electronics

MDPI

Article

Leveraging Feature Extraction and Context Information for Image Relighting

Chenrong Fang [1], Ju Wang [1], Kan Chen [2], Ran Su [1], Chi-Fu Lai [3] and Qian Sun [4,*]

[1] College of Intelligence and Computing, Tianjin University, Tianjin 300072, China; 2021244051@tju.edu.cn (C.F.); wangju@nscc-tj.cn (J.W.); ran.su@tju.edu.cn (R.S.)
[2] Infocomm Technology Cluster, Singapore Institute of Technology, Singapore 138683, Singapore; kan.chen@singaporetech.edu.sg
[3] School of Arts and Social Sciences, Hong Kong Metropolitan University, Hong Kong, China; wcflai@hkmu.edu.hk
[4] School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: sunqian@nuist.edu.cn

Abstract: Example-based image relighting aims to relight an input image to follow the lighting settings of another target example image. Deep learning-based methods for such tasks have become highly popular. However, they are often limited by the geometric priors or suffer from shadow reconstruction and lack of texture details. In this paper, we propose an image-to-image translation network called *DGATRN* to tackle this problem by enhancing feature extraction and unveiling context information to achieve visually plausible example-based image relighting. Specifically, the proposed *DGATRN* consists of a scene extraction, a shadow calibration, and a rendering network, and our key contribution lies in the first two networks. We propose an up- and downsampling approach to improve the feature extraction capability to capture scene and texture details better. We also introduce a feature attention downsampling block and a knowledge transfer to utilize the attention impact and underlying knowledge connection between scene and shadow. Experiments were conducted to evaluate the usefulness and effectiveness of the proposed method.

Keywords: image relighting; upsampling and downsampling; attention; knowledge transfer; neural network

Citation: Fang, C.; Wang, J.; Chen, K.; Su, R.; Lai, C.-F.; Sun, Q. Leveraging Feature Extraction and Context Information for Image Relighting. *Electronics* **2023**, *12*, 4301. https://doi.org/10.3390/electronics12204301

Academic Editor: Fernando De la Prieta Pintado

Received: 17 September 2023
Revised: 8 October 2023
Accepted: 10 October 2023
Published: 17 October 2023

1. Introduction

Example-based image relighting is an important topic in computer vision and graphics. It renders the scene of the input image under new lighting conditions of a given example target image to generate an output. Image relighting has several practical applications and can significantly impact various fields, such as Virtual Reality (VR), Augmented Reality (AR), product visualization, forensics and surveillance, and art restoration and conservation. Specifically, it can create more realistic and immersive environments by adapting virtual lighting to match real-world lighting conditions, enhancing the sense of presence and realism in virtual experiences. In product visualization, image relighting can be applied to product images on e-commerce websites, allowing customers to view products under different lighting conditions and assess their appearance accurately. Furthermore, in the fields of forensics and surveillance, image relighting can enhance security camera footage by adjusting lighting conditions, making it easier to identify suspects or details in low-light or overexposed situations.

In general, scene information is essential for relighting tasks. The conventional relighting methods usually obtain scene information through 3D scene reconstruction. However, this information obtaining process [1] is not easy; it often requires tedious efforts, high computational costs and special acquisition equipment to obtain geometry details, surface

properties, and lighting conditions. Recently, deep learning [2] technology has proven to be successful for a variety of computer vision and graphics tasks, such as neural rendering [3], inverse rendering [4], and scene reconstruction [5]. Introducing deep learning technology to image relighting can reduce the 3D reconstruction efforts while achieving visually pleasing results. The common approaches are leveraging geometric priors and image-to-image translation.

By leveraging scene information such as architectural, geometric, depth, and facial priors, convincing relighting effects under multiple views can be achieved. These methods largely rely on those scene priors and lack the capability of generalization; that is, they are limited by the priors and are mainly suitable for the respective specific building, object, scene, or face scenarios.

Image-to-image translation relighting technology aims to establish an end-to-end rendering process and minimize the dependency on unnecessary scene information that achieves image relighting effects by learning a mapping between the input and output. However, it is still challenging to generate visually appealing results, especially for texture details and shadows. This is mainly because the real-world lighting conditions are more complex and fine, so it is difficult to effectively remove shadows and reconstruct texture details from the original image, which leads to the degrading of realism. The image-to-image translation methods mainly focus on the end image generation. The scene information, especially the features and embedded context information of the features, may not be adequately extracted and used. Moreover, the scene and its shadow information are closely allied; however, the underlying knowledge connection between them is underutilized.

In this paper, we propose an image-to-image translation network called *DGATRN* for image relighting. The proposed *DGATRN* consists of a scene feature extraction, a shadow calibration, and a rendering network; it has the following features:

1. In order to improve the feature extraction capability, we propose new up- and downsampling blocks for a back-projection-based network by integrating dense and global residual blocks for scene and shadow features, respectively.
2. In order to better reconstruct shadows and preserve texture details, we propose a context-aware approach by utilizing spatial information of features and underlying knowledge connection. We introduce a feature attention downsampling module that combines channel attention and spatial attention, as well as introduce a knowledge transfer from the scene extraction to the shadow calibration based on L1 Loss.
3. The proposed techniques are integrated as *DGATRN*. We evaluate it on the representative datasets VIDIT to showcase that *DGATRN* can achieve convincing results.

2. Related Work

2.1. Reconstruction-Based Image Relighting

Image relighting approaches often focused on the reconstruction of the scene or the physical aspects of the problem [1,6,7], neural rendering [3], neural inverse rendering [4], and scene reconstruction [5,8]. While these methods create physically realistic models for relighting, they frequently rely on explicit geometry, illumination parameters, or multiview datasets, which may impede their performance.

