Dispel Darkness for Better Fusion: A Controllable Visual Enhancer based on Cross-modal Conditional Adversarial Learning

Hao Zhang, Linfeng Tang, Xinyu Xiang, Xuhui Zuo, Jiayi Ma*
Electronic Information School, Wuhan University, Wuhan, China
{zhpersonalbox, linfeng0419, jyma2010}@gmail.com, {xiangxinyu, xuhuizuo2001}@whu.edu.cn

Abstract

We propose a controllable visual enhancer, named DDBF, which is based on cross-modal conditional adversarial learning and aims to dispel darkness and achieve better visible and infrared modalities fusion. Specifically, a guided restoration module (GRM) is firstly designed to enhance weakened information in the low-light visible modality. The GRM utilizes the light-invariant high-contrast characteristics of the infrared modality as the central target distribution, and constructs a multi-level conditional adversarial sample set to enable continuous controlled brightness enhancement of visible images. Then, we develop an information fusion module (IFM) to integrate the advantageous features of the enhanced visible image and the infrared image. Thanks to customized explicit information preservation and hue fidelity constraints, the IFM produces visually pleasing results with rich textures, significant contrast, and vivid colors. The brightened visible image and the final fused image compose the dual output of our DDBF to meet the diverse visual preferences of users. We evaluate DDBF on the public datasets, achieving state-of-the-art performances of low-light enhancement and information integration that is available for both day and night scenarios. The experiments also demonstrate that our DDBF is effective in improving decision accuracy for object detection and semantic segmentation. Moreover, we offer a user-friendly interface for the convenient application of our model. The code is publicly available at https://github.com/HaoZhang1018/DDBF.

1. Introduction

Infrared and visible modality fusion is an enhancement technology that aims to integrate the advantages of both modalities to produce fused images with rich textures, significant contrast, and vibrant colors [3, 20, 25, 27, 50]. Due to the outstanding visual characteristics of fused images, infrared and visible modality fusion has found extensive applications in adverse environments such as nighttime. Thus, it is of great interest to develop an effective fusion algorithm to achieve night visual enhancement, thereby promoting the performance of tasks such as video surveillance and vehicle navigation [4, 32, 38, 43, 54].

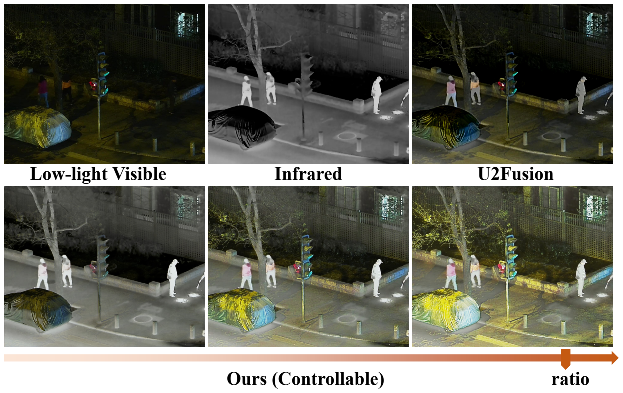

Figure 1. An example of infrared and visible image fusion in low-light environment.

The advancement of deep learning has significantly accelerated progress in the field of infrared and visible modality fusion in recent decades [14, 15, 19, 21, 44, 47, 52, 53]. Deep fusion methods utilize specific network architectures to automatically extract and fuse features and reconstruct images under the guidance of meticulously designed loss functions [17, 28, 43, 46, 48], achieving promising visual performance. However, there are still several challenges that need to be addressed.

Firstly, *the design concept for the loss function used to guide texture and contrast preservation still remains at the level of the multi-modal weighted game (see Eq. (1)), that is, setting multiple optimization objectives in the same domain*. Such a multi-objective optimization [37] will force the network to sacrifice the optimal solutions in each domain during the learning process in favor of minimizing the total loss, resulting in weakened texture and contrast in the fused image [26, 42]. Secondly, nearly all deep fusion

[*] Corresponding author

methods cannot directly handle color visible images [39]. Instead, they adopt a color separation strategy (see Eq. (3)) to process only the luminance component (*e.g.*, Y), while preserving color by duplicating the chrominance components (*e.g.*, Cb and Cr) [34]. However, *a problem with this strategy is that the fixed chrominance components cannot adapt to the changed fused luminance component, leading to color distortions*. Thirdly, existing fusion methods overlook the weakening of information in the visible modality during nighttime imaging. The information mismatch between the infrared and visible modalities results in the loss of beneficial features during the fusion process, ultimately leading to unsatisfactory visual results.

Considering these challenges, we propose DDBF, a controllable visual enhancer based on cross-modal conditional adversarial learning, which aims to dispel the darkness and achieve better fusion. Firstly, we utilize the light-invariant high-contrast characteristic of infrared images as the target prior, and develop a guided restoration module (GRM) that can drive low-light visible images towards having high-illumination distribution. In GRM, we perform data augmentation for infrared images to construct a multi-level conditional adversarial sample set, which facilitates discriminative illumination approximation in the conditional adversarial learning mechanism. Through continuous adversarial learning, the condition input enables the GRM to support a customized enhancement ratio, thus flexibly recovering the information of the visible modality from various degrees of low-light environments. After significantly reducing the information mismatch between infrared and visible modalities at night with GRM, we develop a promising information fusion module (IFM) to solve the problems of texture and contrast weakening as well as color distortion. On the one hand, *we depart from the idea of a multi-modal weighted game and instead formulate a clear optimization objective for preserving sharpened texture and significant contrast*. On the other hand, *we introduce a novel hue fidelity constraint to replace the commonly used color separation strategy, which can adaptively retain satisfying colors*. As a result, our proposed DDBF provides clear and visually pleasing results, allowing for unobstructed viewing in low-light conditions, as shown in Fig. 1.

