A Comparative Study on the Loss Functions for Image Enhancement Networks

Aamir Mustafa, Hongjie You, Rafal K. Mantiuk

Abstract

Image enhancement and image retouching processes are often dominated by global (shift-invariant) change of colour and tones. Most "deep learning" based methods proposed for image enhancement are trained to enforce similarity in pixel values and/or in the high-level feature space. We hypothesise that for tasks, such as image enhancement and retouching, which involve a significant shift in colour statistics, training the model to restore the overall colour distribution can be of vital importance. To address this, we study the effect of a Histogram Matching loss function on a state-of-the art colour enhancement network — HDRNet. The loss enforces similarity of the RGB histograms of the predicted and the target images. By providing detailed qualitative and quantitative comparison of different loss functions on varied datasets, we conclude that enforcing similarity in the colour distribution achieves substantial improvement in performance and can play a significant role while choosing loss functions for image enhancement networks.

Introduction

Traditional image enhancement methods, such as CLAHE [13], are built from handcrafted rules that aim at improving image quality. Finding image enhancement rules that would work for variety of content is a challenging problem. Therefore, the recent deep learning solutions attempt to learn such rules from data, which is typically represented as a collection of input and enhanced image pairs. LLNet [11] introduce a stack of auto-encoders to learn denoising and low-light enhancement jointly on the patch level of the image. Retinex theory [9] based deep learning approaches decompose an image into reflectance and illumination channels and employ deep networks to recover each channel [4, 23, 19, 25, 15]. HDR-Net [6] employed pairwise supervision by incorporating the idea of bilaterial grid filtering and local affine colour transformations for deep neural networks. Several GAN based approaches have been proposed for unpaired low level vision tasks, e.g. dehazing, deraining, super-resolution and photo enhancement [14, 10, 26, 27, 3, 7]. However, most of these works study the effect of network architectures for low-level vision tasks, rather than the importance of loss functions used to train such methods. In this work, we provide a comparative study on the loss functions for the tasks that rely on the overall colour distribution of the image, namely, image enhancement and retouching.

The choice of the loss function for training such methods can play a key role in incorporating the colour distribution of the image. Methods such as HDRNet [6] often use pixel-wise loss functions, such as mean-squared error (MSE or L_2), which enforce pixel similarity rather than a similarity of colour distribution. Recent success of learning-based methods for visual recognition have led to the advent of employing neural networks as feature extractors for loss functions. Most commonly, the networks are trained to minimize the distance between the high level features of a pre-trained VGG network [18] for the predicted and the reference image. Such methods have shown to result in trained models that produce pleasing results by enforcing similarity between the feature space of the reference and the generated image. This class of losses are often referred to as perceptual losses as they are meant to optimize the perceptual quality rather than the pixel differences.

In image enhancement tasks, in addition to the pixel intensities, the true colour distribution of the target image is significantly shifted. To this end, we study the effect of different loss functions for training an Convolutional Neural Network (CNN) based image enhancement network. We show that a loss function that can capture and enforce the inherent colour distributions between the target and the predicted images, in addition to a regular L_2 loss, provides more efficient enhancement of the resulting images. More specifically, we show that a colour distribution loss that enforces *Histogram Matching* [1] between the predicted and the reference images along the three colour channels (RGB) performs significantly better than the conventionally used loss functions. This constraint on model training, alongside the conventional loss functions (like L_2) is more suitable for tasks like image colour enhancement and image retouching rather than perceptual loss functions (like VGG and LPIPS [29]). Moreover, similar to perceptual loss functions, this loss can work in conjunction with pixel-wise losses without any additional changes in the underlying model architecture. In this work, we show that the *Histogram Matching* or *HistMatch* loss function improves the performance of HDRNet [6] for the tasks of image retouching and colour grading on diverse datasets.

Background

When training a deep network for image enhancement, we are provided with a finite set of image pairs $(x_n, y_n) : n \in (1, \ldots, N)$, where each input image $x_n \in \mathscr{X}$ is sampled from an input domain $\mathscr{X}$ and $y_n \in \mathscr{Y}$ belongs to target domain $\mathscr{Y}$. Conventionally, we train a CNN based regression model $g_\theta(\cdot)$, parameterized by θ, to promote an accurate mapping between the input images x and the target images y by minimizing the loss: $\sum_n \mathscr{L}\left(g_\theta(x_n), y_n\right)$. The loss functions used to train such networks can be mainly classified into two categories: pixel-wise losses and feature-wise losses. In image enhancement settings, pixel-wise losses like the mean squared error (L_2), L_1 are most commonly employed and have shown to perform well, however, such losses fail to incorporate the overall image colour distribution while learning a mapping from the input to the target domain. Moreover, colour insensitive image quality metrics like SSIM [22] and MS-SSIM [21] are not suitable for this task as the input-target image pairs often involve significant shift in colour.

