Physically Inspired Dense Fusion Networks for Relighting

Amirsaeed Yazdani
amiryazdani@psu.edu

Tiantong Guo
tiantong@ieee.org

Vishal Monga
vmonga@engr.psu.edu

Abstract

*Image relighting has emerged as a problem of significant research interest inspired by augmented reality applications. Physics-based traditional methods, as well as black box deep learning models, have been developed. The existing deep networks have exploited training to achieve a new state of the art; however, they may perform poorly when training is limited or does not represent problem phenomenology, such as the addition or removal of dense shadows. We propose a model which enriches neural networks with physical insight. More precisely, our method generates the relighted image with new illumination settings via two different strategies and subsequently fuses them using a weight map (w). In the first strategy, our model predicts the material reflectance parameters (**albedo**) and illumination/geometry parameters of the scene (**shading**) for the relit image (we refer to this strategy as **intrinsic image decomposition (IID)**). The second strategy is solely based on the black box approach, where the model optimizes its weights based on the ground-truth images and the loss terms in the training stage and generates the relit output directly (we refer to this strategy as **direct**). While our proposed method applies to both **one-to-one** and **any-to-any** relighting problems, for each case we introduce problem-specific components that enrich the model performance: 1) For one-to-one relighting we incorporate normal vectors of the surfaces in the scene to adjust gloss and shadows accordingly in the image. 2) For any-to-any relighting, we propose an additional multiscale block to the architecture to enhance feature extraction. Experimental results on the VIDIT 2020 and the VIDIT 2021 dataset (used in the NTIRE 2021 relighting challenge) reveals that our proposal can outperform many state-of-the-art methods in terms of well-known fidelity metrics and perceptual loss.*

1. Introduction

Image enhancement problems have experienced significant recent research activity inspired by the proliferation of mobile devices and the availability of training data designed for particular enhancement goals. Image relighting, which is changing the illumination settings of an image, is one of these applications that has attracted significant attention. Another important reason for this growth is the development of **augmented reality** (**AR**), **virtual reality** (VR) based services such as online shopping, online teaching, and games, where the gloss and shadows of the scene should be adjusted based on the change in direction of light and its color temperature. On the other hand, controlling the light source in the level of the photography skills of an amateur user is not trivial, which in turn necessitates the development of techniques for relighting.

From a physical viewpoint, the illumination of an image depends on many factors including the material reflectance property, geometry of the objects in the image, and the number of light sources. For a given light source (L_{ω_i}) and with Lambertian reflectance assumption, the image formation follows the rendering rule [22, 35]:

$$L_{\omega_o} = \int_{\omega_i \in \Omega_o} f(\omega_i, \omega_o) L_{\omega_i} < n, \omega_i > d\omega_i \qquad (1)$$

Here ω_i and ω_o denote the input and output light direction relative to the surface normal n. L_{ω_i} and L_{ω_o} are the incident and reflected lights, and $f(.,.)$ is the bidirectional reflectance distribution function (BRDF) and $< n, \omega_i >$ is the attenuation factor. This equation is usually simplified by assuming: $A = f(\omega_i, \omega_o)$ (constant) and $S = \int_{\omega_i \in \Omega_o} L_{\omega_i} < n, \omega_i > d\omega_i$. Where A denotes albedo and preserves the reflectance properties of the objects and S denotes shading, which holds the illumination properties of the image. The simplified equation is often used as the rendering rule as the original equation is computationally complex.

Deep learning methods have achieved state-of-the-art results for a vast variety of imaging inverse problems. High dynamic range (HDR) imaging algorithms [32, 23] focus on increasing the local contrast of a low dynamic range image. Dehazing algorithms [30, 48] seek for removing the haze artifacts caused by floating particles in the atmosphere. Shadow removal [26] and light enhancement [15] methods focus on enhancing the lighting and removing the artifacts in the image with the existing light source. While all the aforementioned methods deal with adjusting the parameters affected by the lighting of the image, they don't manipulate actual illumination parameters and can not deal with complexities of relighting. Therefore, we are still in the early stages of relighting research. Existing algorithms focus on

particular objects such as portraits or faces [35, 42, 3], hence lacking the versatility to generalize to other classes of objects (e.g buildings). Deep learning methods [13, 21, 9] for relighting are versatile; however, they show poor performance in extreme cases of shadow removal/addition as they often ignore the physics of the problem.

Our central contribution is to generate relighted images with new illumination settings via two different strategies and subsequently fuse them using a weight map (w)

- In the first strategy, the model predicts the material reflectance parameters (**albedo**) and illumination/geometry parameters of the scene (**shading**) for the relit image and constructs the relit image based on the simplified rendering rule Eq. 1. (we refer to this strategy as **intrinsic image decomposition** (**IID**).)
- The second strategy follows a black box approach, where the model optimizes its weights based on the ground-truth images and the loss terms in the training stage and generates the relit output directly (we refer to this strategy as **direct**).

Our proposed method exploits insights from two different sides of the literature. Moreover, since both approaches have a shared encoder, owing to the virtue of joint optimization, they can benefit the shared features the other one induces the encoder to extract as well. In this work, we are addressing the problem of relighting under two categories: 1) **One-to-one**: The objective is to change the color temperature and angle of the light source (referred to as illumination parameters) from one specific setting to another one. 2) **Any-to-any**: The illumination parameters should change from an arbitrary setting according to the illumination of a given guide image. While our proposed method applies to both *one-to-one* and *any-to-any* relighting problems, for each case, we **propose specific innovations** that enrich the model performance:

- For one-to-one relighting we incorporate normal vectors of the surfaces in the scene to adjust gloss and shadows accordingly in the image. This in particular helps boost the performance of the neural network model in the cases where the complicated geometry of the scene requires removing dense shadows or adding shadows to highly glossed regions. We refer to our network for one-to-one relighting as **One-to-one Intrinsic Decomposition-Direct RelightNet** (**OIDDR-Net**).
- For any-to-any relighting, we propose an additional multiscale block to the architecture to enhance feature extraction. This block benefits from analyzing the input RGB image (and depth map) in three different dimension levels. Using dense residual blocks and residual global blocks in each level, it provides multiscale features for the subsequent layers. We refer to our network for any-to-any relighting as **Any-to-any Multiscale Intrinsic-Direct RelightNet** (**AMIDR-Net**).

