Supplementary Material: Weakly-supervised Single-view Image Relighting

Renjiao Yi*, Chenyang Zhu*, Kai Xu†
National University of Defense Technology

In this supplementary material, we introduce additional experiments, discussions, details of relighting video demos, Android app implementation, the Relit dataset, as well as network and training details.

1. Mathematical proofs of Theorem 1

Theorem 1. ***Optimal rank-one approximation.*** *By SVD, $R = U\Sigma V^T$, $\Sigma = diag(\sigma_1, \sigma_2, ...\sigma_k)$, $\Sigma' = diag(\sigma_1, 0, ...)$, $\bar{R} = U\Sigma' V^T$ is the optimal rank-one approximation for R, which meets:*

$$||\bar{R} - R||_F^2 = \min_{b\in\mathcal{R}^N, c\in\mathcal{R}^d} ||bc^T - R||_F^2, \tag{1}$$

where $||\cdot||_F$ denotes the Frobenius norm of a matrix.

Proof. The objective in (1) can be written as following:

$$||bc^T - R||_F^2 = \sum_{i=1}^{d} ||c_i \cdot b - r_i||_2^2. \tag{2}$$

To minimize $||c_i \cdot b - r_i||_2^2$ while b is a fixed unit vector, $c_i \cdot b$ should be the projection of r_i onto b (r_i is the i^{th} column of R). It is equivalent to $c = b^T R$. Then we reduce the optimization problem (1) as:

$$\min_{b\in\mathcal{R}^N, ||b||_2=1} ||bb^T R - R||_F^2. \tag{3}$$

Since b is a unit vector and V are orthonormal, we can rewrite $||bb^T R||_F^2$ as:

$$\begin{aligned} ||bb^T R||_F^2 &= ||b^T R||_2^2 = ||b^T U\Sigma V^T||_2^2 \\ &= ||b^T U\Sigma||_2^2 = \sum_{i=1}^{k} (b^T u_i)^2 \sigma_i^2. \end{aligned} \tag{4}$$

By Pythagorean Theorem, and $||R||_F^2 \geq \sum_{i=1}^{n} ||c_i b||_2$, optimization problem (3) is equivalent to maximizing (4)). Since $\sum_{i=1}^{k} (b^T u_i)^2 = 1$ and $\{\sigma_i\}$ are descending, Equation (4) is maximized when $(b^T u_1)^2 = 1$. It can be accomplished by setting $b = u_1$ and $c = \sigma_1 v_1^T$, i.e., $\bar{R} = U\Sigma' V^T = bc^T$ is the optimal rank 1 approximation for R. □

2. Additional experiments

2.1. Evaluation of Light-Net

To evaluate the performance of Light-Net, we randomly sampled a testing set of 200 images from LIME [9]. It is a synthetic dataset of Bigbird [13] and ShapeNet [2] objects with ground truth normal, albedo, and shading. We use lighting coefficients predicted by Light-Net to render shading with ground truth normal maps. By comparing the rendered shading and ground truths, we can evaluate the accuracy of the estimated lighting coefficients. For quantitative evaluation, we adopt three metrics, including MSE, scale-invariant MSE, and SSIM. The results are in Table 1. We compare to two ablations from Table 1-2 in the main paper, which are "loss+" and "without joint training". We can see that for lighting evaluation, our final model produces the lowest MSE and scale-invariant MSE, and comparable SSIM to "without joint training". From visual examples in Figure 1, our model renders similar shading with ground truths, while the predicted lighting is more directional than ground truths.

Although part of the LIME dataset is used in the pre-training of Normal-Net, here we only use Light-Net for this evaluation, for which the dataset is completely unseen.

Table 1. Quantitative evaluation of Light-Net.

	The final model	loss+	w/o joint training
MSE ↓	**0.0403**	0.0452	0.0414
SMSE ↓	**0.0336**	0.0368	0.0345
SSIM ↑	0.8684	0.8652	**0.8686**

2.2. Evaluation of Spec-Net

To evaluate the performance of specular highlight extraction of Spec-Net, we compare with several prior methods on a real-image dataset from [16]. As shown in Table 2, Spec-Net outperforms other methods in both SMSE and DSSIM. Visual comparisons are in Figure 2. On real im-

*Co-first authors.
†Corresponding authors: kevin.kai.xu@gmail.com.

