SA-AE for Any-to-Any Relighting

Zhongyun Hu, Xin Huang, Yaning Li, and Qing Wang(✉)

School of Computer Science, Northwestern Polytechnical University,
Xi'an 710072, China
{zy_h,hx0817,liyn}@mail.nwpu.edu.cn,qwang@nwpu.edu.cn

Abstract. In this paper, we present a novel automatic model Self-Attention AutoEncoder (SA-AE) for generating a relit image from a source image to match the illumination setting of a guide image, which is called any-to-any relighting. In order to reduce the difficulty of learning, we adopt an implicit scene representation learned by the encoder to render the relit image using the decoder. Based on the learned scene representation, a lighting estimation network is designed as a classification task to predict the illumination settings from the guide images. Also, a lighting-to-feature network is well designed to recover the corresponding implicit scene representation from the illumination settings, which is the inverse process of the lighting estimation network. In addition, a self-attention mechanism is introduced in the autoencoder to focus on the re-rendering of the relighting-related regions in the source images. Extensive experiments on the VIDIT dataset show that the proposed approach achieved the 1st place in terms of MPS and the 1st place in terms of SSIM in the AIM 2020 Any-to-any Relighting Challenge.

Keywords: Any-to-any relighting · Lighting estimation · Deep learning · Autoencoder · Self-attention mechanism

1 Introduction

The goal of this paper is to re-render a source image with a certain illumination setting to match the illumination setting of another guide image. As shown in Fig. 1, the input is a source image of a complex scene and a guide image under a novel lighting, and the output is a relit image of the complex scene under the novel lighting (Fig. 1c). Figure 1d is a relit image generated by the proposed approach. This task is important for a range of applications in augmented reality, visual effects, and production visualization. For example, any-to-any relighting can be used to enhance the underexposed images using an adequate and suitable lighting.

In the past few years, physically-based relighting methods [11,15,17,21] are proposed to explicitly estimate the geometry, reflectance, and lighting of the scene and then re-render this scene using the novel illumination setting. However, this is an ill-posed problem: these scene factors interact in complex ways to form images and multiple combinations of these factors may produce the same image

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12537, pp. 535–549, 2020.
https://doi.org/10.1007/978-3-030-67070-2_32

Fig. 1. An example of any-to-any relighting. In the upper left corner of the guide image, the red arrow and number indicate the direction and color temperature of the light source, respectively. (Color figure online)

[13]. Thus, such approaches have often focused on restricted settings—objects from a specific class (i.e. faces and human bodies). But they are still limited to what is expressible by their estimated physical model, such as a micro-facet SVBRDF model and spherical Gaussian lighting.

In contrast, some other recent approaches [22,24,29] do not have any explicit inverse rendering step for estimating scene properties. Instead, they trained a single neural network to directly render relit images from an implicit scene representation in the latent space. For example, Zhou et al. [29] proposed an hourglass network to capture and consolidate information across all scales of the image for the portrait relighting task. But for any-to-any relighting, the key is how to recast and remove the shadow with a target lighting. Except for the lighting color temperature, most regions of the source image do not need shadow recasting or removal.

In this paper, we propose a self-attention module based autoencoder for any-to-any relighting. Armed with the self-attention module, the relighting-related regions will be carefully distinguished by the autoencoder. Inspired by [22,24,29], in order to reduce the difficulty of learning, we adopt an implicit scene representation learned by the encoder to render the relit images using the decoder. Considering the fact that regressing the exact values is more difficult than the classification, a lighting estimation network is designed as a classification task to predict the lighting settings from the implicit scene representation. Also, a lighting-to-feature network is well designed to recover the corresponding implicit scene representation. In addition, a resize-conv is utilized to replace the transposed-conv to avoid the checkerboard artifacts.

The main contributions of this paper are summarized as follows:

1) We propose a novel automatic model SA-AE for generating a relit image from a source image and a guide image, which is called any-to-any relighting [7]. In addition, a self-attention mechanism is introduced in the autoencoder to focus on the re-rendering of the relighting-related regions in the source images.
2) We tested our proposed method on the VIDIT dataset [7]. Extensive experiments show that the proposed method can achieve the highest MPS (based on the SSIM, LPIPS scores) in the AIM 2020 Any-to-any Relighting Challenge [8].

2 Related Work

Any-to-any relighting can be seen as a special case of image-based relighting, and also relates to inverse rendering.