In summary, we make the following contributions:

- We propose a controllable visual enhancer based on cross-modal conditional adversarial learning to dispel the darkness for better fusing infrared and visible images, which greatly improves the visibility of imaging in low-light environments. To our knowledge, this is the first attempt in the field of image fusion to controllably address the challenge of information loss caused by visible modality degradation at night.
- A guided restoration module is designed to effectively recover the scene information lost in low-light visible modality. By establishing conditional adversarial learning based on the light-invariant contrast of infrared modality, it achieves controllable enhancement while getting rid of the dependence on reference images.
- We develop a novel information fusion module, in which the customized explicit information preservation and hue fidelity constraints can solve the problems of contrast and texture information loss, and color distortion that are common in current methods.

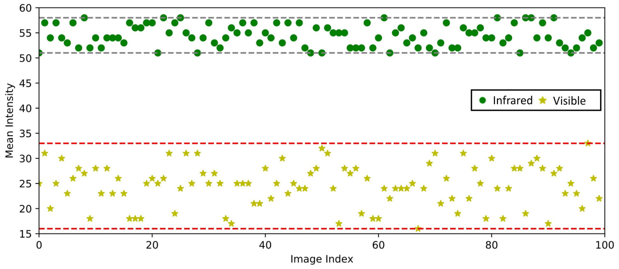

Figure 2. Statistical average intensity of infrared and visible images in different lighting environments that are captured by the same surveillance camera.

2. Background and Motivations

Light-invariant Contrast of Infrared Modal. Infrared images are generated by capturing the thermal radiation emitted by objects, which usually have significant contrast and do not vary with changes in illumination. We randomly select 100 pairs of infrared and visible images, captured by the same fixed multi-mode camera under varying low-light conditions. Average intensities are calculated, and Fig. 2 presents a scatter plot. Infrared images exhibit little intensity variation, indicating that for visible images captured under different lighting conditions, the corresponding infrared images can always provide reliable contrast guidance.

Information Mismatch. There exists an information mismatch between infrared and low-light visible images, as the values used to characterize their appearance attributes are not at the same scale. For example, Fig. 2 demonstrates that the average intensity values of the low-light visible image are significantly lower than those of the infrared image. Therefore, the appearance contrast in the fusion process inevitably deviates, resulting in the loss of some effective visible-modal information hidden in the darkness.

Multi-modal Weighted Game. Owing to the absence of ground truth in the task of infrared and visible modality fusion, most existing deep fusion methods adopt the idea of a multi-modal weighted game to define the loss function:

$$\mathcal{L} = \alpha_1 \|T(F) - T(A)\| + \alpha_2 \|T(F) - T(B)\|, \qquad (1)$$

where F, A, and B represent the fused image, and two different-modal source images, respectively. T is the feature extraction function, which can be specifically defined as intensity, gradient, *etc*. In addition, α_i denotes the weight that controls the optimization tendency, and $\|\cdot\|$ is the matrix norm. Clearly, optimizing F is essentially about finding a balance in the distributions of the source images A and B, which inevitably leads to the loss of useful information. The solution to this limitation is straightforward: define a clear optimization objective instead of multiple ones. Specifically, the improved loss function is defined as follows:

$$\mathcal{L} = \|T(F) - P(T(A), T(B))\|, \tag{2}$$

where P is a custom integration function, responsible for making the clear objective in domain T. We refer to this improved loss function as the explicit information preservation loss.

Color Separation Strategy. Color separation strategy is a commonly employed technique in existing deep fusion methods to achieve color image fusion. Specifically, the visible image is first converted to YCbCr color space, and the luminance (Y) is fused with the infrared image. The resulting fused image is then concatenated with the chrominance channels (Cb and Cr), and transformed back to RGB color space to obtain the final color output. The whole process is formalized as:

$$F = M(C(N(Y_{vis}, I_{ir}), Cb_{vis}, Cr_{vis})), \tag{3}$$

where N is the fusion function, C is the concatenation function, and M is the transformation function from $YCbCr$ to RGB color space. However, the original Cb and Cr do not match the fused Y, which causes color distortion. A possible way to address this problem is to identify a measurement indicator that can describe the distance of intrinsic color properties. Then, even if the fusion affects some apparent color attributes, adaptive color preservation can be achieved by controlling the distance of intrinsic color attributes. Fortunately, the cosine similarity [41] is a good choice to quantify the distance of intrinsic color properties. It effectively eliminates the dimensional differences caused by brightness changes and other factors in most cases, and focuses on the critical color vector angle.

3. Method

We aim to improve visibility in low-light environments through infrared and visible modality fusion. To achieve this goal, we first propose a guided restoration module to enhance the visible modality that suffers from information loss due to poor illumination, reducing the information mismatch with the infrared modality. Then, we introduce an information fusion module to ensure the preservation of texture, contrast, and color fidelity during the information fusion process, thus producing high-quality visual enhancement results. The overall architecture of our DDBF is presented in Fig. 3.

3.1. Guided Restoration Module

As mentioned, the light-invariant high-contrast characteristic of infrared images provides good guidance for improving the illuminating of low-light visible images. Besides, low-light environments are complex and varied, requiring controllable and flexible illumination enhancement. Prompted by these considerations, we propose a GRM that utilizes a conditional generative adversarial network (CGAN) [30] to recover useful information hidden in the darkness.

