Article

Multi-Scale Attention-Guided Non-Local Network for HDR Image Reconstruction

Howoon Yoon, S. M. Nadim Uddin and Yong Ju Jung *

School of Computing, Gachon University, Seongnam 13120, Korea
* Correspondence: yjung@gachon.ac.kr

Abstract: High-dynamic-range (HDR) image reconstruction methods are designed to fuse multiple Low-dynamic-range (LDR) images captured with different exposure values into a single HDR image. Recent CNN-based methods mostly perform local attention- or alignment-based fusion of multiple LDR images to create HDR contents. Depending on a single attention mechanism or alignment causes failure in compensating ghosting artifacts, which can arise in the synthesized HDR images due to the motion of objects or camera movement across different LDR image inputs. In this study, we propose a multi-scale attention-guided non-local network called MSANLnet for efficient HDR image reconstruction. To mitigate the ghosting artifacts, the proposed MSANLnet performs implicit alignment of LDR image features with multi-scale spatial attention modules and then reconstructs pixel intensity values using long-range dependencies through non-local means-based fusion. These modules adaptively select useful information that is not damaged by an object’s movement or unfavorable lighting conditions for image pixel fusion. Quantitative evaluations against several current state-of-the-art methods show that the proposed approach achieves higher performance than the existing methods. Moreover, comparative visual results show the effectiveness of the proposed method in restoring saturated information from original input images and mitigating ghosting artifacts caused by large movement of objects. Ablation studies show the effectiveness of the proposed method, architectural choices, and modules for efficient HDR reconstruction.

Keywords: high-dynamic-range imaging; spatial attention; non-local means; deep learning

1. Introduction

High-dynamic-range (HDR) imaging techniques extend the range of brightness in an image such that a photograph taken by a camera becomes as similar as possible to the scene as it would be observed by the naked eye. HDR images have been used in a wide range of applications such as photography, film, and games [1,2] as they contain more information from a given scene and can provide a better visual experience. However, because camera image sensors generally have a narrow dynamic range, capturing scenes with more widely varying ranges of brightness using a single exposure value is considered relatively difficult [3,4]. Recent advances in imaging technology have enabled the acquisition of HDR images with only a single-sensor camera, but these special cameras are often very expensive for consumers in general. Therefore, the primary method used to obtain HDR imaging is to continuously photograph and overlap multiple images with a low dynamic range (LDR) with different exposure values and then select or fuse the best photo pixels or segments to reconstruct a single HDR image [5–7].

Traditional methods mostly depend on rejection-based approaches [8–17], rigid and non-rigid registration-based approaches [18–23] and patch-based optimization methods [24,25] for reconstructing HDR images. Rejection-based HDR reconstruction methods produce high-quality HDR contents for static scenes with LDR contents for moving regions, thus limiting their scope for static scenes mostly. Registration-based traditional approaches work plausibly well with static scenes and small movements, however, fail in complex
and dynamic scenes with large motions and occluded pixels. Patch-based optimization methods, though capable of high-quality HDR image reconstruction, suffer from high computational complexity and often fail to produce plausible HDR images in run-time.

To alleviate the HDR reconstruction problem, recently, various deep learning methods based on convolutional neural network (CNN) have been utilized to perform multi-exposure-based multi-frame HDR imaging [26–31]. CNN-based methods have shown significant promise in producing HDR contents with fewer LDR images (i.e., typically three LDR images) for comparatively complex scenes. Early studies focused on pre-processing methods to align LDR images beforehand [32–35] using optical flow [22,23,28,29] or homography [29,34,36,37], and then input the data to convolutional neural network (CNN) models. However, although properly coping with small movements of objects in this way has been shown to be possible, the performance of these methods is limited by the accuracy of the alignment. Moreover, these methods often fail in scenes with large motions and saturated regions.

Recent CNN models are focused on performing feature-level attention-based implicit alignment mechanisms that use contextual information from the LDR images to adaptively select pixels/regions for efficient fusion [26,27,30,38]. These methods enabled the movement of an object to be properly processed while maintaining important information for each input image. The key challenge of such a multi-frame HDR imaging approach is to properly handle the “ghosting artifacts” caused by inappropriate image fusion due to objects moving within a scene or camera instability. For example, as shown in Figure 1, if images captured continuously at different times are improperly processed and fused, different objects appear to overlap in the resulting image, i.e., ghosting artifacts, due to the motion of the objects.

![Figure 1. Visual comparison of HDR image restoration results obtained using the proposed method and state-of-the-art methods. From left, Wu [28], AHDR [26], AD [27], NHDRR [30], DAHDR [39], Proposed MSANLnet, and Ground Truth.](image)

In this study, we propose a deep neural network, called MSANLnet, that uses the multi-scale attention mechanism and non-local means technique to effectively alleviate the
ghosting artifact problem in HDR image reconstruction. The proposed MSANLnet has two distinctive parts—(a) multi-scale attention for object motion and saturation mitigation, and (b) non-local means-based fusion. Existing spatial-wise attention only captures important regions on a single scale, which is not effective for distinguishing both object movements and saturated regions in the case of producing HDR contents. In the proposed method, LDR features are aligned based on multi-scale spatial attention modules. During this process, spatial attention is performed for each scale of features to expand the size of the receptive field and progressively correct and fuse important information, such as large movements and saturation, which is found to be effective in producing high-quality HDR content. Again, the non-local means module looks at the whole image and only fuses the correlated global contents with the reference image. Hence, the multi-scale attention mechanism can effectively capture the local movements of the objects and the non-local means-based fusion can reduce ghosting artifacts on a global level by looking at the whole image, which effectively mitigates ghosting artifacts for large object motions and global shifts such as camera translation.

This two-stage design effectively selects different exposure values for each input frame and useful information that is not altered by motion and then fuses images to minimize locality to reduce noise and mitigate the ghosting artifact problem. We validated that the proposed method performs better than the existing methods quantitatively and qualitatively through the experiments on a publicly available dataset that has been mainly used for HDR image restoration. We summarize our contributions as follows.

- We propose MSANLnet, a multi-scale attention-guided non-local network for HDR image reconstruction, that extracts important features from the LDR features using the multi-scale spatial attention and adaptively fuses the contextual features to obtain HDR images.
- We show that the multi-scale spatial attention, combined with the non-local means-based fusion, can effectively alleviate the “ghosting artifact” and produce aesthetic HDR images.
- Our proposed method outperforms the existing methods in both qualitative and quantitative measures, validating the efficacy of the attention modules, non-local means-based fusion, and architectural choices.

