Supplementary Material of "DiffLoss: Unleashing Diffusion Model as Constraint for Training Image Restoration Network"

Jiangtong Tan, Hu Yu, Jie Huang, Zizheng Yang, and Feng Zhao*

MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
{jttan, yuhu520, hj0117, yzz6000}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Section 1 shows the detailed derivations of the closed-form expressions of $q(x_t|x_0)$ and $q(x_{t-1}|x_t, x_0)$.
Section 2 shows more implementation details.
Section 3 shows the comparison between DiffLoss and previous loss functions.
Section 4 shows the details of the dataset.

1 The full derivations

In this section, we shows the detailed derivations of closed-form expressions for the marginal1 distribution $q(x_t|x_0)$ and the reverse diffusion step $q(x_{t-1}|x_t, x_0)$, which is directly given in the main body for conciseness.
Marginal distribution $q(x_t|x_0)$. For $t = 1$, we have $\bar{\alpha}_1 = \alpha_1$, which reduces to be consistent with single-step diffusion transition kernel $q(x_t|x_{t-1})$:

$$q(x_1|x_0) = N(x_1; \sqrt{\alpha_1}x_0, (1-\alpha_1)I). \tag{1}$$

The transition kernel $q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$ can be further written as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \text{where}\varepsilon \sim N(0, I). \tag{2}$$

Then by applying a single-step diffusion transition kernel $q(x_t|x_{t-1})$ to the above, we get

$$\begin{aligned}
x_{t+1} &\overset{(1)}{=} \sqrt{\alpha_{t+1}}x_t + \sqrt{1-\alpha_{t+1}}\varepsilon \\
&\overset{(2)}{=} \sqrt{\alpha_{t+1}}\sqrt{\bar{\alpha}_t}x_0 + \sqrt{\alpha_{t+1}}\sqrt{1-\bar{\alpha}_t}\varepsilon + \sqrt{1-\alpha_{t+1}}\varepsilon \\
&\overset{(3)}{=} \sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{\alpha_{t+1}-\bar{\alpha}_{t+1}}\varepsilon + \sqrt{1-\alpha_{t+1}}\varepsilon \\
&\overset{(4)}{=} \sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{1-\bar{\alpha}_{t+1}}\varepsilon \\
&\sim N(x_t; \sqrt{\bar{\alpha}_{t+1}}x_0, (1-\bar{\alpha}_{t+1})I).
\end{aligned}$$

Reverse diffusion step expressions. Training and derivation are performed via optimizing the usual variational bound on negative log likelihood. Readers

* Corresponding author.

interested in the derivation can refer to Ho *et al*. [6] for a thorough understanding of the formulation.

2 More implementation details.

We train the baseline network from scratch with L1 loss and our DiffLoss, respectively, where our DiffLoss obviously surpasses the baseline with L1 loss. The training performance curve with and without DiffLoss in the second setting is shown in Fig. 1. Obviously, equipped with our DiffLoss, the baseline network converges more quickly and achieves higher performance.

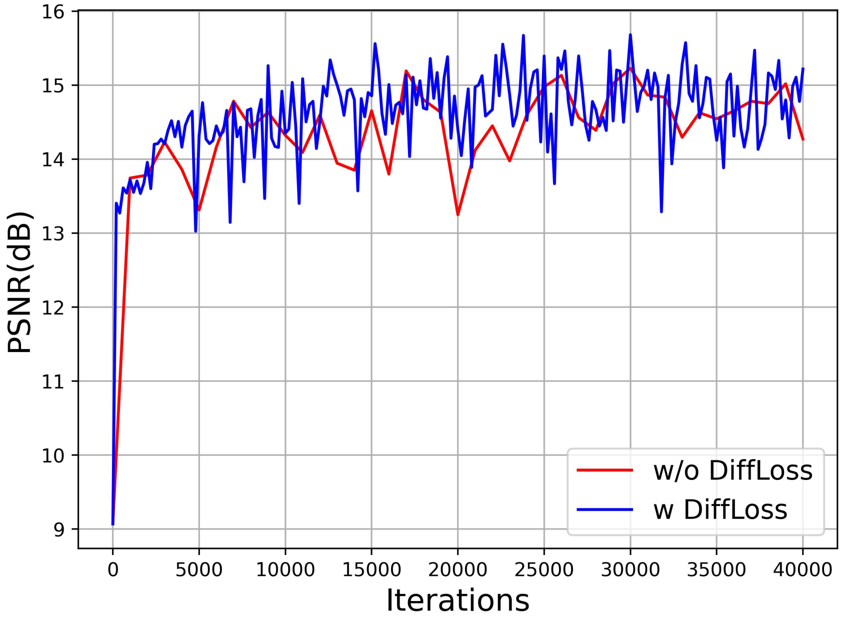

Fig. 1: Graph of PSNR with/without DiffLoss during training process of MSBDN [4] on Dense-Haze [1] dataset.

