DCAAFusion: A Diffusion Model-Based Cross-Attention Adaptive Fusion Network for Nighttime Infrared and Visible Image Fusion

Jinao Li[1,2], Jinyong Cheng[1,2,*], Kening Cui[1,2]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China
10431230216@stu.qlu.edu.cn, cjy@qlu.edu.cn, 10431230146@stu.qlu.edu.cn
*Corresponding Author: Jinyong Cheng Email: cjy@qlu.edu.cn

***Abstract*—Infrared and visible image fusion seeks to combine complementary information from heterogeneous sensors into a unified composite image, thus providing a more comprehensive scene representation. However, fusion in nighttime low-light degradation scenarios remains a significant challenge. Low illumination weakens the brightness and details of visible images, leading to the loss of critical information for effective fusion. Furthermore, existing fusion approaches lack adaptive modeling mechanisms that fully leverage the complementary information from both modalities, often resulting in blurred salient features and the loss of fine textures in the fused image. To address these challenges, we propose a Diffusion Model-Based Cross-Attention Adaptive Fusion Network (DCAAFusion). Specifically, we employ the denoising network of a pre-trained conditional diffusion model to extract visible diffusion features, and that of a pre-trained diffusion model for infrared diffusion features. Subsequently, we design a Degradation Embedding Modulation Module (DEMM), which modulates the visible diffusion features using the degradation embedding extracted by an introduced degradation-aware encoder, compensating for the loss of brightness and details caused by low light at the feature level. Finally, we propose a Cross-Attention Adaptive Fusion Module (CAAFM), which employs multiple attention mechanisms to adaptively fuse infrared structural information and visible textural details, fully exploiting complementary information from both modalities. Extensive experiments demonstrate that DCAAFusion outperforms current state-of-the-art methods, achieving superior performance in both visual quality and quantitative metrics. The source code is available at https://github.com/AochenA/DCAAFusion.**

***Index Terms*—Image fusion, Diffusion model, Low-light degradation scenes, Cross-Attention, Adaptive fusion**

I. Introduction

Image fusion techniques aim to integrate complementary information from different sensors to generate an image with a more comprehensive scene representation [1] [2]. Among various image fusion tasks, infrared–visible image fusion (IVIF) is one of the most prevalent [3]. While visible images are rich in fine textures, they are prone to severe degradation under low illumination. In contrast, infrared imagery captures the thermal radiation emitted by objects, which can highlight prominent targets even under extreme weather conditions, though it lacks fine-grained texture details. By fusing these complementary modalities, high-quality images with enhanced contrast, richer textures, and more salient thermal targets can be obtained, which play an irreplaceable role in many fields, such as military applications, object detection and tracking [4], semantic segmentation [5], and person re-identification [6].

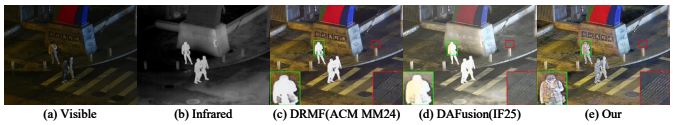

Fig. 1. Fusion results comparison between our method and other methods on the LLVIP dataset under nighttime low-light degradation scenarios.

Despite the generally good results achieved by recent fusion methods for nighttime low-light degradation scenes, several challenges persist. DRMF [7] performs fusion by combining the generative diffusion priors of different modalities in the noise space and learning timestep-dependent combination weights. However, the introduced salient target mask constraint biases the weights toward the infrared modality, thereby weakening the texture and color details of the visible modality, as shown in Fig. 1(c). DAFusion [8] integrates degradation awareness into feature aggregation via implicit degradation estimation and contrastive supervision, and utilizes the Co-refinement Fusion Module and mask constraints for fusion. However, its feature extraction relies on discriminative encoders, making it susceptible to feature representation shift, which leads to attenuated textures or insufficient color restoration, as illustrated in Fig. 1(d).

To address the challenges outlined above, we propose DCAAFusion, an infrared-visible image fusion network specifically designed for nighttime low-light degradation scenes. The network effectively mitigates nighttime low-light degradation in visible images and adaptively fuses visible texture details with infrared structural information, as exhibited in Fig. 1(e). First, we utilize the denoising networks of pre-