The remainder of this study is organized as follows. Related work is discussed in Section 2. The proposed MSANLnet is presented along with detailed descriptions of each module in Section 3. The experimental setup and results are presented in Section 4. Finally, Section 5 presents our conclusions and suggests some possible avenues for further research.

2. Related Work

First proposed by Madden [40] and Mann [41] and popularized by Debevec and Malik [42] for digital imaging, multi-exposure-fusion-based HDR reconstruction methods can be classified into two major classes—(i) traditional approaches, and (ii) deep learning-based approaches. For brevity, we restrict our related work only to multi-exposure-fusion-based approaches.

Rejection-based methods try to find the pixels/regions with motions and select only pixels/regions using a reference image or the static contents of the LDR images [5,8,12], i.e., these methods reject the moving contents. Rejection mechanisms are typically implemented by performing patch-wise comparison [9], illumination consistency and linear relationships among pixels of the LDR images [16,43,44], thresholding [5,11,12], background probability map [45], super-pixel comparison [15] etc. Such approaches can produce fast results, however, are limited to producing LDR contents in the dynamic regions.

Alignment-based traditional methods typically select one reference exposure LDR image and register the remaining LDR images either using rigid registration mechanisms (e.g., SIFT-based [18], SURF-based [10], Median Threshold Bitmap-based [19] methods, etc.) or non-rigid registration mechanisms (e.g., optical flow-based methods [20,22,23,46]).
These methods can produce plausible results for static scenes and small motions, however, perform poorly in recovering saturated pixels.

Patch-based optimization methods typically synthesize LDR patches based on the structure reference patches by finding dense correspondence among the LDR images and fuse the synthesized LDR images to produce HDR contents [24,25,47]. These methods can produce high-quality HDR content but suffer from computational complexity. Though most of the traditional methods produce plausible HDR contents for static scenes, they suffer from poor synthesis/misalignment of LDR contents in dynamic scenes with large motions and recovering saturated pixels/regions from the input LDR images.

Deep learning-based methods are capable of synthesizing novel HDR contents based on the input LDR images, making them suitable for complex dynamic scenes. Early methods employed CNN-based methods, e.g., optical flow estimation [28,29], multi-scale feature extraction [38], etc., to compensate ghosting artifacts. Subsequently, attention-based alignment methods are used to remove ghosting artifacts and recover saturated regions [26,27,39,48].

Deep learning-based multi-scale attention mechanisms have also shown significant promise in various image reconstruction tasks, e.g., semantic image synthesis [49,50], generative image synthesis [51–53], image inpainting [54–57], segmentation [58–64], super-resolution [65–74], image enhancement [75], etc. Though several HDR approaches utilize multi-scale feature fusion [26,38,39,76–81], the removal of object motion and the mitigation of saturation using multi-scale attention mechanisms remain an under-explored research area.

Recently, generative adversarial networks (GANs) have been employed to supervise the HDR reconstruction process, which can effectively correct large motions with higher efficiency [82]. Despite the success of deep learning-based approaches, removing ghosting artifacts in HDR images for large motions is still an active area of research.

3. Methods

3.1. Overview of MSANLnet

The MSANLnet takes three input LDR tensors as inputs. Each tensor consists of an LDR image and the tone-mapped HDR image of the corresponding LDR image. First, the three LDR images in the input LDR set (i.e., $L = \{L_1, L_2, L_3\}$, where $L_i \in \mathbb{R}^{H \times W \times 3}$) are arranged according to their exposure lengths $t_i$, i.e., short exposure ($L_1$), intermediate exposure ($L_2$), and long exposure length ($L_3$). Then, a mapping process is performed on the input LDR image set $L$ to convert them into the input HDR image set, $H = \{H_1, H_2, H_3\}$, $H_i \in \mathbb{R}^{H \times W \times 3}$. Here, the gamma correction ($\gamma = 2.2$, following [26]) is used for the mapping process as follows:

$$H_i = \frac{(L_i)^\gamma}{t_i}; i = 1, 2, 3,$$

Then, a pair of input tensors $X_i$ for each $L_i$ is obtained by concatenating the gamma-corrected $H_i$ and the corresponding $L_i$ in the channel dimension, i.e., $X_i = \{H_i, L_i\}$, where $X_i \in \mathbb{R}^{H \times W \times 6}$. Among these tensors, $X_2$ with the intermediate exposure length is selected as a reference tensor. The set of input tensors $X$ is then fed to MSANLnet to produce an HDR image.

We define the process of generating the final HDR image $H_f$ by fusing the three tensors, $X = \{X_1, X_2, X_3\}$, as follows,

$$H_f = f(X_1, X_2, X_3; \theta),$$

where $f$ represents the proposed MSANLnet and $\theta$ is a parameter of the corresponding network.

The proposed MSANLnet model uses a CNN-based encoder-decoder structure and is largely composed of multi-scale spatial attention modules and non-local means-based fusion modules. Figure 2 shows the overview of the proposed network. As shown...
in Figure 2, the encoder extracts spatial attention features based on the scale of LDR tensor set \( X \) through the multi-scale spatial attention module (please refer to Section 3.2 and Supplementary Materials). The multi-scale spatial attention features contribute to compensating motions and recovering feature values from the saturated regions adaptively. The decoder uses a transposed convolution block to restore the reduced resolution to the original resolution while concatenating spatial attention features and reference image features. Finally, the final HDR image is output through a non-local means-based fusion mechanism (please refer to Section 3.3 and Supplementary Materials) which considers the global relationships among feature values, thus effectively removing ghosting artifacts for larger motions during the HDR image reconstruction.

**Figure 2.** Overview of the proposed MSANLnet for HDR image reconstruction.

### 3.2. Multi-Scale Spatial Attention Module

The proposed approach is designed to effectively identify areas to be used in images with long exposure values and images with short exposure values. Specifically, a multi-scale attention mechanism is devised in the encoder to implicitly align the two images with the reference image, as shown in Figure 2.

First, three 6-channel input tensors \( X_1, X_2, X_3 \) are encoded in spatial multi-scales, respectively, via a convolutional layer to extract features. During this process, max pooling is applied to reduce the spatial size of the feature map extracted at each step by half from the previous step. Thus, the receptive field is increased to capture a larger foreground movement in the multi-scale attention module. Note that when we decrease the spatial size of a feature map by half, we increase its number of channels by double. For example, in our experiment, we used 64 channels for the first scale and 128 channels for the second scale.

