Efficient Multi-exposure Image Fusion via Filter-dominated Fusion and Gradient-driven Unsupervised Learning

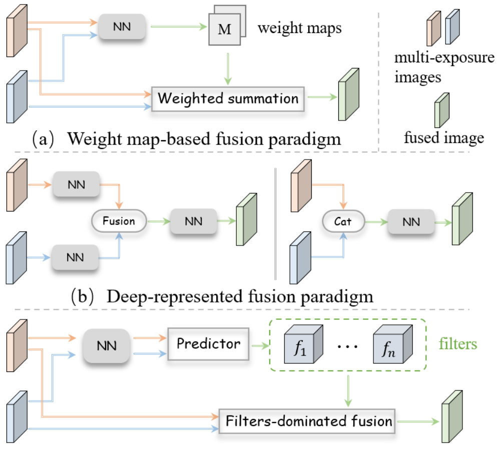
Kaiwen Zheng Jie Huang Hu Yu Feng Zhao*
University of Science and Technology of China
{kezh,hj0117,yuhu520}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract

Multi exposure image fusion (MEF) aims to produce images with a high dynamic range of visual perception by integrating complementary information from different exposure levels, bypassing common sensors' physical limits. Despite the marvelous progress made by deep learning-based methods, few considerations have been given to the innovation of fusion paradigms, leading to insufficient model capacity utilization. This paper proposes a novel filter prediction-dominated fusion paradigm toward a simple yet effective MEF. Precisely, we predict a series of spatial-adaptive filters conditioned on the hierarchically represented features to perform an image-level dynamic fusion. The proposed paradigm has the following merits over the previous: 1) it circumvents the risk of information loss arising from the implicit encoding and decoding processes within the neural network, and 2) it better integrates local information to obtain better continuous spatial representations than the weight map-based paradigm. Furthermore, we propose a Gradient-driven Image Fidelity (GIF) loss for unsupervised MEF. Empowered by the exploitation of informative property in the gradient domain, GIF is able to implement a stable distortion-free optimization process. Experimental results demonstrate that our method achieves the best visual performance compared to the state-of-the-art while achieving an almost 30% improvement in inference time. The code is available at https://github.com/keviner1/FFMEF.

1. Introduction

Due to the physical limitations of common imaging sensors, the dynamic range captured by them is much lower than that of natural scenarios. The low dynamic range (LDR) imaging results frequently produce a poor visual effect. However, a sequence of LDR images with multi-exposure levels generally contains complementary information, especially in pairs of images with extreme overexposure and underexposure property. For this phenomenon, extensive research and application of multi-exposure fusion (MEF) algorithms were initiated in academia decades ago.



(a) Weight map-based fusion paradigm

(b) Deep-represented fusion paradigm

(c) The proposed filter prediction-dominated fusion paradigm

Figure 1. The comparison between the two existing mainstream paradigms and the proposed filter prediction-dominated fusion paradigm. Aiming for a more efficient MEF, we stand on the shoulders of previous paradigms for reaping the benefits of combined deep neural networks and image-level fusion schema.

The long history of the MEF witnesses various sophisticated methods, which can be grouped into traditional algorithms [11, 15, 23, 27, 31] and deep learning-based models [26, 34, 35, 41, 42]. In recent years, there has been a growing emphasis on deep learning-based research, which can be further divided into the weight map-based fusion paradigm (*e.g.* MEFNet [26]) and the deep represented fusion paradigm (*e.g.* DeepFuse [35]), as shown in Fig. 1. The former employs the neural network to generate the weight map for pixel-wise fusion. The latter performs fusion operations in the neural network-encoded domain.

However, both of the aforementioned paradigms have considerable potential for further improvement. First, the deep-represented fusion paradigm confronts a trade-off problem between deeper network design and a higher risk of information distortion and loss. Second, the weight map-

[*] Corresponding author.

based fusion paradigm does not account for the exploitation of local information, and the pixel-wise solution lacks spatial continuity. As illustrated by the survey [48], it is challenging for existing algorithms to possess high-quality visual effects and efficient execution times simultaneously. Under-expected technical designs commonly cause insufficient model capacity utilization, hindering efficient image fusion, and resulting in unacceptable artifacts.

Therefore, in this work, we propose a novel filter prediction-dominated fusion paradigm for efficient multi-exposure image fusion. It is designed to absorb the previous paradigm's nutrients and filter out impurities. Specifically, our paradigm benefits from the powerful representational capabilities of deep neural networks while preventing the risk of information loss in the encoding space. The meticulously designed filter sequence-dominated fusion schema with local information aggregation leads to more efficient image-level fusion. More concretely, our elaborately designed paradigm consists of three phases: Initially, a feature extraction module is employed to obtain the hierarchical representations of each source image. Next, we develop a spatial-adaptive filter predictor that generates a sequence of filters corresponding to the multi-level representation obtained. Lastly, image-level signal processing is performed on the source image through the predicted filters, which adaptively aggregate local information within images and fuse information between images. For optimization terms, a Gradient-driven Image Fidelity (GIF) loss is devised for unsupervised learning. Based on the informative property of the gradient domain, GIF disentangles the reliance between the fused image and the source image, driving effective complementary learning and artifact removal. Compared with the commonly used loss function MEF-SSIM [29] and PMGI [47], GIF emphasizes information fidelity-oriented design, leading to a more stable model convergence. Extension experiments on widely recognized benchmarks [48–50] demonstrate the applicability of our efficient-oriented paradigm and image fidelity-focused GIF to various image fusion challenges.

The main contributions of this work are summarized as:

- We suggest a simple yet effective paradigm, termed the filter prediction-based fusion paradigm, that targets more efficient and high-quality multi-exposure image fusion. Besides, extension experiments involving multi-focus fusion and infrared-visible fusion demonstrate its superiority and application potential.
- We design an image fidelity loss GIF based on in-depth utilization of gradient information to ensure the efficacy of unsupervised complementary learning while reducing distortion artifacts.
- The proposed method presents superior visual performance and delivers a nearly 30% improvement in running time over the previous state-of-the-art on the widely recognized benchmark MEFB.

