mathematics

MDPI

Article

Multi-Task Learning Approach Using Dynamic Hyperparameter for Multi-Exposure Fusion

Chan-Gi Im, Dong-Min Son, Hyuk-Ju Kwon and Sung-Hak Lee *

School of Electronic and Electrical Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daeg 702-701, Republic of Korea; imchmcgi2@knu.ac.kr (C.-G.I.); forhollow@knu.ac.kr (D.-M.S.); olin1223@knu.ac.kr (H.-J.K.)

* Correspondence: shak2@ee.knu.ac.kr; Tel.: +82-53-950-7216

Abstract: High-dynamic-range (HDR) image synthesis is a technology developed to accurately reproduce the actual scene of an image on a display by extending the dynamic range of an image. Multi-exposure fusion (MEF) technology, which synthesizes multiple low-dynamic-range (LDR) images to create an HDR image, has been developed in various ways including pixel-based, patch-based, and deep learning-based methods. Recently, methods to improve the synthesis quality of images using deep-learning-based algorithms have mainly been studied in the field of MEF. Despite the various advantages of deep learning, deep-learning-based methods have a problem in that numerous multi-exposed and ground-truth images are required for training. In this study, we propose a self-supervised learning method that generates and learns reference images based on input images during the training process. In addition, we propose a method to train a deep learning model for an MEF with multiple tasks using dynamic hyperparameters on the loss functions. It enables effective network optimization across multiple tasks and high-quality image synthesis while preserving a simple network architecture. Our learning method applied to the deep learning model shows superior synthesis results compared to other existing deep-learning-based image synthesis algorithms.

Keywords: high dynamic range; multi exposure fusion; image fusion; deep learning

MSC: 68T45

Citation: Im, C.-G.; Son, D.-M.; Kwon, H.-J.; Lee, S.-H. Multi-Task Learning Approach Using Dynamic Hyperparameter for Multi-Exposure Fusion. *Mathematics* **2023**, *11*, 1620. https://doi.org/10.3390/math11071620

Academic Editors: Vladimir V. Arlazarov and Konstantin Bulatov

Received: 9 March 2023
Revised: 24 March 2023
Accepted: 25 March 2023
Published: 27 March 2023

1. Introduction

In general, the dynamic range of the human visual system is approximately 10,000 nits, whereas the dynamic range of existing display devices has a range of approximately 100 nits; therefore, there is a limit to realizing realistic images on a display. Moreover, low-dynamic-range (LDR) images cannot adequately depict a real scene because of the limited response to the high-dynamic-range (HDR) of natural light. To display HDR images on standard dynamic range (SDR) devices, such as computer monitors and televisions, two common techniques are used: tone-mapping operators (TMOs) and multi-exposure fusion (MEF).

TMOs compress the dynamic range of an HDR image into several visible LDR images, allowing HDR scenes to be rendered on an LDR display. There are two types of TMOs: global and local. Global operators use identical nonlinear functions to compress the entire image, making operators fast to compute. Reinhard et al. [1] developed an automated method for mapping HDR world images to LDR images based on a zone system. They used a local averaging logarithmic operator to map the entire image and implemented an automated dodging and burning algorithm to correct bright and dark regions. Duan et al. [2] proposed a histogram adjustment method for displaying HDR images, based on a global TMO, called histogram-adjustment-based linear to equalized quantizer (HALEQ).

However, TMOs can cause a loss of local contrast and detail because they treat the entire image as a whole without considering differences in brightness and contrast within

the image [3]. Local operators can reproduce a better quality of the tone-processed image; however, these algorithms are more complicated than global operators, and the results can be unrealistic because of artifacts.

The MEF technique was developed to create HDR images from multiple LDR images with varying exposures and to realize digital images that resemble real eyes [4]. The MEF technique can implement an image with a wider visual brightness range by appropriately synthesizing several LDR images, and thus, realistic images can be implemented on the display device [5,6]. As shown in Figure 1, unlike under- or over-exposed images that may have saturated areas, the MEF image can depict finer details by combining information from multiple source images.

(**a**) (**b**) (**c**)

Figure 1. Multiple exposure source images and fusion image: (**a**) under-exposed image, (**b**) over-exposed image, and (**c**) fused image by DenseFuse [7].

MEF techniques can be divided into traditional and deep-learning-based methods. Traditional methods include the pixel-based MEF method and the patch-based method. While researchers have developed various traditional methods, with the rise of deep learning, deep-learning-based methods have become mainstream. Deep-learning-based methods have several advantages, such as flexibility and scalability. However, convolutional neural networks (CNNs) have inherent problems in that it is difficult to preserve extracted features due to the defection in establishing long-range dependencies [8], and large amounts of datasets are required to train networks. Moreover, owing to the lack of sufficient multi-exposure training data and ground-truth images for HDR, many unsupervised MEF methods have been proposed.

In this study, we present a method for training MEF networks by setting up multiple tasks based on our image features. Although it is difficult to train due to the difficulty of preserving extracted intricate features with CNNs architectures, it can be overcome by learning several reference images with well-revealed features. These reference images are composed of features that provide useful information for fusion by reflecting the characteristics of the multi-exposure images. Moreover, by applying dynamic hyperparameters to each multitask loss function, the MEF networks can be trained to effectively reproduce HDR images. The contributions of our study are summarized as follows:

- To train MEF networks that require learning, we present a method for setting up a customized dataset;
- To generate multiple tasks based on source features, we perform a procedure that filters unnecessary regions from the source images using multi-exposure image characteristics;
- To reflect the information between multiple tasks, we set dynamic hyperparameters on the loss functions. This helps the network reproduce better-contrast images;
- To produce a high-quality image with a simple network design, we prove that it is possible to utilize multiple tasks and dynamic hyperparameters for the loss function.

The remainder of this paper is organized as follows: Section 2 introduces the existing methods used in our work. Section 3 describes the implementation of the proposed method. Section 4 provides the quantitative and qualitative results of the proposed model as well as other deep-learning-based image fusion algorithms. Finally, we present our conclusions in Section 5.

2. Related Works

2.1. Multi-Exposure Image Fusion

Among the traditional image synthesis methods, there are two major techniques: pixel- and patch-based methods. The pixel-based method synthesizes multi-exposed images according to specific pixel-wise fusion rules. This method selects the best pixels from each source image based on certain criteria and then generates weight maps to combine these pixels. The acquisition of pixel weight maps from source images is typically based on measures of image quality such as sharpness, contrast, and color fidelity. In addition, the weight maps of source images can be obtained in various ways.

Bruce [9] calculated entropy in a circular area centered on each pixel with a radius from the source image. They assigned a weight to each pixel based on its information entropy and merged source images based on these weights. Song et al. [10] proposed an HDR fusion method using a probabilistic model that preserved the calculated image luminance levels and suppressed reversals in the image luminance gradients. Lee et al. [11] calculated the weights for each pixel in the source images by analyzing the relative pixel intensity and global gradient. The relative pixel intensity determines the brightness of each pixel in the image, whereas the global gradient measures the contrast between the adjacent pixels. While these methods have the advantage of obtaining fused images by calculating accurate pixel weight maps, the main drawback of pixel-based methods is the tendency to introduce artifacts in the final fused image, such as noise and halos [12]. Moreover, they cannot properly represent the edge information of images [13].

In the patch-based method, the source images are first divided into small patches or sub-images. Each patch is then processed independently to produce a fused image. This approach allows for better handling of the local differences in brightness and contrast.

Huang et al. [14] introduced a patch-based MEF method that divides an image patch into three independent parts: contrast extraction, structural preservation, and intensity adjustment. They fused these parts separately and reconstructed the desired patch, which was then fed back into the fused image. Wang et al. [15] proposed a method that uses a super-pixel segmentation approach to divide the input images into non-overlapping patches composed of pixels with similar visual properties. This method has the advantages of avoiding the blocking effect and preserving the color attributes of source images. The main advantage of these methods is that the weight map has less noise than pixel-based MEF methods because it combines the neighborhood information of the pixels. However, the performance of patch-based methods is sensitive to the choice of patch size and overlap as well as the characteristics of the source images, making it challenging to achieve consistent and reliable results across different images and applications.

2.2. Deep-Learning-Based Multi-Exposure Image Fusion

As shown in Figure 2, deep-learning-based image fusion networks typically consist of three characteristic stages: feature extraction, feature fusion, and feature reconstruction. Feature extraction involves the extraction of high-level features from source images using deep neural networks. These features represent the source images in high-dimensional space and are used to capture the complex and abstract features of the images. The feature-fusion stage involves combining the extracted features from each source image to obtain a fused feature map. This can be performed in different ways such as using element-wise summation, l_1-norm, or other fusion strategies [7]. Finally, in the feature-reconstruction stage, the fused feature maps are fed into a decoder network to reconstruct the final fused image.

