT can better manage raindrop distortions at multiple scales. The method has high computational overhead, but it provides

good perceptual quality and reconstruction fidelity. Reducing model complexity and improving generalization to a variety of real-world situations will be the focus of future work. Overall, Resfusion and D2W-GAN (when combined with DT-CWT) achieve balanced image restoration accuracy, realism preservation, and structural similarity. Future applications may consider minimizing the computational needs, especially in high-resolution or real-time applications. The methodology can also be further developed to improve thermal cameras' capabilities for removing long-range raindrops during unfavorable weather conditions, in able visibility and detection accuracy are important. It might also be suitable to modify the framework for other image cleaning applications, such as dehazing and defogging, by further fine-tuning the network parameters to expand its scope. Incorporating image fusion techniques could significantly improve the proposed method's efficacy by combining complementary image data to achieve more reliable restoration results. Although the suggested framework focuses on visual quality and structural consistency, it is more computationally complex than other lightweight solutions. In addition to removing raindrops, frequency-conscious representations and multi-scale learning could be used for related image restoration and inspection tasks, including image dehazing and industrial defect detection.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are publicly available on Figshare at <https://doi.org/10.6084/m9.figshare.28433225.v2>. The repository includes the experimental results, simulation outputs, and evaluation data used in this work. The source code and evaluation scripts required to reproduce the proposed D2W-GAN framework are publicly available at <https://github.com/MuthukumarBT/D2W-GAN>.

REFERENCES

- [1] Y. Alami Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 31, 2018. DOI: 10.48550/arXiv.1806.02311.
- [2] S. Chen, T. Ye, J. Bai, E. Chen, J. Shi, and L. Zhu, "Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 106–13 117. DOI: 10.1109/ICCV51070.2023.01205.

- [3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. DOI: 10.48550/arXiv.2006.11239.
- [4] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022. DOI: 10.48550/arXiv.2106.15282.
- [5] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lv, and S. Chen, "Wavedm: Wavelet-based diffusion models for image restoration," *IEEE Transactions on Multimedia*, vol. 26, pp. 7058–7073, 2024. DOI: 10.1109/TMM.2024.3359769.
- [6] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134. DOI: 10.1109/CVPR.2017.632.
- [7] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, 2001. DOI: 10.1006/acha.2000.0343.
- [8] R. Li, R. T. Tan, and L. F. Cheong, "All in one bad weather removal using architectural search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3175–3185. DOI: 10.1109/CVPR42600.2020.00324.
- [9] Z. Li and Q. Gu, "I-vit: Integer-only quantization for efficient vision transformer inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 065–17 075. DOI: 10.1109/ICCV51070.2023.01565.
- [10] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7007–7016. DOI: 10.1109/CVPR.2019.00717.
- [11] J. Nie, J. Xie, J. Cao, and Y. Pang, "Context and detail interaction network for stereo rain streak and raindrop removal," *Neural Networks*, vol. 166, pp. 215–224, 2023. DOI: 10.1016/j.neunet.2023.07.013.
- [12] S. I. Niwas, P. Palanisamy, and K. Sujathan, "Complex wavelet as nucleus descriptors for automated cancer cytology classifier system using ann," in *2010 IEEE International Conference on Computational Intelligence and Computing Research*, 2010, pp. 1–5. DOI: 10.1109/IC-CIC.2010.5705851.
- [13] S. I. Niwas, P. Palanisamy, K. Sujathan, and E. Bengtsson, "Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using complex daubechies wavelets," *Signal Processing*, vol. 93, no. 10, pp. 2828–2837, 2013. DOI: 10.1016/j.sigpro.2012.06.029.
- [14] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. S. Khan, "Promptir: Prompting for all-in-one image restoration," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71 275–71 293, 2023. DOI: 10.52202/075280-3121.
- [15] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2482–2491. DOI: 10.1109/CVPR.2018.00263.
- [16] Y. Qiao, M. Shao, H. Liu, and K. Shang, "Mutual channel prior guided dual-domain interaction network for single image raindrop removal," *Computers & Graphics*, vol. 112, pp. 132–142, 2023. DOI: 10.1016/j.cag.2023.04.006.
- [17] Y. Quan, S. Deng, Y. Chen, and H. Ji, "Deep learning for seeing through window with raindrops," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2463–2471. DOI: 10.1109/ICCV.2019.00255.
- [18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.
- [19] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005. DOI: 10.1109/MSP.2005.1550194.
- [20] Z. Shi, H. Zheng, C. Xu, C. Dong, B. Pan, A. He, T. Li, and H. Fu, "Resfusion: Denoising diffusion probabilistic models for image restoration based on prior residual noise," in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2023, pp. 2773–2783. DOI: 10.52202/079017-4153.
- [21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020. DOI: 10.48550/arXiv.2011.13456.
- [22] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8. DOI: 10.48550/arXiv.1903.12296.
- [23] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780. DOI: 10.48550/arXiv.2201.02973.
- [24] H. Wang, S. Zhao, L. Yu, and X. Yang, "Unsupervised network for single image raindrop removal," in *Computer Vision and Pattern Recognition*, 2024. DOI: 10.48550/arXiv.2412.03019.
- [25] T. Wang, K. Zhang, Z. Shao, W. Luo, B. Stenger, T. Lu, T. K. Kim, W. Liu, and H. Li, "Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions," *International Journal of Computer Vision*, pp. 1–23, 2024. DOI: 10.1007/s11263-024-02056-0.
- [26] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13 095–13 105. DOI: 10.48550/arXiv.2303.09472.
- [27] H. Xia, Y. Lan, S. Song, and H. Li, "Raindrop removal from a single image using a two-step generative adversarial network," *Signal Image and Video Processing*, vol. 16, no. 3, pp. 677–684, 2022. DOI: 10.1007/s11760-021-02007-z.
- [28] J. Xiao, X. Fu, A. Liu, F. Wu, and Z. J. Zha, "Image de-raining transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12978–12995, 2022. DOI: 10.1109/TPAMI.2022.3183612.
- [29] C. Xu, J. Gao, Q. Wen, and B. Wang, "Generative adversarial network for image raindrop removal of transmission line based on unmanned aerial vehicle inspection," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 6668771, 2021. DOI: 10.1155/2021/6668771.
- [30] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739. DOI: 10.48550/arXiv.2111.09881.
- [31] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232. DOI: 10.1109/ICCV.2017.244.
- [32] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 346–10 357, 2023. DOI: 10.1109/TPAMI.2023.3238179.