Unlike conventional image generation models that use random noise as input, GRM treats low-light visible images I_{vis} as samples from the original distribution and defines enhancement ratios r as conditional inputs. Then, the generator G produces enhanced visible images according to $I_{vis}^{en} = G(I_{vis}|r)$. Now, the key lies in specifying the target illumination distributions that correspond to the enhancement ratios, thereby driving brightness adjustment through adversarial learning. Inspired by contrastive learning [1], we can perform data augmentation for original infrared images, constructing a *multi-level infrared sample set* that reflects the desired multiple illumination distributions:

$$I_{ir}^{\gamma} = K(I_{ir}, \{\gamma_1, \gamma_2, \cdots, \gamma_n\}), \tag{4}$$

where K is the data augmentation operation, which in our work refers to gamma transformation [9]. γ is an exponent parameter that stretches or compresses contrast, with an inverse relationship to the augmentation ratio r ($r = \frac{1}{\gamma}$). With all the necessary samples and variables prepared, we are now ready for conditional adversarial learning.

We want the adversarial network to primarily focus on *learning the illumination (or contrast) distribution rather than the differences between the visible and infrared modalities*. A simple yet effective operation is to remove color and introduce blur to reduce modality differences, aligning with the original assumption of early Retinex theory regarding illumination [12, 18, 45]. Therefore, an illumination adversarial loss can be defined for the generator G:

$$\mathcal{L}_{IA-G} = \|D(L(U(I_{vis}^{en}))|r) - a\|_1, \tag{5}$$

where D represents the discriminator function, U is the color removal function, L is the low-pass filtering (LPF) function, and a corresponds to a probability label. In this work, U is specified as YUV color space transformation, while L is defined as a Gaussian filtering. Intuitively, the generator is expected to deceive the discriminator by enhancing visible images with deceptive illumination, so a is set to 1. In contrast, the discriminator aims to distinguish such a deceptive illumination. Therefore, the loss function of the discriminator for illumination adversarial learning is defined as:

$$\mathcal{L}_D = \|D(L(U(I_{vis}^{en}))|r) - b\|_1 + \|D(L(I_{ir}^{\gamma})|r) - c\|_1, \tag{6}$$

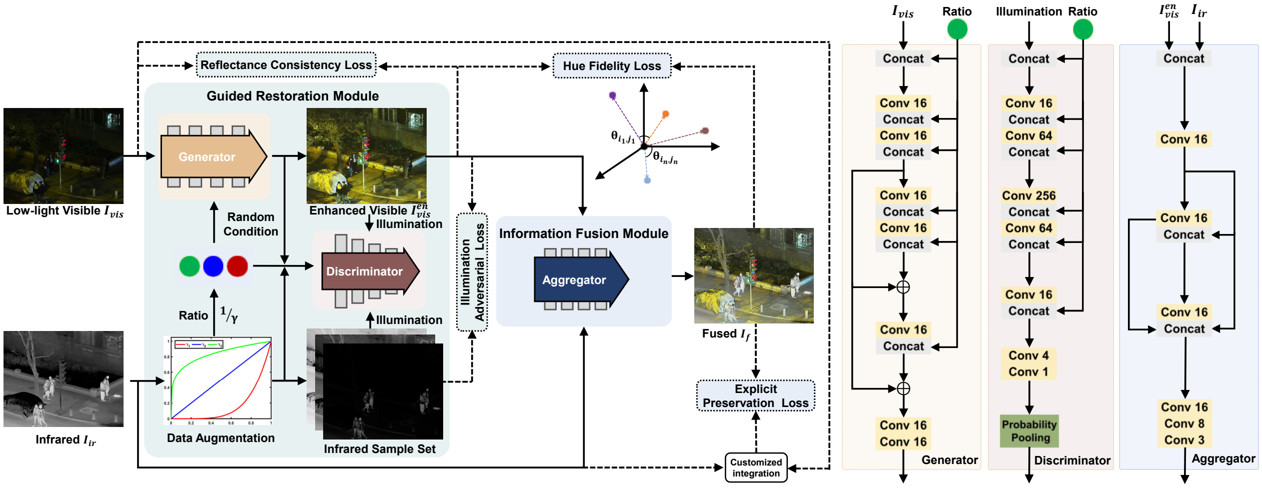

Figure 3. The overview of our DDBF. It consists of a guided restoration module, and an information fusion module. The detailed architectures of sub-networks are on the right, which are lightweight.

where the probability labels b and c should be set as 1 and 0, respectively, to guide towards correct classification. The evolved discriminator forces the generator to improve the quality of adjusted illumination. Besides, skip connections are employed to pass the enhancement ratio r to multiple feature layers of both the generator and discriminator, which provides architectural support for the controllable illumination adjustment. Notably, *the infrared modality is only utilized during the training phase, while GRM can directly enhance the visible modality during the testing phase*.

In addition to illumination adjustment, another important aspect to address in the enhancement process is scene fidelity. Specifically, GRM should ensure that the basic composition (*e.g.*, shape, color, or relative position of objects) of the imaging scene remains unchanged during the illumination adjustment. Fortunately, the reflectance component in Retinex theory [13] captures essential scene information, allowing us to control the reflectance consistency, which is crucial for maintaining scene fidelity. By applying the basic Retinex formula $I = R/S$ (R indicates reflectance, and S denotes illumination), we can use the aforementioned original illumination assumption to estimate reflectance and define a scene fidelity loss for the generator:

$$\mathcal{L}_{SF-G} = \left\| \frac{I_{vis}^{en}}{\max(L(U(I_{vis}^{en})),\delta)} - \frac{I_{vis}}{\max(L(U(I_{vis})),\delta)} \right\|_1, \tag{7}$$

where δ is a small constant ($\delta = 0.01$ in GRM) to avoid the denominator being 0. Moreover, we use residual connections in our generator to facilitate the transfer and preservation of scene information.

3.2. Information Fusion Module

GRM reduces the information mismatch between different modalities by recovering the information from the low-light visible image. This allows us to further develop an IFM that combines the advantages of the infrared and enhanced visible images to generate a visually appealing fused image: $I_f = A(I_{ir}, I_{vis}^{en})$, where A refers to the function of our proposed aggregator. As depicted in Fig. 3, the aggregator module is designed to be lightweight and utilizes skip connections for efficient information integration. Its core lies in the incorporation of specific constraints aimed at preserving texture and contrast, and ensuring color fidelity.

