Multi-modal Bifurcated Network for Depth Guided Image Relighting

Hao-Hsiang Yang[1*], Wei-Ting Chen[1,2*], Hao-Lun Luo[3], and Sy-Yen Kuo[3]

[1] ASUS Intelligent Cloud Services, Asustek Computer Inc, Taipei, Taiwan
[2] Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
[3] Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan.

(islike8399, jimmy3505090)@gmail.com, (r08921051, sykuo)@ntu.edu.tw

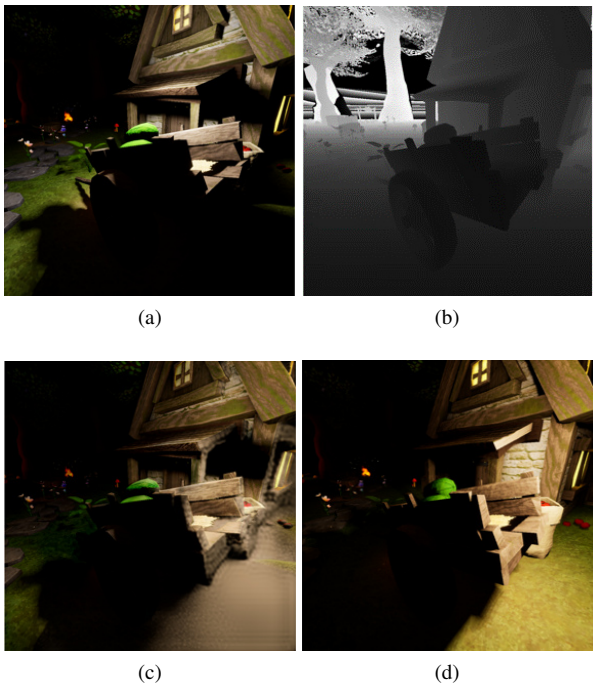
https://github.com/weitingchen83/NTIRE2021-Depth-Guided-Image-Relighting-MBNet

Abstract

Image relighting aims to recalibrate the illumination setting in an image. In this paper, we propose a deep learning-based method called multi-modal bifurcated network (MBNet) for depth guided image relighting. That is, given an image and the corresponding depth maps, a new image with the given illuminant angle and color temperature is generated by our network. This model extracts the image and the depth features by the bifurcated network in the encoder. To use the two features effectively, we adopt the dynamic dilated pyramid modules in the decoder. Moreover, to increase the variety of training data, we propose a novel data process pipeline to increase the number of the training data. Experiments conducted on the VIDIT dataset show that the proposed solution obtains the $\mathbf{1}^{st}$ *place in terms of SSIM and PMS in the NTIRE 2021 Depth Guide One-to-one Relighting Challenge.*

1. Introduction

Given an image, image relighting aims to relight this image into another image with different ambient conditions. In the NTIRE 2021 Depth Guide One-to-one Relighting Challenge, the depth maps are provided. We plot examples including the original image, the corresponding depth map, the relighted image by the proposed method, and the ground truth in Fig. 1. As shown in Fig. 1, there are two inherent challenges for image relighting. First, to generate the image with different ambient conditions, it is necessary to generate shadows into the relighted image. Second, similarly, the shadow from the original image needs to be removed. For example, in Fig. 1 (c), although the region shadows are removed, the texture of the grass cannot be recovered appropriately.



(a) (b) (c) (d)

Figure 1: Example of depth guided image relighting. (a): Original input image. (b): Corresponding depth map. (c): Relighted image by our method. (d): Ground truth.