2. Related Work

2.1. Traditional MEF Methods

Traditional approaches include both spatial domain-based and transform domain-based algorithms. The former algorithms first analyze the information importance of the source images and then utilize the estimated weight maps for spatial-wise fusion. This methodology can be further divided into two types based on the distinct basic units in information evaluation: pixel-based methods [15, 19, 23] and patch-based methods [11,27,28]. In terms of transform domain-based approaches [3,5,31,37,38,43], they generally conduct information fusion over decomposed coefficients of images to exploit the beneficial signals in various domains. Although these traditional methods have achieved impressive results, their manually designed feature extraction and fusion strategies limit the performance.

2.2. Deep Learning-based MEF Methods

Recently, the deep learning-based MEF schemes have received extensive attention and shown promising effects [10, 13, 48, 51]. Based on the convolutional neural network (CNN), Deepfuse [35] first proposes to merge the luminance components in the deep-represented feature domain and fuse the chrominance parts via a traditional weighted average method. In addition to CNN-based designs, both Generative Adversarial Network-based [42, 45] and Transformer-based techniques [24, 34] considerably increase MEF performance in studies. However, these paradigms necessitate more model capacity to prevent the loss of source information during encoding and decoding. Another deep learning-based scheme is to replace the manually constructed information measurement algorithm with neural networks. For example, the MEF-Net [26] utilized CNN and guided filtering to generate pixel-wise weighted maps for image-level fusion. Unfortunately, the fusion results of MEFNet usually suffer from unacceptable artifacts.

In the absence of ground truth images, unsupervised learning-based algorithms [33, 40, 41, 45, 47] have been widely studied and highly concerned, which model the relationship between fused images and source images for driving model optimization. However, the widely adopted loss functions MEF-SSIM and PMGI perform poorly in terms of information fidelity, as previous studies have shown [13, 48]. Besides, Qu *et al.* [34] introduces a self-supervised multi-task learning mechanism for training an encoder-decoder network TransMEF in large natural images dataset *MS-COCO* [22].

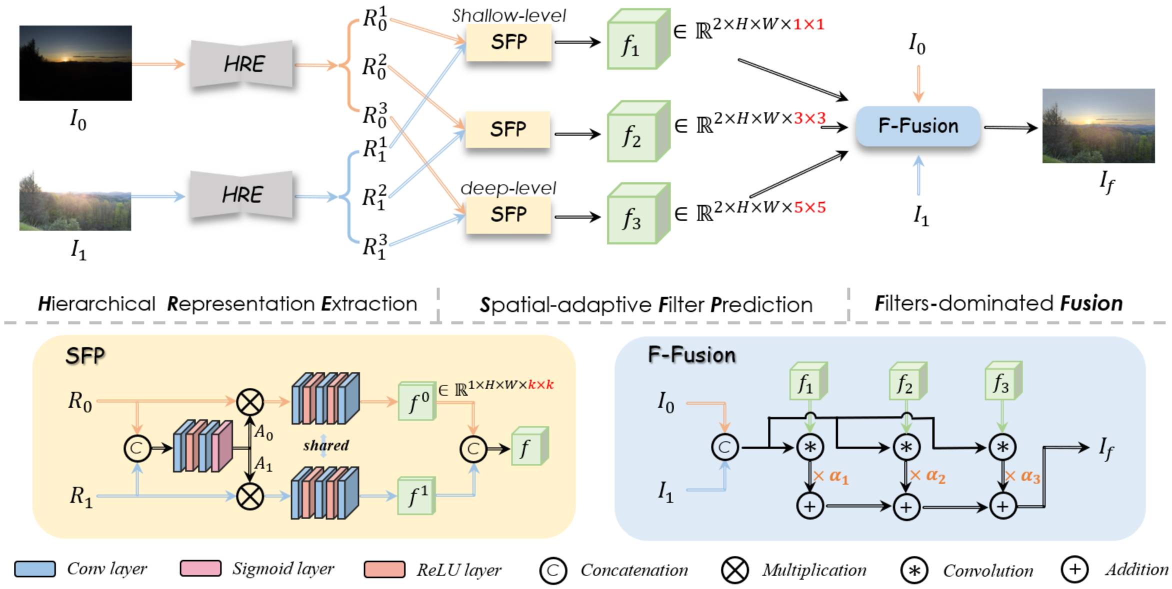


Figure 2. The overall framework of our proposed filter prediction-based fusion paradigm. It is made up of three steps: Hierarchical Representations Extraction, Spatial-adaptive Filter Prediction, and Filters-dominated Fusion. These steps perfectly combine the powerful representation capabilities of deep neural networks with the high efficiency of image-level fusion operations. It should be noted that our method fuses the luminance channel of the image and displays it in RGB form for better visual effects.

3. Method

3.1. Motivation and Overview

Motivation: Most MEF application scenarios involve mobile devices with stringent computational costs and visual effect requirements. However, existing approaches are constrained by suboptimal fusion paradigms, making it hard to balance complexity and performance. In addition, the lack of ground truth images inhibits the acquisition of high-quality images, and previous research [13, 48] has shown that the widely used loss function MEF-SSIM is insufficient in terms of image fidelity. Therefore, this paper offers a more reasonable MEF paradigm to enhance model capacity utilization, thereby reducing overall complexity. In addition, a novel information fidelity-oriented loss function, GIF, is proposed for efficient unsupervised MEF learning, and its outstanding accelerated model convergence effect is demonstrated experimentally.

Problem formulation: Given a pair of source images with different exposure levels, we first convert them to the $YCbCr$ color space. Then the Y channels (I_0, I_1) of source images are fused through the proposed model to obtain a high-quality luminance channel I_f. The color channels Cb and Cr are fused by the traditional weighted average operation as follows:

$$C_f = \frac{C_0|C_0 - 128| + C_1|C_1 - 128|}{|C_0 - 128| + |C_1 - 128|}, \tag{1}$$

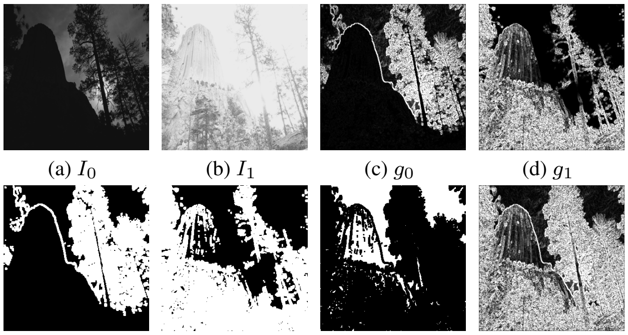
where (C_0, C_1) and C_f represent the C_b or C_r channel of input image pair and the fused image, respectively.

