HMIF:An Efficient Infrared and RGB Image Fusion Network Based on HVI

Yi Li, Huajun Wang

***Abstract*—Infrared and visible image fusion combines the thermal target information from infrared images with the textural details from visible images, significantly enhancing scene perception. However, differences in image acquisition between infrared and RGB sensors introduce key challenges. Firstly, RGB images are susceptible to luminance noise under low-light conditions, degrading fusion quality. Secondly, existing methods often simply superimpose multi-sensor information, neglecting the distinct characteristics of different modalities, which leads to loss of critical information or redundancy. To address these limitations, we propose the HVI Adaptive Brightness Enhancement module to improve fusion stability and accuracy under low-light conditions. Furthermore, based on the existing EMMA framework, we integrate the HVI module with mainstream fusion modules to construct an asymmetric dual-branch fusion architecture. Experimental results demonstrate that the proposed method achieves superior subjective visual quality and objective evaluation metrics in infrared-visible image fusion tasks, effectively enhancing fusion performance in complex scenes.**

***Index Terms*—Image fusion, unsupervised learning,**

I. Introduction

Visible and Infrared Image Fusion (VIF) synthesizes images captured by different sensors, typically infrared and visible spectrum images, into a single composite representation. The core objective is to integrate complementary and redundant information from these source images, generating a fused result richer in information than any individual input. This enhanced imagery significantly benefits downstream processing tasks across diverse fields including digital photography, remote sensing, agriculture, medicine, and biometrics [**?**]. Visible images exhibit high spatial resolution and rich textural detail, closely aligning with human visual perception. However, their quality is highly vulnerable to occlusion, illumination variations, and adverse weather conditions. Infrared imaging, based on thermal radiation capture, offers superior penetration capabilities (through media like smoke or mist) and effectively distinguishes materials through emissivity differences. Infrared images typically suffer from lower spatial resolution, edge blurring due to thermal diffusion effects, and prominent non-uniformity noise and temporal noise.

Consequently, a primary challenge in VIF stems from the fundamentally distinct physical properties captured by each modality: thermal radiation versus light reflection. This divergence places their respective features on separate manifolds [**?**]. Naïve linear fusion strategies, such as weighted averaging, frequently induce feature conflict. For example, a high-temperature region appearing bright in the infrared image might correspond to a low-brightness region in the visible image. Existing VIF approaches fall primarily into traditional methods and deep learning-based techniques. Traditional methods typically follow a three-stage process: feature extraction (e.g., texture from visible images and salient thermal targets from infrared images) using specific transforms like multi-scale decomposition; application of predefined fusion rules within the transformed domain; and reconstruction of the fused image via the inverse transform [**?**]. The reliance on manually designed steps and parameters limits their robustness and ability to produce high-quality fusion results across diverse real-world scenarios [**?**].

Driven by significant advancements in deep learning for image processing, substantial research focuses on leveraging these techniques for VIF [**?**]. Based on network architecture, deep VIF methods are categorized into CNN-based, Autoencoder-based, Generative Adversarial Network-based, Transformer-based, and hybrid approaches. CNN-based methods ([**?**]) effectively capture local features but struggle with long-range dependencies. Autoencoder-based methods, such as DenseFuse [**?**], pre-train an encoder-decoder structure using infrared and visible image pairs. Hand-crafted rules (e.g., addition, L1-norm) then merge the extracted features before the decoder reconstructs the fused image. While yielding reasonable results, dependence on manual fusion rules often leads to suboptimal performance. Generative Adversarial Network-based methods (e.g., DDcGAN [**?**]) employ discriminators to guide generators towards producing fused images matching source distributions. Nevertheless, these methods suffer from interpretability issues, control difficulties, training instability, and potential texture distortion. Transformer-based approaches (e.g., SePT [**?**]) leverage self-attention for effective long-range modeling, albeit often with high computational complexity. Hybrid architectures represent the current mainstream, aiming to combine advantages. For instance, CMFuse [**?**] focuses on optimizing cross-modal interactions, while EMMA [**?**] utilizes outputs from existing fusion algorithms to learn sensor distribution properties for training guidance. Hybrid methods still demonstrate limitations in complex scenarios involving uneven illumination and strong noise.