Previously, many methods [1, 2, 3] based on developing visual priors or capture properties of relighted images have achieved impressive performance. Recently, some deep learning-based methods [4, 5, 6, 7] are proposed without explicit inverse rendering steps for estimating scene properties. However, these methods do not consider complex scenes and various ambient conditions. Moreover, in NTIRE 2021 Depth Guide One-to-one Relighting Chal-

*Equally-contributed first authors.

lenge, there are extra difficulties needed to be addressed. First, the number of training data given in the VIDIT [8] dataset is 300, which is not enough to train the network. To address these issues, our network is based on the backbone pre-trained from ImageNet. Furthermore, we leverage extra data from Track II: Any-to-Any relighting [7] and develop a strategy that can generate new image pairs to increase the number of the training data and the robustness of the network.

Second, depth and image features should be effectively extracted and fused. Depth maps and images contain the different attributes of features. Depth maps present spatial information while the images provide texture, light cues, and dark cues. Thus, we refer to the methods proposed in the RGB-D salient object detection [9, 10, 11] and propose a multi-modal bifurcated network to deal with this issue. This network contains encoder and decoder parts. Our encoder applies two branches without sharing weight to extract features, respectively. Additionally, we apply the dynamic dilated pyramid module to effectively integrate two features. In the decoder parts, we gradually magnify the feature maps and recover the image. Motivated by the U-Net [12, 13], we apply skip connection to connect the feature maps with identical size from the encoder and decoder parts to obtain better relighted images.

We make the following contributions in this paper:

1. The multi-modal bifurcated network is proposed for depth guided image relighting. This structure can extract image and depth features by two branches. Then, these two features are fused by dynamic dilated pyramid modules effectively.

2. We propose a new strategy to leverage additional images from Track II and construct more input-output pairs as the training data.

3. Several experiments performed on the VIDIT [8] dataset demonstrate that our solution achieves the $\mathbf{1}^{st}$ place in terms of MPS and SSIM in the NTIRE 2021 Depth Guided One-to-one Relighting Challenge.

2. Related Works

Two tasks are very similar to depth guided image relighting: RGB-D saliency object detection and image relighting. In this section, we briefly describe several works related to these tasks.

2.1. RGB-D Salient Object Detection

Salient Object Detection (SOD) aims to imitate the human visual system and detect certain regions or objects that attract human attention. RGB-D SOD is known as combining the extra depth maps to fulfill salient detection. Some models [10, 11] extract features from images and depth maps independently, and conducts feature maps fusion of the two modalities in the decoder. In [11], Asymmetric Two-Stream Architecture (ATST) is proposed which considers the inherent differences between the RGB and the depth data for the salient detection. In [10], Guided Residual (GR) blocks are proposed to feed the RGB image and the depth image alternately to reduce the mutual degradation. They also address progressive guidance in the stacked GR blocks within each side-output to remedy the false detection and the missing parts. On the other hand, in [14], the single stream network is designed that concatenates the RGB image and depth image across the channel dimension and directly applies the depth map to guide both the early fusion and the middle fusion between the RGB information and the depth information, which saves the feature encoder of the depth stream. Additionally, in [15], authors propose the RD3D that 3D convolutional neural networks are introduced to address the RGB-D SOD. This network adopts the progressive fusion involving both the encoder and the decoder stages. Since the multi-modal fusion methods from RGB-D SOD can be leveraged for the depth-guided image relighting, in this paper, the proposed MBNet is based on the HDFNet [9].

2.2. Image Relighting

The algorithms for image relighting can be categorized as the physical-based method and the deep learning-based method. Traditional methods [1, 2, 3] depend on the physical observation to further estimate the ambient conditions, reflectance, and lighting of the scene in the image and then re-render this scene by another illumination setting. In [2], an algorithm treating the complex scene as a linear system that transforms the original light into the reflected light is proposed. This algorithm progressively refines the approximation of the reflectance field until the required precision is reached. In [3], authors prove that the light transport of diffuse scenes under the spatially varying illumination can be decomposed into the direct, near-range and far-range transports. They separate these three components in the frequency space. On the other hand, many deep learning-based methods for image relighting [4, 5, 7] are proposed. Image relighting can be seen as an image-to-image translation problem. Several low-level image processing problems like haze/smoke removal [16, 17, 18], underwater enhancement [19], reflection removal [20], image deraining [21] and image desnowing [22] are very similar to image relighting. Generally speaking, the encoder-decoder structure like U-Net [12, 13] can deal with these tasks. Some methods that deal with the ambient conditions especially for the image relighting are developed. Gray loss described in [5] can drive the network to learn the illumination gradient in target domain images. Xu *et.al* propose a CNN-based method [4] to relight a scene under a new illumination based on five im-