Framework overview: As shown in Fig. 2, our proposed paradigm consists of three well-designed core components. Specifically, a Hierarchical Representation Extractor (HRE) is first introduced to employ the powerful nonlinear mapping capability of neural networks to obtain multi-level critical information representations of the source image. This deep-represented information is beneficial for analyzing and evaluating source images to guide the spatial-adaptive fusion procedure. Inspired by Kernel Prediction Network [30], we adopt a Spatial-adaptive Filter Predictor (SFP) series to predict filter sequences composed of different kernel sizes. Based on the fact that deeper features have larger receptive fields, the multi-level SFPs correspond to hierarchical representations. Finally, the predicted filter sequence dominates the developed F-Fusion module's flexible and efficient image-level multi-exposure fusion. The details of SFP and F-Fusion will be described in Sec. 3.2 and Sec. 3.3. HRE is a Unet-like module composed of residual blocks (see Supplementary Materials for the details).

3.2. Spatial-adaptive Filter Predictor (SFP)

The highlights of the proposed novel paradigm are well articulated in SFP, which leverages represented features to predict filters for image-level fusion. Since multi-exposure fusion is a complementary learning process, cross-image interaction is required to assist information analysis. Inside SFP, a spatial attention mechanism is first devised to obtain the attention maps (A_0, A_1) corresponding to the input feature pair (R_0, R_1) as:

$$A_0, A_1 = SpatialAttention([R_0, R_1]), \tag{2}$$


(a) I_0 (b) I_1 (c) g_0 (d) g_1

(e) m_0 (f) m_1 (g) m_{sth} (h) g_{max}

Figure 3. Visual display of gradient maps, where g_0, g_1 are the raw gradient maps of images I_0, I_1. After processing, m_0, m_1, m_{sth} and g_{max} are obtained to complete GIF loss (See Fig. 4).

where $[\cdot]$ means concatenation operation. Following information scaling by the attention maps, we offer a convolutional neural network-based prediction head to produce image-private filters (f^0, f^1) as:

$$\begin{aligned} f^0 &= PredictionHead(R_0 \times A_0), \\ f^1 &= PredictionHead(R_1 \times A_1). \end{aligned} \tag{3}$$

Then the filters f^0 and f^1 are concatenated to get the current-level final filter $f \in \mathbb{R}^{2 \times H \times W \times k^2}$. It denotes that SFP predicts a $k \times k$-sized filter for every position, thus enabling spatial-adaptive processing over source images.

3.3. Filters-dominated Fusion (F-Fusion)

We implement our F-Fusion as a special fusion rule that embraces learnable parameters and local information aggregation, thus leading to more powerful image fusion. In concrete terms, inside F-Fusion, the predicted spatial-adaptive filters $\{f_i\}_{i=1}^3$ are employed to dominate the image-level fusion as follows:

$$I_f = \sum_{i=1}^{3} \alpha_i (f_i * [I_0, I_1]), \tag{4}$$

where $*$ represents the convolution operation and the learnable factors $\{\alpha_i\}_{i=1}^3$ are introduced for achieve flexible information fusion. Commonly used static fusion operations (*e.g.*, element-wise addition, averaging, and standard convolution) actually deviate from the dynamic properties required for MEF tasks, which are well addressed in our approach. In contrast to the weight map-based fusion paradigm, F-Fusion employs a multi-level schema with local information aggregation.

3.4. Gradient-driven Image Fidelity (GIF) loss

In the absence of ground truth images, ensuring that the fused information-rich image suffers less distortion of source information is a critical problem. A simple and straightforward design that facilitates information fidelity is to establish the relationships $\{w_i\}_{i=0}^1$ between the fused image I_f and the source images $\{I_i\}_{i=0}^1$, constructing the optimization function as follows:

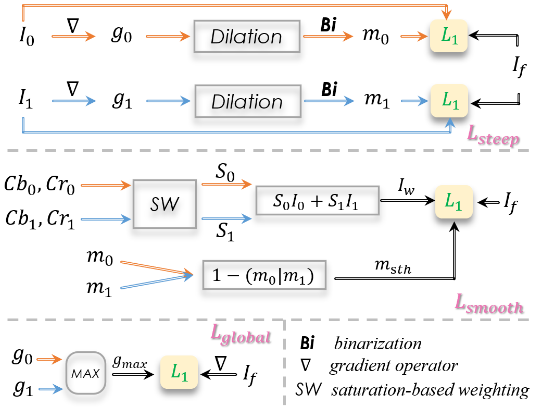


Figure 4. The workflow of three core components of the devised unsupervised loss function GIF.

$$arg\min \sum_{i=0}^{1} w_i \left\| \Phi(I_f) - \Phi(I_i) \right\|_1, \tag{5}$$

Where $\|\cdot\|_1$ represents the L_1 distance used as an example, and the function $\Phi(\cdot)$ usually stands for identity transformation, and VGG16 (or ResNet101) encoding in previous work [40, 41, 47, 51]. Some works [47, 51] that assign constant values to $\{w_i\}_{i=0}^1$ depart from the spatially dynamic complementary properties of MEF, while others [40, 41] pre-train an Image Quality Assessment (IQA) model for estimating $\{w_i\}_{i=0}^1$ and achieve a more reasonable fashion.

Without the introduction of pre-trained models, we propose a simple yet effective loss function, GIF (See Fig. 4), based on gradient information analysis. Although the raw gradient map can successfully depict the richness of information (See Fig. 3), there are several drawbacks to directly employing it as the $\{w_i\}_{i=0}^1$: 1) Sparse representations that lack local continuity are not conducive to constraining high-quality fusion. 2) The constraint for gradient-smooth regions needs separate consideration. 3) Different regions are constrained by distinct optimization targets, thus suffering the stitching artifacts. To address these issues, the $\mathcal{L}_{steep}$, $\mathcal{L}_{smooth}$ and $\mathcal{L}_{global}$ components are elaborately constructed in GIF as:

$$\mathcal{L}_{GIF} = \lambda_1 \mathcal{L}_{steep} + \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{global}, \tag{6}$$

where λ_1 and λ_2 are weighting factors set to 1.25 and 2.

