Unsupervised Multi-Exposure Image Fusion Breaking Exposure Limits via Contrastive Learning

Han Xu*, **Haochen Liang***, **Jiayi Ma**†

Electronic Information School, Wuhan University, Wuhan 430072, China
xu_han@whu.edu.cn, njnu_lianghaochen@163.com, jyma2010@gmail.com

Abstract

This paper proposes an unsupervised multi-exposure image fusion (MEF) method via contrastive learning, termed as *MEF-CL*. It breaks exposure limits and performance bottleneck faced by existing methods. MEF-CL firstly designs similarity constraints to preserve contents in source images. It eliminates the need for ground truth (actually not exist and created artificially) and thus avoids negative impacts of inappropriate ground truth on performance and generalization. Moreover, we explore a latent feature space and apply contrastive learning in this space to guide fused image to approximate normal-light samples and stay away from inappropriately exposed ones. In this way, characteristics of fused images (*e.g.*, illumination, colors) can be further improved without being subject to source images. Therefore, MEF-CL is applicable to image pairs of any multiple exposures rather than a pair of under-exposed and over-exposed images mandated by existing methods. By alleviating dependence on source images, MEF-CL shows better generalization for various scenes. Consequently, our results exhibit appropriate illumination, detailed textures, and saturated colors. Qualitative, quantitative, and ablation experiments validate the superiority and generalization of MEF-CL. Our code is publicly available at https://github.com/hanna-xu/MEF-CL.

Introduction

The dynamic range of a natural scene is usually much larger than that of a digital camera (Kou et al. 2017). The camera can only capture scenes within a limited dynamic range, resulting in unsatisfactory image illumination, content, and other characteristics. As shown in the source images in Fig. 1, it is difficult for a single image to describe all the scene contents. These improperly exposed images can be enhanced by single-image enhancement methods. However, under-exposed regions suffer from not only poor visualization but also low signal-to-noise ratio (SNR). Besides, the missing contents in over-exposed regions cannot be recovered by correction. Thus, the single-image based enhancement still shows limited performances. By comparison, multi-exposure image fusion (MEF) can merge the information in multi-exposure images into a fused image. The fused image preserves the details and colors of source images and exhibits higher dynamic range and better scene description. Thus, MEF has been widely used in many fields, such as industrial welding, remote sensing, medical imaging, unmanned driving and other fields (Xing et al. 2018).

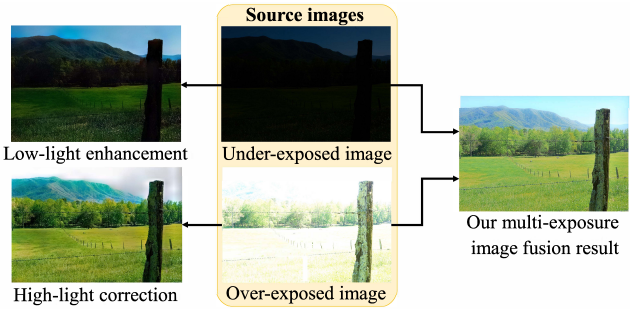

Figure 1: Example of single-image enhancement by single-image enhancement method (Afifi et al. 2021) and the proposed multi-exposure image fusion method. By integrating more information, the fusion result shows better scene description than the single-image enhancement result.

In the past few decades, a variety of MEF algorithms have been proposed, which can be divided into traditional and deep learning-based methods. The traditional methods use the relevant digital transformation to analyze and design fusion rules in the spatial or transformation domain. However, these methods extract features manually and thus show *limited generality to complex scenes and various illumination environments*. Moreover, considering the fusion feasibility, they forcibly use the same transformation for images with different exposures without considering differences, resulting in *limited expression ability of extracted features*.

Recently, deep learning has driven the development of image fusion. The deep learning-based methods can rely on distinctive network branches to achieve differentiated and targeted feature extraction. Moreover, based on deep learning, the methods can also learn adaptive feature fusion strategies (Xu, Zhang, and Ma 2021). These methods can be divided into supervised and unsupervised methods. Supervised methods rely on ground truth (GT) for supervision, while *GT in MEF does not actually exist and is created artificially, which maybe inaccurate and unrealistic*. The un-

*These authors contributed equally.
†Corresponding author

supervised methods usually depend on some image characteristics to design constraints and fusion rules. Thus, the characteristics of the fused image have *strong correlation with source images, resulting in performance bottlenecks*. For this reason, existing methods *forcibly require a pair of over-exposed and under-exposed images* for fusion.

