AIM 2020: Scene Relighting and Illumination Estimation Challenge - Supplementary Material -

Majed El Helou[1], Ruofan Zhou[1], Sabine Süsstrunk[1], Radu Timofte[2], Mahmoud Afifi*, Michael S. Brown*, Kele Xu*, Hengxing Cai*, Yuzhong Liu*, Li-Wen Wang*, Zhi-Song Liu*, Chu-Tak Li*, Sourya Dipta Das*, Nisarg A. Shah*, Akashdeep Jassal*, Tongtong Zhao*, Shanshan Zhao*, Sabari Nathan*, M. Parisa Beham*, R. Suganya*, Qing Wang*, Zhongyun Hu*, Xin Huang*, Yaning Li*, Maitreya Suin*, Kuldeep Purohit*, A. N. Rajagopalan*, Densen Puthussery*, Hrishikesh P S*, Melvin Kuriakose*, Jiji C V*, Yu Zhu*, Liping Dong*, Chenghua Li*, and Cong Leng*

[1] EPFL, Switzerland
[2] ETHZ, Switzerland

Abstract. We review the AIM 2020 virtual image relighting and illumination estimation challenge [4], based on the VIDIT data [3]. This supplementary material covers the details of the remaining methods that are not included in the main paper due to limited space.

Keywords: Image Relighting, Illumination Estimation, Style Transfer

1 Track 1 methods

1.1 DeepRelight: Deep Relighting Network (DRN) for Image Light Source Manipulation

The proposed Deep Relighting Network (DRN) shown in Fig. 1 tackles the single image relighting task with three parts. First, it recovers the structure information of the scene. Second, it estimates the lighting effects (especially the shadows) for the target light source. Finally, a renderer combines the results and gives the final estimation. To improve the representation power of the auto-encoders, the down- and up-sampling processes are designed based on the back-projection theory [6, 15]. The manipulation of lighting effects needs to inpaint the shadows of the input and recast shadows following the target lighting. DRN uses the idea of generative

Majed El Helou, Ruofan Zhou, Sabine Süsstrunk *(majed.elhelou, ruofan.zhou, sabine.susstrunk)@epfl.ch*, and Radu Timofte *radu.timofte@vision.ee.ethz.ch*, are the challenge organizers, and the other authors are challenge participants.
*The main paper's appendix lists all the teams and affiliations.

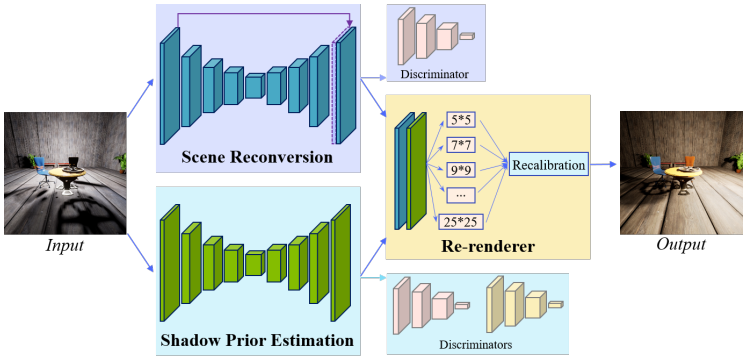

Fig. 1. The architecture of Deep Relighting Network (DRN).

adversarial learning [8, 16] that measures the light effects though a shadow-region discriminator. The renderer works through a multi-scale perception, which aggregates the global and local information for high-quality estimations.

1.2 Withdrawn: Enhanced Unsupervised Image-to-Image Translation Networks (EUnit)

The proposed EUnit is based on the structure of Unit [9], as shown in Fig. 2. In order to better perceive illumination information, we modified the dilated residual dense blocks (DRDB) [19] and combined it with Unit [9]. The full architecture consists of two encoders, one for content and another one for guidance, and two decoders. There are three discriminators for supervising the sub-networks while in the training stage. Besides, considering that the input size can affect the extraction of illumination information, multi-scale prediction was adopted for inference.