2.1 Image-Based Relighting

Debevec et al. [6] proposed to relight the scene by densely sampling the light transport function using thousands of images. Furthermore, the coherence of the light transport function [16,19,20] is utilized to relight the scene using fewer samples. However, these approaches still require hundreds of images, and this acquisition process is very time-consuming. In addition, special acquisition systems need to be designed to simulate the desired illumination. Driven by the success of deep learning, Xu et al. [24] used a non-linear CNN-based representation that exploits correlations in light transport across scenes to relight the scene with only five images. But Sun et al. [22] and Zhou et al. [29] argued that the utility is usually limited due to requirements of multiple images of the scene under controlled or known illuminations, two deep neural networks with similar structure are proposed to relight the face using a single RGB image of a portrait taken in an unconstrained environment. Different from the face with the symmetric structure or single objects of a specific class, the VIDIT dataset [7] contains complex indoor and outdoor scenes. Besides, not all regions of the image are equally important in contributing to the relighting, only the task-related regions are of concern. In this paper, an attention mechanism is introduced to make the network focus on the relighting-related regions.

2.2 Inverse Rendering

Inverse rendering is to estimate the illumination, reflectance properties, and geometry from observed appearance (i.e. one or more images). Once these scene properties are all estimated, relighting can be viewed as a natural extension of inverse rendering, which is performed by the physically based rendering (PBR) pipeline. Traditional inverse rendering [1–4] is usually to jointly optimize the scene properties to achieve the set of values that best explain the observed image. For example, Barron et al. [4] used a complex combination of generic priors to recover shape, albedo, and illumination in an optimization-based framework. In the past few years, researchers have concentrated on data-driven approaches for learning priors instead of handcrafted priors. Sengupta et al. [21] presented a residual block-based architecture SfSNet to disentangle normal and albedo into separate subspaces. Yu et al. [25] used multiview stereo supervision to train an hourglass-based neural network with skip connections to predict normal and albedo from a single image. Although such approaches have an explicit physically meaningful representation, they are sometimes limited to what is expressible by their estimated physical model. In contrast, we utilize an implicit learning-based scene representation to render the target image, which will greatly reduce the difficulty of learning.

3 Method

3.1 Problem Formulation

Given an image I and its corresponding illumination setting L, we formulate the general relighting problem as follows,

$$\phi_1 : I \rightarrow \left(Z^i, Z^l\right), \psi_1 : \left(Z^i, Z^l\right) \rightarrow \hat{I} \tag{1}$$

$$\phi_2 : Z^l \rightarrow L, \psi_2 : L \rightarrow \hat{Z}^l \tag{2}$$

$$\hat{\phi}_1, \hat{\psi}_1, \hat{\phi}_2, \hat{\psi}_2 = \underset{\phi_1, \psi_1, \phi_2, \psi_2}{\arg\min} \left\| I - \left(\psi_1 \circ \phi_1\right) I \right\| + \left\| Z^l - \left(\psi_2 \circ \phi_2\right) Z^l \right\| \tag{3}$$

where Z^i (Z^l) is the implicit representation of the intrinsic property (lighting setting) of the scene, ϕ_1 and ψ_1 are the encoder and decoder for the input image I respectively, ϕ_2 and ψ_2 are the encoder and decoder for the Z^l respectively. In particular, based on the above general relighting problem formulation, any-to-any relighting can be done as shown in the Fig. 2. A source image I_S and a guide image I_G are both fed into the encoder ϕ_1 to get their own Z^i and Z^l. The $\hat{Z}^l$, which is reconstructed by the encoder ϕ_2 and the decoder ψ_2, and the Z^i are concatenated and fed into the decoder ψ_1 to obtain the relit image I_R. Compared with common style transfer [10], the advantage of this design is that the source image can be relighted with a user-controlled illumination setting even if the guide image is missing (more details in Fig. 7 and Fig. 8).

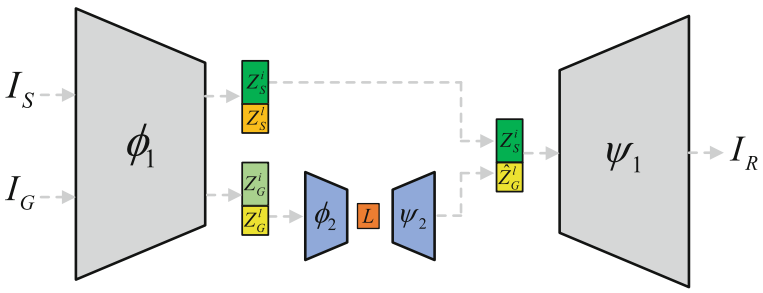

Fig. 2. Schematic diagram.

We model both the encoder ϕ and the decoder ψ as a convolutional neural network. The details of the network architecture are described in Sect. 3.2. We train the network on the VIDIT dataset.