Firstly, we consider the preservation of texture and contrast, aiming to alleviate the limitation of traditional multimodal weighted game idea that often lead to information weakening. As discussed in Section 2, we propose to construct a clear objective for preserving significant contrast and sharpened texture to tackle this challenge. Formally, we introduce the explicit information preservation constraints: the significant contrast loss $\mathcal{L}_{SC-A}$ and the sharpened texture loss $\mathcal{L}_{ST-A}$. The significant contrast loss $\mathcal{L}_{SC-A}$ for the aggregator is defined as:

$$\mathcal{L}_{SC-A} = \|I_f - \max(I_{ir}, I_{vis}^{en})\|_1, \tag{8}$$

here, we use the maximum function to determine the most salient pixel intensity in each spatial location, forming the basis of the contrast optimization objective. Eq. (8) is a specialization of Eq. (2), where T is defined as the intensity domain (*i.e.*, identity map), and P is specified as the maximum function. Similarly, the sharpened texture loss $\mathcal{L}_{ST-A}$ for the aggregator is defined as:

$$\mathcal{L}_{ST-A} = \|\nabla I_f - \max(\nabla I_{ir}, \nabla I_{vis}^{en})\|_1. \tag{9}$$

The inclusion of above two loss terms effectively achieves explicit information preservation, addressing the problem of information weakening presented in existing methods.

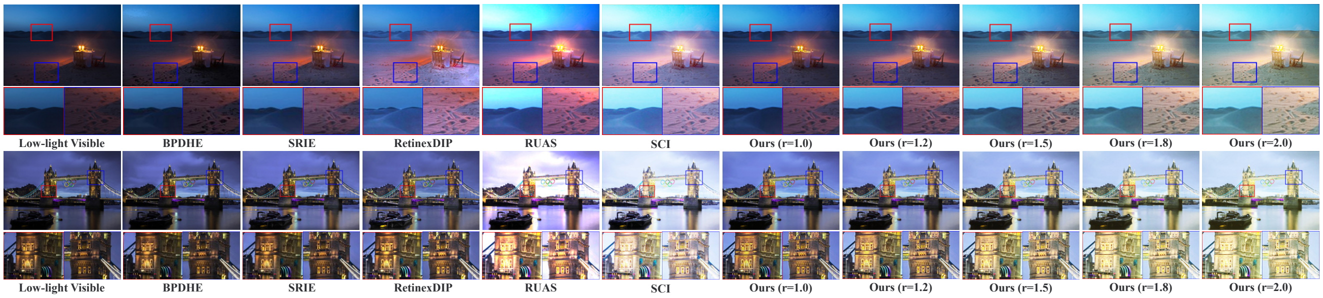

Figure 4. Visualization of low-light enhancement on the ExDark dataset.

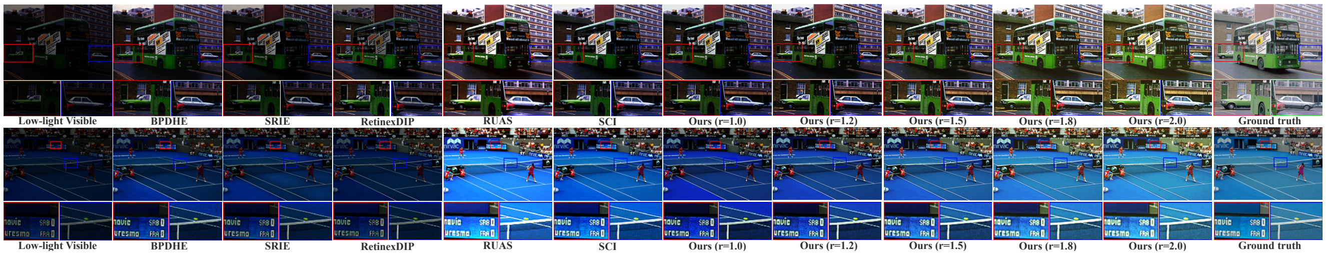

Figure 5. Visualization of low-light enhancement on the AGLIE dataset.

Secondly, we address the limitation of existing fusion methods that are unable to directly handle color images by incorporating a constraint for color fidelity. As mentioned in Section 2, we propose to utilize cosine similarity to measure the difference in intrinsic color properties. Therefore, we define a hue fidelity loss $\mathcal{L}_{HF-A}$ to achieve color fidelity, given by:

$$\mathcal{L}_{HF-A} = 1 - \sum_i \sum_j \frac{\sum_k I_{f_{i,j,k}} \times I^{en}_{vis_{i,j,k}}}{\sum_k \sqrt{I^2_{f_{i,j,k}}} \times \sqrt{I^{en^2}_{vis_{i,j,k}}}}, \qquad (10)$$

where i, j, k represent the pixels in the i-th row, j-th column, and k-th channel, respectively. On the one hand, the color vectors are ℓ_2-normalized along the channel, which helps eliminate dimensional differences caused by external factors such as illumination. On the other hand, by using color vector angles as the evaluation criterion, we can better preserve the intrinsic color property.

3.3. Interactive Executable Interface

We integrate all the functions of our DDBF into an interactive executable interface, which provides complete actionable function buttons and output visualization. In this way, users can easily achieve low-light enhancement and multi-modal image fusion, and obtain enhanced and fused visualizations that meet their visual preferences in a WYSIWYG (What You See Is What You Get) manner. Please refer to the *Suppl. Material* for more details.