Recent success of Convolutional Neural Networks (CNNs) in image classification tasks [18] have led to the advent of another class of loss functions known as the feature-wise losses. In such settings, the similarity between the reference and the predicted image is computed in the feature space of deep CNNs. Johnson

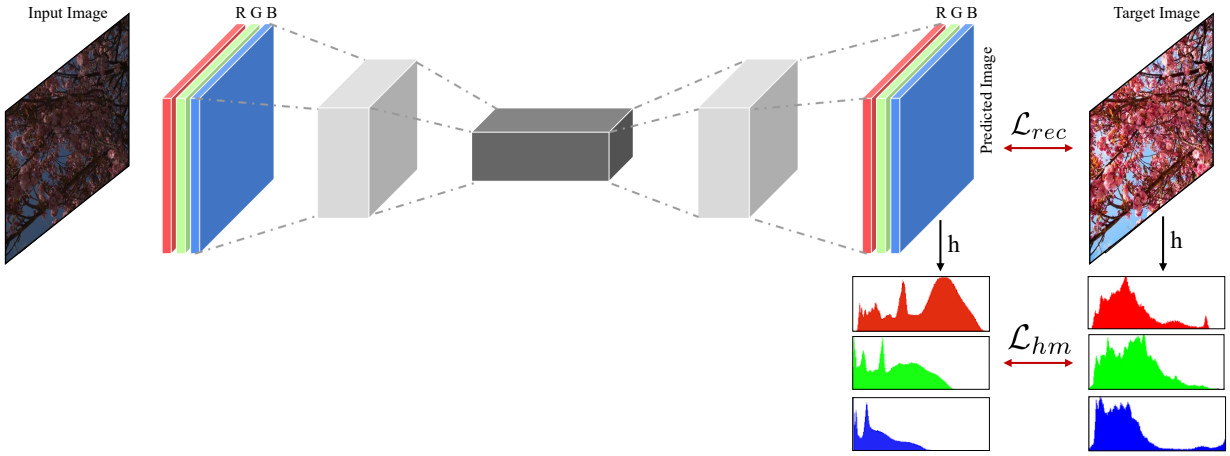

Figure 1: Figure shows the working of the proposed combination of loss functions for the task of image enhancement. Please note that the loss function is independent of the network architecture employed. The total loss function used to train the model is a weighted combination of the L_2 and the Histogram Matching Loss L_{hm}.

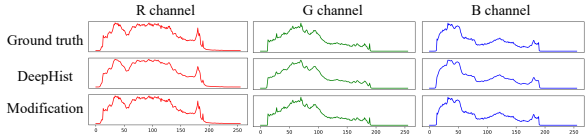

Figure 2: The plot shows the comparison of histogram estimation for DeepHist[1] with and without the modifications. Using the derivative of a function closer to a unit step performs sharper and closer estimation of the color histograms.

et al. [8] employed a pre-trained VGG-Net[18] on the ImageNet dataset [17] to extract the latent features from the target and the predicted image. L_2 norm is then computed between the VGG-Net features.

Methods such as HDRNet [6] train the model with a weighted combination of the L_2 pixe-wise and the VGG feature-wise losses [18]. Following the protocol set by [8], the feature maps for computing the perceptual loss are extracted after the relu1_2, relu2_2, relu3_3, and relu4_3 of the VGG-16 network. However, all those losses enforce similarity of local rather than global image statistics. For example, VGG features at the output of relu4_3 have the receptive field of the size 150×150 pixels. Furthermore, the extracted VGG features have been optimized for the task of image classification, not the colour representations.

Differentiable Histogram

In this section, we explain the differentiable approximation of a histogram from [1] with our modifications that allow more precise estimations for spiky and discrete RGB histograms. The differentiable histogram estimation lets us formulate the loss function that combines Earth Mover's Distance (EMD) between the histograms of the input and target images and a regular pixel-wise loss. Later we show a detailed comparison between the network optimized for learning the colour distribution and other loss functions for the tasks of image enhancement and retouching.