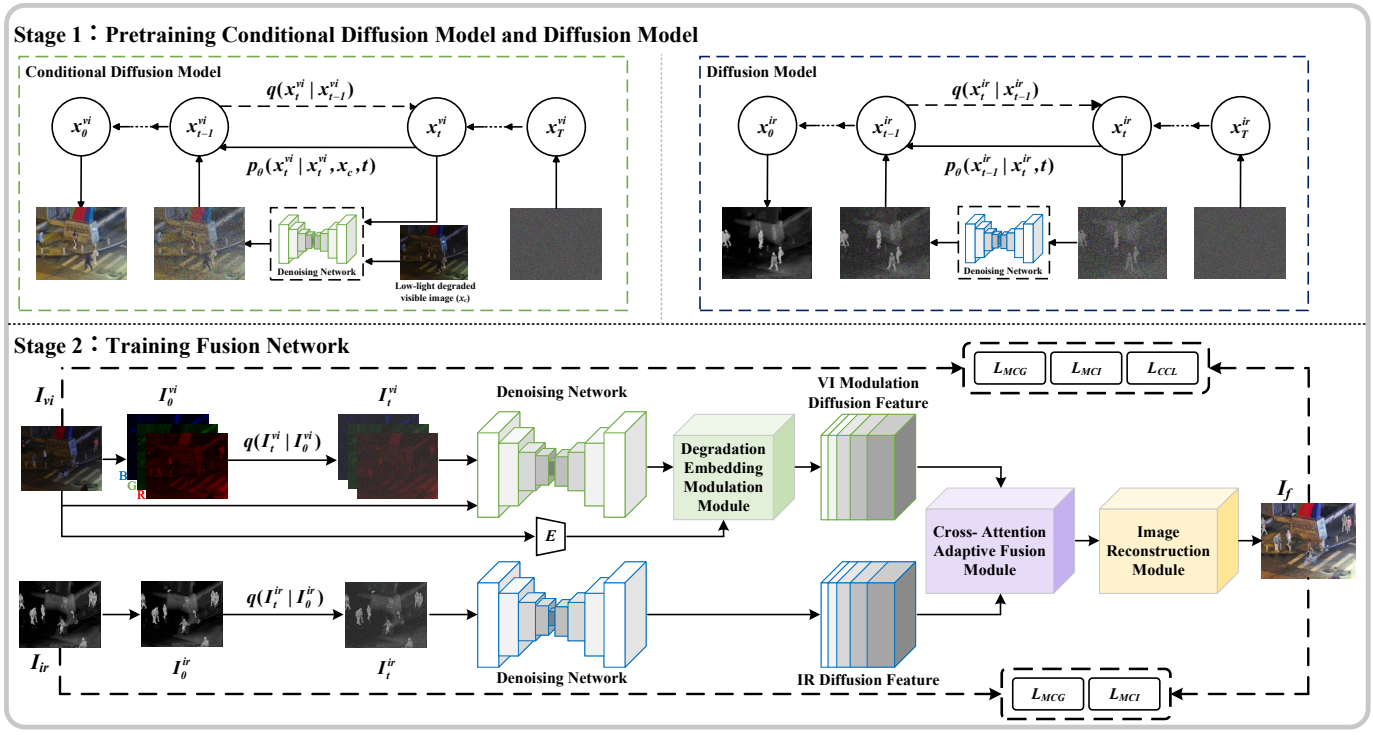

Fig. 2. Framework of the proposed DCAAFusion. $q(\cdot|\cdot)$ and $p_\theta(\cdot|\cdot)$ stand for the forward process and reverse process. I_0^{vi} and I_0^{ir} represent the initial states of the diffusion process corresponding to the inputs I_{vi} and I_{ir}. I_t^{vi} and I_t^{ir} denote the noisy versions in the forward process at selected timesteps t. E denotes the degradation-aware encoder. L_{MCG}, L_{MCI}, and L_{CCL} represent the multi-channel gradient loss, multi-channel intensity loss, and color consistency loss.

trained conditional diffusion models and diffusion models to extract visible and infrared features at selected timesteps. Next, we design a Degradation Embedding Modulation Module (DEMM), which uses a degradation embedding extracted by the introduced degradation-aware encoder to modulate the visible diffusion features, compensating for the loss of brightness and detail caused by low-light degradation. Finally, we present Cross-Attention Adaptive Fusion Module (CAAFM), which models cross-modal dependencies through cross-attention and then applies channel and spatial attention to dynamically allocate weights and adaptively fuse the features. The primary contributions of our work are summarized below:

- We propose a dual-branch feature extraction framework that leverages the frozen denoising networks of pre-trained conditional diffusion models and diffusion models to extract diffusion features from visible and infrared images at selected timesteps.
- We propose a Degradation Embedding Modulation Module (DEMM), which leverages the degradation embedding extracted by the degradation-aware encoder to adaptively modulate visible diffusion features, explicitly compensating for the loss of brightness and detail.
- We design a Cross-Attention Adaptive Fusion Module (CAAFM) that first employs cross-attention to model cross-modal dependencies, and then performs adaptive fusion through dynamic channel and spatial attention weighting, enhancing salient infrared structures while preserving detailed visible textures.