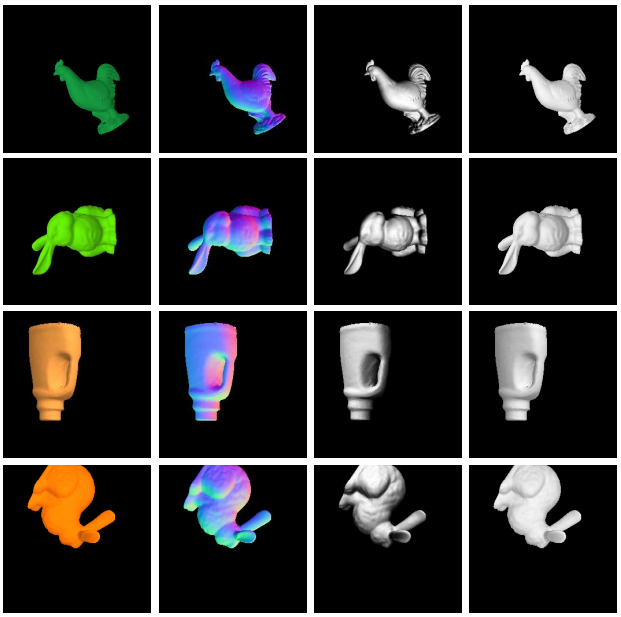

Figure 1. Visual examples of Light-Net evaluation. We render shading from ground truth normal and predicted lighting, and produce close results with ground truth shading.

Table 2. Quantitative evaluation of specular highlight separation on real images.

	[11]	[12]	[15]	Ours
MSE	0.0334	0.0305	0.0334	**0.0148**
DSSIM	0.1745	0.2087	0.1743	**0.1500**

Figure 2. Qualitative comparisons on four data from real-image specularity separation dataset from [16], captured by cross-polarization. From left to right, there are input images and diffuse components after removing specular highlights by [11,12,15], ours, and ground truths.

ages where highlights are strong, and highlight regions are saturated, most methods tend to over-extract specular highlights, while Spec-Net performs well due to the training on a large scale of real images.

2.3. Additional results of experiments in the main paper

Due to the page limit of the main paper, we show additional results for experiments in the main paper. In Figure 3, there are visual comparisons of two data from MIT intrinsics [4]. Here all methods are not fine-tuned on MIT dataset. Here SIRFS [1] and DI [10] are supervised methods. Yi [16] and ours are self-supervised, while they predict shading by a Shading-Net, and our shading is rendered from predicted normal and lighting. In Figure 5, there are visual comparisons of normal estimation to several state-of-the-art methods, for the quantitative evaluation in Table 2 of the main paper, on unseen data from Janner et al. [5]. Our method produces more details in normal maps. In Figure 4, we compare with a full relighting pipeline RelightNet [18] on real object insertion.

We provide supplementary results of Table 3 and Figure 6 of the main paper in Figure 9. Our diffuse and non-Lambertian rendering layers produce similar results with GT renderers. GT renderers are implemented by Monte-Carlo sampling of point lights following the Blinn-Phong model.

2.4. More discussions

Multi-view stereo as normal supervision. Previous method [19] uses multi-view stereo to reconstruct normal maps on outdoor building images in MegaDepth dataset [6], where ground truth depth maps are also available. Features on outdoor buildings are rich, which are suitable for multi-view stereo to reconstruction.

For object images, we explored similar approaches and found it not working for our scenarios. we use a reconstruction pipeline of adopting VisualSFM [14] to reconstruct sparse point clouds, then PMVS2 [3] to further reconstruct dense point clouds. Applying the pipeline needs multi-view images as inputs, which would introduce a heavy workload for capturing multi-view images for all objects. For demonstration, we capture additional multi-view images and test the pipeline on several objects. For each object, we capture about 50 multi-view images as inputs. From the results, we find the point clouds are very sparse due to lack of features. A example is shown in Figure 6, textureless regions are quite common on natural objects, where the features are sparse, and reconstruction results have many holes on the resulting dense point clouds. For some other object, due to the lack of features, VisualSFM even fails to reconstruct a initial point cloud. Thus, adopting SFM and MVS to reconstruct geometry is not an option for our cases.