Our experimental results on VIDIT 2020/2021 dataset prove that our proposed method can outperform state-of-the-art in terms of fidelity metrics and perceptual loss. **Our OIDDR-Net ranked second and AMIDR-Net ranked among top five teams in NTIRE 2021 depth guided image relighting challenge [12].**

2. Related Works

The existing works in the area of image relighting can generally be divided into two groups of deep learning-based methods and conventional image processing methods. In the line of conventional methods and starting with models proposed for inverse rendering [3, 25] and shape estimation [46], there have been relighting works based on decomposing the image into its reflectance, illumination, and geometry components. Duchêne *et al.* [10] utilize a set of outdoor multiview scenes along with the sunlight direction to achieve albedo and shading decomposition for relighting. Wen *et al.* [52] develop a technique in which the estimated radiance environment maps, along with spherical harmonics, are used for face relighting. Other algorithms [29, 37, 40] treat relighting as approximating the light transport function of the scene to generate the new illumination settings using the input lighting parameters. While these methods construct physically realistic models for relighting, they rely on explicit illumination parameters of the scene or multiview datasets. This is considered a bottleneck for them.

While conventional methods mostly focus on physical aspects of the problem, deep learning-based algorithms rely on the capability of neural networks, along with typically large training data set, in developing a mapping function between two image domains. Methods proposed in [13, 21] view the relighting as an image-to-image translation problem and make use of Generative Adversarial Networks (GANs) for image relighting. Xu *et al.* [47] use five images to manipulate the illumination under predefined light direction. Inspired by inverse rendering, numerous works incorporate neural networks to estimate image illumination and geometry parameters. Yu *et al.* [49, 34] view relighting as a fruit of regressing the albedo, shading, and light coefficients of the input RGB image using fully convolutional networks. Face relighting methods [42, 35, 53] combine the capabilities of neural networks and the physics of relighting, which is customized for face images. Although introducing neural networks has shown promising results for the relighting problem, the appropriate dataset yet seems to be a challenging aspect. In the past years, IIW [2] and MIP SINTEL [4] have been introduced for intrinsic decomposition and optical flow analysis, respectively. More recently, Helou *et al.* proposed a virtual image dataset for illumination transfer [11], which formulates relighting into one-to-one and any-to-any problems. Along the line of scene re-

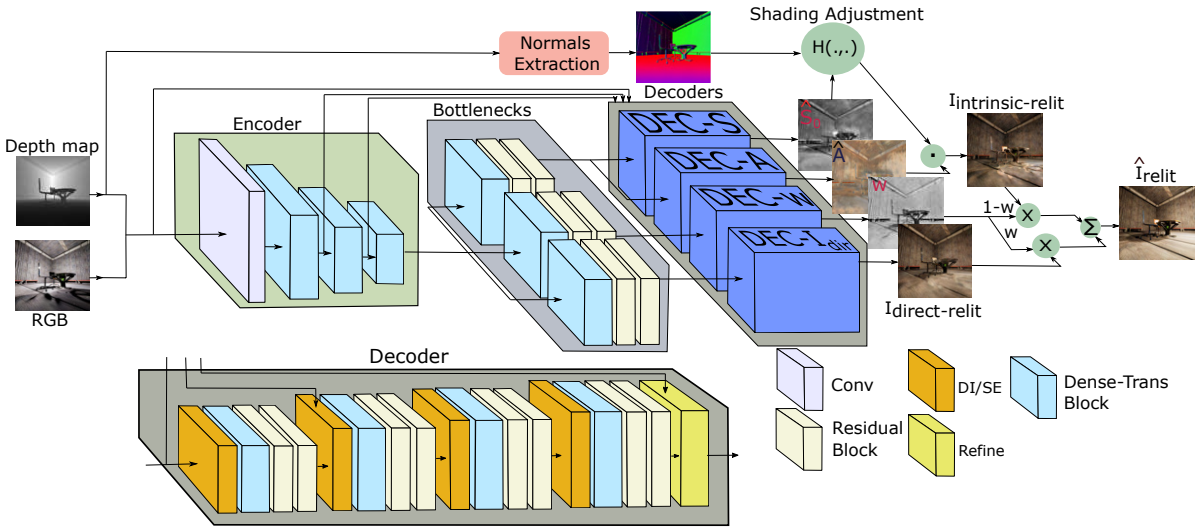

Figure 1: Our proposed OIDDR-Net architecture.

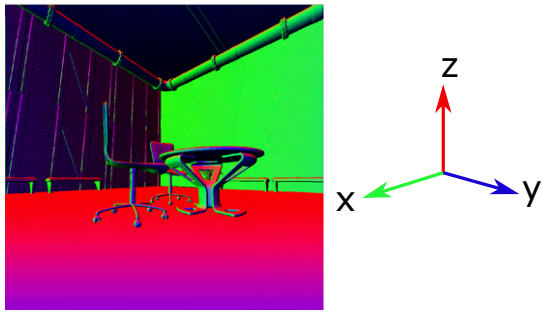

Figure 2: The visualization of the normal vectors in an example image. Each of RGB channels corresponds to x-y, y-z, and x-z planes, respectively.

lighting and illumination estimation challenge in AIM 2020 [16], Puthessery *et al.* [38] propose a U-net [41] model for one-to-one relighting where Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT) are attached to downsampling and upsampling layers, respectively. SA-AE [18] also develops a U-Net-based architecture for any-to-any relighting, where two auxiliary networks are incorporated for estimating the lighting of the guide image and providing lighting features to the decoder.

