LTNet: Light Transfer Network for Depth Guided Image Relighting

Yu Zhu[3*] Bosong Ding[1*] Chenghua Li[1*†] Wanli Qian[6] Fangya Li[7]
Yiheng Yao[1] Ruipeng Gang[2] Chunjie Zhang[4] Jian Cheng[1,5†]

[1]Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[2]Academy of Broadcasting Science, NRTA, Beijing 100866, China
[3]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[4]Beijing Key Laboratory of Advanced Information Science and Network Technology, BJTU, China
[5]Nanjing Artificial Intelligence Chip Research, CASIA, Nanjing 211100, China
[6]Georgia Institute of Technology, GA 30318, USA
[7]State Key Laboratory of Media Convergence and Communication, Beijing 100024, China

zhuyu.cv@gmail.com, lichenghua2014@ia.ac.cn, jcheng@nlpr.ia.ac.cn
https://github.com/lchia/relighting_track1_ntire2021

Abstract

Relighting is an interesting yet challenging low-level vision problem, which aims to re-render the scene with new light sources. In this paper, we introduce LTNet, a novel framework for image relighting. Unlike previous methods, we propose to solve this challenging problem by decoupling the enhancement process. Specifically, we propose to train a network that focuses on learning light variations. Our key insight is that light variations are the critical information to be learned because the scene stays unchanged during the light transfer process. To this end, we employ a global residual connection and corresponding residual loss for capturing light variations. Experimental results show that the proposed method achieves better visual quality on the VIDIT dataset in the NTIRE2021 relighting challenge.

1. Introduction

Light is an integral part of photography, which can directly affect the aesthetics of an image. Therefore, proficiency in the use of dimming tools is a must for becoming a professional photographer. In recent years, the post-adjustability of digital tools has significantly simplified the dimming process, making the tools become an indispensable asset for photographers. However, the post-modification is limited to modifying small-scale changes in light intensity or hue. Due to these limitations, photographers still have to spend a lot of time in preparation to adjust the appropriate direction of light. Fortunately, benefit from the rapid development of computational photography, we already have the possibility of modifying the scene light with one click, that is, image relighting.

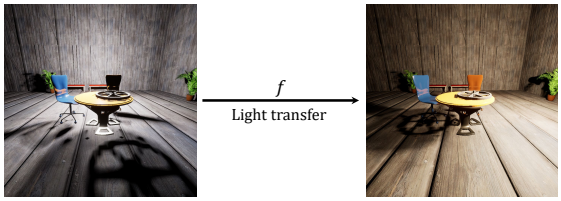


Figure 1. An illustration of the relighting process. For the input image x, the function f modifies its illumination conditions and outputs relit image y. It is worth noting that lighting variations are the critical information that f needs to master because the scene stays unchanged during the lighting transfer process.

Image relighting aims at automatically enhancing images with specific light modifications. Specifically, it re-renders the scene by simulating custom light intensities and light angles in the post-editing of the image, as shown in Figure 1. Attracted by its interesting and practical application, researchers have conducted multiple studies, especially in portrait relighting [22], human relighting [11, 3, 19, 22]. However, their datasets are collected from complex photographic equipment, which means that their future research potential is greatly limited. Recently, in the AIM2020 [1] and NTIRE2021 [2] competitions, Helou et

