Zero-shot Single Image Restoration through Controlled Perturbation of Koschmieder's Model
(Supplementary Material)

Aupendu Kar*, Sobhan Kanti Dhara*, Debashis Sen, Prabir Kumar Biswas
Indian Institute of Technology Kharagpur, WB, India

1. Zero-shot Network Architecture and Training

The zero-shot learning framework proposed in the main paper is shown in its Figure 1. It consists of a transmission map estimator M_T and an atmospheric light /global background light estimator M_A. In the main paper, their architectures are introduced briefly in Section 3.3 and the loss functions used for the training are given Sections 3.4 and 3.5. Here, we elaborate these in detail.

1.1. Network Architecture

The detailed architectures of M_T and M_A are shown in Figure 1(a). The input image is fed into the M_T network at multiple scales to perform multi-scale feature selection (see Section 1.1.2) on features extracted color channel-wise (see Section 1.1.1) to obtain the relevant features. These features are then processed by a few convolution layers to yield the transmission map estimate. [Conv2D, a, b, c] in the figure depicts a convolution layer with 'a' number of output channels, 'b' is the kernel size and 'c' is the stride. The value of 'Ch' in Figure 1(a) is discussed in Section 1.1.3. The input image is also fed into the M_A network, where it is passed through a few convolution layers that use intermediate features from the M_T network as multi-scale feature attention. The features thus obtained are subjected to global average pooling followed by a fully connected layer to yield one scalar atmospheric light /global background light estimate for each color channel. [MaxPool, 15, 7] in the figure represents a maxpool layer with a 2D kernel of size 15×15 and a stride of 7.

1.1.1 Channel-wise Feature Extraction

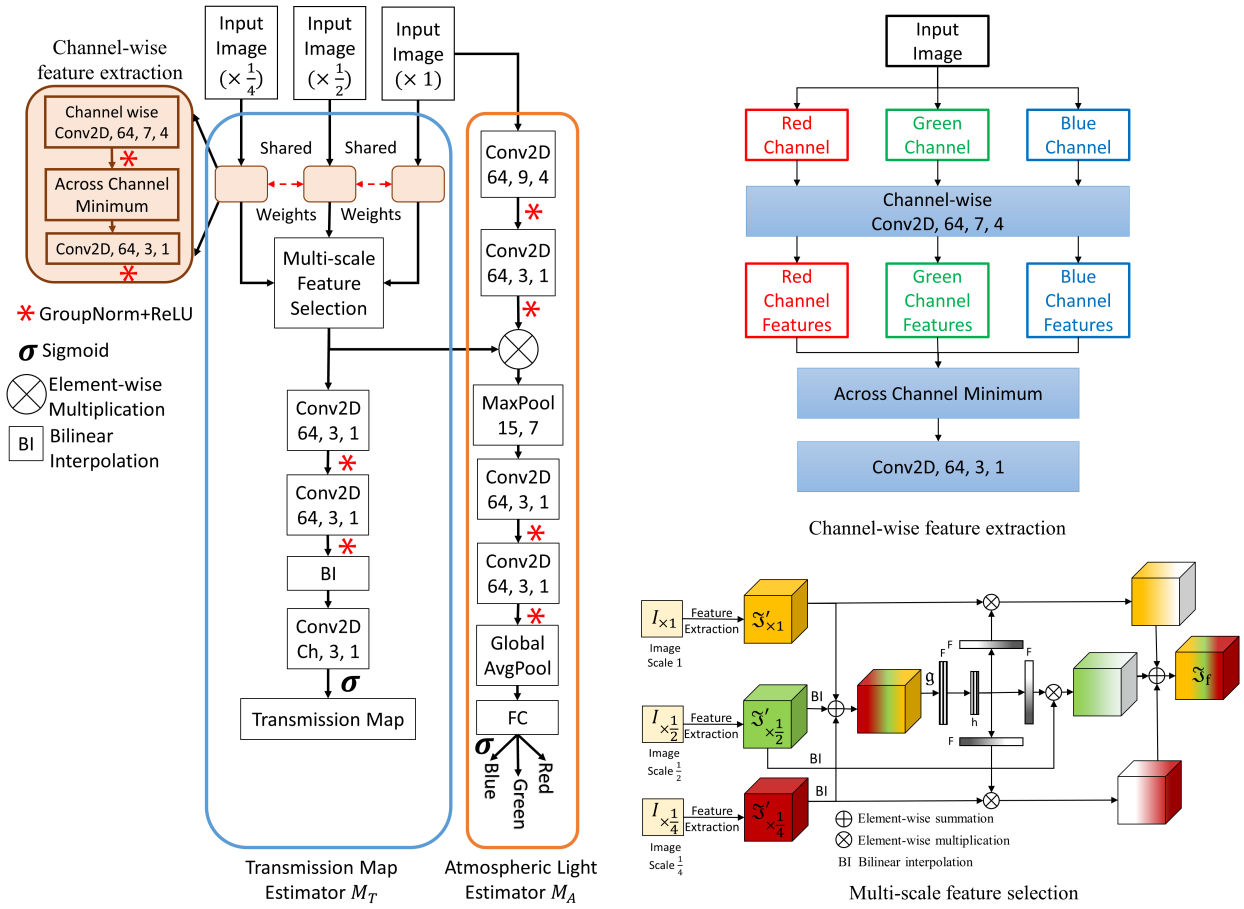
As shown in Figure 1(a), each color channel of the input image at a particular scale is processed by a convolution layer, following which the across-channel minimum of the channel-wise extracted features are retained. This across-channel min-pooling is motivated by our aim to reduce a degradation (like haze) that usually increases the intensities in the color channels. The features obtained after the across-channel minpooling is then subjected to another convolution layer. The concept of channel-wise feature extraction is inspired by the proposal in [1]. A detailed block diagram based representation of the channel-wise feature extraction is shown in Figure 1(b). The channel-wise feature extraction is applied for each scale at which the input image is considered for the multi-scale feature selection explained in Section 1.1.2.