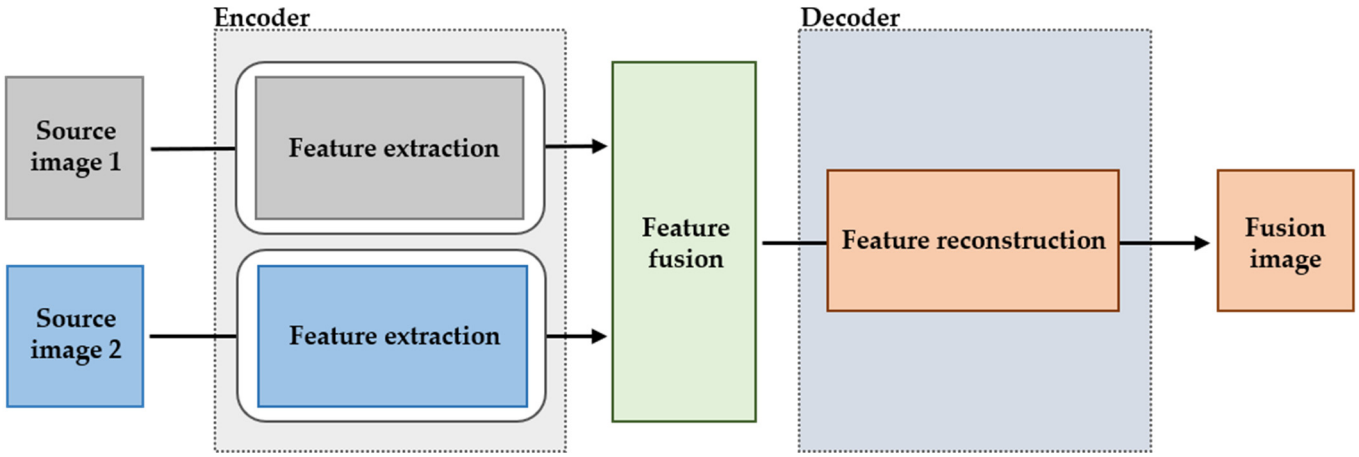

Figure 2. Traditional deep-learning-based image fusion structure.

The main advantage of deep-learning-based image fusion networks is that they can learn the optimal feature representations for image fusion directly from the data without the need for handcrafted features or explicit rules. Additionally, deep-learning-based approaches can be trained end-to-end, which enables the joint optimization of all stages of the image fusion process. However, the performance of deep-learning-based image fusion networks depends on the training dataset's quality and size and the network architecture's complexity.

In the MEF field, deep-learning-based image fusion methods are divided into two categories: supervised learning and unsupervised learning. In supervised learning, a large number of ground truth images are required for training even though the MEF ground truth is insufficient. To overcome the lack of dataset images, the generation of the ground truth was studied together. Kalantari et al. [16] combined three exposed LDR images to generate a ground-truth HDR image. Subsequently, they trained a deep CNN model to reproduce the HDR image from a set of images aligned with the optical flow. However, the generated ground truth images are not real, and their use may reduce the fusion performance. Therefore, many researchers have attempted to train networks in an unsupervised manner.

Unsupervised learning MEF methods usually modify network structures to extract informative features from source images. For example, Xu et al. [17] proposed U2Fusion, which extracts features with a pre-trained VGGNet-16 and utilizes DenseNet for fusion. The authors obtained various shapes of feature maps using a convolutional layer before the max-pooling layer. This could preserve deep-level features such as spatial structures as well as shallow features such as textures and details.

Qu et al. [8] introduced a transformer-based MEF framework called the TransMEF. They designed a transformer module in addition to a CNN module for training to address the inherent limitations of CNNs. By combining the CNN and transformer modules, the network can extract both local and global information from a pair of source images. Although these models have demonstrated good fusion performance and the ability to produce high-quality fused images, training their complex network architecture still requires numerous datasets. This can lead to slow training and increased memory use. Hence, finding the correct balance between network complexity and performance is crucial for developing effective fusion networks.

3. Proposed Methods

3.1. Framework Overview

As shown in Figure 3a, our network is based on the DenseFuse [7] architecture. We focused on optimizing the network by setting up multiple informative target images.

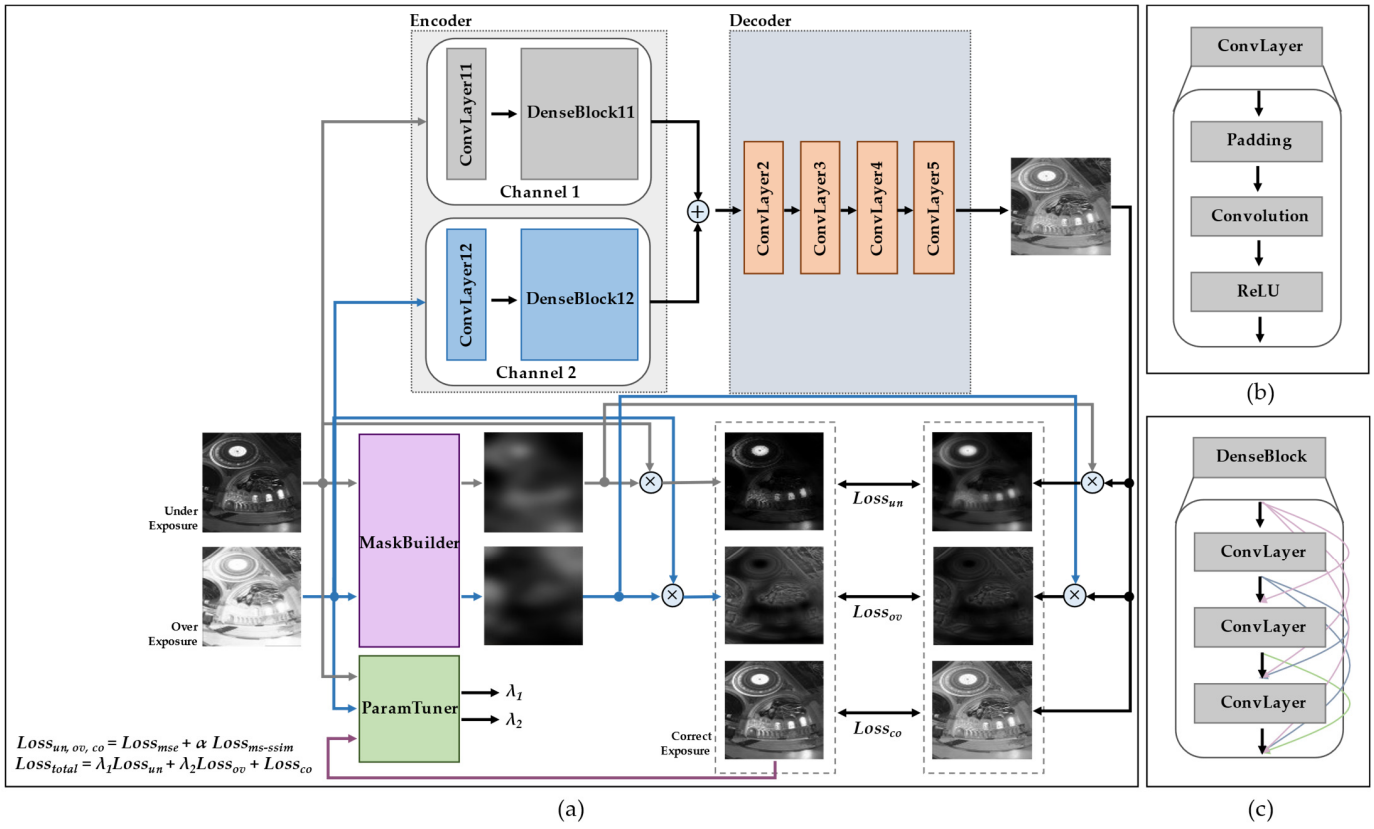

Figure 3. Framework of the proposed model: (**a**) the proposed training network with multi-task images; (**b**) detailed structure of ConvLayer; (**c**) detailed structure of DenseBlock.

During the training phase, the classified multi-exposed images were resized to 256 × 256 pixels, and only the luminance images of the LDR images entered each channel in the encoder. At the same time, new target images that contain useful information of the input images are generated using the MaskBuilder block. In addition, ParamTuner creates weight values that are multiplied by $Loss_{un}$ and $Loss_{ov}$ to optimize the network effectively.