3.2 Network Architecture

As shown in Fig. 3, our proposed SA-AE consists of four parts, which are a scene encoder ϕ_1, a scene decoder ψ_1, a lighting estimation network ϕ_2, and a lighting-to-feature network ψ_2. The scene representation Z in the latent space can be

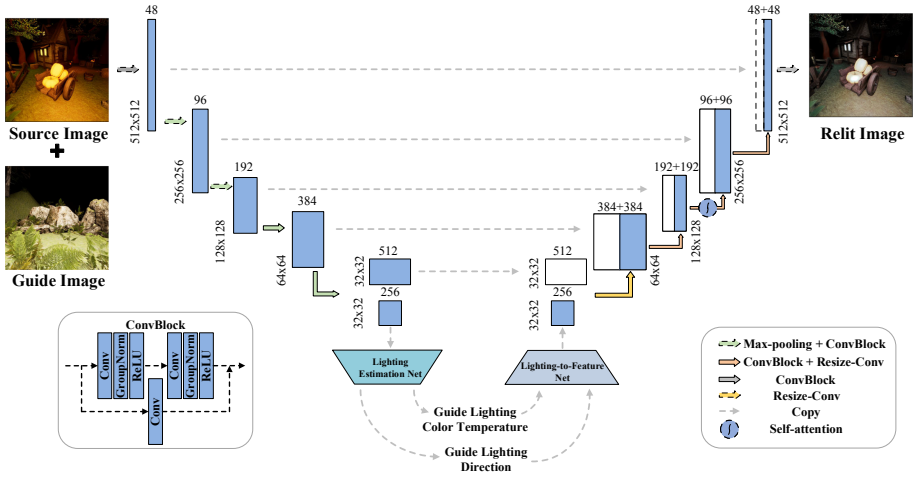

Fig. 3. An overview of the proposed SA-AE.

divided into a lighting-dependent representation Z^l and an intrinsic property-dependent representation Z^i, as discussed in Sect. 3.1. In the scene encoder ϕ_1, Z^l and Z^i are obtained through five ConvBlocks and four max-pooling layers that gradually decrease the spatial resolution and increase the number of channels by a factor 2. In the lighting-to-feature network ψ_2, the lighting directions and color temperatures, which both are represented as one-hot vectors in our work, are processed by two fully connected layers to output a 256-dimensional lighting-dependent feature vector $\hat{Z}^l$. The combination of the two disentangled implicit representation ($\hat{Z}^l$ and Z^i) is then fed into the scene decoder ψ_1 to get the relit image. The scene decoder ψ_1 consists of four ConvBlocks and four Resize-convs that gradually increase the spatial resolution and decrease the number of channels by a factor 2, which is the inverse process of the scene encoder ϕ_1.

As a result, the relit image can be rendered from the lighting-dependent representation of the guide image and the intrinsic property-dependent representation of the source image. Note that the intrinsic property-dependent representation of the source image can be directly obtained by feeding the source image into the scene encoder. But for the lighting-dependent representation of the guide image, a lighting estimation network that consists of two fully connected layers is designed to acquire the illumination setting of the guide image, and then the guide illumination setting is fed into a lighting-to-feature network to recover the corresponding lighting-dependent representation.

Self-attention Mechanism. Recent works [23,26] suggest that the self-attention mechanism helps with modeling long-range, multi-level dependencies across image regions. For any-to-any relighting, the key is how to recast and remove the shadow in the source image with a guide lighting. But recent deep

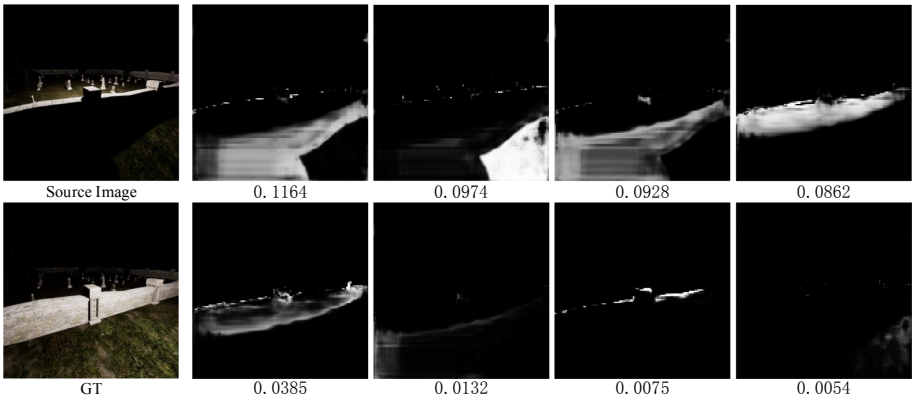

Fig. 4. Attention map visualization. The four columns on the right are the attention maps. The numbers below each attention map show the attention scores.

learning-based relighting approaches [22,29] only focused on extracting and fusing feature maps of different scales, they ignore that the importance of different regions in the source images is different. Therefore, a self-attention mechanism is introduced to focus on the re-rendering of the relighting-related regions in the source images. For feature maps $\mathbf{x}$, the calculation of the corresponding self-attention feature maps $\mathbf{y}$ can be divided into two steps. Firstly, the attention map $\beta_{j,i}$, which indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region, is defined as follows,

$$\beta_{j,i} = \frac{\exp\left(\left(\mathbf{W}_f \mathbf{x}_i\right)^T \left(\mathbf{W}_g \mathbf{x}_j\right)\right)}{\sum_{i=1}^{N} \exp\left(\left(\mathbf{W}_f \mathbf{x}_i\right)^T \left(\mathbf{W}_g \mathbf{x}_j\right)\right)} \tag{4}$$

where $\mathbf{W}_f$ and $\mathbf{W}_g$ are the learned weight matrices. Then the self-attention feature map $\mathbf{y}_j$ is calculated as follows,

$$\mathbf{y}_j = \mathbf{W}_v \left(\sum_{i=1}^{N} \beta_{j,i} \mathbf{W}_h \mathbf{x}_i\right) \tag{5}$$

where $\mathbf{W}_v$ and $\mathbf{W}_h$ are the learned weight matrices. See Sect. 4.2 for more details.