To solve these issues, we propose a novel end-to-end unsupervised MEF model based on contrastive learning, named as *MEF-CL*. The proposed method has three advantages: **i)** As an unsupervised method, MEF-CL does not require pseudo (artificially created) GT for supervision and thus can avoid the remaining inappropriate under-/over-exposed regions in fusion results; **ii)** Compared with existing unsupervised methods, the proposed method can break exposure limits and then broaden its application. Specifically, unlike existing methods only applicable to a pair of under-exposed and over-exposed images to generate a properly exposed fused image, MEF-CL can also fuse two over-exposed images or two under-exposed images. As shown in Fig. 2, when fusing a pair of over-exposed or under-exposed images, the fused image of the existing MEF method exhibits the illumination between those of two source images. In other words, the fused image is still over-exposed or under-exposed. By comparison, in both cases, our fused images in Fig. 2 show significantly appropriate and satisfactory illumination and scenes; **iii)** When focusing on the common multi-exposure image fusion (fusing a pair of over-exposed and under-exposed images), our fused image also shows clear details and advantageous illumination.

These advantages benefit from the framework of MEF-CL. We firstly design a fusion network and similarity constraint to generate the fused image, which preserves scene contents of source images. Then, we apply contrastive learning to adaptively adjust the illumination and other apparent characteristics of the fused image for better performance. Specifically, a network maps images into a latent feature space, where we design a contrastive learning block. This block uses the contrastive loss guide fused images to approach the characteristics of normal-light images and stay away from those of improperly exposed images. The contributions are summarized as the following three aspects:

- It is the first time that contrastive learning is employed to multi-exposure image fusion. We apply contrastive learning to adaptively adjust the characteristics of fused images. In this way, the characteristics of fused images are guided to approach those of normal-light images for better performance, such as richer texture details, superior illumination, brighter colors, *etc*.
- Compared with existing unsupervised MEF methods, our method breaks the exposure limit and thus broadens the application. It can fuse any two different-exposure images rather than being limited to an over-exposed image and an under-exposed image in existing MEF methods.
- It is a novel end-to-end unsupervised MEF network. It mitigates the stumbling block caused by the lack of ground truth. Compared with supervised MEF methods taking pseudo ground truth for supervision, our fusion results are not restricted by the inappropriate influence of pseudo ground truth. Thus, the proposed method shows better generalization for various scenes.

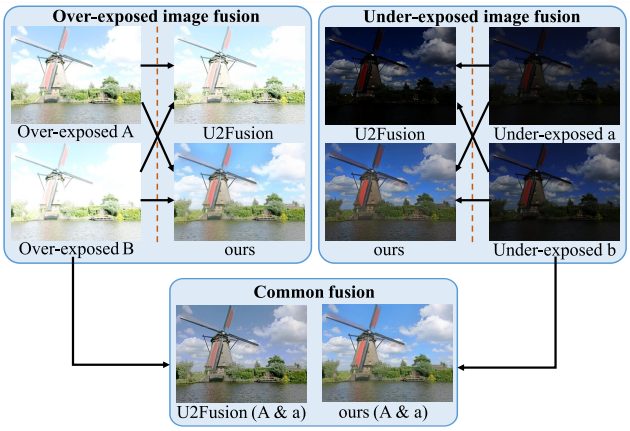

Figure 2: Comparison of fusion performances for a pair of over-exposed images, a pair of under-exposed images, and the common over-exposed and under-exposed images (state-of-the-art competitor: U2Fusion (Xu et al. 2022)).

Related Work

Multi-exposure Image Fusion. i) *Traditional methods.* They mainly include pixel-based, patch-based and sparse representation-based methods. Mertens (Mertens, Kautz, and Van Reeth 2009) built a pyramid and set the weight distribution by measuring the quality evaluation methods of contrast, saturation and exposure per pixel. The patch-based method takes full account of the relationship between pixels and balances the illumination change in the neighborhood. Sparse representation has been widely used because of its clear components and texture details and no artifacts (Wang, Liu, and He 2014; Sakai et al. 2015). Almost all these methods manually design features, but the generality is limited for complex scenes. Moreover, to ensure the feasibility of subsequent feature fusion, traditional methods are forced to use the same transformation for different source images, without considering their feature differences, which may lead to unsatisfactory feature expression ability. ii) *Deep learning-based methods.* These methods have made great achievements in performance and visual effect recently. Prabhakar *et al*. (Ram Prabhakar, Sai Srikar, and Venkatesh Babu 2017) introduced deep learning into the MEF task for the first time. This method uses MEF-SSIM as the training loss function and constructs a new CNN architecture to realize unsupervised learning. Xu *et al*. (Xu, Ma, and Zhang 2020) applied the generative adversarial networks (GAN) to MEF for the first time. They introduced a self attention mechanism into the proposed architecture to further correct artifacts according to the fusion results.