Despite progress in integrating key complementary information from infrared and visible images, significant challenges persist: (1) Visible images suffer severe detail loss in underexposed/overexposed regions under extreme illumination variations, which existing algorithms inadequately address; (2) Predominant frameworks typically apply symmetric feature extraction paths [**?**], hindering integration with adaptive enhancement strategies for visible inputs;

To address these limitations, this paper presents three core contributions:

- **HVI Adaptive Brightness Enhancement Strategy:** We propose an HVI adaptive brightness enhancement strategy. This strategy effectively enhances the stability of the fusion process and the robustness of the results under low-light and uneven illumination conditions.
- **Asymmetric Dual-Branch Fusion Framework:** We construct an asymmetric dual-branch fusion framework. This framework, termed the Asymmetric Dual-Branch Fusion Framework , integrates the HVI module with a mainstream fusion module. It employs a two-stage fusion approach to separate the enhanced visible image and the original infrared image.
- **Experimental Results:** Experimental results demonstrate that our fusion model significantly improves the color richness of visible images, exhibits stronger detail preservation capabilities, mitigates noise interference, and markedly enhances the visibility of target objects within scenes.

II. METHODOLOGY

A. HVI Transformation

Low-Light Image Enhancement (LLIE) aims to improve visual quality under poor illumination conditions, with core challenges involving effective noise suppression and color distortion mitigation. Current methodologies predominantly process images in standard RGB space [**?**]. However, natural scene imagery exhibits high sensitivity to brightness fluctuations due to illumination variations, occlusions, and shadows. The fundamental limitation of RGB representation lies in the tight coupling of luminance and chrominance information: brightness variations induce correlated changes across all three channels (R, G, B), preventing independent manipulation of lightness and color.

To address this limitation, conversion to HSV color space is frequently employed. This perceptual model decouples color attributes into three orthogonal components:

$$\text{Hue (H)}: \text{ dominant wavelength} \tag{1}$$

$$\text{Saturation (S)}: \text{ color purity} \tag{2}$$

$$\text{Value (V)}: \text{ brightness intensity} \tag{3}$$

The separation facilitates improved color denoising and luminance restoration. RGB-to-HSV conversion proceeds as follows. Given $R, G, B \in [0, 255]$, first normalize components:

$$R' = \frac{R}{255}, \quad G' = \frac{G}{255}, \quad B' = \frac{B}{255} \tag{4}$$

Compute extrema and differential:

$$C_{\text{max}} = \max(R', G', B') \tag{5}$$

$$C_{\text{min}} = \min(R', G', B') \tag{6}$$

$$\Delta = C_{\text{max}} - C_{\text{min}} \tag{7}$$

Hue (H) is then derived:

$$H = \begin{cases} 0^\circ & \Delta = 0 \\ 60^\circ \times \left(\frac{G'-B'}{\Delta} \mod 6 \right) & C_{\text{max}} = R' \\ 60^\circ \times \left(\frac{B'-R'}{\Delta} + 2 \right) & C_{\text{max}} = G' \\ 60^\circ \times \left(\frac{R'-G'}{\Delta} + 4 \right) & C_{\text{max}} = B' \end{cases} \tag{8}$$

Saturation (S) and Value (V) follow:

$$S = \begin{cases} 0 & C_{\text{max}} = 0 \\ \Delta / C_{\text{max}} & \text{otherwise} \end{cases}, \quad V = C_{\text{max}} \tag{9}$$

Despite advantages in luminance-chrominance separation, HSV introduces noticeable red-band artifacts and dark-region noise during enhancement, particularly degrading perceptual quality in saturated reds and shadows. To overcome these artifacts, we introduce the HVI color space [**?**], which decomposes sRGB images into:

$$\text{Intensity map } (I_{\text{max}}): \text{ scene illuminance} \tag{10}$$

$$\text{HV color map } (\hat{H}, \hat{V}): \text{ chromatic structure} \tag{11}$$

where I_{max} is computed per-pixel as:

$$I_{\text{max}}(x) = \max_{c \in \{R,G,B\}} I_c(x) \tag{12}$$

HVI eliminates hue discontinuities through polar encoding:

$$H = \cos\left(\frac{\pi h}{3}\right), \quad V = \sin\left(\frac{\pi h}{3}\right) \tag{13}$$

where $h \in H$ and $v \in V$ represent orthogonal components. This formulation ensures continuous representation across the hue spectrum (notably eliminating red-region artifacts) while minimizing Euclidean distances between perceptually similar colors.