Using median or mean reflectance vs. the singular reflectance. One may wonder whether using median or mean images of reflectance predictions in one batch will have similar results with our low-rank constraint. Firstly, losses between the median or mean reflectance of one batch

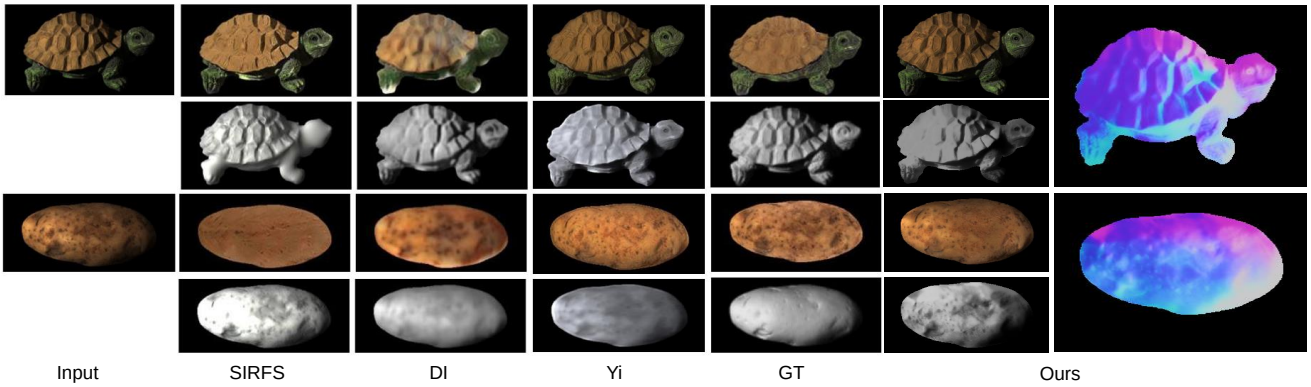

Figure 3. Qualitative comparisons on two data from MIT Intrinsics. Odd rows are input images, albedos and even rows are shadings. The normal predicted by our method is shown at right.

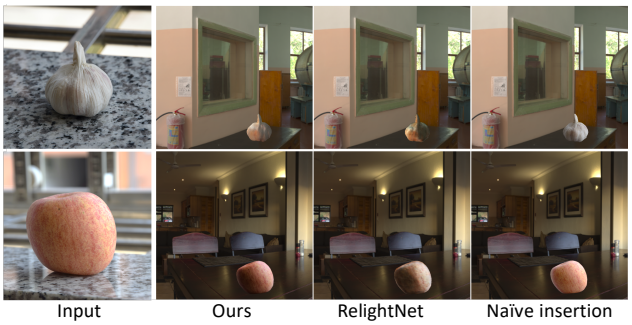

Figure 4. Qualitative comparisons on object insertion by ours, RelightNet [18] and naive insertion without relighting.

and predicted reflectance are not scale-invariant. Secondly, the median image is not differentiable. Thirdly, we perform a large amount of testing on our Relit dataset and found that singular reflectance is more robust to shadows, intensity saturations and uneven lighting, which are common cases in natural images. Some visual comparisons are shown in Figure 7, we can see that mean image may generate incorrect reflectance in some regions due to the above reasons while dominant singular reflectance generates much more reasonable reflectance maps. It is because SVD solves the dominant direction of reflectance maps, better than naive averaging. Note that we show cases on input images in Figure 7 because at the beginning of joint training, the network initializes from predicting reflectance the same as input images. We can see that using singular reflectance is much better visually, with convergence proven.

Comparisons to other low-rank losses. As mentioned in the main paper, our definition of low-rank constraint is more robust and easy to converge. We evaluate the robustness of our low-rank loss with losses from [17] and [16]. Previous low-rank losses have more than one local optima as mentioned in [16]. Thus they have to use a pretraining phase to initialize the training, and the learning rates are hand-picked to make sure the final models converge to the local optima near the pretraining results. In Table 3, we found the learning rate has to be tuned carefully. For loss$^+$ in the table, a learning rate smaller than 10^{-8} would work. For loss*, we test learning rates from 10^{-2} to 10^{-8}, and all cases degenerate to predict all-white or all-zero shadings. Setting a small learning rate also makes the training time much longer. Our loss has only one global and local optima, and it is promised to converge, and it does not suffer from degenerating.

	10^{-2}	10^{-4}	10^{-6}	10^{-8}
loss$^+$ (σ_2)	✗	✗	✗	✓
loss* (σ_2/σ_1)	✗	✗	✗	✗
Ours	✓	✓	✓	✓

Table 3. The robustness of different loss formulations. ✗ means the training degenerates to an invalid shading and ✓ means the training is converging.

