NTIRE 2021 Depth Guided Image Relighting Challenge

Majed El Helou[1] Ruofan Zhou[1] Sabine Süsstrunk[1] Radu Timofte[1] Maitreya Suin*
A. N. Rajagopalan* Yuanzhi Wang* Tao Lu* Yanduo Zhang* Yuntao Wu*
Hao-Hsiang Yang* Wei-Ting Chen* Sy-Yen Kuo* Hao-Lun Luo* Zhiguang Zhang*
Zhipeng Luo* Jianye He* Zuo-Liang Zhu* Zhen Li* Jia-Xiong Qiu*
Zeng-Sheng Kuang* Cheng-Ze Lu* Ming-Ming Cheng* Xiu-Li Shao*
Chenghua Li* Bosong Ding* Wanli Qian* Fangya Li* Fu Li* Ruifeng Deng*
Tianwei Lin* Songhua Liu* Xin Li* Dongliang He* Amirsaeed Yazdani*
Tiantong Guo* Vishal Monga * Ntumba Elie Nsampi * Zhongyun Hu *
Qing Wang * Sabari Nathan* Priya Kansal* Tongtong Zhao* Shanshan Zhao*

Abstract

Image relighting is attracting increasing interest due to its various applications. From a research perspective, image relighting can be exploited to conduct both image normalization for domain adaptation, and also for data augmentation. It also has multiple direct uses for photo montage and aesthetic enhancement. In this paper, we review the NTIRE 2021 depth guided image relighting challenge.

We rely on the VIDIT dataset for each of our two challenge tracks, including depth information. The first track is on one-to-one relighting where the goal is to transform the illumination setup of an input image (color temperature and light source position) to the target illumination setup. In the second track, the any-to-any relighting challenge, the objective is to transform the illumination settings of the input image to match those of another guide image, similar to style transfer. In both tracks, participants were given depth information about the captured scenes. We had nearly 250 registered participants, leading to 18 confirmed team submissions in the final competition stage. The competitions, methods, and final results are presented in this paper.

1. Introduction

Due to its broad utility in research and in practice, image relighting is gaining increasingly more attention. An image relighting method would enable extended data augmentation, by providing images with various illumination settings, and similarly enables illumination domain adaptation as it can transform test images into a standard unique illumination setup. In practice, image relighting is also useful for photo montage and other aesthetic image retouching applications. The task of image relighting is however challenging. The relighting method needs to understand the geometry and illumination of the scene, be able to remove shadows, recast shadows, and occasionally inpaint totally dark areas, on top of the transformation of the illuminant.

The objective of this image relighting challenge is to push forward the state-of-the-art in image relighting, and provide a benchmark to assess competing solutions. We rely on the novel dataset **V**irtual **I**mage **D**ataset for **I**llumination **T**ransfer (VIDIT) [1], which we briefly discuss in the following section. We refer the reader to some previous solutions on this dataset [2, 3, 4, 5, 6, 7, 8, 9], and to the first edition of this challenge [10] for an extensive overview of related datasets and a more detailed discussion about the use of VIDIT. In contrast with the first edition, we exploit depth map information to improve the scene understanding of the methods and in turn improve the final results. Depth is important as it relates to different image degradations such as chromatic aberrations [11, 12, 13], which can cause various depth-dependent blur effects that should be properly synthesized across spectral channels [14], especially if the solutions should be extended to multi-spectral data in future work. More importantly, on top of the importance of geometry understanding, having auxiliary information can generally improve the overall learning [15] and interpretability of the solutions. This additional internal learning is exploited by the participating teams, as discussed in the following sections.

This challenge is one of the NTIRE 2021 associated

Majed El Helou, Ruofan Zhou, Sabine Süsstrunk *(majed.elhelou, sabine.susstrunk)@epfl.ch*, and Radu Timofte *radu.timofte@vision.ee.ethz.ch*, are the challenge organizers, and the other authors are challenge participants.
*Appendix A lists all the teams and affiliations.

challenges: nonhomogeneous dehazing [16], defocus deblurring using dual-pixel [17], depth guided image relighting [18], image deblurring [19], multi-modal aerial view imagery classification [20], learning the super-resolution space [21], quality enhancement of heavily compressed videos [22], video super-resolution [23], perceptual image quality assessment [24], burst super-resolution [25], high dynamic range [26].

2. Depth guided image relighting

2.1. Dataset

All challenge tracks exploit the novel VIDIT [1] dataset that contains 300 training scenes and 90 scenes divided equally between validation set and test set, all being mutually exclusive. Scenes are each captured 40 times: from 8 equally-spaced azimuthal angles, and with 5 color temperatures for the illumination. Image resolution is 1024×1024, and the full resolution is used in both tracks. Additionally, for this edition of the challenge, the associated depth maps are used by the participants. The full dataset can be found online[1], with the exception of the ground-truth test data that is kept private. For reporting purposes, in research papers outside of the challenge, authors typically provide their results on the validation set.

2.2. Challenge tracks

The tracks 1 and 2 are similar to the tracks 1 and 3 of our first edition [10], respectively, but with the addition of depth map information to guide the relighting, and the use of full 1024×1024 resolution images.

Track 1: One-to-one relighting. In this first track, the illumination settings of both the input and the output images are pre-determined and fixed. The objective is therefore to transform an image from its original illumination settings to a known output illumination setup.

Track 2: Any-to-any relighting. The second track allows for more flexibility in the target illumination. More specifically, the target illumination settings are dictated by a guide image similar to style transfer applications.

Evaluation protocol. We evaluate the results using the standard PSNR and SSIM [27] metrics, and the self-reported run-times and implementation details are also provided in Tables 1 and 2. For the final ranking, we define a Mean Perceptual Score (MPS) as the average of the normalized SSIM and LPIPS [28] scores, themselves averaged across the entire test set of each submission

$$0.5 \cdot (S + (1 - L)), \tag{1}$$

where S is the SSIM score, and L is the LPIPS score.