As observed in Figure 3a, the encoder consists of two channels. The DenseBlock contains three ConvLayer blocks, in which the output of each layer is connected to the other layers, and each ConvLayer of the network contains one reflection padding layer, namely a convolutional layer with a kernel size of 3 × 3, and a ReLU activation function layer, as shown in Figure 3a,b. This structure preserves the information of the source images as much as possible and allows the network to be trained easily. The features of the under-/over-exposed images are extracted from the encoder block through the above convolution layers, and they are simply added. The fused feature maps enter the decoder, and the fused image is finally reconstructed using four ConvLayer blocks. The reconstructed image learns the correct-exposed image of the pair. In addition, we applied masks generated from MaskBuilder to the reconstructed image to perform multi-task self-supervised learning.

3.2. Dataset Acquisition

Networks that synthesize multi-exposure images can be trained in a supervised manner using ground truth images or in an unsupervised manner using similarity metric-based loss functions to retain the features of the source images [18]. Recently, owing to the lack of sufficient ground truth images in the MEF field, many researchers have investigated the unsupervised reconstruction of high-quality images [8]. To reconstruct higher-quality fusion images without reference images, it is important to learn the various features of each multi-exposure image.

We categorized multi-exposure images into three groups, i.e., under-exposure, over-exposure, and correct exposure, and paired these three types of exposure images. As shown in Figure 4, to acquire large amounts of exposure pairs, we first fixed the appropriate correct-exposed image from the correct exposure group; second, we set the image that was captured with less exposure time than the correct-exposed image as an under-exposed image, and the others that were captured with a higher exposure time than the correct-exposed image as an over-exposed image. Because the correct-exposed image contains some information about the under-exposed and over-exposed images as well as its own information, we initially aimed to converge a reconstructed image of LDR images into a correct-exposed image.

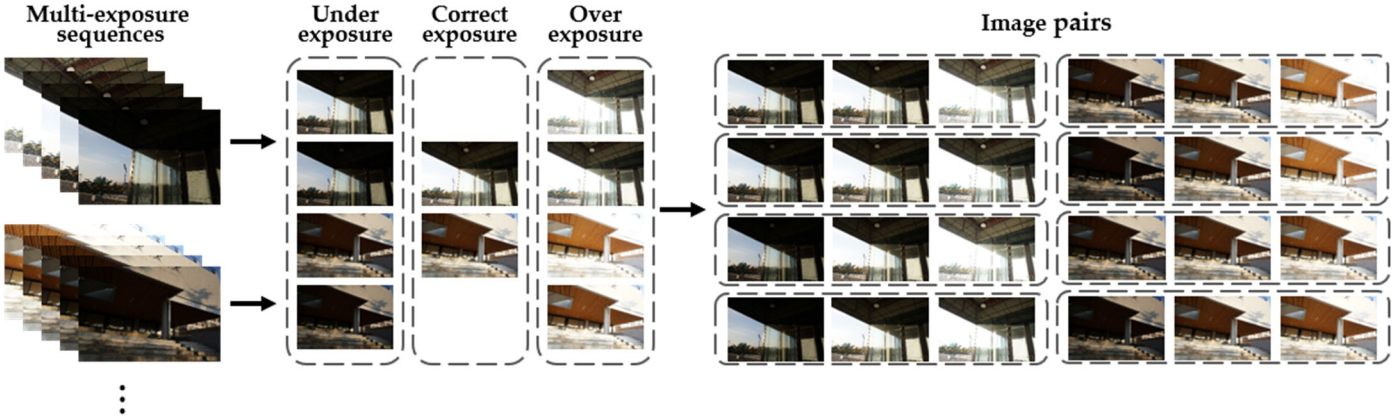

Figure 4. Training image pairs acquisition from multi-exposure datasets.

3.3. MaskBuilder

MaskBuilder is designed to create multi-task images because only the correct-exposed image is not sufficient to learn sufficient multi-exposure image features. The under-exposed image contains meaningful information in a relatively bright area, as the dark portion of the image is saturated; conversely, the over-exposed image contains useful information on a relatively dark area, as the bright portion of the image is saturated, as shown in Figure 5. Based on this characteristic, we devised masks that filter unnecessary information from the input LDR images.

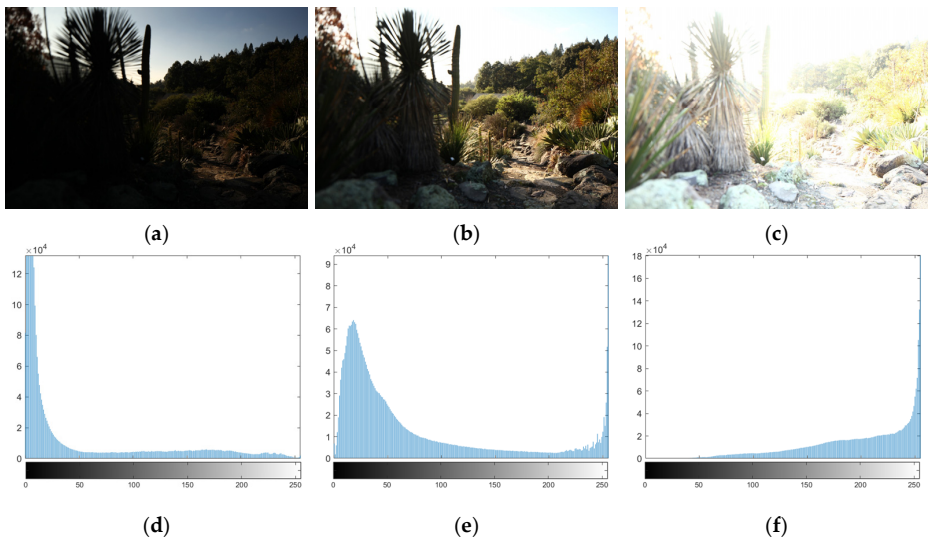

Figure 5. Multi-exposure images and histograms: (**a**) under-exposure image; (**b**) correct-exposure image; (**c**) over-exposure image; (**d**) histogram of under-exposure image; (**e**) histogram of correct-exposure image; (**f**) histogram of over-exposure image.

To determine whether the regions of the inputs are instructive, we blurred the images of each input image as follows:

$$G(x,\ y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{1}$$

$$I_g(i,\ j) = \sum_{y=-m}^{m}\sum_{x=-m}^{m} I(i+x,\ j+y)G(x,\ y) \tag{2}$$

where $G(\cdot)$ represents a two-dimensional Gaussian function, $I(i,j)$ represents the pixel value of the input image at $(i,\ j)$, σ is the standard deviation, m represents the kernel size of the Gaussian filter, and I_g is the blurred image. The higher the σ values that blur over a wider radius, the greater the amount of regions of the input images that can be removed. Thus, it is required to find an appropriate σ value. In Section 4, we arbitrarily set the σ value to 10, 20, and 50 and compared the results to determine the most suitable σ value. Lastly, the kernel size was conventionally set using Equation (3),

$$m = 6\sigma + 1 \tag{3}$$

Since the final mask should remove saturated areas from incorrect-exposed images, it should have a value of 0 for saturated areas and a value of 1 for unsaturated areas. Thus, min-max normalization was utilized to generate the final mask $\hat{I}_g$ using Equation (4),

$$\hat{I}_g(i,\ j) = \frac{I_g(i,\ j) - min_{I_g}}{max_{I_g} - min_{I_g}} \tag{4}$$

where max_{I_g} and min_{I_g} are the maximum value of I_g and the minimum values of I_g, respectively.

The generated mask eliminates dark regions of the inputs or outputs through multiplication. However, in the case of over-exposed images, the mask eliminates bright regions of the images because there is useful information in the dark region. Thus, we obtained the other filter mask $\hat{I}'_g$ for the over-exposed images using Equation (5).

$$\hat{I}'_g(i,\ j) = 1 - \frac{I'_g(i,\ j) - min_{I'_g}}{max_{I'_g} - min_{I'_g}} \tag{5}$$

where I'_g indicates the blurred over-exposed image in Equation (5), and $max_{I'_g}$ and $min_{I'_g}$ indicate the maximum value of I'_g and the minimum values of I'_g, respectively. A diagram of the mask building is shown in Figure 6.

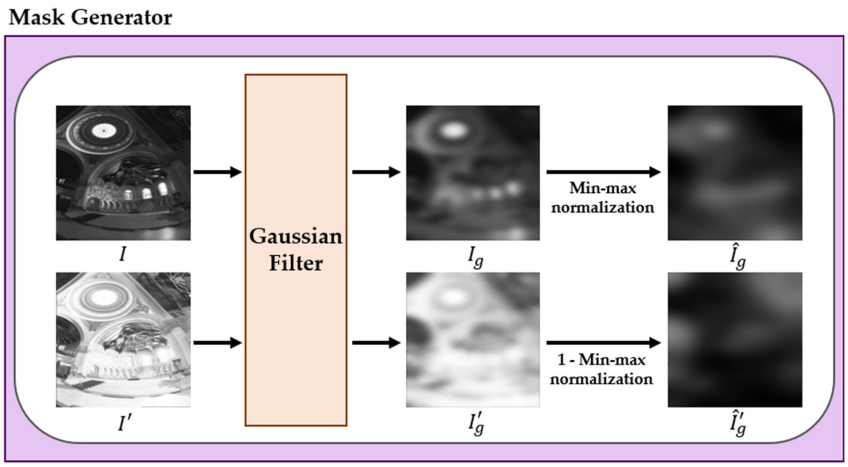

Figure 6. Detailed structure of the proposed MaskBuilder.