Noise amplification in low-light regions is suppressed via adaptive intensity compression:

$$C_k(x) = \sqrt[k]{\sin\left(\frac{\pi I_{\text{max}}(x)}{2}\right) + \varepsilon} \tag{14}$$

where the learnable parameter k controls compression strength. This function implements radial mapping, with small C_k values concentrating low-intensity regions. The transformed components are then obtained by:

$$\hat{H} = C_k \odot S \odot H, \quad \hat{V} = C_k \odot S \odot V \tag{15}$$

where $\odot$ denotes the Hadamard product. Final HVI representation concatenates $\hat{H}$, $\hat{V}$, and I_{max}.

In multispectral fusion frameworks, visible images transformed to HVI yield HV maps $(\hat{H}, \hat{V})$ and I_{max}, while infrared images provide complementary intensity data. These components are channel-concatenated and fed into enhancement modules, leveraging HVI's artifact suppression and signal fidelity preservation properties.

B. Enhance HVI Fusion Module

The enhancement fusion module comprises an HV feature branch and an intensity (I) feature branch. Initial feature extraction is performed using 3×3 convolutions. Subsequently, a U-Net architecture-based fusion network (Fig. **??**) integrates multimodal features. This network embeds multiple Cross-Attention Blocks (CABs) and incorporates skip connections between its encoder and decoder.

The cross-attention mechanism establishes bidirectional guidance between the two branches. Features extracted by the intensity (I) branch dynamically modulate feature responses

in the HV branch via attention weights, enabling brightness-adaptive noise suppression (enhanced denoising in low-light regions while preserving original color in bright regions). Crucially, the optimized HV features simultaneously feed back structural and textural color information to the I branch, constraining the enhancement process to mitigate halo artifacts. This design is grounded in the physical observation that noise intensity varies inversely with illumination intensity, necessitating region-specific processing.

The decoder outputs two enhanced feature maps: the enhanced HV features ($\hat{HV}_e$) and the enhanced intensity features ($\hat{I}_e$). These features are concatenated along the channel dimension and fed into the Post-HVI Inverse Transform (PHVIT) module to reconstruct the enhanced sRGB output image.

The PHVIT module reverses the transformation. It decomposes $\hat{HV}_e$ into its orthogonal components $\hat{H}_e$ and $\hat{V}_e$ using learned parameters C_k and ϵ, computing intermediate variables $\hat{h}$ and $\hat{v}$:

$$\begin{aligned} \hat{h} &= \frac{\hat{H}_e}{C_k + \epsilon} \\ \hat{v} &= \frac{\hat{V}_e}{C_k + \epsilon} \end{aligned} \tag{16}$$

Subsequently, $\hat{h}$, $\hat{v}$, and $\hat{I}_e$ are transformed back to the standard HSV color space:

$$\begin{aligned} H_{new} &= \arctan(\frac{\hat{v}'}{\hat{h}'}) \mod 1 \\ S_{new} &= \alpha_s \sqrt{(\hat{v}')^2 + (\hat{h}')^2} \\ V_{new} &= \alpha_i \hat{I}_e \end{aligned} \tag{17}$$

where α_s and α_i are learnable linear parameters scaling saturation and intensity weights, respectively. Finally, the components H_{new}, S_{new}, and V_{new} are converted from the HSV color space to the standard sRGB color space, yielding the enhanced output image.

C. Enhanced EMMA Framework with Equivariant Imaging Prior

To effectively leverage multi-modal fusion information, we propose the Enhanced Multi-modal self-supervised fusion frAmework (EMMA), which integrates an equivariant imaging prior. EMMA adapts the inherent pattern of its loss function by simulating the sensing imaging process based on natural imaging principles. The core workflow of EMMA is illustrated in Figure **??**.

During training, the original infrared (i) and visible (v) images are first input into the HVI Fusion network (Fig. **??**(a)). This network acts as the fusion backbone, extracting and integrating information from both modalities to produce a preliminary fused feature map f.

Subsequently, f is processed by the Equivariant Module (Fig. **??**(b)), which applies predefined geometric transformations (e.g., rotation, translation, reflection) to generate a transformed fused image f_t. This step is grounded in the Equivariant Imaging Prior, which posits that the representation of a fusion result should exhibit corresponding consistency under these transformations. The augmentation enhances model robustness and generalization in the absence of direct supervision.