3. Proposed Method

3.1. Fusion Strategy

As mentioned earlier, our model is based on two approaches to the relighting problem:

1) Intrinsic Image Decomposition (**IID**): From the physical standpoint, every light image can be decomposed into two main parameters [1, 25]: **albedo** and **shading**. Albedo, which is the light independent parameter of the image, preserves the reflectance property of the material in the scene, while shading holds properties corresponding to illumination and the geometry of the image. Based on this, the relit image can be expressed as: $I_{\text{intinsic-relit}} = \hat{A} \odot \hat{S}$. Here, $\hat{A}$ and $\hat{S}$ denote estimated albedo and shading, respectively. $\odot$ is the element wise product operation. This method has been shown as an effective way to relight the image scene using the input RGB image (and depth map) [27, 5, 43, 49, 34]. Following this approach, the model is guided toward a systematic way of learning to relight by which it can distinguish between features associated with material reflectance property and features for the illumination and geometry of the scene.

2) Direct Relighting (**DR**): In addition to the intrinsic decomposition of the images, we also follow the end-to-end learning method as in state of the art [19, 7, 38] for learning a mapping function between the two lighting settings: $f(I) = I_{\text{direct-relit}}$. Where f denotes the mapping function learned by neural network model. This way the model, in addition to the physically inspired insight, constructs an auxiliary insight that complements the other one in terms of the extracted discriminative features.

Next, we generate a spatially varying weight map (w) to fuse the estimates:

$$\hat{I}_{relit} = w I_{direct\text{-}relit} + (1 - w) I_{intrinsic\text{-}relit} \quad (2)$$

This fusion strategy helps the model benefit from both aspects of the problem simultaneously. Furthermore, owing to the usage of a couple of shared structures and the virtue of joint optimization, each approach aids the other one through the insight it lends to the model.

3.2. Network Architecture

OIDDR-Net (Fig. 1) and AMIDR-Net (Fig. 3) are similar in terms of the general architecture which is inspired by U-Net [41, 31]. An encoder is shared by three bottlenecks followed by four decoders. Two decoders designed for the IID strategy share a bottleneck, while the other two

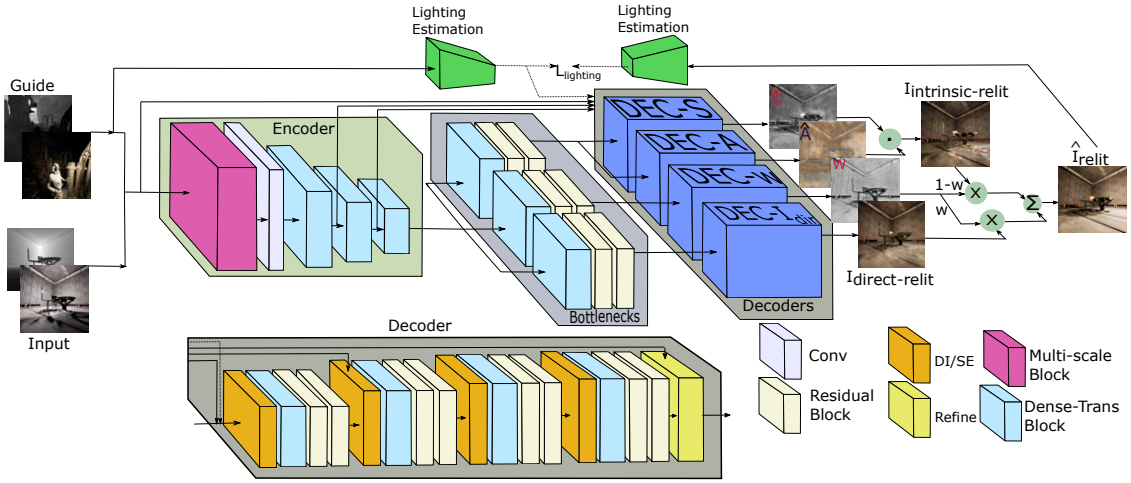

Figure 3: Our proposed AMIDR-Net architecture.

Table 1: Encoder Structure for AMIDR-Net. (*OIDDR-Net doesn't have the multiscale block and guide inputs.)

	Multi-Scale block	Base	Dense-Trans.1	Dense-Trans.2	Dense-Trans.3
Input	[Input image, Guide image*, Input depth map, Guide depth map*]	Multi-scale Output	Base	Dense-Trans.1	Dense-Trans.2
structure	See Fig. 4	$\begin{bmatrix} 7\times 7 \text{ conv.} \\ 3\times 3 \text{ max-pool} \end{bmatrix}$	$\begin{bmatrix} 1\times 1 \text{ conv.} \\ 3\times 3 \text{ conv.} \end{bmatrix} \times 6$ $\begin{bmatrix} 1\times 1 \text{ conv.} \\ 2\times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 1\times 1 \text{ conv.} \\ 3\times 3 \text{ conv.} \end{bmatrix} \times 12$ $\begin{bmatrix} 1\times 1 \text{ conv.} \\ 2\times 2 \text{ avg-pool} \end{bmatrix}$	$\begin{bmatrix} 1\times 1 \text{ conv.} \\ 3\times 3 \text{ conv.} \end{bmatrix} \times 24$ $\begin{bmatrix} 1\times 1 \text{ conv.} \\ 2\times 2 \text{ avg-pool} \end{bmatrix}$
Output	$384\times 384\times 8$	$96\times 96\times 64$	$48\times 48\times 128$	$24\times 24\times 256$	$24\times 24\times 512$

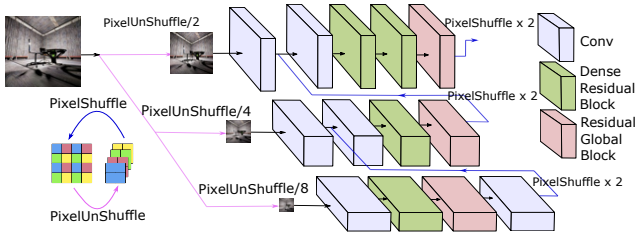

Figure 4: The multiscale block.