The above deep learning-based methods can be divided into supervised and unsupervised methods. Supervised learning exhibits higher model fitting accuracy. However, the key factor that limits its performance is the lack of GT. At present, the GT is pseudo which is actually selected from the results of existing 13 representative methods. The

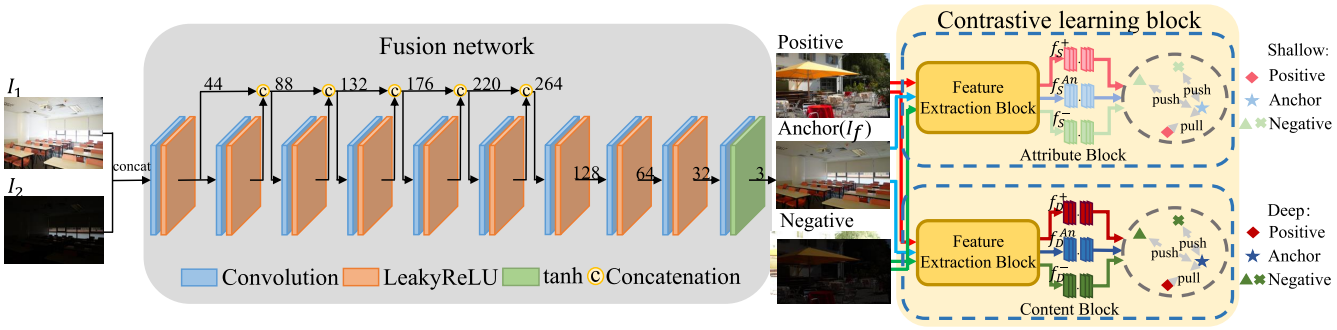

Figure 3: Overall framework of the proposed unsupervised multi-exposure image fusion network based on contrastive learning.

performance of supervised methods is restricted by the existing methods. In addition, since it is not taken directly by the HDR camera, the artifacts in the pseudo GT will also directly affect the final fusion result.

The unsupervised learning is based on gradients, weights, or other characteristics to maintain similarity between fusion result and source images. For example, PMGI (Zhang et al. 2020a) and SDNet (Zhang and Ma 2021) use gradient and intensity. U2Fusion (Xu et al. 2022) uses gradient of deep features. However, the strong correlation between the fused image and source images results in performance bottlenecks and limited dynamic range. Moreover, they set the threshold of loss functions artificially in advance, restricting the upper limit of fusion performance in different scenes.

Contrastive Learning. According to whether the data is labeled or not, it can be divided into supervised and unsupervised learning. Unsupervised learning is mainly divided into generative learning and contrastive learning. Contrastive learning focuses on learning common representations of similar instances and distinguishing differences between dissimilar instances. The Moco method (He et al. 2020) has achieved good results by increasing the proportion of negative samples. Chen *et al*. (Chen et al. 2020) proposed the SimCLR method to obtain two transformed images belonging to the same image by randomly transforming the input image. The two transformed images have the same high-level semantic information. Therefore, by maximizing the similarity of the two modal data, contrastive learning has achieved remarkable results. Therefore, contrastive learning has great potential in the unsupervised field. Through training, the network can push the anchor close to the positive sample and away from the negative sample in the deep feature space. In MEF, we aim to make the fused image close to normal-light images and away from extreme exposure images to achieve better fusion performance.

Proposed Method

Problem Formulation

This paper studies the unsupervised MEF method and proposes an unsupervised MEF network breaking exposure limits via contrastive learning. Taking two source images I_1 and I_2 with different exposures as inputs, the network aims to learn the mapping function F corresponding to the fusion network. The fusion network is expected to generate a fused image I_f with rich details and appropriate illumination and colors. The overall process is defined as follows:

$$I_f = F\left(I_1, I_2, \theta\right), \tag{1}$$

where θ represents the parameters in the fusion network.

To realize $F(\cdot)$, the main framework is shown in Fig. 3. It mainly consists of a fusion network and a contrastive learning block. The fusion network takes source images as input, extracts features from source images, fuses the extracted features and finally maps the fused features back to the image domain to generate the fused image. Following this process, it is possible to adjust other apparent characteristics of the fused image while preserving the contents of source images.

To preserve the contents of source images into the fused image, we define a loss function to optimize the fusion network which constrains the similarity between the fused image and source images. The self-supervised loss function retains information from both source images for more detailed scene descriptions and thus does not require ground truth for guidance. However, the similarity constraint alone will result in the high correlation between the fused image and source images and place an upper limit on the fused image. For this reason, the fusion results of existing methods are subject to source images, making them only applicable to a pair of under-exposed and over-exposed images.

Therefore, on the basis of content preservation, we introduce contrastive learning into our method by designing the subsequent contrastive learning block for further optimization. In our method, taking any two source images with different exposures, we expect to produce a moderately exposed fused image similar to normal-light images. As there is no ground truth for supervision, it is impossible to perform pixel-level similarity constraint. Alternatively, we map multi-exposure images into some latent feature space. In this space, the fused image is expected to be close to normal-light images of other scenes to exhibit some common characteristics of normal-light images, such as appropriate illumination, saturate colors, abundant details, and some other apparent characteristics. With the features of normal-light images of other scenes as positive samples for guidance, the features of fused image can get rid of the dependence on source images and be pulled to normal exposure and colors. Moreover, the source images are improperly exposed images. Their features can be additionally set as negative samples for contrastive learning. The features of the fused image

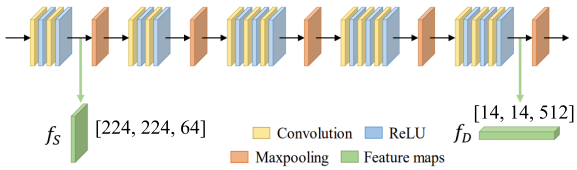

Figure 4: Extracted shallow-level and deep-level features.

should be far apart from those of under-/over-exposed images. Then, the fused image can be further pushed to normal exposure and colors. To realize it, we design a contrastive loss function to guide the process in the feature space.