II. Related Work

A. Fusion of Infrared and Visible Images

The rapid advancement of deep learning technologies has significantly propelled the field of multimodal image fusion. The powerful feature extraction ability of Auto-Encoder (AE)-based fusion frameworks has led to their widespread adoption. Li et al. employed auto-encoders with dense blocks for feature extraction and integration through handcrafted fusion rules [9]. Although this approach has inspired numerous subsequent studies [10], [11], its fixed fusion strategy struggles to adapt to varying scene characteristics, thereby limiting further improvements in fusion performance. In response, Convolutional Neural Network (CNN)-based architectures have enhanced the flexibility and interpretability of the fusion process [12]–[14]. To address the needs of downstream high-level vision tasks, Tang et al. developed a fusion method driven by semantic information [15]. This model incorporates a segmentation loss to constrain the network, ensuring the preservation of critical semantic details.

With the successful application of Generative Adversarial Network (GAN) in vision tasks, Ma et al. were the first to apply adversarial learning techniques to the task of IVIF [16]. Subsequently, Liu et al. introduced a dual-adversarial learning architecture with target awareness [17]. By utilizing saliency masks to separately constrain the generation of the foreground

and background, this approach achieves high-quality fusion beneficial for object detection. To overcome the limitations of CNN in capturing contextual information, Transformer architectures have gained increasing attention. Ma et al. developed a unified fusion framework leveraging the Swin Transformer [18], using a cross-attention mechanism to exploit multi-modal complementary features, thereby achieving the dual preservation of texture details and salient targets [19].

B. Diffusion Model

Recently, Diffusion Model (DM) [20] have emerged as a powerful class of deep generative models. Their exceptional ability to sample high-fidelity images from learned distributions has significantly advanced a wide range of applications, including image generation, super-resolution, and inpainting [21]. Building on diffusion models, Conditional Diffusion Model (CDM) [22] incorporate external conditions to guide the denoising process and have shown remarkable potential in low-light enhancement [23]. Extending this approach, recent developments have applied diffusion models to multi-modal image fusion tasks. For instance, DDFM [24] builds on pre-trained diffusion models and introduces score matching, repurposing the generative prior of natural images for multi-modal fusion. Similarly, Dif-Fusion [25] pioneers the use of denoising networks as feature extractors, achieving high-fidelity color fusion by leveraging the prior knowledge acquired during pre-training.

III. Method

A. Network Overview

The overall workflow of DCAAFusion is illustrated in Fig. 2. In Stage 1, we pretrain the conditional diffusion model on high-quality visible images x_0^{vi} conditioned on low-light degraded visible images x_c, and pretrain the diffusion model on infrared images x_0^{ir}. In Stage 2, given a low-light degraded visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$ and an infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$, we generate their noisy counterparts I_t^{vi} and I_t^{ir} at selected timesteps t via the forward diffusion process. We then concatenate I_{vi} and I_t^{vi} and feed them into the frozen denoising network ϵ_θ^{vi} to extract visible (VI) diffusion features F_{vi}, while feeding I_t^{ir} into ϵ_θ^{ir} to extract infrared (IR) diffusion features F_{ir}. In parallel, the degradation-aware encoder takes I_{vi} as input to produce a degradation embedding. The degradation-embedding modulation module (DEMM) uses this embedding to adaptively modulate F_{vi}, yielding the modulated VI diffusion features $\tilde{F}_{vi}$. Next, $\tilde{F}_{vi}$ and F_{ir} are fused by the cross-attention adaptive fusion module (CAAFM) to produce the fused representation F_f. Finally, the image reconstruction module maps F_f to the fused image I_f through a sequence of upsampling and convolutional layers.

B. Diffusion-Based Feature Extraction

Both the diffusion model and the conditional diffusion model characterize the underlying data distribution through a forward diffusion process that progressively adds noise, followed by a reverse denoising process. While both share the same forward process, the conditional diffusion model distinguishes itself by incorporating conditional information in the reverse process to guide generation.

1) Forward Diffusion Process: For each modality $m \in \{vi, ir\}$, the forward diffusion process is modeled as a fixed Markov chain:

$$q(x_t^m|x_{t-1}^m) = \mathcal{N}(x_t^m; \sqrt{\alpha_t}x_{t-1}^m, \beta_t\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, β_t denotes the variance schedule, and $\mathbf{I}$ denotes the identity matrix. The marginal distribution at timestep t can be derived using the reparameterization trick as:

$$q(x_t^m|x_0^m) = \mathcal{N}(x_t^m; \sqrt{\bar{\alpha}_t}x_0^m, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$.

2) Reverse Denoising Process: The reverse process follows a parameterized Markov chain, progressively denoising standard Gaussian noise to reconstruct the data. In the conditional diffusion model, the low-light degraded visible image x_c is incorporated as a conditional input during the reverse denoising process. The reverse processes are defined as follows:

$$p_\theta(x_{t-1}^{ir}|x_t^{ir}) = \mathcal{N}(x_{t-1}^{ir}; \mu_\theta(x_t^{ir}, t), \sigma_t^2\mathbf{I}), \tag{3}$$

$$p_\theta(x_{t-1}^{vi}|x_t^{vi}, x_c) = \mathcal{N}(x_{t-1}^{vi}; \mu_\theta(x_t^{vi}, x_c, t), \sigma_t^2\mathbf{I}), \tag{4}$$

where μ_θ is the predicted mean and σ_t^2 is derived from a fixed variance schedule.