Resize-Conv. Due to the checkerboard artifacts caused by the transposed-conv overlap and random initialization, resize-conv [18] is utilized to replace the common transposed-conv for upsampling. Resize-conv first upscales the low-resolution feature maps using bilinear interpolation and then employs a standard convolutional layer with a kernel size 3×3.

3.3 Supervision for Training SA-AE

In our work, lighting estimation and relighting are seen as a classification task and a regression task respectively. Because the direction and color temperature are two different properties of the light source, we apply the cross-entropy loss function H to supervise the learning of the lighting estimation network:

$$L_c = H\left(p_{temp}, q_{temp}\right) + H\left(p_{dir}, q_{dir}\right) \tag{6}$$

Where p_{temp} and p_{dir} are the expected color temperature and lighting direction respectively, q_{temp} and q_{dir} are the actual color temperature and lighting direction respectively. For the relighting task, a MSE loss is used to supervise the proposed SA-AE, which gives a relatively high weight to large errors and helps the network to pay more attention to relighting-related regions. Inspired by [28], the loss function $SSIM$ is used to make the network learn to produce visually pleasing images. In addition, we also minimize the difference between the gradients of the relit image and Ground-truth to reduce noise effects. Thus, the loss for the relighting task is defined as:

$$L_r = \lambda_1 \left\| \hat{I} - I \right\|_2 + \lambda_2 \left\| Grad\left(\hat{I}\right) \text{-} Grad\left(I\right) \right\|_1 + \lambda_3 \left(1 - SSIM\left(\hat{I}, I\right)\right) \tag{7}$$

Where $Grad$ is a function to calculate the gradient of the image, and λ is the weight coefficient. Finally, the total loss is a linear combination of the lighting estimation loss and the relighting loss:

$$L_{total} = L_c + L_r \tag{8}$$

4 Experimental Results

4.1 Training Details

We conduct our experiments using Pytorch on 8 NVIDIA GTX1080Ti GPUs. The parameters of the network are initialized using Kaiming uniform initialization [9]. We optimize the parameters by the Adam optimizer [12] with learning rate $= 1e-4$, betas $= (0.9, 0.999)$. Consequently, the batch size is set to be 16 to maximize GPU memory utilization. Except that the weight λ_3 is set to 0.1, the other weights are all set to 1.

4.2 Model Analysis

Attention Maps Visualization. As discussed in Sect. 3.2, a self-attention module is introduced to focus on the re-rendering of the relighting-related regions in the source images, which gives a high weight to the regions that need shadow removal and recasting. We sum the attention scores of each attention map and sort them from largest to smallest, and the first eight attention maps with high attention scores are visualized in Fig. 4. In addition, the attention scores of the other attention maps are almost 0. It suggests that the relighting-related regions in the source images are well distinguished by the proposed SA-AE.

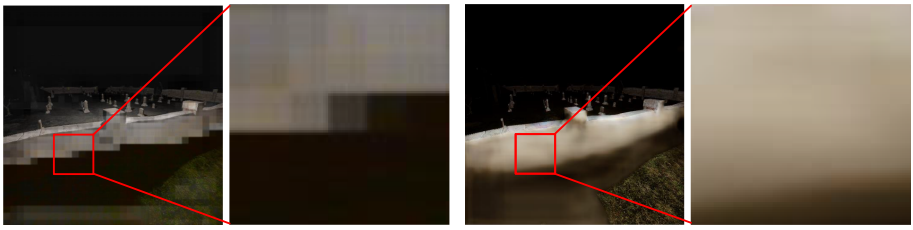

Fig. 5. Transposed-conv v.s. Resize-conv

Resize-conv v.s. Transposed-conv. In our initial experiment, transposed convolution was used for upsampling. However, it can be seen from the left image in Fig. 5 that a large number of checkerboard artifacts appear in the generated images. Thus, resize convolution is used to replace the transposed convolution for upsampling. The right image in Fig. 5 shows a pleasant visual effect.