The transformed image f_t is then fed into two independent, pre-trained pseudo-sensor modules: an infrared pseudo-sensor module A_i and a visible pseudo-sensor module A_v. Both modules utilize a standard U-Net architecture. These modules are trained with pseudo ground truth generated from state-of-the-art fusion algorithms to learn the mapping from the fused image back to the original source images. During EMMA's main training phase, the weights of these pseudo-sensor modules remain frozen; they function solely as "inverse decoders" performing forward inference to assess whether f_t retains sufficient information for reconstructing the source images. Consequently, A_i infers a pseudo-infrared image $\hat{i}_t$ and A_v infers a pseudo-visible image $\hat{v}_t$.

Finally, the pseudo-infrared $\hat{i}_t$ and pseudo-visible $\hat{v}_t$ images are cycled back as input into the same HVI Fusion network, producing the final fused output $\hat{f}_t$.

D. Loss Function

Our framework incorporates loss functions for infrared images, visible images, and the final fused output. The total loss L_{total} is computed as a weighted combination of these three components:

$$\begin{aligned} L_{\text{total}} = L(A_i(f), i) + \alpha_1 L(A_v(f), v) \\ + \alpha_2 L(f_{\hat{t}}, f_t) \end{aligned} \tag{18}$$

where $L(\hat{x}, x) = \ell_1(x, \hat{x}) + \ell_1(\nabla x, \nabla \hat{x})$. Here, ∇ denotes the **Sobel operator** for computing image gradients, and α_1, α_2 are tuning parameters that balance the contribution of each loss term.

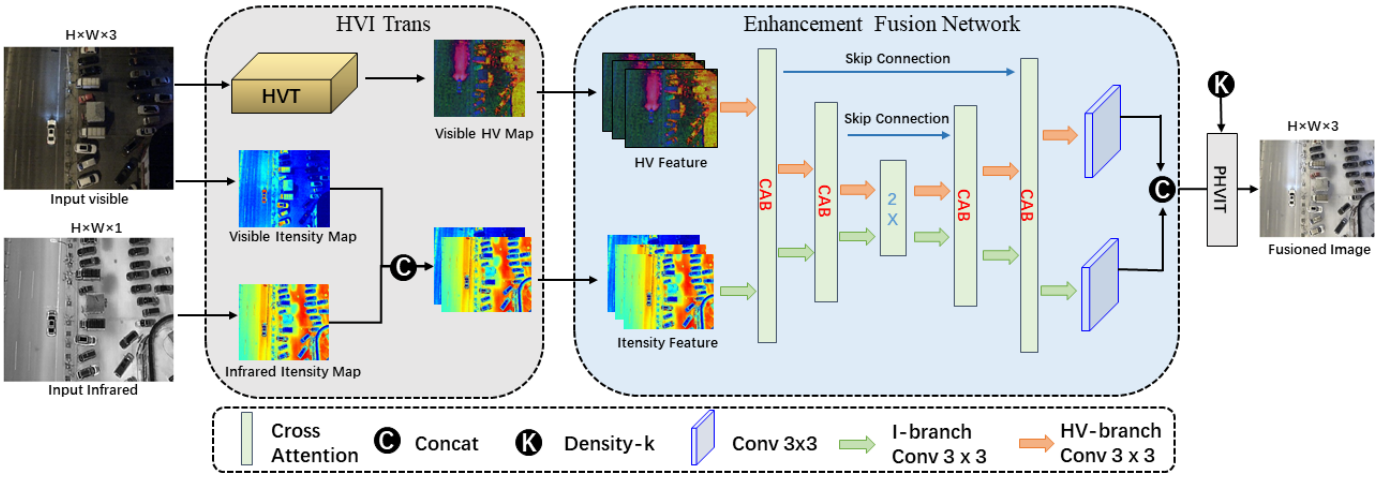

Fig. 1: Overview of the Enhanced Fusion Network Architecture. (a) HVI Color Transformation Module (HVIT): Transforms visible light into the HVI color space, obtaining its HV channels and intensity map, alongside the infrared image's intensity map. (b) Enhanced Fusion Network: Employs a dual-branch UNet architecture as its core, integrating six lightweight Cross-Attention Block (CAB) modules. (c) Perceptual Inverse HVI Transformation (PHVIT): Converts the enhanced HVI representation back into the RGB format, producing the final enhanced fused image.