3) Loss Function of Diffusion Process: The loss functions, denoted as $\mathcal{L}_{\text{DM}}$ and $\mathcal{L}_{\text{CDM}}$, are defined as:

$$\mathcal{L}_{\text{DM}} = ||\epsilon - \epsilon_\theta^{ir}(\sqrt{\bar{\alpha}_t}x_0^{ir} + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2, \tag{5}$$

$$\mathcal{L}_{\text{CDM}} = ||\epsilon - \epsilon_\theta^{vi}(\sqrt{\bar{\alpha}_t}x_0^{vi} + \sqrt{1-\bar{\alpha}_t}\epsilon, x_c, t)||^2, \tag{6}$$

where ϵ denotes standard Gaussian noise, ϵ_θ^{vi} refers to the denoising network designed for the visible modality, while ϵ_θ^{ir} corresponds to that of the infrared modality.

4) Denoising Network: To estimate the noise introduced during the forward diffusion process, the denoising networks ϵ_{vi} and ϵ_{ir} employ the U-Net architecture used in SR3 [22]. The network consists of a contracting encoder and an expansive decoder, each comprising five residual blocks, along with a diffusion head implemented by a single convolutional layer. Specifically, the diffusion head generates the final predicted noise, while the output of the decoder serves as high-dimensional feature representations transformed by the neural network. After pre-training ϵ_{vi} and ϵ_{ir}, their parameters are frozen and utilized feature extractors. During the subsequent training phase of the fusion network, feature maps are extracted from the intermediate decoder layers of the denoising network at selected timesteps t. These maps are then concatenated along the channel dimension to generate the visible (VI) diffusion features F_{vi} and infrared (IR) diffusion features F_{ir}.

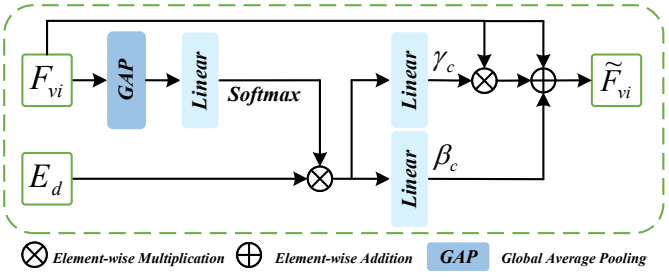

Fig. 3. Framework of our proposed Degradation Embedding Modulation Module.

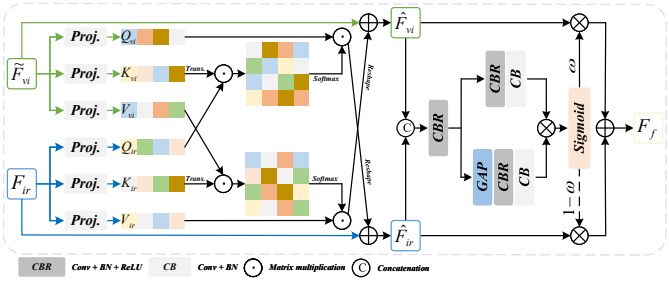

Fig. 4. Framework of our proposed Cross-Attention Adaptive Fusion Module.

C. Degradation Embedding Modulation Module

The VI diffusion features F_{vi} extracted by the conditional diffusion model suffer from inadequate representation of brightness information and texture details in low-light degradation scenarios. Inspired by [26], we design a Degradation Embedding Modulation Module (DEMM), aimed at adaptively modulating F_{vi} using the degradation embedding E_d to enhance details in low-light images and improve brightness in darker regions.

To obtain E_d for guiding the modulation, we introduce a degradation-aware encoder. Specifically, we employ the Latent Map Encoder [27] to obtain the latent map, which is mapped into the degradation embedding via Global Average Pooling (GAP) and a Multi-Layer Perceptron (MLP). The specific process of modulating F_{vi} using the degradation embedding is illustrated in Fig. 3. The input feature F_{vi} is fed into GAP, a linear transformation, and softmax activation to yield the weight α. Subsequently, α is element-wise multiplied with E_d to obtain the weighted embedding $\tilde{E}_d = \alpha \otimes E_d$. Then, $\tilde{E}_d$ is passed through linear layers to generate modulation factors γ_c and β_c. These modulation factors rectify the dynamic range and baseline intensity of the feature, respectively, thereby compensating for the loss of detail and brightness at the feature level. The final adaptive modulation process is formulated as:

$$\tilde{F}_{vi} = F_{vi} \otimes \left(1 + \gamma_c\right) + \beta_c. \tag{7}$$

D. Cross-Attention Adaptive Fusion Module

Visible and infrared images exhibit significant differences in the representation of information, and the absence of adaptive cross-modal modeling often results in inadequate utilization of their complementary features. To address this, we propose the Cross-Attention Adaptive Fusion Module (CAAFM), which employs cross-attention to explicitly model cross-modal relationships, followed by channel and spatial attention to dynamically assign weights for the effective fusion of complementary information.