4.3 Quantitative Evaluation

Challenge Results. A Mean Perceptual Score [7] (MPS) is defined as the average of the normalized SSIM and LPIPS [27] scores to determine the final ranking in the AIM2020 Relighting Challenge,

$$MPS = 0.5 \cdot (S + (1 - L)) \tag{9}$$

where S is the SSIM score, and L is the LPIPS score. The any-to-any relighting track of the AIM 2020 Challenge had 56 participants, with 6 finalists submitting results for the test stage. Table 1 shows that our results obtained the $1^{\mathbf{st}}$ **place** in terms of MPS based on the perceptual quality, which verifies the effectiveness of our proposed model. Our result also achieved 18.54 dB on the test set, which is 0.81 dB lower than the 3^{nd} method and 0.30 dB higher than the 2^{nd} method. Note that our model takes 0.15 s on average to process 512 512 images used in the test phase.

Table 1. AIM 2020 Image Relighting Challenge Track 3 (Any-to-any relighting) results on the test set.

Rank	Method	**MPS**	SSIM	LPIPS	PSNR	Run-time	GPU
1	Ours	**0.6484 (1)**	**0.6353 (1)**	0.3386 (3)	18.5436 (2)	0.15 s	1080TI
2	the 2^{nd} method	0.6428 (2)	0.6195 (2)	0.3338 (2)	18.2384 (4)	6 s	1080TI
3	the 3^{rd} method	0.6424 (3)	0.6042 (3)	**0.3194 (1)**	**19.3559 (1)**	0.3 s	Titan X
4	the 4^{th} method	0.6213 (4)	0.5881 (4)	0.3455 (4)	17.6314 (5)	13 s	
5	the 5^{th} method	0.5258 (5)	0.4451 (5)	0.3936 (5)	18.3493 (3)		Titan Xp
6	the 6^{th} method	0.3465 (6)	0.4123 (6)	0.7192 (6)	10.4483 (6)	0.0289 s	

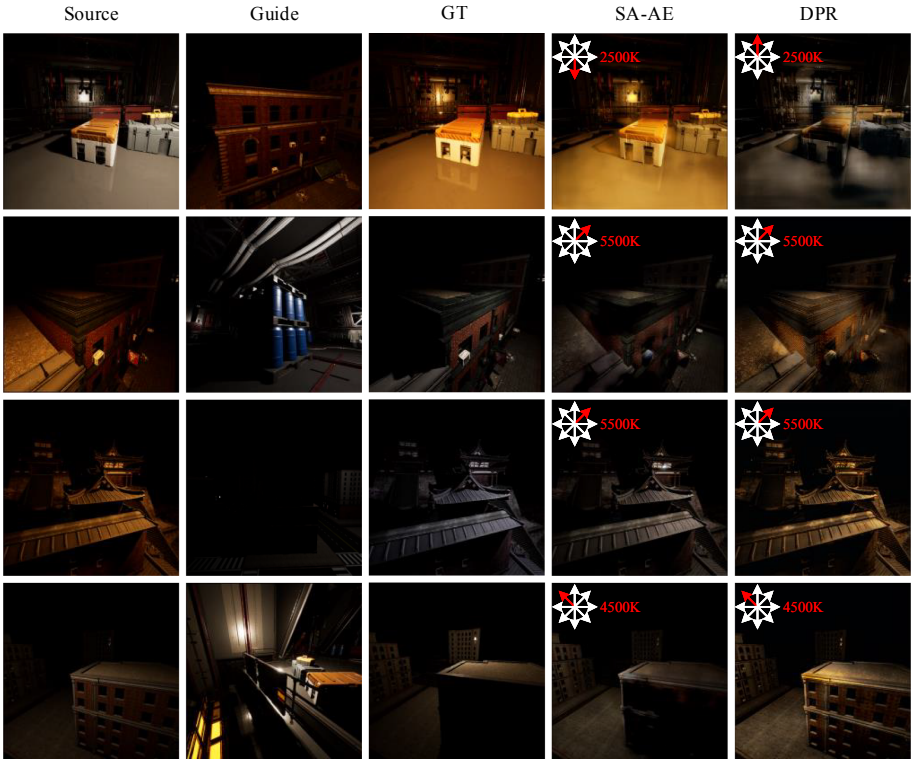

Fig. 6. Visual comparison of different approaches on the validation set.

Comparison to DPR [29]**.** We compared a deep single-image portrait relighting method DPR to our method on the validation set. Note that the VIDIT dataset only provides 8 different azimuthal angles of the light sources and the corresponding zenith angles of the light sources are unknown. Therefore, a spherical harmonics lighting, which is used to represent the illumination of the environment in DPR, can't be directly applied to the VIDIT dataset. For a fair comparison, we modified their lighting estimation network and loss functions to ours in order to train the DPR on the VIDIT dataset, while the others remain unchanged. Table 2 shows that our method outperforms the DPR in terms of SSIM and PSNR.

