

# Complementary Intrinsic from Neural Radiance Fields and CNNs for Outdoor Scene Relighting

Siqi Yang<sup>1,2,3,#</sup> Xuanning Cui<sup>1,2,4,#</sup> Yongjie Zhu<sup>5</sup> Jiajun Tang<sup>1,2</sup> Si Li<sup>5</sup> Zhaofei Yu<sup>3</sup> Boxin Shi<sup>1,2,3,4,\*</sup>

<sup>1</sup> National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

<sup>2</sup> National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

<sup>3</sup> Institute for Artificial Intelligence, Peking University

<sup>4</sup> AI Innovation Center, School of Computer Science, Peking University

<sup>5</sup> School of Artificial Intelligence, Beijing University of Posts and Telecommunications

{yousiki, cxn, jiajun.tang, yuzf12, shiboxin}@pku.edu.cn

yongjie.zhu.96@gmail.com lisi@bupt.edu.cn

## 6. Additional implementation details

**Hierarchical sampling.** We follow the hierarchical sampling strategy of NeRF [3] to implement Equation (3) and (10) through two models with the same structure: We first select  $N_c = 64$  points along each ray using stratified sampling and then draw  $N_f = 128$  points according to the piecewise-constant PDF:

$$\hat{w}_i = \frac{w_i}{\sum_{j=1}^{N_c} w_j}, w_i = T_i(1 - \exp(-\sigma_i \delta_i)). \quad (20)$$

Note that all the intrinsic components estimated ( $\tilde{\mathbf{A}}, \tilde{\mathbf{S}}, \tilde{\mathbf{N}}$ ) are accumulated through the values at the  $N_c + N_f$  points, predicted by the “fine” network, corresponding to the “fine” rendering results in NeRF [3]. Our “coarse” network uses only Equation (3) and (4) for rendering and supervision, as we find Equation (3) and (4) are sufficient to produce rough density.

**Network architectures.** We introduce the detailed network architectures of our modules. The design of our IntrinsicCNN ( $G_{\text{intrinsic}}$ ) is shown in Figure 8: The input image  $\mathbf{I}$  is downsampled with three convolutional layers, then proceeds with nine residual blocks and is finally upsampled with the skip connections from the first three convolutional layers. The first convolution kernel is in the shape of  $7 \times 7$  and others are in the shape of  $3 \times 3$ . We use instance normalization and ReLU activation after each convolution, and use sigmoid and tanh to adjust the value range of final outputs ( $\tilde{\mathbf{A}}, \tilde{\mathbf{S}}, \tilde{\mathbf{N}}$ ). The design of our LightingCNN ( $G_{\text{lighting}}$ ) is much simpler. As shown in Figure 9, it consists of four convolutional layers, a global average pooling, and a fully-connected layer, with instance normalization and ReLU activation. Our SkyMLP ( $F_{\text{sky}}$ ) has five fully-connected layers in total, while only one layer observes the lighting representation ( $\mathbf{L}$  and  $\tilde{\mathbf{L}}$ ).  $F_{\text{density}}$  consists of eight fully-connected

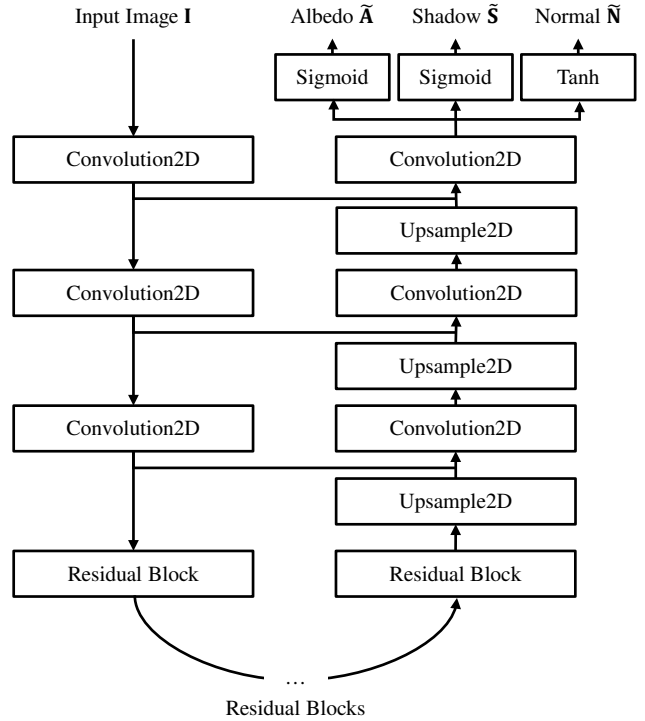


Figure 8. Network architecture of IntrinsicCNN.

layers and other MLPs ( $F_{\text{color}}, F_{\text{albedo}}, F_{\text{shadow}}$ ) consist of only one layer, keeping similar structure with NeRF [3].