Challenge phases. (1) Development phase: registered participants have access to the full training data (input and ground-truth images and depth maps), and to the input data of the validation set. A leader board enables the participants to get immediate automated feedback on their performance and their ranking relative to the other competing teams, by uploading their validation set output results to our server. (2) Testing: registered participants have access to the input of the test sets, and can upload their results in a similar way as during the development phase. However, the difference is that results are not shown to the participants to counter any potential overfitting. We only accept a final submission of the test set outputs when it is accompanied by reproducible open-source code, and a fact sheet containing all the details of the proposed solution.

3. Challenge results

The results of our tracks 1 and 2 are collected in Tables 1 and 2, respectively. Challenge solutions are described in the following section for each team. We also show some visual results from top-performing solutions with the associated input, depth, and ground-truth information in Fig. 1 for track 1, and in Fig. 2 for track 2. Generally, all solutions outperform the results we had in the first edition when depth information was not used. This strongly supports the importance of depth maps for the image relighting tasks.

4. Track 1 methods

4.1. AICSNTU-MBNet: Multi-modal Bifurcated Network for depth guided image relighting (MBNet)

The method details are described in [29]. As shown in Fig. 3, we rely on the HDFNet [30] to fulfill the depth guided image relighting task. As shown in Fig. 7, the proposed network consists of two structures to extract the depth and image features. We apply the two ResNet50 networks as backbones. Three depth and image features from conv3, conv4 and conv5 are fused to achieve representative features. To fuse these two features with multi-receptive fields, we leverage the densely connected architecture to generate the combined features with rich texture and structure information. Then, these features are fed to the dynamic dilated pyramid module (DDPM) [30] that can generate a more discriminative result. Then, the output of DDPM combines with the output of the decoder by convolving with the multi-scale convolution kernels [31, 32]. In the decoder part, similar to U-net, we gradually magnify the feature maps and implement a skip connection to concatenate the identical-size feature maps. Furthermore, we make our network learn

[1] https://github.com/majedelhelou/VIDIT

Team	Author	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time	Platform	GPU
AICSNTU-MBNet	HaoqiangYang	0.7663	0.6931	0.1605	19.1469	2.88s	PyTorch	Tesla V100
iPAL-RelightNet	auy200	0.7620	0.6874	0.1634	18.8358	0.53s	PyTorch	Titan XP
NTUAICS-ADNet	aics	0.7601	0.6799	0.1597	18.8639	2.76s	PyTorch	Tesla V100
VUE	lifu	0.7600	0.6903	0.1702	19.8645	0.23s	PyTorch	P40
NTUAICS-VGG	jimmy3505090	0.7551	0.6772	0.1670	18.2766	2.12s	PyTorch	Tesla V100
DeepBlueAI	DeepBlueAI	0.7494	0.6879	0.1891	19.8784	0.17s	PyTorch	Tesla V100
usuitakumi	usuitakumi	0.7229	0.6260	0.1801	16.8249	0.04s	PyTorch	Tesla V100
MCG-NKU	NK_ZZL	0.7147	0.6191	0.1896	19.0856	0.33s	PyTorch	RTX TITAN
alphaRelighting	lchia	0.7101	0.6084	0.1882	15.8591	0.04s	PyTorch	Tesla K80
Wit-AI-lab	MDSWYZ	0.6966	0.6113	0.2181	17.5740	0.9s	PyTorch	RTX 2080Ti
Couger AI	Sabarinathan	0.6475	0.5469	0.2518	18.2938	0.015s	Tensorflow	GTX 1070

Table 1. NTIRE 2021 Depth-Guided Image Relighting Challenge Track 1 (One-to-one relighting) results. The MPS, used to determine the final ranking, is computed following Eq. (1).

Team	Author	**MPS** ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Run-time	Platform	GPU
DeepBlueAI	DeepBlueAI	0.7675	0.7087	0.1737	20.7915	0.17s	PyTorch	Tesla V100
VUE	lifu	0.7671	0.6874	0.1532	19.8901	0.3s	PyTorch	P40
AICSNTU-SSS	HaoqiangYang	0.7609	0.6784	0.1566	19.2212	2.04s	PyTorch	Tesla V100
NPU-CVPG	elientumba	0.7423	0.6508	0.1661	18.6039	0.674s	PyTorch	TITAN RTX
iPAL-RelightNet	auy200	0.7341	0.6711	0.2028	20.1478	0.51s	PyTorch	Titan XP
IPCV_IITM	ms_ipcv	0.7172	0.6052	0.1708	18.7472	0.3s	PyTorch	TitanX
Wit-AI-lab	MDSWYZ	0.6976	0.5985	0.2032	17.5222	6s	PyTorch	RTX 2070

Table 2. NTIRE 2021 Depth-Guided Image Relighting Challenge Track 2 (Any-to-any relighting) results. The MPS, used to determine the final ranking, is computed following Eq. (1).

Figure 1. A challenging example image from the NTIRE 2021 Image Relighting Challenge Track 1 (One-to-one relighting) with the output results of some top submission methods.

Figure 2. A challenging example image from the NTIRE 2021 Image Relighting Challenge Track 2 (Any-to-any relighting) with the output results of some top submission methods.

the residual instead of the full images: the final output is the difference of the original image and the relit image.

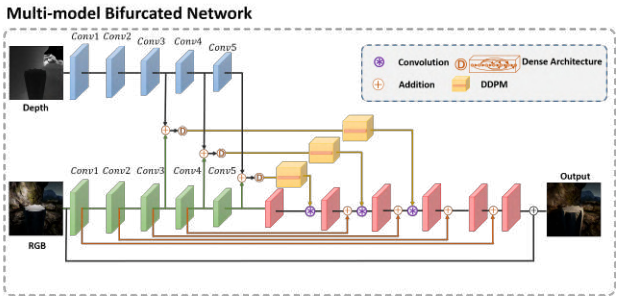

Figure 3. The architecture of the proposed multi-modal bifurcated network. The network consists of two streams: depth stream and RGB-image stream. We use the dense architecture and DDPM for better feature extraction.