As shown in Figure 2, a spatial attention block is executed for each scale to extract the spatial attention maps of the non-reference images. A spatial attention block is shown in Figure 3. In the spatial attention block, the input is the concatenation of the non-reference image features, \( X_i; i \neq 2 \), and the reference image \( X_2 \) features. Then, the spatial attention map is extracted at each scale \( s \) as follows,

\[
A_i^{(s)} = g(X_i^{(s)}, X_2^{(s)}); i = 1, 3; s = 1, 2, \ldots, N,
\]

where \( s \) denotes the number of spatial scales and \( g() \) is the spatial attention map extractor. The attention map extractor consists of two consecutive convolutions and sigmoid function, as used in a previous study [26]. Note that, in our experiment, we used \( N=2 \) (i.e., two scales as depicted in Figure 2). An element of the spatial attention map \( A_i \) is in a range of \([0 – 1]\) and the spatial size of the generated \( A_i^{(s)} \) is equal to \( X_i^{(s)} \).
Finally, an element-wise multiplication between the input feature $X_i^{(s)}$ and the extracted $A_i^{(s)}$ is performed to generate the final spatial attention feature $X'_i^{(s)}$ of the images.

$$X'_i^{(s)} = A_i^{(s)} \cdot X_i^{(s)}; i = 1, 3; s = 1, 2, \ldots, N. \quad (3)$$

As seen in Figure 4, two spatial attention maps focus on important regions for implicit alignment of the input LDR images to the reference image. As the network is provided with explicit multi-scale contextual information, it can compensate for large foreground motions and can effectively mitigate the ghosting artifacts. Moreover, as the spatial attention maps represent the differences among the LDR images with varying exposures, the network implicitly learns to recover saturated pixel values.

3.3. Non-Local Means-Based Fusion

In some local areas of the image feature, sufficient information is not available due to occlusion or saturation caused by the movement of objects [26]. Therefore, in image fusion networks, useful information must be extracted into a large receptive field. However, a single CNN convolution filter can be used in only a limited area of the receptive field. Previous studies have shown the importance of short and long-range dependencies among the pixel/feature values for various computer vision tasks [54,83–86]. Non-local means is a method of calculating long-range dependencies and restoring pixels through weighted averages based on these dependencies, intuitively beneficial for removing ghosting artifacts. To this end, we applied a non-local block to image fusion so that global information is utilized and locality is reduced to effectively alleviate ghosting artifacts.

First, we utilize the residual dense convolutional block [87] before applying the non-local block, such that the feature $Z$ recovered to the original size through the decoding.
process learns sufficient local information. By concatenating all layers, useful local information is adaptively extracted through the features of the previous layer and the features of the current layer. The extracted $X$ enters a non-local block input and outputs $Y$ of the same size, $H \times W \times C$, as a result. Figure 5 illustrates the structure of a residual dense convolutional block composed of three convolutional blocks.

Figure 5. A residual dense convolutional block is composed of dense concatenations based on skip-connections between a series of convolutional blocks consisting of convolutions and ReLU activation functions.

For the non-local based fusion, we adopt the asymmetric pyramid non-local block (APNB) [88] to construct the non-local blocks. In APNB, spatial pyramid pooling is applied to a standard non-local block to reduce computation by sampling part of the feature map based on meaningful global information. Figure 6 shows the structure of an asymmetric pyramidal non-local block, including spatial pyramid pooling.

Figure 6. Structure of the asymmetric pyramidal non-local block with spatial pyramid pooling [88].

First, the input feature map $X$ is converted to different embeddings, $\Phi$, $\theta$, and $\gamma$, respectively, through three $1 \times 1$ convolutions, $W_\phi$, $W_\theta$ and $W_\gamma$.

$$\Phi = W_\phi(X); \theta = W_\theta(X); \gamma = W_\gamma(X).$$  \hspace{1cm} (4)

The spatial pyramid pooling module is composed of four pooling layers that derive the results of different sizes in parallel. Such a pyramid pooling mechanism improves the expressive power of global features as previous studies have shown the effectiveness of the global and multi-scale representations for capturing scene semantics. The sampling process through pyramid pooling $P_\theta$ and $P_\gamma$ are represented by the following equation.

$$\theta_P = P_\theta; \gamma_P = P_\gamma.$$  \hspace{1cm} (5)

The number of spatially sampled anchor points $S$ can be expressed as,
where $S$ denotes the number of sampled anchor points and $n$ denotes the width of the output features processed through the pooling layer. Thereafter, based on the standard non-local block method [89], the pseudo-matrix $V_p$ of $\theta_p$ and $\Phi$ is calculated as follows,

$$V_p = \Phi^T \times \theta_p.$$  

Instance normalization is then applied to $V_p$ to generate a normalized pseudo-matrix $V^p$, and the Softmax function of the self-attention mechanism is used to derive the attention result as follows,

$$Z = V^p \times \gamma^T_p.$$  

The final output of the asymmetric pyramidal non-local block is obtained as follows,

$$Y = cat(W_o(Z), X),$$

where $W_o$ is composed of a $1 \times 1$ convolution and can be learned by weighting the parts where parameters are important in the non-local operation. Subsequently, the network generates a final 3-channel HDR image through a convolutional layer.

4. Experimental Results

4.1. Dataset, Objective Function, Implementation Details, and Evaluation Design

4.1.1. Dataset

We used an open dataset [28] for HDR performance comparison to train and evaluate the proposed method. Specifically, out of a total of 89 image sets, 74 image sets were used as a training set, 5 were used as a validation set, and the remaining 10 were used as a testing set. Each image set comprised three LDR images and one HDR ground-truth image. Here, three LDR images were captured using exposure bias $\{-2, 0, +2\}$ or $\{-3, 0, +3\}$. As in an existing study on HDR [26], in the training stage, a total of 1775 image patches were used to perform training by cropping LDR images with a size of $1000 \times 1500$ into $256 \times 256$ patch sizes and using them with a stride of 128 through data augmentation to mitigate over-fitting.