Visual comparisons to previous low-rank losses (loss+ and loss*) from [16, 17] are in Figure 8. We can see that loss+ gives similar results to ours, while albedo by our method is more smooth in color, and our normal is more accurate from Table 2 in the main paper. Note that here loss+ is trained in a small learning rate of 10^{-8} to prevent degeneration. It also benefits from our large-scale Relit dataset. However, even by a small learning rate of 10^{-8}, loss* still degenerates and starts to predict all black albedo maps, as in Figure 8.

3. Limitations

There are several limitations, as well as future directions of the proposed method. One limitation is that, cast shadows (visibility) are not considered, which can further nar-

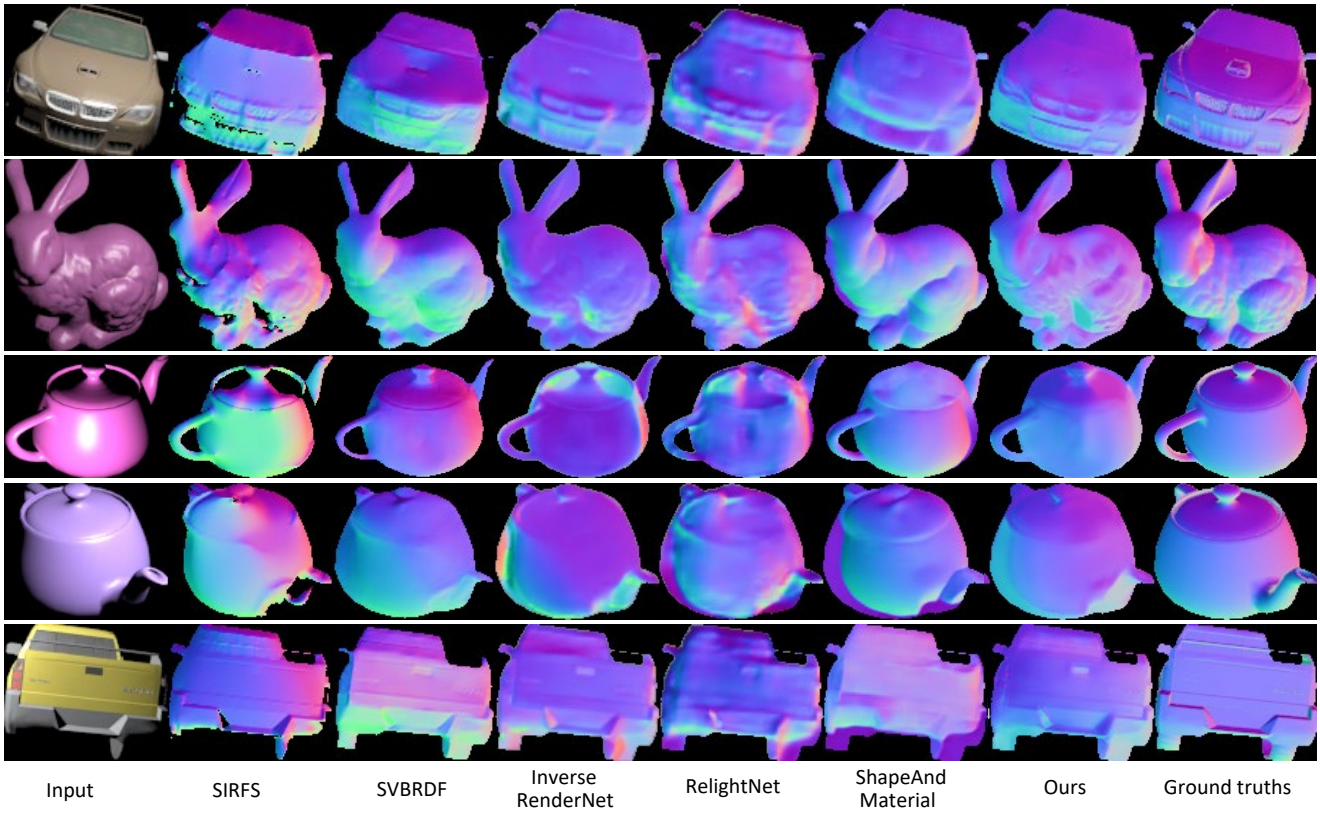

Figure 5. Normal estimation comparisons with SIRFS [1], SVBRDF [7], InverseRenderNet [19], RelightNet [18]and ShapeAndMaterial [8] on selected data from Janner et al. [5]. The reference color map can be found in Figure 5 in the main paper, where the red channel is x-axis pointing right, green channel corresponds to y-channel pointing down, and the blue channel is z-axis pointing out from the image plane.