As shown in Fig. 4, we employ a projection function consisting of convolution and reshaping operations to project the VI modulated diffusion feature $\tilde{F}_{vi}$ and the IR diffusion feature F_{ir} into the corresponding Query (Q), Key (K), and Value (V), respectively, as presented in the following:

$$\{Q_{vi}, K_{vi}, V_{vi}\} = Reshape(Conv^{vi}_{qkv}(\tilde{F}_{vi})), \tag{8}$$

$$\{Q_{ir}, K_{ir}, V_{ir}\} = Reshape(Conv^{ir}_{qkv}(F_{ir})), \tag{9}$$

where $Conv(\cdot)$ and $Reshape(\cdot)$ correspond to the convolutional layer and the reshape operation, respectively. Attention weights for the visible and infrared modalities are computed using the following equations:

$$A_{vi} = \text{softmax}(Q_{vi}K_{ir}^{\top}),\ A_{ir} = \text{softmax}(Q_{ir}K_{vi}^{\top}). \tag{10}$$

Subsequently, we perform matrix multiplication between the attention weights and the value to aggregate cross-modal information. The aggregated features are then reshaped back to the spatial domain and added to the original features of the corresponding query branch via residual connections. This process is formulated as:

$$\hat{F}_{vi} = \tilde{F}_{vi} + Reshape(A_{vi}V_{ir}), \tag{11}$$

$$\hat{F}_{ir} = F_{ir} + Reshape(A_{ir}V_{vi}). \tag{12}$$

Then, $\hat{F}_{vi}$ and $\hat{F}_{ir}$ are concatenated channel-wise. The resulting tensor is passed through the parallel channel and spatial attention module to yield the fusion weights ω. The process of generating the fusion weights is as follows:

$$A_c = CB(CBR(GAP(CBR(C(\hat{F}_{vi}, \hat{F}_{ir}))))), \tag{13}$$

$$A_s = CB(CBR(CBR(C(\hat{F}_{vi}, \hat{F}_{ir})))), \tag{14}$$

$$\omega = \sigma(A_c \otimes A_s), \tag{15}$$

where $CBR(\cdot)$ denotes a convolutional block consisting of a 1×1 convolution, batch normalization, and ReLU activation, and $CB(\cdot)$ implies the same structure but omits the ReLU activation. $C(\cdot)$ and $\sigma(\cdot)$ denote the channel concatenation operation and sigmoid function, respectively. Finally, the adaptive fusion is realized via a gating mechanism controlled by the fusion weight ω, which can be expressed as:

$$F_f = \omega \otimes \hat{F}_{vi} + (1 - \omega) \otimes \hat{F}_{ir}. \tag{16}$$

E. Loss Function

With the aim of preserving thermal features, texture details, and accurate color reproduction in the fused results, we use a loss function composed of multiple loss terms to supervise the training of the fusion network. Specifically, we introduce

a multi-channel gradient loss [25] to preserve the rich texture details inherent in visible images, which is formulated as:

$$L_{MCG} = \frac{1}{HW} \sum_{i=1}^{3} \left\| \nabla I_f^i - \max(\nabla I_{vi}^i, \nabla I_{ir}) \right\|_1, \quad (17)$$

where ∇ represents the gradient operator. H and W represent the spatial dimensions (height and width) of the input image. i denotes the color channel index. $\|\cdot\|_1$ stands for the l_1-norm, while $\max(\cdot)$ is used to calculate the element-wise maximum. Thermal radiation is commonly represented by the intensity of pixels. To guarantee that the pixel intensity distribution of the fusion result aligns with that of the source images, we introduce a multi-channel intensity loss [25]. This can be expressed as:

$$L_{MCI} = \frac{1}{HW} \sum_{i=1}^{3} \left\| I_f^i - \max(I_{vi}^i, I_{ir}) \right\|_1. \quad (18)$$

To ensure color fidelity, we introduce a color consistency loss operating in the YCbCr space. It achieves natural color reproduction by constraining the discrepancy between the fused image and visible image in the Cb and Cr chrominance channels, defined as:

$$L_{CCL} = ||Cb_{I_f} - Cb_{I_{vi}}||_1 + ||Cr_{I_f} - Cr_{I_{vi}}||_1. \quad (19)$$

Finally, we define the total fusion loss L_{total} as:

$$L_{total} = \lambda_1 L_{MCG} + \lambda_2 L_{MCI} + \lambda_3 L_{CCL}, \quad (20)$$

where λ_1, λ_2, and λ_3 are hyperparameters that balance the losses.