*equal contribution
†corresponding author

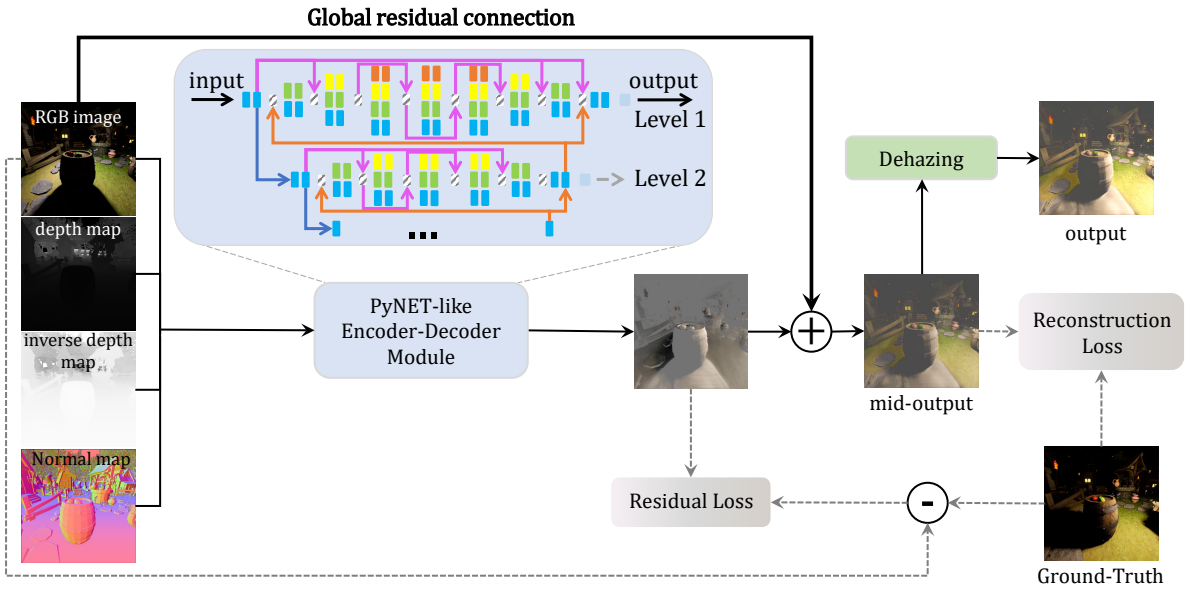


Figure 2. The whole pipeline of our proposed LTNet. The input data consists of four parts, which are the original RGB image, depth map, normal map and inverse normal map. And we take PyNET [9] as our base encoder-decoder model. A global residual connection is employed on it for making the LTNet focus on capturing the illumination variations. During training, we optimize LTNet with the reconstruction loss and the proposed residual loss.

al. proposed to simulate the realistic training data, Virtual Image Dataset for Illumination Transfer (VIDIT) [7]. They take the images by Unreal Engine 4 and provide depth information in the latest competition, which greatly diminishes the difficulty of data collection.

Currently, several studies based on VIDIT have been presented. DeepRelight [20] tries to utilize five images of a scene to reconstruct its appearance under the new illumination conditions. MSRNet [18] takes a two-stage network to accomplish the relighting task step by step. Besides, among the AIM2020 competition [1], previous works mainly adopt an encoder-decoder framework to rebuild the scene and re-render it with new light settings jointly. However, scene reconstruction is not a priority for the relighting task. Relatively, the light transformation is the critical information to capture. It is natural to derive this perception that the change in the scene is not considerable before and after relighting.

In this paper, we present LTNet, focusing on learning the light variation for the image lighting task. To achieve this goal, we consider two aspects. Firstly, we reduce the reconstruction difficulty. It is worth noting that, unlike the natural light conditions, some of the image shadow regions in the simulated VIDIT are absolutely dark, i.e., the pixel values equal zero, which greatly increases the reconstruction difficulty. To solve this problem, we take the original RGB images and depth maps to generate normal maps and the corresponding inverse depth maps as additional scene information. Specifically, they co-operate the original image and the depth image as inputs to assist the network in reconstructing the scene. The enhanced input data provides a wealth of information for the following network. The depth map provides a weak 3D structure of the input image scene in order to render new shadows under varied lighting conditions. And the normal map depicts the shape of the dark regions well. Secondly, we use the global residual structure to make the network capture illumination changes. Correspondingly, we present a novel residual loss to further enhance the sensitivity of the network to light variations. As shown in Figure 2, we utilize an encoder-decoder network architecture to make full use of multi-scale information that assists the reconstruction process and perceive illumination changes. Besides, due to the fog artifact caused by the complexity of the reconstruction process itself, we also employ a defogging algorithm as post-processing to further improve the visual effect.

In summary, our main contributions are as follows:

1) We point out that the light variance on an unchanged scene is the key to the relighting task.
2) We present LTNet, focusing on light variations learning. Benefiting from our complementary reconstruction information and residual network structure design, it can effectively learn the light variation.
3) Extensive results demonstrate that our method achieves state-of-the-art performance in terms of objective metrics and also has a significant improvement in visual quality.

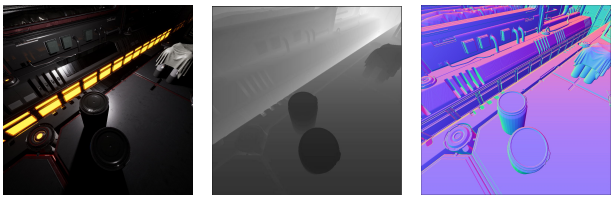
2. Related Work

In this section, we review the relevant works including image relighting and residual learning methods.

Image relighting aims at re-rendering the captured image, specifically modifying its original light source settings (light source position, direction and color temperature). It enables us to relight images in customized lighting conditions with better artistic aesthetics on demand.