4.2. iPAL-RelightNet: One-to-One Intrinsic Decomposition-Direct RelightNet (OIDDR-Net)

This approach [33] exploits two different strategies for generating the relit image with new illumination settings (Fig. 4). In the first strategy, the relit image is estimated by first predicting the albedo (material reflectance properties) and shading (illumination and geometry properties) of the scene. The intial shading estimation is refined using the normal vectors of the scene which in particular leads to better addition or removal of deep shadows. The rendering rule [34] is then used to estimate the relit image ($I_{intrinsic\text{-}relit}$). In the second strategy, the relit image ($I_{direct\text{-}relit}$) is generated based on a black box approach, in which the model learns to predict the output based on the ground-truth images and loss terms in the training stage. The two estimates are fused to generate the final relit output ($I_{final\text{-}relit}$) using a spatial weight map (w), which is learned during the training stage. The source code for the two tracks is made available online[2].

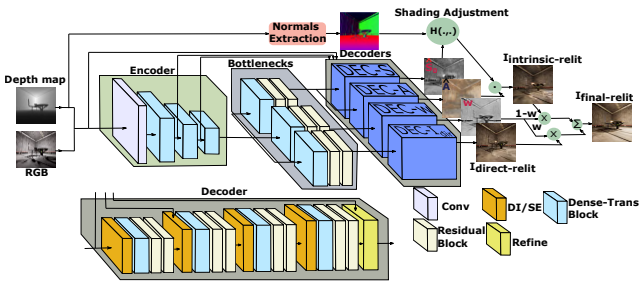

Figure 4. OIDDR-Net architecture.

[2] https://github.com/yazdaniamir38/Depth-guided-Image-Relighting

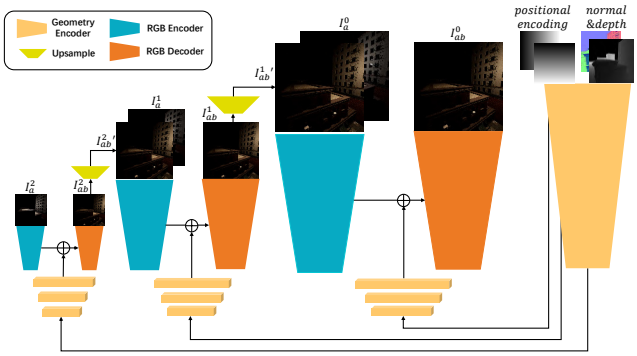

Figure 5. Overview of the proposed pyramidal depth guided relighting network.

4.3. MCG-NKU: Pyramidal Depth Guided Relighting Network

This method handles the relighting task in a coarse-to-fine manner, which has two exclusive branches designed for RGB input and depth. The depth branch has a cascade structure. It can provide diverse information for different scales and attempts to extract high-level information. The RGB branch has three encoders to tackle the input at different scales and three corresponding decoders that take features from the RGB encoders and the depth branch as input and generates the relit images. We find that the surface normal is an important prior that induces the network to learn local illumination. Additionally, we utilize an explicit approach with intuition borrowed from Transformer networks. Specifically, we emphasize the relative positional information by using linear positional encoding that encodes the x-axis/y-axis relative position through two feature maps. The overall structure of our proposed method is illustrated in Fig. 5.

4.4. VUE: Deep RGB and Frequency Domain CNNs for Image Relighting (DRFNet)

We make use of both the RGB domain [35] and the frequency domain [2, 36, 37] CNNs for one-to-one image relighting by end-to-end training, and the overall pipeline of our solution is illustrated in Fig. 6. Intuitively, the frequency domain and the RGB domain of images are two distinctive domains. Building mappings from one lighting setup to the other illumination settings in the RGB space and frequency space could be complementary. These CNNs are all designed as U-Net encoder-decoder architectures with skip connections. Specifically, for RGB CNNs, the encoder is composed of N "Conv-IN-ReLU-Pooling" blocks and its decoder contains N "Upsample-Conv-IN-ReLU" building blocks; for frequency domain CNNs, the corresponding building blocks in its encoder and decoder are "DWT-Spatial2Depth-Conv-IN-ReLU" and "Depth2Spatial-IDWT-Conv-IN-ReLU", re-

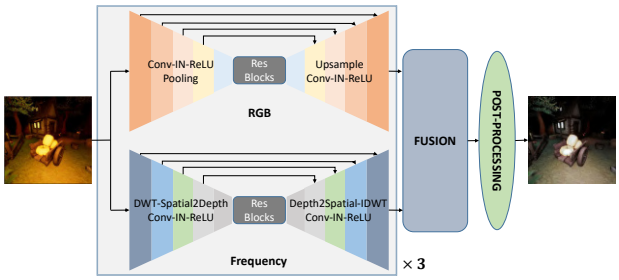

Figure 6. Overall architecture of the DRFNet in track 1.

spectively. There are also M ResBlocks between the encoder and decoder. The fusion module simply performs per-pixel averaging. As validated in the developing phase of this challenge, slightly lowering the luminance and enhancing the color saturation are effective post-processing strategies. Note that depth images are provided, thus in our implementation the CNNs take as input the concatenation of the RGB image and its depth image in both training and testing phases.

4.5. DeepBlueAI: Deep Fusion Network for Image Relighting (DFNIR)

The team uses the method and model weights trained in track 2 and simply selects one image with the pre-defined illumination settings (East, 4500K) from the training set as the guide image to get the results for track 1. More details can be found in the description of the track 2 solution 5.1.