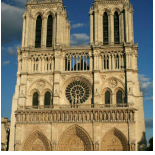

**Random jitter.** We use random jitter during our training stage to improve the stabilization of our model, by adding random Gaussian noise to the predicted  $s$  and randomly flipping the images horizontally. We empirically find that adding small Gaussian noise to  $s$  before volumetric accumulation can improve the robustness of our shadow predictor  $F_{\text{shadow}}$ :

$$s'(\mathbf{x}) = F_{\text{shadow}}(\gamma_x(\mathbf{x}), \mathbf{L}) + \varepsilon_s, \varepsilon_s \sim N(0, 0.25). \quad (21)$$






<sup>#</sup> Contributed equally to this work as first authors

<sup>\*</sup> Corresponding author

Table 2. Sample images and the numbers of training and test imaging in our dataset.

Scenes	BRANDENBURG GATE	BUCKINGHAM PALACE	GRAND PLACE	NOTRE DAME	HOUSES OF PARLIAMENT
					
Training	1000	1000	900	1000	800
Test	363	314	164	235	183

Scenes	PANTHEON EXTERIOR	TAJ MAHAL	TODAJI TEMPLE	SACRÉ COEUR	TREVI FOUNTAIN
					
Training	900	1000	800	900	1000
Test	260	312	104	279	491

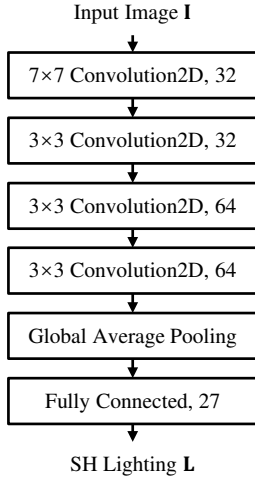


Figure 9. Network architecture of LightingCNN.

To enhance the generalization ability of IntrinsicCNN, we randomly crop the images and pseudo labels into patches with the size of  $256 \times 256$ , and then randomly flip these patches horizontally as data augmentation.

**Image collection.** We collect our outdoor landmark image dataset from the MegaDepth dataset [2] and the Phototourism dataset [1]. We keep the camera parameters and point coordinates generated by COLMAP [5]. We also cal-

culate the near and far planes according to the horizontal distance between each camera and the 3D point. After filtering out low-quality images, we create a dataset with 10 outdoor scenes and about 10,000 images in total to train our NeRF module. The detailed information of each scene is illustrated in Table 2. We have provided results on 3 scenes (SACRÉ COEUR, TODAJI TEMPLE, PANTHEON EXTERIOR) in Figures 2, 3, 5 of the main paper, and results from the remaining scenes are included in Section 7 of this supplementary material. Apart from these 10 scenes, we also collect 2 outdoor scenes from the MegaDepth dataset [2] with about 100 images in total for generalization analysis (COLLEGE OF THE FOUR NATIONS and TRIUMPHAL ARCH OF THE STAR, whose results have been provided in Figure 4 of the main paper.).

## 7. Additional results

We present more relighting and intrinsic decomposition results of our method and compare with [6, 8] in Figures 12, 13, 14.

With the help of our LightingCNN, our method can synthesize more reasonable shading according to the surface appearance of the reference image. As shown in the seventh row of Figure 12, our method relights the scene with globally bright illumination and avoids baking the blue of the sky into shading.

With the supervision of pseudo labels, our IntrinsicCNN captures more accurate intrinsic components and enhances

the relighting performance. As shown in the second row of Figure 13, our method distinguishes the effect of the sunset from the albedo of SACRÉ COEUR, while RelightingNet [6] keeps this effect in relighting results.

