CFMW: Cross-modality Fusion Mamba for Multispectral Object Detection under Adverse Weather Conditions

Haoyuan Li[1,*], Qi Hu[1,*], You Yao[2], Kailun Yang[3,†], and Peng Chen[1,†]
[1]Zhejiang University of Technology, [2]Google, [3]Karlsruhe Institute of Technology

ABSTRACT

Cross-modality images that integrate visible-infrared spectra cues can provide richer complementary information for object detection. Despite this, existing visible-infrared object detection methods severely degrade in severe weather conditions. This failure stems from the pronounced sensitivity of visible images to environmental perturbations, such as rain, haze, and snow, which frequently cause false negatives and false positives in detection. To address this issue, we introduce a novel and challenging task, termed visible-infrared object detection under adverse weather conditions. To foster this task, we have constructed a new Severe Weather Visible-Infrared Dataset (SWVID) with diverse severe weather scenes. Furthermore, we introduce the Cross-modality Fusion Mamba with Weather-removal (CFMW) to augment detection accuracy in adverse weather conditions. Thanks to the proposed Weather Removal Diffusion Model (WRDM) and Cross-modality Fusion Mamba (CFM) modules, CFMW is able to mine more essential information of pedestrian features in cross-modality fusion, thus could transfer to other rarer scenarios with high efficiency and has adequate availability on those platforms with low computing power. To the best of our knowledge, this is the first study that targeted improvement and integrated both Diffusion and Mamba modules in cross-modality object detection, successfully expanding the practical application of this type of model with its higher accuracy and more advanced architecture. Extensive experiments on both well-recognized and self-created datasets conclusively demonstrate that our CFMW achieves state-of-the-art detection performance, surpassing existing benchmarks. The dataset and source code will be made publicly available at https://github.com/lhy-zjut/CFMW.

KEYWORDS

Cross-modality object detection, Image restoration, Denoising Diffusion models, Mamba, Pedestrian detection

1 INTRODUCTION

In an open and dynamic environment, object detection faces challenging weather conditions such as rain, haze, and snow. The rapid advancement of deep-learning-based object detection methods has significantly improved the ability to identify and classify objects. Benefiting from the advanced feature extraction and fusion strategies, cross-modality object detection methods have achieved high accuracy, *e.g.*, CFT [34], GAFF [56], and CFR_3 [54]. However, as shown in Fig. 1, the performance of these methods is often challenged by adverse weather conditions, which can severely impact the visibility and quality of visual data. Although the infrared image could provide complementary cues to some extent, it cannot repair the appearance distortion or information loss of visual images. Thus, traditional cross-modality object detection methods still face severe performance degradation under adverse weather.

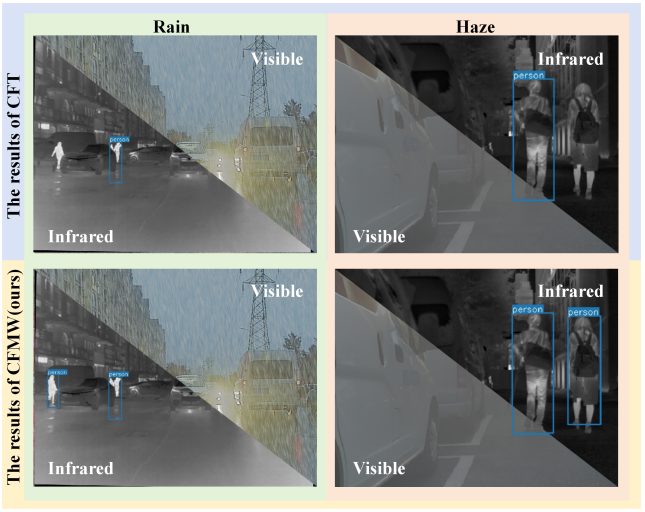

Figure 1: The proposed method can achieve high-precision cross-modality object detection under adverse weather conditions. The top two examples are results from CFT [34], while the bottom two examples are results from CFMW (ours).

Existing methods cannot be directly applied to adverse weather conditions, since the color gamut of visible images is weakened by environmental disturbance and the existing fusion methods are difficult to fully fuse visible and infrared spectra, nor have they made sufficient training under corresponding datasets. To make up the blank in this research area, we construct and release a new dataset, named Severe Weather Visible-Infrared Dataset (SWVID), as well as propose a novel framework named Cross-modality Fusion Mamba with Weather-removal (CFMW).

To facilitate research in this area, we propose a new visible-infrared dataset, named SWVID, which is designed to encompass diverse severe weather scenarios by mathematically formalizing the impact of various weather phenomena on images. Specifically, SWVID comprises 20,000 aligned visible-infrared image pairs, spanning three weather conditions and two scenes, with each condition and scene evenly distributed. Motivated by the critical research gap highlighted in Fig. 1, where current methods falter in adverse weather, we introduce CFMW for multispectral object detection under adverse weather conditions. Our CFMW leverages a Weather Removal Diffusion Model (WRDM) and Cross-modality Fusion Mamba (CFM) to enhance detection accuracy amid adverse weather

*Equal contribution. †Corresponding authors (e-mail: chenpeng@zjut.edu.cn, kailun.yang@kit.edu).

conditions while minimizing computational burden. Specifically, WRDM is employed to restore affected visible images before fusion with infrared counterparts, offering plug-and-play compatibility with image fusion networks. Based on learning reversal to increase the order of noise and disrupt the process of data samples, the WRDM model is advantageous to minimize the impact of adverse weather conditions. Additionally, CFM can be integrated into the feature extraction backbone, effectively integrating global contextual information from diverse modalities. Recent research shows that Mamba [10] achieves higher inference speed and overall metrics than the equivalent-scale transformer. To our knowledge, this study represents the first endeavor to employ Diffusion models and Mamba for multispectral object detection.

Extensive experiments on both well-established and self-created datasets demonstrate that our CFMW method achieves superior detection performance compared to existing benchmarks. Specifically, we achieved about 17% performance improvement compared with the current state-of-the-art image restoration methods. The proposed method achieves about 8% accuracy improvement while saving 51.2% GPU memory compared with CFT [34], a state-of-the-art cross-modality object detection method.

At a glance, we summarize the main contributions as follows:

- We introduce a novel task focusing on visible-infrared object detection under adverse weather conditions and develop a new dataset called the Severe Weather Visible-Infrared Dataset (SWVID), which simulates real-world conditions. SWVID comprises 60, 000 paired visible-infrared images and labels, encompassing weather conditions such as rain, haze, and snow;
- We propose a novel approach, Cross-modality Fusion Mamba with Weather-removal (CFMW) for multispectral object detection under adverse weather conditions;
- We introduce a novel Weather Removal Diffusion Model (WRDM) and Cross-modality Fusion Mamba (CFM) modules to tackle image de-weathering and visible-infrared object detection tasks simultaneously;
- Extensive experiments demonstrate that this integration achieves the best task migration capacity, resulting in state-of-the-art performance for both tasks.

2 RELATED WORK

In this section, we briefly review previous related works about cross-modality object detection, state space model, and multi-weather image restoration.

Cross-modality Object Detection The existing cross-modality object detection methods can be divided into two categories: feature level and pixel level fusion, distinguished through feature fusion methods and timing. Recently, dual stream object detection models based on convolutional neural networks have made great progress in improving recognition performance [4, 34, 37, 54, 55], while pixel level fusion methods have also achieved good performance [5, 44, 59]. Other works employing methods such as GAN to effective integration also have achieved good results [51, 58, 59]. Those works can be integrated into downstream tasks such as object detection. Traditional convolutional neural networks have limited receptive fields that the information is only integrated into a local area when using the convolution operator, where the self-attention operator of the transformer can learn long-range dependencies [43]. Thus, a transformer-based method, named Cross-Modality Fusion Transformer (CFT) [34], was presented and achieved state-of-the-art detection performance. Differing from these works, we first introduce Mamba into cross-modality object detection to learn long-range dependencies with gating mechanisms, achieving high accuracy and low computation overhead simultaneously.