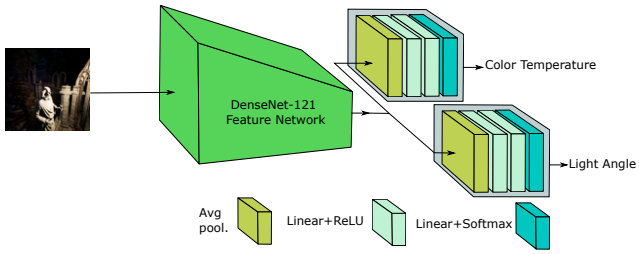

Figure 5: The lighting estimation network used for illumination regularization and extraction of illumination features from the guide image.

decoders designated for direct strategy and generating the weight map (w) are fed by a bottleneck, each. The details of these components are as follows:

1) Encoder: The encoder construction (Table 1) follows DenseNet-121 [20] feature extraction part, which is originally proposed for classification tasks. It consists of a feature extraction part followed by classification layers. We borrow the input convolutional layer, the first three dense layers, and their following transitional blocks in the feature extraction part. The main advantage of using these pretrained layers for our model is that since they're trained over ImageNet dataset [8], they provide our model an initial representation capability. This in turn helps for the faster convergence in the training. It is worth mentioning that the first convolutional layer in DenseNet accepts three channels RGB images and inputs to OIDDR-Net and AMIDR-Net are of 4 and 8 channels, respectively. To address this, for OIDDR-Net, we modify the convolutional block by keeping the first three channels and initializing the fourth one as a grayscale transformation of the other three. For AMIDR-Net, we substitute it by an eight channel convolutional block initialized randomly.

2) Bottlenecks: There are three bottlenecks consisting of a dense transitional block and two residual blocks (see Table 1,2 for details of these blocks). The main function of bottlenecks is to connect the encoder and decoders by compiling the extracted features of the encoder based on the characteristics of the decoders. So, Decoder-A and Decoder-S share a bottleneck as they both contribute on $I_{\text{intrinsic-relit}}$.

3) Decoders: Our network features four decoders for predicting the components: albedo ($\hat{A}$), shading ($\hat{S}$), weight map (w), and directly relit image ($I_{\text{direct-relit}}$). Table 2 details the structure of decoders. Each decoder includes four levels of cascaded attention module (Squeeze and excitation [17] or dilation inception modules [30]), a dense transitional block, and two residual blocks. It means that there

Table 2: Decoder Structure. (for OIDDR-Net the input to the decoder doesn't include lighting estimation outputs.) C depends on the functionality of the decoder.

	Dense-Trans.5	Res.5	Dense-Trans.6	Res.6	Dense-Trans.7
Input	[bottleneck output, Dense.Trans.2, Lighting Estimation]	Dense-Trans.5	[Trans.1, Res.5]	Dense-Trans.6	Res.6
structure	[SE/Dilation (R=16), batch norm, 3×3 conv.] $\times 7$, [1×1 conv., upsample 2]	[3×3 conv., 3×3 conv.] $\times 2$	[SE/Dilation (R=16), batch norm] $\times 7$, [1×1 conv., upsample 2]	[3×3 conv., 3×3 conv.] $\times 2$	[batch norm, 3×3 conv.] $\times 7$, [1×1 conv., upsample 2]
output	$48\times48\times128$	$48\times48\times128$	$96\times96\times64$	$96\times96\times64$	$192\times192\times32$
	Res.7	Dense-Trans.8	Res.8	Refine.9	Refine.10
Input	Dense-Trans.7	Res.7	Dense-Trans.8	[Input, Res.8]	Refine.9
structure	[3×3 conv., 3×3 conv.] $\times 2$	[batch norm, 3×3 conv.] $\times 7$, [1×1 conv., upsample 2]	[3×3 conv., 3×3 conv.] $\times 2$	[SE/Dilation (R=3), batch norm, 3×3 conv.]	[32×32 avg-pool, 1×1 conv., upsample]
output	$192\times192\times32$	$384\times384\times16$	$384\times384\times16$	$384\times384\times20$	$384\times384\times1$
	Refine.11	Refine.12	Refine.13	Output.14	
Input	Refine.9	Refine.9	Refine.9	[Refine9.10.11.12.13]	
structure	[16×16 avg-pool, 1×1 conv., upsample]	[8×8 avg-pool, 1×1 conv., upsample]	[4×4 avg-pool, 1×1 conv., upsample]	3×3 conv.	
Output	$384\times384\times1$	$384\times384\times1$	$384\times384\times1$	$384\times384\times C$	

is an analogy between decoders' structure and the encoder except that the channel attention modules are incorporated midway. This in particular helps each decoder give weight to feature maps based on its functionality, while benefiting the shared encoder. Meanwhile, skip connections from the encoder assist the decoders in reconstructing the scene as in U-net.

4) Lighting Estimation Network: As our main objective in this work is to change the illumination of the images, we need custom blocks for extracting illumination features. Moreover, it is worth mentioning that neural networks usually need either ground-truth or custom loss terms in order to extract our expected features. To this end, we train a lighting estimation network to predict the light angle and color temperature of the images using the training set for the any-to-any problem. We make use of the pretrained feature extraction part of this network (Fig. 5) to compute a perceptual loss for comparing the illumination features of the relit output. Furthermore, in AMIDR-Net (Fig. 3), this network is incorporated for feeding the decoders with the illumination features of the guide image.