2.2. Image-to-Image Deep Learning Image Relighting

Isola et al. [2] achieved significant success in many different types of image-to-image transformation tasks using the Pix2Pix method, which uses a U-Net type network structure. Wei et al. [9] studied Retinex-Net, which is based on the Retinex theory for low-light images. It first decomposes the low-light image into reflectance and illumination elements and then adjusts sub-networks to refine the illumination to light up the input image. Xu et al. [10] used a U-Net network to approximate the light transmission function in the image, thereby achieving the relighting process. Their relighting method can achieve good results with only five reference images, but it is still a relighting method for a specific scene. Wang et al. [11] proposed a scene relighting method called DRN based on the VIDIT

datasets [12]. This method consists of three sub-networks that complete the tasks of scene reconstruction, shadow prior estimation, and re-rendering. Both the scene reconstruction and shadow estimation sub-networks use a structure similar to the U-Net network. Based on DRN, Wang et al. [13] proposed MCN, which used a new downsampling feature self-calibration block (DFSB) and upsampling feature self-calibration block (UFSB) as the basic module units for the feature encoder and decoder in the scene reconstruction and shadow estimation tasks. In addition, Yang et al. [14] proposed S3Net, a single-stream network architecture that uses a depth map to guide the relighting task using attention modules and enhancement modules.

Recent results such as [15,16] can achieve convincing image relighting results; however, preserving texture and shadow details is still challenging. Some feature details and context information are underutilized; these can be helpful in improving the relighting results. Our proposed *DGATRN* considers feature extraction and feature space information on the basis of the methods DRN [11], MCN [13], and S3Net [14].

3. Overview

For input image X under lighting condition, L_Φ, we would like to generate relit output image Y under the target lighting condition L_Ψ. Inspired by [11,13], according to Retinex theory, the intrinsic scene information (S) would not change under different lighting conditions. Thus the relighting problem can be formulated based on image-to-image translation:

$$Y = R(L_\Phi^{-1}(X), L_{\Phi\to\Psi}(X)), \tag{1}$$

where $L_\Phi^{-1}(X)$ is the scene S, and it represents the scene extraction operation. $L_{\Phi\to\Psi}(X)$ represents the lighting condition transferring operation that migrates lighting condition L_Φ to L_Ψ. This is approximated by the shadow calibration network in *DGATRN*. R represents the rendering operation using the extracted scene and the new lighting condition to generate the final output. That is, *DGATRN* consists of three operations: scene extraction, shadow calibration, and rendering networks (Figure 1). The detailed structures of the networks are as shown in Figure 2:

1. The scene structure extraction network is a GAN network. The generator consists of a 7×7 convolutional layer combined with a feature attention downsampling module (Section 5.1), four upsampling blocks, a residual block, four downsampling blocks, and a 3×3 convolutional layer. The up- and downsampling blocks will be discussed in Section 4. The scene extraction network also uses skip connections to fuse the feature information of the first, second, third, and fourth upsampling blocks together, and to fuse the feature information outputs by the 7×7 and 3×3 convolutional layers. The discriminator structure is the same as PatchGAN [17].
2. The shadow calibration network is similar to the scene extraction network. The main differences are that (1) it removes skip connections, which helps the shadow calibration network focus on global lighting effects; (2) the loss functions and discriminators are different (discussed later); and (3) it accepts the knowledge transferred from the scene structure extraction network (Section 5.2).
3. The rendering synthesis network uses a multiscale perception structure. First, the input information of the rendering synthesis network is inputted into convolutional layers with different kernel sizes $N \times N$ ($N = 2i + 1, i \in [1, 12]$) to extract feature information of different scales. Large feature scales help to obtain global lighting information about the scene, and small feature scales help to obtain local texture information about the scene. The feature information of different scales is fused together, and then the fused feature information is processed by average pooling, fully connected layers, and activation layers. This part mainly learns the weights of different channels of the features. The obtained weights and fused feature information are multiplied and added element-wise and processed by convolutional layers to output the relit image.

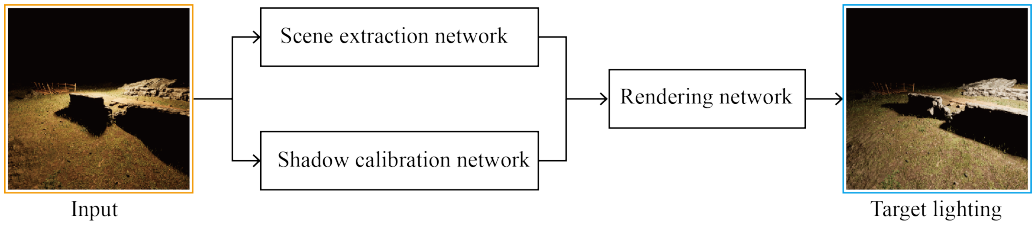

Figure 1. Three module structure of *DGATRN*. The input image is fed into the scene extraction network and the shadow calibration network separately, resulting in output results. After combining the outputs of these two networks, they are input into the rendering and compositing network to obtain the final relit image.