State Space Model The concept of the State Space Model was initially introduced in the S4 model [11], presenting a distinctive architecture capable of effectively modeling global information, compared with traditional convolutional neural networks and transformers. Based on S4, the S5 model [38] reduces complexity to a linear level, with H3 [31] introducing it into language model tasks. Mamba [10] introduced an input-activate mechanism to enhance the State Space model, achieving higher inference speed and overall metrics compared with equivalent-scale transformers. With the introduction of Vision Mamba [61] and Vmamba [30], the application of the State Space Model has been extended into visual tasks. Currently, existing research does not consider effectively generalizing the State Space Model to cross-modality object detection.

Multi-Weather Image Restoration Recently, some attempts have been made to unity multiple recovery tasks in a single deep learning framework, including generating modeling solutions to recover superimposed noise types [9], recovering superimposed noise or weather damage with unknown test time, or especially unfavorable multi-weather image fading [3, 22, 42]. All in One [23] unified a weather restoration method with a multi-encoder and decoder architecture. It is worth noting that diffusion-based conditional generative models have shown state-of-the-art performance in various tasks such as class-conditional data synthesis with classifier guidance [7], image super-resolution [14], image deblurring [48]. Denosing diffusion restoration models (DDRM) [21] were proposed for general linear inverse image restoration problems, exploiting pro-trained denoising diffusion models for unsupervised posterior sampling. Generally, diffusion models were so far not considered to be generalized to adverse weather scenes in the cross-modality image fusion field. Unlike existing works, we expand the multi-weather restoration to the field of cross-modality fusion.

3 PROPOSED FRAMEWORK

3.1 Overview

As shown in Fig. 2, CFMW comprises two main stages. In the multi-weather image restoration stage, we aim to achieve image restoration of three types of adverse weather conditions (rain, snow, and haze) and implement it using a unified framework with only one pre-trained weight. In the cross-modality fusion stage, we aim to integrate unique features of different modalities. Inspired by CFT [34], to show the effectiveness of our proposed CFM fusion model, we extend the framework of YOLOv5 to enable multispectral object detection. We present our carefully designed loss functions and training procedure for WRDM and CFM in the last subsection.

3.2 Weather Removal Diffusion Model (WRDM)

Denoising diffusion models [13, 39] are a class of generative models, that learn a Markov chain that gradually transforms a Gaussian

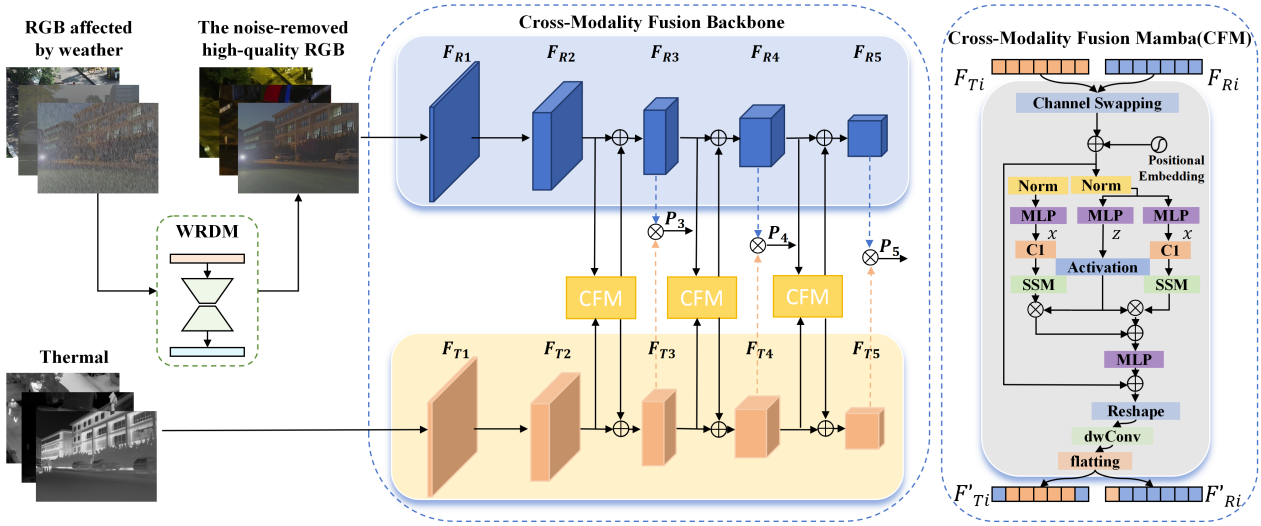

Figure 2: Framework of Cross-Modality Fusion Mamba backbone. It has three parts: a Weather Removal Diffusion Model (WRDM), a two-stream feature extraction network (our baseline), and three Cross-Modality Fusion Mamba (CFM) modules. $\oplus$ represents element-wise add, $\otimes$ represents element-wise multiply, and C1 is short of 1-dimension convolutions.

noise distribution into the data distribution trained by the models. The original denoising diffusion probabilistic models (DDPMs)[13] diffusion process (data to noise) and generative process (noise to data) are based on a Markov chain process, resulting in a large number of steps and huge time consumption. Thus, denoising diffusion implicit models (DDIMs) [40] were presented to accelerate sampling, providing a more efficient class of iterative implicit probabilistic models. DDIMs define the generative process via a class of non-Markovian diffusion processes that lead to the same training objective as DDPMs but can produce deterministic generative processes, thus speeding up sample generation. In DDIMs, implicit sampling refers to the generation of samples from the latent space of the model in a deterministic manner. Implicit sampling using a noise estimator network can be performed by:

$$X_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \left(\frac{X_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(X_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(X_t, t). \tag{1}$$

where X_t and X_{t-1} represent the data $X_0 \sim q(X_0))$ in different diffusion time steps, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon_\theta(X_t, t)$ can be optimized as: $\mathbb{E}_{X_0,t}, \epsilon_t \sim N(\mathbf{0}, \boldsymbol{I}), [\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t\|^2]$.

Conditional diffusion models have shown state-of-the-art image-conditional data synthesis and editing capabilities [6, 7]. The core idea is to learn a conditional reverse process without changing the diffusion process. Our proposed WRDM is a conditional diffusion model, adding reference images (clear images) in the process of sampling to guide the reconstructed image to be similar to reference images. As shown in Fig. 3, we introduce a new parameter $\widetilde{X}$, which represents the weather-degraded observation. A Markov chain is defined as a diffusion process, and Gaussian noise is gradually added to simulate the gradual degradation of data samples until reaching time point T. We ground our model hyper-parameters via a U-Net architecture based on WideResNet [52]. For the input images conditional reflection, we connect patch x_T and $\widetilde{x}$, to obtain the six-dimensional input image channel. Conditioning the reverse process on $\widetilde{X}$ can maintain its compatibility with implicit sampling, so we could expand Eq. (1) as:

$$X_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \left(\frac{X_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(X_t, \widetilde{X}, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(X_t, \widetilde{X}, t). \tag{2}$$

The sampling process starts from $X_T \sim N(\mathbf{0}, \boldsymbol{I})$, following a deterministic reverse path towards X_0 with fidelity. See more derivation details in the supplementary material.

Our proposed WRDM is a patch-based conditional diffusion model, guiding the reverse sampling process toward smoothness across neighboring patches. During training, we randomly sample the pxp patch location for P_i within the compute of image dimensions. Under any given time step T, we reverse-sample the average estimated noise of each pixel in the overlapping patch area according to Fig. 3, which effectively controls the reverse sampling process to ensure that all adjacent patches have higher fidelity.