3.3. Exploiting Normals for One-to-One Relighting

While the model is guided toward learning a physics-based solution for relighting, it may not necessarily be successful in changing the lighting parameters of a given scene. This could be due to the complicated geometry of the scene, which causes the presence of dense shadows needed to be removed or the presence of highly glossed objects needed to be shadowed. This is a challenging aspect of relighting for a neural network model, as neural networks usually fail in regressing outputs that lie on one side of the extreme since their share in training data is typically small. Therefore, the model in order to keep its generalization over the whole data distribution would typically show artifacts on these extreme cases. To address this issue specifically, for the case of one-to-one relighting, we propose to incorporate the information associated with the normal vectors of the surfaces present in the scene. It is shown that the shading of an image can be derived as a nonlinear function of 9-dimensional spherical harmonics coefficients and the normal vectors [35, 43, 1, 39]. The normal vectors of a scene indicate the orientation of pixels associated with each surface in the image. Fig. 2 shows an example in which the colors red, green, and blue indicate surfaces parallel to x-y, y-z, and x-z plane, respectively. Since our model learns to predict the shading directly from the information provided during the training stage in the form of ground-truth, instead of carrying out the non-linear calculations, we incorporate normal vectors into the problem as weight adjustments. More precisely, in the case of one-to-one relighting in which we know the target lighting direction, the normal vectors are used as adjustment weights for the surfaces facing toward or against the light direction: $\hat{S} = H(n_{\textit{light-dir}}, \hat{S}_0)$. Where H, $n_{\textit{light-dir}}$, and $\hat{S}_0$ are the linear adjustment function, the normal vector component corresponding to the light direction of the target, and the shading output by the model, respectively.

3.4. Any-to-Any Relighting and Multiscale Features

In any-to-any relighting there is no meaningful pixel-wise correspondence between the guide and input RGB image/depth map. Therefore unlike one-to-one relighting, training the network on image patches is not feasible. On the other hand, training the network on whole images limits the representation power of the model as the model may not necessarily extract features from lower fields of view. To prevent that, we equip AMIDR-Net with a multiscale feature extraction block [51]. Fig. 4 shows the details of this block. Using PixelUnShuffle operations, it downsamples the input to three different levels. In each level dense residual and global residual blocks extract the features. Subsequently, the extracted feature maps are **customly** upsampled and fed to the next level using PixelShuffle operation.

Finally, in the highest level the multiscale feature maps are processed and fed to the main pipeline. The main advantage of using this multiscale block over the ones, which make use of traditional upsampling modules, is how it guides the network to learn the upsampling while optimizing the feature maps. Simply put, the model learns how to extract patches from the input while keeping the correspondence between the feature maps from the guide and input.

3.5. Customized Loss Function

To train our model so that every part of it functions based on our expectation, we need to define custom loss terms for each part. Our overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{total} + \lambda_1 \mathcal{L}_{IID} + \lambda_2 \mathcal{L}_{direct} + \lambda_3 \mathcal{L}_{SSIM} + \lambda_4 \mathcal{L}_{lighting} \quad (3)$$

$$\mathcal{L}_{total} = ||\hat{I}_{relit} - Y_{relit}||_2^2 \quad (4)$$

$$\mathcal{L}_{IID} = ||\hat{A} \odot \hat{S} - Y_{relit}||_2^2 + ||\hat{A} - A||_2^2 + ||\hat{S} - S||_2^2 \quad (5)$$

$$\mathcal{L}_{direct} = ||I_{direct\text{-}relit} - Y_{relit}||_2^2 \quad (6)$$

$$\mathcal{L}_{SSIM} = 1 - SSIM(\hat{I}_{relit}, Y_{relit}) \quad (7)$$

$$\mathcal{L}_{lighting} = ||g(\hat{I}_{relit}) - g(Y_{relit})||_2^2 - \sum_{i=1}^{8} Y^i_{dir\text{-}guide} log(\hat{Y}^i_{dir}) - \sum_{j=1}^{5} Y^j_{color\text{-}guide} log(\hat{Y}^j_{color}) \quad (8)$$

Where $\mathcal{L}_{total}$, $\mathcal{L}_{IID}$, and $\mathcal{L}_{direct}$ are terms to ensure the decoder outputs $I_{intrinsic\text{-}relit}$, $I_{direct\text{-}relit}$, and $\hat{I}_{relit}$ match the ground-truth Y_{relit}. Of note, $\hat{I}_{relit}$ is the fused output (eq. 2). To help the model predict physically feasible estimates for albedo and shading, we use a pretrained intrinsic decomposition network [34], which is trained on SINTEL dataset [4], to generate pseudo ground-truths A and S. $\mathcal{L}_{SSIM}$ is used to maximize the structural similarity index (SSIM) of the relit output and ground-truth. We also define $\mathcal{L}_{lighting}$ to minimize the difference between the relit output and the ground-truth in terms of illumination parameters. In eq. 8, in the first term, the intermediate features generated by lighting estimation network (denoted by g) are compared for the relit output and ground-truth. The second and third terms (blue) are specifically incorporated for AMIDR-Net, where we minimize the negative log-likelihood of the light direction and color temperature in the relit output ($\hat{Y}_{dir}$ and $\hat{Y}_{color}$) based on guide image parameters ($Y_{dir\text{-}guide}$ and $Y_{color\text{-}guide}$). λ_1, λ_2, λ_3, and λ_4 are hyperparameters adjusting the contribution of each term in the overall loss term.

4. Implementation Details[1]

4.1. Dataset

We use VIDIT dataset [11] generated by the Unreal gaming engine [14] and consisting of two subsets for one-to-one and any-to-any relighting.