Without understanding the whole outdoor scene, synthesizing the cast shadow under specific lighting condition from only one input image is challenging and usually leads to unrealistic results. As shown in the sixth row of Figure 13, RelightingNet [6] produces unrealistic cast shadows as there are actually no buildings surrounding TODAII TEMPLE. Shadow generation from a single image with a limited view is out of the scope of our method and should be left to future work.

In Figure 14, we show additional results of estimated intrinsics on BRANDENBURG GATE and TAJ MAHAL scenes. Our method recovers smoother and cleaner intrinsic components than RelightingNet [6] and InverseRenderNet [8]. As shown in the fifth and sixth rows of Figure 14, our NeRF module is especially effective in recovering the shape of the dome. And with the supervision of pseudo labels generated by our NeRF module, our IntrinsicCNN can also estimate normal and shading maps more precisely.

We follow the idea of NeRF-OSR [4] and conduct another quantitative evaluation. We collect a subset of image pairs from each scene among the outdoor image collections, and project the output images to the same view point as the target images (using camera parameters recovered by COLMAP and depth maps provided by MegaDepth) to roughly align them (as shown in Figure 10). Then we evaluate the relighting capabilities of our CNN module and RelightingNet [6]. The metrics are reported in Table 3 and note that the sky regions are eliminated in the evaluation.



Figure 10. Example of image projection.

Method	MSE ↓	MAE ↓
Our CNN	<b>0.1919</b>	<b>0.3565</b>
RelightingNet	0.2165	0.3746

Table 3. Quantitative evaluation by projection.

## 8. Failure cases

There are failure cases when our method cannot tell apart shadow and albedo correctly in extremely dark regions due to the ambiguity between them. As shown in Figure 2, some regions such as the spaces between pillars, are predicted to have extremely dark shadows by InverseRenderNet [7] and RelightingNet [6], while our IntrinsicCNN prefers darker albedo and brighter shadow. Such ambiguity is mainly from NeRF module. For regions that are black in almost all observed images, our NeRF module can reconstruct these regions by setting either their albedo or shadow to black. We experimentally find that our NeRF module usually recovers these regions with darker albedo. Our IntrinsicCNN inherits this ambiguity with the generated pseudo labels so that it fools our method into confusing them. Another example is shown in Figure 11 and marked with red bounding boxes. In this example, our NeRF module predicts undesirable shadow in the red boxes, our IntrinsicCNN recovers a slightly better result, and RelightingNet [6] provides more reasonable intrinsic decomposition. However, preferring darker shadow is not always a good strategy. As shown in Figure 11, the darkness of the black windows marked with green bounding boxes comes from their albedo rather than shadow. But RelightingNet [6] produces dark shadows for these windows, while our results are more reasonable in this case.

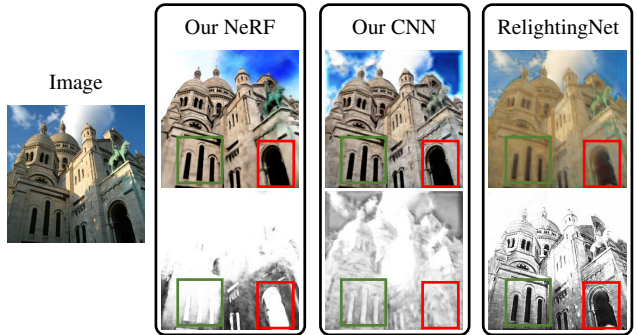


Figure 11. Example of failure case.





Figure 12. Relighting results (the target image in the middle rendered with lighting condition of the reference image in the leftmost and rightmost column) with cast shadows from our CNN and RelightingNet (RLN) [6].





Figure 13. Relighting results (the target image in the middle rendered with lighting condition of the reference image in the leftmost and rightmost column) with cast shadows from our CNN and RelightingNet (RLN) [6].