Furthermore, WRDM can be regarded as a plug-in, embedded into other works such as visible-infrared image fusion to remove the influence of multi-weather conditions, which is demonstrated experimentally in Fig. 5.

3.3 Cross-modality Fusion Mamba (CFM)

The goal of Cross-modality Fusion Mamba (CFM) is to introduce the advanced state space model (SSM), or Mamba [10], to cross-modality object detection. Structured state space sequence models (S4) and Mamba are inspired by the continuous system, mapping a 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t)$ through a hidden

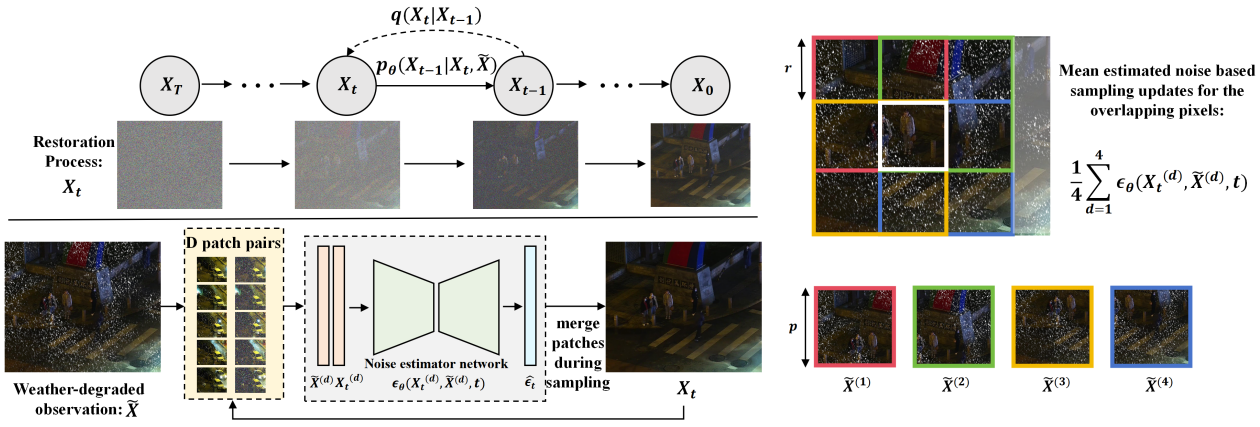

Figure 3: Schematic diagram of WRDM training and reasoning process. The left side is the framework of WRDM. We use a paired data distribution $(\widetilde{X}, X_t)$, splitting into $(\widetilde{X}^{(d)}, X_t^{(d)})$ for model-training. The right side is the illustration of the patch-based diffusive image restoration pipeline (4 patches for example here).

state $h(t) \in \mathbb{R}^N$. This system uses $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ as the evolution parameter and $\boldsymbol{B} \in \mathbb{R}^{N\times 1}, \boldsymbol{C} \in \mathbb{R}^{1\times N}$ as the projection parameters, so that $y(t)$ could evolve as follows:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h'(t). \end{aligned} \tag{3}$$

Notice that S4 and Mamba are the discrete versions of the continuous system, including a timescale parameter Δ to transform the continuous parameters A, B to discrete parameters $\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}$ as follows:

$$\begin{aligned} \bar{\boldsymbol{A}} &= exp(\Delta \boldsymbol{A}), \\ \bar{\boldsymbol{B}} &= (\Delta \boldsymbol{A})^{-1}(exp(\Delta \boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \boldsymbol{B}. \end{aligned} \tag{4}$$

After that, Eq. (3) could be rewritten as:

$$\begin{aligned} h_t &= \bar{\boldsymbol{A}}h_{t-1} + \bar{\boldsymbol{B}}x_t, \\ y_t &= \boldsymbol{C}h_t. \end{aligned} \tag{5}$$

Finally, the models compute output through a global convolution as follows:

$$\begin{aligned} \bar{\boldsymbol{K}} &= \boldsymbol{C}\bar{\boldsymbol{B}}, \boldsymbol{C}\bar{\boldsymbol{A}}\bar{\boldsymbol{B}}, ..., \boldsymbol{C}\bar{\boldsymbol{A}}^{\boldsymbol{M-1}}\bar{\boldsymbol{B}}, \\ y &= x * \bar{\boldsymbol{K}}. \end{aligned} \tag{6}$$

where $\boldsymbol{M}$ is the length of the input sequence x, and $\bar{\boldsymbol{K}} \in \mathbb{R}^M$ is a structured convolution kernel.

Standard Mamba is designed for the 1-D sequence. As shown in Vision Mamba (Vim), 2-D multispectral images $t \in \mathbb{R}^{H\times W\times C}$ could be transformed into the flattened 2-D patches $x_p \in \mathbb{R}^{J\times(P^2\cdot C)}$, where (H, W) represents the size of input images, C is the channels, and P is the size of image patches. Similarly, we linearly project the x_p to the vector with size D and add position embeddings $\boldsymbol{E}_{pos} \in \mathbb{R}^{(J+1)\times D}$ as follows:

$$\boldsymbol{T}_0 = [t_{cls}; t_p^1\boldsymbol{W}; t_p^2\boldsymbol{W}; ...; t_p^J\boldsymbol{W}] + \boldsymbol{E}_{pos}. \tag{7}$$

where t_p^j is the $j-th$ path of $\boldsymbol{t}$, $\boldsymbol{W} \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the learnable projection matrix.

Here are more details of the proposed CFM. As mentioned in the introduction section, the RGB modality and the Thermal modality show different features under different lighting and weather conditions, which are complementary and redundant. Therefore, we aim to design a block to suppress redundant features and fuse complementary to efficiently harvest essential cross-modal cues for object detection against adverse weather conditions. Motivated by the concept of Cross-Attention [1], we introduce a new cross-modality Mamba block to fuse features from different modalities. As shown in Fig. 2, to encourage feature interaction between RGB and Thermal modalities, we use a Channel Swapping Mamba block (CS) [12], which incorporates information from different channels and enhances cross-modality correlations. Given RGB features F_{R_i} and Thermal features F_{T_i}, the first half of channels from F_{R_i} will be concatenated with the latter half of F_{T_i} and processed through the Mamba block for feature extraction. The obtained features are added to F_{R_i}, creating a new feature $F_{R_i}{}'$. Meanwhile, the first half of F_{T_i} is concatenated with the latter half of F_{R_i}, then passes through the Mamba block. The obtained features are added to F_{T_i}, creating a new feature $F_{T_i}{}'$.