In digital image processing, a pixel intensity for each colour channel lies within a discrete range of K intensity values. Colour statistics of an image can be described by computing the image histogram by counting the total number of pixels in each intensity value a.k.a bins. However, considering that the image space can take any value in a continuous intensity range $[0,1]$, we can define the pixel intensity of an image pixel $p \in \Omega$ as $I(p) \in [0,1]$. As in [1], we use the Kernel Density Estimation (KDE) for approximating the channel wise density for each colour channel I as follows:

$$\hat{f}_I(i) = \frac{1}{|\Omega| B} \sum_{p \in \Omega} \mathscr{K}\left(\frac{I(p) - i}{B/\alpha}\right), \tag{1}$$

where, $i \in [0,1]$, $\mathscr{K}(\cdot)$ is the kernel, B is the bandwidth and $|\Omega|$ is the total number of pixels in the image. Due to the spiky and discrete nature of the image histograms, we choose the kernel $\mathscr{K}(\cdot)$ as a derivative of a function closer to a unit step rather than a conventional logistic regression function $\sigma(z)$, as in [1]. For this we set an additional hyper-parameter α, which is empirically set at 1000. This leads to a closer estimation of the image histogram than a sigmoid function as in [1] (see Fig. 2). The kernel is defined as follows:

$$\mathscr{K}(z) = \frac{d}{dz}\sigma(z) = \frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right), \tag{2}$$

For an efficient performance as a loss function, we divide the interval $[0,1]$ into K intervals $\{B_k\}_{k=0}^{K-1}$, with each interval of equal length of $L = 1/K$ with center $\mu_k = -1 + L(k+1/2)$. For our experiments, we conducted a grid search for the hyper-parameter K from the following set of values $K \in \{50, 100, 256\}$ and found the best results for $K = 256$. Each bin k contains the pixel intensities between the interval $B_k = [\mu_k - L/2, \mu_k + L/2]$. We can then define the probability of a pixel belonging to a certain bin, with center μ_k, for a colour channel I as follows:

$$P_I(k) = Pr(i \in B_k) = \int_{B_k} \hat{f}_I(i)\, di. \tag{3}$$

Given the kernel function $\mathscr{K}(\cdot)$ as the derivative of $\sigma(\alpha z)$ (see Eq. 2), we calculate the the function $P_I(k)$, which provides the value for the k^{th} bin in a differentiable histogram. Finally for a specific colour channel I_j, the image histogram for all bin centers and the respective probabilities is given as:

$$\mathbf{h}_j = \left\{\mu_k, P_I(k)\right\}_{k=0}^{K-1}. \tag{4}$$

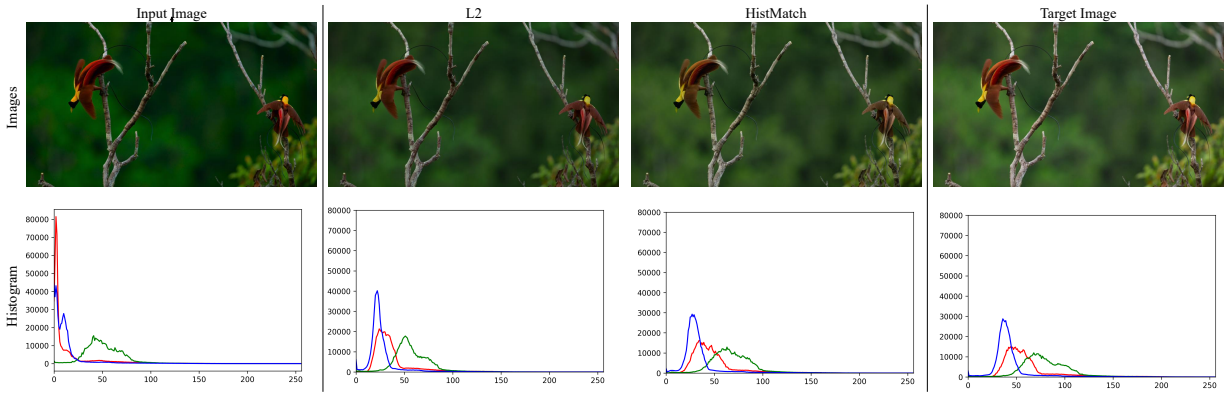

Figure 3: The figure shows a comparison of the image histogram of the images predicted using models trained with L_2 and our loss. It can be seen that our loss function results in final predicted image's colour distribution to be closer to the target image.