For the generated masks $\hat{I}_g$ and $\hat{I}'_g$ in Figure 6, grayscale-normalized images were used because it is difficult to visually examine the original images.

3.4. Loss Function

The proposed model has three different types of target images. To learn each feature of the targets, multi-loss functions were designed, and $Loss_{total}$ is defined as follows:

$$Loss_{total} = \lambda_1 Loss_{un} + \lambda_2 Loss_{ov} + Loss_{co} \tag{6}$$

where $Loss_{un}$, $Loss_{ov}$, and $Loss_{co}$ are the losses between the different multiexposed images and outputs, and λ_1 and λ_2 are the hyperparameters that control the ratio of $Loss_{un}$ and $Loss_{ov}$, respectively.

Each loss consists of two loss functions that can preserve the salient information of the target image as follows:

$$Loss = Loss_{mse} + \alpha Loss_{ms-ssim} \tag{7}$$

$$Loss_{mse} = \|I_{out} - I_{target}\|_2 \tag{8}$$

$$Loss_{ssim} = 1 - MS\text{-}SSIM(I_{out},\ I_{target}) \tag{9}$$

where $Loss_{mse}$ denotes the mean square error (MSE) loss, $Loss_{ms-ssim}$ denotes the multi-scale structural similarity (MS-SSIM) loss [19], $MS\text{-}SSIM(\cdot)$ denotes the MS-SSIM operator between the output and target, and α is a hyperparameter used to match the magnitude of the MSE and MS-SSIM losses. The α was experimentally set to 1000 according to [7].

3.5. ParamTuner

Although we classified multi-exposure datasets into under-exposed, over-exposed, and correct-exposed images, we realized that the exposure value (EV) difference between the under-exposed and correct-exposed images and the difference between the over-exposed and correct-exposed images are not consistent. This is because LDR images have different exposure times even though they belong to the same category. Thus, we assumed that it is necessary to set the dynamic λ_1 and λ_2 values for each multi-exposed image pair. For example, let us assume that five multi-exposed images, which consist of 0 EV, $\pm$1 EV, and $\pm$2 EV, can be paired, as shown in Figure 7.

There is an identical EV difference between the correct-exposed and incorrect-exposed images in Figure 7a. However, in Figure 7b,c, the difference between the correct-exposed and under-exposed images is larger than that between the correct-exposed and over-exposed images or vice versa. We expected that the larger EV difference between the correct-exposed and incorrect-exposed images means that the incorrect-exposed image has more significant information than the other incorrect-exposed image; thus, the network can be trained more effectively by setting a higher lambda value on the incorrect-exposed image loss.

However, because of the challenge of measuring the relative exposure value difference between the correct-exposed image and the under-exposed or over-exposed image, we employed an adaptive gamma value inspired by adaptive gamma correction (AGC) [20]. The AGC method can adequately evaluate the gamma value of an image by classifying the image's contrast as follows:

$$g(I) = \begin{cases} \varrho_1, & D \leq 1/\tau \\ \varrho_2, & otherwise \end{cases} \tag{10}$$

where I is an input image, and ϱ_1 and ϱ_2 are low-contrast class and high- (or moderate-) contrast class, respectively. D is defined as $diff((\mu + 2\sigma), (\mu - 2\sigma))$, τ is a parameter used to define the contrast of an image, and σ and μ are the standard deviation and mean of the image intensity, respectively. The input image can be categorized differently based on the τ, and we determined τ = 3 to be the appropriate value by referring to [20].

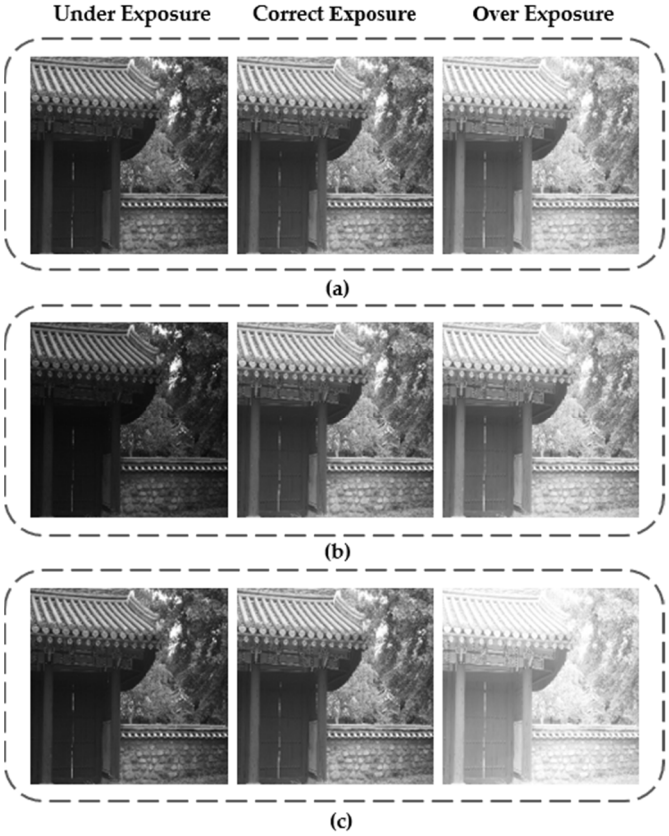

Figure 7. Image pairs by different multi-exposed images (left–right): (**a**) −1 EV, 0 EV, and +1 EV; (**b**) −2 EV, 0 EV, and +1 EV; (**c**) −1 EV, 0 EV, and +2 EV.

After classifying the constant class, the gamma value of the image was calculated according to its group. Equations (11) and (12) show γ of the image belonging to ϱ_1 group and belonging to ϱ_2, respectively.

$$\gamma = -\log_2(\sigma) \tag{11}$$

$$\gamma = exp[\frac{1-(\mu+\sigma)}{2}] \tag{12}$$

In ParamTuner, each adaptive gamma value of the multi-exposed images was calculated using the AGC algorithm. Next, the difference in gamma values between the correct-exposed image and the under-exposed or over-exposed image was computed separately. Finally, λ_1 and λ_2 were determined using the normalized ratio function. A diagram of this strategy is shown in Figure 8.

In Figure 8, I_{un}, I_{co}, and, I_{ov} are under-exposed, correct-exposed, and over-exposed images, respectively. γ_{ov}, γ_{co}, and γ_{ov} are the gamma values of the multi-exposed images, and D_1 and D_2 are defined as $diff(\gamma_{un}, \gamma_{co})$ and $diff(\gamma_{ov}, \gamma_{co})$, respectively. The final value of $\lambda_i (i = 1,\ 2)$ was calculated using Equation (13).

$$\lambda_i = \beta \frac{D_i}{\sum_{j=1}^{k} D_j} \tag{13}$$

where β is a hyperparameter that controls the ratio of each incorrect-exposed image, and $k = 2$ indicates the number of lambda. The value of β was set to two to account for the entire loss ratio.

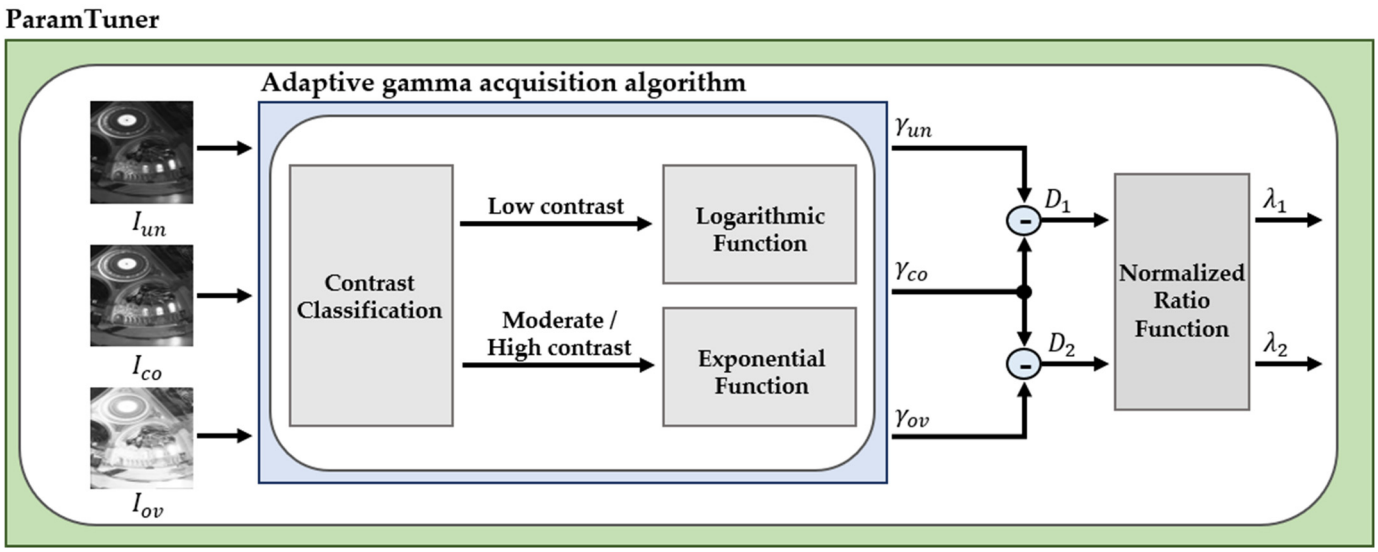

Figure 8. Detailed structure of the proposed ParamTuner.