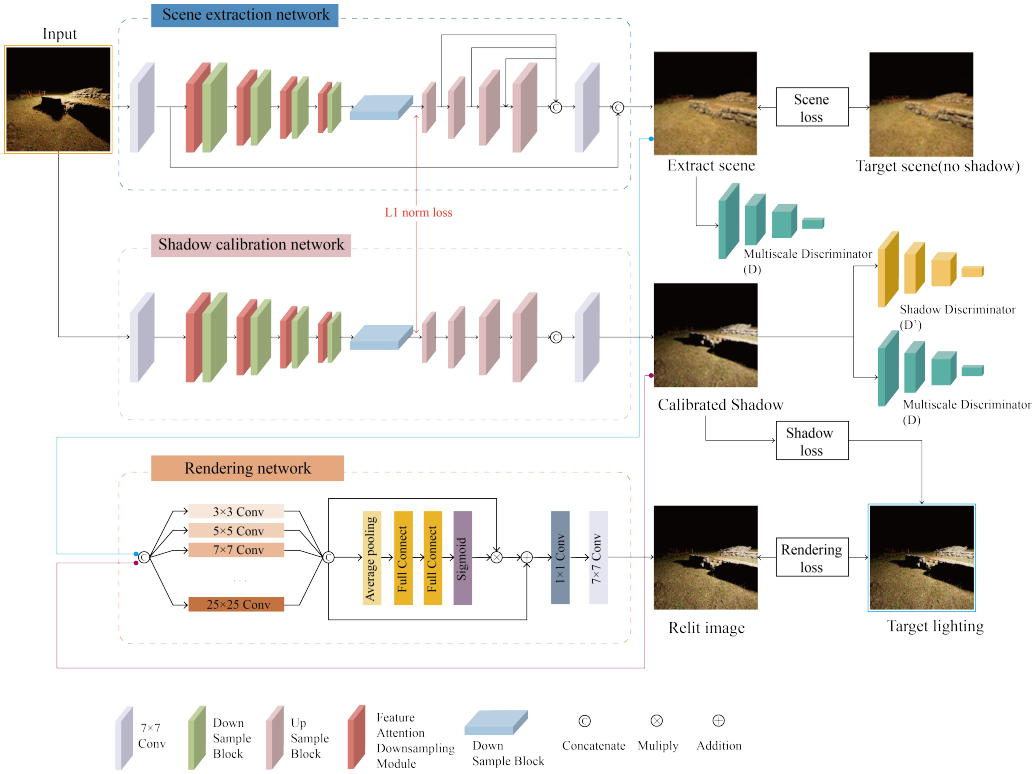

Figure 2. The overall structure of *DGATRN*. The scene extraction network (**top**) is trained first and is followed by the shadow estimation network (**middle**) with the same input. Their outputs, the extracted scene, and calibrated shadow images are used as the input for the rendering network (**bottom**). The shadow estimation and rendering networks use the same 'target lighting' image, which is the given example for lighting the target image. Its shadows are removed for use as the target image ('target scene') for the scene extraction network.

Note that the scene extraction network is trained first and is followed by the shadow estimation network. Their results are used for training the rendering network. As mentioned in Section 5.2, in order to let the shadows be aware of the contextual information from the scene, the knowledge of the scene extraction network is transferred to the intermediate feature map of the shadow calibration network through an L1 norm loss function.

Loss Functions

For the scene structure extraction network, there is no explicit geometric prior information in the datasets. Thus, we adopt the exposure fusion method [18] as in DRN [11] to obtain the no-shadow image of the input, as the target image, namely $Y_{noShadow}$. The loss

function for generating a no-shadow image (G) is based on the L1 norm and minimization of the sum of the per-pixel absolute difference.

$$\begin{array}{l} L_G(X, Y_{noShadow}) \\ = E(\|Y_{noShadow} - G(X)\|) \\ = \sum_{i=1}^{N} \mid Y^i_{noShadow} - G(X^i) \mid, \end{array} \tag{2}$$

where $X^i, i = 1, \ldots, N$ represents the ith of pixel in X (N pixels in total). The adversarial loss function with generator D and discriminator G is

$$\begin{array}{l} min_G max_D L_{Adv}(G, D) \\ = \sum_{i=1}^{N} log D(Y^i_{noShadow}) \\ + \sum_{i=1}^{N} log(1 - D(G(X^i))). \end{array} \tag{3}$$

It is combined with L_G as the loss function for the scene extraction network:

$$\begin{array}{l} L_{scene} = \gamma L_G(X, Y_{noShadow}) \\ + (1 - \gamma) min_G max_D L_{Adv}(G, D), \end{array} \tag{4}$$

where γ is the weight to balance the two losses. In this shadow calibration network, the loss function will incorporate another loss with a shadow discriminator:

$$\begin{array}{l} L_{shadow} = \gamma_1 L_G(X, Y) \\ + \gamma_2 min_G max_D L_{Adv}(G, D) \\ + \gamma_3 min_G max_{D'} L_{Adv}(G, D'), \end{array} \tag{5}$$

where D' is the shadow discriminator with the focus on low-intensity regions. It is realized by applying low-pass filtering on the generated pixel intensity x, using $min(\alpha, x)$, where α is a shadow sensitivity threshold (0.059 in our experiments). Similarly, γ_1, γ_2, and γ_3 are weights to balance the losses and they sum to one. The loss function of the rendering network combines the L1 norm loss and perceptual loss using the features using the pre-trained VGG-19 network $feat$:

$$\begin{array}{l} L_{render} = \gamma_r(\|Y - Y'\|) \\ + (1 - \gamma_r)(\|feat(Y) - feat(Y')\|), \end{array} \tag{6}$$

where Y is the final target, Y' is the output from the render network, and γ_r balances the losses.

4. Feature Extraction Improvement

The up- and downsampling blocks use a sequence of encoding/decoding operations for input and latent information mapping. We adopt a similar up- and downsampling back projection structure as MSN. In *DGATRN* (Figure 2), the scene extraction aims to capture more scene details. Thus, we incorporate dense residual block (DR_Block) to facilitate better extraction of scene feature details. On the other hand, shadow calibration requires paying more attention to the influence of the overall scene on the shadow features. Thus, we introduce global residual block (GR_Block), which helps to better capture the holistic scene information to improve shadow feature extraction. We also improve its performance by using LeakyReLu in place of ReLu (with α as 0.01). This adjustment enhances the feature learning ability.

4.1. Dense Residual Block and Global Residual Block

DR_Block is inspired by DenseNet [19] as shown in Figure 3. Each layer (to compute features X_n) connects to the previous $n - 1$ layers and incorporates their results. The computed X_n is then added with weight λ_{DR} element-wise to the input features X to obtain the output features Y. It enables the network to effectively propagate and retain features.

In this paper, we use a sequence of 3×3 convolutional layers with a stride of 2, λ_{DR} (the weight of X_n) is set as 0.1, and n is set to 2 due to its frequent usage.