Figure 6. Object reconstructed by VisualSFM and PMVS2. Selected multi-view image inputs are shown on the left and reconstructed dense point clouds are on the right.

row the gap between relighting results and reality. Furthermore, parametric models such as Blinn-Phong and Phong are difficult to model semitransparent and transparent materials, which are also common in real scenarios. Spherical harmonics are also limited to model high-frequency lighting components. We plan to explore these directions in the future.

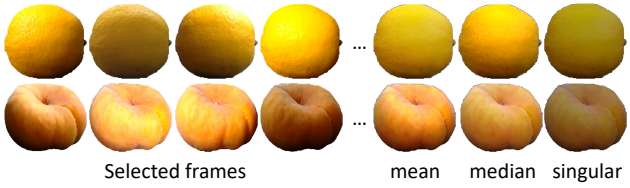

Figure 7. On each row, selected images from one batch are shown at the left. Corresponding mean image, median image and singular image are at the right.

4. Relighting demos

On the project page [1], we include many relighting videos under changing backgrounds. Relit images are inserted to target scenes to show a seamless AR object insertion effect. We demonstrate single-object insertion and multi-object insertion where multiple objects are from different input images. We also demonstrate editing the materials of objects. Object insertion is quite popular in AR applications, and most AR Apps simply adopt naive insertion without relight-

[1] https://renjiaoyi.github.io/relighting/

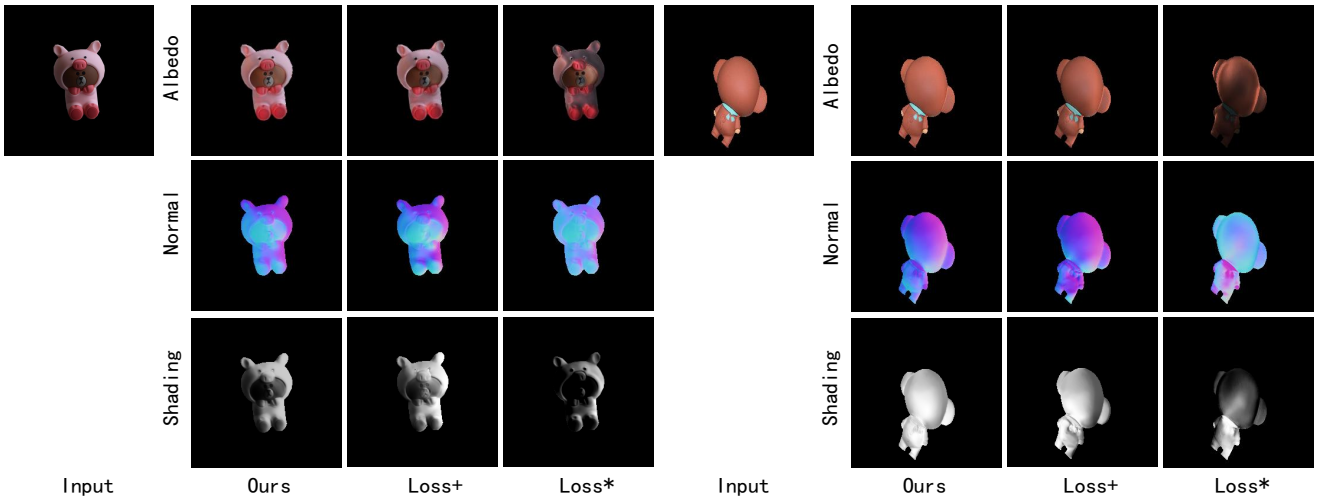

Figure 8. Visual comparison of three low-rank losses on two unseen images. Benefiting from the Relit dataset, loss* produces similar results while setting a small learning rate.

ing, such as the dancing hotdog in SnapChat, and furniture in Ikea Place. From the video, we can see our method generates much better object insertion results than naive insertion, demonstrating the importance of this problem.