In particular, we first design $\mathcal{L}_{steep}$ to drive the information fidelity of gradient-steep regions of the source image. To deal with the sparse representation of gradient maps, we dilate them using 3x3 max-pooling. Next, employ the binarization operation with a threshold of 0.5 to generate the

0-1 masks $\{m_i\}_{i=0}^{1}$ as follows:

$$\{m_i\}_{i=0}^{1} = Binarization(Dilation(\{\bigtriangledown I_i\}_{i=0}^{1})), \quad (7)$$

where $\bigtriangledown$ stands the $scharr$ gradient operator.With the selected masks, $\mathcal{L}_{steep}$ is calculated as:

$$\mathcal{L}_{steep} = m_0 \left\| I_f - I_0 \right\|_1 + m_1 \left\| I_f - I_1 \right\|_1 . \quad (8)$$

Since it is hard to measure the value of gradient-smooth regions in the luminance channel, we further introduce the color information from the Cb and Cr channels. Higher weights are assigned to pixels with better color saturation, and the weight maps $\{S_i\}_{i=0}^{1}$ for each source image are constructed as follows:

$$\begin{aligned} s_i &= \frac{exp(|Cb_i - 128| + |Cr_i - 128|)}{\sum_{j=0}^{1} exp(|Cb_j - 128| + |Cr_j - 128|)}, \\ S_i &= \frac{exp(s_i)}{\sum_{j=0}^{1} exp(s_j)}. \end{aligned} \quad (9)$$

The mask m_{sth} denoting the gradient-smooth area is jointly calculated from m_0 and m_1. Then, the loss function $\mathcal{L}_{smooth}$ is defined as:

$$\begin{aligned} \mathcal{L}_{smooth} &= m_{sth} \left\| I_f - I_w \right\|_1 , \\ &= (1 - m_0 \mid m_1) \left\| I_f - (S_0 I_0 + S_1 I_1) \right\|_1 . \end{aligned} \quad (10)$$

Although $\mathcal{L}_{steep}$ and $\mathcal{L}_{smooth}$ constrain the fidelity of image information in various areas, the stitching artifacts between the regions are unavoidable. A significant phenomenon attracts our attention: extremely discontinuous values near the edges of the stitching artifacts will create anomalous gradient responses. Consequently, we develop a globally constrained loss function $\mathcal{L}_{global}$ based on the selection of maximal gradient information as:

$$\mathcal{L}_{global} = \left\| \bigtriangledown I_f - max(\bigtriangledown I_0, \bigtriangledown I_1) \right\|_1 , \quad (11)$$

where $max(\cdot)$ means the element-wise maximum operation. To sum up, the three terms that makeup GIF is complementary to each other and achieve efficient unsupervised learning through in-depth analysis of gradient information.

4. Experiments

4.1. Datasets and Implementation Details

Since the recently proposed benchmark MEFB [48] collects a test set of 100 multi-exposure image pairs from multiple sources, the fusion results are able to comprehensively measure the performance of the algorithms, especially the generalization capability. For a fair comparison, our unsupervised training is performed on the external dataset SICE [6], which contains 482 well-aligned training samples (list will be available). Specifically, the training images are resized to the size of 128 × 128 with data augmentation performed (random flipping). The proposed network is implemented with the PyTorch framework. We conduct all our experiments on an NVIDIA 2080Ti GPU. During training for a total of 30 epochs, we set the batch size to 2 and the AdamW optimizer with a learning rate of 0.1×10^{-2}. Beneficial from the high model capacity utilization of the suggested new paradigm, the number of channels is set at 4.

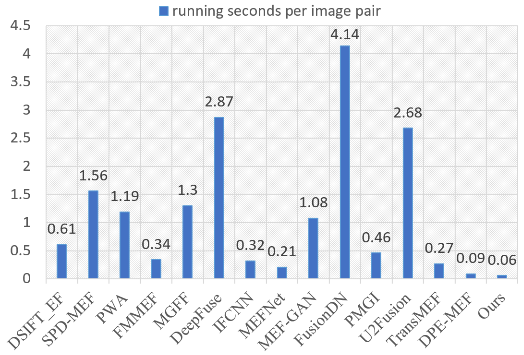


Figure 5. Average running time comparison over MEFB, a dataset containing 100 image pairs of average size 551 × 707.

4.2. Comparison with Previous Methods

To evaluate the proposed MEFNET, we compare it with several state-of-the-art algorithms: 1) traditional methods, including DSIFT_EF [23], FMMEF [20], MEFOpt [25], PWA [28], MGFF [3], and SPD_MEF [27]; 2) deep learning-based methods, including DeepFuse [35], MEFNet [26], IFCNN [51], MEF-GAN [42], FusionDN [41], PMGI [47], U2Fusion [40], TransMEF [34], and DPE-MEF [13]. The qualitative and quantitative experiments are detailed as follows.

Quantitative evaluation. Following previous works [13, 34, 48], we chose several widely acknowledged Image Quality Assessment (IQA) metrics for quantitative measurement. Specifically, structural similarity-based metrics (Q_W [32], *MEF-SSIM* [29]), information theory-based metrics (*EN [36]*, *FMI* [12], *NMI* [16], *CE* [4], *PSNR* [17], Q_{NCIE} [39]), image feature-based metrics (*AG* [9], $Q^{AB/F}$ [44], Q_P [52]), and human perception-inspired metrics (Q_{CB} [8], Q_{CV} [7], *VIF* [14]) are selected for a comprehensive comparison. Details about the metrics are presented in the Supplementary Material. The average evaluation results of all metrics are presented in Tab. 1, where ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

As can be seen, each method achieves superior performance under different evaluation metrics, but no one can succeed on all the metrics. Therefore, it is reasonable to conduct a comprehensive analysis based on the performance of all evaluation indicators. Traditional methods