GR_Block is inspired by OIDDR-Net [20]. As shown in Figure 4, it uses a multiscale feature extraction method to guide the network to optimize the feature map while maintaining the correspondence between the image and the input features. First, the input X is connected to a 3×3 convolutional layer with a stride of 2 to extract features, then a pooling layer is applied for dimension reduction from (n, c, h, w) to (n, c). The output is then added to a fully connected layer to further extract features. Next, they are upsampled from (n, c) to $(n, c, 1, 1)$, and the obtained features are multiplied by the input features X. The resulting output is then inputted to a 3×3 convolutional layer with a stride of 2 and added element-wise (with weights $\lambda_{GR} = 0.1$) with the input feature X to obtain the output feature Y of the GR_Block.

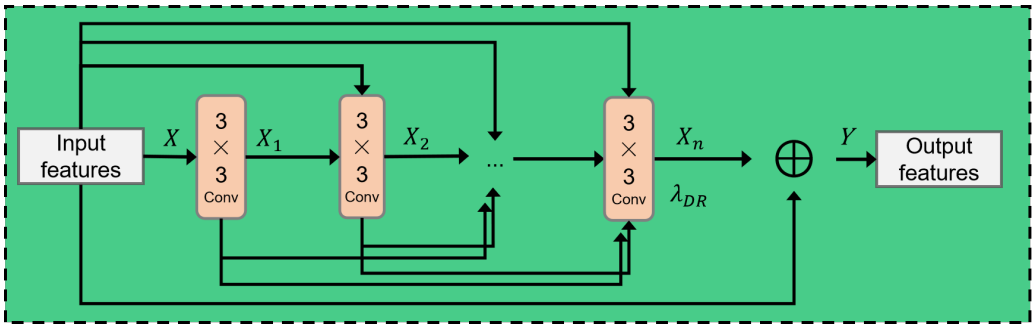

Figure 3. The structure of DR_Block.

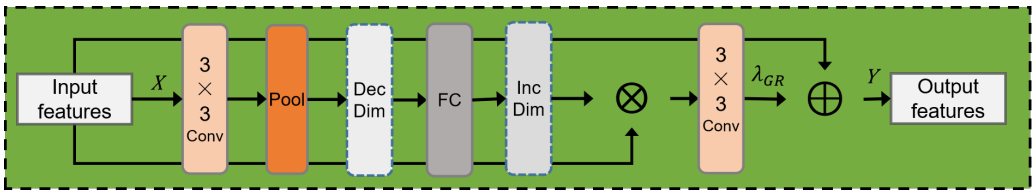

Figure 4. The structure of GR_Block.

4.2. Up- and Downsampling Blocks

We use the downsampling block of the scene extraction network as an example to present our method (shown at the top of Figure 5). The downsampling block first passes the input features into a 3×3 convolutional layer with a stride of 2, mapping them to a smaller scale as F_{small}. Then, the F_{small} is inputted into a 4×4 transpose convolutional layer with a stride of 2 and mapped back to the input scale space F_{normal}. Meanwhile, another branch generates calibration weight w_1 by passing the input features through a DB_Block and a LeakyReLU function. w_1 calibrates F_{normal} by element-wise multiplication as the calibrated features $F_{normal}^{w_1}$. They are then remapped to the small-scale space by a second 3×3 convolutional layer with a stride of 2, obtaining the feature $F_{small}^{w_1}$. F_{small} will be also put into a branch that includes a DB_Block to obtain calibration weight w_2. The calibrated weights w_1 and w_2 are applied and added to obtain the output features F_{out}, which are half-sized features. This can be represented as follows:

$$F_{out} = Conv_3^2(w_1 DeConv_4(Conv_3^1(F))) + w_2, \tag{7}$$

where Con_3^1 and Con_3^2 represent two different 3×3 convolutional layers, $DeCon_4$represents a 4×4 transpose convolutional layer, and the w_1 and w_2 are the new calibration weights. They can be defined as

$$\begin{aligned} w_1 &= LeakyReLu(DR\_Block(F)) \\ w_2 &= DR\_Block(Conv_3^1(F)). \end{aligned} \tag{8}$$

Likewise, the upsampling block of the scene extraction network (shown at the bottom of Figure 5) with the output features F_{out} double-sized is defined as follows:

$$F_{out} = DeConv_4^2(w_1 Conv_3(DeConv_4^1(F))) + w_2. \quad (9)$$

The scene extraction network focuses more on the detailed information of features, while the shadow calibration network needs to consider the overall impact of global lighting, so it pays more attention to the global information of features. Thus GR_Block learns global information and calculates calibration weights in the shadow calibration network. As shown in Figure 5, the up- and downsampling structure in the shadow calibration network follows the same network structure as in the scene extraction network. The difference lies in the calculation of the two calibration weights; in Equation (8), *DR_Block* is replaced with *GR_Block*.

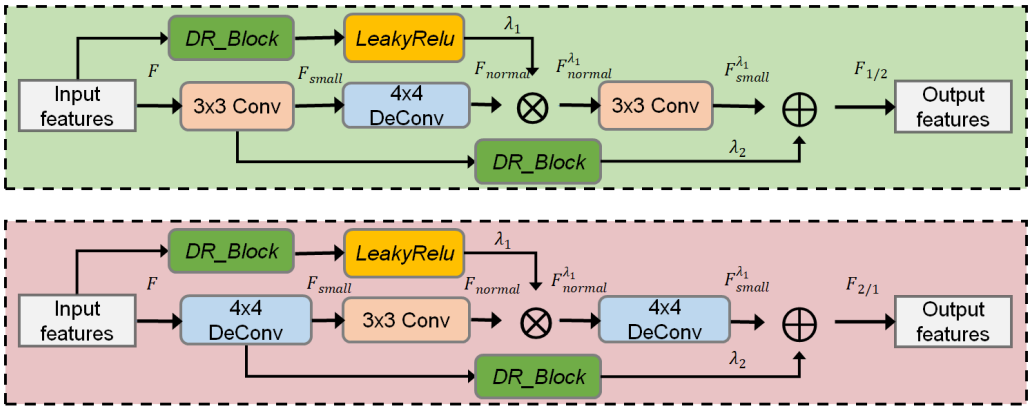

Figure 5. The down- (**top**) and upsampling (**bottom**) blocks for the scene extraction network. The shadow calibration network uses a similar structure with DR_Block replaced by GR_Block.