4.1.2. Objective Function

As proposed in a previous work [28], HDR images were mainly presented after being tone-mapped, and thus, it was more effective to optimize the network in the tone-mapping domain than in the HDR domain. A tone-mapping process using a $\mu$-law was performed on the HDR image $H$ and ground-truth image $GT$ derived in this study as follows.

$$T(X) = \log \left( \frac{1 + \mu X}{1 + \mu} \right),$$

where $\mu$ is a parameter indicating the degree of compression, and $T(X)$ denotes a tone-mapped image. We set $\mu = 5000$ [28], and the range of the resulting $T(X)$ is $[0, 1]$. Subsequently, to minimize the error value between $T(H)$ and $T(GT)$ by applying per-pixel $L_1$-Loss, the network was trained after defining the loss function $L$ as follows,

$$L = |T(GT) - T(H)|_1.$$  

4.1.3. Implementation Details

In the training stage of the proposed method, the number of epochs and batch size were set to 200 and 4, respectively. The learning rate was initially set to 0.0001, and then the learning rate was reduced by multiplying by 0.1 at 100 epochs. We used the Adam optimizer [90], and both learning and evaluation were implemented using the Pytorch.
framework [91]. Our experiments are performed on a single NVIDIA 2080Ti GPU. The number of model parameters is 2.8 million (2,772,707) and the average inference time per image is 0.0181 s.

4.1.4. Evaluation Design

To evaluate the efficiency of the proposed method, we compared it with five state-of-the-art methods, both in qualitative and quantitative manner. Specifically, we compared our method with AHDR [26], Wu [28], AD [27], NHDRR [30], and DAHDR [39]. AHDR [26] introduced spatial attention modules in HDR imaging to guide the merging according to the reference image. AD [27] used spatial attention module for feature-level attention and a multi-scale alignment module (i.e., PCD alignment [92]) to align the images in the feature-level. DAHDR [39] exploits both spatial attention and feature channel attention to achieve ghost-free merging. NHDRR [30] is based on U-Net-based architecture with non-local means to produce HDR images. All of the selected models are deep learning-based methods for generating an HDR image using multiple LDR images. In addition, all of these models used LDR images with large motions as training datasets, which were considered appropriate to objectively compare performance with our method. All models were trained in the same way and in the same environment as the proposed approach for a fair comparison.

4.2. Quantitative Evaluation

To quantitatively evaluate the performance of the model, we measured the peak signal-to-noise ratio (PSNR) and structural similarity index map (SSIM) on the testing set. PSNR is used to evaluate the impact of lost information on the quality of a generated or compressed image and represents the power of received noise with respect to the maximum power of the received signal. SSIM is a method designed to evaluate human visual quality differences rather than numerical loss and evaluates quality by comparing luminance, contrast, and structure values that make up images, rather than only performing comparisons between pixels. We used the PSNR-$l$ and PSNR-$\mu$ indicators, which measure PSNR in linear and tone-mapped domains, respectively. For the SSIM, we used values measured in tone-mapped domains.

Table 1 shows the average value of the quantitative indicators, and the larger values correspond to the higher quality. The proposed model achieved higher values than other models in all three quantitative performance indicators. The proposed method achieved better performance in terms of PSNR-$l$, PSNR-$\mu$, and SSIM measurements compared to AHDR [26], Wu [28], AD [27], NHDRR [30], and DAHDR [39]. This occurred because the proposed method can restore details and effectively eliminate ghosting artifacts through non-local means-based fusion modules. NHDRR [30] also merges images using non-local means modules, but MSANLnet, which extracts features using scale-specific spatial attention, exhibited numerically better performance.

### Table 1. Quantitative comparison of the performance of the proposed method (MSANLnet) and state-of-the-art models.

| Method/Metric   | PSNR-$l$ ↑ | PSNR-$\mu$ ↑ | SSIM ↑ |
|-----------------|------------|--------------|--------|
| NHDRR [30]      | 33.6970    | 39.2461      | 0.9382 |
| Wu [28]         | 36.4266    | 41.9838      | 0.9618 |
| AHDR [26]       | 38.0948    | 39.5937      | 0.9757 |
| AD [27]         | 37.9287    | 40.8102      | 0.9812 |
| DAHDR [39]      | 36.5915    | 40.0708      | 0.9753 |
| MSANL (proposed) | 40.4370    | 42.7466      | 0.9821 |

4.3. Qualitative Evaluation

Figures 7 and 8 visually compare the results of HDR image generation via the proposed method and existing state-of-the-art models [26–28,30,39]. In particular, Figure 7
qualitatively compares the experimental results of images with saturated backgrounds and large foreground movements. Input images referenced by the network by exposure value are shown in Figure 7a, and the tone-mapped HDR result images produced by the proposed method are shown in Figure 7b. The corresponding testing set image had a saturated background area and large foreground movements. Areas with large movements and saturation levels were cut into patches as shown in Figure 7b,c, respectively. The details of the background were obscured by the movement of an object in an image with a small exposure length, and the information was lost even in an image with long or medium exposure lengths. Because the image was saturated, learning models were likely to use artifacts and distorted information from the first image with the background occluded for HDR imaging, as seen in Figure 7c.

As seen in Figure 7, Wu [28] produced an excessively smooth image and failed to restore the details of the saturated area and removed ghosting artifacts. Although AHDR [26] selectively incorporated useful areas through spatial attention modules to restore obscured or saturated details, they still exhibited ghosting artifacts. NHDRR [30] removed ghosting artifacts to some extent, but output some blurring artifacts and missed the details of many areas. AD [27] uses PCD alignment modules [92], which reduce ghosting artifacts to some extent. However, the over-exposed area was not fully recovered. In DAHDR [39], the afterimage of the finger remains intact, which is not effective in removing ghosting artifacts. Furthermore, it can be seen that the hidden area cannot be restored as well. The results of the proposed MSANLnet expressed more details compared to the existing methods, while also eliminating ghosting artifacts, as may be observed from the figures. The resulting images demonstrate that the proposed MSANLnet expresses more details of saturated regions through the scale-specific spatial attention modules, and effectively reduces ghosting artifacts through non-local means-based fusion.

As seen in Figure 7, Wu [28] produced an excessively smooth image and failed to restore the details of the saturated area and removed ghosting artifacts. Although AHDR [26] selectively incorporated useful areas through spatial attention modules to restore obscured or saturated details, they still exhibited ghosting artifacts. NHDRR [30] removed ghosting artifacts to some extent, but output some blurring artifacts and missed the details of many areas. AD [27] uses PCD alignment modules [92], which reduce ghosting artifacts to some extent. However, the over-exposed area was not fully recovered. In DAHDR [39], the afterimage of the finger remains intact, which is not effective in removing ghosting artifacts. Furthermore, it can be seen that the hidden area cannot be restored as well. The results of the proposed MSANLnet expressed more details compared to the existing methods, while also eliminating ghosting artifacts, as may be observed from the figures. The resulting images demonstrate that the proposed MSANLnet expresses more details of saturated regions through the scale-specific spatial attention modules, and effectively reduces ghosting artifacts through non-local means-based fusion.