Subsequently, we project the features: $F_{R_i}{}'$ and $F_{T_i}{}'$ into the shared space during the feature fusion process, using the gating mechanism to encourage complementary feature learning while restraining redundant features. As shown in Fig. 2, we first normalize every token sequence in $F_{R_i}{}'$ and $F_{T_i}{}'$ with **Norm** block, which helps to improve the convergence speed and performance of the model. Then project the input sequence through linear layers and apply SiLu as the activation function. $\bar{\boldsymbol{A}}_{\boldsymbol{o}}, \bar{\boldsymbol{B}}_{\boldsymbol{o}}$, and $\boldsymbol{C}_{\boldsymbol{o}}$ can be generated by the Parameters Function:

$$\bar{\boldsymbol{A}}_{\boldsymbol{o}}, \bar{\boldsymbol{B}}_{\boldsymbol{o}}, \boldsymbol{C}_{\boldsymbol{o}} = ParametersFunction(x_o{}'), \tag{8}$$

where $x_o' = \boldsymbol{Linear}(x_o^x \boldsymbol{Norm}(F_i^{o'}))$. After that, we apply State Space Model **(SSM)**:

$$y_o = \boldsymbol{SSM}(\bar{\boldsymbol{A}}_{\boldsymbol{o}}, \bar{\boldsymbol{B}}_{\boldsymbol{o}}, \boldsymbol{C}_{\boldsymbol{o}})(x_o{}'), \tag{9}$$

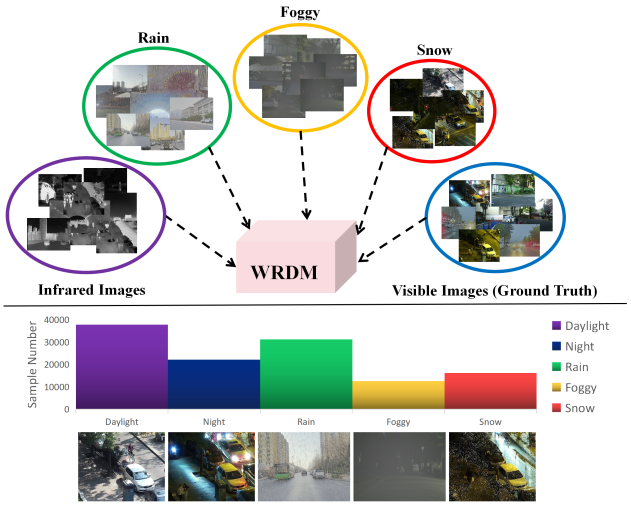

Figure 4: Overview of the established SWVID benchmarks. The dataset includes three weather conditions (*i.e.*, Rain, Foggy, and Snow), and two scenarios (*i.e.*, Daylight and Night), providing 60, 000 images in total.

Then we apply the gating operation, followed by residual connection:

$$z = \textbf{\textit{Linear}}^{z}(F_{T_i}{}'), \tag{10}$$

$$y_R{}' = y_R \odot \textbf{\textit{SiLU}}(z), \tag{11}$$

$$y_T{}' = y_T \odot \textbf{\textit{SiLU}}(z), \tag{12}$$

$$F_i = \textbf{\textit{Reshape}}(\textbf{\textit{Linear}}^{T}(y_R{}' + y_T{}') + F_i{}'). \tag{13}$$

Finally, we get the fused 2-D feature F_i successfully.

Different from CFT [34], our fusion block improves computational efficiency while inheriting the components of global receptive field and dynamic weight. Comparing the state space model (SSM) in our CFM block with the self-attention mechanism of transformers in CFT [34], both of them play an important role in providing global context adaptively, but self-attention is quadratic to sequence length while SSM is linear to sequence length [61]. To achieve lower memory usage when dealing with long-sequence works, CFM chooses the recomputation method as the same as Mamba. Experiment on the SWVID and LLVIP dataset, whose resolution is 1080×720, shows that CFT requires 21.88GB GPU memory while CFM only requires 10.72GB, saving 11.16GB in the same configuration.

3.4 Loss Functions

As a two-stage pre-training model, we carefully design the training loss functions to produce enhanced results with minimum blurriness and the closest details to ground-truth images and to extract the differences between RGB and thermal modalities.

For training WRDM, the goal of the loss function in this stage is to maximize the data log-likelihood $log_{p_\theta(x_0)}$. Since maximizing this target directly is very challenging, we use variational inference to approximate this target. Variational inference approximates the true posterior distribution $p_\theta(x_0 : T)$ by introducing a variational distribution $q(x_1 : T|x_0)$ and then minimizes the difference between these two distributions. Here we define $\mathcal{L}_\theta = -log_{p_\theta}(x_0)$, we have:

$$\mathcal{L}_\theta = \sum_{t=1}^{T} \mathbb{E}_q[log_{p_\theta}(x_0|x_T)] - \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t-1}|x_t)} [D_{KL}(q(x_{t-1}|x_t, x_0))||p_\theta(x_{t-1}|x_t)]. \tag{14}$$

where the second term is the expected value of the Kullback-Leibler divergence between $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$.

In alignment with the prevalent practices in this field, the overall loss function ($\mathcal{L}_{total}$) is a sum of the bounding-box regression loss ($\mathcal{L}_{box}$), the classification loss ($\mathcal{L}_{cls}$), and the confidence loss ($\mathcal{L}_{conf} = \mathcal{L}_{noobj} + \mathcal{L}_{obj}$).

$$\mathcal{L}_{total} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{noobj} + \mathcal{L}_{obj}, \tag{15}$$

Details of the loss function for CFMW are elucidated in the supplementary material.

Table 1: Comparisons of SWVID benchmark with existing visible-infrared datasets. ✓means available while ✗denotes the opposite.

Dataset	Year	Resolution	Publication	Scene		
				Daylight	Night	Weather
KAIST [16]	2015	640 × 512	CVPR	✓	✓	✗
FLIR [8]	2018	640 × 512	-	✓	✓	✗
RoadScene [50]	2020	640 × 512	AAAI	✓	✓	✗
LLVIP [18]	2021	1080 × 720	ICCV	✓	✓	✗
MSRS [41]	2022	640 × 480	Info. Fusion	✓	✓	✗
M3FD [27]	2022	640 × 512	CVPR	✓	✓	✗
VTUAV [32]	2022	1920 × 1080	CVPR	✓	✓	✗
SWVID	2024	1080 × 720	**Proposed**	✓	✓	✓

4 EXPERIMENTS

4.1 Established SWVID benchmark

Dataset. The color gamut of visible images is weakened by environmental disturbance in dynamic environments, and the existing fusion methods make it difficult to fully fuse visible and infrared spectra because of a deficiency of sufficient training under corresponding datasets. As shown in Fig. 4, we established the benchmark, SWVID, which is constructed from the public datasets (*i.e.* LLVIP [18], M3FD [27], MSRS [41]) collected in the real scene. It contains a variety of uniformly distributed scenes (daylight, night, rain, foggy, and snow), simulating real environments through the combination of different scenes. Furthermore, we provide the corresponding ground-truth images for each visible image affected by adverse weather conditions for image fusion and image restoration network training. As shown in Table 1, compared with previous visible-infrared datasets, SWVID is the first one that considers weather conditions. Specifically, we have constructed the dataset from public visible-infrared datasets as follows:

$$\mathcal{D}_{rain}(J(x)) = J(x)(1 - M_r(x)) + R(x)M_r(x), \tag{16}$$

$$\mathcal{D}_{snow}(J(x)) = J(x)(1 - M_s(x)) + S(x)M_s(x), \tag{17}$$

$$\mathcal{D}_{foggy}(J(x)) = J(x)e^{-\int_0^{d(x)} \beta dl} + \int_0^{d(x)} L_\infty \beta e^{-\beta l} dl. \tag{18}$$