[1]Please find implementation details and results at: github/Relighting

One-to-One Relighting: VIDIT'21 training set for one-to-one relighting, provides 300 1024×1024 images with one particular light direction and color temperature and their corresponding ground-truth with the target particular illumination settings. Additionally, in VIDIT'21, unlike VIDIT'20, depth maps are provided for each image. The validation set includes 45 samples. We augment the training set to a set with about 37000 samples by 1) cropping 256×256 patches. 2) Resizing the whole images to 256×256. 3) Rotating the patches and resized images slightly with small angles (0-12 degrees). We don't incorporate flipping or rotation with large angles as the light direction in this problem should be fixed across the training samples.

Any-to-Any Relighting: The diversity across the training set is higher in the case of any-to-any relighting. For training VIDIT provides 12000 samples consisting of 300 scenes with a combination of 8 different light angles and 5 different color temperatures (40 for each scene) and 1 depth map for each scene. The validation set includes 90 images. As mentioned earlier, we cannot crop patches in this case; however, training on whole 1024×1024 images is not possible due to memory limits. Therefore, we resize the images to 384×384. We create the training set following two steps: 1) For each sample in the set, we randomly choose three different guide samples. The guide samples, obviously, are not from the same scene as the original sample. 2) For each of the chosen guide images, we find the version of original scene having the same illumination setting as the guide image. This leads to a training set with 36000 samples.

4.2. Training

We use Adam optimizer [24] with an initial learning rate of 10^{-4}, which decreases by a rate of 0.7 every 10 epochs. Owing to pretrained weights of DenseNet and incorporation of pseudo ground-truths, both OIDDR-Net and AMIDR-Net show fast convergence (optimally 20 epochs), but we train the models for 25 epochs (with batch sizes of 8 and 2, respectively) to ensure the complete stability of them. λ_1, λ_2, λ_3, and λ_4 are set to 0.4, 0.4, 0.8, and 0.03, respectively using cross validation [33].

4.3. Testing

One-to-One Relighting: We observe that our model shows better performance by the following ensemble method: i) 1024×1024 RGB image and depth map are fed to the model. ii) 384×384 RGB image and depth map are input to the model. The output is fed to a **bicubic** interpolation module (implemented in pytorch [36]) and scaled to the original size. The final output is the average of the two estimates.

Any-to-Any Relighting: In order to get the best performance of AMIDR-Net during the test phase, we resize the

Table 3: One-to-one-VIDIT'21 validation's ablation study.

Model	PSNR	SSIM	LPIPS	MPS
OIDDR-Net	**18.39**	**0.6980**	**0.2591**	**0.7194**
w/o Normals	17.49	0.6805	0.2647	0.7079
w/o $L_{lighting}$	17.59	0.6669	0.2741	0.6964

input to 384×384 so the model has the same observation as in training. The outputs then will be upsampled using bicubic interpolation.

5. Experimental Results

In this section we present experimental results of our proposed OIDDR-Net and AMIDR-Net. We provide ablation studies to show the effect of loss terms and novel network components. We also compare our models with state of the art. Our evaluation metrics are Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [54], Learned Perceptual Image Patch Similarity (LPIPS) [50], and Mean Perceptual Score (MPS) which is: $MPS = 0.5(SSIM + 1 - LPIPS)$.

5.1. Ablation Study

Effect of Exploiting Normal Vectors and $L_{lighting}$ on OIDDR-Net: To observe how incorporating normal vectors for adjusting the shading estimation of the network would affect its performance, we conduct an experiment on VIDIT'21 dataset through which we train OIDDR-Net with initial shading estimation. Additionally, to study the effect of comparing the illumination in network relit output to illumination in ground-truth (through $L_{lighting}$), we train OIDDR-Net without applying $L_{lighting}$ in training loss. Table 3 shows the performance results of the three models. While both factors play an important role in minimizing the fidelity and perceptual loss, we can see how $L_{lighting}$ contributes to structural similarity.

Effect of Multiscale Feature Extraction and $L_{lighting}$ on AMIDR-Net: To see the importance of the multiscale block as well as $L_{lighting}$, we train three models: 1) AMIDR-Net with full architecture and $L_{lighting}$ being activated during training, 2) AMIDR-Net with the multiscale block removed from the architecture, and 3) AMIDR-Net trained with $L_{lighting}$ dropped from the training loss. Table 4 shows the results of our experiments on validation data, whereby one can infer the effect of different components on the network performance. The noticeable drop in SSIM, after removing the multiscale block, proves its impact in helping the network extract more discriminative features for constructing the structural similarity between the output and the ground-truth.

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

All the existing works for image-based relighting (applicable to VIDIT) have been proposed for the dataset without depth information. Therefore, to have a fair comparison, we trained and evaluated our OIDDR-Net and AMIDR-Net on VIDIT'20 training and validation set.

Table 4: Any-to-any-VIDIT'21 validation's ablation study.

Model	PSNR	SSIM	LPIPS	MPS
AMIDR-Net	**19.83**	**0.6940**	**0.3381**	**0.6779**
w/o the multiscale block	19.09	0.6685	0.3421	.6632
w/o $L_{lighting}$	19.22	0.6721	0.3403	0.6659

Table 5: Comparison with state of the arts for one-to-one relighting on VIDIT'20 validation set.