Note that the backgrounds are cropped from HDR lighting panoramas, after Gamma corrections with γ as 2.2. Codes for pre-computation of Spherical Harmonic coefficients, and end-to-end inverse rendering and relighting will be released on the project page.

5. App implementation

To implement the object relighting app in the Android mobile system, we convert the network models to Pytorch Mobile and package them inside the application as assets. For object photos captured from the camera, an on-device GrabCut in OpenCV is applied to obtain the object mask. To ensure acceptable automatic segmentation results, we require users to capture the objects under a background of solid colors. For photos loading from memory, the object mask is required as an additional input. We can insert and relight single or multiple objects from different photos into the same scene, and manipulate the layouts and sizes through simple dragging, tailored for amateur users.

The application is implemented in Java, using the Android Gradle plugin of version 3.5.0 with several additional Gradle and Pytorch dependencies. The app demo video is also on the project page.

6. The Relit dataset

To capture foreground-aligned videos of objects under changing illuminations, we design an automatic device for data capture, as shown in Figure 3 in the main paper. The main part is an electric turntable painted black to avoid strong reflections. While capturing data, objects and the camera are fixed on the turntable. The turntable rotates at a uniform angular velocity of 12.6 rad/s, controlled by a remote to avoid shaking. For each video, the device is rotated by 360° for 50 seconds.

The device is chargeable and portable, enabling us to capture data under arbitrary scenes easily. The target object stays static in the image coordinate system in captured videos, with changing illuminations and backgrounds. These foreground-aligned videos can facilitate many tasks, such as image relighting, segmentation, and inverse rendering.

In summary, the Relit dataset consists of 500 videos for more than 100 objects under different indoor and outdoor lighting. Each video is 50 seconds, resulting in 1500 foreground-aligned frames under various lighting. In total, the Relit dataset consists of $750K$ images. In preprocessing, we segment the mask for one frame of each video and apply it to all frames to remove the changing backgrounds. Selected objects are shown in Figure 3 in the main paper. The objects cover a wide variety of shapes, materials, and textures.

Some foreground-aligned images in Relit dataset are shown in Figure 11-14. These are selected frames from some videos after preprocessing. Sample videos from the dataset are shown on the project page, where the device is very stable, making sure the foreground objects are staying well-aligned among all frames. The dataset is released on the project page.