5. Attention and Knowledge Transfer

5.1. Feature Attention Downsampling

To consider the importance of the spatial information associated with the features, inspired by S3Net [14], we introduce the feature attention downsampling module (*FADM*) as the attention emphasizes the contextual information. As shown in Figure 6, the FADM combines a channel attention block (CA) and a pixel attention block (PA) to form a network module [21].

This is introduced to both the scene extraction and shadow calibration networks. The scene information needs to be preserved; moreover, the shadow calibration network needs to change the color temperature and lighting direction of the image while preserving the inherent scene information of the image.

FADM is added before the downsampling block. Since the output feature resolution of the network's shallow layers is higher, it contains more detail and positional information and is closer to the input image [22]. The addition of attention in *DGATRN* is shown in Figure 2. We also conducted an ablation study to verify this, which is referred to in the next Section.

5.2. Knowledge Transfer

Shadows are closely correlated to the scene information. Therefore, it is beneficial to transfer and fuse the knowledge of the scene extraction network and the shadow correction network. Conceptually, we would like to let the shadows be aware of the contextual information from the scene. This paper refers to the design ideas of KTDN [12].

When training the shadow calibration network, the knowledge of the scene extraction network is transferred to the intermediate feature map of the shadow calibration network through an L1 norm loss function, in order to help the shadow correction network utilize

the prior knowledge of the trained scene extraction network to help to reconstruct shadows under the target lighting conditions.

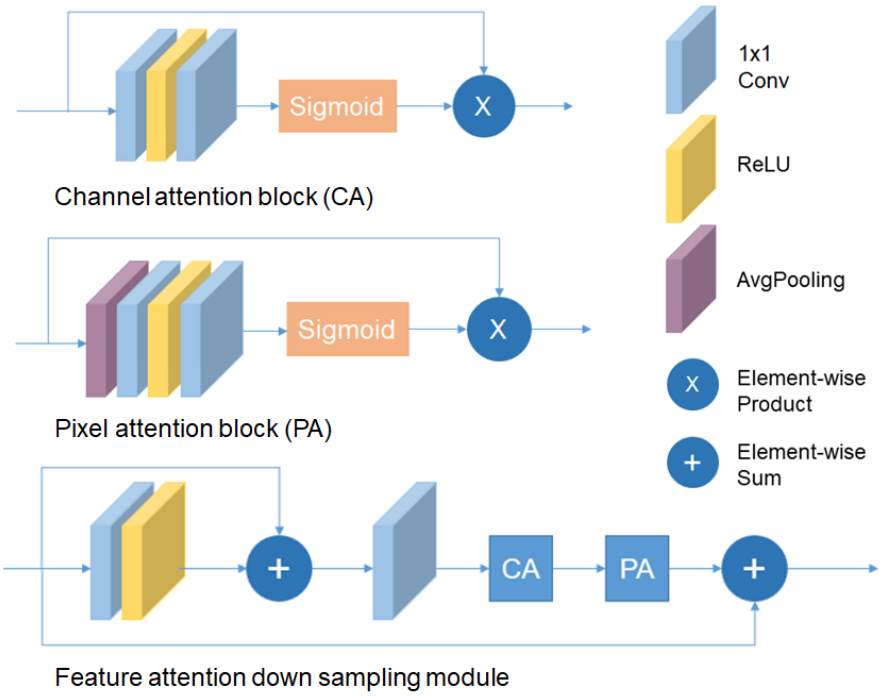

Figure 6. Feature attention down sampling.

Our experiments demonstrated that by using the knowledge transfer method, texture details can be better preserved and be more realistic in the reconstructed image. The L1 norm loss function minimizes the sum of the absolute pixel-wise differences between the intermediate layer features of the scene extraction network and the shadow estimation network:

$$L_{KT}(X_{mid-scene}, X_{mid-shadow}) = \sum_{i=1}^{N} \mid X_{mid-scene} - X_{mid-shadow} \mid . \quad (10)$$

6. Experiments

Experiments were conducted to train the three sub-networks in *DGATRN*: the scene extraction network, the shadow calibration network, and the rendering synthesis network. First, the scene extraction network was trained using the designed loss function and paired input images and target images without shadows. Next, the shadow calibration network was trained using paired input images and target images. Finally, with the scene extraction network and shadow calibration network fixed and their discriminators and the last convolutional layer removed, the rendering network was trained using the designed loss function. During training, all image sizes were resized from 1024×1024 to 512×512, and the mini-batch size was set to 6. The Adam optimizer was used to adjust the network parameters, with a momentum of 0.5 and a learning rate of 0.0001. Each network was trained for 20 epochs. All experiments were conducted using the PyTorch deep learning training framework on a workstation with three NVIDIA GTX3060Ti GPUs. The VIDIT datasets [23] were used for training and testing, and the evaluation metrics used were PSNR, SSIM, LPIPS, and MPS.

6.1. Comparison

We compare the results using our method *DGATRN* with these representative methods: U-Net [24], Pix2Pix [2], Retinex-Net [9], DRN [11], MCN [13], and S3Net [14]. We also include the results of the method by taking away the attention and transfer modules from

DGATRN, namely *DGRN*, as well as results from the baseline method with attention added to the residual sampling blocks, namely *BP + AD*.

The results (Table 1) demonstrate that by using the proposed method to leverage the detailed and global information, the image relighting results can be improved.

Table 1. Comparison based on VIDIT datasets [23].