Recently, there are several works on the image relighting task. DPR [19] focuses on portrait relighting by embedding the target illumination setting into the encoder-decoder bottleneck for encoding. DeepRelighting [20] encodes new illumination information through a Multilayer Perceptron (MLP) and provides the code to the intermediate representation for light transfer. The base model in DeepRelighting [20] is the UNet [16]. The SA-AE network [1] adopts an implicit scene representation learned by the encoder to render the relit images using the decoder. NRUNet [1] decomposes the relighting process by using two sub-networks, normalization subnetwork and relighting network. DRNIR [1] presents a residual network based on the hourglass network for the image relighting task. Wang et al. [4] proposed the Deep Relighting Network (DRN) for the image relighting task. It consists of three subnetworks, which are respectively responsible for scene reconversion, shadow prior estimation and re-rendering. The three subnetworks are employed with the same UNet network. Densen et al. [15] proposed a wavelet decomposed-based model, named WDRN. MSRNet [18] takes a two-stage network to accomplish the relighting task step by step. We can summarize that these methods are all based on the multi-scale structure and try to reconstruct the scene as well as the light transformation. In this paper, we follow the multi-scale scheme. Different from previous methods, we decouple the relighting process as scene reconstruction and light transfer. And we present LTNet that focuses on learning the light variation.

Residual learning plays an important role in deep learning-based methods. It is first proposed in ResNet [6] for the classification task, which aims to address the problems in deep networks, such as gradient disappearance gradient explosion. Recently, residual learning has also proven to be very effective in low-level vision tasks. For the denoising task, DnCNN [23] demonstrated the effectiveness of residual learning. The later FFDNet [24] also follows this schema to learn the noise by utilizing the global residual connection. Besides, residual learning also is employed in the super-resolution task. Zhang et al. [26] propose a dense network with multiple residual connection and achieves impressive performance. The following ESRGAN [21] further improves the performance by residual learning. Besides, residual learning also shows great reconstruction capability in the image restoration task. Mao et al. [14] propose an autoencoder with residual connection. Jiao et al. [10] propose Formresnet to tackle the image restoration problem by learning the structured residual.



(a) original image (b) depth map (c) normal map

Figure 3. An example of a normal map. Compared to the original image and its depth map, the normal map can provide rich detail information, such as texture.

For the relighting task, residual learning is also practical. As described in AIM2020 challenge report [1], networks such as CFRN, DRNIR, and NRUNet all use residual connection to learn efficiently and effectively. In this paper, we follow the residual learning and use the global residual structure to enforce the network capture illumination changes. And we present a novel residual loss to further enhance the sensitivity of the network to light variations.

3. Proposed Method

In this section, we presented the whole pipeline of LTNet, as shown in Figure 2. To make the LTNet focus on capturing the light variation, we firstly reduce the reconstruction difficulty by offering more input information and then use the global residual structure with the corresponding residual loss. In the following, we elaborate on them in order.

3.1. Complementary Scene Information

To reduce the difficulty in the scene reconstruction process, we first provide the network additional information, especially the normal map and inverse depth map.

In 3D computer graphics, normal mapping is a texture mapping technique used for faking the lighting of bumps and dents, which is an implementation of bump mapping [17]. In practical applications, such as console games, this technique is often used to enhance the appearance and detail of low polygon models by generating normal maps from high polygon models or height maps. And normal maps are usually stored as common RGB images, where the R, G, and B components correspond to the X, Y, and Z coordinates of the normal maps, respectively. According to the above accessibility and usefulness, the normal map can still complement detailed information in the relighting task, as shown in Figure 3.

In real scenes, the normal map is obtained directly from the RGB image and has a moderate effect. However, it is

worth noting that, unlike the natural light conditions, some of the image shadow regions in the simulated VIDIT are absolutely dark, i.e., the pixel values equal zero, which greatly increases the reconstruction difficulty. To solve this problem, we take the original RGB image and depth map to generate the corresponding normal map. As shown in Figure 4, we first convert the RGB image to grayscale. Then, we utilize the sobel operator to extract the edge features, assigning the gradient values in the horizontal and vertical directions to the spatial coordinates x, y respectively. And z is fixed at 0.1. For the depth map, we only operate the second step as above. At last, we fuse the two normal maps in a mean way as the final output.

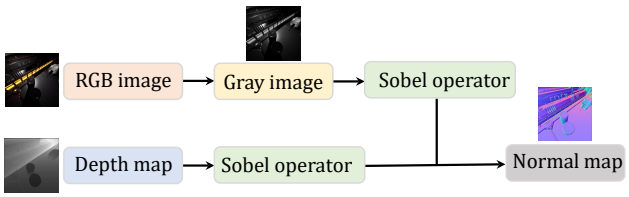


Figure 4. Generation pipeline of normal map.