3 Comparison between DiffLoss and previous loss functions

In this section, we briefly introduce the commonly-used loss functions (L1, Perceptual, and Adversarial loss) in image restoration task, analyse their difference with our DiffLoss, and present quantitative and qualitative comparisons.For the setting of these commonly-used loss functions, we follow [3, 5].

3.1 Commonly-used loss functions.

L1 Loss. The L1 loss is used to constrain the per pixel distance of the output dehazed image and its ground-truth.

$$\mathcal{L}_1 = \|x - z\|_1 , \tag{3}$$

$$L_{total} = \mathcal{L}_1, \tag{4}$$

Perceptual Loss. The perceptual loss function is defined using high-level features extracted from a pre-trained convolutional network. Instead of encouraging the output dehazed image to be exactly the same as its ground-truth in the pixel domain, the perceptual loss aims to encourage it to have similar a feature representation in the backbone network. Specifically, following [3, 5], we use the VGG16 [7] pre-trained on ImageNet [2] as the loss network to measure perceptual similarity.

$$\mathcal{L}_{\text{per}} = \sum_{j=1}^{3} \frac{1}{C_j H_j W_j} \|\phi_j(x) - \phi_j(z)\|_2^2 , \tag{5}$$

$$L_{total} = \|x - z\|_1 + \gamma \mathcal{L}_{\text{per}}, \tag{6}$$

where H_j, W_j and C_j denote the height, width, and channel of the feature map in the j-th layer of VGG16, ϕ_j is the activation of the j-th layer. x and z are respectively the ground truth image and our dehazed result. γ is set to 0.001.
Adversarial Loss. The adversarial loss is defined based on the probabilities of the discriminator $D(z)$ over the dehazed image z as:

$$\mathcal{L}_{adv} = \sum_{n=1}^{N} -\log D(z) , \tag{7}$$

$$L_{total} = \|x - z\|_1 + \gamma \mathcal{L}_{adv}, \tag{8}$$

where, $D(z)$ is the probability of reconstructed image z to be a haze-free image, and γ is set to 0.005.

3.2 Their difference with the DiffLoss.

Previous loss functions have the following limitations: (1) L1 works in pixel space, which may produce images deviated from natural distribution. (2) The VGG16 used in perceptual loss is pre-trained for high-level task, instead of low-level image restoration task. (3) Adversarial loss treats restoration network as generator and inserts an additional discriminator network. While, adversarial loss needs to train a discriminator for every restoration dataset, which is troublesome and time-consuming. In contrast to these optimization functions, our DiffLoss is tailored for image restoration task and works in a plug-and-play manner to existing image restoration networks and datasets. DiffLoss can empower restoration networks with better perceptual quality as well as better semantic recovery.

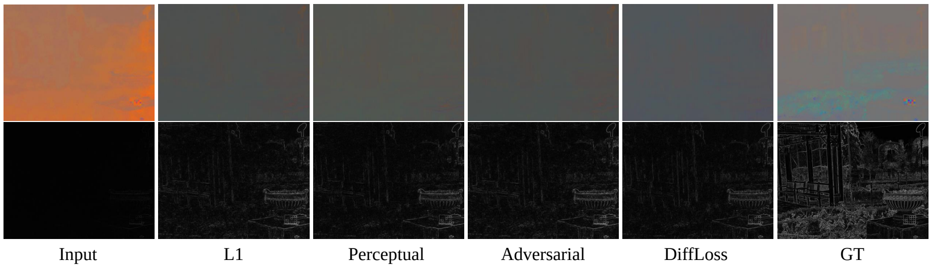

Fig. 2: Visualization of the color map (top) and structure map (bottom) for samples from Dense-Haze [1] with MSBDN [4] as baseline.

4 Details of Datasets

We use LOL dataset for low-light image enhancement, which consists of 450 pairs of training images and 50 pairs of testing images. For image derain, we use Rain13K dataset for training, which includes 13,712 pairs of training images collected from multiple datasets, and use Rain100H for testing, which includes 100 pairs of testing images. We use Dense-Haze dataset for image dehazing, which consists of 50 pairs of training images and 5 pairs of testing images.

The CUB datasets, which is used for image classification, contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing.

References

1. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In: Proceedings of the IEEE International Conference on Image Processing. pp. 1014–1018 (2019)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Deng, Q., Huang, Z., Tsai, C.C., Lin, C.W.: HardGAN: A haze-aware representation distillation gan for single image dehazing. In: Proceedings of the European Conference on Computer Vision. pp. 722–738. Springer (2020)
4. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2157–2167 (2020)
5. Fu, M., Liu, H., Yu, Y., Chen, J., Wang, K.: DW-GAN: A discrete wavelet transform gan for nonhomogeneous dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–212 (2021)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)