Model	PSNR	SSIM	LPIPS	MPS	Runtime(s)
OIDDR-Net (ours)	**17.62**	**0.6645**	**0.2733**	**0.6956**	0.53
WDRN [38]	17.45	0.6642	0.2771	0.6935	0.05
DRN [45]	17.59	0.596	0.440	0.578	0.5
DMSHN [6]	17.20	0.5696	0.3712	0.5992	0.0058
SRN [44]	16.94	0.5660	0.4319	0.5670	0.87
Dense-GridNet [28]	16.67	0.2811	0.3691	0.9120	0.9326
Dong *et al.* [9]	17.14	0.6132	0.2764	0.6684	—

One-to-one Relighting: We modify our OIDDR-Net for accepting only the RGB image (so normals are not exploited) and train it over the training set. We compare our modified OIDDR-Net with existing methods in Table 5. While [44] and [28] are proposed for deblurring and dehazing, all other methods have been proposed for the same exact problem and dataset. Table 5 shows that our OIDDR-Net outperforms state of the art w.r.t. all metrics as a result of fusing the power of neural networks and the physics of the problem. We can also qualitatively confirm this in Fig. 6, where OIDDR-Net's output successfully mimics the illumination settings of the ground-truth without artifacts.

Any-to-any Relighting: We modify our AMIDR-Net by changing the number of input channels and removing the skip connections corresponding to the depth maps and train it on VIDIT 2020 dataset. We compare our modified AMIDR-Net with state of the art in Table 6, where SA-AE [19] is the winner of AIM 2020 any-to-any relighting track and [9] is an encoder-decoder network proposed by another participant of the same challenge. We also compare our method with an adapted version of [53], which is originally proposed for portrait relighting. According to Table 6, AMIDR-Net outperforms others w.r.t. all evaluation metrics. Fig. 7 visualizes three outputs from different methods where we see how our AMIDR-Net changes the illumination of the input according to guide image without artifacts.

Comparison with NTIRE 2021 Relighting Methods: Additionally, we compare our models with two methods from the top 5 methods of the NTIRE 2021 relighting challenge. As Table 7 confirms, OIDDR-Net and AMIDR-Net are among the top-performing methods. OIDDR-Net ranked second in one-to-one relighting in terms of MPS and AMIDR-Net ranked second in terms of PSNR.

6. Conclusion

We develop a physically inspired dense fusion network for image relighting. Our method benefits from the capabil-

Figure 6: Qualitative comparison between different methods on one-to-one relighting. From left to right: SRN [44], DeepGridNet [28], DRN [45], DMSHN [6], WDRN [38], OIDDR-Net (ours) and ground-truth.

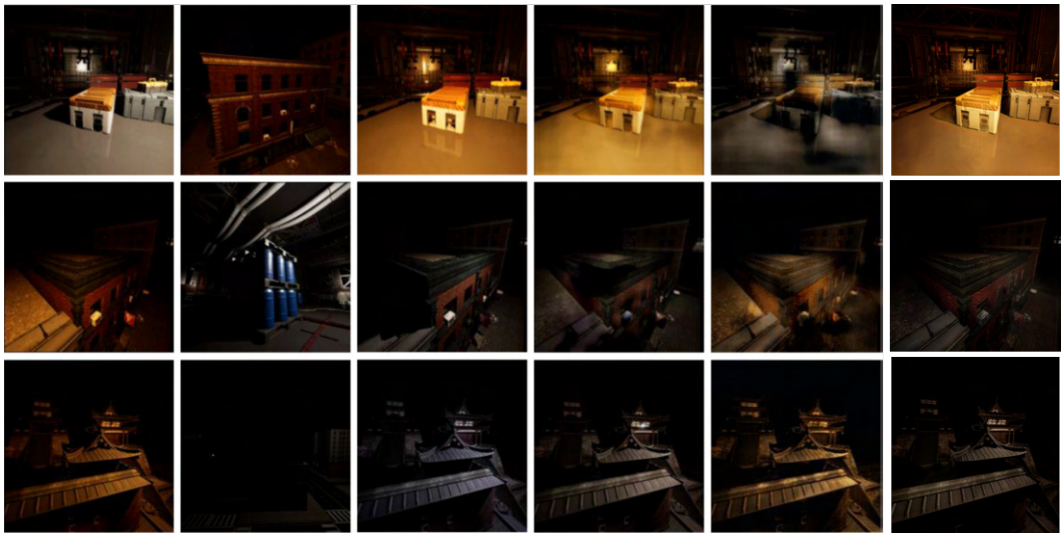

Figure 7: Qualitative comparison between different methods. From left to right: input image, guide image, ground-truth, SA-AE [19], DPR [53], and AMIDR-Net (ours).

Table 6: Comparison with state of the arts for any-to-any relighting on VIDIT'20 validation set.[2]

Model	PSNR	SSIM	Runtime(s)
AMIDR-Net (ours)	**19.16**	**0.6621**	0.51
SA-AE [19]	18.06	0.6480	0.15
DPR [53]	16.40	0.5238	0.095
Dong *et al.* [9]	18.07	0.5994	—

Table 7: Comparison with other methods in NTIRE2021 relighting challenge on VIDIT'21 test set.

Track	Model	PSNR	SSIM	LPIPS	MPS	Runtime(s)
One-to-one	OIDDR-Net (ours)	18.83	0.6874	0.1634	0.7620	**0.53**
	Method 1	19.14	0.6931	0.1605	0.7663	2.88
	Method 2	18.27	0.6772	0.1670	0.7551	2.12
Any-to-any	AMIDR-Net (ours)	**20.14**	0.6711	0.2028	0.7341	**0.51**
	Method 1	19.22	0.6784	0.1566	0.7609	2.04
	Method 2	18.60	0.6508	0.1661	0.7423	0.6740

ity of dense networks in extracting representative features, while simultaneously estimating albedo and shading – key components of the relighting physical model. The simultaneous intrinsic image decomposition and direct relighting help the model refine its feature extraction by joint optimization. This leads to physically more feasible results in terms of illumination parameters and therefore less artifacts in the obtained relighted images. Ablation studies explain the role of each component in models proposed both for one-to-one and any-to-any relighting. Comparisons with existing literature on benchmark datasets and competing methods in the NTIRE'21 relighting challenge show our proposal achieves state-of-the-art results.