4.6. NTUAICS-ADNet: Depth Guided Image Relighting by Asymmetric Dual-stream Network

The overview of our solution ADNet is illustrated in Fig. 7. As shown in Fig. 7, we first use a powerful network to extract the RGB-image feature representation. Both the encoder and the decoder are based on the Res2net [38] network and the decoder contains an attention mechanism (Attention) [39] and enhanced modules (EM), which is motivated by [40]. The details of the EM and Attention approaches are demonstrated in Fig. 7 (the orange box and the blue box). The attention block consists of both spatial and channel attention mechanisms. For the feature extraction of the depth map, we utilize the smaller backbone, that is, the ShuffleNet V2 [41] to extract the depth map feature representation. Moreover, we combine the features in Conv 2, Conv 3, and Conv 4 with the multi-modal fusion (MMF) module. This module contains a convolutional layer to make the size of the depth feature representation identical to that of the RGB-image feature representation. And these two feature representations are multiplied and summarised to become refined features. The latter features are concatenated and combined with the representation from the decoder features. It is noted that the target output of the network is designed so as to learn the residual components between the output and input instead of the whole image.

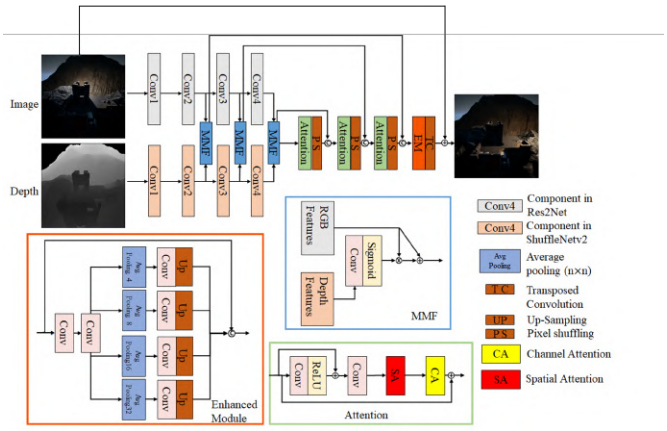

Figure 7. The proposed ADNet. This model applies the Res2Net and ShuffleNetv2 as the encoder to extract the image and depth map feature representations. We design the Multi-Modal Fusion block to integrate these two feature representations. In the decoder parts, we use the attention mechanism and the enhanced module to refine features and relight the image.

4.7. NTUAICS-VGG: Two Birds, One Stone: VGG-based Network with VGG-based Loss for Depth Guided Image Relighting

Inspired by [42], we design two branches to extract the RGB-image feature representation and the depth feature representation, respectively. For the image features, we apply the VGG-16 to extract multi-scale representations. Then, these features are fed into the multi-scale residual block (MSR) to generate the initial estimation of the relit image. The architecture of MSR is presented in Fig. 8. Then, the initial estimation of the relit image is fed into a series of guided residual blocks (GR) [42] for further refinement. The GR blocks use the features extracted by each layer in the RGB branch and depth branch to reconstruct the details of the final output. For the depth feature representation, we simply apply a single layer of convolution (3×3 convolutional kernels with 64 channels) to extract three kinds of features. Then, with the feature extraction of the two branches and the refinement process, the final relit output is generated.

We rely on the $\ell 1$ Chabonnier loss [43], the Wavelet SSIM loss [44], and the visual perceptual loss [45] to train our model. We use the Adam optimizer with a learning rate of 0.0001 and the learning rate decreases by 0.1 every 50 epochs. The total number of training epochs is 200 and the training takes 6 hours when setting the batch size to 3 and using V100 GPUs.

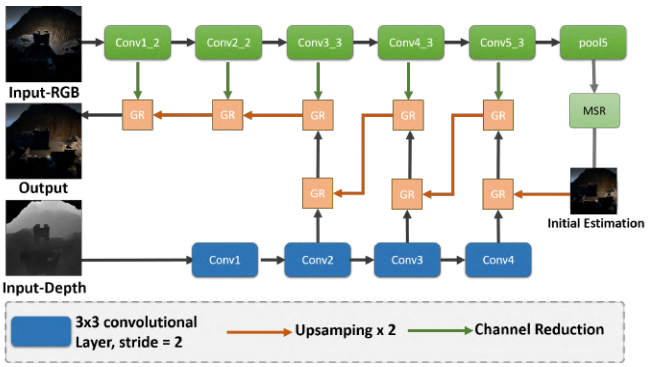

Figure 8. The architecture of the proposed relighting network. The features of RGB and depth information are fed into GR blocks for generating the final output.

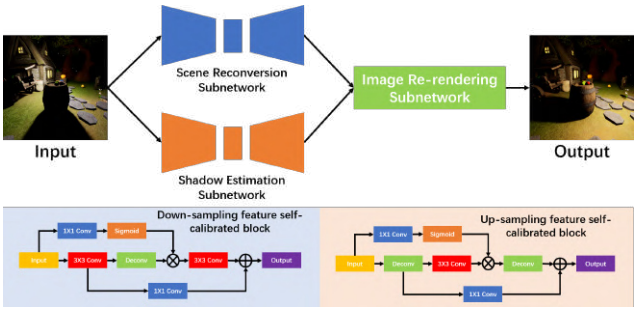

Figure 9. Overall architecture of the MCN.

4.8. Wit-AI-lab: Multi-scale Self-calibrated Network for Image Light Source Transfer (MCN)

Inspired by [46], the architecture of MCN is shown in Fig. 9, which consists of three parts: scene reconversion subnetwork, shadow estimation subnetwork, and image re-rendering subnetwork. First, the input image is processed in the scene reconversion subnetwork to extract primary scene structures. At the same time, the shadow estimation subnetwork aims to the change of the lighting effect. Finally, the image re-rendering subnetwork learns the target color temperature and reconstructs the image with primary scene structure information and predicted shadows. Considering that image light source transfer is a task of recalibrating the light source settings, the team proposes a novel downsampling feature self-calibrated block (DFSB) and upsampling feature self-calibrated block (UFSB) as the basic blocks for scene reconversion and shadow estimation tasks to calibrate feature information iteratively, thereby improving the performance of light source transfer. In addition, the team designs the multi-scale feature fusion method for the scene reconversion task, which provides more accurate primary scene structure for the image re-rendering task.