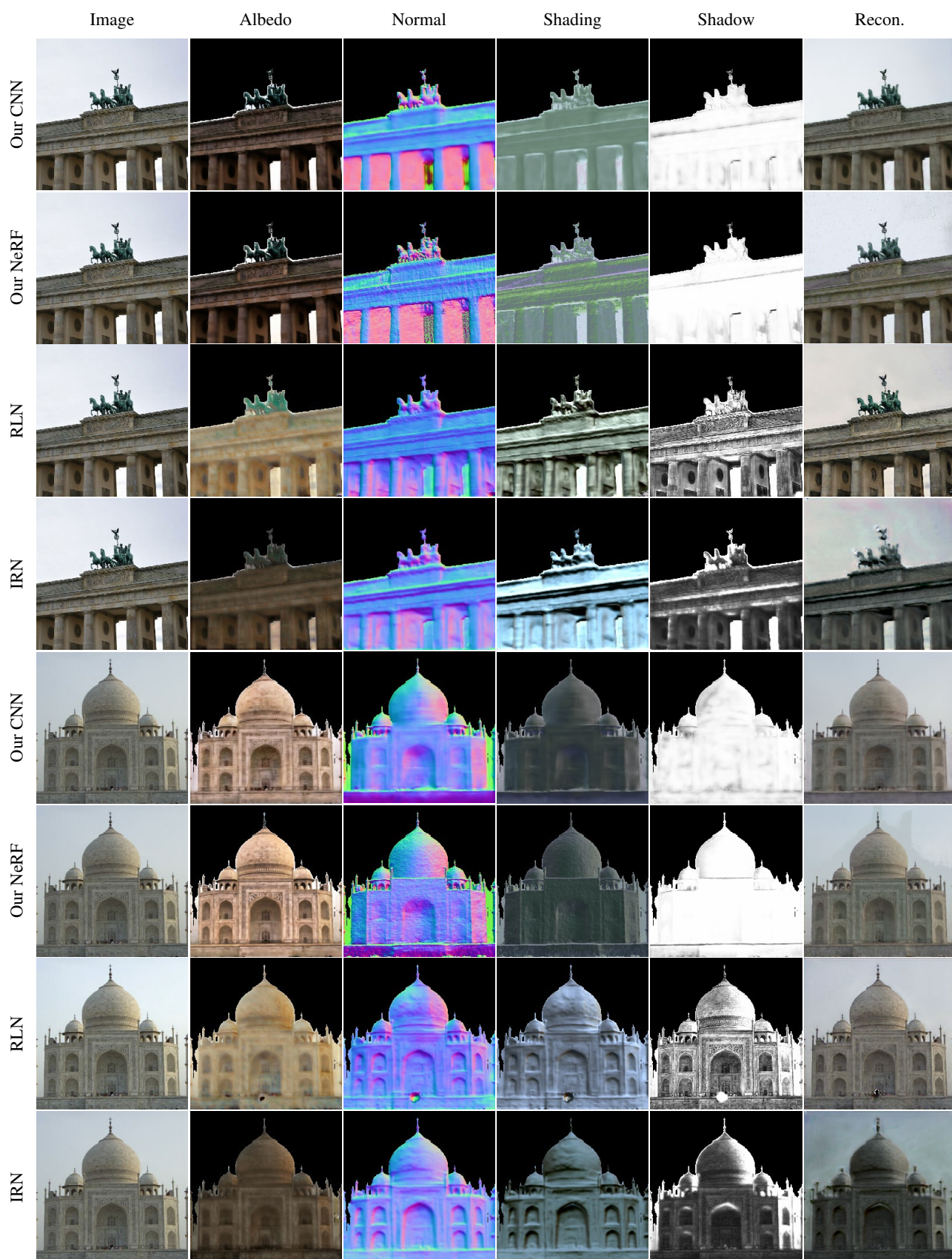


Figure 14. Estimated intrinsics (albedo, normal, shading, and shadow) and reconstruction results with shadow from our CNN module, our NeRF module, RelightingNet (RLN) [6], and InverseRenderNet (IRN) [8].

## References

- [1] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 129(2):517–547, 2021. [2](#)
- [2] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. [2](#)
- [3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [1](#)
- [4] Viktor Rudnev, Mohamed Elgharib, William Alfred Peter Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. [3](#)
- [5] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. [2](#)
- [6] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. [2](#), [3](#), [4](#), [5](#), [6](#)
- [7] Ye Yu and William AP Smith. InverseRenderNet: Learning single image inverse rendering. In *CVPR*, 2019. [3](#)
- [8] Ye Yu and William AP Smith. Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE TPAMI*, 44(7):3659–3675, 2021. [2](#), [3](#), [6](#)