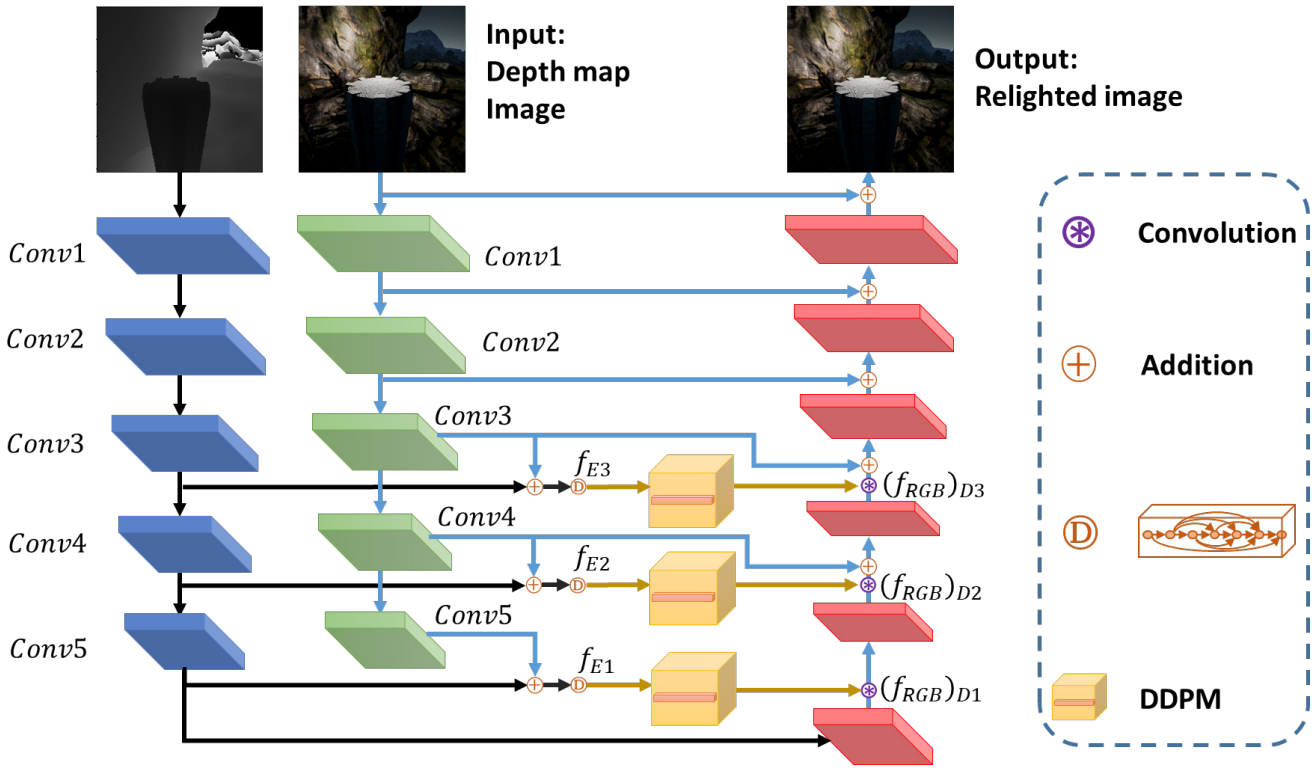


Figure 2: The architecture of the proposed multi-modal bifurcated network. The network consists of two streams as encoder parts: depth stream and RGB-image stream. We use the dense architecture ,DDPM and skip connection for better feature extraction as decoder parts.

ages captured under a pre-defined illumination setting. This method tries to estimate a non-linear function that generalizes the estimation from the optimal sparse samples.