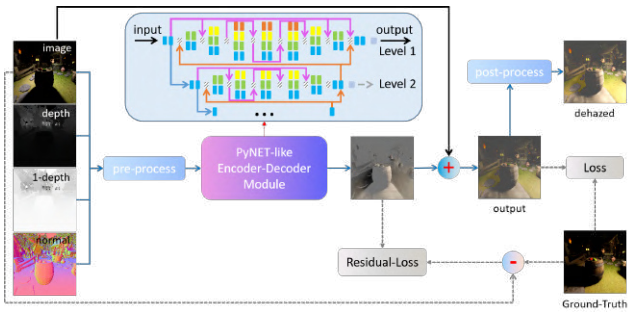

Figure 10. Overall architecture of the LTNet.

4.9. alphaRelighting: Light Transfer Network for Depth Guided Image Relighting (LTNet)

As shown in Fig. 10, the key modules of the proposed LTNet are the global skip connection, the PyNET-like encoder-decoder network and the additional residual loss L_{RES}. The global skip connection enables the input to be added to the output of the encoder-decoder network directly. Then, the encoder-decoder network is set to learn the residual between the ground-truth image (I_{GT}) and the input image (I_{input}). Besides, we add a residual ground-truth ($I_{GT} - I_{input}$) to supervise the training as well as the original loss supervised by I_{GT}. LTNet uses a PyNET-like encoder-decoder network [47] to learn the residual of the ground-truth image and the input image, which adopts a slightly dense connection and a number of convolution blocks in parallel with convolution filters of different sizes. In LTNet, the global skip connection plays a key role because it forces the network to encode the variance of the input image and the target.

5. Track 2 methods

5.1. DeepBlueAI: Deep Fusion Network for Image Relighting (DFNIR)

The team designed a U-Net style encoder-decoder structure with RGB-D image maps as input and directly outputs the relit RGB image, as shown in Fig. 11. In the encoder network, inspired by [48], a fusion network is designed to extract feature representations of the input and guide images. To better recognize the illumination of the input and guide images, a Dilated Residual Module (DRM) is designed and used at each level, with three convolutional layers with dilation of 1,3, and 5. Based on successful previous work[35], the team also designed and added an illumination-estimator network and an illumination-to-feature network. The difference is that in our solution we completely replace the features of the input image with the features of the guide image output obtained from the illumination-to-feature net. In addition, we use in the skip connections and the decoder network a Residual block with

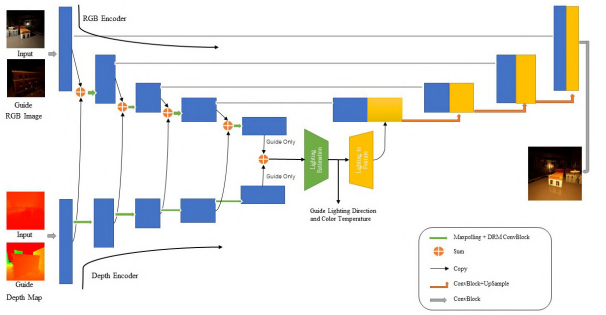

Figure 11. Overall architecture of the DFNIR.

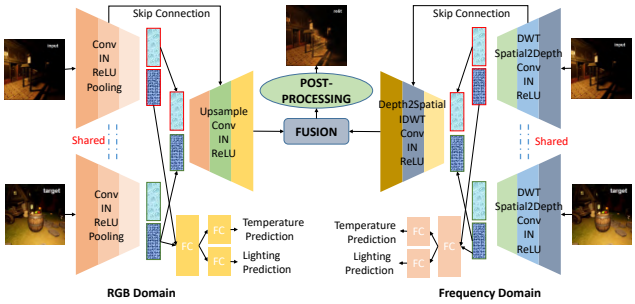

Figure 12. Overall architecture of the DRFNet in track 2.

two convolutional layers.

5.2. VUE: Deep RGB and Frequency Domain CNNs for Image Relighting (DRFNet)

We rely on the same network architecture as in track 1. We utilize both the RGB domain and the frequency domain CNNs for any-to-any image relighting by end-to-end training, and the overall pipeline of our solution is illustrated in Fig. 12. The difference is that we take both input image and guide image as the inputs of our network. To support any-to-any mapping, we explicitly split feature maps generated by the encoder into "content feature" and "lighting feature". The "lighting feature" is trained with extra supervision from the temperature prediction and illumination-direction prediction branches. Then, the learned lighting feature of the guide image and the content feature of the input image are combined for decoding the desired relighting output. The fusion module simply performs per-pixel averaging.

5.3. AICSNTU-SSS: A Single Stream structure for depth guided image relighting (S3Net)

The method directly combines the original image, original depth map, guide image, and guide depth map as the input. This input is seen as the 8-channel tensor and the output is the 3-channel image, refered to as the relit image. We rely on the Res2Net101 [38] network for our backbone. After the input is passed through this backbone, multi-scale features are obtained. We use the bottom feature to connect to the decoder. The decoder consists of the stacking of convolutional layers to recover the size of the feature maps. We use skip connect to merge the last three feature maps from the backbone to their corresponding feature maps. At the last two layers, we also add an attention module that contains both spatial and channel attention and an enhanced module [40, 49, 50] to refine the features. Finally, the decoder outputs the three-channel tensor referred to as the relit image. The model is illustrated in Fig. 13. The details of this method are thoroughly explained in [51].