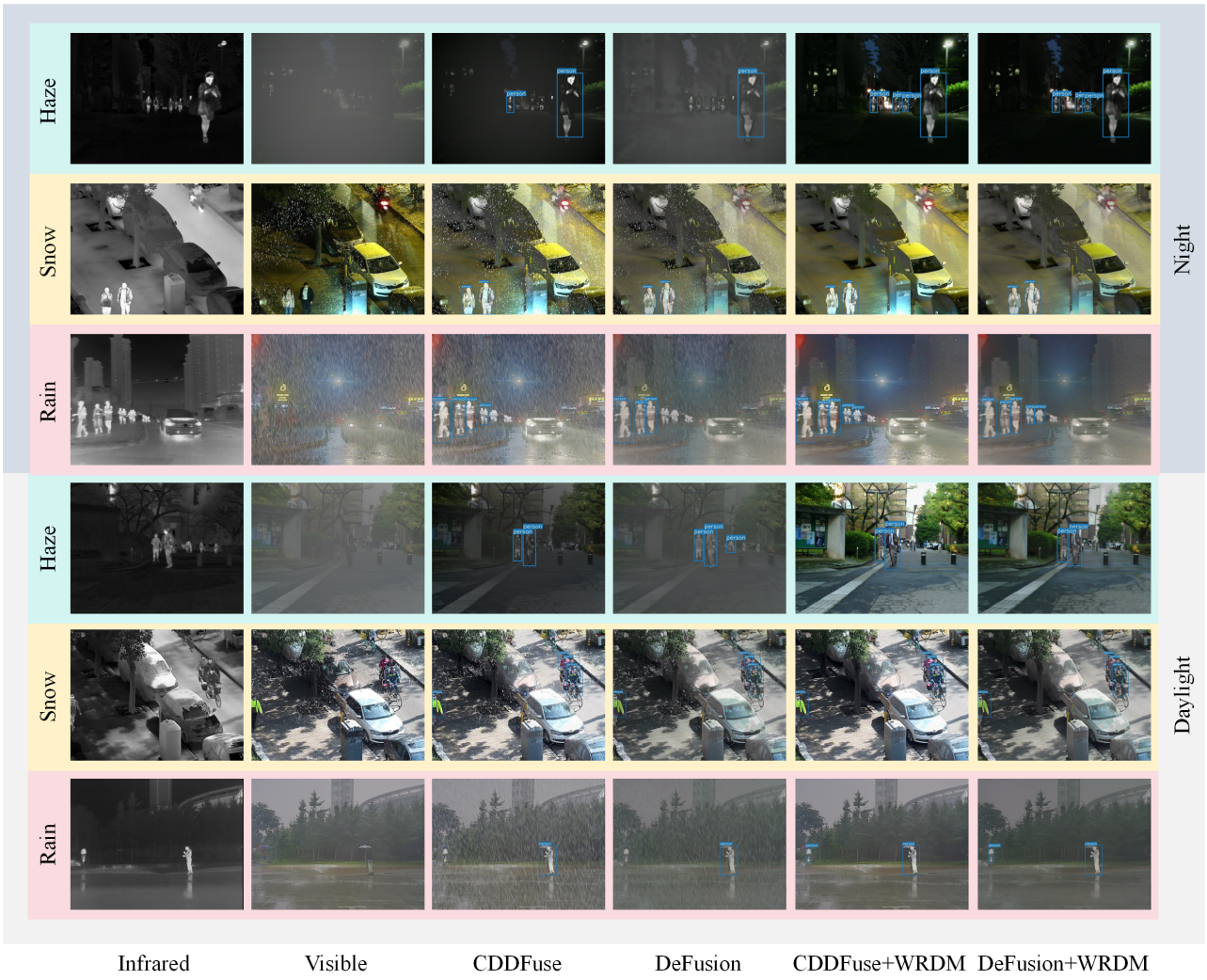

Figure 5: Examples of daylight and night scenes for multimodal fusion and object detection visualization, including three kinds of adverse weather conditions (rain, haze, and snow). We embed WRDM into two state-of-the-art visible-infrared fusion methods (*i.e.*, CDDFuse [59] and DeFusion [25]) to mitigate the adverse impact of weather conditions.

where x represents the spatial location in an image, $\mathcal{D}_{rain}(J(x))$, $\mathcal{D}_{snow}(J(x))$ and $\mathcal{D}_{foggy}(J(x))$ represent a function that maps a clear image to one with rain, snow, and fog particle effects, $J(x)$ represents the clear image with no weather effects, $M_r(x)$ and $M_s(x)$ represent rain and snow equivalents, $R(x)$ represents a map of the rain masks, $S(x)$ represents a chromatic aberration map of the snow particles. Considering scattering effects, $d(x)$ represents the distance from the observer at a pixel location x, β is an atmospheric attenuation coefficient, and L_∞ is the radiance of light.

We divide SWVID into the training set (34, 280 images), validation set (17, 140 images), and test set (8, 570 images), each folder contains three parts: pairs of visible-infrared images and corresponding weather-influenced visible images. Notice that weather-influenced visible images contain three kinds of weather conditions, classified as SWVID-snow, SWVID-rain, and SWVID-foggy. During the training period, we use the pairs of images (weather-influenced and ground-truth) to train WRDM in the first stage, then use the pairs of images (ground-truth and infrared) with corresponding labels to train CFM in the second stage. During the validating and testing period, we use the pairs of images (weather-influenced and infrared) directly, verifying and testing the performance of CFMW under real conditions. Also, we use the same way when evaluating other networks in comparative experiments.

Evaluation metrics. We adopt the conventional peak signal-to-noise ratio (PSNR) [15] and structural similarity (SSIM) [47] for quantitative evaluations between ground truth and restored images. PSNR is mainly used to evaluate the degree of distortion after image processing, while SSIM pays more attention to the

Table 2: Quantitative comparisons in terms of PSNR and SSIM (higher is better) with state-of-the-art image deraining, dehazing, and desnowing methods. For the sake of fairness, we uniformly use the visible light part of the established SWVID dataset as the evaluation dataset.

Image-Deraining Task	SWVID-rain (RGB)		Image-Dehazing Task	SWVID-foggy (RGB)		Image-Desnowing Task	SWVID-snow (RGB)	
	PSNR↑	SSIM↑		PSNR↑	SSIM↑		PSNR↑	SSIM↑
pix2pix [17]	19.95	0.7270	pix2pix [17]	25.12	0.8359	SPANet [46]	29.92	0.8260
CycleGAN [60]	17.65	0.6452	DuRN [29]	31.44	0.9256	DDMSNet [57]	34.87	0.9462
PCNet [19]	27.13	0.8546	AttentiveGAN [33]	32.56	0.9331	DesnowNet [2]	32.15	0.9416
MPRNet [53]	29.14	0.9022	IDT [49]	34.14	0.9412	RESCAN [24]	30.57	0.9003
de-rain (ours)	**36.78**	**0.9464**	**de-haze (ours)**	**36.53**	**0.9795**	**de-snow (ours)**	**42.23**	**0.9821**
All-in-One [23]	25.13	0.8856	All-in-One [23]	31.24	0.9122	All-in-One [23]	28.12	0.8815
TransWeather [42]	29.77	0.9107	TransWeather [42]	33.85	0.9388	TransWeather [42]	35.15	0.9417
WRDM (ours)	**35.02**	**0.9322**	**WRDM (ours)**	**35.88**	**0.9602**	**WRDM (ours)**	**40.98**	**0.9578**

Table 3: Comparison of performances with other networks on the SWVID-snow dataset.

Model	Data	Backbone	mAP50↑	mAP75↑	mAP↑
mono-modaltiy networks					
Faster R-CNN [36]	RGB	ResNet50	82.3	34.6	30.7
Faster R-CNN [36]	Thermal	ResNet50	90.6	63.7	55.4
SDD [28]	RGB	VGG16	73.6	37.8	38.6
SDD [28]	Thermal	VGG16	88.6	55.6	50.2
YOLOv3 [35]	RGB	Darknet53	78.3	29.4	24.4
YOLOv3 [35]	Thermal	Darknet53	84.6	50.7	47.4
YOLOv5 [20]	RGB	CSPD53	80.7	38.2	30.7
YOLOv5 [20]	Thermal	CSPD53	90.5	65.2	57.6
YOLOv7 [45]	RGB	CSPD53	85.3	41.8	34.9
YOLOv7 [45]	Thermal	CSPD53	91.8	67.6	60.4
multi-modality networks					
Baseline	RGB+T	CSPD53	92.2	68.4	59.3
CFT [34]	RGB+T	CFB	92.4	71.1	58.4
CFMW (ours)	RGB+T	CFM	**97.2**	**76.9**	**63.4**

Table 4: Comparison of performances with other networks on the LLVIP [18] dataset.