3.6. Image Fusion Process

When two images with different exposures are captured with a single sensor camera, the boundaries of the two images may not match because of changes in the shooting conditions, such as movement of the subject, which may cause blending problems (e.g., ghost, double boundary, blurring, etc.). Therefore, we aligned two LDR images using scale-invariant feature transform (SIFT) [21] and homography [22] before image fusion. The SIFT algorithm is a widely adopted technique for feature extraction, and it can extract distinctive invariant features between two LDR images, which can be aligned by homography based on the extracted features.

In image processing, homography matching involves matching two images with different viewpoints to obtain the same viewpoint through homography conversion. Homography transformation refers to a series of processes for obtaining a transformation matrix that maps points on a two-dimensional plane to another plane in a three-dimensional space and transforms the points.

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \times \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{14}$$

Equation (14) shows the equation for homographic transformation that maps a point $(x, y, 1)$ in a three-dimensional space to $(x', y', 1)$. The h_{11} through h_{33} $3\times$ 3 matrices denote a homography matrix. To homographically match two images from different viewpoints, a process of finding keypoints of each image and matching these keypoints through a descriptor is required, which can be performed using the SIFT algorithm [21]. The SIFT algorithm mainly consists of four parts: scale-space extrema detection, keypoint localization, orientation assignment, and keypoints and descriptors [23]. After calculating each keypoint and descriptor in two different images, keypoint pairs with the highest similarity were obtained by feature mapping. The final aligned image was computed by a homography matrix using these keypoint pairs.

However, it is sometimes difficult to extract features between under-exposed and over-exposed images because of the dissimilarity in the exposure values. Thus, we applied AGC [20] to correct the image dissimilarity. A diagram of the proposed image-fusion scheme is shown in Figure 9.

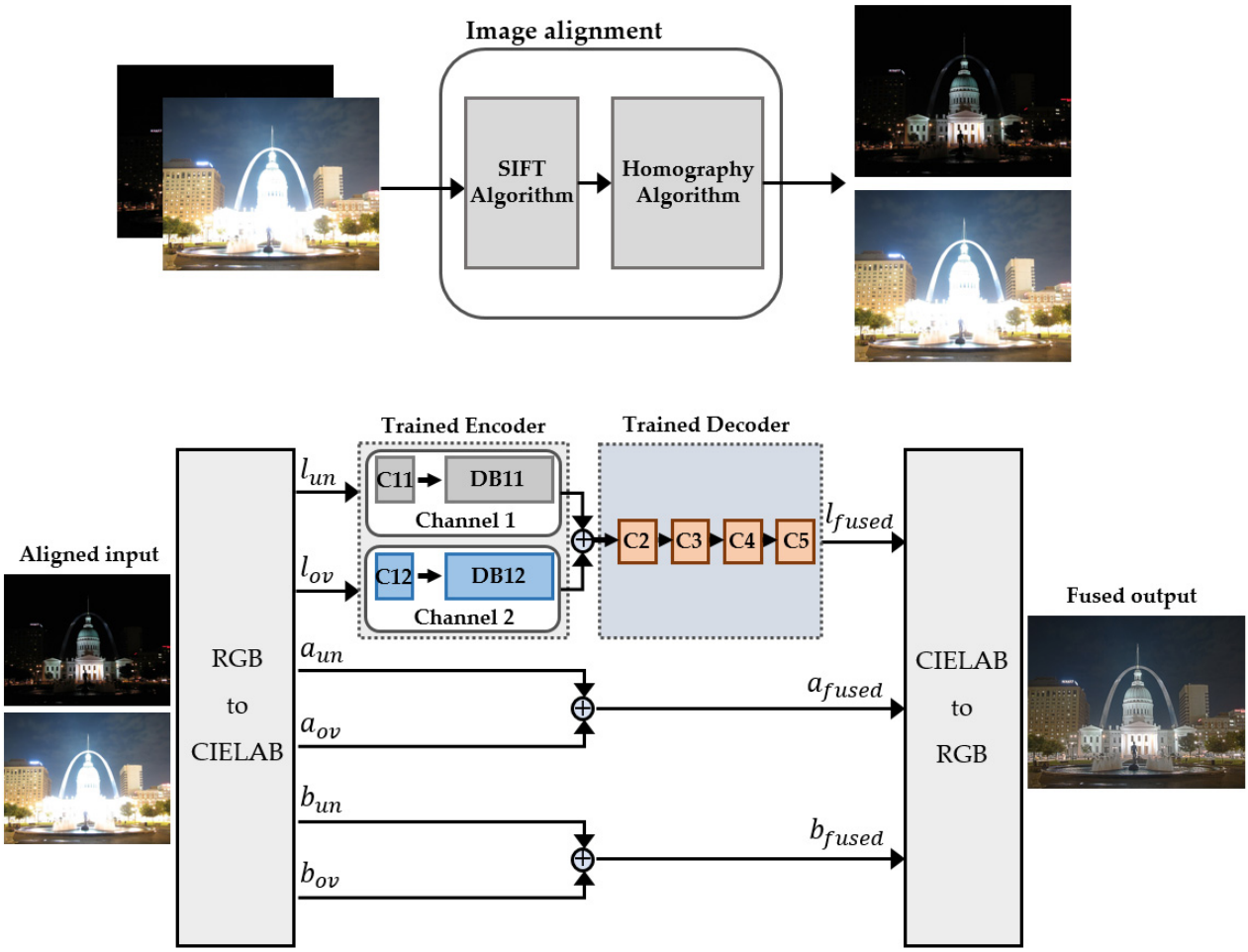

Figure 9. Image fusion scheme of the proposed model.

After aligning the LDR images, the color space of the images was converted from RGB to CIELAB. Because CIELAB has excellent color separation based on human vision, it preserves the color components of each input image. In Figure 9, l_{un} and l_{ov} denote the luminance channels of the under-exposed and over-exposed images, respectively. C11, C12, C2, C3, C4, and C5 are the trained ConvLayers. DB11 and DB12 are the trained DenseBlocks. l_{fused} is determined from l_{un} and l_{ov} by trained neural networks; a_{fused} and b_{fused} are simply calculated by adding half of a_{un} and a_{ov} or b_{un} and b_{ov}, which are the color channels of the under-exposed and over-exposed images. Finally, the fused image was obtained by converting the color space into an RGB space.

4. Experimental Results

In this section, we analyze our approach and compare it with other deep-learning-based image fusion methods. A total of 460 image pairs (under-/over-/correct-exposed images) from the open dataset [24,25] and our acquired dataset were used as the training datasets. All training images were resized to 256 × 256. The network was trained for 40 epochs with a batch size of 1 and a learning rate of 1×10^{-4}. The training networks and all experiments were implemented with an NVIDIA RTX 2060 GPU and Intel Core i5-6500 CPU.

To confirm the effect of the sigma value of the Gaussian filter in MaskBuilder, we compared each $Loss_{total}$ as shown in Figure 10. As observed, a larger value of sigma leads to a relatively unstable convergence loss, and the best loss is also the lowest when the sigma value is 10. Therefore, our method set the sigma value to 10.