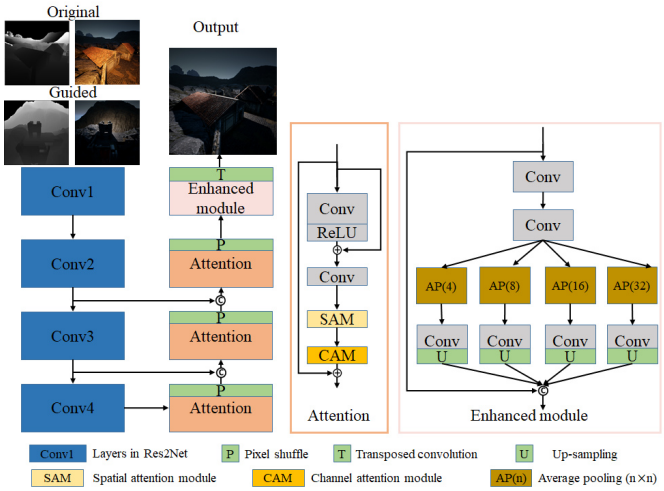

Figure 13. The proposed network for depth guided image relighting. This model applies the Res2Net and an encoder to extract both the image and depth map features. All images and depth maps are concatenated as input. In the decoder parts, we use an attention mechanism including channel and spatial attention and the enhanced module to refine features and relight the image.

5.4. iPAL-RelightNet: Any-to-any Multiscale Intrinsic-Direct RelightNet (AMIDR-Net)

Similar to one-to-one relighting, the proposed method [33] makes use of two strategies (intrinsic decomposition and black box) to generate the final relit output. The proposed network architecture (Fig. 14) is based on the U-Net [52], while the encoder exploits the DenseNet [53] pretrained layers. AMIDR-Net is different from OIDDR-Net in that it benefits from a multiscale block [54] for extracting more representative features from the input and the guide. Additionally, a lighting estimation network, which is trained separately over the training set and the corresponding illumination parameters, is incorporated first to provide the decoders with the illumination features of the guide, and second to compare the illumination of the relit output and the guide (via calculating a loss term: $L_{Lighting}$).

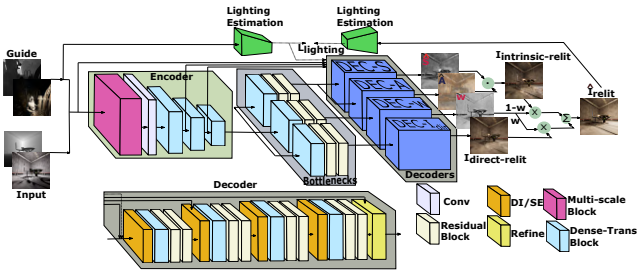

Figure 14. AMIDR-Net architecture overview.

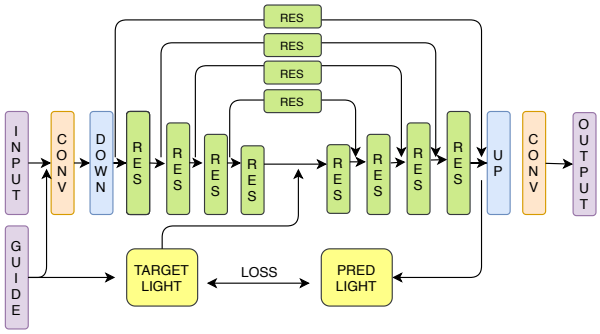

Figure 15. DRNIR network architecture overview.

5.5. IPCV IITM: Deep Residual Network for Image Relighting (DRNIR)

We illustrate in Fig. 15 the structure of the proposed residual network with skip connections, based on the hourglass network [55]. The network has an encoder-decoder structure [56, 57, 58] with skip connections similar to [59]. Residual blocks are used in the skip connections, and BatchNorm and ReLU non-linearity are used in each of the blocks. The encoder features are concatenated with the decoder features of the same level. The network takes the input image and directly produces the target image. Our solution converts the input RGB images to LAB for better processing. To reduce the memory consumption without harming the performance, the team uses a pixel-shuffle block [60] to downsample the image. In addition, the depth map is concatenated with the input image. The network is fist trained using the $\ell 1$ loss, then fine-tuned with the MSE loss. Experiments with adversarial loss did not lead to stable training. The learning rate of the Adam optimizer is 0.0001 with a decay cycle of 200 epochs. Data augmentation is used to make the network more robust.

5.6. Wit-AI-lab : Self-calibrated Relighting Network for any-to-any Relighting (SRN)

We show in Fig. 16 the architecture of our SRN method, which contains a normalization network and a relighting network. The normalization network aims to produce uniformly-lit white-balanced images, and the relighting network uses the latent feature representation of the guide image and uniformly-lit image produced by the normalization network to re-render the target image. Considering that any-to-any relighting is a task of recalibrating the light source settings, the team proposes a novel self-calibrated block as the basic block of the feature encoder to calibrate the feature information for the normalization network and the relighting network iteratively, thereby improving the performance of any-to-any relighting.

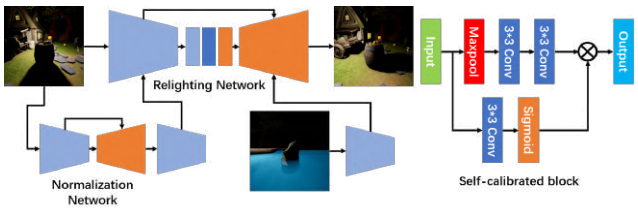

Figure 16. Overall architecture of the SRN.

5.7. NPU-CVPG : Shadow Guided Refinement Network for Any-to-any Relighting

Our proposed framework uses the self attention auto-encoder (SA-AE) of Hu *et al.* [35] as its backbone architecture. Two additional modules are added; namely a shadow synthesis module, and a shadow refinement module. The shadow synthesis module is designed as a conditional auto-encoder where the illumination direction is used as the conditional variable to control the depth for the shadow image generation process. This is done by first transforming the input image's depth and the predicted illumination direction into a common feature space, then fusing those features using an element-wise sum. The shadow refinement module is a computational unit used to further refine the relighting features. The entire framework is trained end-to-end.