For the latent feature space, we expect the extracted features to be a comprehensive representation of images. We select the VGG-16 network (Simonyan and Zisserman 2014) pre-trained on large-scale natural images to map source images into the feature space. As shown in Fig. 4, VGG-16 can extract a variety of features from the image from shallow to deep. The information in different levels represents various characteristics. For example, shallow-level features contain characteristics such as textures and shape details; while deep-level features characterize the overall style of the image, such as the overall illumination, color saturation, *etc*. In this work, we select both shallow-level and deep-level features as the feature space for contrastive learning.

Loss Function

To preserve the scene contents in source images, we implement the similarity constraint from two aspects: intensity distribution and structural similarity. Thus, the loss function of the proposed method consists of three items, including the mean square error (MSE) loss (L_{MSE}), the structural similarity loss (L_{SSIM}) and the contrastive loss ($L_{InfoNCE}$). The general loss function is defined as follows:

$$L = L_{MSE} + \lambda_1 L_{SSIM} + \lambda_2 L_{InfoNCE}, \tag{2}$$

where λ_1, λ_2 are hyper-parameters to control the trade-off.

Mean Square Error Loss The mean square error loss L_{MSE} is used to constrain the intensity distribution differences of images at the pixel level. The calculation formula is defined with L_2 loss:

$$L_{MSE} = \frac{1}{2} \left\| I_f - I_1 \right\|_2 + \frac{1}{2} \left\| I_f - I_2 \right\|_2 . \tag{3}$$

Structural Similarity Loss The structural similarity metric (SSIM) models distortion according to the similarity in light, contrast and structural information. We use the structural similarity loss L_{SSIM} to constrain the fused image in terms of structure information so that the fused image can fully learn the structure information of source images. The calculation formula is as follows:

$$L_{SSIM} = 1 - \frac{SSIM(I_f, I_1) + SSIM(I_f, I_2)}{2}. \tag{4}$$

Contrastive Loss $L_{InfoNCE}$ (Hjelm et al. 2018; Hu et al. 2021) represents the contrastive loss. NCE stands for noise contrastive estimation, which is widely used in contrastive learning networks. Contrastive learning is to make similar samples closer and push different samples away. One way to achieve this is to use similarity metric to measure the similarity between the features of two samples. In this paper, we use the mean of pixel level cosine similarity as the similarity between feature maps. With the feature maps f^A and f^B of size $H \times W \times C$, the definition of their similarity is:

$$S\left(f^A, f^B\right) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \frac{f_{hw}^A f_{hw}^B}{\left\| f_{hw}^A \right\| \left\| f_{hw}^B \right\|}. \tag{5}$$

Most contrastive learning focuses on comparing features with noise contrastive estimation (NCE) (Gutmann and Hyvärinen 2010) function. NCE performs nonlinear logistic regression to distinguish between observation data and some artificially generated noise. This function is defined as:

$$L_{NCE} = -\log \frac{\exp\left(S\left(q, k^+\right)/\tau\right)}{\exp(S(q,k^+)/\tau) + \exp(S(q,k^-)/\tau)}, \tag{6}$$

where $S(\cdot)$ represents the cosine similarity between feature maps as mentioned above. τ represents the temperature super parameter, k^+, k^- and q represent the characteristics of positive, negative and anchor samples, respectively.

Moco (He et al. 2020) proved that the employ of more negative samples can make fuller use of contrastive learning. Thus, the contrastive loss function $L_{InfoNCE}$ extends the calculation method from logistic regression to cross entropy by improving the NCE loss function, making full use of negative samples. For the i-th sample, the calculation formula L_i is as follows:

$$L_i = -\log \frac{\exp\left(S\left(f_i^T, f_i^+\right)/\tau\right)}{\exp\left(S\left(f_i^T, f_i^+\right)/\tau\right) + \sum_{j=1}^{N} \exp\left(S\left(f_i^T, f_j^-\right)/\tau\right)}, \tag{7}$$

where f_i^+, f_j^- and f_i^T represent the features of positive samples, negative samples and fused images respectively, and N represents the number of negative samples. In this work, N is set to 360. Finally, the total contrastive loss is as follows:

$$L_{InfoNCE} = \frac{1}{N} \sum_{i=1}^{N} L_i. \tag{8}$$

Network Architecture

Fusion Network We use the fully connected DenseNet to make full use of the feature information of the source image to generate the fused image. It solves the problem of traditional encoder-decoder networks with many parameters, low feature utilization, and severe information distortion. Thus, it can avoid the fused image from losing feature information in source images. The framework adopts DenseNet (Huang et al. 2017) architecture, as shown in Fig. 3. It consists of 10 layers. The kernel size is 3×3 and the stride is 1. The activation function of the first 9 layers is LeakyReLU with the slope set to 0.2 and that of the last layer is tanh.