4. Experiments

4.1. Datasets and Implementation

Datasets. We train our DDBF on the LLVIP dataset [11], manually selecting 400 high-quality image pairs from 10 street scenarios as the training data due to imperfect registration. During training, we adopt a cropping and expanding strategy to obtain a large number of patches, and randomly apply one of the 7 data augmentation strategies to them, *e.g.*, reverse, rotate, flip, and their combinations. In the testing phase, evaluation is done on ExDark [23], AGLIE [24], LLVIP [11], MFNet [7], and RoadScene [44] datasets.

Implementation Details. GRM and IFM are iteratively trained using the Adam optimizer with a batch size of 10, and the training lasts for $1,500$ epochs. To improve the training stability of CGAN, we employ a soft label strategy where labels a and c are relaxed to random numbers within the range of [0.8,1.0], while label b is assigned to a random number between 0 and 0.2. All experimental work is carried out using an NVIDIA RTX 2080Ti GPU with 11GB memory and an Intel CPU i7-8750H.

4.2. Comparative Experiments

Our DDBF offers two output modes: low-light enhancement and low-light multi-modal fusion. We compare it with specialized approaches for these two tasks. See the *Suppl. Material* for more visual results with high-quality images.

Low-light Enhancement. In the testing phase, *our method can enhance low-light images directly without the need for inputting infrared images, so our method can be easily deployed to low-light scenarios with only visible modality.*

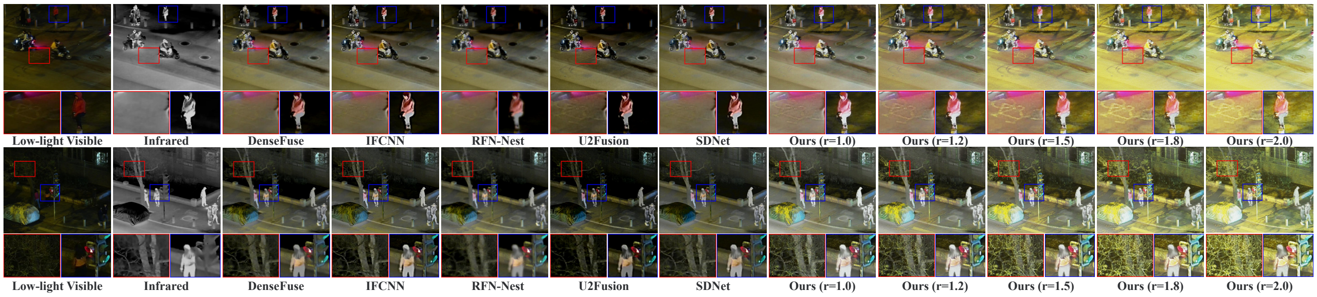

Figure 6. Visualization of multi-modal fusion methods on the LLVIP dataset.

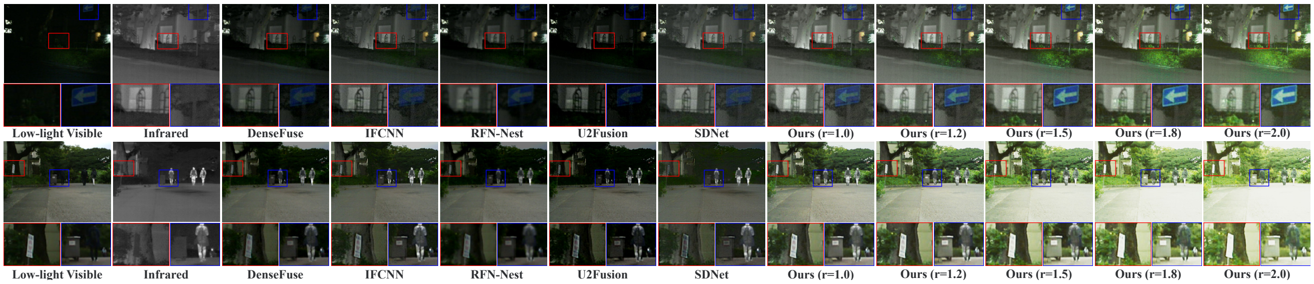

Figure 7. Visualization of multi-modal fusion methods on the MFNet dataset.

Table 1. Statistical results of low-light enhancement.

Dataset	Metric	BPDHE	SRIE	RetinexDIP	RUAS	SCI	Ours ($r=1.0$)	Ours ($r=1.2$)	Ours ($r=1.5$)	Ours ($r=1.8$)	Ours ($r=2.0$)
ExDark	NIQE ↓	3.800	3.516	3.382	3.843	3.621	3.657	3.473	**3.372**	3.348	<u>3.382</u>
	PIQE ↓	41.611	38.105	34.193	36.343	36.567	<u>33.832</u>	**33.533**	33.861	34.400	35.065
AGLIE	SSIM ↑	0.536	0.456	0.512	0.572	0.623	0.527	0.598	0.675	**0.706**	<u>0.704</u>
	PSNR ↑	11.407	10.234	10.535	13.961	14.845	12.754	14.230	15.770	**16.434**	<u>16.427</u>

We compare it with five state-of-the-art techniques, including BPDHE [10], SRIE [5], RetinexDIP [51], RUAS [22], and SCI [29]. The test set consists of 100 images from the ExDark dataset [11] and 40 images from the AGLIE dataset [24]. For the ExDark dataset lacking ground truth, we utilize non-reference metrics NIQE [31] and PIQE [40], while for the AGLIE dataset with ground truth, we use well-known SSIM and PSNR. Fig. 4 shows the visual results of different methods on the ExDark dataset. It can be seen that our DDBF produces fine desert textures and tower structures with a very natural look. Consistently, our DDBF achieves visually pleasing results that are more consistent with the ground truth on the AGLIE dataset, as shown in Fig. 5. Importantly, our method can generate enhanced images with progressive exposures, flexibly allowing users to customize their preferred results. Furthermore, the quantitative evaluation in Table 1 shows that our DDBF achieves the best score on non-reference NIQE and PIQE, and attains the highest SSIM and PSNR scores, demonstrating its effective naturalness maintenance, texture preservation, and noise reduction.