4.4 Qualitative Evaluation

Comparison to DPR. Figure 6 shows the qualitative results of DPR and the proposed SA-AE. In the upper left corner of the relit images, the red arrows and numbers indicate the direction and color temperature of the light source, respectively. In the first row, we can see that the predicted light direction by our method is consistent with the actual light direction of the guide image, while the

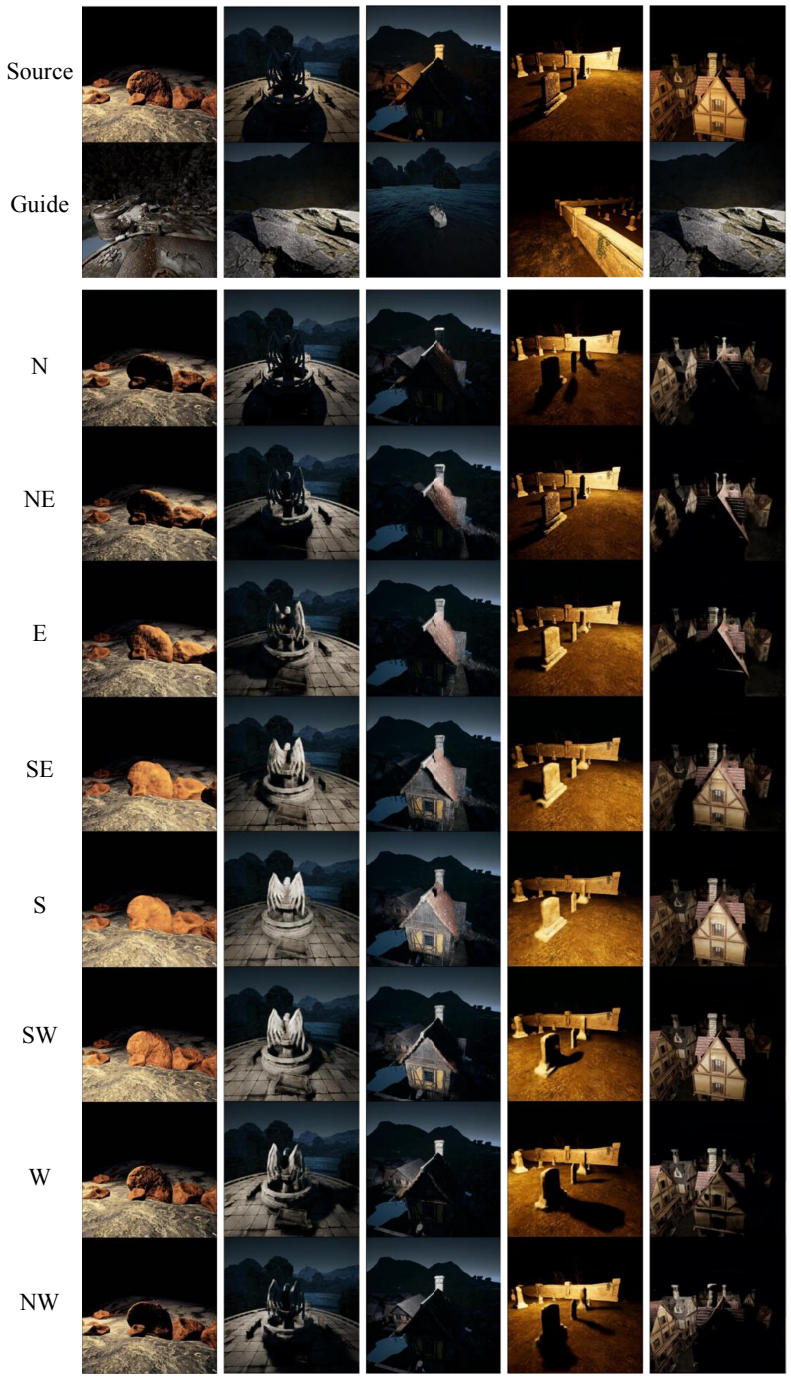

Fig. 7. More relit images with 8 different azimuth angles of the light source.

DPR predicted a completely opposite direction. This may be due to the overall darkness of the guide image, which interferes with the prediction of the DPR. In the third row, although both methods predicted the color temperature of the guide image as 5500K, the relit image generated by the DPR still looks like warm colors. Similarly, in the last row, both methods predicted the direction of the guide image as northwest, but the front of the building is still very bright in the relit image relighted by the DPR, which implies that the DPR doesn't perform recasting shadow very well. In contrast, our method correctly casts the shadow to the front of the building.

Table 2. Quantitative comparison of our method to DPR on the validation set.

Method	PSNR	SSIM
DPR [29]	16.4079	0.5238
SA-AE	**18.0695**	**0.6480**

A User-controlled Relighting. As discussed in Sect. 3.1, even if a guide image is missing, we can still perform relighting according to user demands. As illustrated in Fig. 7, the eight rows of images at the bottom are relighted using 8 different light source directions and the light color temperature of the guide image. As the direction of the light source moves, the shadows caused by the scene geometry are removed and recasted very well. Figure 8 also demonstrated that the five rows of images at the bottom are relighted using 5 different light color temperatures and the light source direction of the guide image. When the light color temperature keeps increasing, the relit images generated by our method changes from warm colors to cool colors gradually.