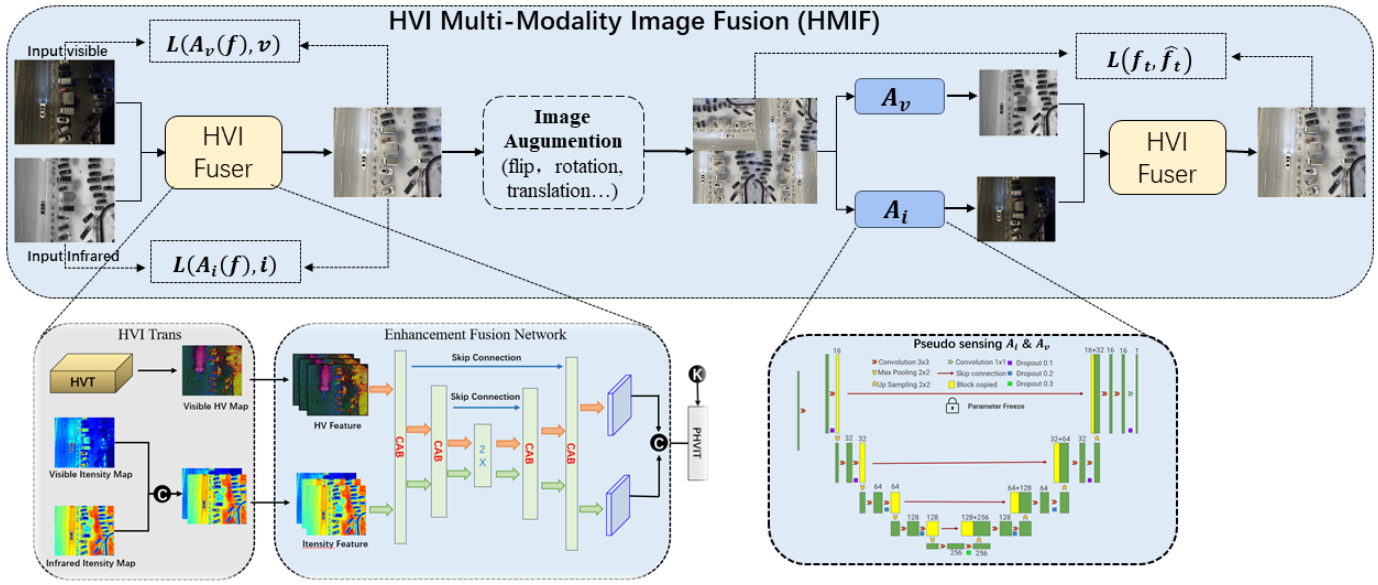

Fig. 2: EMMAaHVI FusionEnhanced HVI Fusion NetworkbcUNetfreeze

III. Experiments

A. Evaluation Metrics

To quantitatively assess algorithm performance, we employ seven metrics: Average Gradient (AG), Edge Intensity (EI), Entropy (EN), Mutual Information (MI), Standard Deviation (SD), Spatial Frequency (SF), and Structural Similarity Index (SSIM). Their mathematical formulations are detailed below.

Average Gradient (AG) quantifies image sharpness using spatial derivatives:

$$\mathrm{AG} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \sqrt{\frac{(\partial I/\partial i)^2 + (\partial I/\partial j)^2}{2}} \tag{19}$$

where M and N represent image width and height, and $I(i, j)$ is the pixel intensity at location (i, j).

Edge Intensity (EI) measures edge strength based on Sobel filtering:

$$EI(F) = \frac{1}{MN} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} [(F * h_x)(i, j)^2 + (F * h_y)(i, j)^2]} \tag{20}$$

Here, $*$ denotes convolution, F is the fused image, and h_x, h_y are Sobel kernels in the horizontal and vertical directions.

Entropy (EN) evaluates the information content of an image:

$$\mathrm{EN} = -\sum_{k=0}^{L-1} p(k) \log_2 p(k) \tag{21}$$

where $p(k)$ is the probability of occurrence of gray level k in

the image histogram, and L is the number of possible gray levels.

Mutual Information (MI) assesses the dependency between source images A, B and the fused image F:

$$\begin{aligned} \text{MI} = &\sum_{a,f} p_{A,F}(a,f) \log\left(\frac{p_{A,F}(a,f)}{p_A(a)p_F(f)}\right) \\ &+ \sum_{b,f} p_{B,F}(b,f) \log\left(\frac{p_{B,F}(b,f)}{p_B(b)p_F(f)}\right) \end{aligned} \tag{22}$$

where $p_{A,F}(a,f)$ and $p_{B,F}(b,f)$ are joint probability distributions, and $p_A(a)$, $p_B(b)$, $p_F(f)$ are marginal distributions.