Low-light Multi-modal Fusion. Our method offers a novel alternative in cases where enhancing only the visible modality cannot provide satisfactory results, named multi-modal fusion. We compare it with five state-of-the-art methods, including DenseFuse [14], IFCNN [49], RFN-Nest [16], U2Fusion [44], and SDNet [46]. The test data consists of 100 image pairs from the LLVIP dataset and 100 image pairs from the MFNet [7] dataset. As shown in Figs. 6 and 7, our DDBF effectively restores the objects hidden in the darkness, and naturally presents significant contrast and rich textures. On the contrary, other competitors provide relatively poor visibility. Besides, our DDBF naturally preserves the scene colors, while other methods suffer color distortion due to the use of the color separation strategy. Furthermore, we provide quantitative results in Table 2, where four popular image fusion metrics are selected, including MI [33], VIF [8], AG [2], and SD [35]. Our DDBF ranks first in all objective metrics for both the LLVIP and MFNet datasets. These findings highlight two key advantages of our method. First, the designed GRM reduces information mismatch between infrared and visible modalities, enabling better information integration. Second, the proposed explicit information preservation constraints and hue fidelity constraint address information weakening and color distortion issues during fusion. Besides, the controllable fusion paradigm also provides users with the opportunity to choose preferences for their observations.

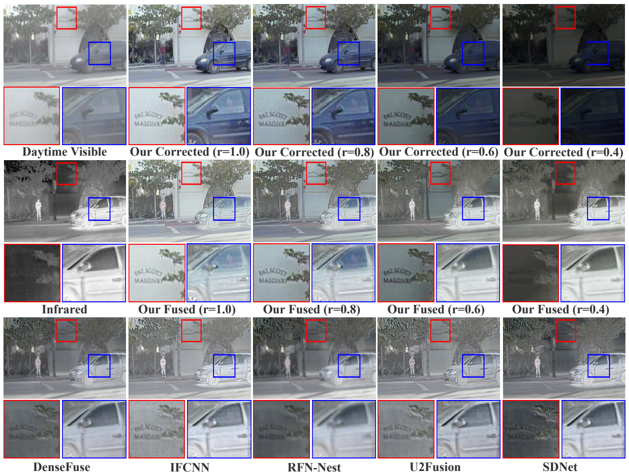

Figure 8. Generalization to the daylight RoadScene dataset.

Table 2. Statistical results of multi-modal fusion methods.

Dataset	LLVIP				MFNet			
Metric	MI ↑	VIF ↑	AG ↑	SD ↑	MI ↑	VIF ↑	AG ↑	SD ↑
DenseFuse	2.683	0.400	3.899	0.152	2.945	0.423	2.426	0.126
IFCNN	2.797	0.421	5.901	0.162	2.877	0.423	3.517	0.131
RFN-Nest	2.318	0.373	2.964	0.152	2.715	0.407	1.950	0.130
U2Fusion	2.239	0.352	4.682	0.148	2.515	0.405	3.267	0.132
SDNet	2.880	0.355	4.847	0.144	2.787	0.388	3.211	0.108
Ours ($r=1.0$)	**2.976**	0.473	8.020	0.189	3.365	0.485	4.133	0.162
Ours ($r=1.2$)	2.914	0.486	8.634	0.197	3.378	**0.490**	4.398	0.176
Ours ($r=1.5$)	2.904	0.499	9.150	**0.204**	3.392	0.485	4.591	0.189
Ours ($r=1.8$)	2.910	**0.504**	**9.252**	0.200	3.405	0.480	**4.649**	**0.193**
Ours ($r=2.0$)	2.927	0.503	8.951	0.191	**3.435**	0.476	4.568	0.190

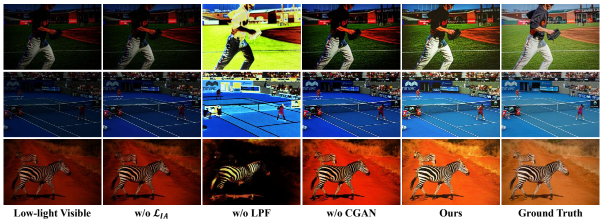

Figure 9. Visual analysis of GRM.

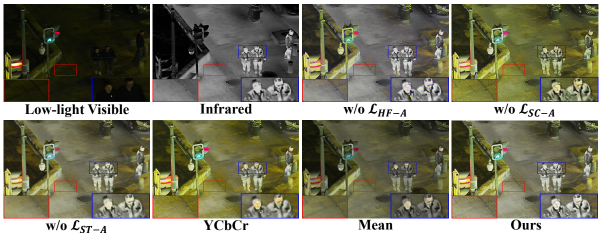

Figure 10. Visual analysis of IFM.

Table 3. Results of generalization to daylight.

RoadScene	MI ↑	VIF ↑	AG ↑	SD ↑
DenseFuse	2.435	0.293	4.952	0.154
IFCNN	2.338	0.313	6.647	0.143
RFN-Nest	2.308	0.307	3.860	**0.175**
U2Fusion	2.080	0.287	6.533	0.139
SDNet	2.467	0.225	**6.708**	0.174
Ours ($r=0.4$)	3.183	0.199	4.000	0.150
Ours ($r=0.6$)	2.805	0.254	4.644	0.121
Ours ($r=0.8$)	2.968	0.364	5.403	0.118
Ours ($r=1.0$)	**3.288**	**0.434**	5.451	0.119

Generalization to Daylight. Because our DDBF models the dependence of illumination changes on the enhancement ratio, our method can be applied not only to low-light environments but also to daytime scenarios. The generalization performance is evaluated using 104 daytime image pairs from the RoadScene dataset [44]. The visual results in Fig. 8 demonstrate that our DDBF can correct the low contrast of visible modality caused by overexposure, which implies that *our method can achieve both low-light enhancement and overexposure correction, providing attractive dual restoration capabilities*. Moreover, our method also excels in daylight multi-modal fusion, preserving text clarity on walls and maintaining the saliency of the car. Results in Table 3 further confirm the advantages of our DDBF, with the highest scores on MI and VIF, indicating strong information correlation and visual fidelity preservation.