4.5 Limitations and Future Work

Although our method won the first place in the competition, Table 1 indicates that SSIM, LPIPS, and PSNR of different approaches on the test set are still very low compared to other similar image manipulation tasks [5,14], which means that any-to-any relighting is a very challenging task. As far as our proposed SA-AE is concerned, there are still several problems. First, Fig. 9(a) shows that the shadow is not completely recasted on the locomotive, mainly because of the wrong perception of the scene structure. More geometric information about the scene (i.e. depth maps or normal maps) can be provided in the future to improve shadow recasting. Second, the relit regions where the shadows are removed often lose texture details, as shown in Fig. 9(b). A flow vector can be used to select

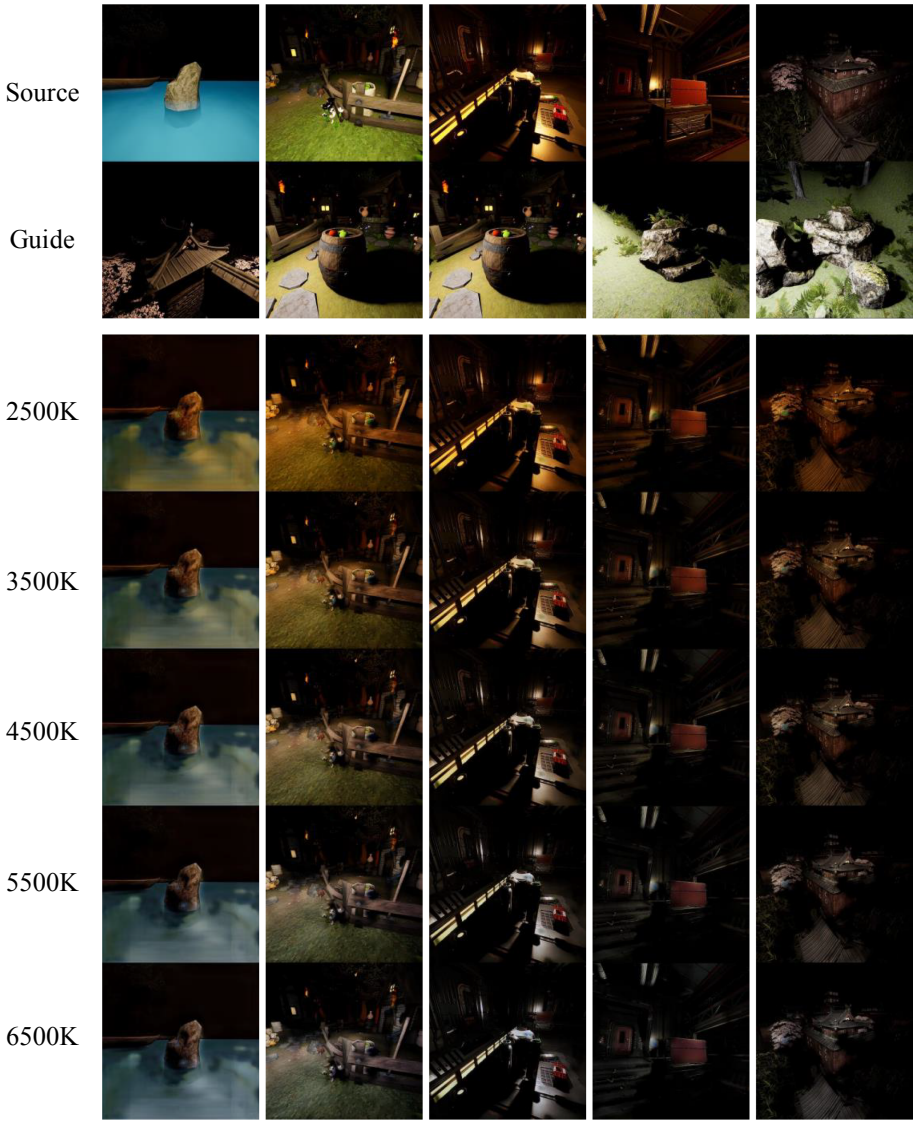

Fig. 8. More relit images with 5 different color temperatures of the light source.

similar textures from adjacent regions to fill in the region where shadow needs to be removed. Third, what we call halo artifacts often appears in the relit images, and imposing more penalties on the region of halo artifacts is one of the potential solutions in the future.