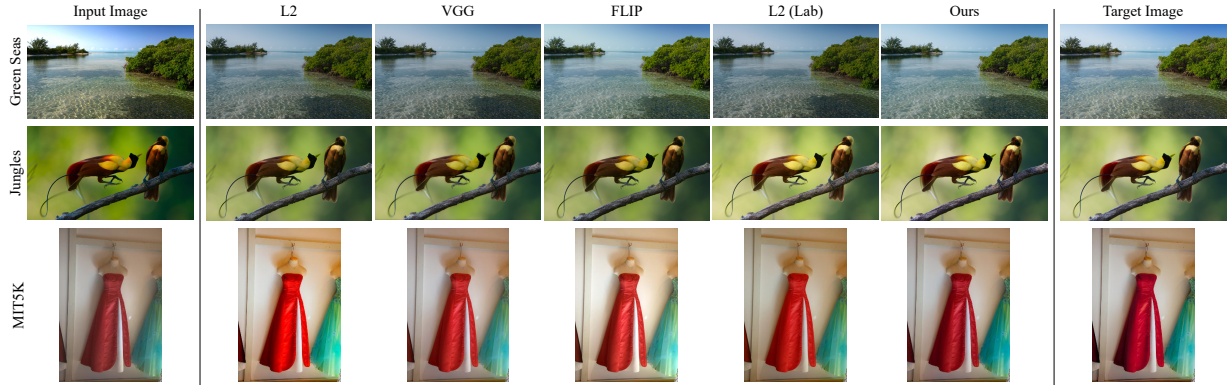

Figure 4: Qualitative comparisons of HDRNet models trained with different loss functions on three datasets. The target for MIT5k dataset is the expert C retouched image. It can be seen that addition of a constraint over the image histogram while training enforces the predicted image colour distribution to be closer to that of the target image.

In Fig. 3, we show a comparison of the computed histogram for a sample target domain image with the predicted output using different loss functions. It can be seen that addition of a histogram matching loss in the combined loss function provides enhanced colour representation at inference.

Loss Function

To compute the distance between two image histograms $\mathbf{h}_1$ and $\mathbf{h}_2$, we use the EMD [16]. Werman et al. [24] showed that the EMD between two 1D histogram vectors can be computed as the L_1 distance between the cumulative histograms. The use of L_1 loss over the cumulative density function (CDF) allows for faster convergence and easier optimization of a network. The final histogram matching loss used in our model training can be defined as follows:

$$\mathscr{L}_{hm}(\mathbf{h}_1,\mathbf{h}_2) = \sum_{i=0}^{K-1} \left| \mathrm{CDF}_i(\mathbf{h}_1) - \mathrm{CDF}_i(\mathbf{h}_2) \right|, \tag{5}$$

where, $\mathrm{CDF}_i(\mathbf{h}_1)$ is the i-th element of the cumulative density function of the channel-wise estimated histogram $\mathbf{h}_1$. The overall loss function is composed of the original objective of the CNN augmented with the histogram matching loss for each of the three RGB channels as:

$$\mathscr{L}(x,y) = \mathscr{L}_{rec} + \frac{\lambda}{3} \sum_{c=1}^{3} \mathscr{L}_{hm}(x_c,y_c), \tag{6}$$

where c is each colour channel and the scalar λ is to control the weightage given to the histogram matching loss term. Here, the reconstruction loss $\mathscr{L}_{rec}$ is the L_2 distance between the predicted image and the target image.

Data

For experimentation, two real-world datasets with input-target image pairs were used. Firstly, we explore the performance of our method on MIT-Adobe FiveK dataset [2]. The dataset consists of 5000 RAW images, and each input image is retouched by 5 colour artists according to their specific choice of image enhancement. In this work, we follow the common practice [28, 6, 3, 20, 12] of using the retouched images by expert C as the ground truth target images. We choose a random sample of 4000 images for the train set and the rest 1000 as the test set following a 80–20 train/test ratio. The images are rescaled to $480p$ resolution.

