SILT: Self-supervised Lighting Transfer Using Implicit Image Decomposition (Supplementary Material)

Nikolina Kubiak[1]
n.kubiak@surrey.ac.uk
Armin Mustafa[1]
armin.mustafa@surrey.ac.uk
Graeme Phillipson[2]
graeme.phillipson@bbc.co.uk
Stephen Jolly[2]
stephen.jolly@bbc.co.uk
Simon Hadfield[1]
s.hadfield@surrey.ac.uk

[1] University of Surrey
Centre for Vision, Speech and Signal Processing (CVSSP)
Guildford, Surrey, UK

[2] BBC Research & Development
Salford, Greater Manchester, UK

Appendix A: Implicit image decomposition

In this section we discuss the benefits of implicit image decomposition. We also show a few examples of reflectance R and shading S maps generated by G_λ during the implicit decomposition step, as well as the new shading maps $\hat{S}$ acquired by processing the original shading map by generator G. The images are shown in Fig. 1.

As can be seen in the examples in Fig. 1, the new shading maps processed by generator G do present a unified shading representation that differs from the original shading maps which correspond to different input light directions. This aligns with our goal of translating all images, regardless of initial lighting style, to the same domain. Additionally, we can observe clear shadows in the input shading maps, particularly distinct in the bottom two rows. This shows that the varying factors are recognised as such and correctly assigned to the shading ('changing-factors') maps.

We note that the reflectance and shading images have a visible cyan tint, but, as demonstrated by all examples shown in the main paper body, this is not reflected in the final, lighting-corrected images. We believe this is due to the white balance of the indoor lighting and the multiplicative colour space. Regardless, we wish to reiterate that achieving top performance in this domain was never our objective. We only care about performing decomposition in a way that simplifies the lighting transfer task for the generator G and, thus, improves the overall performance of our model (as shown in the ablation study).

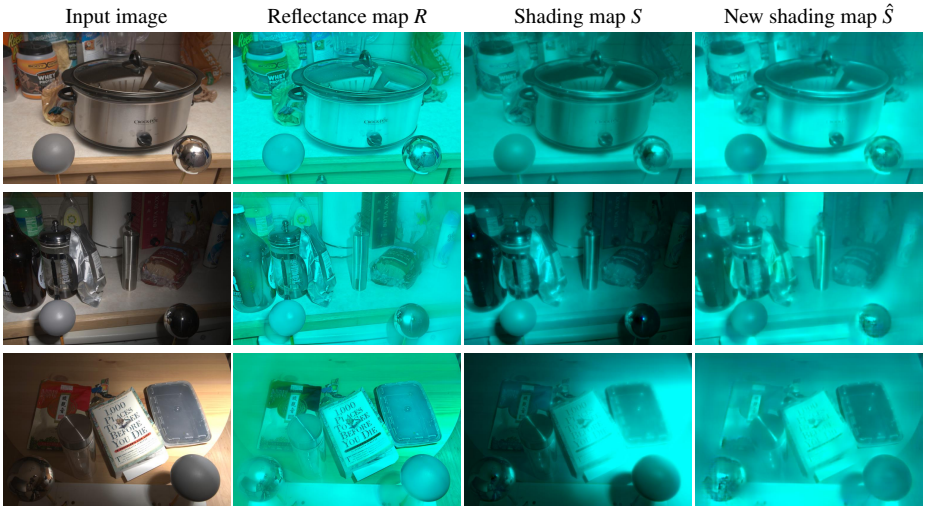

Figure 1: Examples of inputs, their implicitly decomposed reflectance and shading maps, and the new shading maps. Each row depicts a different scene and lighting direction.

Appendix B: Higher definition examples

The max. size of images used for training models discussed in the main body of the paper was 786×512 pixels. In case of the Multi Illumination [4] data, the produced outputs were additionally cropped to highlight the most interesting image areas. In this section, however, we show that SILT (trained on the Multi Illumination dataset) can produce visually pleasing results across different image sizes. We demonstrate this by applying SILT to a number of higher quality images and then performing a sweep across different image resolutions. The samples are presented in Fig. 2 and 3 and come from the OpenSurfaces [1] and Cityscapes [2] datasets, respectively.

Even though SILT was trained on lower-definition images, the model can apply lighting changes to larger images without serious artefacts. In the largest re-styled images, the titles of larger books (OpenSurfaces) and vehicle registration plates (Cityscapes) are still as clearly visible as in the input image. In Fig. 2, the two smallest images have the most visible artefacts, particularly in the top right corner. In Fig. 3, however, this effect appears to be worse for the two larger images, where a road pole on the right hand side is surrounded by white 'haze'.

The pix2pixHD paper [5] recommends using a LocalEnhancer addition to the regular generator, to allow for HD image processing. In case of 4k data, two of such enhancers would need to be trained and finetuned with the best model. However, we find that even without these add-ons our model does not visibly degrade the input image quality and performs well regardless of input size.

Appendix C: Model complexity

We report the complexity of our SILT model, specifically - the number of total vs trainable parameters, and GFLOPs. These are measured for both datasets and image sizes used for

training SILT, and shown below in Table 1.

dataset + image size	GFLOPs	# model params (of which trainable)
Multi Illumination [4] (786×512)	2792.768	79,328,552 (66,383,592)
VIDIT [3] (512×512)	1862.846	

Table 1: SILT model complexity for Multi Illumination and VIDIT datasets and their corresponding training image sizes.

Appendix D: Output similarity loss

In the main paper body we report the results obtained using the output similarity loss $\mathcal{L}_{os}$. The loss is defined in Eq. 2 and composed of the standard L1 loss and the L1 loss between spatial gradients calculated over the generated images (referred to as L1(sg) in Table 2).

Below we show that adding L1(sg) to the commonly used L1 error metric improves the model performance. We start with just the core losses (*i.e.* GAN losses $\mathcal{L}_g$ and $\mathcal{L}_d$ and the decomposition loss $\mathcal{L}_{dcp}$). When we add the L1 loss, we can observe a significant decrease across all performance metrics. When L1 is combined with L1(sg), the decline is smaller and the PSNR value actually improves slightly.

#	losses used	SSIM ↑	PSNR ↑	VGG ↓
1	core	0.730	17.075	0.428
2	core + L1	0.632	15.352	0.547
3	core + L1 + L1(sg)	0.682	17.142	0.495

Table 2: $\mathcal{L}_{os}$: The effect of adding the L1(spatial gradient) sub-loss to the L1 error metric.

Figure 2: Lighting transfer results for an OpenSurfaces image re-styled using SILT trained on the Multi Illumination dataset. The numbers next to 'Output' show the size of the input image used to create the output image (of the same size).

Figure 3: Lighting transfer results for a Cityscapes image re-styled using SILT trained on the Multi Illumination dataset. The numbers next to 'Output' show the size of the input image used to create the output image (of the same size).

References

[1] S. Bell, P. Upchurch, N. Snavely, and K. Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (SIGGRAPH)*, 2013.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[4] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A multi-illumination dataset of indoor object appearance. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.