3. Proposed methods

3.1. Overall Neural Network

As shown in Fig. 2, to leverage the depth information for depth guided image relighting task effectively, we refer the dual-stream bifurcated architecture in [9] as the backbone. Our network consists of two streams to extract the depth and the image features. We apply the two ResNet 50 [23] pre-trained from the ImageNet as the backbones. In the network design, the depth and image features from three intermediate layers are fused to achieve the representative features. We fuse the features from the conv3, the conv4 and the conv5 layers to balance the effectiveness and the efficiency of the network. Specifically, for the features in the shallower layers, they are generally noisier and the high-resolution of these features may increase the computational burden. However, the features in the conv3 to the conv5 may still preserve the valid information [9] and with lower resolution. To fuse these two features with multiple receptive fields, we leverage the densely connected architecture to generate the combined features with the fruitful texture and the structural information. Then, these features are fed to the Dynamic Dilated Pyramid Module (DDPM) [9] that can generate a more discriminative result. We will describe this module in the following section. The output of the DDPM is combined with the output of the decoder by convolving with the multi-scale convolution kernels [24, 22, 25]. In the decoder part, similar to the U-net [13, 12], we gradually magnify the feature maps and implement the skip connection to concatenate the identical size feature maps. Furthermore, we make our network learn the residual [23] instead of the whole images. That is, the final output is the difference between the original image and the relighted image.

3.2. Dynamic Dilated Pyramid Module

In this part, we illustrate the detail of the DDPM. As shown in Fig. 3, the input of DDPM is the fused feature f_E and the feature $(f_{RGB})_D$ from encoders, respectively. First, the $(f_{RGB})_D$ passes through the convolution kernel to reduce the dimension of the feature which is termed as ${}^*(f_{RGB})_D$. Second, the kernel generation units (KGUs) [9] are adopted on the fused feature f_E to generate different weight tensors (i.e, f^1_{KGU}, f^2_{KGU}, f^3_{KGU}) which can cover

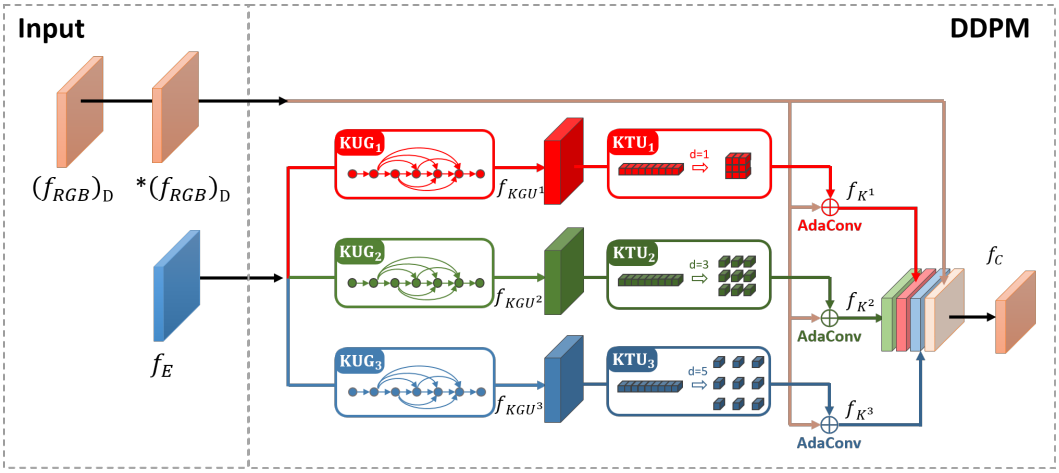


Figure 3: The architecture of the dynamic dilated pyramid module (DDPM). There are two modules in the network, that is, the kernel generation units (KGUs) and the kernel transformation units (KTUs).