6. Conclusion

We presented the competition methods along with the benchmark results for our one-to-one and any-to-any depth guided image relighting challenges. The participating teams obtained high-quality results that are very competitive across the top-scoring solutions. We noted a significant improvement in the overall results compared to our first edition, and this was the case for all participating teams. This significant improvement strongly supports the use of depth information in relighting tasks. Hence, we strongly encourage future work to exploit depth information, and generally scene geometry information, for image relighting.

Acknowledgements

We thank all NTIRE 2021 sponsors https://data.vision.ee.ethz.ch/cvl/ntire21/. We also note that all tracks were supported by the CodaLab infrastructure https://competitions.codalab.org.

A. Teams and affiliations

NTIRE challenge organizers
Members: Majed El Helou, Ruofan Zhou, Sabine Süsstrunk ({*majed.elhelou,sabine.susstrunk*}*@epfl.ch*, EPFL, Switzerland), and Radu Timofte (*radu.timofte@vision.ee.ethz.ch*, ETH Zürich, Switzerland).

– AICSNTU-MBNet –
Members: Hao-Hsiang Yang (*islike8399@gmail.com*), Wei-Ting Chen, Hao-Lun Luo, Sy-Yen Kuo.
Affiliation: ASUS Intelligent Cloud Services, ASUSTeK Computer Inc; Graduate Institute of Electronics Engineering, National Taiwan University; Department of Electrical Engineering, National Taiwan University.

– AICSNTU-SSS –
Members: Hao-Hsiang Yang (*islike8399@gmail.com*), Wei-Ting Chen, Sy-Yen Kuo.
Affiliation: ASUS Intelligent Cloud Services, ASUSTeK Computer Inc; Graduate Institute of Electronics Engineering, National Taiwan University; Department of Electrical Engineering, National Taiwan University.

– alphaRelighting –
Members: Chenghua Li (*lichenghua2014@ia.ac.cn*), Bosong Ding, Wanli Qian, Fangya Li.
Affiliation: Institute of Automation, Chinese Academy of Sciences, China.

–Couger AI–
Members:Sabari Nathan (*sabari@couger.co.jp*), Priya Kansal.
Affiliation: Couger Inc.

– DeepBlueAI –
Members: Zhipeng Luo (*luozp@deepblueai.com*), Zhiguang Zhang, Jianye He.
Affiliation: DeepBlue Technology (Shanghai) Co., Ltd, China.

– iPAL-RelightNet –
Members: Amirsaeed Yazdani (*amiryazdani@psu.edu*), Tiantong Guo, Vishal Monga.
Affiliation:
The Pennsylvania State University
School of Electrical Engineering and Computer Science
The Information Processing and Algorithms Laboratory (iPAL).

– IPCV_IITM –
Members: Maitreya Suin (*maitreyasuin21@gmail.com*), A. N. Rajagopalan.
Affiliation: Indian Institute of Technology Madras, India.

– MCG-NKU –
Members: Zuo-liang Zhu (*nkuzhuzl@gmail.com*), Zhen Li, Jia-Xiong Qiu, Zeng-Sheng Kuang, Cheng-Ze Lu, Ming-Ming Cheng, Xiu-Li Shao.
Affiliation: Nankai University, China.

– NPU-CVPG –
Members: Ntumba Elie Nsampi (*elientumba@mail.nwpu.edu.cn*), Zhongyun Hu, Qing Wang.
Affiliation:
Computer Vision and Computational Photography Group, School of Computer science, Northwestern Polytechnical University, China.

– NTUAICS-ADNet –
Members: Wei-Ting Chen (*f05943089@ntu.edu.tw*), Hao-Hsiang Yang, Hao-Lun Luo, Sy-Yen Kuo.
Affiliation: Graduate Institute of Electronics Engineering, National Taiwan University; ASUS Intelligent Cloud Services, ASUSTeK Computer Inc; Department of Electrical Engineering, National Taiwan University.

– NTUAICS-VGG –
Members: Hao-Lun Luo (*r08921051@ntu.edu.tw*), Hao-Hsiang Yang, Wei-Ting Chen, Sy-Yen Kuo.
Affiliation: Department of Electrical Engineering, National Taiwan University; ASUS Intelligent Cloud Services, ASUSTeK Computer Inc; Graduate Institute of Electronics Engineering, National Taiwan University.

– usuitakumi –
Members: Tongtong Zhao (*daitoutiere@gmail.com*), Shanshan Zhao.
Affiliation: Dalian Maritime University; China Everbright Bank.

– VUE –
Members: Fu Li (*lifu@baidu.com*), Ruifeng Deng, Tianwei Lin, Songhua Liu, Xin Li, Dongliang He.
Affiliation: Department of Computer Vision (VIS), Baidu Inc.

– Wit-AI-lab –
Members: Yuanzhi Wang (*w906522992@gmail.com*), Tao Lu, Yanduo Zhang, Yuntao Wu.
Affiliation: Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, China.

References

[1] M. El Helou, R. Zhou, J. Barthas, and S. Süsstrunk, "VIDIT: Virtual image dataset for illumination transfer," *arXiv preprint arXiv:2005.05460*, 2020. 1, 2

[2] D. Puthussery, M. Kuriakose, J. C V *et al.*, "WDRN: A wavelet decomposed relightnet for image relighting," *arXiv preprint arXiv:2009.06678*, 2020. 1, 4

[3] P. Gafton and E. Maraz, "2D image relighting with image-to-image translation," *arXiv preprint arXiv:2006.07816*, 2020. 1

[4] A. P. Dherse, M. N. Everaert, and J. J. Gwizdała, "Scene relighting with illumination estimation in the latent space on an encoder-decoder scheme," *arXiv preprint arXiv:2006.02333*, 2020. 1

[5] Z. Hu, X. Huang, Y. Li, and Q. Wang, "SA-AE for any-to-any relighting," in *European Conference on Computer Vision*. Springer, 2020, pp. 535–549. 1