Standard Deviation (SD) indicates overall contrast:

$$\text{SD}(F) = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I(i,j)-\mu_F\right)^2} \tag{23}$$

with μ_F representing the mean intensity of the fused image F.

Spatial Frequency (SF) captures global activity level using row and column frequencies:

$$\text{SF} = \sqrt{\text{RF}^2 + \text{CF}^2} \tag{24}$$

$$\text{RF} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=2}^{N}\left[I(i,j)-I(i,j-1)\right]^2} \tag{25}$$

$$\text{CF} = \sqrt{\frac{1}{MN}\sum_{i=2}^{M}\sum_{j=1}^{N}\left[I(i,j)-I(i-1,j)\right]^2} \tag{26}$$

Structural Similarity Index (SSIM) quantifies perceptual similarity between reference image x and fused image y:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{27}$$

where $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$ are mean intensities, σ_x^2 and σ_y^2 are variances, $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$ is the covariance, $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$. The constants are set as $K_1 = 0.01$, $K_2 = 0.03$, and L is the dynamic range ($2^{\text{bits per pixel}} - 1$).

B. Datasets

VisDrone-DroneVehicle: ?] contains 56,878 drone-captured images. This dataset comprises 28,439 visible RGB images and their corresponding 28,439 infrared images, providing a comprehensive paired set of visible and infrared modality pairs for cross-modality analysis.

M3SVD Dataset: ?] offers 220 pairs of spatiotemporally aligned infrared and visible video sequences. These sequences can be decomposed into 153,797 temporally and spatially synchronized RGB-IR image pairs. The dataset encompasses diverse scenarios, including daytime and nighttime environments, along with challenging conditions such as low illumination, overexposure, and camera jitter.

C. Experimental Results

To comprehensively evaluate the performance of our proposed method, we conducted rigorous quantitative comparisons and generalization validation across multiple challenging datasets. For quantitative assessment, four state-of-the-art multimodal fusion methods were selected as baselines: DDFM [**?**], QuadPrior [**?**], GSAD [**?**], and EMMA [**?**]. Experimental results demonstrate that our HMIF model outperforms the baseline methods across multiple quantitative metrics, confirming its effectiveness in fusion quality.

As shown in Tables **??** and **??**, the HMIF model exhibits significantly superior overall performance compared to existing advanced models. Specifically (Table **??**), while HMIF shows a slight decrease in SD (-0.08) compared to EMMA, it achieves substantial improvements in key perceptual quality metrics: AG increased by 0.9 and EI improved by 4.29. This notable performance improvement primarily results from the proposed HVI adaptive brightness enhancement strategy. This strategy robustly adaptively adjusts low-light and unevenly illuminated regions, effectively enhancing the stability of the fusion process. Consequently, it improves the contrast and clarity of the fused images, thus significantly boosting metrics like AG and EI which reflect structural detail and sharpness.

To thoroughly assess the model's generalization capability, we performed testing on the M3SVD dataset. Representative samples, frame_00754.jpg from video "0118_1904" and frame_00000.jpg from video "1208_1717", were selected for detailed analysis. The results (Tables **??** and **??**) confirm that HMIF consistently maintains superior performance in these new scenarios, demonstrating its robustness and broad applicability. As detailed in Table **??**, although HMIF slightly trails EMMA in EN, it surpasses all existing state-of-the-art methods across all other key metrics, including AG, EI, and SD. This strong generalization capability stems from the adaptability of the HVI strategy to diverse illumination conditions and the effectiveness of our designed asymmetric dual-branch fusion framework. This framework specifically optimizes the processing of HV (Chroma and Luma) features from the enhanced visible image branch and intensity features from the fused infrared and visible image branch. This dual optimization effectively enhances the model's fusion performance under complex and varied real-world conditions. Collectively, the quantitative evaluations and cross-dataset experiments validate the superiority of the HMIF fusion model.