IV. Experiments

A. Experimental Settings

To validate the fusion performance of the proposed DCAA-Fusion, we conducted extensive qualitative and quantitative experiments on the LLVIP [28], M^{3}FD [17], and MSRS [13] datasets, and compared our results with those of seven state-of-the-art methods, including FusionGAN [16], U2Fusion [12], DIVFusion [14], Dif-Fusion [25], DRMF [7], Crossfuse [29], and DAFusion [8]. We quantitatively evaluated the fusion performance using six metrics, including Entropy (EN), Mutual Information (MI), Average Gradient (AG), Spatial Frequency (SF), Standard Deviation (SD), and Visual Information Fidelity (VIF).

The training process was conducted in two stages. In Stage 1, both the visible conditional diffusion model and the infrared diffusion model were pretrained on the EMS dataset [26]. This dataset is designed for multi-degradation tasks, and contains 970 pairs of low-light degraded and normal-light visible images, along with 970 corresponding infrared images. The source images were randomly cropped to 160×160 and the batch size was set to 64. Except for these settings, all other key diffusion training configurations (including the noise schedule, the total diffusion steps T, the EMA setting, and the number of training iterations) were kept consistent with DRMF [7]. In Stage 2, 2,000 pairs of low-light degraded images were selected from the LLVIP dataset to train the fusion network. Specifically, intermediate feature maps were extracted from the frozen denoising networks ϵ_{vi} and ϵ_{ir} at timesteps $t = 50, 100,$ and 200 to serve as prior information for the training of subsequent fusion modules. Similarly, the source images were randomly cropped to 160×160, and the model was trained for 200 epochs with a batch size of 8 and a learning rate of 1×10^{-5}. The Adam optimizer was used for both training stages. Hyperparameters were set to $\lambda_1 = 1.5$, $\lambda_2 = 1$, and $\lambda_3 = 10$. The experiments were implemented using the PyTorch framework and executed on a workstation equipped with a single NVIDIA RTX 4090 GPU.

B. Results and Analysis

1) Qualitative Comparison: As illustrated in the first row of Fig. 5, FusionGAN and CrossFuse are unable to recover the details of the flowerbed in the green box and the fine textures of the pedestrians in the red box in low-light environments. Although U2Fusion and Dif-Fusion retain the fine texture of the pedestrians, they still struggle to adequately recover the details of the flowerbed. DIVFusion and DAFusion alleviate low-light degradation to some extent, but the flowerbed details remain blurred. While DRMF recovers clear flowerbed details, it loses the texture information of the pedestrians. Our approach effectively retains the details of the flowerbed while maintaining superior texture preservation of the pedestrians.

In addition, as shown in the second row of Fig. 5, U2Fusion, FusionGAN, Dif-Fusion, and CrossFuse fail to mitigate low-light degradation, resulting in the inability to recover the tree textures in the green box and the background details in the red box. Although DIVFusion and DAFusion alleviate low-light degradation, they exhibit color deviations, and the texture details remain blurred. While DRMF effectively recovers tree textures and background details, it shows limited performance in maintaining the saliency of infrared thermal targets. In contrast, our method simultaneously alleviates low-light degradation and effectively preserves both infrared thermal information and fine-grained visible textures and details.

2) Quantitative Comparison: Table I presents the average performance metrics of different fusion methods evaluated on 200 randomly selected image pairs from the LLVIP dataset. which are strictly non-overlapping with the training set. Our method outperforms the comparative methods, ranking first in four metrics and second in both EN and MI. The superior AG and SF scores demonstrate that our method captures richer spatial details and more effectively retains edge and texture information. The top VIF score shows that our fused results are closer to human visual perception. A larger SD value highlights that our method achieves higher contrast, resulting in a clearer distinction between thermal targets and the background in the fused results.

C. Generalization Experiment

1) Qualitative comparison: The third to sixth rows of Fig. 5 provide the visual comparisons between the proposed method and seven other comparative methods on the M^{3}FD and

Fig. 5. Visual comparison of our method with seven other methods on the LLVIP, M^{3}FD and MSRS datasets.

TABLE I
QUANTITATIVE COMPARISON OF OUR METHOD WITH SEVEN OTHER METHODS ON THE LLVIP, M^{3}FD AND MSRS DATASETS. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN **BOLD RED** AND **BOLD BLUE**, RESPECTIVELY.