1.1.2 Multi-scale Feature Selection

In our architecture, we consider multi-scale feature selection inspired by the kernel selection network of [2]. Relevant features that drive image restoration may correspond to different scales at different image regions, which has motivated us to consider the multi-scale operation where the network is allowed to ignore features at unimportant scales by learning small weights. The multi-scale feature selection network used in our approach is shown in Figure 1(b).

For simplicity, let us explain the network considering two different scales. Features $\mathfrak{J}'_{\times 1}$ and $\mathfrak{J}'_{\times \frac{1}{2}}$ are extracted using the channel-wise feature extraction block explained in Section 1.1.1 from the input image $I_{\times 1}$ and down-scaled input image

*Aupendu Kar and Sobhan Kanti Dhara share equal contribution
Webpage and Code: aupendu.github.io/zero-restore
Correspondence to: {*mailtoaupendu, dhara.sk*}*@gmail.com*



(a) Architecture of the zero-shot network

(b) Detailed architectures two pivotal blocks

Figure 1. Detailed network architecture of our zero-shot single image restoration framework

$I_{\times \frac{1}{2}}$, respectively. The extracted features $\mathfrak{J}'_{\times 1}$ and $\mathfrak{J}'_{\times \frac{1}{2}}$ are combined and fed into a fully connected network, which generates weights corresponding to the features. Those weights are then used to perform a weighted summation of the features at the different scales, where bilinear interpolation is used for upscaling the features at the lower scales. We perform global average pooling $\mathfrak{g}$ before feeding the features into the fully connected layer. In the figure, F is the number of vector elements which equals the number of spatial features before the average pooling, and h is the number of hidden layers in the fully connected network. The feature $\mathfrak{J}_f$ obtained is used to estimate the transmission map, and also used as an attention map for atmospheric /global background light estimation.

1.1.3 Transmission Map for Image Dehazing and Underwater Image Restoration

As discussed in the main paper, our zero-shot image restoration network is targeted for image dehazing and underwater image restoration. In this context, the value of 'Ch' in Figure 1(a) is 1 for image dehazing and 3 for underwater image restoration. This allows the M_T network to estimate a single pixel-wise transmission map for image dehazing and a three-channel pixel-wise transmission map for underwater image restoration, which is in line with the characteristics of the corresponding degradations as discussed in Section 3.1 of the main paper. The estimates of the atmospheric light for image dehazing and the global background light for underwater image restoration are obtained using the exactly same M_A network topology.