Contrastive Learning Block The contrastive learning block consists of three main components, including feature extraction block, attribute block and content block. The framework is shown in Fig. 3.

Feature Extraction Block. The levels of image features increase with the network depth. By increasing the layer numbers of convolutional neural network, we can extract higher-level and richer image features. Based on it, we use the pre-trained VGG-16 network (Simonyan and Zisserman

2014) to extract multi-level image features, as shown in Fig. 4. The numbers are the sizes of the extracted feature maps. For an input RGB image I_1, we separate it into RGB channels respectively. Then, we send the channels into the VGG-16 network and extract the feature maps from them. The extracted feature maps are concatenated along the channel dimension as f_S and f_D.

Attribute Block and Content Block. The deeper the convolutional neural network is, the more deep-level features that people cannot perceive in the image can be extracted. In Fig. 4, f_S belongs to shallow-level feature, which retains a large number of texture and shape details, while f_D mainly retains deep-level features, such as content or spatial structure. Therefore, the attribute block uses the shallow-level features to improve the texture of the fused image, while the content block uses the deep-level features to improve the scene content information of the fused image.

In the contrastive learning task, we take the naturally captured normal-light images as positive samples, and the images with over-exposed or under-exposed regions as negative samples. We extract the shallow-level or deep-level features of these samples and the fused images. The contrastive loss function promotes the features of the fused images in the feature space, making them close to the positive samples and away from the negative samples and changing the weight of the fusion network. Finally, the fused image has rich texture details and visual effects.

Experiment Results and Discussions

Implementation Details

We conduct experiments on the SICE dataset (Cai, Gu, and Zhang 2018)[1] and perform the verification on different scenes, including indoor and outdoor scenes.We randomly selected 479 image sequences as the training set. The remaining 80 image sequences are as the test set. We select the brightest and darkest images in each sequence as the over-exposed and under-exposed negative sample images, respectively. It can significantly divide the over-exposed, under-exposed, and normal-light domains in the latent feature space. In the training phase, all the images are cropped into patches of size 64×64. The hyper-parameters are set as: $\lambda_1 = 10$, $\lambda_2 = 20$, $\tau = 0.01$. The batch size is set to 20, the training epoch is 2, and the learning rate is 0.0001. We use the RMSProp optimizer for optimization. The overall framework is implemented in TensorFlow. The experiments are performed on an NVIDIA Geforce GTX Titan V GPU.

Performance Evaluation

We use nine state-of-the-art methods for comparison, including two traditional methods: GFF (Li, Kang, and Hu 2013), FMMEF (Li et al. 2020), and seven deep learning-based methods: Deepfuse (Ram Prabhakar, Sai Srikar, and Venkatesh Babu 2017), MEF-Net (Ma et al. 2019), IFCNN (Zhang et al. 2020b), MEF-GAN (Xu, Ma, and Zhang 2020), U2Fusion (Xu et al. 2022), AGAL (Liu et al. 2022), and TransMEF (Qu et al. 2022). IFCNN and MEF-GAN are supervised. MEF-Net, Deepfuse, U2Fusion, AGAL and TransMEF are unsupervised.

Qualitative Comparison To validate the effectiveness and generalization of MEF methods, the results on outdoor and indoor scenes are compared in Figs. 5 and 6, respectively. By comparison, our method shows two advantages.

First, our methods show better generalization for complex and various scenes. As shown in Figs. 5 and 6, when dealing with indoor and outdoor scenes, many competitors exhibit large differences in the fusion performance. For example, in Fig. 5, AGAL not only recovers the color in the sky, but also restores the texture details of the house clearly. It reflects the powerful feature extraction ability and color recovery ability of its self-attention mechanism. However, in Fig. 6, when dealing with the indoor scene where the source images show large exposure differences, AGAL shows obvious texture and color distortion. Similarly, GFF and MEF-Net show satisfied performances in the indoor scene, especially shown in the color card. As shown in Fig. 6, their fused images show bright colors and high saturation. However, in Fig. 5, their fused images suffer from serious artifacts and dark regions. By comparison, our MEF-CL shows good fusion performance in both indoor and outdoor scenes. The reason is that the supervised method essentially forces the fusion results to be close to the pseudo GT while the pseudo GT may have color distortion, halo artifacts and other drawbacks. These drawbacks ultimately limit the fusion performance and affect the fusion results. In comparison, unsupervised methods usually use texture or luminance loss to generate the ideal fused image. The pre-defined loss function places an upper limit on the image quality of the fused image artificially. When dealing with complex multi-exposure scenes, the parameters in the network cannot be flexibly adjusted, resulting in performance gaps. In our method, we train the fusion network with a large number of positive and negative samples. Through contrastive learning, over-exposed, under-exposed, and normal-light domains can be divided in the latent feature space. It does not need the pseudo GT for supervision as supervised methods and is more flexible and accurate than artificial setting as in existing unsupervised methods. Thus, MEF-CL can achieve superior generalization.