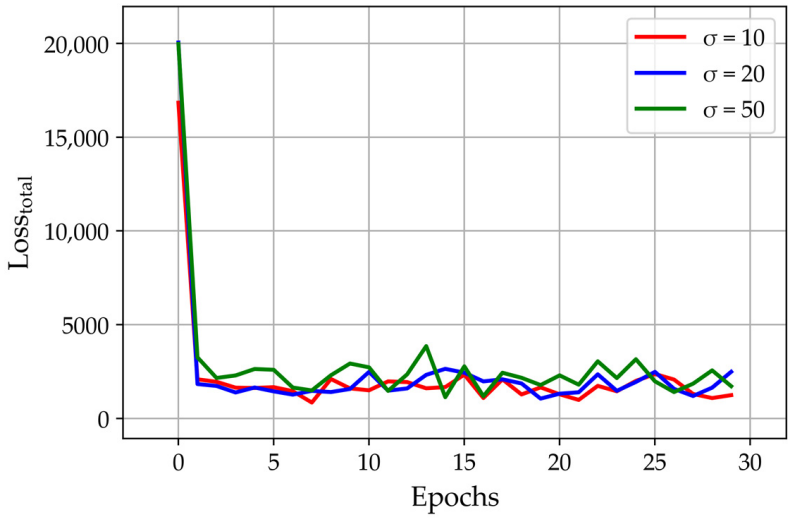

Figure 10. The graph plot of total loss per epoch according to sigma.

We compared the fusion results of our method with those of three deep-learning-based image fusion methods [7,8,17]. Fourteen under-/over-exposed image pairs were selected from the previous available dataset [26,27] for the experiment, and the fused images of each method were evaluated using 10 image-quality metrics: average gradient (AG), edge intensity (EI), feature mutual information (FMI), LPC-SI, S3, spatial frequency (SF), perceptual image quality evaluator (PIQE), MS SSIM, Qabf, and visual information fidelity for fusion (VIFF). The AG metric computes the gradient, which contains details and texture information in the image [28]. EI is a metric that represents the image quality and sharpness based on the edge intensity value [29]. FMI determines the mutual feature information between the source and fused images [30]. LPC-SI can detect the sharpness of visual images in the complex wavelet transform domain [31]. S3 shows how images have sharper areas considering human cognitive characteristics [32]. SF reflects the distribution of gradients to compute the amount of detail and texture in the fused image [33]. PIQE calculates image quality using a spatial mask, with a smaller score indicating better perceptual quality [34]. The MS SSIM is a structural similarity index measure combined with multiscale information of the source image [19]. The Qabf metric reflects the quality of visual information from source images, and a larger score indicates a better fusion effect [35]. The VIFF uses VIF models to assemble visual information from the source and fused images and measure the effective visual information of fusion [36].

As shown in Table 1, our method achieved the best performance and showed excellent results, particularly in terms of image sharpness and human perception-based quality metrics. The proposed method has an average improvement of about 32.8% over the existing synthesis methods in the image sharpness matrix (AG, EI, LPC-SI, S3, SF) and an average improvement of about 27.35% in the image-quality evaluation metrics (PIQE, Qabf, VIFF).

Table 1. Comparison of image-quality metrics score with existing image fusion methods: HyperP_MB denotes a deep learning model that applies our training methods.

Method	AG	EI	FMI	LPC-SI	S3	SF	PIQE	MS-SSIM	Qabf	VIFF
DenseFuse	4.1796	41.3714	0.3366	0.9369	0.1830	13.7378	26.0896	0.8781	0.4208	0.4285
U2Fusion	4.7719	48.9770	0.2891	0.9438	0.1798	15.3116	35.4318	0.8468	0.3898	0.4181
TransMEF	3.9983	40.3294	0.2910	0.9360	0.1627	13.1360	37.8373	0.8636	0.3520	0.4084
HyperP_MB	6.1319	61.6277	0.3447	0.9543	0.2369	19.7974	25.8651	0.8916	0.5059	0.5477

Figures 11–16 show the results of each fusion method. Overall, the proposed method has better contrast and clarity than other methods. Whereas our method has detailed features from source images in the highlighted region, the other fusion methods have low

distinctness. Because our method preserves useful information of under-/over-exposed images, the boundaries of the object and texture are distinguishable.

Figure 11. Input and result images (1): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

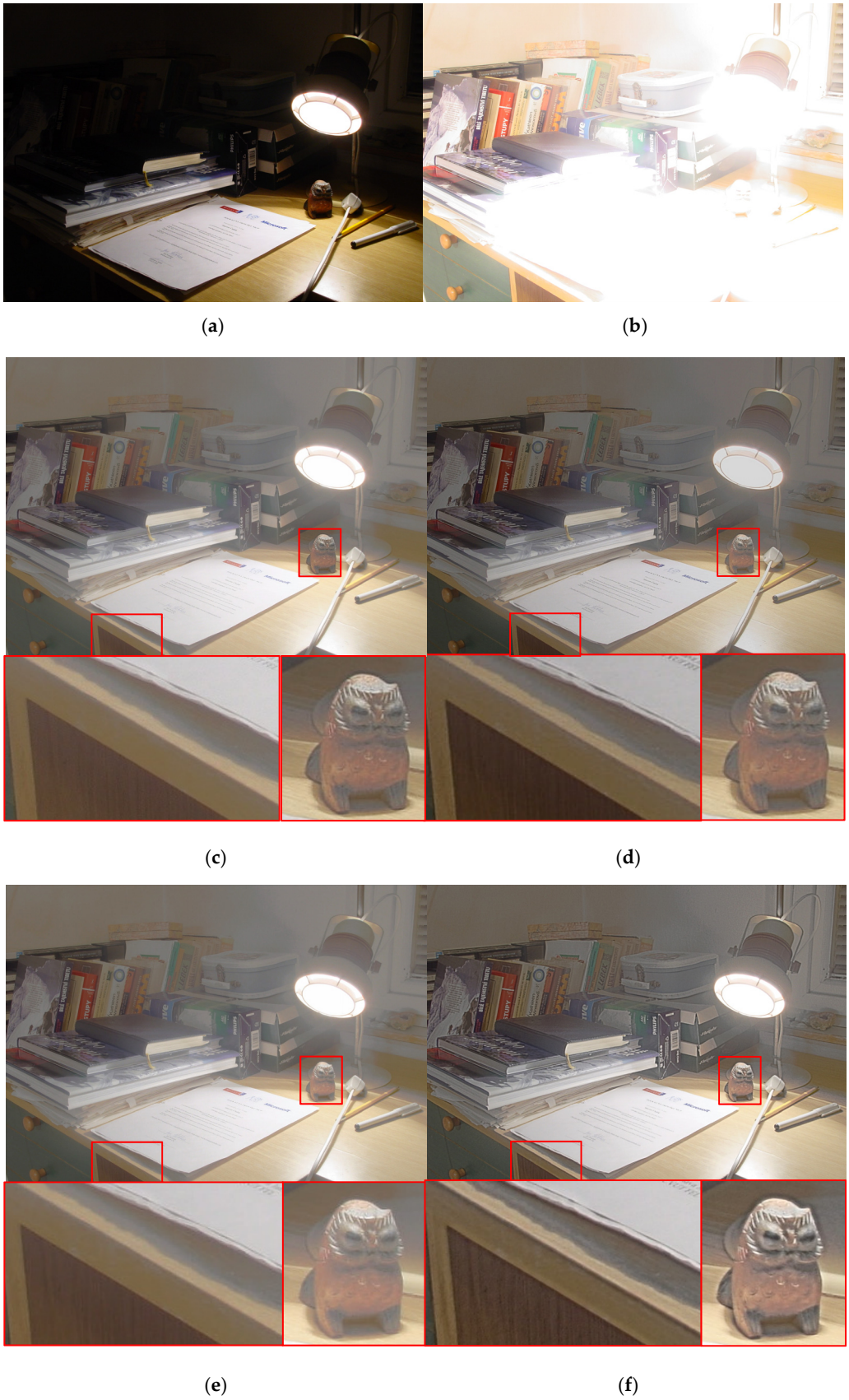

Figure 12. Input and result images (2): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