Efficiency. Efficiency is also an important factor for evaluating the performance of methods. Therefore, we conduct an efficiency analysis for our DDBF. First, we count the number of parameters of DDBF, in which GRM and IFM consist of a total of 0.406 M parameters. Then, we measure the average running time of DenseFuse, IFCNN, RFN-Nest, U2Fusion, SDNet, and our DDBF on test images of size about 820 × 1024, which are 0.327, 0.144, 0.372, 0.223, 0.148, and 0.124 seconds, respectively. Our DDBF achieves the fastest running speed. Besides, Parameter counts for them are 0.074, 0.084, 7.524, 0.659, 0.070, and 0.406 M, and their FLOPs are 148.006, 109.189, 2199.958, 1107.901, 112.963, and 46.778 G, respectively.

4.3. Ablation Studies

Guided Restoration Module. We conduct ablation experiments on AGLIE to analyze the effectiveness of specific designs in GRM: illumination adversarial loss $\mathcal{L}_{IA}$, low-pass filtering function (LPF), and conditional generative adversarial network (CGAN). According to Table 1, we set $r = 1.8$ as the baseline for the ablation. The visual results in Fig. 9 show the impact of removing these components. More concretely, removing $\mathcal{L}_{IA}$ results in the inability to adjust the illumination. The absence of LPF leads to the persistence of modality differences between infrared and visible images, causing the network to overly focus on contrast rather than global illumination. Besides, after replacing CGAN with a conventional GAN, the module losses the capability of controlling the enhancement ratio. As a result, GRM cannot effectively restore information in some scenarios. The quantitative results in Table 4 further demonstrate the negative impacts caused by removing these designs in

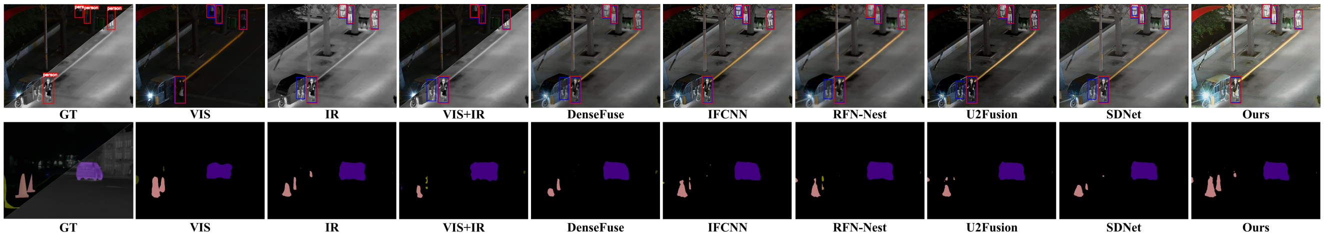

Figure 11. Application to high-level vision tasks. Red boxes in the top row represent ground truth, while blue represent detection results.

Table 4. Quantitative analysis of GRM.

AGLIE	w/o $\mathcal{L}_{IA}$	w/o LPF	w/o CGAN	Ours
SSIM ↑	0.405	0.123	0.526	**0.706**
PSNR ↑	5.205	9.841	13.133	**16.434**

Table 5. Quantitative analysis of IFM.

LLVIP	w/o $\mathcal{L}_{HF-A}$	w/o $\mathcal{L}_{SC-A}$	w/o $\mathcal{L}_{ST-A}$	YCbCr	Mean	Ours
MI ↑	2.891	2.769	**3.123**	2.773	2.211	2.904
VIF ↑	0.483	0.496	0.489	0.474	0.408	**0.499**
AG ↑	7.473	9.086	6.165	8.899	5.414	**9.150**
SD ↑	0.186	0.197	0.187	0.194	0.150	**0.204**

Table 6. Results of application to high-level vision tasks.

	Detection				Segmentation	
	Precision	Recall	mAP@0.5	mAP@0.95	mIOU	mACC
VIS	0.976	0.946	0.764	0.667	40.292	43.774
IR	0.966	**0.992**	0.913	0.753	40.274	43.828
VIS+IR	0.977	0.977	0.927	**0.762**	40.956	44.792
DenseFuse	0.973	0.981	0.889	0.752	38.978	42.441
IFCNN	0.977	0.977	0.896	0.754	39.750	43.139
RFN-Nest	0.977	0.981	0.893	0.736	39.990	43.551
U2Fusion	0.973	0.973	0.903	0.757	40.249	43.608
SDNet	0.959	**0.992**	0.913	0.752	39.353	42.707
Ours	**0.984**	0.969	**0.934**	0.760	**41.628**	**45.704**

GRM. Overall, these designs collectively ensure the high performance, flexibility, and reliability of our GRM.