and advanced driver assistance systems.

Muthukumar Balamurugan received the B.E. degree from Banari Amman Institute of Technology, India, and the M.E. degree from Anna University, India. He is currently pursuing a Ph.D. degree at the National Institute of Technology Tiruchirappalli (NIT Trichy), India. He is presently working as a Senior Technical Architect in the Advanced Driver Assistance Systems (ADAS) Division at Continental Autonomous Mobility (India) Pvt. Ltd., Bengaluru–560100, India. His research interests include computer vision, machine learning, deep learning,



University, Gwangju, South Korea. He is currently working as a Senior Technical Architect at AUMOVIO Autonomous Mobility India Private Limited, Bengaluru, Karnataka, India. His research interests include computer vision, video processing, deep learning, and intelligent automotive systems.

Shivarama Holla K. received the M.Sc. degree in Computer Science from Mangaluru University, Mangaluru, Karnataka, India, in 2015, and the Ph.D. degree from the Department of Mathematical and Computational Sciences, National Institute of Technology Karnataka, Surathkal, India, in 2018. He worked as a Senior Research and Development Engineer at Continental Automotive Components India Pvt. Ltd., Bengaluru, India, from 2018 to 2020. He subsequently worked as a postdoctoral fellow in image and video reconstruction at Chosun



Varun P. Gopi received the B.Tech. degree in Electronics and Communication Engineering from Amal Jyothi College of Engineering, Kanjirappally, India, in 2007, and the M.Tech. degree from the College of Engineering Trivandrum, Kerala, India, in 2009. He obtained the Ph.D. degree in Electronics and Communication Engineering from the National Institute of Technology, Tiruchirappalli, India, in 2014. He is currently an Associate Professor in the Department of Electronics and Communication Engineering at the National Institute of Technology

Tiruchirappalli, India. He is a Senior Member of IEEE. His research interests include artificial intelligence, machine learning, signal and image processing, medical image analysis, biomedical signal processing, and edge-AI systems for healthcare applications.