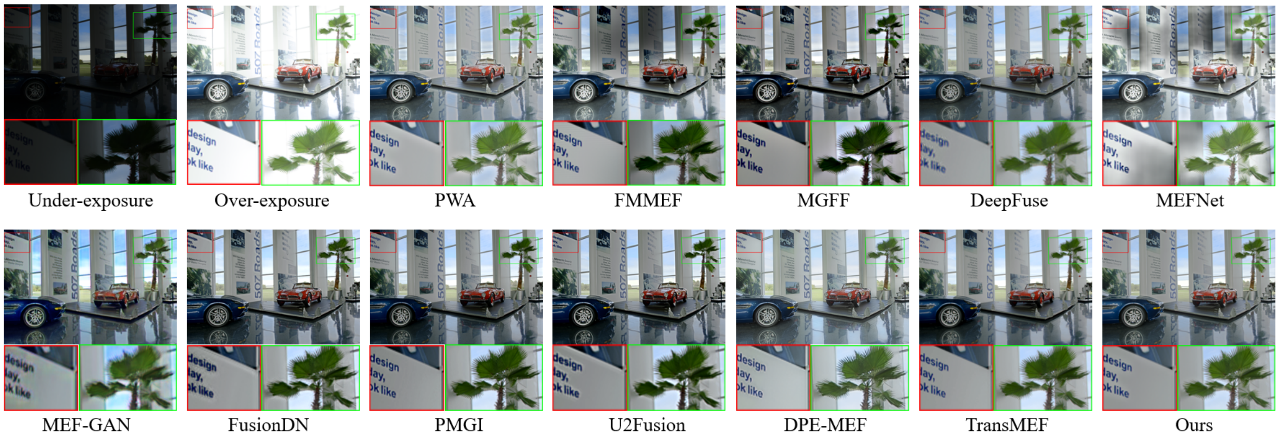


Figure 6. Qualitative comparison with the selected representative methods on the MEFB dataset [48]. As can be observed, although the conventional techniques provide high detail accuracy, the fused images lack global exposure consistency. Together with the recently proposed DPE-MEF and TransMEF, our method achieves the most advanced visual effects. More rigorous, our approach is better at balancing high and low exposure regions, as shown in outdoor sky vistas and indoor-exhibited cars. Please zoom in for the details.

Table 1. Average evaluation metric values of all methods on the MEFB dataset. The top three values are denoted in red, blue and green.

Method	*EN*↑	*FMI*↑	*NMI*↑	*PSNR*↑	Q_{NCIE}↑	*AG*↑	$Q^{AB/F}$↑	*CE*↓	Q_P↑	Q_W↑	*MEF-SSIM*↑	Q_{CB}↑	Q_{CV}↓	*VIF*↑
DSIFT_EF	7.3495	0.8947	0.5399	56.6056	0.8141	5.1000	0.6792	2.7297	0.6456	0.8619	0.9470	0.4663	825.4635	0.7265
FMMEF	7.3719	0.8970	0.4736	56.8570	0.8119	5.6441	0.7058	2.9333	0.6593	0.9072	0.9737	0.4525	650.8868	0.8876
MEFOpt	7.1864	0.8943	0.5855	56.7196	0.8157	5.7658	0.6884	3.2660	0.6098	0.8860	0.9355	0.4674	708.9612	0.7089
MGFF	7.1203	0.8897	0.6069	57.1132	0.8139	6.5461	0.6406	2.9861	0.6379	0.8463	0.9549	0.4678	308.4949	1.0184
PWA	7.0595	0.8971	0.7521	57.1722	0.8198	5.5585	0.6963	2.9955	0.6425	0.8949	0.9654	0.4460	288.0862	0.7249
SPD_MEF	7.1238	0.8871	0.6948	57.1051	0.8178	5.9832	0.6327	3.2146	0.6177	0.8039	0.9377	0.4551	353.4474	0.7772
DeepFuse	6.8504	0.8727	0.7408	57.1035	0.8177	3.4920	0.3884	3.0852	0.3517	0.5478	0.8968	0.3892	362.9800	0.5114
MEFNet	7.3899	0.8896	0.5967	56.5941	0.8166	6.0104	0.6746	3.0300	0.5954	0.8655	0.9139	0.4816	593.4327	0.8470
IFCNN	7.0347	0.8824	0.7708	57.1951	0.8186	6.0123	0.5960	3.4098	0.5616	0.8336	0.9432	0.4112	247.7693	0.7016
FusionDN	7.3293	0.8770	0.7251	56.9770	0.8178	6.7934	0.5363	2.9357	0.5044	0.7761	0.9240	0.4386	325.1348	0.9363
MEF-GAN	6.9547	0.8456	0.5727	56.9474	0.8132	4.6702	0.2836	2.8222	0.1239	0.3002	0.7722	0.3844	618.6932	0.5810
PMGI	7.0846	0.8854	0.7909	57.1165	0.8192	5.4189	0.5684	3.0084	0.5254	0.8035	0.9360	0.4208	293.9210	0.8077
U2Fusion	6.7392	0.8821	0.7675	57.0550	0.8179	5.5829	0.5356	2.9761	0.5046	0.7874	0.9304	0.4174	253.7540	0.8358
DPE_MEF	7.2383	0.8788	0.6120	57.1051	0.8141	6.6607	0.5995	4.1311	0.5612	0.8304	0.9452	0.3942	257.3125	0.7885
TransMEF	6.8603	0.8910	0.9229	57.1319	0.8237	4.5949	0.6035	2.8038	0.5649	0.8059	0.9499	0.4001	253.3766	0.7658
Ours	6.9942	0.8880	0.8311	57.1918	0.8206	5.0976	0.6584	2.7933	0.6073	0.8357	0.9621	0.4102	248.0949	0.7119

generally perform well in quantitative evaluation due to the excellent information preservation brought by their lossless image-level processing. Among them, FMMEF ranked in the top three in seven indicators, four of which won the first prize, giving the most dazzling answer. Our deep learning-based solution also performs admirably on six assessment measures, including five terms second places, thanks to the proposed effective paradigm and GIF loss.

Qualitative evaluation. As an ill-posed problem without ground truth images, evaluating the algorithm's visual performance is essential. It can be observed from Fig. 6, traditional methods that work well in quantitative comparison present results with an uneven exposure level, resulting in unrealistic perceptions. From the zoomed-in area in the red box, the results of DeepFuse, MEF-GAN, FusionDN, and U2Fusion all suffer from different degrees of degradation. The MEF-Net, designed based on the weight-map fusion paradigm, achieves good information fidelity but exhibits poor exposure consistency. DPE-MEF, TransMEF, and our method all exhibit excellent performance in terms of information fidelity and overall exposure consistency. A closer look reveals that DPE-MEF achieves higher brightness while sacrificing fidelity in high-exposure areas (the outdoor sky), and TransMEF lacks brightness in low-exposure areas (chassis of the indoor cars). Our method attains a relatively balanced effect, and the presented visual perception is closer to the real scenarios.