[2]LPIPS and MPS are not made available by other works. The runtime for [9] is not reported.

References

[1] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 3, 5

[2] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. 33(4), July 2014. 2

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 2, 6

[5] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *2013 IEEE International Conference on Computer Vision*, pages 241–248, 2013. 3

[6] Sourya Das, Nisarg Shah, Saikat Dutta, and Himanshu Kumar. Dsrn: an efficient deep network for image relighting, 02 2021. 7, 8

[7] Sourya Dipta Das, Nisarg A. Shah, Saikat Dutta, and Himanshu Kumar. Dsrn: an efficient deep network for image relighting, 2021. 3

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[9] Liping Dong, Yu Zhu, Zhuolong Jiang, Xiangyu He, Zhaohui Meng, Chenghua Li, Cong Leng, and Jian Cheng. An ensemble neural network for scene relighting with light classification. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 581–595, Cham, 2020. Springer International Publishing. 2, 7, 8

[10] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.*, 34(5), Nov. 2015. 2

[11] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020. 2, 6

[12] 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. 2

[13] Paul Gafton and Erick Maraz. 2d image relighting with image-to-image translation, 2020. 2

[14] Epic Games. Unreal engine | the most powerful real-time 3d creation platform,https://www.unrealengine.com/en-us/. 6

[15] W. He, Y. Liu, J. Feng, W. Zhang, G. Gu, and Q. Chen. Low-light image enhancement combined with attention map and u-net network. In *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 397–401, 2020. 1

[16] Majed Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, Mahmoud Afifi, Michael Brown, Kele Xu, Hengxing Cai, Yuzhong Liu, Li-Wen Wang, Zhi-Song Liu, Chu-Tak Li, Sourya Das, Nisarg Shah, Akashdeep Jassal, Tongtong Zhao, Shanshan Zhao, Sabari Nathan, Dr.M.Parisa Beham, and Jian Cheng. Aim 2020: Scene relighting and illumination estimation challenge, 09 2020. 3

[17] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 4

[18] Zhongyun Hu, Xin Huang, Yaning Li, and Qing Wang. Saae for any-to-any relighting. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 535–549, Cham, 2020. Springer International Publishing. 3

[19] Zhongyun Hu, Xin Huang, Yaning Li, and Qing Wang. Saae for any-to-any relighting. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 535–549, Cham, 2020. Springer International Publishing. 3, 7, 8

[20] 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)*, pages 2261–2269, 2017. 4

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018. 2

[22] James T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150, Aug. 1986. 1

[23] Z. Khan, M. Khanna, and S. Raman. Fhdr: Hdr image reconstruction from a single ldr image using feedback network. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, 2019. 1

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[25] Edwin Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 02 1971. 2, 3

[26] H. Le and D. Samaras. Shadow removal via shadow image decomposition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8577–8586, 2019. 1

[27] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+depth video. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 327–340, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 3

[28] X. Liu, Y. Ma, Z. Shi, and J. Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7313–7322, 2019. 7, 8

[29] Wojciech Matusik, Matthew Loper, and Hanspeter Pfister. Progressively-refined reflectance functions from natural illumination. EGSR'04, page 299–308, Goslar, DEU, 2004. Eurographics Association. 2

[30] K. Metwaly, X. Li, T. Guo, and V. Monga. Nonlocal channel attention for nonhomogeneous image dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1842–1851, 2020. 1, 4

[31] K. Metwaly, X. Li, T. Guo, and V. Monga. Nonlocal channel attention for nonhomogeneous image dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1842–1851, 2020. 3

[32] Kareem M. Metwaly and V. Monga. Attention-mask dense merger (attendense) deep hdr for ghost removal. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2623–2627, 2020. 1

[33] Vishal Monga. *Handbook of Convex Optimization Methods in Imaging Science*. Springer International Publishing, Cham, 2018. 6

[34] Takuya Narihira, Michael Maire, and Stella Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. pages 2992–2992, 12 2015. 2, 3, 6

[35] T. Nestmeyer, J. F. Lalonde, I. Matthews, and A. Lehrmann. Learning physics-guided face relighting under directional light. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5123–5132, 2020. 1, 2, 5

[36] Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. 2019. 6

[37] Pieter Peers, Dhruv K. Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. Compressive light transport sensing. 28(1), Feb. 2009. 2

[38] 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, 7, 8

[39] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 18:2448–59, 11 2001. 5

[40] Dikpal Reddy, Ravi Ramamoorthi, and Brian Curless. Frequency-space decomposition and acquisition of light transport under spatially varying illumination. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, page 596–610, Berlin, Heidelberg, 2012. Springer-Verlag. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015. 3

[42] S. Sengupta, D. Lichy, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[43] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5444–5453, 2017. 3, 5

[44] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 7, 8

[45] Li-Wen Wang, Wan-Chi Siu, Zhi-Song Liu, Chu-Tak Li, and Daniel Lun. Deep relighting networks for image light source manipulation, 08 2020. 7, 8

[46] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven Gortler, David Jacobs, and Todd Zickler. From shading to local shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 10 2013. 2

[47] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Trans. Graph.*, 37(4), July 2018. 2

[48] M. Yu, V. Cherukuri, T. Guo, and V. Monga. Ensemble dehazing networks for non-homogeneous haze. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1832–1841, 2020. 1

[49] Y. Yu and W. A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3150–3159, 2019. 2, 3

[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[51] 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)*, pages 1695–1704, 2020. 5

[52] Zhen Wen, Zicheng Liu, and T. S. Huang. Face relighting with radiance environment maps. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–158, 2003. 2

[53] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs. Deep single-image portrait relighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7193–7201, 2019. 2, 7, 8

[54] Zhou 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*, 13(4):600–612, 2004. 7