1.2. Loss function weights

As mentioned in Section 3.4 of the main paper, there are six losses in our proposed zero-shot network. The losses are the transmission relation loss $\mathcal{L}_{TR}$, the light similarity loss $\mathcal{L}_{LS}$, the pure white saturation penalty $\mathcal{L}_{SPW}$, the pure

black saturation penalty $\mathcal{L}_{SPB}$, the Gray-world assumption Loss $\mathcal{L}_{GW}$, and the total variation loss $\mathcal{L}_{TV}$. As mentioned in Section 3.5, a weighted sum of these losses are used for the training. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{TR} + \mathcal{L}_{LS} + 0.001\mathcal{L}_{SPW} + 0.001\mathcal{L}_{SPB} + 1000\mathcal{L}_{GW} + 0.001\mathcal{L}_{TV} \quad (1)$$

The weights are assigned to bring the losses to the same range so that the training takes place in a balanced manner with respect to the individual losses.

2. Adaptation for Low-light Enhancement

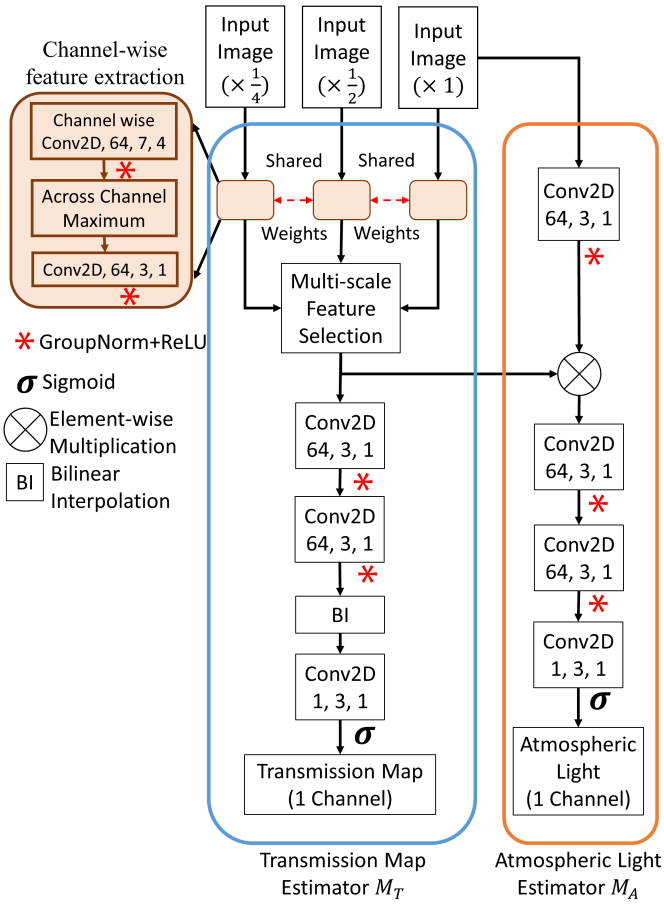


Figure 2. Network architecture for zero-shot low-light image enhancement

In Section 6.3 of the main paper, we have shown the applicability of our zero-shot framework to low-light image enhancement mentioning a few modifications required for the adaptation. We elaborate those modifications here.

2.1. Modification in the network architecture

Figure 2 shows the zero-shot architecture for low-light image enhancement. Comparing the architecture in this figure to that in Figure 1(a), we see there are a few significant changes. The across-scale maximum of the channel-wise extracted features are retained instead of the minimum, as the degradation is due to low-light which decreases the intensities in the color channels and our aim is to increase the illumination.

As discussed in Section 6.3, both the transmission and atmospheric light estimates for low-light image enhancement have been found to be pixel-wise achromatic maps in literature. Hence, the M_T network architecture is modified to estimate a single-channel pixel-wise transmission map replacing 'Ch' with 1, and a completely different M_A network architecture is used to estimate a single-channel pixel-wise atmospheric light map. This new network does not have the max pooling, the global average pooling and the fully connected blocks. It is a hierarchical set of convolution layers, where the intermediate features from M_T network is used as multi-scale feature attention.