Information Fusion Module. We evaluate the effectiveness of specific designs in our IFM on the LLVIP dataset. According to Figs. 6 and 7, we set $r = 1.5$ as the baseline for the ablation on these designs, because it leads to relatively better visual results. These specific designs include hue fidelity and explicit information preservation constraints. They correspond to hue fidelity loss $\mathcal{L}_{HF-A}$, significant contrast loss $\mathcal{L}_{SC-A}$, and sharpened texture loss $\mathcal{L}_{ST-A}$. In ablation experiments, we directly remove $\mathcal{L}_{HF-A}$, and replace $\mathcal{L}_{SC-A}$ and $\mathcal{L}_{ST-A}$ with the commonly used loss functions based on the multi-modal weighted game. Besides, we try to use the color separation strategy (YCbCr) to preserve the color, and tailor the integration function P to the mean operation. The visual results in Fig. 10 show that the removal of $\mathcal{L}_{HF-A}$ leads to unnatural colors with fragmented tone distribution. Excluding $\mathcal{L}_{SC-A}$ reduces the saliency of thermal objects, while the absence of $\mathcal{L}_{ST-A}$ results in local structural smoothing. Additionally, the YCbCr strategy leads to an overly yellowish tint, especially in highlighted thermal object regions. Using the mean function leads to the fused image suffering from brightness neutralization The objective metrics reported in Table 5 further support the significance of these designs to our IFM.

4.4. Application to High-level Vision Tasks

Furthermore, we apply DDBF to high-level vision tasks, *i.e.*, object detection and semantic segmentation. Notably, we use the $r = 1.5$ version of our DDBF due to its excellent visualization performance. We retrain YOLOv5 [36] and SegNeXt [6] on source infrared and visible images and the fused images of different methods. The visual results are presented in Fig. 11. The detection and segmentation results based on our fused images are more accurate, while others suffer from false detections and incomplete segmentations. The quantitative results in Table 6 demonstrate that our method achieves the best scores on most metrics. Notably, the higher decision accuracy based on our fused image results compared to VIS+R may be attributed to our method's ability to restore scene information. Overall, these results prove that our method can effectively aggregate scene information and provide high-quality semantic guidance.

5. Conclusion

This paper proposes a controllable visual enhancer using cross-modal conditional adversarial learning. First, we design a guided restoration module to recover the scene information lost in low-light visible modality. It constructs a multi-level sample set for conditional learning, enabling users to customize the enhancement ratio according to actual circumstances. Then, a novel information fusion module with explicit information preservation and hue fidelity constraints is developed to deliver enhanced visualization characterized by significant contrast, rich textures, and faithful colors. Extensive results reveal DDBF's advantages with a user-friendly interface for practical application.

Acknowledgments

This work was supported by NSFC (62276192).

References

[1] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 33:8765–8775, 2020. 3

[2] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015. 6

[3] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021. 1

[4] Wang Di, Liu Jinyuan, Fan Xin, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3508–3515, 2022. 1

[5] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2782–2790, 2016. 6

[6] Menghao Guo, Chengze Lu, Qibin Hou, Zhengning Liu, Mingming Cheng, and Shimin Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35, 2022. 8

[7] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multispectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017. 5, 6

[8] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. 6

[9] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE Transactions on Image Processing*, 22(3):1032–1041, 2013. 3

[10] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 6

[11] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3496–3504, 2021. 5, 6

[12] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, 1997. 3

[13] Edwin H Land. The retinex. *American Scientist*, 52(2):247–264, 1964. 4

[14] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 1, 6

[15] Hui Li, Xiao-Jun Wu, and Tariq Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12):9645–9656, 2020. 1

[16] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. 6

[17] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23:1383–1396, 2021. 1

[18] Zhetong Liang, Weijian Liu, and Ruohe Yao. Contrast enhancement by nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, 25(2):673–686, 2016. 3

[19] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan 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*, pages 5802–5811, 2022. 1

[20] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, pages 1–28, 2023. 1

[21] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8115–8124, 2023. 1

[22] Risheng Liu, Long Ma, Tengyu Ma, Xin Fan, and Zhongxuan Luo. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5953–5969, 2023. 6

[23] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 5

[24] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7):2175–2193, 2021. 5, 6

[25] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1

[26] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 1

[27] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with

multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:5005014, 2021. 1

[28] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 1

[29] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 6

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 6

[32] Amanda C Muller and S Narayanan. Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, 10(2):137–149, 2009. 1

[33] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313–315, 2002. 6

[34] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4714–4722, 2017. 2

[35] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement Science and Technology*, 8(4):355, 1997. 6

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 8

[37] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[38] Jin Tang, Dongzhe Fan, Xiaoxiao Wang, Zhengzheng Tu, and Chenglong Li. Rgbt salient object detection: Benchmark and a novel cooperative ranking approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4421–4433, 2020. 1

[39] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022. 2

[40] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *Proceedings of the National Conference on Communications*, pages 1–6, 2015. 6

[41] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information Sciences*, 307:39–52, 2015. 3

[42] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12484–12491, 2020. 1

[43] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021. 1

[44] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Mmachine Intelligence*, 44(1):502–518, 2022. 1, 5, 6, 7

[45] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 3

[46] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021. 1, 6

[47] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12797–12804, 2020. 1

[48] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[49] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020. 6

[50] Fan Zhao, Wenda Zhao, Libo Yao, and Yu Liu. Self-supervised feature adaption for infrared and visible image fusion. *Information Fusion*, 76:189–203, 2021. 1

[51] Zunjin Zhao, Bangshu Xiong, Lei Wang, Qiaofeng Ou, Lei Yu, and Fa Kuang. Retinexdip: A unified deep framework for low-light image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1076–1088, 2022. 6

[52] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc 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*, pages 5906–5916, 2023. 1

[53] Zixiang Zhao, Jiangshe Zhang, Haowen Bai, Yicheng Wang, Yukun Cui, Lilun Deng, Kai Sun, Chunxia Zhang, Junmin Liu, and Shuang Xu. Deep convolutional sparse coding networks for interpretable image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2368–2376, 2023. 1

[54] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. In *Proceedings of the ACM International Conference on Multimedia*, pages 465–472, 2019. 1