three square neighborhoods (i.e., 3×3, 7×7, and 11×11) . It is noted that KGUs are with four densely connected layers [26] which can improve the feature propagation and feature reuse effectively. Then, we leverage the kernel transformation units (KTUs) to yield regular convolution kernels with various dilation rate (i.e., 1, 3, 5) by reorganizing kernel tensors and inserting various numbers of zeros. Then, the three parallel output features combine with the $^*(f_{RGB})_D$ by the convolution kernels, respectively. We term these combined features as f_K^1, f_K^2, f_K^3. Finally, we combine f_K^1, f_K^2, f_K^3 and $^*(f_{RGB})_D$ to generate the output of the DDPM f_C.

3.3. Extra Data Usage

In this section, we propose two strategies to increase the training data so that our model can learn the mapping function of the depth guided image relighting robustly. In this challenge, the input image (6500-N) is with 6500K color temperature and the north illumination angle while the output image (4500-E) is with 4500K color temperature and with the east illumination angle. In order to improve the robustness of our network, we leverage the images in Depth Guided Image Relighting: Track II Any-to-Any relighting [7]. This track provides the images with various illumination temperature and different illumination angles. Specifically, we apply the images with the different illumination angle (6500-NE) but the identical color temperature as the input and the ground truths are the same ones (4500-E). As shown in Fig. 4, the additional image (6500-NE) is very similar to the original one (6500-N). With this additional data, the model can understand the direction information comprehensively.

Moreover, We adopt additional images which contain the same scene but with the different illuminating angle (4500-W). We can flip horizontally the image with the west illumination angle to achieve the new image with the east illumination angle. Thus, we develop a new strategy to further increase training data and illustrate it in Fig. 5. As shown in Fig. 5 (d) and Fig. 5 (e), we flip the original RGB (6500-N) images and the corresponding depth map horizontally. The horizontally flipped image of (4500-W) is the output of the flipped input (6500-N). With this operation, the training data can be increased.

3.4. Loss Functions

In this paper, we leverage three loss functions to measure the differences between the relighted images and the ground truth. The three loss functions are the Charbonnier loss [27], the SSIM loss [28], and the perceptual loss [29]. The Charbonnier loss can be expressed as:

$$L_{Cha}(x, \hat{x}) = \frac{1}{T}\sum_{i}^{T}\sqrt{(x_i - \hat{x_i})^2 + \epsilon^2} \tag{1}$$

where x and $\hat{x}$ are the ground truth and relighted images, respectively. e is a tiny constant for the stable and the robust convergence.

The second loss function is the SSIM loss function. SSIM loss is expressed as:

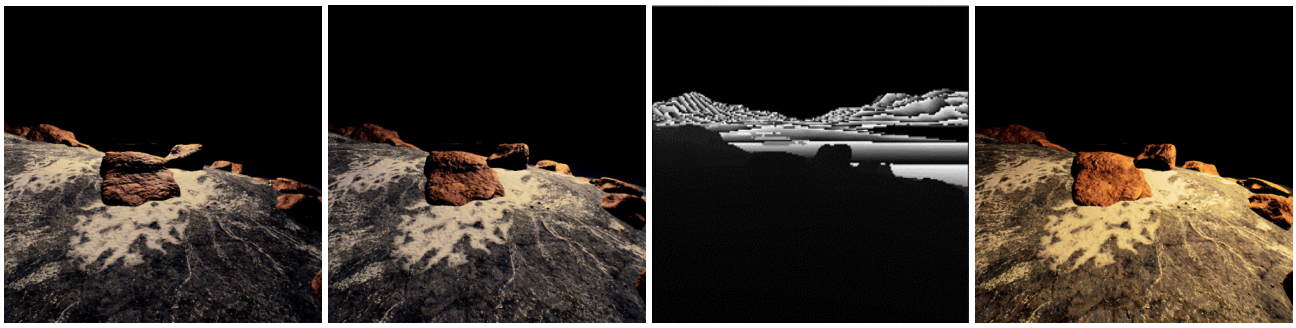
$$L_{SSIM}(x, \hat{x}) = -SSIM(x, \hat{x}) \tag{2}$$

SSIM loss is beneficial to reconstruct local structures and details.