Figure 8 also shows a comparison of images with motion and supersaturation regions in the testing set. The proposed MSANLnet restored details more clearly compared to other models, and also removed visual artifacts to produce higher-quality HDR images.
Figure 8. Visual comparison of HDR images restored from the dataset provided in [28]. From left, Wu [28], AHDR [26], NHDRR [30], AD [27], DAHDR [39], Proposed MSANLnet, and Ground Truth.
4.4. Ablation Study

4.4.1. Ablation on the Network Structure

We performed an ablation study of the proposed network structure and analyzed the results. We compared the proposed MSANLnet with the following variant models to identify the importance of the individual components. Table 2 shows the quantitative comparison among the baselines and the proposed MSANLnet.

- **ANLnet**: The multi-scale attention module was removed from this version; i.e., this was a variant model using a single-scale attention module of the baseline model AHDR [26] and a non-local means-based fusion module.
- **MSANet**: The non-local fusion module was removed from this version. That is, this variant adopted a multi-scale attention module and dilated residual dense block (DRDB)-based fusion as used in baseline models.
- **MSANLnet**: The proposed MSANLnet model, which includes a multi-scale attention module and a non-local means-based fusion module.

**Table 2.** Quantitative comparison of the performance of the proposed method (MSANLnet) and its baseline variants.

| Method/Metric   | PSNR-l | PSNR-μ | SSIM  |
|-----------------|--------|--------|-------|
| MSAnet          | 37.1958| 39.5646| 0.9699|
| ANLnet          | 39.0193| 40.6810| 0.9781|
| MSANLnet (Proposed) | 40.4370| 42.7466| 0.9821|

As shown in Table 2, when a multi-scale attention module was used, the performance in terms of PSNR was improved by 1.41 dB compared to the use of a single-scale module. As seen in Figure 9, an afterimage of arm movement remains in MSAnet, and the difference cannot be repaired naturally, creating boundaries. However, it can be seen that the proposed model with multi-scale attention effectively eliminates such artifacts. These results confirm that expanding the receptive field helps reduce ghosting artifacts and restore details.

![Figure 9](image.png)

**Figure 9.** Qualitative comparison of the performance of the proposed method (MSANLnet) and its baseline variants.

4.4.2. Comparison of Degradation of HDR Restoration for Global Motion

We have performed additional experiments with the translation applied to the input LDR images and visually compared the HDR restoration performance between the proposed and existing methods. Specifically, we globally shifted the first input LDR image to...
the right by 50 pixels, kept the reference LDR image (i.e., the second image) unchanged, and shifted the third input LDR image to the left by 50 pixels, to simulate the global translation operation. The visual results are summarized in Figure 10.

As seen in Figure 10, most of the existing methods fail to mitigate ghosting artifacts in the case of such global motion caused by the camera’s translation. However, our proposed method can effectively produce HDR image contents without severe quality degradation, even though we do not employ any explicit alignment mechanisms. This is because the multi-scale attention mechanism can effectively capture the local movements of the objects and the non-local means-based fusion reduces ghosting artifacts on a global level by looking at the whole image, which essentially mitigates ghosting artifacts for large object motions and global shifts such as translation.
4.4.3. Generalization Ability on Different Dataset

We performed additional experiments on the generalization ability of the proposed method on a different test dataset. Specifically, we used test dataset provided in [93] and summarized the visual results in Figure 11. As seen in the figure, our method can generalize to unseen test datasets and performs comparatively better than the existing methods in restoring HDR contents, even though our method was not trained on the dataset.

![Figure 11. Visual comparison of additional HDR images from the dataset provided in [93]. From left, Wu [28], AHDR [26], NHDRR [30], AD[27], DAHDR [39], and Proposed MSANLnet.](image)

5. Conclusions

In this study, we have proposed MSANLnet as a non-local network based on multi-scale attention for effective HDR image restoration. In the encoder part, implicit alignment of features is performed at various resolutions with multi-scale spatial attention modules. In the decoder part, image restoration is performed by adaptively incorporating useful information utilizing a long-range dependency with a non-local means-based fusion module. The results show that the proposed method exhibited better performance than existing deep learning methods. The results of this study have confirmed the importance and impact of the expansion of the receptive field in HDR image restoration based on CNN models. In the future, we plan to further improve the quality of HDR images by applying and developing vision transformer-based [94] modules with long-range dependencies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/s22187044/s1, Table S1: Architecture of the Encoder. Table S2: Architecture of the Decoder.

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

Funding: This work was supported by the Gachon University research fund of 2022 (GCU-202206860001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anderson, E.F.; McLoughlin, L. Critters in the classroom: A 3D computer-game-like tool for teaching programming to computer animation students. In Proceedings of the International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2007, San Diego, CA, USA, 5–9 August 2007; Educators Program; Swanson, J., Ed.; ACM: New York, NY, USA, 2007; p. 7. [CrossRef]

2. Rempel, A.G.; Heidrich, W.; Li, H.; Mantiuk, R. Video viewing preferences for HDR displays under varying ambient illumination. In Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization, APGV 2009, Chania, Crete, Greece, 30 September–2 October 2009; Mania, K., Riecke, B.E., Spencer, S.N., Bodenheimer, B., O’Sullivan, C., Eds.; ACM: New York, NY, USA, 2009; pp. 45–52. [CrossRef]

3. Nayyar, S.K.; Mitsunaga, T. High dynamic range imaging: Spatially varying pixel exposures. In Proceedings of the 2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000), Hilton Head, SC, USA, 13–15 June 2000; Computer Society: Washington, DC, USA, 2000; pp. 1472–1479. [CrossRef]

4. Tumblin, J.; Agrawal, A.K.; Raskar, R. Why I want a gradient camera. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–26 June 2005; Computer Society: Washington, DC, USA, 2005; pp. 103–110. [CrossRef]

5. Reinhard, E.; Ward, G.; Pattanaik, S.N.; Debevec, P.E.; Heidrich, W. High Dynamic Range Imaging— Acquisition, Display, and Image-Based Lighting, 2nd ed.; Academic Press: Cambridge, MA, USA, 2010.