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	MPS (↑)
U-Net [24]	13.06	0.561	0.307	0.627
Pix2Pix [2]	16.27	0.621	0.274	0.678
Retinex-Net [9]	12.28	0.162	0.657	0.253
DRN [11]	16.08	0.633	0.283	0.675
MCN [13]	16.19	0.641	0.280	0.680
S3Net [14]	16.56	0.621	0.282	0.669
DGRN	16.38	0.649	0.273	0.688
DGATRN	**16.70**	**0.660**	**0.260**	**0.700**

Figure 7 shows the evaluation results of our proposed method, *DGATRN*, and other comparative methods on a test set of 800 images. This graph provides a visual representation of the performance of different methods in terms of evaluation metrics across the entire test set. As shown in Figure 7, our method, *DGATRN*, outperforms other comparative methods comprehensively in terms of PSNR, SSIM, LPIPS, and MPS on the entire test set of 800 images. Performance-wise, our method is similar as MCN [13], since both are founded on the same basis.

In terms of visual quality, as shown in Figure 8, comparing to the representative methods DRN, MCN, and S3Net, we observed that all methods can relight the scene; however, our method can better preserve the details of the scene and shadow, such the rock, shadow, and terrain details. Such details are relatively fine details and the subtle shadows (occlusion) help to depict such bumpy surface details. As our method explicitly addresses feature details as well as the connection between the scene and shadow, thus it can preserve these details better.

We also tested our method in the AIM track1 challenge [25], as shown in Figure 9, the target images are unfortunately not available. We observed that all methods, in general, can achieve similar relighting results with some artifacts, especially for the regions that are very dark, as these features are challenging to capture and recover. Compared to other methods, some details including the fire, shadow, and hole can be better preserved using our method.

6.2. Ablation Study

We also conducted an ablation study for both the scene extraction (Table 2) and shadow calibration (Table 3) to demonstrate the usefulness of each component of our method.

BP is the baseline, AD, AUD, AU, and ARes indicate that the attention is added to the down- (AD), down-/up- (AUD), up- (AU), and residual sampling (ARes) blocks, respectively. DG indicates adding DR_Block and GR_Block for improving feature extraction. *DGRN* means only the improved feature extraction part is included but the attention and knowledge transfer parts are taken away. Based on this study, we found that adding DR_Block and GR_Block leads to an enhancement in the results.

We observed that adding knowledge transfer (+K) can also enhance the results. Moreover, it is optimal to add the attention *FADM* to the downsampling part, such that we can make full use of the semantic and detailed information on the shallow features as much as possible, whereas adding attention to other parts may degrade the overall performance as this may cause undesired conflicts.

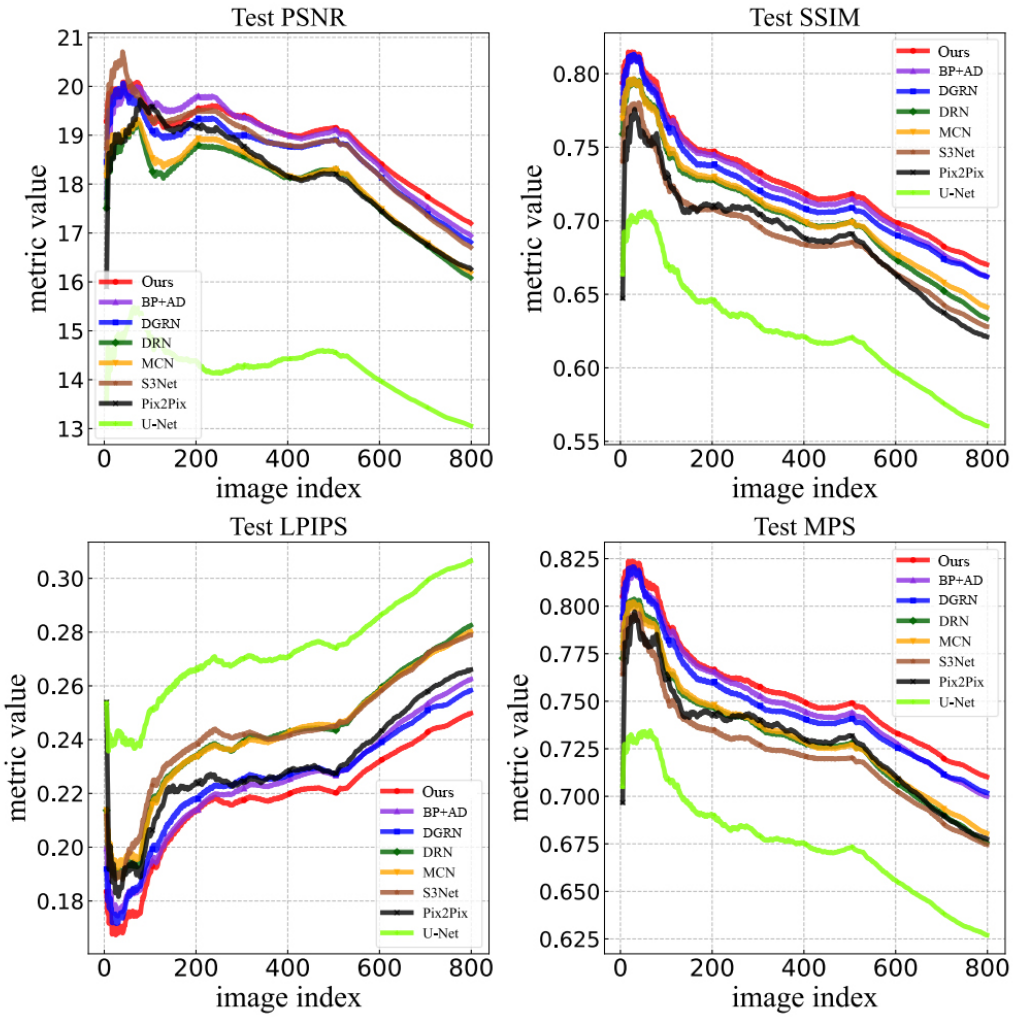

Figure 7. Comparison with other representative methods on the test set of 800 images. The horizontal axis of the graph represents the sequence number of the 800 test images, while the vertical axis represents the values of the evaluation metrics.