1.3 Hertz: Fast Deep Multi-scale Hierarchical Network for Single View Image Relighting

The proposed network is based on the DMSHN Model [1]. Here, the architecture consists of 3 levels where each level has a pair of encoder and decoder. The input image is downsampled by a factor of 2 and 4 to create an image pyramid. The model architecture is shown in Fig. 3. We use the Adam optimizer with an initial learning rate of 0.00001. The learning rate was decreased to 0.00005 while we trained our model for 400 epochs. The training is done with 512×512 resized input images with their corresponding resized target images from the training set. We use a batch size of 2 and no data augmentation. The chosen loss function is a weighted sum of the $\ell 1$ loss, the SSIM loss, the perceptual loss and the TV Loss. During inference, we feed the whole 1024×1024 image to the model. The proposed method is very lightweight, fast in runtime and efficient (requires less resources and can run on CPU).

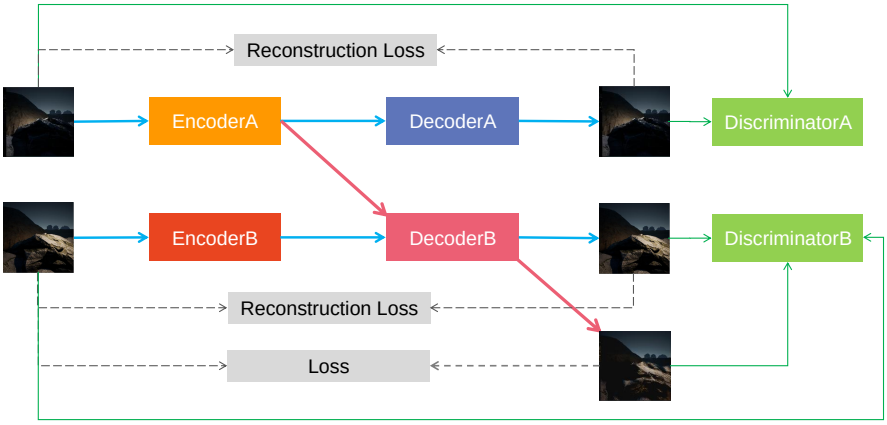

Fig. 2. Overall architecture of the EUnit network.

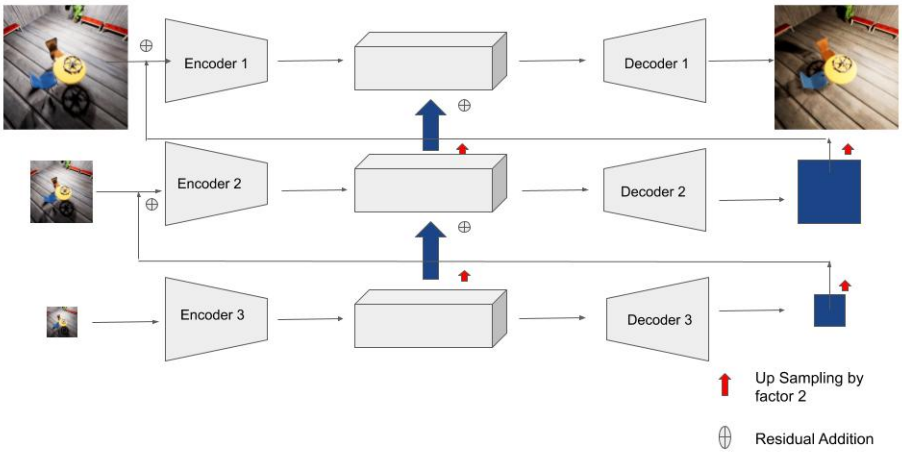

Fig. 3. Architecture diagram of Deep Multi-Scale Hierarchical Network.