TABLE I: Comparative evaluation on DroneVehicle dataset (Image: 17647.jpg)

	AG	EI	EN	MI	SD	SF	SSIM
DDFM	5.23	90.02	**7.64**	1.72	52.37	10.42	0.87
QuadPrior	5.24	97.68	7.44	1.79	55.85	10.43	**1.07**
GSAD	5.98	102.48	7.48	1.84	50.86	11.55	0.99
EMMA	6.27	103.64	7.61	2.03	**61.72**	12.02	0.89
gray!40 Ours	**6.36**	**107.93**	7.62	**2.13**	61.64	**12.24**	1.03

TABLE II: Comparative evaluation on DroneVehicle dataset (Image: 17874.jpg)

	AG	EI	EN	MI	SD	SF	SSIM
DDFM	5.00	109.16	7.33	1.62	51.16	9.69	0.82
QuadPrior	5.47	114.04	7.61	1.61	55.63	10.64	0.92
GSAD	5.33	116.53	**7.62**	1.89	57.15	10.59	0.79
EMMA	5.83	126.88	7.48	1.54	**62.09**	11.23	**1.13**
gray!40 Ours	**6.09**	**127.67**	7.58	**1.92**	61.94	**11.65**	1.10

TABLE III: Comparative evaluation on M3SVD dataset (Video: 0118_1904, Frame: 00754)

	AG	EI	EN	MI	SD	SF	SSIM
DDFM	2.44	55.82	6.87	2.14	53.19	6.28	1.13
QuadPrior	2.31	56.32	6.88	**3.53**	59.55	6.46	0.95
GSAD	2.94	63.02	6.55	2.12	52.39	7.15	1.03
EMMA	2.97	63.31	**7.35**	2.97	63.93	7.13	**1.25**
gray!40 Ours	**3.07**	**63.45**	7.27	3.44	**64.75**	**7.43**	1.19

D. Visualization

Visual comparisons between HMIF and state-of-the-art methods are presented in Figures **??**, **??**, and **??**. The results demonstrate that our method effectively integrates thermal radiation information from infrared images with the fine structural details of visible images, while more faithfully preserving color information. Figure **??** illustrates the comparative results on samples ”17605.jpg” and ”17611.jpg” from the DroneVehicle dataset. Observations reveal that fusion outputs from DDFM and QuadPrior exhibit noticeable edge blurring and noise interference. Although EMMA shows improved fusion performance, its ability to preserve color fidelity remains comparatively weaker. In contrast, by incorporating the HVI adaptive brightness enhancement strategy, our model significantly enhances the color richness of the fused results while preserving finer details, leading to substantially improved visibility of target objects within the scene.

TABLE IV: Comparative evaluation on M3SVD dataset (Video: 1208_1717, Frame: 00000)

	AG	EI	EN	MI	SD	SF	SSIM
DDFM	5.11	181.15	6.83	2.59	72.42	11.00	0.87
QuadPrior	5.17	215.24	6.92	2.90	73.33	11.05	0.73
GSAD	5.36	211.55	7.72	2.46	68.07	11.20	0.88
EMMA	5.34	211.09	**7.34**	3.27	78.28	11.10	0.91
gray!40 Ours	**5.42**	**219.72**	7.29	**3.34**	**80.57**	**11.22**	**0.99**

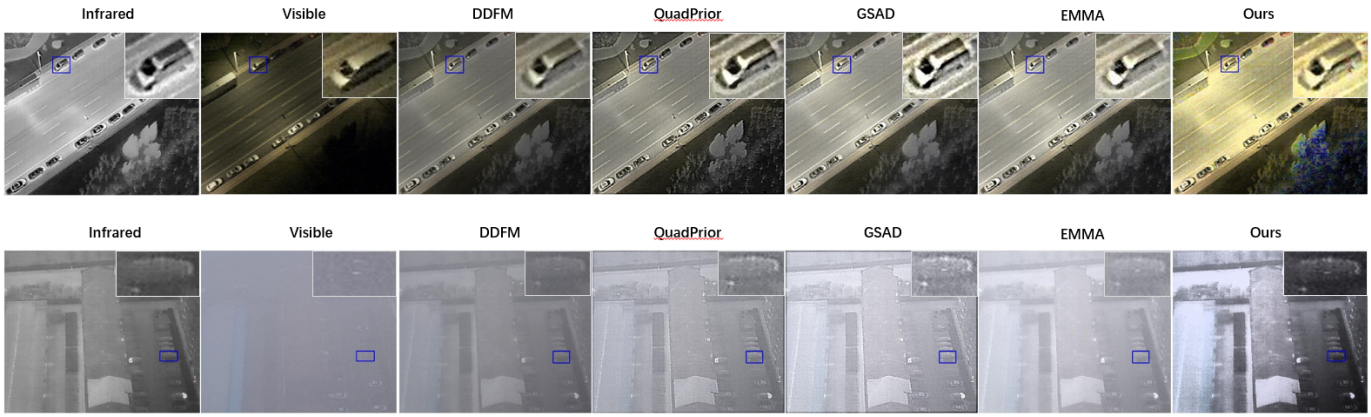

Fig. 3: Comparative results for image pairs ”17605.jpg” and ”17611.jpg” from the DroneVehicle dataset