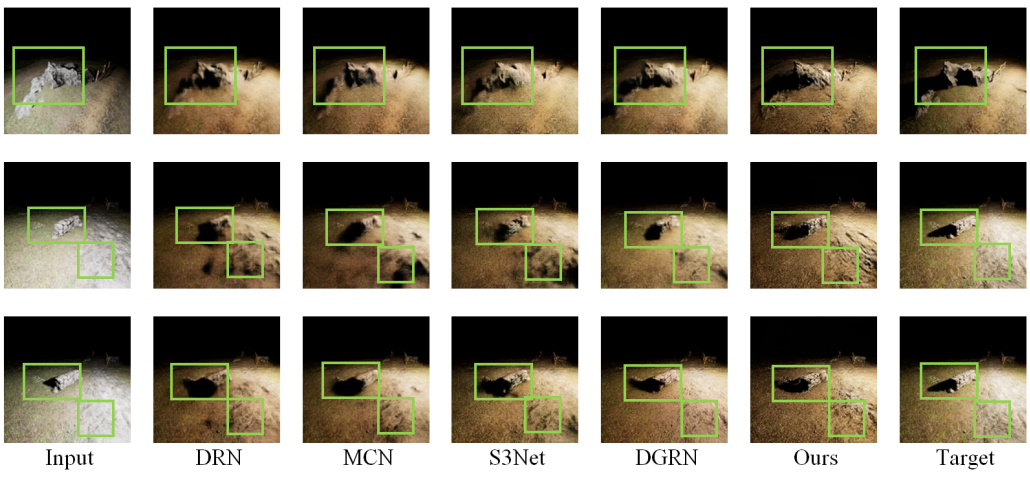

Figure 8. Comparison with other representative methods. The details can be better preserved as shown in the regions highlighted using the green boxes. Comparing the proposed methods *DGRN* and *DGATRN* (labeled as 'Ours'), *DGATRN* performs slightly better.

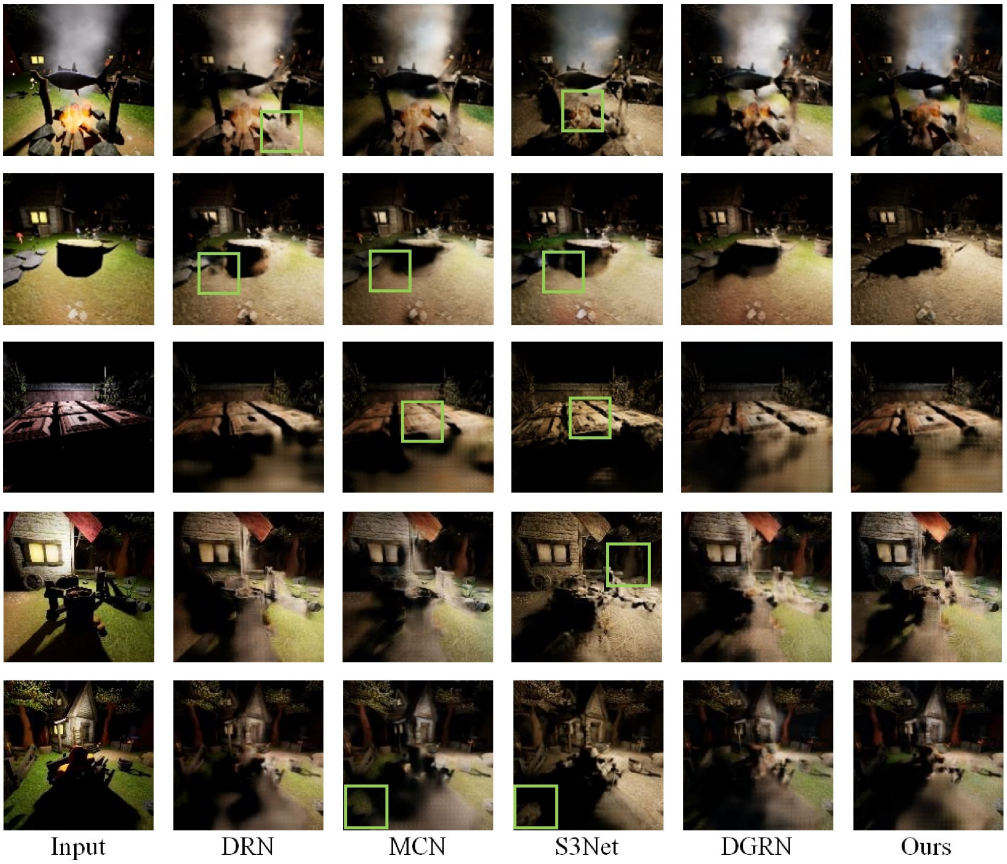

Figure 9. Comparison with other representative methods on the AIM track1 challenge [25]; note that the target example image is not available. As shown in the regions highlighted in green, some key details such as the fire, shadow, and hole might not be well preserved. The proposed methods *DGRN* and *DGATRN* (labeled as 'Ours') achieve similar results.

Table 2. Ablation study for scene extraction network.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	MPS (↑)
BP	18.81	0.638	0.256	0.691
BP + AD	18.93	0.655	0.255	0.700
BP + DG(DGRN)	19.38	0.668	0.245	0.712
DGRN + AUD	18.42	0.607	0.288	0.659
DGRN + AU	19.33	0.675	0.248	0.714
DGRN + ARes	19.32	0.660	0.260	0.700
DGRN + AD	**19.57**	**0.681**	**0.243**	**0.720**

Table 3. Ablation study for shadow calibration network.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	MPS (↑)
BP	15.24	0.475	0.463	0.506
BP + K	15.60	0.510	0.438	0.536
BP + AD	15.70	0.504	0.448	0.528
BP + AD + K	15.74	0.525	0.431	0.547
DGRN	15.76	0.521	0.429	0.546
DGRN + AUD	15.68	0.499	0.458	0.520
DGRN + AU	15.76	0.521	0.429	0.546
DGRN + ARes	**15.78**	0.523	0.415	0.554
DGRN + AD	15.76	0.537	0.413	0.562
DGRN + AD + K	15.66	**0.539**	**0.406**	**0.566**

6.3. Limitations

As previously noted, our technique might exhibit the same artifacts as other methods, particularly in extremely dark regions where features are difficult to capture and restore. Utilizing prior information could aid in minimizing these artifacts.