Model	Data	Backbone	mAP50↑	mAP75↑	mAP↑
mono-modaltiy networks					
Faster R-CNN [36]	RGB	ResNet50	91.4	48.0	49.2
Faster R-CNN [36]	Thermal	ResNet50	96.1	68.5	61.1
SDD [28]	RGB	VGG16	82.6	31.8	39.8
SDD [28]	Thermal	VGG16	90.2	57.9	53.5
YOLOv3 [35]	RGB	Darknet53	85.9	37.9	43.3
YOLOv3 [35]	Thermal	Darknet53	89.7	53.4	52.8
YOLOv5 [20]	RGB	CSPD53	90.8	51.9	50.0
YOLOv5 [20]	Thermal	CSPD53	94.6	72.2	61.9
YOLOv7 [45]	RGB	CSPD53	91.4	58.4	53.6
YOLOv7 [45]	Thermal	CSPD53	94.6	70.6	62.4
multi-modality networks					
Baseline	RGB+T	CSPD53	95.2	71.4	62.3
CFT [34]	RGB+T	CFB	97.5	72.9	63.6
CFMW (ours)	RGB+T	CFM	**98.8**	**77.2**	**64.8**

structural information and visual quality of the images.

$$PSNR = 10 \times lg(\frac{(2^n - 1)^2}{MSE}), \tag{19}$$

$$SSIM = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}, \tag{20}$$

As for object detection quantitative experiments, we introduced three object detection metrics: mean Average Precision (mAP, mAP50, and mAP75) to evaluate the accuracy of the object detection models. For more calculation details, please refer to the supplementary material.

4.2 Implantation Details

As for WRDM, we performed experiments both in specific-weather conditions and multi-weather conditions image restoration settings. We denote our specific-weather restoration models as de-rain, de-snow, and de-foggy to verify the general WRDM model under specific weather conditions. We trained the 128×128 patch size version of all models. We use NVIDIA RTX 4090 cards to perform all the experiments. We use Adam as an optimizer while training all the models we compare. During the training process, we trained WRDM for 3×10^6 iterations. As for CFM, we did not perform task-specific parameter tuning or modifications to the network architecture. For better performance, we select the YOLOv5 model's public weight initialization (yolov5s.pt), which is pre-trained on the COCO dataset [26].

4.3 Comparative Experiments

In this section, we make comparisons with several state-of-the-art methods in image deweathering and cross-modality object detection separately. In Table 2, we perform comparisons with methods for image desnowing (*i.e.* SPANet [46], DDMSNet [57], DesnowNet [2], RESCAN [24]), deraining (*i.e.* pix2pix [17], CycleGAN [60], PCNet [19], MPRNet [53]), and dehazing (*i.e.* pix2pix [17], DuRN [29], Attentive-GAN [33], IDT [49]), as well as two state-of-the-art multi-weather image restoration methods: All in One [23] and TransWeather [42]. In Table 3 and Table 4, to prove the consistent improvements of CFMW, we compare with several base single-modality object detection methods (*i.e.*, Faster R-CNN [36], SDD [28], YOLOv3 [35], YOLOv5 [20], YOLOv7 [45]) and several multi-modality object detection methods (*i.e.*, our baseline, standard two-stream YOLOv5 object detection network, and CFT [34]).

Table 5: Ablation experiments on SWVID-snow dataset. To present the general effectiveness of our CFMW, we further combine the WRDM and CFM module with other classical detectors (*i.e.*, YOLOv7, YOLOv5, Faster R-CNN).

Modality	Method	Detector	mAP50↑	mAP75↑	mAP↑
RGB	CSPDarknet53	YOLOv7 [45]	85.3	41.8	34.9
Thermal	CSPDarknet53		95.8	72.6	60.4
RGB+T	+two stream		95.4	68.1	60.4
	+CFM		95.5	68.6	63.3
	+WRDM		96.5	70.9	63.1
	+CFM&WRDM		**96.6**	**75.1**	**64.1**
RGB	CSPDarknet53	YOLOv5 [20]	80.7	38.2	30.7
Thermal	CSPDarknet53		90.5	65.2	57.6
RGB+T	+two stream		92.2	68.4	59.3
	+CFM		96.5	70.6	63.3
	+WRDM		96.4	71.2	62.8
	+CFM&WRDM		**97.2**	**76.9**	**63.4**
RGB	Resnet53	Faster R-CNN [36]	82.3	34.6	30.7
Thermal	Resnet53		90.6	63.7	55.4
RGB+T	+two stream		93.7	62.8	55.4
	+CFM		96.7	69.5	61.9
	+WRDM		96.2	69.4	61.6
	+CFM&WRDM		**96.2**	**69.7**	**62.2**

Comparison of image deweathering. As shown in Table 2, we use the single RGB modality of the SWVID dataset (including rain, foggy, and haze weather conditions) as a comparative dataset to measure the performance of different models under different weather conditions. The top of the table contains results from specific-weather image restoration, where we show $S = 50$ sampling time steps. For image-deraining, image-dehazing, and image-desnowing tasks, the proposed solution consistently achieves the best results (36.78/0.9464 on SWVID-rain, 36.53/0.9795 on SWVID-foggy, and 42.23/0.9821 on SWVID-snow). Especially, in the image de-rain task, the performance improvement is about 24% compared with the current state-of-the-art method (MPRNet [53]). For multi-weather image restoration, although the results are not as good as the specific-weather model due to the complexity of the task, the proposed method also reaches the best results (35.02/0.9322 on SWVID-rain, 35.88/0.9602 on SWVID-foggy, and 40.98/0.9578 on SWVID-snow) compared with All in One [23] and TransWeather [42], with about 17% performance improvement compared against TransWeather [42] and about 25% performance improvement compared against All in One [23].

Comparison of cross-modality object detection. As shown in Table 3 and Table 4, we use LLVIP [18] and SWVID-snow as the comparative datasets. Compared with SWVID-rain and SWVID-foggy, the size of pedestrians in these two datasets is more in line with the general object detection standards. There are more complex cases of pedestrian overlap in these two datasets, which can better measure the accuracy of the object detection networks. The top of the table contains results from single-modality networks, each network uses the RGB modality or the thermal modality for detection. The bottom of the table shows results from multi-modality networks, including our baseline, CFT [34] and the proposed CFMW. According to Table 3, it can be observed that with the integration of WRDM and CFM, CFMW achieves an overwhelming performance improvement on each metric (mAP50:2.3↑, mAP75:4.3↑, mAP:3.0↑) on SWVID-snow compared with the best existing network on each metric, which shows that it has preferable adaptability under adverse weather conditions. Also, CFMW can achieve a more accurate detection (mAP50:98.8, mAP75:77.2, mAP:64.8) with lower computational consumption, as shown in Table 4, which demonstrates the commonality of CFWM.

4.4 Ablation Study

In this section, we analyze the effectiveness of CFMW. We first validate the importance of WRDM and CFM modules in performance improvement in a parametric form through detailed ablation experiments, then visually show the role of WRDM in cross-modality fusion and object detection tasks to highlight its versatility as a weather-restoration plug-in.

Ablation experiments To understand the impact of each component in our method, we have performed a comprehensive set of ablation experiments. As shown in Table 5, we further combine the CFM and WRDM with other classical detectors, *i.e.* YOLOv7 [45], YOLOv5 [20] and Faster R-CNN [36] to present the general effectiveness of our CFMW. The proposed CFMW improves the performance of cross-modality object detection using either a one-stage or two-stage detector under complex weather conditions. Specifically, CFM achieves an 11.3% gain on mAP50, an 81.6% gain on mAP75, and a 78.3% gain on mAP (on YOLOv5 [20]). After adding WRDM, we achieved a 12.1% gain on mAP50, an 88.2% gain on mAP75, and an 80.4% gain on mAP. CFM and WRDM provide non-negligible gains for all the considered evaluation metrics.