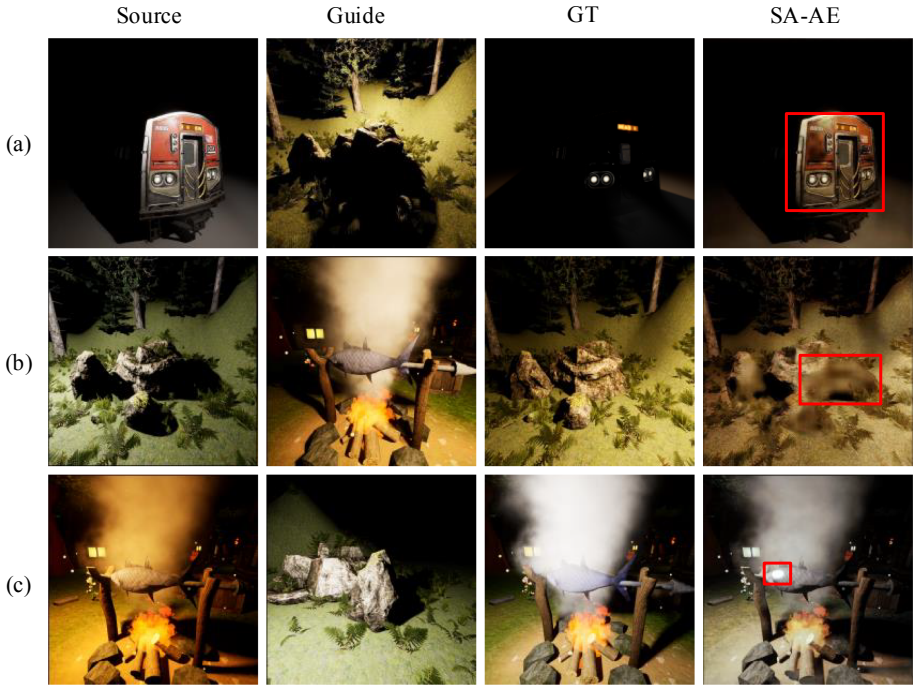

Fig. 9. Failure cases.

5 Conclusions

In this paper, we have presented a novel automatic model SA-AE for any-to-any relighting. A U-shape convolutional neural network-based autoencoder is well designed to learn an implicit scene representation to reduce the difficulty of the learning. In order to focus on the re-rendering of the relighting-related regions in the source images, an attention mechanism is also introduced into autoencoder. In addition, a lighting estimation network and a lighting-to-feature network, which are inverse processes of each other, are proposed to clearly control the relighting of the source image with a given illumination setting. The experiments show that the proposed SA-AE achieved the 1st place in terms of MPS and the 1st place in terms of SSIM in the AIM 2020 Any-to-any Relighting Challenge.

Acknowledgements. This work is supported by NSFC under Grant 61531014.

References

1. Barron, J.T., Malik, J.: Color constancy, intrinsic images, and shape estimation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 57–70. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_5
2. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 334–341. IEEE (2012)
3. Barron, J.T., Malik, J.: Intrinsic scene properties from a single RGB-D image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 17–24 (2013)
4. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1670–1687 (2014)
5. Bau, D., et al.: Semantic photo manipulation with a generative image prior. ACM Trans. Graph. (TOG) **38**(4), 1–11 (2019)
6. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 145–156 (2000)
7. El Helou, M., Zhou, R., Johan, B., Süsstrunk, S.: VIDIT: virtual image dataset for illumination transfer. arXiv preprint arXiv:2005.05460 (2020)
8. El Helou, M., Zhou, R., Süsstrunk, S., Timofte, R., et al.: AIM 2020: scene relighting and illumination estimation challenge. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020 Workshops. LNCS, vol. 12537, pp. 499–518. Springer, Cham (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
10. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
11. Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. ACM Trans. Graph. (TOG) **37**(6), 1–11 (2018)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Knill, D.C., Richards, W.: Perception as Bayesian inference. Chapter The Perception of Shading and Reflectance. Cambridge University Press, New York (1996)
14. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
15. Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Inverse rendering for complex indoor scenes: shape, spatially-varying lighting and SVBRDF from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2475–2484 (2020)
16. Matusik, W., Loper, M., Pfister, H.: Progressively-refined reflectance functions from natural illumination. In: Rendering Techniques, pp. 299–308 (2004)
17. Nestmeyer, T., Lalonde, J.F., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5124–5133 (2020)

18. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
19. Peers, P., et al.: Compressive light transport sensing. ACM Trans. Graph. (TOG) **28**(1), 1–18 (2009)
20. Reddy, D., Ramamoorthi, R., Curless, B.: Frequency-space decomposition and acquisition of light transport under spatially varying illumination. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 596–610. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_43
21. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: SfSNET: learning shape, reflectance and illuminance of faces 'in the wild'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6296–6305 (2018)
22. Sun, T., et al.: Single image portrait relighting. ACM Trans. Graph. **38**(4), 79:1–79:12 (2019)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
24. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Trans. Graph. (TOG) **37**(4), 1–13 (2018)
25. Yu, Y., Smith, W.A.: InverseRenderNet: learning single image inverse rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3155–3164 (2019)
26. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363 (2019)
27. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
28. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imag. **3**(1), 47–57 (2016)
29. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7194–7202 (2019)