We also noted that when there is a significant difference in the lighting directions between the input and output images both our method in this paper and existing methods experience a decrease in effectiveness. In future work, we consider predicting depth maps to further extract scene information and optimize the training datasets to enhance the results.

7. Conclusions

This paper presents a method for the problem of attaining visually realistic image relighting by introducing an image-to-image translation network called *DGATRN*. The proposed method focuses on enhancing feature extraction and utilizing context information. *DGATRN* integrates a sequence of methods including the up- and downsampling approach for improved feature extraction and the feature attention downsampling block and knowledge transfer for better utilization of the attention impact and scene–shadow correlation. Experiments were conducted to evaluate the proposed method and demonstrate that *DGATRN* can achieve convincing results.

In the future, we plan to generate our method for more computer vision and graphics tasks such as correction of overexposed images, texture synthesis, and scene reconstruction. We would also like to improve our method for dark scenes by considering the prior information.

Author Contributions: Conceptualization , Q.S.; Methodology, C.F., J.W., K.C. and Q.S.; Software, C.F. and J.W.; Investigation, C.F., J.W., K.C., R.S. and C.-F.L.; Resources, R.S.; Writing—original draft, C.F., J.W. and K.C.; Writing—review & editing, K.C., R.S., C.-F.L. and Q.S.; Visualization, C.-F.L.; Supervision, R.S. and Q.S.; Funding acquisition, C.-F.L., R.S. and Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61702363, and Grant 62222311, the Smart Traffic Fund of Hong Kong under Grant PS/48/2209/RA and the Startup Foundation for Introducing Talent of NUIST, (2022r075).

Data Availability Statement: 3rd Party Data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Marschner, S.R.; Greenberg, D.P. Inverse lighting for photography. In Proceedings of the IST/SID Fifth Colort Imaging Conference, Scottsdale, AZ, USA, 17–20 November 1997; Volume 11.
2. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
3. Tewari, A.; Fried, O.; Thies, J.; Sitzmann, V.; Lombardi, S.; Sunkavalli, K.; Martin-Brualla, R.; Simon, T.; Saragih, J.; Nießner, M.; et al. State of the art on neural rendering. In *Computer Graphics Forum*; Wiley Online Library: New York, NY, USA, 2020; Volume 39, pp. 701–727.
4. Yu, Y.; Smith, W.A. Inverserendernet: Learning single image inverse rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3155–3164.
5. Srinivasan, P.P.; Mildenhall, B.; Tancik, M.; Barron, J.T.; Tucker, R.; Snavely, N. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8080–8089.
6. Loscos, C.; Frasson, M.C.; Drettakis, G.; Walter, B.; Granier, X.; Poulin, P. Interactive virtual relighting and remodeling of real scenes. In Proceedings of the Rendering Techniques' 99: Proceedings of the Eurographics Workshop, Granada, Spain, 21–23 June 1999; Springer: Berlin/Heidelberg, Germany, 1999; pp. 329–340.

7. Yu, Y.; Debevec, P.; Malik, J.; Hawkins, T. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; pp. 215–224.
8. Choe, J.; Im, S.; Rameau, F.; Kang, M.; Kweon, I.S. Volumefusion: Deep depth fusion for 3d scene reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16086–16095.
9. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv* **2018**, arXiv:1808.04560.
10. Xu, Z.; Sunkavalli, K.; Hadap, S.; Ramamoorthi, R. Deep image-based relighting from optimal sparse samples. *ACM Trans. Graph. (ToG)* **2018**, *37*, 1–13. [CrossRef]
11. Wang, L.W.; Siu, W.C.; Liu, Z.S.; Li, C.T.; Lun, D.P. Deep relighting networks for image light source manipulation. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer: Cham, Switzerland, 2020; pp. 550–567.
12. Wu, H.; Liu, J.; Xie, Y.; Qu, Y.; Ma, L. Knowledge transfer dehazing network for nonhomogeneous dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 478–479.
13. Wang, Y.; Lu, T.; Zhang, Y.; Wu, Y. Multi-scale self-calibrated network for image light source transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 252–259.
14. Yang, H.H.; Chen, W.T.; Kuo, S.Y. S3Net: A single stream structure for depth guided image relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 276–283.
15. Zhu, Z.L.; Li, Z.; Zhang, R.X.; Guo, C.L.; Cheng, M.M. Designing an illumination-aware network for deep image relighting. *IEEE Trans. Image Process.* **2022**, *31*, 5396–5411. [CrossRef] [PubMed]
16. Yi, R.; Zhu, C.; Xu, K. Weakly-supervised Single-view Image Relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8402–8411.
17. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
18. Mertens, T.; Kautz, J.; Van Reeth, F. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer Graphics Forum*; Wiley Online Library: New York, NY, USA, 2009; Volume 28, pp. 161–171.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Yazdani, A.; Guo, T.; Monga, V. Physically inspired dense fusion networks for relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 497–506.
21. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
22. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
23. Helou, M.E.; Zhou, R.; Barthas, J.; Süsstrunk, S. VIDIT: Virtual image dataset for illumination transfer. *arXiv* **2020**, arXiv:2005.05460.
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
25. El Helou, M.; Zhou, R.; Süsstrunk, S.; Timofte, R.; Afifi, M.; Brown, M.S.; Xu, K.; Cai, H.; Liu, Y.; Wang, L.W.; et al. AIM 2020: Scene relighting and illumination estimation challenge. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer: Cham, Switzerland, 2020; pp. 499–518.