In particular, the spatial relationship of objects in the scene is critical during the reconstruction process. And the depth map D has 3D information about the input image in one direction, which helps the model to encode the relative positions of objects. Therefore, Moreover, we also use D as complementary input information. Moreover, due to the instability of the depth information, i.e., D may have large or small values, we uniformly normalize it to $[0, 1]$. With the equipment of the RGB image, depth map and normal map, we mine more information to provide to the network. Intuitively, we invert the normalized depth map, i.e., $1 - D$. Because there is the fact that depth information is both useful and difficult to learn for the network [13].

After obtaining the normal and inverse depth maps, we simple concatenate them as the LTNet's input, as shown in Figure 2. Specifically, the enhanced input has 8 channels, which are respectively the original RGB image (3 channels), depth map (1 channel), inverse depth map (1 channel), and the normal image (3channels).

Here, we have obtained all the network inputs, and next, we will elaborate on the structure of LTNet.

3.2. LTNet and Residual Loss

After decreasing the difficulty of scene reconstruction, we enforce the network to focus on learning the light variation. To ensure this, we propose to utilize a global residual structure to make the network capture subtle illumination changes. Correspondingly, we present a novel residual loss to further enhance the sensitivity of the network to light

Figure 5. An example of dehazing. The left image is the output of LTNet. The right image is the result of dehazing operation [5].

variations. We expand them in detail below, from the internal network structure to the loss function.

Firstly, to make full use of multi-scale information that assists the reconstruction process and perceives illumination changes, we take the PyNET [9] as our base model, a typical encoder-decoder framework. PyNET [9] is proposed in the RAW2RGB task, which plays a role as the camera ISP. As shown in Figure 2, it has an inverted pyramidal shape and processes the image at different scales. Moreover, PyNET adopts a slightly dense connection and a number of convolution blocks in parallel with convolution filters of different sizes. In our reimplementation, we reduce the model size of the PyNET for obtaining larger input sizes due to the GPU memory constraints.

Benefit from the strong ability to represent multi-scale information, PyNET can easily reconstruct sophisticated scenes. However, scene reconstruction is not a priority for the relighting task because the scene stays unchanged during the light transfer process. Relatively, the light transformation is the critical information to capture. In order to keep the key idea in the model's mind, we propose LTNet, which is PyNET equipped with a global residual connection.

To further enhance LTNet's sensitivity to illumination variations, we propose a residual loss L_r, as following:

$$L_r = \frac{1}{n}\sum_{i=1}^{n}(|I_i^{gt} - I_i^{pred}| + (I_i^{gt} - I_i^{pred})^2) \quad (1)$$

where I^{gt} and I^{pred} are respectively the ground truth image and the prediction of our LTNet. While training, L_r works together with the final composite loss to optimize LTNet.

3.3. Training Loss

To obtain better visual quality of the LTNet's output, we supplement the residual loss with additional perceptual loss L_p, as following:

$$L_p = 1 \times L_C + 0.1 \times L_{LPIPS} \quad (2)$$

where L_C is the Charbonnier color loss [12] implemented by MAE, which assists the optimization of the deep learning model to be fast and steady. And L_{LPIPS} is the LPIPS

loss [25]. It aims at helping the deep generative model acquire better visual quality of the output images. And we simply combine L_r and L_p as the full training loss.

$$L_{train} = L_p + L_r + L_{rc} \tag{3}$$

where L_{rc} is the reconstruction loss, which is identical to the sum of L_r and L_p.

3.4. Post-process

According to the above elaborate design, the LTNet has greatly improved the effect of relighting. However, the output image may have a foggy artifact due to the challenging reconstruction process. To eliminate this phenomenon, we use a public dehazing tool [5] as a post-process to further improve the visual quality of the output. It should be noticed that the post-process operation is important for reducing the uniform distributed lighting. As shown in Figure 5, the post-process greatly improves the sharpness and the contrast of the original output.

During the competition, we apply the snap-shot ensemble [8] technique in order to obtain better experimental results. And we choose five models, especially trained with 90 epochs, 123 epochs, 124 epochs, 128 epochs, and 134 epochs respectively, in our final test phase. We collect the corresponding five different predicted image sets and calculate the average of them as our submitted results.