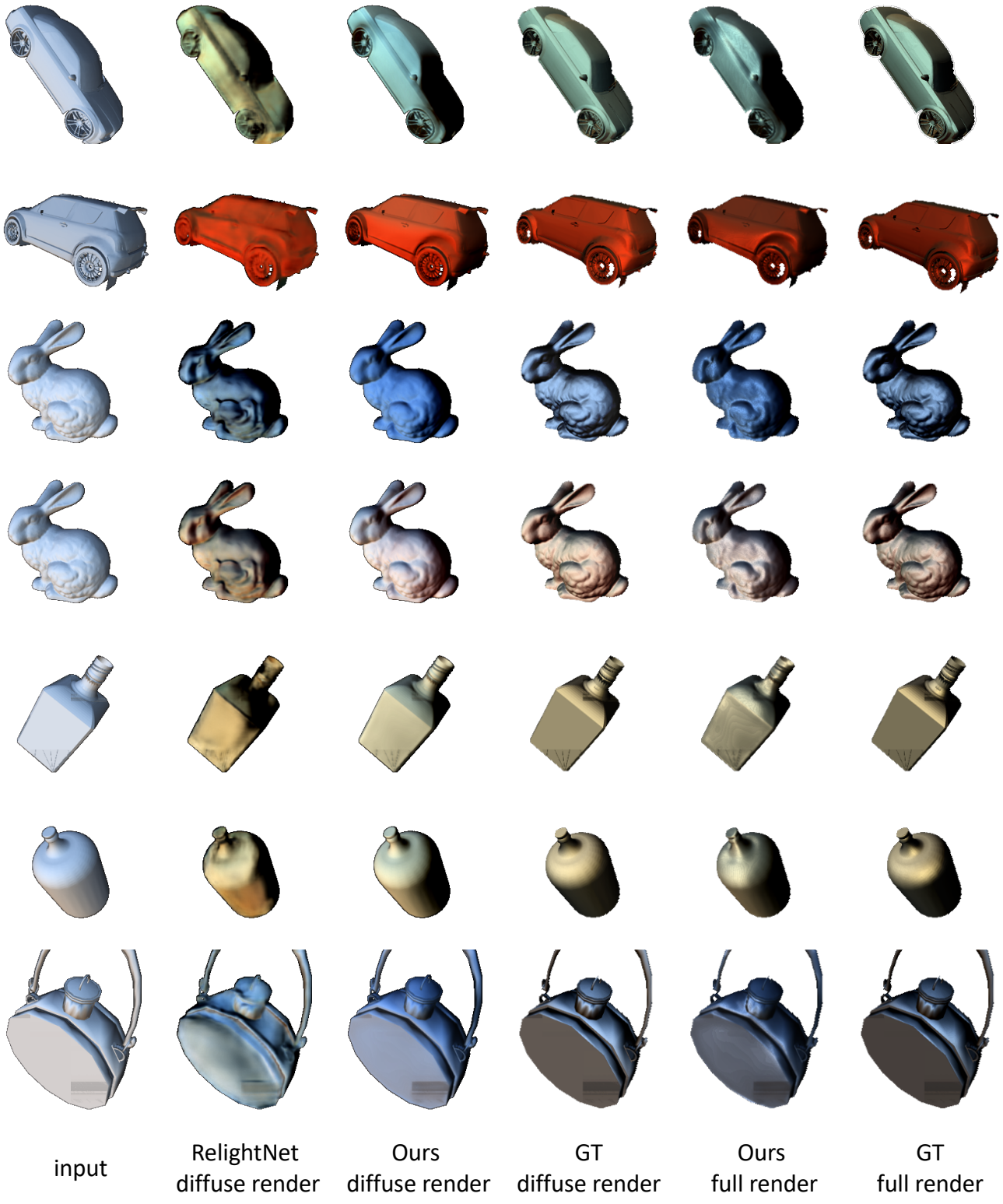

Figure 9. Quantitative evaluation on relighting with and without specularity. RelightNet [18] can only provide diffuse relighting. Baseline* denotes the images under the original lighting.

7. Network structure and training details

Normal-Net and Light-Net are the only two learnable modules in our diffuse pipeline, and an optional specular branch may be used depending on the materials of target objects. The structures are in Figure 10. Spec-Net shares the same structure with [16]. The network to regress specular reflectance S_p and smoothness α shares the same structure of Light-Net, while changing the output to 4 channels (3 for specular reflectance and 1 for smoothness).

In pretraining of Normal-Net, $50K$ synthetic images from LIME [9] are used for training. The learning rate is 10^{-4} without further adjustments. The training lasts for 50 epochs, by Adam optimizer.

In our joint training, we use the large-scale foreground-aligned images from Relit dataset. Light-Net is initialized from scratch and Normal-Net is initialed by the pre-trained model. The learning rate is 10^{-6} without further adjustments. Each round of joint training last for 3 epochs, taking 60 minutes per epoch on Tesla P40 GPU. The joint training process driven by the proposed low-rank loss converges rapidly, which takes 6 hours in total, thanks to the convergence proven in the main paper.

References

[1] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 2, 4

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[3] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 2

[4] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, 2009. 2

[5] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *NIPS*, pages 5936–5946, 2017. 2, 4

[6] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 2

[7] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 4

[8] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W. Jacobs. Shape and material capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6123–6133, June 2021. 4

[9] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6315–6324, 2018. 1, 7

[10] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2992–2992, 2015. 2

[11] Hui-Liang Shen and Zhi-Huan Zheng. Real-time highlight removal using intensity ratio. *Applied optics*, 52(19):4483–4493, 2013. 2

[12] Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[13] Arjun Singh, James Sha, Karthik S Narayan, Tudor Achim, and Pieter Abbeel. Bigbird: A large-scale 3d database of object instances. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 509–516. IEEE, 2014. 1

[14] Changchang Wu et al. Visualsfm: A visual structure from motion system. 2

[15] Takahisa Yamamoto and Atsushi Nakazawa. General improvement method of specular component separation using high-emphasis filter and similarity function. *ITE Transactions on Media Technology and Applications*, 7(2):92–102, 2019. 2

[16] Renjiao Yi, Ping Tan, and Stephen Lin. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12685–12692, 2020. 1, 2, 3, 7

[17] Renjiao Yi, Chenyang Zhu, Ping Tan, and Stephen Lin. Faces as lighting probes via unsupervised deep highlight extraction. In *ECCV*, September 2018. 3