Additionally, we evaluate the efficacy of our method on video dataset, where the input and target movies are separately colour graded for Standard Dynamic Range (SDR) and High Dynamic Range (HDR) displays by colour artists. For this task, we decode two Blue-ray movies, namely, "BBC Planet Earth II Episode 3 - Jungles" and "BBC Blue Planet II Episode 5 - Green Seas" which constitute our dataset. From each input (HDR) and target (SDR) movie, we choose a sequence of every 120^{th} frame. Each frame pair was added to the dataset after computing the cross-correlation to ensure time-synchronization. We choose the

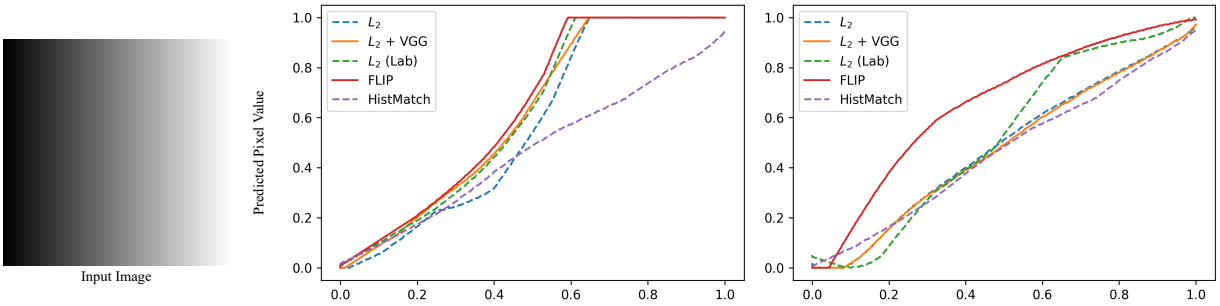

Figure 5: The plot shows the comparison of HDRNet model's prediction trained with different loss functions on a synthetic test image. The left and the right plots are for the model trained on the Green Seas movie and the MIT-Adobe FiveK dataset respectively.

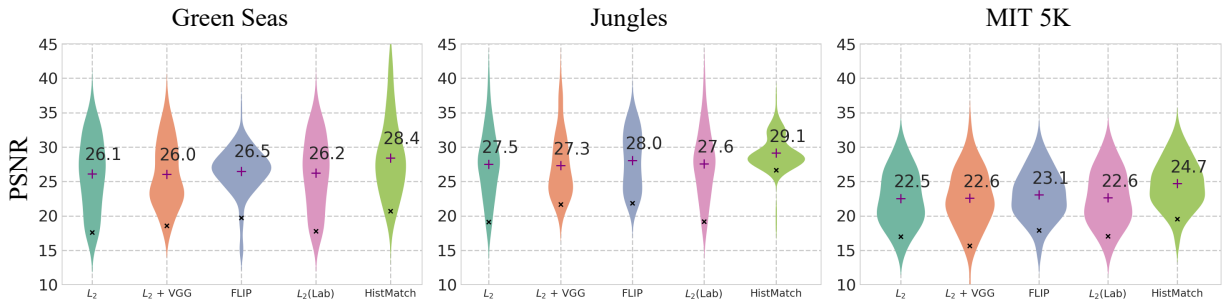

Figure 6: Quantitative performance comparison of HDRNet trained using different loss functions for the task of image enhancement on different datasets. The purple '+' in the plots show the mean and black '×' show the lowest 5^{th} percentiles. Note that our method provides an enhanced average PSNR as compared to other loss functions.

first 80% of frames for training and the rest 20% for testing. To this end, we have a total of around 500 image pairs for training and another 100 for testing for each movie. Each movie frame is rescaled to a resolution of half HD ($960p \times 540p$). Both SDR and HDR RGB pixel values are display encoded (BT.2020 + PQ for HDR, BT.709 + sRGB for SDR).

Experiments

In this section, we evaluate the effect of incorporating image histogram matching loss for image enhancements tasks where a CNN is employed. In this work, we choose the state-of-the art HDRNet [6] architecture to train our input to target domain image mapping for two applications: image tone mapping and image retouching. For image tone mapping, we evaluate the performance of HDRNet on movies dataset where the input and the target frames have been separately colour graded for HDR and SDR displays. For the task of image retouching, we use the camera RAW and the expert C retouched images from the MIT-Adobe FiveK dataset. We provide the results for baseline models (trained using L_2 loss), perceptually trained models (trained with L_2+ VGG loss, as done in the original paper [6]) and the models trained with the proposed loss function for each application. The codes for our experiments are based on PyTorch [5]. Each network was trained for 500 epochs with an initial learning rate of $1e-4$ with an exponential learning rate scheduling. For histogram construction, we use $K = 256$ bins and the width of the bin $L = 1/256$. To achieve the best performance, we need to select the values for the hyper-parameter λ, which controls the split between the L_2 loss and the histogram matching loss while training. For this we conducted a grid search from the following set of values $\lambda \in \{0.1, 1, 10\}$ and found the best results for $\lambda = 1$.