Method	EN	AG	SF	SD	VIF	MI
LLVIP Dataset						
FusionGAN	6.614	3.085	10.342	31.007	0.426	2.549
U2Fusion	6.582	4.977	15.558	37.030	0.530	2.266
DIVFusion	**7.592**	5.798	16.164	**52.876**	0.752	2.159
Dif-Fusion	7.146	5.695	17.448	43.118	0.680	2.921
DRMF	7.357	**7.972**	**21.280**	52.061	**0.918**	**3.575**
CrossFuse	6.499	4.167	15.231	27.494	0.741	2.336
DAFusion	7.334	5.784	18.639	50.476	0.809	2.859
Ours	**7.489**	**8.510**	**23.245**	**53.412**	**1.021**	**2.975**
M^{3}FD Dataset						
FusionGAN	6.972	3.916	12.539	37.156	0.425	3.032
U2Fusion	7.102	6.103	17.853	39.812	0.620	2.861
DIVFusion	**7.558**	5.478	16.627	**52.950**	0.622	2.908
Dif-Fusion	6.876	5.733	17.795	34.911	0.527	3.286
DRMF	6.953	5.502	19.326	40.544	**0.837**	**5.117**
CrossFuse	6.677	4.587	15.274	32.360	0.719	3.531
DAFusion	7.206	**6.295**	**19.841**	51.832	0.778	3.463
Ours	**7.396**	**7.802**	**22.159**	**53.714**	**0.904**	**3.546**
MSRS Dataset						
FusionGAN	5.637	1.724	5.117	20.012	0.477	1.939
U2Fusion	5.314	2.606	8.288	23.558	0.525	2.030
DIVFusion	**7.386**	4.772	12.588	53.516	0.776	2.365
Dif-Fusion	6.687	3.951	11.722	42.989	0.835	3.346
DRMF	7.258	4.144	12.756	**56.361**	0.987	**4.768**
CrossFuse	6.533	3.039	9.681	37.620	0.838	3.130
DAFusion	7.170	**5.165**	**13.973**	47.808	**1.015**	2.636
Ours	**7.513**	**4.951**	**14.459**	**56.821**	**1.068**	**3.658**

MSRS datasets, respectively. Our method provides brighter scenes and effectively alleviates the low-light degradation issue. As observed in the red box, the pedestrian edges in our method are clearer and the contours more complete. The region within the green box shows that our method retains the edges of building window frames, as well as the fine texture details of billboards and stair handrails, more effectively than comparative methods.

2) Quantitative comparison: We randomly selected 200 image pairs from the M^{3}FD and MSRS datasets, respectively, for quantitative comparison. Table I shows the average performance metrics of our method and seven other different fusion methods. The results indicate that our method obtained the highest scores in SD, VIF, and SF, demonstrating that the fused images generated by the proposed method possess higher contrast, richer texture details, and improved visual information fidelity. To conclude, the proposed approach demonstrates favorable performance on both the M^{3}FD and MSRS datasets, validating its excellent generalization ability.

D. Ablation Study

We performed an ablation analysis to evaluate the DEMM and CAAFM modules, as well as the diffusion process and timesteps. These experiments were conducted on a subset of 200 image pairs randomly selected from the LLVIP dataset.

1) Ablation Study on the Proposed Modules: First, we removed the DEMM from the original model, as shown in Fig. 6(c). Without utilizing the degradation embedding extracted by the degradation-aware encoder to modulate visible diffusion features, the network cannot effectively mitigate low-light degradation, resulting in the fused image failing to fully represent scene information. Second, we replaced CAAFM with an average weighted fusion strategy, as illustrated in Fig. 6(d). Because of the absence of cross-attention for modeling cross-modal dependencies, and the absence of channel and spatial attention for calculating fusion weights, the fused image cannot fully integrate complementary information, leading to

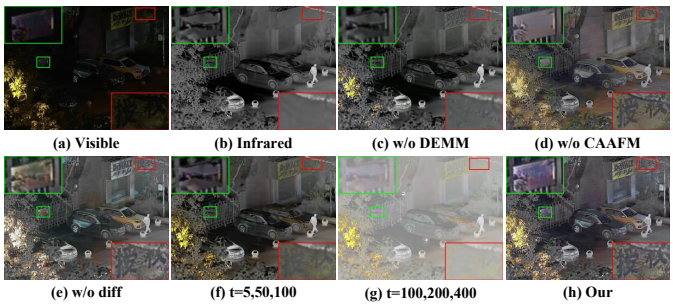

Fig. 6. Visual comparison of the ablation study on the LLVIP dataset.

TABLE II
QUANTITATIVE ABLATION ANALYSIS OF THE PROPOSED MODULES ON THE LLVIP DATASET.

Method	EN	AG	SF	SD	VIF	MI
w/o DEMM	6.972	5.749	16.548	42.865	0.748	2.917
w/o CAAFM	7.315	8.157	20.853	51.812	0.927	2.961
Ours	**7.489**	**8.510**	**23.245**	**53.412**	**1.021**	**2.975**

TABLE III
QUANTITATIVE ABLATION ANALYSIS OF THE DIFFUSION PROCESS AND TIMESTEPS ON THE LLVIP DATASET.