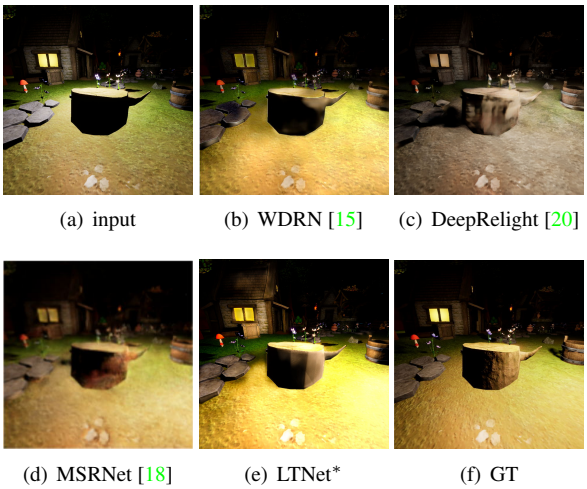
4. Experimental Results

In this section, we evaluate our LTNet on the relighting benchmark dataset VIDIT [7] and compare it with other competition methods. We first describe the implementation details and report the experimental results on NTIRE2021 VIDIT dataset. Extensive results show that our method achieves state-of-the-art performance. Finally, we conduct ablation studies on complementary scene information, light variation learning, the encoder-decoder network architecture, and the model size of the LTNet.

4.1. Implement Details

In the NTIRE2021 competition, the latest VIDIT [7] dataset is presented with depth information. As described in Section 3.1, we take the original RGB image and depth map, combined with our supplementary inverse depth map and normal map, as inputs. The input size is the same as the primitive image size, that is, 1024×1024.

In the training stage, we randomly initialize the weights of the whole network. We use Adam optimizer with momentum terms (0.9,0.999). The initial learning rate is 0.0001, which is decayed exponentially as the number of iterations increases. We train the LTNet on VIDIT's training dataset for 300 epochs with 2 instances stacking a mini-batch. Random rotation and flipping to the images are employed for data augmentation. We implement the experiments separately on Tesla K80s and NVIDIA RTXs depending on the model size. For benchmark evaluation, we compare the results on validation datasets of previous state-of-the-art methods. For competition, we report the results on the test dataset.



(a) input (b) WDRN [15] (c) DeepRelight [20]

(d) MSRNet [18] (e) LTNet* (f) GT

Figure 6. Qualitative comparison with the state-of-the-art methods.

4.2. Comparisons with State-of-the-Arts

In this section, we use the VIDIT benchmark to verify the performance of our LTNet. We compare our method with the previous state-of-the-art relighting methods, WDRN [15], DeepRelight [20] and MSRNet [18]. Table 1 illustrates a quantitative comparison between previous methods and ours. Compared to MSRNet, we provide 0.7242 MPS compared to MSRNet's 0.5905. For other metrics, our LTNet outperforms MSRNet considerably. This growth explains that our model concentrates further on light variations and perceptual quality. For visual comparison, we provide visualization results from different models, as shown in Figure 6. It is worth noting that WDRN and Deeprelight perform worse in light transferring. The MSRNet does better than the above two methods but still suffers from blurry artifacts. Our proposed LTNet shows superior performance on visual quality, especially the relit area. Both the evaluation metrics and the visualization results demonstrate that our LTNet outperforms the state-of-the-art methods.

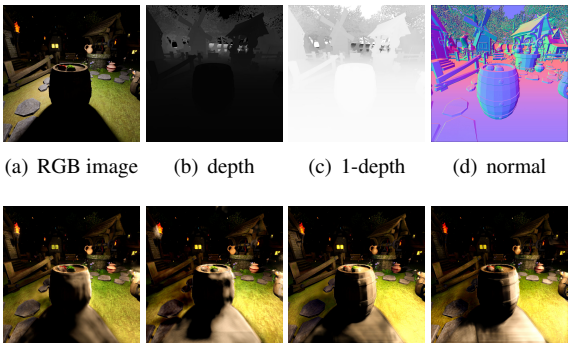
In the middle part of the Table 1, we present the metrics of LTNet with and without dehazing operation. Despite the fact that we could have achieved better results in terms of MPS and PSNR without post-processing. In our experience, although the objective metrics of the images with post-processing are relatively inferior, they have better visual results. Figure 7 shows more comparisons.

At the bottom of the Table 1, we report our competition results in the NTIRE2021 challenge [2] compared with

Table 1. Comparison of our LTNet against the state-of-the-art methods on the VIDIT validation dataset and the results of the test dataset on NTIRE2021 Challenge (the bottom chart). We directly cite the best results reported in [15, 20, 18, 2]. The best quantitative results (MPS, LPIPS, PSNR and SSIM) are in bold. LTNet* is the submitted version in NTIRE2021 Challenge with small model size, while the LTNet† is a larger version of LTNet with more channels in the PyNET-like encoder-decoder module.