Running time comparison. The complexity of the model is an important criterion to measure its potential application in practical situations with constrained computing resources. In Fig. 5, we report the average inference time of the algorithms over the MEFB test set. As can be seen, we improved inference time by 30% and 77%, respectively, compared to previous state-of-the-art methods, DPE-MEF and TransMEF. A comprehensive evaluation combining results in quantitative and qualitative experiments demonstrates that our proposed method achieves the optimal balance between model efficiency and performance.

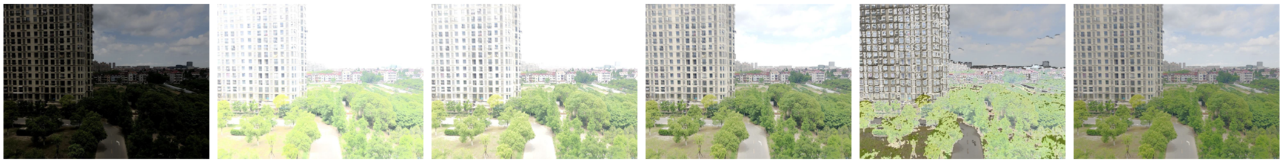
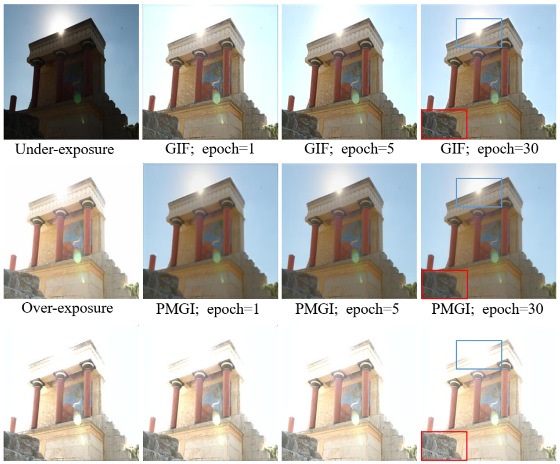

Under-exposure Over-exposure w/ $\mathcal{L}_{raw}$ w/o SW w/o $\mathcal{L}_{global}$ GIF

Figure 7. Visualization of the ablation results over the core components within the GIF. Three groups of control experiments are carried out: 1) w/ $\mathcal{L}_{raw}$: utilize the unprocessed raw gradient map as $\{w_i\}_{i=0}^{1}$; 2) w/o SW: substitute the saturation-based weighting (SW) in $\mathcal{L}_{smooth}$ with a simple weighted average operation; and 3) w/o $\mathcal{L}_{global}$: without the devised artifacts removal-oriented global constraint.

Table 2. Results of ablation experiments for various filter combinations. The best values are highlighted in bold.

Filter Combinations				*EN*↑	*FMI*↑	*NMI*↑	*PSNR*↑	Q_{NCIE}↑	*AG*↑	$Q^{AB/F}$↑	*CE*↓	Q_P↑	Q_W↑	*MEF-SSIM*↑	Q_{CB}↑	Q_{CV}↓	*VIF*↑
1 × 1	3 × 3	5 × 5	7 × 7														
✓				6.2271	0.8911	**0.9362**	57.0160	0.8232	4.7126	0.6023	**1.7070**	0.5769	0.7201	0.9051	0.3530	272.6144	0.7661
✓	✓			6.8272	0.8928	0.8846	57.1869	0.8229	4.6485	0.6368	2.5867	0.5848	0.7840	0.9368	0.3710	**239.0854**	0.6870
	✓	✓		6.1858	0.8912	0.9119	57.0075	0.8222	4.9193	0.6030	1.7781	0.5827	0.7231	0.9042	0.3523	271.0377	**0.7978**
✓	✓	✓	✓	6.7303	**0.8943**	0.9207	57.1467	**0.8243**	4.9066	0.6354	2.0139	0.5857	0.7717	0.9314	0.3641	251.8706	0.7359
✓	✓	✓		**6.9942**	0.8880	0.8311	**57.1918**	0.8206	**5.0976**	**0.6584**	2.7933	**0.6073**	**0.8357**	**0.9621**	**0.4102**	248.0949	0.7119


Under-exposure GIF; epoch=1 GIF; epoch=5 GIF; epoch=30

Over-exposure PMGI; epoch=1 PMGI; epoch=5 PMGI; epoch=30

MEF-SSIM; epoch=1 MEF-SSIM; e.=5 MEF-SSIM; e.=30 MEF-SSIM; e.=100

Figure 8. Visually compare the optimization process of the proposed GIF loss with the PMGI and MEF-SSIM loss functions.

4.3. Ablation Studies

In this section, we conduct a series of ablation studies about the proposed filter-dominated fusion paradigm and unsupervised loss function GIF.

Effects of various filter combinations. We conduct a series of ablation experiments to investigate whether filters with local information aggregation outperform pixel-level fusion. From Tab. 2, it can be observed that the combinations composed of a single 1×1 size filter or without 1×1 size filter are indeed inferior to the others in structural similarity-based indexes (*MEF-SSIM*, Q_W). And high mutual information metrics (*FMI*, *NMI*) indicate that their fusion results are biased toward one of the source images. In addition, there is no substantial performance benefit when the size of the fusion filter reaches 7×7. The final adopted scheme combines filters with sizes of 1×1, 3×3, and 5×5 to give the optimal performance.

Influence of three components within the GIF. In Fig. 7, we investigate the significance of the information fidelity-oriented designs Within GIF. In particular, we first replace $\mathcal{L}_{steep}$ and $\mathcal{L}_{smooth}$ based on gradient map preprocessing with a $\mathcal{L}_{raw}$ that builds the correlation $\{w_i\}_{i=0}^{1}$ from the raw gradient maps. It is visible that the lack of pre-processing in the gradient domain decreases the detail quality of the fused image. To illustrate the benefit of incorporating color saturation information, the saturation-based weighting (SW) operation used to generate smooth region optimization targets is replaced by weighted averaging. Under such conditions, the overall contrast of the produced fusion result is inferior to that of GIF. Besides, the fusion results exhibit severe stitching artifacts due to the removal of $\mathcal{L}_{global}$. In conclusion, the ablation studies indicate the efficacy of GIF's fundamental components and their complimentary functions during optimization.