2.2. Loss function weights

As explained in Section 6.3 of the main paper, leaving the Gray-world assumption loss out, we use the remaining five losses in our proposed zero-shot network for low-light image enhancement. Those losses are the transmission relation loss $\mathcal{L}_{TR}$, the light similarity loss $\mathcal{L}_{LS}$, the pure white saturation penalty $\mathcal{L}_{SPW}$, the pure black saturation penalty $\mathcal{L}_{SPB}$, and the total variation loss $\mathcal{L}_{TV}$. The weighted sum of these losses used for training the zero-shot low-light image enhancement network is

$$\mathcal{L} = \mathcal{L}_{TR} + \mathcal{L}_{LS} + 0.01\mathcal{L}_{SPW} + 0.01\mathcal{L}_{SPB} + 0.001\mathcal{L}_{TV} \tag{2}$$

Further, in the case of pure white and pure black saturation penalties, it is found that the ranges of these penalties originating from the red ($\mathcal{L}^{\mathcal{R}}$) and green ($\mathcal{L}^{\mathcal{G}}$) channels are higher compared to that from the blue channel ($\mathcal{L}^{\mathcal{B}}$). Hence, for balanced training, these penalties are obtained as follows

$$\mathcal{L}_{SPW} = \mathcal{L}^{\mathcal{R}}{}_{SPW} + \mathcal{L}^{\mathcal{G}}{}_{SPW} + 10\mathcal{L}^{\mathcal{B}}{}_{SPW} \tag{3}$$

$$\mathcal{L}_{SPB} = \mathcal{L}^{\mathcal{R}}{}_{SPB} + \mathcal{L}^{\mathcal{G}}{}_{SPB} + 10\mathcal{L}^{\mathcal{B}}{}_{SPB} \tag{4}$$

instead of a simple summation, yielding better enhancement results.

2.3. Ablation Study: L_1 loss vs L_2 loss in $\mathcal{L}_{TR}$ for Zero-shot Low-light Enhancement

Loss	PSNR in dB	SSIM	CIEDE2000
L_2	14.91	0.627	55.50
L_1	**17.50**	**0.695**	**46.88**

Table 1. Ablation study related to $\mathcal{L}_{TR}$ loss computation for low-light image enhancement

For reasons explained in Section 6.3, in the case of low-light image enhancement, we consider L_1 loss instead of L_2 loss used in the computation of the transmission relation loss $\mathcal{L}_{TR}$. Here, in Table 1, we compare the enhancement performance obtained using the two losses in $\mathcal{L}_{TR}$. As evident, the use of L_1 loss clearly turns out to be superior. The image dataset used for the above study is the LOL dataset of [3].

2.4. Effect of Perturbation Factor

Figure 3 shows the sensitivity of perturbation factor α. We vary α from 0.1 to 0.9 and observe average PSNR values on LOL dataset. It is observed that output performance is not sensitive to the selection of α. The standard deviation between PSNR values for different selection of α is 0.48dB.

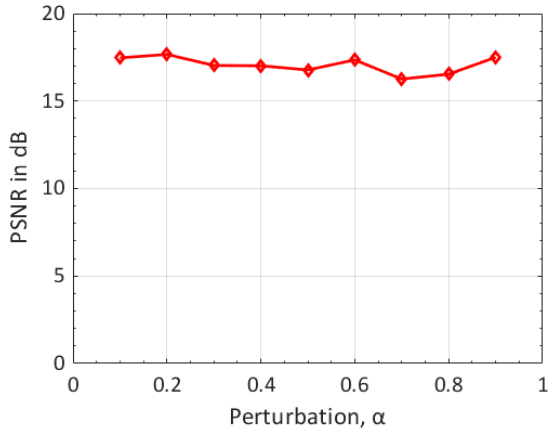


Figure 3. Sensitivity of perturbation factor α

2.5. Additional Results on Lowlight Images

Figure 4 shows more results on lowlight images. All the results are from DarkFace dataset [3].

(a) Lowlight Image (b) Enhanced Image

Figure 4. Additional results of our method on lowlight images

3. Running Time Analysis

We perform running time analysis on an NVIDIA 2080Ti GPU. Testing and training time for an image of size 256×256 are 4.41 milliseconds and 6.5 minutes respectively and for an image of size 512×512 are 4.45 milliseconds and 10.7 minutes respectively. We train the model for $10,000$ iterations and use the dehazing model to calculate training time.

References

[1] Akshay Dudhane, Kuldeep M. Biradar, Prashant W. Patil, Praful Hambarde, and Subrahmanyam Murala. Varicolored image de-hazing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[2] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 510–519, 2019. 1

[3] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 4, 5