Model	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
WDRN [15]	0.6453	0.6310	0.3405	17.0717	0.0300 (P100)
DeepRelight [20]	0.5892	0.5928	0.4144	17.4252	0.5000 (2080TI)
MSRNet [18]	0.5905	0.5899	0.4088	17.8900	0.0116 (1080 Ti)
LTNet*	0.6754	0.6369	0.2861	16.8836	0.0422 (Tesla K80)
LTNet† (with dehaze)	0.7063	0.6730	0.2604	17.7711	0.0611 (Tesla K80)
LTNet† (without dehaze)	**0.7242**	**0.6955**	**0.2470**	**19.1853**	0.0546 (Tesla K80)
AICSNTU-MBNet	**0.7663**	**0.6931**	**0.1605**	**19.1469**	2.88s (Tesla V100)
iPAL-RelightNet	0.7620	0.6874	0.1634	18.8358	0.53s (Titan XP)
VUE	0.7671	0.6874	0.1532	19.8901	0.23s (Tesla V100)
LTNet*	0.7101	0.6084	0.1882	15.8591	**0.0422s (Tesla K80)**

Figure 7. Results with or without dehazing compared to the ground truth. The images are sampled from the validation dataset of the VIDIT [7].



(a) RGB image (b) depth (c) 1-depth (d) normal

(e) I.D. (f) I.D.V. (g) I.D.N (h) I.D.V.N

Figure 8. Visual quality of the outputs of models trained with different augmented input data.

other teams. We list the results of top three solutions and our submitted LTNet*. Compared with their sophisticated models, our LTNet achieves good results in a short inference time.

4.3. Ablation Study

In this section, we separately conduct ablation studies on complementary scene information, light variation learning, the base encoder-decoder network, the model size of the LTNet, and training loss. In particular, as described in Section 3, we enforce the LTNet to focus on the acquisition of illumination variations from two aspects, i.e., the complementary scene information and the light variation learning. In the following, we demonstrate the effectiveness of the proposed method in detail.

4.3.1 Supplementary Scene Information

To reduce the difficulty of scene reconstruction, we propose supplementary scene information as network input. As described in Section 3, the input data consists of four parts, which are the original RGB image, depth map, normal map and inverse normal map. In this experiment, we compare the results by considering four kinds of inputs: (i) I, D: the concatenation of the RGB image and the depth image, (ii) I, D, V: the concatenation of the RGB image, the depth image, and the inverse depth image $1 - D$, (iii) I, D, N: the concatenation of the RGB image, the depth image, and the normal image, (iv) I, D, V, N: the concatenation of the RGB image, the depth image, the inverse depth image, and the normal image.

Table 2. Ablation study for different augmented inputs.

Input	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
I, D	0.6756	0.6308	0.2795	16.8453	0.0448
I, D, V	0.6666	0.6177	0.2846	16.4589	0.0464
I, D, N	**0.6855**	**0.6490**	**0.2780**	**16.7863**	0.0470
I, D, V, N	0.6828	0.6473	0.2817	16.6719	0.0467

As shown in Table 2, The model trained with I, D, N achieves the best performance. This confirms that our pro-

posed normal map is the most important factor to improve quantitative metrics. Figure 8 demonstrates that the output of LTNet trained with normal maps has better visual quality than other models. Besides, the inverse depth map V can also improve the visual quality. Therefore, we finally proposed to take $I.D.V.N$ as network input in the following experiments.

4.3.2 Light Variation Learning

In this subsection, we conduct the ablation study on the light variation learning, which is proposed to make the LTNet be more sensitive to the illumination variation. In addition to the augmented inputs, three other factors play a key role in the overall pipeline, which are respectively the global skip connection s, the residual loss L_r and the original reconstruction loss L_{rc}.

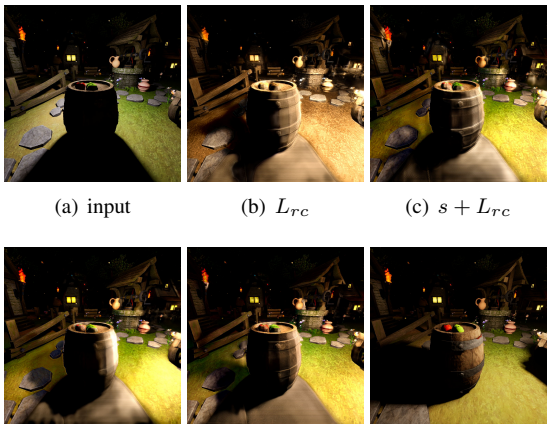
Table 3. Ablation study on the model settings for light variation learning.

s	L_r	L_o	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
×	×	✓	0.6525	0.6248	0.3197	16.5672	0.0426
✓	×	✓	0.6701	0.6255	0.2852	**16.9150**	0.0469
✓	✓	×	0.6809	0.6425	**0.2807**	16.7779	0.0488
✓	✓	✓	**0.6828**	**0.6473**	0.2817	16.6719	0.0467

As shown in Table 3, the first model is trained without global skip connection s and residual loss L_r. And it is inferior to all the models trained with global skip connection s. This demonstrates the validity of the global skip connection in the proposed LTNet. Besides, considering the second and third rows of the Table 3, the residual loss L_r can effectively improve the outputs' visual quality according to the visual quantitative metrics, MPS and SSIM. As shown in Figure 9, we visualize the outputs of models trained with different model settings. The model trained with L_{rc} and L_r achieves relatively better visual results than others.