Comparison with the MEF-SSIM and PMGI. To demonstrate the superiority of the proposed GIF loss in unsupervised multi-exposure fusion optimization, we compare it with the MEF-SSIM [29] loss function and the PMGI [47] loss function. MEF-SSIM is inspired by structural similarity (SSIM) and decomposes images into structural, intensity, and strength terms at the patch level for the calculations. As shown in Fig. 8, it is challenging for MEF-SSIM to handle the extremely overexposed or underexposed locations. The design of PMGI is consistent with Equ. (5), where the function $\Phi(\cdot)$ adopts the gradient operator and the identity transformation, and $\{w_i\}_{i=0}^{1}$ is assigned a constant value. Since no independent correspondence between the fused image and the source image is established, PMGI lacks spatial adaptability in its constraints, limiting its performance. Our proposed GIF achieves the best information fidelity, and the optimization process is more stable than the commonly used MEF-SSIM.

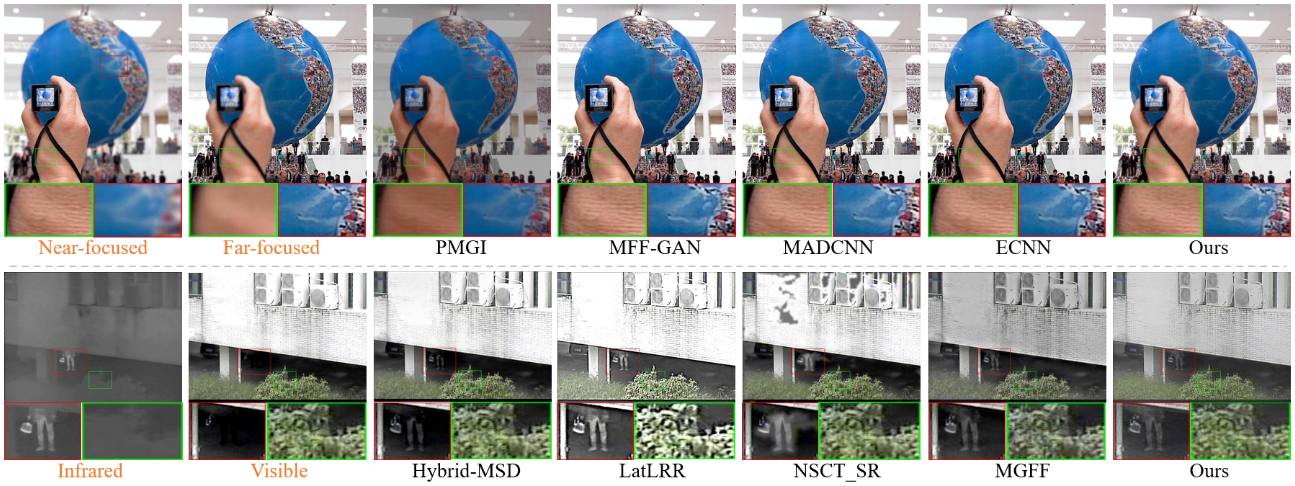


Figure 9. Visually demonstrate the effectiveness of the proposed efficient-oriented fusion paradigm on a broad range of image fusion tasks, using MFF and VIF as examples. As can be observed, our method achieves comparable visual performance with representative methods.

Table 3. Quantitative results on the multi-focus image fusion benchmark [49]. The best values are highlighted in bold.

Method	*EN*↑	*FMI*↑	Q_{NCIE}↑	*AG*↑	$Q^{AB/F}$↑	Q_W↑	Q_{CV}↓	*VIF*↑
ECNN	7.1437	0.8778	0.8168	7.7970	0.5800	0.8682	111.3797	0.7964
MADCNN	7.1553	**0.8817**	0.8212	7.7907	**0.6634**	**0.8898**	121.12008	0.8281
PMGI	7.0272	0.8663	0.8154	5.8564	0.4553	0.6826	304.6180	0.8773
MFF-GAN	7.1101	0.8703	0.8158	**8.3671**	0.5357	0.8284	161.1865	0.8583
Ours	**7.3846**	0.8707	**0.8221**	7.1355	0.6265	0.8470	**77.5857**	**0.9156**

Table 4. Quantitative results on the visible-infrared image fusion benchmark [50]. The best values are highlighted in bold.

Method	*AG*↑	*EI*↑	*EN*↑	*PSNR*↑	Q_{CV}↓	*RMSE*↓	*SF*↑	*SSIM*↑
MGFF	5.839	60.607	7.114	58.212	676.9	0.1092	17.916	1.406
Hybrid_MSD	6.126	63.491	7.304	58.173	**510.9**	0.1102	19.659	1.405
LatLRR	**8.962**	**92.813**	6.909	56.180	697.3	0.1686	**29.537**	1.184
NSCT_SR	6.492	67.956	**7.396**	57.435	1447	0.1314	19.389	1.277
Ours	5.1668	53.0491	6.8486	**58.3247**	837.89	**0.1068**	14.9646	**1.4202**

4.4. Extension Experiments

To investigate the feasibility of the proposed fusion paradigm in a spacious range of image fusion problems, we conduct extension experiments over Multi-Focus image Fusion (MFF) and Visible-Infrared image Fusion (VIF). It should be noted that other experimental settings are kept consistent except for dataset changes. Specifically, on the MFF benchmark [49], we report our performance compare to representative algorithms PMGI [47], MFF-GAN [46], MADCNN [18], and ECNN [1]. Experiments on the VIF task are compared with Hybrid-MSD [53], LatLRR [21], NSCT_SR [2] and MGFF [3] on benchmark [50]. Fig. 9 and Tab. 3,4 depict the results of qualitative and quantitative experiments. As can be observed in the top column of Fig. 9, our method combines the near-focus and far-focus images to generate the resulting image with global sharpening properties. The color channels are derived from the visible image in the visible and infrared image fusion procedure. The comparison results in VIF illustrated that our method achieves better information fidelity, merging the valuable information of the source image without introducing artifacts. In conclusion, the comparable performance to previous state-of-the-art methods highlights the applicability of our efficient-focused fusion paradigm and information fidelity-targeted unsupervised loss function GIF to a vast array of image fusion problems.