Visual interpretation To verify the applicability of WRDM as a plug-in intuitively, we visually show the application scenario of WRDM in the field of visible-infrared image fusion and object detection. As shown in Fig. 5, we perform comparisons with methods of visible-infrared image fusion methods (*i.e.* CDDFuse [59], DeFusion [25]). It can be seen from the figure that compared with the original images, the image fusion effects of the two methods before and after using WRDM are quite different, more people at the far end of images could be detected successfully after deweathering. In cross-modality object detection, rich image details can provide great assistance for feature extraction and fusion, with direct fusion without removing the weather influence causing the loss and interference of image details.

5 CONCLUSION

In this work, we introduce a novel approach to visible-infrared object detection under severe weather conditions, namely the Severe Weather Visible-Infrared Dataset (SWVID). We have provided a valuable resource for training and evaluating models in realistic and challenging environments. The Cross-modality Fusion Mamba with Weather-removal (CFMW) model, has proven to be highly effective in enhancing detection accuracy while managing computational efficiency. Our extensive experiments have shown that CFMW outperforms existing benchmarks, achieving state-of-the-art on both tasks: multi-weather image restoration and cross-modality object detection. This work opens up new possibilities for cross-modality object detection in adverse weather.

REFERENCES

[1] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention Multi-scale Vision Transformer for Image Classification. In *Proceedings of the the IEEE/CVF International Conference on Computer Vision (ICCV).*

[2] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Chen-Che Tsai, and Sy-Yen Kuo. 2020. JSTASR: Joint Size and Transparency-Aware Snow Removal Algorithm Based on Modified Partial Convolution and Veiling Effect Removal. In *Proceedings of the the European Conference on Computer Vision (ECCV).*

[3] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. 2022. Learning Multiple Adverse Weather Removal via Two-stage Knowledge Learning and Multi-contrastive Regularization: Toward A Unified Model. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[4] Yi-Ting Chen, Jinghao Shi, Christoph Mertz, Shu Kong, and Deva Ramanan. 2021. Multimodal Object Detection via Bayesian Fusion. *arXiv preprint arXiv:2104.02904* 3, 6 (2021).

[5] Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. 2023. MUFusion: A General Unsupervised Image Fusion Network Based on Memory Unit. *Information Fusion* 92 (2023), 80–92.

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat Gans on Image Synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[8] Team. F. 2018. Free Flir Thermal Dataset for Algorithm Training. (2018). https://www.flir.com/oem/adas/adas-dataset-form/.

[9] Xin Feng, Wenjie Pei, Zihui Jia, Fanglin Chen, David Zhang, and Guangming Lu. 2021. Deep-masking Generative Network: A Unified Framework for Background Restoration from Superimposed Images. *IEEE Transactions on Image Processing* 30 (2021), 4867–4882.

[10] Albert Gu and Tri Dao. 2023. Mamba: Linear-time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752* (2023).

[11] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv preprint arXiv:2111.00396* (2021).

[12] Xuanhua He, Ke Cao, Ke Ren Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. 2024. Pan-Mamba: Effective pan-sharpening with State Space Model. *arXiv preprint arXiv 2402.12192* (2024).

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[14] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research* 23, 47 (2022), 1–33.

[15] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of Validity of PSNR in Image/Video Quality Assessment. *Electronics Letters* 44 (2008), 800–801.

[16] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baselines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[18] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In *Proceedings of the the IEEE/CVF International Conference on Computer Vision (ICCV).*

[19] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. 2021. Rain-Free and Residue Hand-in-Hand: A Progressive Coupled Network for Real-Time Image Deraining. *IEEE Transactions on Image Processing* 30 (2021), 7404–7418.

[20] Glenn Jocher. 2020. *YOLOv5 by Ultralytics.* https://doi.org/10.5281/zenodo.3908559

[21] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising Diffusion Restoration Models. In *Proceedings of the the International Conference on Advances in Neural Information Processing Systems (NIPS).*

[22] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. 2022. All-in-one Image Restoration for Unknown Corruption. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[23] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. 2020. All in One Bad Weather Removal Using Architectural Search. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[24] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. 2018. Recurrent Squeeze-and-excitation Context Aggregation Net for Single Image Deraining. In *Proceedings of the European conference on computer vision (ECCV).* 254–269.

[25] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. 2022. Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion. In *Proceedings of the the European Conference on Computer Vision (ECCV).*

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the the European Conference on Computer Vision (ECCV).*

[27] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. Target-aware Dual Adversarial Learning and A Multi-scenario Multi-modality Benchmark to Fuse Infrared and Visible for Object Detection. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[28] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. SSD: Single Shot MultiBox Detector. In *Proceedings of the the European Conference on Computer Vision (ECCV).*

[29] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. 2019. 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 (CVPR).*

[30] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. Vmamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166* (2024).

[31] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2022. Long Range Language Modeling via Gated State Spaces. *arXiv preprint arXiv:2206.13947* (2022).

[32] Zhang Pengyu, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. 2022. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline. In *Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[33] Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive Generative Adversarial Network for Raindrop Removal from A Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[34] Fang Qingyun and Wang Zhaokui. 2022. Cross-modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery. *Pattern Recognition* 130 (2022), 108786.

[35] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv 1804.02767* (2018).

[36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[37] Manish Sharma, Mayur Dhanaraj, Srivallabha Karnam, Dimitris G Chachlakis, Raymond Ptucha, Panos P Markopoulos, and Eli Saber. 2020. YOLOrs: Object Detection in Multimodal Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), 1497–1508.

[38] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified State Space Layers for Sequence Modeling. *arXiv preprint arXiv:2208.04933* (2022).

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *Proceedings of the the International Conference on Machine Learning (ICML).*

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. *arXiv preprint arXiv:2010.02502* (2020).

[41] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. 2022. PIAFusion: A Progressive Infrared and Visible Image Fusion Network Based on Illumination Aware. *Information Fusion* 83 (2022), 79–92.

[42] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. 2022. Transweather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems* 30 (2017).

[44] Vibashan VS, Poojan Oza, and Vishal M Patel. 2023. Towards Online Domain Adaptive Object Detection. In *Proceedings of the the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).*

[45] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7:Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors. In *Proceedings of the IEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[46] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W. H. Lau. 2019. Spatial Attentive Single-Image Deraining With a High Quality Real Rain Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[47] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[48] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. 2022. Deblurring via Stochastic Refinement. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[49] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zhengjun Zha. 2022. Image De-Raining Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2022), 12978–12995.
[50] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. 2020. FusionDN: A Unified Densely Connected Network for Image Fusion. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
[51] Meilong Xu, Linfeng Tang, Hao Zhang, and Jiayi Ma. 2022. Infrared and Visible Image Fusion via Parallel Scene and Texture Learning. *Pattern Recognition* 132 (2022), 108929.
[52] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. *arXiv preprint arXiv 1605.07146* (2016).
[53] Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-Stage Progressive Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[54] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. 2020. Multispectral Fusion for Object Detection with Cyclic Fuse-and-refine Blocks. In *Proceedings of the IEEE International conference on image processing (ICIP)*.
[55] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. 2021. Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. In *Proceedings of the the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
[56] Heng ZHANG, Élisa Fromont, Sébastien Lefèvre, Bruno Avignon, and Universite de Rennes. 2021. Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.
[57] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. 2021. Deep Dense Multi-Scale Network for Snow Removal Using Semantic and Depth Priors. *IEEE Transactions on Image Processing* 30 (2021), 7419–7431.
[58] Fan Zhao, Wenda Zhao, and Huchuan Lu. 2023. Interactive Feature Embedding for Infrared and Visible Image Fusion. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
[59] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. 2023. Cddfuse: Correlation-driven Dual-branch Feature Decomposition for Multi-modality Image Fusion. In *Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
[61] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv preprint arXiv:2401.09417* (2024).