4.3.3 Base Encoder-decoder Network

As shown in Figure 2, we adopt a PyNET-like encoder-decoder network as our base model. In this subsection, we make an ablation study on the base model to verify that PyNET [9] has powerful reconstruction ability. As a comparison, we evaluate LTNet with two different base models: (i) WDRN, (ii) PyNET. WDRN [15] is a UNet-like encoder-decoder network used for image relighting in AIM2020 [1], which is the winner of the image relighting competition. It adopts the wavelet decomposition module to replace the upsample and the downsample operation in the original UNet, which has better results in several low-level vision tasks.



(a) input (b) L_{rc} (c) $s + L_{rc}$

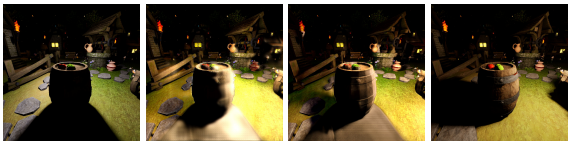
(d) $s + L_r$ (e) $s + L_{rc} + L_r$ (f) GT

Figure 9. Visual results of models trained with different settings. The global skip connection, the original reconstruction loss L_{rc}, or the residual loss L_r are considered in this experiment.

Table 4. Ablation study on base encoder-decoder networks. $E\&D$ represents the base encoder-decoder networks. All the models are evaluated on the validation dataset of VIDIT [7].

$E\&D$	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
WDRN	0.6490	0.5996	0.3016	15.7180	0.0704
PyNET	**0.6828**	**0.6473**	**0.2817**	**16.6719**	0.0467

As reported in Table 4, the LTNet with PyNET significantly outperforms the one with WDRN in the relighting task with less inference time. Besides, as shown in Figure 10, given the same input image, the result of the LTNet with PyNET is much more natural than the other one. LTNet with PyNET-like encoder-decoder network handles better in both scene reconstruction and generating new images with new light conditions.



(a) input (b) +WDRN (c) +PyNET (d) GT

Figure 10. Visual samples of the LTNet with different encoder-decoder networks. +WDRN means LTNet is based on WDRN, which is same to +PyNET.

4.3.4 Model Size

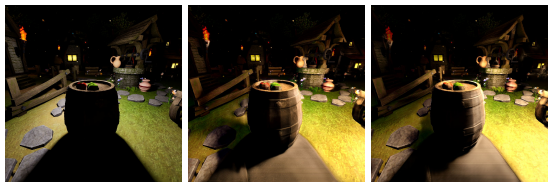
To further explore the potential of LTNet, we implemented ablation experiments on the model size. For ease of illustration, we use n_c and n_f to respectively indicate the base channel number in the first level of the PyNET-like network

and the other levels' channel number. In the following, we modify the model size by replacing these two hyperparameters, n_c and n_f.

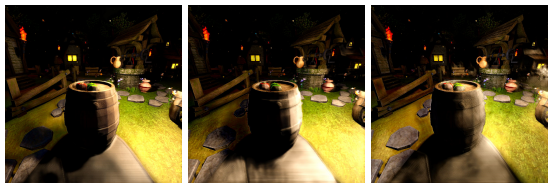
Table 5. Ablation study on the model size of the LTNet.

n_c, n_f	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
8,32	**0.7063**	**0.6730**	**0.2604**	**17.7711**	0.0611
4,32	0.7030	0.6668	0.2608	16.9733	0.0556
4,16	0.6920	0.6573	0.2733	16.9539	0.0550
4,16	0.6828	0.6473	0.2817	16.6719	0.0467
4,4	0.6462	0.5910	0.2987	16.1319	0.0604

As represented in Table 5, the quantitative performance gets better as long as the model size increases, which is in line with our basic perception. However, the visual quality may not be consistent with the objective metrics. As shown in Figure 11, the output image even has redundant shading. After weighing the computational cost and the visual quality, we follow the setting, $n_f = 16, n_{f1} = 2$, in all experiments, including in the NTIRE2021 challenge.



(a) input (b) 32&8 (c) 32&4



(d) 16&4 (e) 16&2 (f) 4&2

Figure 11. Visual samples of the LTNet with the different model sizes.