Second, our fusion results show more appropriate illumination, more detailed textures, and more saturated colors. On the one hand, as shown in Figs. 5 and 6, the fusion results of GFF, Deepfuse, IFCNN, MEF-GAN, and U2Fusion show dark illumination in some regions. The scenes in these regions suffer from poor visibility due to inappropriate illumination. By comparison, our fusion results show more appropriate illumination for these regions. On the other hand, as shown in the first example in Fig. 5 and the second example in Fig. 6, our fusion results show more detailed and clearer textures than all the competitors. And the artifacts can also be alleviated in our results. Moreover, our fusion results exhibit more saturated colors. As for the traditional methods, GFF and FMMEF produce serious color distortion and artifacts in the fusion results. In the deep learning-based methods, some unsupervised methods still use traditional methods to fuse the colors of source images. It inevitably

[1]https://github.com/csjcai/SICE

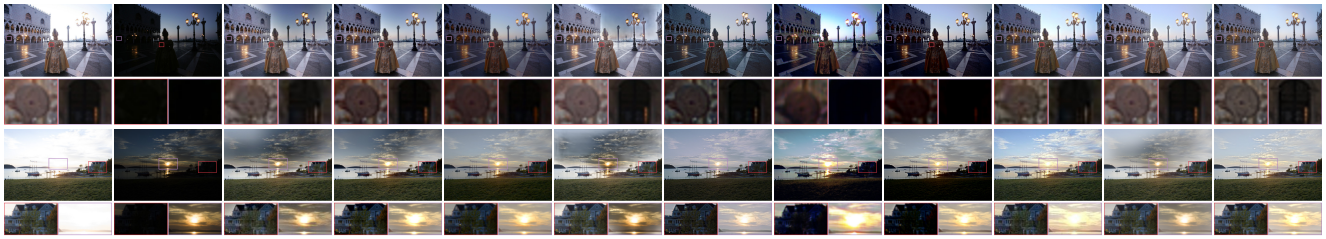

Figure 5: Qualitative comparison results on outdoor multi-exposure image pairs.

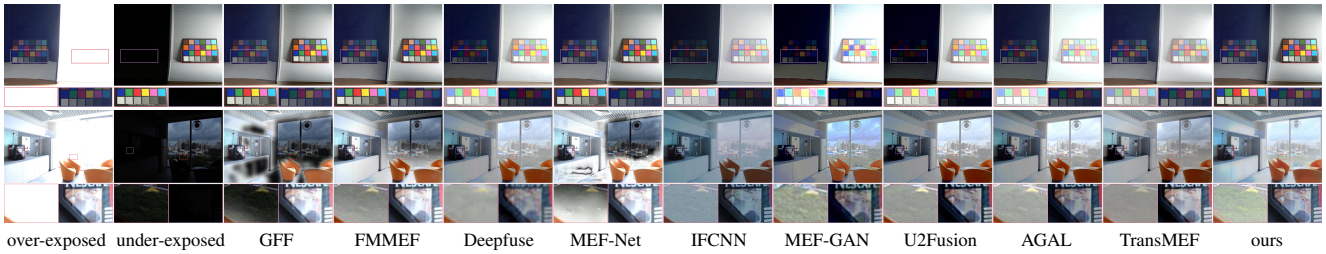

Figure 6: Qualitative comparison results on indoor multi-exposure image pairs.

results in colors with low saturation, represented by Deepfuse, IFCNN, U2Fusion, and TransMEF. Compared with all competitors in general, our results exhibit the superior colors with the best visual effect.

Quantitative Comparison For objective evaluation, we take peak signal-to-noise ratio (PSNR), MEF-SSIM (Ma, Zeng, and Wang 2015), edge intensity (EI) (Rajalingam and Priya 2018), correlation coefficient (CC) (Shah, Merchant, and Desai 2013) and DeltaE for evaluation. PSNR represents the distortion in the fusion process. MEF-SSIM and CC focus on the structural similarity and linear correlation between the fused image and source images, respectively. EI measures the sharpness of edges in the fused image. DeltaE evaluates the color differences between the fused image and GT. The quantitative results tested on the test set from SICE dataset are reported in Tab. 1. Our MEF-CL is superior in MEF-SSIM, CC and DeltaE. It shows that our fusion results are closest to source images in structure and correlation with the minimal color differences. For EI and PSNR, our MEF-CL achieves the suboptimal performance. It proves that our algorithm can generate fusion results with richer edge information and produce less distortion.