A DERIVATION OF DENOISING DIFFUSION MODELS

The forward process is a fixed Markov Chain that corrupts the data $x_0 \sim q(x_0)$ at T time steps, by injecting Gaussian noise according to a variance schedule $\beta_1 \sim \beta_T$, which could be expressed using the following formula:

$$q(x_t|x_t-1) = N(x_t; \sqrt{1-\beta_t}x_t - 1, \beta_t \mathbf{I}), \tag{21}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{22}$$

The reverse process defined by the joint distribution $p_\theta(x_{0:T})$ is a Markov Chain with learned Gaussian denoising transitions starting at a standard normal prior $p(x_T) = N(x_T; \boldsymbol{O}; \boldsymbol{I})$, which could be expressed using the following formula:

$$p_\theta(x_0 : T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{23}$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)). \tag{24}$$

The reverse process is parameterized by a neural network that estimates $\prod_{t=1}^{T}$ and $\sum_\theta(x_t, t)$.

Denoising diffusion implicit models provide a new method to accelerate deterministic sampling for pre-trained diffusion models, which can generate consistent and better-quality image samples. Following implicit sampling utilizes a generalized non-Markov forward process formula:

$$q_\lambda(x_{1:T}|x_0) = q_\lambda(x_T|x_0) \prod_{t=2}^{T} q_\lambda(x_{t-1}|x_t, x_0), \tag{25}$$

$$q_\lambda(x_T|x_0) = N(\sqrt{1-\beta_T}x_0, \beta_T \boldsymbol{I}). \tag{26}$$

for $t \geq 2$:

$$q_\lambda(x_{t-1}|x_t, x_0) = N(x_{t-1}; \sqrt{1-\beta_{t-1}} + \sqrt{\beta_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{1-\beta_t}x_0}{\sqrt{\beta}}, \sigma_t^2 \boldsymbol{I}). \tag{27}$$

where σ_t^2 is a real number. We can prove by mathematical induction that for all t:

$$q_\lambda(x_T|x_0) = N(x_t; \sqrt{1-\beta_T}x_0, \beta_T \boldsymbol{I}). \tag{28}$$

So we could rewrite the distribution in (24) in terms of a particular choice of its standard deviation λ_t as:

$$p_\lambda(x_{t-1}|x_t, x_0) = N(x_{t-1}; \widetilde{\boldsymbol{\mu}}_t(x_t, x_0), \lambda_t^2 \boldsymbol{I}), \tag{29}$$

$$\widetilde{\boldsymbol{\mu}}_t = \sqrt{\bar{\alpha_{t-1}}} \cdot x_0 + \sqrt{1 - \bar{\alpha_{t-1}} - \lambda_t^2} \cdot \boldsymbol{\epsilon}_t. \tag{30}$$

where $\alpha_t = 1 - \beta_t, \bar{\alpha_t} = \prod_{i=1}^{t} \alpha_i, \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{I})$. Each of them has the same dimensionality as data x_0 and latent variable x_t.

B MORE DETAILS OF LOSS FUNCTIONS

During training WRDM, specifically, for each time step t, we have:

$$\mathcal{L}_\theta = \mathbb{E}_q[log_{p_\theta}(x_0|x_T)] - \mathbb{E}_{q(x_{t-1}|x_t)} [D_{KL}(q(x_{t-1}|x_t, x_0))||p_\theta(x_{t-1}|x_t)]. \tag{31}$$

where the first term is the expected value of $log_{p_\theta}(x_0|x_T)$ under the variational distribution $q(x_T)$, and the second term is the expected value of the Kullback-Leibler divergence between $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$. The Kullback-Leibler divergence measures the difference between two probability distributions. Summing up the variational bounds for all time steps, we obtain the variational bound for the entire diffusion process:

$$\mathcal{L}_\theta = \sum_{t=1}^{T} \mathbb{E}_q[log_{p_\theta}(x_0|x_T)] - \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t-1}|x_t)} [D_{KL}(q(x_{t-1}|x_t, x_0))||p_\theta(x_{t-1}|x_t)]. \tag{32}$$

Then we show more calculation details about $\mathcal{L}_{total}$, $\mathcal{L}_{box}$, $\mathcal{L}_{cls}$, $\mathcal{L}_{noobj}$ and $\mathcal{L}_{obj}$ here:

$$\mathcal{L}_{total} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{noobj} + \mathcal{L}_{obj}, \tag{33}$$

$$\mathcal{L}_{box} = \sum_{i=0}^{S^2} \sum_{j=0}^{N} \boldsymbol{l}_{i,j}^{obj} \cdot [1 - GIoU_i], \tag{34}$$

$$\mathcal{L}_{cls} = \sum_{i=0}^{S^2} \sum_{j=0}^{N} \boldsymbol{l}_{i,j}^{obj} \cdot \sum_{c \in classes} p_i(c) log(\hat{p}_i(c)), \tag{35}$$

$$\mathcal{L}_{noobj} = \sum_{i=0}^{S^2} \sum_{j=0}^{N} \boldsymbol{l}_{i,j}^{noobj} \cdot (c_i - \hat{c}_i)^2, \tag{36}$$

$$\mathcal{L}_{obj} = \sum_{i=0}^{S^2} \sum_{j=0}^{N} \boldsymbol{l}_{i,j}^{obj} \cdot (c_i - \hat{c}_i)^2. \tag{37}$$

where Generalized Intersection over Union (GIoU) is employed as the predicted regression loss. S^2 and N represent the number of image grids during prediction and the number of predicted boxes. $p(c)$ and $\hat{p}(c)$ represent the probability that the real sample is class c and the probability that the network predicts the sample to be class c. $\boldsymbol{l}_{i,j}^{obj}$ represent whether the j^{th} predicted box of the i^{th} grid is a positive sample, with $\boldsymbol{l}_{i,j}^{noobj}$ represent whether the j^{th} predicted box of the i^{th} grid is a negative sample.

C MORE DETAILS OF METRICS CALCULATION

PSNR could be calculated as follows:

$$PSNR = 10 \times lg(\frac{(2^n - 1)^2}{MSE}), \tag{38}$$

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X(i,j) - Y(i,j))^2. \tag{39}$$

where H and W represent the height and width of the images, n is the number of bits per pixel (generally taken as 8), $X(i,j)$ and $Y(i,j)$ respectively represent the pixel values at the corresponding coordinates.

SSIM could be calculated as follows:

$$SSIM = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}, \tag{40}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1}, \tag{41}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C2}{\sigma_x^2 + \sigma_y^2 + C2}, \tag{42}$$

$$s(x,y) = \frac{\sigma_{xy} + C3}{\sigma_x\sigma_y + C3}. \tag{43}$$

where $l(x,y)$ measures brightness, $c(x,y)$ measures contrast ratio, $s(x,y)$ measures structure, μ and σ represents mean and standard deviation.

mAP, mAP50 and mAP75 could be calculated as follows:

$$mAP = \frac{1}{n} \sum_{i=1}^{N} AP_i, \tag{44}$$

$$AP_i = \int_0^1 Precision\ d(Recall). \tag{45}$$

mAP50 computes the mean of all the AP values for all categories at IoU=0.50, and mAP75 computes the mean at IoU=0.75, similarly.