6. Granados, M.; Ajdin, B.; Wand, M.; Theobalt, C.; Seidel, H.; Lensch, H.P.A. Optimal HDR reconstruction with linear digital cameras. In Proceedings of the The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010; Computer Society: Washington, DC, USA, 2010; pp. 215–222. [CrossRef]

7. Yan, Q.; Sun, J.; Li, H.; Zhu, Y.; Zhang, Y. High dynamic range imaging by sparse representation. Neurocomputing 2017, 269, 160–169. [CrossRef]

8. Eden, A.; Uyttendaele, M.; Szeliski, R. Seamless image stitching of scenes with large motions and exposure differences. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, USA, 17–22 June 2006; IEEE Computer Society: Washington, DC, USA, 2006; pp. 2498–2505. [CrossRef]

9. Gallo, O.; Gelfandz, N.; Chen, W.C.; Tico, M.; Pulli, K. Artifact-free high dynamic range imaging. In Proceedings of the 2009 IEEE International Conference on Computational Photography (ICCP), San Francisco, CA, USA, 16–17 April 2009; pp. 1–7. [CrossRef]

10. Granados, M.; Kim, K.I.; Tompkin, J.; Theobalt, C. Automatic noise modeling for ghost-free HDR reconstruction. ACM Trans. Graph. 2013, 32, 1–10. [CrossRef]

11. Grosch, T. Fast and robust high dynamic range image generation with camera and object movement. In Proceedings of the Vision, Modeling and Visualization, RWTH Aachen, Aachen, Germany, 22–24 November 2006; pp. 277–284.

12. Min, T.; Park, R.; Chang, S. Histogram based ghost removal in high dynamic range images. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, New York, NY, USA, 28 June–2 July 2009; pp. 530–533. [CrossRef]

13. Pecce, F.; Kautz, J. Bitmap movement detection: HDR for dynamic scenes. J. Virtual Real. Broadcast. 2013, 10, 1–13.

14. Prabhakar, K.R.; Babu, R.V. Ghosting-free multi-exposure image fusion in gradient domain. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, 20–25 March 2016; pp. 1766–1770. [CrossRef]

15. Raman, S.; Chaudhuri, S. Reconstruction of high contrast images for dynamic scenes. Vis. Comput. 2011, 27, 1099–1114. [CrossRef]

16. Wu, S.; Xie, S.; Rahardja, S.; Li, Z. A robust and fast anti-ghosting algorithm for high dynamic range imaging. In Proceedings of the International Conference on Image Processing, ICIP 2010, Hong Kong, China, 26–29 September 2010; pp. 397–400. [CrossRef]

17. Zhang, W.; Cham, W. Reference-guided exposure fusion in dynamic scenes. J. Vis. Commun. Image Represent. 2012, 23, 467–475. [CrossRef]

18. Tomaszewskas, A.; Mantiuk, R. Image Registration for Multi-Exposure High Dynamic Range Image Acquisition. In Proceedings of the 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2007, Bory, Czech Republic, 29 January–1 February 2007; pp. 49–56.

19. Ward, G. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. J. Graph. G实务 Game Tools 2003, 8, 17–30. [CrossRef]

20. Gallo, O.; Troccoli, A.J.; Hu, J.; Pulli, K.; Kautz, J. Locally non-rigid registration for mobile HDR photography. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 48–55. [CrossRef]

21. Hu, J.; Gallo, O.; Pulli, K. Exposure stacks of live scenes with hand-held cameras. In Proceedings of the Computer Vision—ECCV 2012—12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Lecture Notes in Computer Science; Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7572, pp. 499–512. [CrossRef]
22. Kang, S.B.; Uyttendaele, M.; Winder, S.A.J.; Szeliski, R. High dynamic range video. *ACM Trans. Graph.* 2003, 22, 319–325. [CrossRef]
23. Zimmer, H.; Bruhn, A.; Weickert, J. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Comput. Graph. Forum* 2011, 30, 405–414. [CrossRef]
24. Hu, J.; Gallo, O.; Pulli, K.; Sun, X. HDR deghosting: How to deal with saturation? In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE Computer Society, 2013; pp. 1163–1170. [CrossRef]
25. Sen, P.; Kalantari, N.K.; Yasesoubi, M.; Darabi, S.; Goldman, D.B.; Shechtman, E. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.* 2012, 31, 1–11. [CrossRef]
26. Yan, Q.; Gong, D.; Shi, Q.; van den Hengel, A.; Shen, C.; Reid, I.D.; Zhang, Y. Attention-guided network for ghost-free high dynamic range imaging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 1751–1760. [CrossRef]
27. Liu, Z.; Lin, W.; Li, X.; Rao, Q.; Jiang, T.; Han, M.; Fan, H.; Sun, J.; Liu, S. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, 19–25 June 2021; pp. 463–470. [CrossRef]
28. Kalantari, N.K.; Ramamoorthi, R. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* 2017, 36, 1–12. [CrossRef]
29. Wu, S.; Xu, J.; Tai, Y.; Tang, C. Deep high dynamic range imaging with large foreground motions. In Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018; Proceedings, Part II; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2018, Volume 11206, pp. 120–135. [CrossRef]
30. Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.; Zhang, Y. Deep HDR imaging via a non-local network. *IEEE Trans. Image Process.* 2020, 29, 4308–4322. [CrossRef] [PubMed]
31. Ma, K.; Duanmu, Z.; Zhu, H.; Fang, Y.; Wang, Z. Deep guided learning for fast multi-exposure image fusion. *IEEE Trans. Image Process.* 2020, 29, 2808–2819. [CrossRef] [PubMed]
32. Tursun, O.T.; Akyüz, A.O.; Erdem, A.; Erdem, E. The state of the art in HDR deghosting: A survey and evaluation. *Comput. Graph. Forum* 2015, 34, 683–707. [CrossRef]
33. Yan, Q.; Zhu, Y.; Zhang, Y. Robust artifact-free high dynamic range imaging of dynamic scenes. *Multim. Tools Appl.* 2019, 78, 11487–11505. [CrossRef]
34. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision;* Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
35. Baker, S.; Scharstein, D.; Lewis, J.P.; Roth, S.; Black, M.J.; Szeliski, R. A database and evaluation methodology for optical flow. *Int. J. Comput. Vis.* 2011, 92, 1–31. [CrossRef]
36. Szpak, Z.L.; Chojnacki, W.; Eriksson, A.P.; van den Hengel, A. Sampson distance based joint estimation of multiple monographs with uncalibrated cameras. *Comput. Vis. Image Underst.* 2014, 125, 200–213. [CrossRef]
37. Szpak, Z.L.; Chojnacki, W.; van den Hengel, A. Robust multiple homography estimation: An ill-solved problem. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2132–2141.
38. Yan, Q.; Gong, D.; Zhang, P.; Shi, Q.; Sun, J.; Reid, I.D.; Zhang, Y. Multi-scale dense networks for deep high dynamic range imaging. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, 7–11 January 2019; pp. 41–50. [CrossRef]
39. Yan, Q.; Gong, D.; Shi, Q.J.; van den Hengel, A.; Shen, C.; Reid, I.D.; Zhang, Y. Dual-attention-guided network for ghost-free high dynamic range imaging. *Int. J. Comput. Vis.* 2022, 130, 76–94. [CrossRef]
40. Madden, B.C. *Extended Intensity Range Imaging;* Technical Reports (CIS); University of Pennsylvania, Philadelphia, PA, USA, 1993; p. 248. [CrossRef]
41. Mann, S.; Picard, R. *Beingundigital'with Digital Cameras;* MIT Media Lab Perceptual: Cambridge, MA, USA, 1994; Volume 1, p. 2. [CrossRef]
42. Debevec, P.E.; Malik, J. Recovering high dynamic range radiance maps from photographs. In Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, Los Angeles, CA, USA, 3–8 August 1997; ACM Press/Addison-Wesley Publishing Co.: New York, NY, USA; pp. 369–378. [CrossRef]
43. Kao, W.; Hsu, C.; Chen, L.; Kao, C.; Chen, S. Integrating image fusion and motion stabilization for capturing still images in high dynamic range scenes. *IEEE Trans. Consumer Electron.* 2006, 52, 735–741. [CrossRef]
44. SidiBé, D.; Puech, W.; Strauss, O. Ghost detection and removal in high dynamic range images. In Proceedings of the 17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, Scotland, UK, 24–28 August 2009; pp. 2240–2244.
45. Khan, E.A.; Akyüz, A.O.; Reinhard, E. Ghost removal in high dynamic range images. In Proceedings of the International Conference on Image Processing, ICIP 2006, Atlanta, GA, USA, 8–11 October 2006; pp. 2005–2008. [CrossRef]
46. Bogoni, L. Extending dynamic range of monochrome and color images through fusion. In Proceedings of the 15th International Conference on Pattern Recognition, ICPR-2000, Barcelona, Spain, 3–7 September 2000; IEEE Computer Society, 2000; pp. 3007–3016. [CrossRef]
47. Ma, K.; Li, H.; Yong, H.; Wang, Z.; Meng, D.; Zhang, L. Robust multi-exposure image fusion: A structural patch decomposition approach. *IEEE Trans. Image Process.* 2017, 26, 2519–2532. [CrossRef]
48. Yu, G.; Zhang, J.; Ma, Z.; Wang, H. Efficient Progressive High Dynamic Range Image Restoration via Attention and Alignment Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, 19–20 June 2022; pp. 1123–1130. [CrossRef]