Results and Discussion

In this section, we test the importance of using underlying colour statistics while designing the loss function. For this, we evaluate the performance of the HistMatch loss function on CNN based architectures for the task of image tone mapping and colour retouching. We further compare the performance of the loss for HDRNet [6] training, with the most widely used loss functions. In Fig. 6, we provide a PSNR comparison of HDRNet trained with various loss functions as a violin plot. Furthermore, in Fig. 4, we provide a qualitative result comparison for an image sample from each dataset. It can be seen that the proposed loss function provides enhanced colour reconstruction of HDRNet.

In Fig. 5, we plot the tone curves for the HDRNet models trained with different loss functions on "Green Seas" movie and the MIT-FiveK datasets. It can be seen that the additional histogram matching loss term acts as a regularizer to prevent drastic shifts in the colour of the predicted image. We observe the same results in qualitative results shown in Fig. 4, where the additional loss term prevents over enhancement of the predicted image. The resulting images produced using the HistMatch loss are thereby closer to the target image.

Conclusion

There has been an immense surge in the methods proposed for image colour enhancement and colour retouching. However, none of the models explicitly enforce similarity in the colour statistics of the predicted and the target image. We in this work, study the effect of a different loss functions on the performance of HDRNet model for tasks that involve significant shift in the overall colour statistics. Finally, we conclude that a loss function that explicityly enforces similarity in the RGB histogram of the predicted and the target image provides much superior performance to its counterparts.

References

[1] Avi-Aharon, M., Arbelle, A., Raviv, T.R.: Deephist: Differentiable joint and color histogram layers for image-to-image translation. arXiv preprint arXiv:2005.03995 (2020)

[2] Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (Jun 2011). https://doi.org/10.1109/CVPR.2011.5995413

[3] Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6306–6314 (2018)

[4] Fu, X., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2782–2790 (2016)

[5] Ge, J.: HDRnet-PyTorch. https://github.com/gejinchen/HDRnet-PyTorch (2021)

[6] Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 118 (2017)

[7] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021)

[8] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)

[9] Land, E.H.: The retinex theory of color vision. Scientific american **237**(6), 108–129 (1977)

[10] Li, R., Pan, J., Li, Z., Tang, J.: Single image dehazing via conditional generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8202–8211 (2018)

[11] Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)

[12] Park, J., Lee, J.Y., Yoo, D., Kweon, I.S.: Distort-and-recover: Color enhancement using deep reinforcement learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5928–5936 (2018)

[13] Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing **39**(3), 355–368 (sep 1987). https://doi.org/10.1016/S0734-189X(87)80186-X, https://linkinghub.elsevier.com/retrieve/pii/S0734189X8780186X

[14] Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2482–2491 (2018)

[15] Ren, X., Li, M., Cheng, W.H., Liu, J.: Joint enhancement and denoising method via sequential decomposition. In: 2018 IEEE international symposium on circuits and systems (ISCAS). pp. 1–5. IEEE (2018)

[16] Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)

[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

[18] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

[19] Wang, J., Tan, W., Niu, X., Yan, B.: Rdgan: Retinex decomposition based adversarial learning for low-light enhancement. In: 2019 IEEE international conference on multimedia and expo (ICME). pp. 1186–1191. IEEE (2019)

[20] Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6849–6857 (2019)

[21] Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003)

[22] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (April 2004). https://doi.org/10.1109/TIP.2003.819861

[23] Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)

[24] Werman, M., Peleg, S., Rosenfeld, A.: A distance metric for multidimensional histograms. Computer Vision, Graphics, and Image Processing **32**(3), 328–336 (1985)

[25] Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. IEEE Transactions on Image Processing **30**, 2072–2086 (2021)

[26] Yang, X., Xu, Z., Luo, J.: Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[27] Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 701–710 (2018)

[28] Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

[29] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep networks as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)