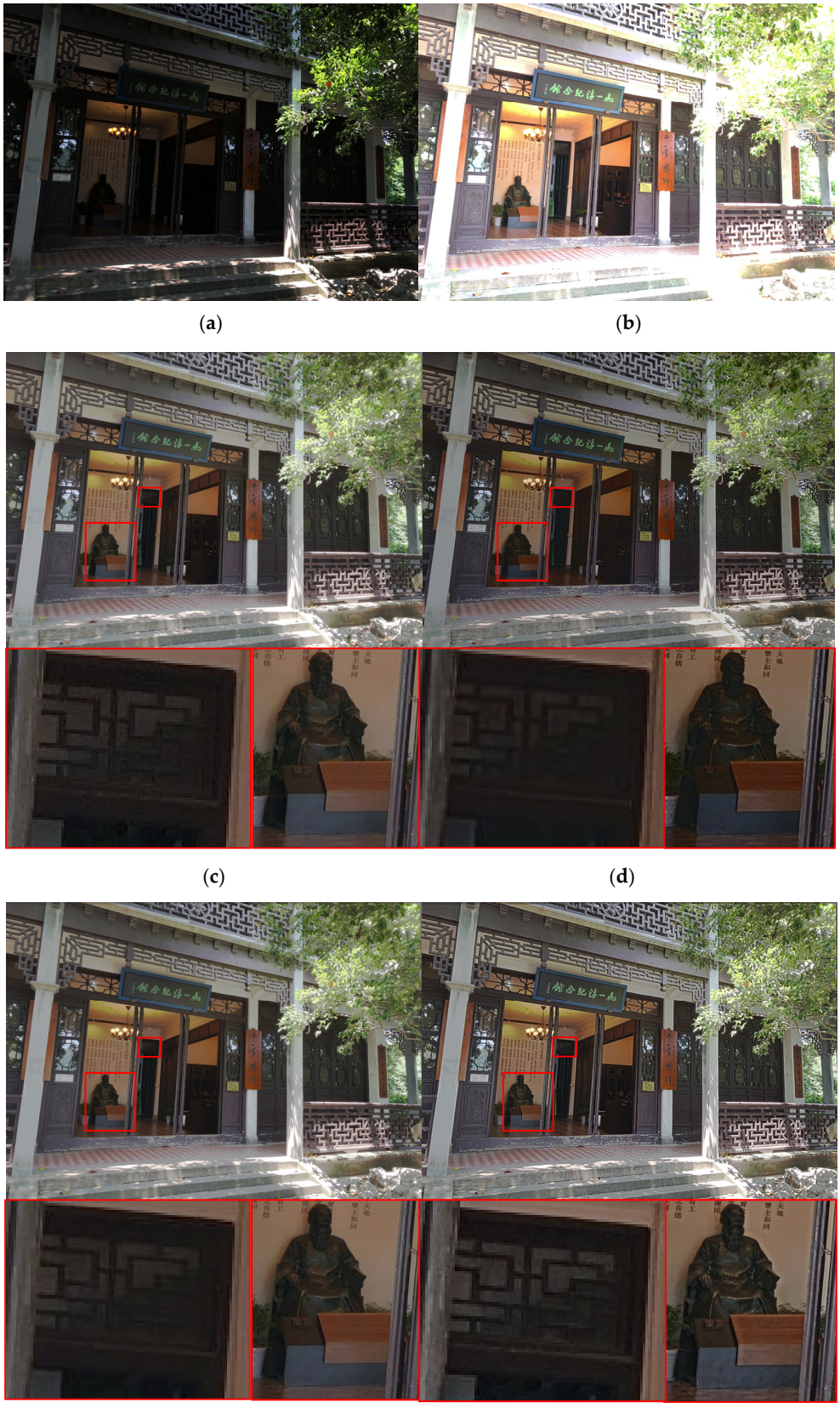

Figure 13. Input and result images (3): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

Figure 14. Input and result images (4): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

Figure 15. Input and result images (5): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

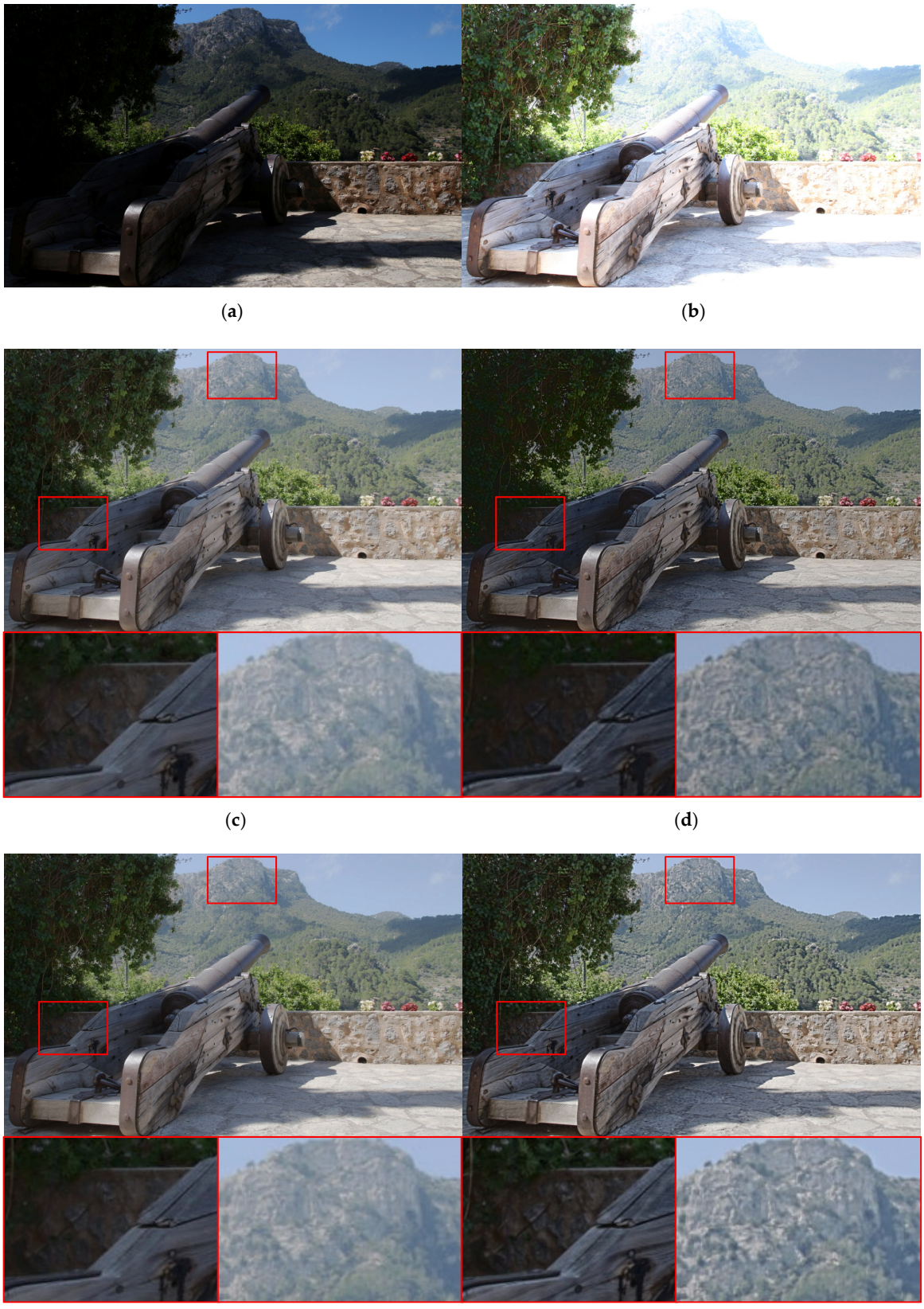

Figure 16. Input and result images (6): (**a**) Under-exposed image, (**b**) Over-exposed image, (**c**) DenseFuse, (**d**) U2Fusion, (**e**) TransMEF, (**f**) Proposed model.

The proposed model was trained using a correct-exposed image from incorrectly exposed inputs. However, it does not reflect useful information from under-or over-exposed images. Thus, we added the MaskBuilder and ParamTuner to robustly learn the information from the inputs. To verify the effectiveness of the MaskBuilder and ParamTuner, we trained and evaluated the networks by ablating our blocks, as listed in Table 2.

Table 2. Comparison of image-quality metrics score between the proposed methods: Co-target denotes correct-exposed image target, and MB and PT denote the MaskBuilder and the ParamTuner, respectively.

Co-Target	MB	PT	AG	EI	FMI	LPC-SI	S3	SF	PIQE	MS-SSIM	Qabf	VIFF
✔			4.8802	49.2164	0.3491	0.9526	0.2021	16.0696	27.4676	0.8809	0.4622	0.4604
✔	✔		5.9602	60.0486	0.3459	0.9549	0.2294	19.3997	27.9720	0.8962	0.5184	0.5421
✔	✔	✔	6.1319	61.6277	0.3447	0.9543	0.2369	19.7974	25.8651	0.8916	0.5059	0.5477

As we observed, the network with MaskBuilder and ParamTuner had superior scores, which indicate image sharpness and human perceptual quality. However, in terms of the quality of fusion with the source images, it showed slightly lower quality scores than networks without ParamTuner. We speculate that this is because the features of the source images are not synthesized equally, allowing the dynamic hyperparameter to learn more information from a more useful source image. As a result, the outcome of the network with ParamTuner can be considered effective because the objective of our approach is to synthesize a better quality HDR image using useful details from the LDR images.