49. Tang, H.; Bai, S.; Sebe, N. Dual Attention GANs for Semantic Image Synthesis. In Proceedings of the MM '20: The 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R., Eds.; ACM: New York, NY, USA, 2020; pp. 1994–2002. [CrossRef]

50. Lv, Z.; Li, X.; Niu, Z.; Cao, B.; Zuo, W. Semantic-shape Adaptive Feature Modulation for Semantic Image Synthesis. arXiv 2022, arXiv:2203.16898.

51. Sun, W.; Wu, T. Learning Spatial Pyramid Attentive Pooling in Image Synthesis and Image-to-Image Translation. arXiv 2019, arXiv:1901.06322.

52. Ren, Y.; Wu, Y.; Li, T.H.; Liu, S.; Li, G. Combining Attention with Flow for Person Image Synthesis. In Proceedings of the MM ’21: ACM Multimedia Conference, Virtual Event, 20–24 October 2021; Shen, H.T., Zhuang, Y., Smith, J.R., Yang, Y., Cesar, P., Metze, F., Prabhakaran, B., Eds.; ACM: New York, NY, USA, 2021; pp. 3737–3745. [CrossRef]

53. Zheng, H.; Liao, H.; Chen, L.; Xiong, W.; Chen, T.; Luo, J. Example-Guided Image Synthesis Using Masked Spatial-Channel Attention and Self-supervision. In Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV; Lecture Notes in Computer Science; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds., Springer: Berlin/Heidelberg, Germany, 2020; Volume 12359, pp. 422–439. [CrossRef]

54. Uddin, S.M.N.; Jung, Y.J. SIFNet: Free-form image inpainting using color split-inpaint-fuse approach. Comput. Vis. Image Underst. 2022, 221, 103446. [CrossRef]

55. Wang, N.; Li, J.; Zhang, L.; Du, B. MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; Kraus, S., Ed.; pp. 3748–3754. [CrossRef]

56. Qin, J.; Bai, H.; Zhao, Y. Multi-scale attention network for image inpainting. Comput. Vis. Image Underst. 2021, 204, 103155. [CrossRef]

57. Wang, X.; Chen, Y.; Yamashiki, T. Spatially adaptive multi-scale contextual attention for image inpainting. Multim. Tools Appl. 2022, 81, 31831–31846. [CrossRef]

58. Xie, B.; Yang, Z.; Yang, L.; Luo, R.; Wei, A.; Weng, X.; Li, B. Multi-Scale Fusion with Matching Attention Model: A Novel Decoding Network Cooperated with NAS for Real-Time Semantic Segmentation. IEEE Trans. Intell. Transp. Syst. 2022, 23, 12622–12632. [CrossRef]

59. Wang, W.; Wang, S.; Li, Y.; Jin, Y. Adaptive multi-scale dual attention network for semantic segmentation. Neurocomputing 2021, 460, 39–49. [CrossRef]

60. Chen, J.; Tian, Y.; Ma, W.; Mao, Z.; Hu, Y. Scale channel attention network for image segmentation. Multim. Tools Appl. 2021, 80, 16473–16489. [CrossRef]

61. Sagar, A.; Soundrapandiyan, R. Semantic Segmentation with Multi Scale Spatial Attention For Self Driving Cars. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021; pp. 2650–2656. [CrossRef]

62. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. arXiv 2020, arXiv:2005.10821.