Efficiency Comparison We compare the efficiency on the 120 test image pairs. The runtime is shown in Tab. 1. Traditional methods are tested on a laptop with 3.2 GHz AMD Ryzen 7 5800H CPU. Due to the small model size, our MEF-CL achieves comparable runtime efficiency, only legging behind MEF-Net by a narrow margin.

Validation on Breaking Exposure Limits Due to the high dynamic range of the real scene, the exposure ratios between image sequences sometimes varies greatly. In unsupervised deep learning methods, because they often use weights or gradients as loss functions to constrain the fusion. The common method is to fuse an over-exposed image with an under-exposed image, which limits the range of fusion. Through the strong deep-level feature guidance of contrastive learning, our MEF-CL can enhance the dark areas in the under-exposed image and correct the bright areas in the over-exposed image. In other words, we break through the exposure limitation of unsupervised deep learning methods. As shown in Fig. 7, our results are more excellent in the over-exposed image fusion and under-exposed image fusion.

Ablation Study

Ablation Study of Loss Functions In this section, we perform the ablation experiment to verify effectiveness of the contrastive loss $L_{InfoNCE}$. In addition, to improve the quality of fused images, we use L_{MSE} and L_{SSIM} for similarity constraints from two aspects, *i.e.*, intensity distribution and structural similarity. To verify the effectiveness of these losses, we remove them respectively and retrain the network. The other settings are the same as those of MEF-CL.

The qualitative results are shown in Fig. 8. By comparing (e) and (f), it shows that $L_{InfoNCE}$ plays a major role in enhancing the details and improving colors. Comparing (c) and (f) or comparing (d) and (f), it can be seen that the absence of L_{MSE} results in large differences in the illumination, causing unreasonable intensity distribution. Without L_{SSIM}, the image is blurred and lacks texture details. Therefore, L_{SSIM} and L_{MSE} complement each other. In addition, quantitative experiments are also conducted. The results are shown in Tab. 2. It proves the effectiveness of the contrastive loss and the advantages of contrastive learning.

Ablation Study of Blocks The contrastive learning block includes three blocks: feature extraction, attribute, and content. Attribute and content blocks complement each other. In the above section, by ablating $L_{InfoNCE}$, we verified the role of the contrastive learning block in improving the details and enhancing the colors. To verify the functions of shallow-level and deep-level features, we perform ablation

	MEF-SSIM	CC	PSNR	EI	DeltaE	Runtime (second)
GFF	0.826 ± 0.01	0.567 ± 0.04	58.035 ± 0.65	68.535 ± 0.64	17.045 ± 0.74	1.356 ± 0.03
FMMEF	0.897 ± 0.01	0.731 ± 0.03	58.426 ± 0.79	69.042 ± 0.71	19.490 ± 0.46	1.196 ± 0.02
Deepfuse	0.890 ± 0.01	0.944 ± 0.02	**58.687 ± 0.71**	58.324 ± 0.60	13.790 ± 0.52	0.182 ± 0.01
MEF-Net	0.905 ± 0.01	0.710 ± 0.04	58.166 ± 0.74	74.351 ± 0.69	15.470 ± 0.62	**0.023 ± 0.01**
IFCNN	0.901 ± 0.01	0.926 ± 0.03	58.414 ± 0.64	**98.784 ± 0.75**	16.293 ± 0.71	0.038 ± 0.01
MEF-GAN	0.872 ± 0.01	0.928 ± 0.03	58.394 ± 0.66	62.584 ± 0.66	14.467 ± 0.66	1.002 ± 0.03
U2Fusion	0.835 ± 0.01	0.937 ± 0.03	58.586 ± 0.70	75.296 ± 0.64	18.807 ± 0.73	0.951 ± 0.05
AGAL	0.902 ± 0.01	0.933 ± 0.03	58.487 ± 0.67	73.578 ± 0.72	14.900 ± 0.38	0.044 ± 0.01
TransMEF	0.907 ± 0.01	0.937 ± 0.04	58.602 ± 0.75	75.349 ± 0.65	13.922 ± 0.64	0.051 ± 0.02
MEF-CL (ours)	**0.911 ± 0.01**	**0.946 ± 0.03**	58.604 ± 0.73	80.131 ± 0.68	**13.572 ± 0.42**	0.033 ± 0.01

Table 1: Quantitative comparison results of different MEF methods on the SICE dataset (mean and standard deviation of four metrics are shown; bold: optimal, underline: suboptimal).

Figure 7: Fusion results of different MEF methods for fusing two over-exposed images and two under-exposed images.

Figure 8: Ablation study of loss functions.