5. Conclusion

In this work, we provide a new solution for unsupervised multi-exposure image fusion (MEF) by redesigning the fusion paradigm and loss function. The hierarchical representation-aware fusion rule and local information aggregation-based image-level fusion empower the most efficient fusion paradigms, i.e., the filters-prediction dominated fusion. Inspired by the informative properties of the source images in the gradient domain, a gradient-driven image fidelity loss GIF is meticulously designed for unsupervised learning. Consequently, we attain the best visual performance with a 30% reduction in inference time compared to previous state-of-the-art MEF approaches. Furthermore, the superiority of our approach for a wide variety of image fusion problems is demonstrated in extension experiments involving multi-focus image fusion and visible-infrared image fusion. In future work, we will make further improvements to achieve a unified framework for multi-modal image fusion. Code will be available.

6. Acknowledgments

This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Mostafa Amin-Naji, Ali Aghagolzadeh, and Mehdi Ezoji. Ensemble of cnn for multi-focus image fusion. *Information Fusion*, 51:201–214, 2019.

[2] Mostafa Amin-Naji, Ali Aghagolzadeh, and Mehdi Ezoji. Ensemble of cnn for multi-focus image fusion. *Information Fusion*, 51:201–214, 2019.

[3] Durga Prasad Bavirisetti, Gang Xiao, Junhao Zhao, Ravindra Dhuli, and Gang Liu. Multi-scale guided image and video fusion: A fast and efficient approach. *Circuits, Systems, and Signal Processing*, 38(12):5576–5605, 2019.

[4] DM Bulanon, TF Burks, and V Alchanatis. Image fusion of visible and thermal images for fruit detection. *Biosystems Engineering*, 103(1):12–22, 2009.

[5] Peter J Burt and Raymond J Kolczynski. Enhanced image capture through fusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 173–182, 1993.

[6] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[7] Hao Chen and Pramod K Varshney. A human perception inspired quality metric for image fusion based on regional information. *Information Fusion*, 8(2):193–207, 2007.

[8] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image and Vision Computing*, 27(10):1421–1432, 2009.

[9] 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.

[10] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Transactions on Image Processing*, 30:3098–3112, 2021.

[11] A Ardeshir Goshtasby. Fusion of multi-exposure images. *Image and Vision Computing*, 23(6):611–618, 2005.

[12] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5):744–756, 2011.

[13] Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Information Fusion*, 79:248–262, 2022.

[14] 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.

[15] Naila Hayat and Muhammad Imran. Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter. *Journal of Visual Communication and Image Representation*, 62:295–308, 2019.

[16] M Hossny, S Nahavandi, and D Creighton. Comments on ”information measure for performance of image fusion”. *Electronics Letters*, 44(18):1066–1067, 2008.

[17] P Jagalingam and Arkal Vittal Hegde. A review of quality metrics for fused image. *Aquatic Procedia*, 4:133–142, 2015.

[18] Rui Lai, Yongxue Li, Juntao Guan, and Ai Xiong. Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. *IEEE Access*, 7:114385–114399, 2019.

[19] Sang-hoon Lee, Jae Sung Park, and Nam Ik Cho. A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1737–1741, 2018.

[20] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:5805–5816, 2020.

[21] Hui Li and Xiao-Jun Wu. Infrared and visible image fusion using latent low-rank representation. *arXiv preprint arXiv:1804.08992*, 2018.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[23] Yu Liu and Zengfu Wang. Dense sift for ghost-free multi-exposure fusion. *Journal of Visual Communication and Image Representation*, 31:208–224, 2015.

[24] 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.

[25] Kede Ma, Zhengfang Duanmu, Hojatollah Yeganeh, and Zhou Wang. Multi-exposure image fusion by optimizing a structural similarity index. *IEEE Transactions on Computational Imaging*, 4(1):60–72, 2017.

[26] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:2808–2819, 2019.

[27] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017.

[28] Kede Ma and Zhou Wang. Multi-exposure image fusion: A patch-wise approach. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1717–1721, 2015.

[29] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015.

[30] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018.

[31] Sujoy Paul, Ioana S Sevcenco, and Panajotis Agathoklis. Multi-exposure and multi-focus image fusion in gradi-

ent domain. *Journal of Circuits, Systems and Computers*, 25(10):1650123, 2016.

[32] Gemma Piella and Henk Heijmans. A new quality metric for image fusion. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages III–173, 2003.

[33] Ying Qi, Shangbo Zhou, Zihan Zhang, Shuyue Luo, Xiaoran Lin, Liping Wang, and Baohua Qiang. Deep unsupervised learning based on color un-referenced loss functions for multi-exposure image fusion. *Information Fusion*, 66:18–39, 2021.

[34] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2126–2134, 2022.

[35] 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.

[36] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[37] Jian Sun, Hongyan Zhu, Zongben Xu, and Chongzhao Han. Poisson image fusion based on markov random field fusion model. *Information Fusion*, 14(3):241–254, 2013.

[38] Qiantong Wang, Weihai Chen, Xingming Wu, and Zhengguo Li. Detail-enhanced multi-scale exposure fusion in yuv color space. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2418–2429, 2019.

[39] Qiang Wang, Yi Shen, and Jian Qiu Zhang. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, 200(3-4):287–295, 2005.

[40] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[41] 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*, volume 34, pages 12484–12491, 2020.

[42] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020.

[43] Jianbo Xu, Youjun Huang, and Jianli Wang. Multi-exposure images of wavelet transform fusion. In *Proceedings of the International Conference on Digital Image Processing*, volume 8878, pages 67–71. SPIE, 2013.

[44] Costas S Xydeas, Vladimir Petrovic, et al. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

[45] Zhiguang Yang, Youping Chen, Zhuliang Le, and Yong Ma. Ganfuse: a novel multi-exposure image fusion method based on generative adversarial networks. *Neural Computing and Applications*, 33(11):6133–6145, 2021.

[46] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021.

[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*, volume 34, pages 12797–12804, 2020.

[48] Xingchen Zhang. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74:111–131, 2021.

[49] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[50] Xingchen Zhang, Ping Ye, and Gang Xiao. Vifb: A visible and infrared image fusion benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 104–105, 2020.

[51] 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.

[52] Jiying Zhao, Robert Laganiere, and Zheng Liu. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control*, 3(6):1433–1447, 2007.

[53] Zhiqiang Zhou, Bo Wang, Sun Li, and Mingjie Dong. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters. *Information Fusion*, 30:15–26, 2016.