Finally, the perceptual loss is written as:

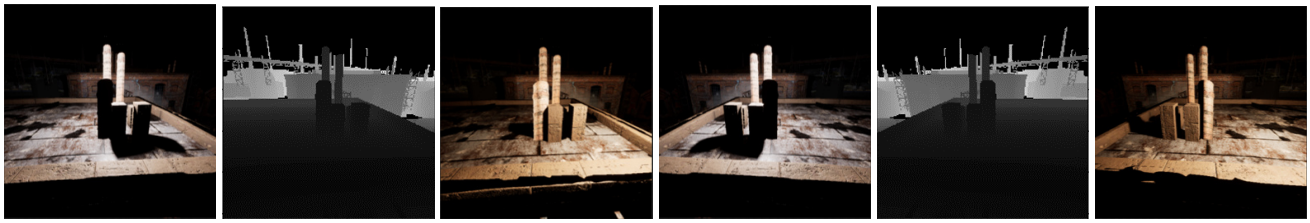
$$L_{Per}(x, \hat{x}) = |(VGG(x) - VGG(\hat{x})| \tag{3}$$

where VGG means the VGG19 network [30]. In our work, we use the features from conv3-3 layer. The overall loss



(a) (b) (c) (d)

Figure 4: Additional images are used in the training phase. (a) Original image (6500-N). (b) Different illuminating angle image (6500-NE). (c) Guided depth map. (d) Output image.



(a) (b) (c) (d) (e) (f)

Figure 5: Using (4500-W) images to increase the diversity of the training data. (a): Original input image. (b): Original depth map. (c): Output image. (d): Flipped image of (a). (e): Flipped depth map of (b). (f): Flipped image of (4500-W).

function containing three terms is expressed as:

$$L_{Total} = \lambda_1 L_{cha} + \lambda_2 L_{SSIM} + \lambda_3 L_{Per} \quad (4)$$

where λ_1, λ_2 and λ_3 are weights to control the final objective functions. These three weights are empirically adjusted as hyper-parameters.

4. Experiments

4.1. Experimental Setting

In the NTIRE 2021 Depth Guide One-to-one Relighting Challenge, the novel Virtual Image Dataset for Illumination Transfer (VIDIT) [8] is provided as the training and the validation data. This dataset consists of 390 various scenes that are captured at 40 different illumination conditions including 8 different azimuthal angles and five color temperatures such as 2500K, 4500K, etc. Furthermore, the corresponding depth maps are provided. In track I - depth guided one-to-one relighting, the input images are depth map and a pre-defined illumination condition θ = North, temperature = 6500K (e.g., (6500-N)) and the output image is set at a different illumination setting θ = East, temperature = 4500K. Though in track I, only two conditions of images are used as the input and the output pairs, it is allowed to utilize the extra data to improve the accuracy of the model. During the training and the evaluation phases, the image size is 1024 $\times$ 1024, and we do not use any data augmentation like random flip and random crop. The Adam optimizer [31] is utilized with a batch size of 3 to train the network. We train the network for 200 epochs with the momentum β_1 = 0.5, β_2 = 0.999. The learning rate is initialed as 10^{-4} and divided by ten after 50 epochs. The λ_1, λ_2 and λ_3 in (4) are set as 1, 1.1 and 0.1, respectively. We perform our experiments on a single Nvidia V100 graphic card and the PyTorch platform. We spend about 11 hours finishing the model training. In the testing phase, we take 2.8867 seconds to predict a single image. The source code will be available in our project page.