[18] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision*, pages 84–101. Springer, 2020. 2, 3, 4, 6

[19] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 2, 4

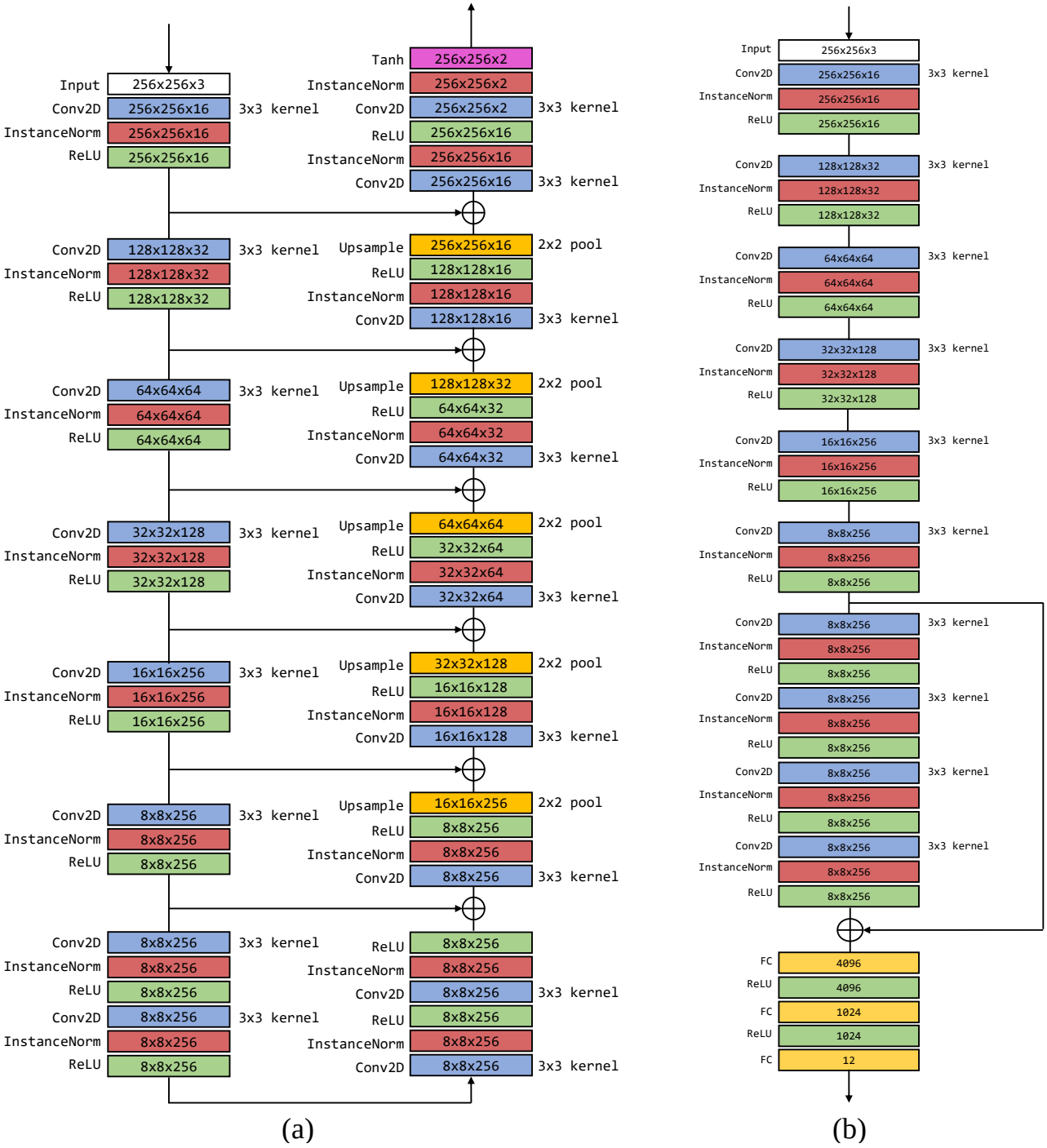

Figure 10. (a) Structure of Normal-Net. (b) Structure of Light-Net.

Figure 11. Selected frames from one video in Relit dataset.

Figure 12. Selected frames from one video in Relit dataset.

Figure 13. Selected frames from one video in Relit dataset.

Figure 14. Selected frames from one video in Relit dataset.