63. Liu, M.; Zhang, C.; Zhang, Z. Multi-Scale Deep Convolutional Nets with Attention Model and Conditional Random Fields for Semantic Image Segmentation. In Proceedings of the 2nd International Conference on Signal Processing and Machine Learning, SPML 2019, Hangzhou, China, 27–29 November 2019; ACM: New York, NY, USA, 2019; pp. 73–78. [CrossRef]

64. Yang, S.; Peng, G. Attention to Refine Through Multi Scales for Semantic Segmentation. In Proceedings of the Advances in Multimedia Information Processing—PCM 2018—19th Pacific-Rim Conference on Multimedia, Hefei, China, 21–22 September 2018; Proceedings, Part II; Lecture Notes in Computer Science; Hong, R., Cheng, W., Yamashiki, T., Wang, M., Ngo, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11165, pp. 232–241. [CrossRef]

65. Zhang, M.; Wang, H.; Zhang, Z.; Chen, Z.; Shen, J. Lightweight Multi-Scale Asymmetric Attention Network for Image Super-Resolution. Micromachines 2021, 13, 54. [CrossRef]

66. Mehta, N.; Murala, S. MSAR-Net: Multi-scale attention based light-weight image super-resolution. Pattern Recognit. Lett. 2021, 151, 215–221. [CrossRef]

67. Pang, S.; Chen, Z.; Yin, F. Lightweight multi-scale aggregated residual networks for image super-resolution. Multim. Tools Appl. 2022, 81, 4797–4819. [CrossRef]

68. Sun, Y.; Qin, J.; Gao, X.; Chai, S.; Chen, B. Attention-enhanced multi-scale residual network for single image super-resolution. Signal Image Video Process. 2022, 16, 1417–1424. [CrossRef]

69. Li, W.; Li, J.; Li, J.; Huang, Z.; Zhou, D. A lightweight multi-scale channel attention network for image super-resolution. Neurocomputing 2021, 456, 327–337. [CrossRef]

70. Wang, L.; Shen, J.; Tang, E.; Zheng, S.; Xu, L. Multi-scale attention network for image super-resolution. J. Vis. Commun. Image Represent. 2021, 80, 103300. [CrossRef]

71. Liu, H.; Cao, F.; Wen, C.; Zhang, Q. Lightweight multi-scale residual networks with attention for image super-resolution. Knowl. Based Syst. 2020, 203, 106103. [CrossRef]
72. Xiong, C.; Shi, X.; Gao, Z.; Wang, G. Attention augmented multi-scale network for single image super-resolution. *Appl. Intell.* 2021, 51, 935–951. [CrossRef]
73. Soh, J.W.; Cho, N.I. Lightweight Single Image Super-Resolution with Multi-Scale Spatial Attention Networks. *IEEE Access* 2020, 8, 35383–35391. [CrossRef]
74. Cao, F.; Liu, H. Single image super-resolution via multi-scale residual channel attention network. *Neurocomputing* 2019, 358, 424–436. [CrossRef]
75. Zhang, X.; Wang, X. MARN: Multi-Scale Attention Retinex Network for Low-Light Image Enhancement. *IEEE Access* 2021, 9, 50939–50948. [CrossRef]
76. Choi, S.; Cho, J.; Song, W.; Choe, J.; Yoo, J.; Sohn, K. Pyramid inter-attention for high dynamic range imaging. *Sensors* 2020, 20, 5102. [CrossRef]
77. Chen, J.; Yang, Z.; Chan, T.N.; Li, H.; Hou, J.; Chau, L. Attention-Guided Progressive Neural Texture Fusion for High Dynamic Range Image Restoration. *IEEE Trans. Image Process.* 2022, 31, 2661–2672. [CrossRef][PubMed]
78. Xiao, J.; Ye, Q.; Liu, T.; Zhang, C.; Lam, K. Multi-scale Sampling and Aggregation Network For High Dynamic Range Imaging. *arXiv* 2022, arXiv:2208.02448. https://doi.org/10.48550/arXiv.2208.02448.
79. Ye, Q.; Suganuma, M.; Xiao, J.; Okatani, T. Learning Regularized Multi-Scale Feature Flow for High Dynamic Range Imaging. *arXiv* 2022, arXiv:2207.02539. https://doi.org/10.48550/arXiv.2207.02539.
80. Lee, J.; Shin, J.; Lim, H.; Paik, J. Deep High Dynamic Range Imaging without Motion Artifacts Using Global and Local Skip Connections. In Proceedings of the IEEE International Conference on Consumer Electronics, ICCE 2022, Las Vegas, NV, USA, 7–9 January 2022; pp. 1–2. [CrossRef]
81. Hu, J.; Wu, L.; Li, N. High dynamic range imaging with short- and long-exposures based on artificial remapping using multiscale exposure fusion. *J. Vis. Commun. Image Represent.* 2022, 87, 103585. [CrossRef]
82. Niu, Y.; Wu, J.; Liu, W.; Guo, W.; Lau, R.W.H. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Trans. Image Process.* 2021, 30, 3885–3896. [CrossRef][PubMed]
83. Uddin, S.M.N.; Jung, Y.J. Global and Local Attention-Based Free-Form Image Inpainting. *Sensors* 2020, 20, 3204. [CrossRef][PubMed]
84. Nadim Uddin, S.M.; Ahmed, S.H.; Jung, Y.J. Unsupervised Deep Event Stereo for Depth Estimation. *IEEE Trans. Circuits Syst. Video Technol.* 2022, 1. [CrossRef]
85. Ahmed, S.H.; Jang, H.W.; Uddin, S.M.N.; Jung, Y.J. Deep event stereo leveraged by event-to-image translation. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, AAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021; AAAI Press: Palo Alto, CA, USA, 2021; pp. 882–890. [CrossRef]
86. Jang, H.W.; Jung, Y.J. Deep color transfer for color-plus-mono dual cameras. *Sensors* 2020, 20, 2743. [CrossRef]
87. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481. [CrossRef]
88. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 593–602. [CrossRef]
89. Wang, X.; Girshick, R.B.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803. [CrossRef]
90. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
91. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
92. Wang, X.; Chan, K.C.K.; Yu, K.; Dong, C.; Loy, C.C. EDVR: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 1954–1963. [CrossRef]
93. Zheng, J.; Li, Z.; Zhu, Z.; Wu, S.; Rahardja, S. Hybrid Patching for a Sequence of Differently Exposed Images with Moving Objects. *IEEE Trans. Image Process.* 2013, 22, 5190–5201. [CrossRef]
94. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, 3–7 May 2021.