Table 1: The ablation experiment of applying the different data and the residual learning.

Description	PSNR	SSIM
Baseline	18.0215	0.6834
+ Extra data	18.9677	0.7103
+ Extra data + Residual learning	19.3558	0.7175

(a) Original image (b) Depth map (c) jimmy3505090 (d) aics (e) Ours (f) Ground Truth

Figure 6: Visual comparison for the relighted results recovered by the MBNet and other solutions.

4.2. Ablation Experiments

We conduct the ablation experiment to verify that each module applied in this paper can benefit the proposed relighting network. In all experiments, the image size is set as 1024$\times$ 1024. We test each module and report the results under the validation set. We select the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) as objective metrics for the quantitative evaluation. Overall, the ablation studies consist of three different experimental scenarios: 1) We apply the original training data to train the MBNet as the baseline. 2) We apply the method described in Section

Table 2: Relighted results by some state-of-the-art RGB-D salient object methods.

Methods	PSNR	SSIM
ATST [11]	18.2678	0.665
DANET [14]	18.3341	0.6805
RD3D [15]	17.7763	0.6668
PGAR [10]	17.6436	0.6748
Ours	19.3558	0.7175

3 to increase the training data to train our MBNet. 3) We increase the training data and also apply the residual learning [23] strategy. We summary the ablation experiments in Table 1. One can see that both PSNR and SSIM scores of setting 2 are increased compared with setting 1. It can show that increasing training data is beneficial for better performance and robustness. Additionally, compared with setting 2, the performance of setting 3 is improved effectively. It demonstrates that the residual learning can further improve the accuracy of the relighting.

4.3. Comparison with State-of-the-art Methods

First, we compare the MBNet with four state-of-the-art RGB-D SOD methods including ATST [11], DANET [14], RD3D [15] and PGAR [10] as described in the Section 2. Note that, we use the same training set to train these methods. We replace the final convolutional layers of RGB-D SOD networks, so their output is three-channel tensors as relighted images. As shown in Table 3, the MBNet outperforms other methods with a large margin. Our method achieves the best performance on both PSNR and SSIM, which surpasses the second place 1.02 dB and 0.037 in SSIM.

Furthermore, we report the performances of some submissions in the NTIRE 2021 Depth Guide One-to-one Relighting Challenge [32]. The performance is evaluated on the validation and the test dataset and the results are shown in Table 3. Additionally, The Mean Perceptual Score (MPS) is used for final evaulation. The MPS is defined as the average of the normalized SSIM and LPIPS [33] scores. The MBNet produces moderate quality outputs and the 1^{st} place performance in both MPS and SSIM metrics. We also plot some relighted images generated by our method and other participants in Fig. 6. Compared to other methods, our images can remove more shadows and present clear outlines of objects, though some results are not satisfactory.

5. Conclusion

In this paper, to address depth guided image relighting, we develop the multi-modal bifurcated network. This network extracts both depth and image features by a dual bifurcated backbone. To fuse multi-modal features, the dynamic dilated pyramid module is introduced. This module contains densely connected layers and multi-scale kernels to fuse and refine features from a dual bifurcated backbone. Furthermore, to improve the robustness and performance, we propose a new strategy to increase the training image pairs by leveraging extra images. Several experiments implemented on the novel VIDIT [8] dataset proves that our solution achieves the 1^{st} place in terms of MPS and SSIM in the NTIRE 2021 Depth Guided One-to-one Relighting Challenge.