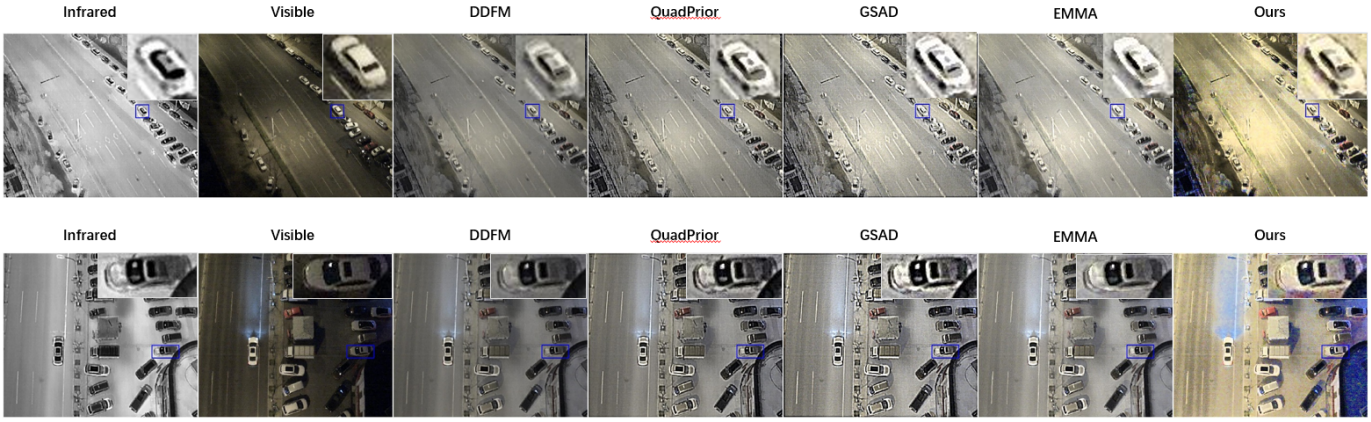

Fig. 4: Comparative results for image pairs ”17647.jpg” and ”17874.jpg” from the DroneVehicle dataset

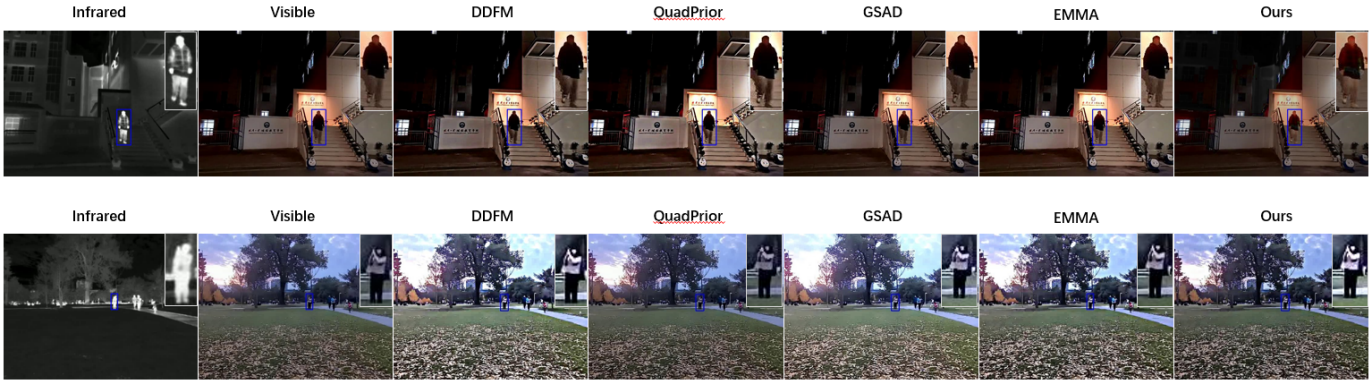

Fig. 5: Comparative results for frame 00177 (video 0118_1904) and frame 00000 (video 1208_1717) from the M3SVD dataset

IV. Conclusion