[6] L. Dong, Y. Zhu, Z. Jiang, X. He, Z. Meng, C. Li, C. Leng, and J. Cheng, "An ensemble neural network for scene relighting with light classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 581–595. 1

[7] L.-W. Wang, W.-C. Siu, Z.-S. Liu, C.-T. Li, and D. P. Lun, "Deep relighting networks for image light source manipulation," in *European Conference on Computer Vision*, 2020. 1

[8] S. D. Das, N. A. Shah, and S. Dutta, "MSR-Net: Multi-scale relighting network for one-to-one relighting," in *NeurIPS Workshops*. 1

[9] S. D. Das, N. A. Shah, S. Dutta, and H. Kumar, "DSRN: an efficient deep network for image relighting," *arXiv preprint arXiv:2102.09242*, 2021. 1

[10] 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, 2

[11] B. Llanos and Y.-H. Yang, "Simultaneous demosaicing and chromatic aberration correction through spectral reconstruction," in *IEEE Conference on Computer and Robot Vision (CRV)*, 2020, pp. 17–24. 1

[12] M. El Helou, F. Dümbgen, and S. Süsstrunk, "AAM: An assessment metric of axial chromatic aberration," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2486–2490. 1

[13] J. Zhao, Y. Hou, Z. Liu, H. Xie, and S. Liu, "Modified color CCD moiré method and its application in optical distortion correction," *Precision Engineering*, 2020. 1

[14] M. El Helou, Z. Sadeghipoor, and S. Süsstrunk, "Correlation-based deblurring leveraging multispectral chromatic aberration in color and near-infrared joint acquisition," in *International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1402–1406. 1

[15] M. El Helou and S. Süsstrunk, "Blind universal Bayesian image denoising with Gaussian noise level learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 4885–4897, 2020. 1

[16] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, R. Timofte *et al.*, "NTIRE 2021 nonhomogeneous dehazing challenge report," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[17] A. Abuolaim, R. Timofte, M. S. Brown *et al.*, "NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[18] M. El Helou, R. Zhou, S. Süsstrunk, R. Timofte *et al.*, "NTIRE 2021 depth guided image relighting challenge," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[19] S. Nah, S. Son, S. Lee, R. Timofte, K. M. Lee *et al.*, "NTIRE 2021 challenge on image deblurring," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[20] J. Liu, O. Nina, R. Timofte *et al.*, "NTIRE 2021 multi-modal aerial view object classification challenge," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[21] A. Lugmayr, M. Danelljan, R. Timofte *et al.*, "NTIRE 2021 learning the super-resolution space challenge," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[22] R. Yang, R. Timofte *et al.*, "NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[23] S. Son, S. Lee, S. Nah, R. Timofte, K. M. Lee *et al.*, "NTIRE 2021 challenge on video super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[24] J. Gu, H. Cai, C. Dong, J. S. Ren, Y. Qiao, S. Gu, R. Timofte *et al.*, "NTIRE 2021 challenge on perceptual image quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[25] G. Bhat, M. Danelljan, R. Timofte *et al.*, "NTIRE 2021 challenge on burst super-resolution: Methods and results," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[26] E. Pérez-Pellitero, S. Catley-Chandar, A. Leonardis, R. Timofte *et al.*, "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 2

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595. 2

[29] H.-H. Yang, W.-T. Chen, H.-L. Luo, and S.-Y. Kuo, "Multi-modal bifurcated network for depth guided image relighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 2

[30] 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

[31] 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, pp. 11 681–11 689. 2

[32] 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. 2

[33] A. Yazdani, T. Guo, and V. Monga, "Physically inspired dense fusion networks for relighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 4, 7

[34] J. T. Kajiya, "The rendering equation," *SIGGRAPH Comput. Graph.*, vol. 20, no. 4, p. 143–150, Aug. 1986. [Online]. Available: https://doi.org/10.1145/15886.15902 4

[35] Z. Hu, X. Huang, Y. Li, and Q. Wang, "Sa-ae for any-to-any relighting," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 535–549. 4, 6, 8

[36] M. El Helou, R. Zhou, and S. Süsstrunk, "Stochastic frequency masking to improve super-resolution and denoising networks," in *ECCV*, 2020. 4

[37] R. Zhou, F. Lahoud, M. El Helou, and S. Süsstrunk, "A comparative study on wavelets and residuals in deep super resolution," *Electronic Imaging*, no. 13, pp. 135–1, 2019. 4

[38] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019. 5, 7

[39] 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*, vol. 29, pp. 6773–6788, 2020. 5

[40] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8160–8168. 5, 7

[41] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131. 5

[42] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 520–538. 5

[43] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[44] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2628–2632. 5

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016. 5

[46] L.-W. Wang, W.-C. Siu, Z.-S. Liu, C.-T. Li, and D. P. Lun, "Deep relighting networks for image light source manipulation," in *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2020, pp. 550–567. 6

[47] A. Ignatov, L. Van Gool, and R. 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. 6

[48] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, November 2016. 6

[49] 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*. Springer, 2020, pp. 754–770. 7

[50] 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. 7

[51] 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. 7

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015. 7

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. 7

[54] Y. Zhao, L. Po, Q. Yan, W. Liu, and T. Lin, "Hierarchical regression network for spectral reconstruction from rgb images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1695–1704. 7

[55] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs, "Deep single-image portrait relighting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7194–7202. 8

[56] K. Purohit, M. Suin, P. Kandula, and R. Ambasamudram, "Depth-guided dense dynamic filtering network for bokeh effect rendering," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3417–3426. 8

[57] M. Suin, K. Purohit, and A. Rajagopalan, "Degradation aware approach to image restoration using knowledge distillation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 162–173, 2020. 8

[58] ——, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3606–3615. 8

[59] 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, pp. 4700–4708. 8

[60] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883. 8