1.4 Image Lab: Multilevel Attention-Based One-to-One Image Relighting

The proposed attention-based one-to-one image relighting model is shown in Fig. 4. In this architecture, the input image is passed to the coordinate convolutional layer to map the pixels to the Cartesian coordinate space [10]. The output of the coordinate convolution layers is moved to the encoder block. The proposed network is inspired by the multilevel hyper vision net [2] and has an encoder-decoder structure with supervision layers. The encoder block contains 3×3 convolution layers with two res2net [5] blocks and downsampling layers as a convolution layer with stride 2. The decoder block contains the same blocks and downsampling layers replaced with subpixel scaling layers [13]. In the skip connection, the encoder features are fed to the convolution block attention [17](CBAM).The output of CBAM is concatenated with encoder features and the upsampled layer output. The decoder block's output is directly connected to the output layer and supervised by the loss function. Except for the first decoder, previous supervision layers' outputs are concatenated with the corresponding decoder block features and fed to the next decoder block. In the last decoder, All the decoder features are upsampled and concatenated together. These features are fed to CBAM and supervised by the loss function. The shared AIM 2020 dataset consists of 1200 images, which we randomly split into 70 percent for training and 30 percent for validation. We normalize the training images to the range [0,1]. We use the Adam optimizer with a learning rate of 0.001 to 0.00001 and 500 epochs for training the model. The proposed network is trained with the IntelCore i7 processor, GTX 1080 GPU, 8GB RAM, Platform keras.

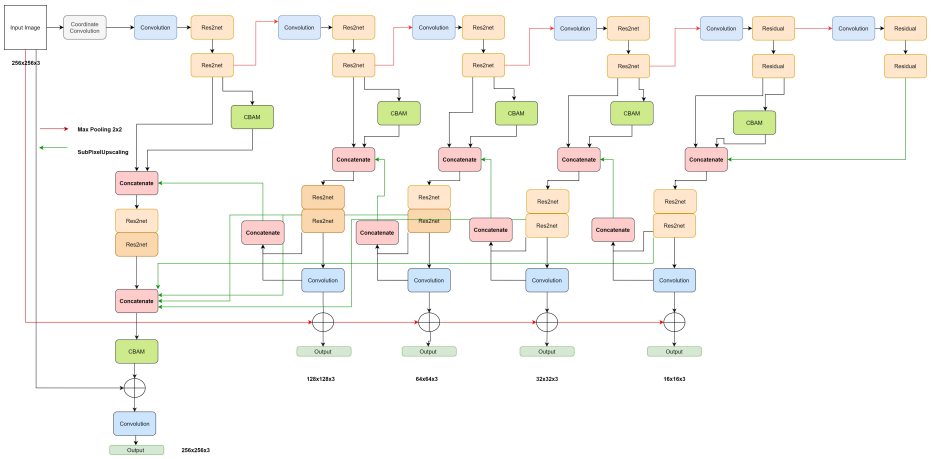

Fig. 4. The proposed light-weight multi-level supervision model. We show all the details of the architecture for better reproducibility (best viewed zoomed in on screen).

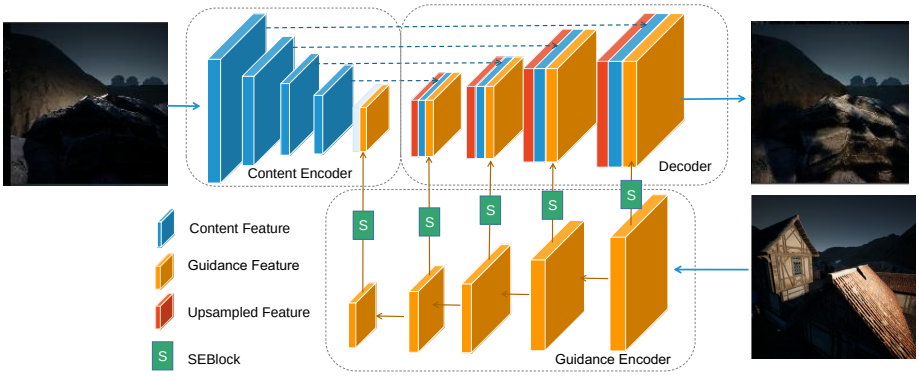

Fig. 5. The overall structure of the proposed DEDNet.