	MEF-SSIM	CC	PSNR	EI
w/o L_{MSE}	0.889	0.920	58.527	76.873
w/o L_{SSIM}	0.907	0.934	58.495	79.845
w/o $L_{InfoNCE}$	0.875	0.914	58.443	72.966
w/o attribute block	0.876	0.902	58.455	74.465
w/o content block	0.899	0.937	58.583	78.523
MEF-CL	**0.911**	**0.946**	**58.604**	**80.131**

Table 2: Quantitative comparison of ablation study with mean and standard deviation (bold: optimal).

experiments on the attribute block and content block, respectively. The qualitative results are shown in Fig. 9. Comparing Figs. 9(c) and (e), when the attribute block is removed, the fused image becomes blurred and lacks texture details. Comparing Figs. 9(d) and (e), it can be found that although the texture details have been improved, local areas suffer from color distortion. By combining the attribute and content blocks, the fused image obtained by the complete network (Fig. 9(e)) makes up for the previous defects, with rich texture details and good visual perception. In addition, we

Figure 9: Ablation study of blocks.

also conduct quantitative experiments and results are shown in Tab. 2. The method without attribute block ranks the second. By combining them, MEF-CL shows the best results on all the metrics. It shows that the combination of attribute and content blocks can achieve higher similarity with source images and exhibit higher contrast.

Conclusion

In this paper, we propose an unsupervised multi-exposure image fusion network based on contrastive learning, termed as MEF-CL. In our method, we fully extract the shallow-level and deep-level features in the source images. Then, the network uses contrastive learning to guide the fused image to learn the characteristics of normal-light images, so that the fused image can independently learn the texture and structural features in the source image, so as to achieve generality in different fused scenes. Also because of contrastive learning, our method can be extended to image pairs of any different exposures. Compared with nine state-of-the-art multi-exposure image fusion methods, our method can achieve advanced performance both qualitatively and quantitatively.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62276192), and the Key Research and Development Program of Hubei Province (2020BAB113).

References

Afifi, M.; Derpanis, K. G.; Ommer, B.; and Brown, M. S. 2021. Learning multi-scale photo exposure correction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 9157–9167.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.*, 27(4): 2049–2062.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, 1597–1607.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 9729–9738.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hu, Q.; Wang, X.; Hu, W.; and Qi, G.-J. 2021. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1074–1083.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4700–4708.

Kou, F.; Wei, Z.; Chen, W.; Wu, X.; Wen, C.; and Li, Z. 2017. Intelligent detail enhancement for exposure fusion. *IEEE Transactions on Multimedia*, 20(2): 484–495.

Li, H.; Ma, K.; Yong, H.; and Zhang, L. 2020. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Trans. Image Process.*, 29: 5805–5816.

Li, S.; Kang, X.; and Hu, J. 2013. Image fusion with guided filtering. *IEEE Trans. Image Process.*, 22(7): 2864–2875.

Liu, J.; Shang, J.; Liu, R.; and Fan, X. 2022. Attention-guided Global-local Adversarial Learning for Detail-preserving Multi-exposure Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.*, 32(8): 5026–5040.

Ma, K.; Duanmu, Z.; Zhu, H.; Fang, Y.; and Wang, Z. 2019. Deep guided learning for fast multi-exposure image fusion. *IEEE Trans. Image Process.*, 29: 2808–2819.

Ma, K.; Zeng, K.; and Wang, Z. 2015. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.*, 24(11): 3345–3356.

Mertens, T.; Kautz, J.; and Van Reeth, F. 2009. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Computer Graphics Forum*, 28(1): 161–171.

Qu, L.; Liu, S.; Wang, M.; and Song, Z. 2022. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proc. AAAI Conf. Artif. Intell.*, 2126–2134.

Rajalingam, B.; and Priya, R. 2018. Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis. *International Journal of Engineering Science Invention*, 2: 52–60.

Ram Prabhakar, K.; Sai Srikar, V.; and Venkatesh Babu, R. 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4714–4722.

Sakai, T.; Kimura, D.; Yoshida, T.; and Iwahashi, M. 2015. Hybrid method for multi-exposure image fusion based on weighted mean and sparse representation. In *Proceedings of the European Signal Processing Conference*, 809–813.

Shah, P.; Merchant, S. N.; and Desai, U. B. 2013. Multifocus and multispectral image fusion based on pixel significance using multiresolution decomposition. *Signal, Image and Video Processing*, 7(1): 95–109.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, J.; Liu, H.; and He, N. 2014. Exposure fusion based on sparse representation using approximate K-SVD. *Neurocomputing*, 135: 145–154.

Xing, L.; Cai, L.; Zeng, H.; Chen, J.; Zhu, J.; and Hou, J. 2018. A multi-scale contrast-based image quality assessment model for multi-exposure image fusion. *Signal Processing*, 145: 233–240.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2022. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1): 502–518.

Xu, H.; Ma, J.; and Zhang, X.-P. 2020. MEF-GAN: Multi-exposure image fusion via generative adversarial networks. *IEEE Trans. Image Process.*, 29: 7203–7216.

Xu, H.; Zhang, H.; and Ma, J. 2021. Classification saliency-based rule for visible and infrared image fusion. *IEEE Trans. Comput. Imaging*, 7: 824–836.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10): 2761–2785.

Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; and Ma, J. 2020a. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proc. AAAI Conf. Artif. Intell.*, 12797–12804.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020b. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.