Timesteps	EN	AG	SF	SD	VIF	MI
w/o diff	6.987	7.812	20.370	49.012	0.796	2.561
5,50,100	7.150	7.920	21.981	52.192	0.891	2.706
100,200,400	6.154	3.071	11.512	31.756	0.503	2.361
Ours	**7.489**	**8.510**	**23.245**	**53.412**	**1.021**	**2.975**

the loss of detail within the green box and texture within the red box. The quantitative results of the ablation study in Table II show that only by incorporating both DEMM and CAAFM can the best performance be achieved.

2) Ablation on Diffusion Process and Timesteps: To investigate the impact of the diffusion process and different timesteps on fusion performance, we first removed only the diffusion process while maintaining the original network architecture, as demonstrated in Fig. 6(e). In the absence of diffusion prior guidance, although the fused results can alleviate low-light degradation to a certain extent, color information is severely lost. Subsequently, while retaining the diffusion process, we selected different combinations of timesteps to extract features, with the results illustrated in Fig. 6(f) and Fig. 6(g). When $t = \{5, 50, 100\}$ is chosen, the fusion results struggle to fully recover high-frequency texture details; when $t = \{100, 200, 400\}$ is chosen, significant information loss occurs in the fused results due to excessive noise intensity. The quantitative ablation results for the diffusion process and timesteps are presented in Table III, demonstrating that the fusion results achieve optimal performance only when the diffusion process is maintained with $t = \{50, 100, 200\}$.

E. Cost Comparison

To evaluate the computational cost of various diffusion model-based methods, we conducted extensive tests on a 24GB RTX 4090 GPU, with all experiments performed at a consistent resolution of 480×640. The detailed results are presented in Table IV. We compared key metrics, including the number of parameters, FLOPs, and runtime, where lower values indicate higher computational efficiency. The experimental results show that the proposed method offers favorable computational efficiency and is suitable for practical deployment.

TABLE IV
COMPARISON OF COMPUTATIONAL COSTS FOR DIFFERENT DIFFUSION MODEL-BASED METHODS

Method	Params.	FLOPs	running time
Dif-Fusion	416.47M	845.27G	0.91 s
DRMF	170.98M	246.32G	5.57 s
Ours	376.62M	277.83G	3.87 s

V. CONCLUSION

In this work, we present DCAAFusion, an infrared and visible image fusion network for nighttime low-light scenarios, aiming to address the loss of visible information induced by low-light degradation and the insufficient utilization of complementary modal information. Specifically, we utilize a pretrained conditional diffusion model and a diffusion model to extract visible and infrared diffusion features. The Degradation Embedding Modulation Module utilizes the degradation embedding extracted by the introduced degradation-aware encoder to adaptively modulate the visible diffusion features, explicitly compensating for the loss of brightness and details at the feature level. The Cross-Attention Adaptive Fusion Module utilizes cross-attention to model cross-modal dependencies and adaptively fuses infrared structural information and visible texture details through dynamic weighting via channel and spatial attention. Experimental results indicate that, when compared to seven state-of-the-art methods, our method achieves superior performance in both visual quality and quantitative metrics.

REFERENCES

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[2] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5906–5916.

[3] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information fusion*, vol. 45, pp. 153–178, 2019.

[4] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 808–823.

[5] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[6] X. Song and P. Dai, "Cman: Compact modality alignment network with dual stream transformer for visible-infrared person re-identification," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–9.

[7] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma, "Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8546–8555.

[8] X. Wang, Z. Guan, W. Qian, J. Cao, R. Ma, and C. Bi, "A degradation-aware guided fusion network for infrared and visible image," *Information Fusion*, vol. 118, p. 102931, 2025.
[9] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.
[10] H. Li, X.-J. Wu, and J. Kittler, "Rfn-nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.
[11] K. Cui, J. Cheng, and Y. Pan, "Fdafusion: A joint frequency decomposition and adaptive fusion network for nighttime infrared and visible image fusion," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.
[12] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 502–518, 2020.
[13] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, 2022.
[14] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "Divfusion: Darkness-free infrared and visible image fusion," *Information Fusion*, vol. 91, pp. 477–493, 2023.
[15] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 2022.
[16] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," *Information fusion*, vol. 48, pp. 11–26, 2019.
[17] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5802–5811.
[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
[19] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.
[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.
[22] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
[23] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9935–9946.
[24] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "Ddfm: denoising diffusion model for multi-modality image fusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8082–8093.
[25] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Transactions on Image Processing*, vol. 32, pp. 5705–5720, 2023.
[26] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 026–27 035.
[27] T. Wang, K. Zhang, Y. Zhang, W. Luo, B. Stenger, T. Lu, T.-K. Kim, and W. Liu, "Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement," *Pattern Recognition*, vol. 166, p. 111628, 2025.
[28] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.
[29] H. Li and X.-J. Wu, "Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Information Fusion*, vol. 103, p. 102147, 2024.