2 Track 2 methods

2.1 debut_kele: Multi-Task Learning-Based Illumination Settings Estimation

Unlike previous attempts, which aim to build two classifiers separately, the solution employs multi-task learning for the task [12], thus predicting the orientation and temperature at the same time. For the backbone network, the team uses the EfficientNet [14] pre-trained on ImageNet, and fine-tines the parameters during the training phase. For the training, they implement the framework using PyTorch and use the Adam optimizer with an $\ell 2$ penalty multiplied by 10^{-3} for the training phase. The learning rate is set to 10^{-4} and it is decayed linearly by a factor of 0.99 after each epoch. The team sets the maximum number of epochs to 30, with a patience of 5. The total running time is about 12 hours for the training while the inference can be real-time. Neither data augmentation nor external data are used in this method. A bagging-based ensemble approach is used, and the solution average the 5 folds' prediction to obtain the final submission.

3 Track 3 methods

3.1 AiRiA_CG: Dual Encoder with Decoder Network

The proposed model consists of two encoders, respectively, a content encoder and guidance encoder, and an enhanced feature fusion decoder. The overall structure is presented in Fig. 5. In order to get illumination information, a single ResNext50-32×4d [18] is chosen as the guidance encoder, and is fine-tuned on the dataset of track 2. The content encoder is the encoder of a UNet [11], which is pre-trained with the dataset of track 1. The information of the two encoders is merged in the shared decoder. The SE attention block [7] is applied to each skip-connection, with the aim of removing redundant content information.

References

1. Das, S.D., Dutta, S.: Fast deep multi-patch hierarchical network for nonhomogeneous image dehazing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020) 2
2. D.Sabarinathan, Beham, M., Roomi, S.: Moire image restoration using multi level hyper vision net. Image and Video Processing (2020), arXiv:2004.08541 4
3. El Helou, M., Zhou, R., Barthas, J., Süsstrunk, S.: VIDIT: Virtual image dataset for illumination transfer. arXiv preprint arXiv:2005.05460 (2020) 1
4. El Helou, M., Zhou, R., Süsstrunk, S., Timofte, R., et al.: AIM 2020: Scene relighting and illumination estimation challenge. In: European Conference on Computer Vision Workshops (2020) 1
5. Gao, Shang-Hua, Cheng, Ming-Ming, Zhao, Kai, Zhang, Xin-YuYang, Ming-Hsuan, Torr, Philip: Res2Net a new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019), 10.1109/TPAMI.2019.2938758 4
6. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: IEEE conference on computer vision and pattern recognition (CVPR). pp. 1664–1673 (2018) 1
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition (CVPR). pp. 7132–7141 (2018) 5
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1125–1134 (2017) 2
9. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. arXiv preprint arXiv:1703.00848 (2017) 2
10. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. Advance Neural Information Processing Systems pp. 9605—-9616 (2018) 4
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015) 5
12. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Advances in Neural Information Processing Systems. pp. 527–538 (2018) 5
13. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1874–1883 (2016) 4
14. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019) 5
15. Wang, L.W., Liu, Z.S., Siu, W.C., Lun, D.P.: Lightening network for low-light image enhancement. IEEE Transactions on Image Processing **29**, 7984–7996 (2020) 1
16. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2
17. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV) pp. 1–17 (2018) 4
18. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE conference on computer vision and pattern recognition (CVPR). pp. 1492–1500 (2017) 5

19. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1751–1760 (2019) 2