References

[1] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000. 1, 2

[2] W. Matusik, M. Loper, and H. Pfister, "Progressively-refined reflectance functions from natural illumination." in *Rendering Techniques*, 2004. 1, 2

[3] D. Reddy, R. Ramamoorthi, and B. Curless, "Frequency-space decomposition and acquisition of light transport under spatially varying illumination," in *European Conference on Computer Vision*, 2012. 1, 2

[4] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi, "Deep image-based relighting from optimal sparse samples," *ACM Transactions on Graphics (ToG)*, 2018. 1, 2

[5] D. Puthussery, M. Kuriakose, J. C V *et al.*, "Wdrn: A wavelet decomposed relightnet for image relighting," *arXiv preprint arXiv:2009.06678*, 2020. 1, 2

[6] M. El Helou, R. Zhou, S. Süsstrunk, R. Timofte *et al.*, "AIM 2020: Scene relighting and illumination estimation challenge," in *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2020. 1

[7] H.-H. Yang, W.-T. Chen, and S.-Y. Kuo, "S3Net: A single stream structure for depth guided image relighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 1, 2, 4

[8] M. E. Helou, R. Zhou, J. Barthas, and S. Süsstrunk, "Vidit: Virtual image dataset for illumination transfer," *arXiv preprint arXiv:2005.05460*, 2020. 2, 5, 7

[9] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," *arXiv preprint arXiv:2007.06227*, 2020. 2, 3

[10] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European Conference on Computer Vision*, 2020. 2, 7

[11] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 374–390. 2, 7

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2, 3

Table 3: The average SSIM, PSNR, MPS and LPIPS of some submissions over NTIRE 2021 depth guided image relighting validation and testing dataset.

User name	Validation		Testing			
	SSIM	PSNR	MPS	SSIM	LPIPS	PSNR
auy200	0.6937	18.4492	0.7620	0.6874	0.1634	18.8358
aics	0.7069	19.1026	0.7601	0.6799	0.1597	18.8639
lifu	0.7104	19.0048	0.7600	0.6903	0.1702	19.8645
jimmy3505090	0.7069	18.2937	0.7551	0.6772	0.1670	18.2766
DeepBlueAI	0.6757	18.6800	0.7494	0.6879	0.1891	19.8784
Ours	0.7175	19.3558	0.7663	0.6931	0.1605	19.1469

[13] H.-H. Yang and Y. Fu, "Wavelet U-net and the chromatic adaptation transform for single image dehazing," in *IEEE International Conference on Image Processing (ICIP)*, 2019. 2, 3

[14] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-d salient object detection," in *European Conference on Computer Vision*, 2020. 2, 7

[15] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," *arXiv preprint arXiv:2101.10241*, 2021. 2, 7

[16] H.-H. Yang, C.-H. H. Yang, and Y.-C. J. Tsai, "Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2

[17] W.-T. Chen, S.-Y. Yuan, G.-C. Tsai, H.-C. Wang, and S.-Y. Kuo, "Color channel-based smoke removal algorithm using machine learning for static images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2855–2859. 2

[18] W.-T. Chen, H.-Y. Fang, J.-J. Ding, and S.-Y. Kuo, "PMHLD: patch map-based hybrid learning dehazenet for single image haze removal," *IEEE Transactions on Image Processing*, 2020. 2

[19] H.-H. Yang, K.-C. Huang, and W.-T. Chen, "LAFFNet: A lightweight adaptive feature fusion network for underwater image enhancement," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 2

[20] G.-C. Tsai, W.-T. Chen, S.-Y. Yuan, and S.-Y. Kuo, "Efficient reflection removal algorithm for single image by pixel compensation and detail reconstruction," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018. 2

[21] H.-H. Yang, C.-H. H. Yang, and Y.-C. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in *IEEE International Conference on Image Processing (ICIP)*, 2020. 2

[22] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *European Conference on Computer Vision*, 2020. 2, 3

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3, 7

[24] W.-T. Chen, J.-J. Ding, and S.-Y. Kuo, "PMS-net: Robust haze removal based on patch map for single images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[25] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169. 3

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 4

[27] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[28] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, 2016. 4

[29] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711. 4

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] M. El Helou, R. Zhou, S. Süsstrunk, R. 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. 7

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 7