4.3.5 Training Loss Configurations

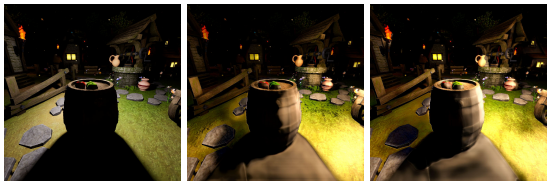
At last, we perform ablation study on training losses. In our experiments, four losses are involved, which are the Charbonnier color loss [12] L_C implemented by MAE, the LPIPS loss [25] L_{LPIPS}, and residual loss L_r. We consider the following four losses for training: (i) $A = L_C$; (ii) $B = L_C + 0.5 \times L_{LPIPS}$; (iii) $C = L_C + L_r$; (iv) $D = L_{1C} + 0.1 \times L_{LPIPS} + L_r$.

As reported in Table 6, the LTNet trained with L_C has the best quantitative results. However, as shown in Figure 12, it is inferior for the visual quality of the outputs

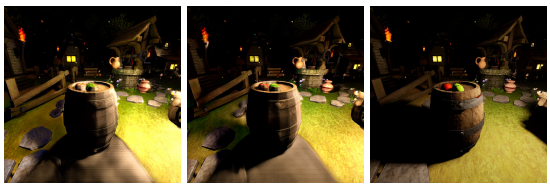
Table 6. Ablation study on different training losses.

Loss	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time
A	**0.6861**	**0.6486**	**0.2763**	**16.9595**	0.0552
B	0.6791	0.6365	0.2784	16.6524	0.0570
C	0.6788	0.6342	0.2765	16.4459	0.0608
D	0.6828	0.6473	0.2817	16.6719	0.0467

of the LTNet trained with L_C. Compared with the four settings, the performance of the LTNet trained with setting D is relatively considerable.



(a) input (b) A (c) B



(d) C (e) D (f) GT

Figure 12. Visual samples of the outputs of the LTNet trained different losses.

5. Conclusion

In this paper, we point out that the light variance on an unchanged scene is the key to the relighting task. And we present LTNet, focusing on light variations learning. Benefiting from our complementary reconstruction information and residual network structure design, it can effectively learn the light variation. Extensive results demonstrate that our method achieves state-of-the-art performance in terms of objective metrics and also has a significant improvement in visual quality. We believe the light variation learning is a promising direction for the relighting task, which worth further explorations.

6. Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61906193; No. 62072026; No. 61906195; No. 62076235), Beijing Municipal Science & Technology Commission (No. Z191100003419003), and Beijing Natural Science Foundation (JQ20022).

References

[1] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. AIM 2020: Scene relighting and illumination estimation challenge. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2020. 1, 2, 3, 7

[2] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021: Depth-guided image relighting challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 1, 5, 6

[3] Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deferred neural lighting: Free-viewpoint relighting from unstructured photographs. *ACM Trans. Graph.*, 39(6), Nov. 2020. 1

[4] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[5] K. He, S. Jian, Fellow, IEEE, and X. Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(12):2341–2353, 2011. 4, 5

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE*, 2016. 3

[7] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. Vidit: Virtual image dataset for illumination transfer, 2020. 2, 5, 6, 7

[8] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017. 5

[9] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 4, 7

[10] Jianbo Jiao, Wei-Chih Tu, Shengfeng He, and Rynson WH Lau. Formresnet: Formatted residual learning for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–46, 2017. 3

[11] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019. 1

[12] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 4, 8

[13] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260. IEEE, 2010. 4

[14] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *arXiv preprint arXiv:1603.09056*, 2016. 3

[15] Densen Puthussery, Hrishikesh Panikkasseril Sethumadhavan, Melvin Kuriakose, and Jiji Charangatt Victor. Wdrn: A wavelet decomposed relightnet for image relighting. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 519–534, Cham, 2020. Springer International Publishing. 3, 5, 6, 7

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[17] Peter-Pike Sloan. Normal mapping for precomputed radiance transfer. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 23–26, 2006. 3

[18] Saikat Dutta Sourya Dipta Das, Nisarg A Shah. Msr-net: Multi-scale relighting network for one-to-one relighting. In *Differentiable computer vision, graphics, and physics in machine learning, in NeurIPS 2020 Workshop*, 2020. 2, 3, 5, 6

[19] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4), July 2019. 1, 3

[20] Li-Wen Wang, Wan-Chi Siu, Zhi-Song Liu, Chu-Tak Li, and Daniel P. K. Lun. Deep relighting networks for image light source manipulation. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 550–567, Cham, 2020. Springer International Publishing. 2, 3, 5, 6

[21] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[22] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018. 1

[23] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 3

[24] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 3

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 8

[26] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3