5. Conclusions

In this paper, we propose a novel training method for MEF using multi-task learning. First, we classified the multi-exposure images into three categories: under-exposure, over-exposure, and correct exposure. The images that belong to the under-exposure or over-exposure category were used as source images and as reference images. This allowed the MEF networks to compensate for the saturated regions of the LDR images. Moreover, we suggest MaskBuilder to reproduce advanced reference images that contain useful information from source images. Thus, the ground truth of the LDR images is not necessary. Finally, our ParamTuner has the effect of fusing high-quality images by applying dynamic hyperparameters to the incorrect-exposed image losses. The quantitative and qualitative experimental results demonstrate that the proposed method can produce images with sharper and enhanced quality as perceived by humans. However, since this study only tested the learning method of the proposed approach on one CNNs model, it is necessary to evaluate the validity of the proposed method on various networks in future studies. In addition, future research should explore how to optimize a deep-learning-based fusion model by applying our proposed learning method so that it can be used in practical applications such as surveillance systems or autonomous driving, which require diverse types of image information.

Author Contributions: Conceptualization, S.-H.L.; methodology, S.-H.L. and C.-G.I.; software, C.-G.I.; validation, S.-H.L. and C.-G.I.; formal analysis, S.-H.L. and C.-G.I.; investigation, S.-H.L. and C.-G.I.; resources, S.-H.L., D.-M.S. and C.-G.I.; data curation, S.-H.L., H.-J.K. and C.-G.I.; writing—original draft preparation, C.-G.I.; writing—review and editing, S.-H.L.; visualization, C.-G.I.; supervision, S.-H.L.; project administration, S.-H.L.; funding acquisition, S.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Korea (NRF-2021R1I1A3049604) and supported by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-

RS-2022-00156389) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest regarding the publication of this paper.

References

1. Reinhard, E.; Stark, M.; Shirley, P.; Ferwerda, J. Photographic Tone Reproduction for Digital Images. In Proceedings of the SIGGRAPH 2002: 29th Annual Conference on Computer Graphics and Interactive Techniques, San Antonio, TA, USA, 23–26 July 2002; pp. 267–276. [CrossRef]
2. Duan, J.; Bressan, M.; Dance, C.; Qiu, G. Tone-Mapping High Dynamic Range Images by Novel Histogram Adjustment. *Pattern Recognit.* **2010**, *43*, 1847–1862. [CrossRef]
3. Jung, T.; Kwon, H.J.; Hahn, J.; Lee, S.H. Enhanced HDR Image Reproduction Using Gamma-Adaptation-Based Tone Compression and Detail-Preserved Blending. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2020**, *E103A*, 728–732. [CrossRef]
4. Burt, P.J. *The Pyramid as a Structure for Efficient Computation*; Springer: Berlin/Heidelberg, Germany, 1984; pp. 6–35. [CrossRef]
5. Jinno, T.; Okuda, M. Multiple Exposure Fusion for High Dynamic Range Image Acquisition. *IEEE Trans. Image Process.* **2012**, *21*, 358–365. [CrossRef] [PubMed]
6. An, J.; Lee, S.H.; Kuk, J.G.; Cho, N.I. A Multi-Exposure Image Fusion Algorithm without Ghost Effect. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 22–27 May 2011; pp. 1565–1568. [CrossRef]
7. Li, H.; Wu, X.J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]
8. Qu, L.; Liu, S.; Wang, M.; Song, Z. TransMEF: A Transformer-Based Multi-Exposure Image Fusion Framework Using Self-Supervised Multi-Task Learning. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2126–2134. [CrossRef]
9. Bruce, N.D.B. ExpoBlend: Information Preserving Exposure Blending Based on Normalized Log-Domain Entropy. *Comput. Graph.* **2014**, *39*, 12–23. [CrossRef]
10. Song, M.; Tao, D.; Chen, C.; Bu, J.; Luo, J.; Zhang, C. Probabilistic Exposure Fusion. *IEEE Trans. Image Process.* **2012**, *21*, 341–357. [CrossRef]
11. Lee, S.H.; Park, J.S.; Cho, N.I. A Multi-Exposure Image Fusion Based on the Adaptive Weights Reflecting the Relative Pixel Intensity and Global Gradient. In Proceedings of the 2018 IEEE International Conference on Image Processing (ICIP 2018), Athens, Greece, 7–10 October 2018; pp. 1737–1741. [CrossRef]
12. Xu, F.; Liu, J.; Song, Y.; Sun, H.; Wang, X. Multi-Exposure Image Fusion Techniques: A Comprehensive Review. *Remote Sens.* **2022**, *14*, 771. [CrossRef]
13. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-Level Image Fusion: A Survey of the State of the Art. *Inf. Fusion* **2017**, *33*, 100–112. [CrossRef]
14. Huang, F.; Zhou, D.; Nie, R.; Yu, C. A Color Multi-Exposure Image Fusion Approach Using Structural Patch Decomposition. *IEEE Access* **2018**, *6*, 42877–42885. [CrossRef]
15. Wang, S.; Zhao, Y. A Novel Patch-Based Multi-Exposure Image Fusion Using Super-Pixel Segmentation. *IEEE Access* **2020**, *8*, 39034–39045. [CrossRef]
16. Kalantari, N.K.; Ramamoorthi, R. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* **2017**, *36*, 1–12. [CrossRef]
17. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518. [CrossRef]
18. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732.
19. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multi-Scale Structural Similarity for Image Quality Assessment. *Conf. Rec. Asilomar Conf. Signals Syst. Comput.* **2003**, *2*, 1398–1402. [CrossRef]
20. Rahman, S.; Rahman, M.M.; Abdullah-Al-Wadud, M.; Al-Quaderi, G.D.; Shoyaib, M. An Adaptive Gamma Correction for Image Enhancement. *Eurasip J. Image Video Process.* **2016**, *2016*, 35. [CrossRef]
21. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
22. Sukthankar, R.; Stockton, R.G.; Mullin, M.D. Smarter Presentations: Exploiting Homography in Camera-Projector Systems. *Proc. IEEE Int. Conf. Comput. Vis.* **2001**, *1*, 247–253. [CrossRef]
23. Son, D.-M.; Kwon, H.-J.; Lee, S.-H. Visible and Near Infrared Image Fusion Using Base Tone Compression and Detail Transform Fusion. *Chemosensors* **2022**, *10*, 124. [CrossRef]
24. Debevec, P.E.; Malik, J. Recovering High Dynamic Range Radiance Maps from Photographs. In Proceedings of the ACM SIGGRAPH 2008 Classes, Los Angeles, CA, USA, 11–15 August 2008; Volume 31. [CrossRef]
25. HDRsoft Gallery. Available online: http://www.hdrsoft.com/examples2.html (accessed on 26 November 2015).

26. Cai, J.; Gu, S.; Zhang, L. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Trans. Image Process.* **2018**, *27*, 2049–2062. [CrossRef]
27. Multi-Exposure HDR Capture. Wikipedia. Available online: https://en.wikipedia.org/wiki/Multi-exposure_HDR_capture (accessed on 3 January 2023).
28. Cui, G.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Detail Preserved Fusion of Visible and Infrared Images Using Regional Saliency Extraction and Multi-Scale Image Decomposition. *Opt. Commun.* **2015**, *341*, 199–209. [CrossRef]
29. Rajalingam, B.; Priya, R. Hybrid Multimodality Medical Image Fusion Technique for Feature Enhancement in Medical Diagnosis. *Int. J. Eng. Sci. Invent.* **2018**, *2*, 52–60.
30. Haghighat, M.; Razian, M.A. Fast-FMI: Non-Reference Image Fusion Metric. In Proceedings of the 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 15–17 October 2014; pp. 1–3.
31. Hassen, R.; Wang, Z.; Salama, M.M.A. Image Sharpness Assessment Based on Local Phase Coherence. *IEEE Trans. Image Process.* **2013**, *22*, 2798–2810. [CrossRef] [PubMed]
32. Vu, C.T.; Phan, T.D.; Chandler, D.M. S_3: A Spectral and Spatial Measure of Local Perceived Sharpness in Natural Images. *IEEE Trans. Image Process.* **2012**, *21*, 934–945. [CrossRef] [PubMed]
33. Eskicioglu, A.M.; Fisher, P.S. Image Quality Measures and Their Performance. *IEEE Trans. Commun.* **1995**, *43*, 2959–2965. [CrossRef]
34. Venkatanath, N.; Praneeth, D.; Maruthi Chandrasekhar, B.H.; Channappayya, S.S.; Medasani, S.S. Blind Image Quality Evaluation Using Perception Based Features. In Proceedings of the 2015 21st National Conference on Communications (NCC 2015), Mumbai, India, 27 February–1 March 2015. [CrossRef]
35. Xydeas, C.S.; Petrović, V. Objective Image Fusion Performance Measure. *Electron. Lett.* **2000**, *36*, 308. [CrossRef]
36. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A New Image Fusion Performance Metric Based on Visual Information Fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [CrossRef]