Supplemental Material:
DiffIR: Efficient Diffusion Model for Image Restoration

Bin Xia 1,3, Yulun Zhang 2, Shiyin Wang 3, Yitong Wang 3, Xinglong Wu 3, Yapeng Tian 3, Wenming Yang 1, and Luc Van Gool 2
1 Tsinghua University, 2 ETH Zürich, 3 ByteDance Inc, 4 University of Texas at Dallas

1. Evaluation on Real-world SR

We train and validate our DiffIR S2 on real-world SR using the same settings of Real-ESRGAN [15]. Specifically, we adopt the same loss functions of Real-ESRGAN [16], which further introduce perceptual loss and adversarial loss to the basic L1 loss. We set the learning rate of the KDSR GAN to 2 × 10^{-4}. We further validate the effectiveness of KDSR on Real-World datasets. For optimization, we use Adam with \( \beta_1 = 0.9, \beta_2 = 0.99 \). In both two stages of training, we set the batch size to 64, with the input patch size being 64. We evaluate all methods on the dataset provided in the challenge of Real-World Super-Resolution: NTIRE2020 Track1 and Tracks [9]. In addition, we also validate our DiffIR on RealSRSet [2]. Since NTIRE2020 Track1 and RealSRSet datasets provide a paired validation set, we use the LPIPS [22], DIoU [4], and PSNR for the evaluation.

The quantitative results are shown in Tab. 1. We can see that DiffIR S2 outperforms SOTA real-world SR method KDSR GAN on LPIPS, DIoU, and PSNR, consuming fewer computational costs. In addition, we can see that DiffIR S2 outperforms classic real-world SR method Real-ESRGAN on LPIPS, DIoU, and PSNR, only consuming its 63% Mult-Adds. Furthermore, compared with DM-based LDM [11], DiffIR S2 achieves much better performance consuming only 2% Mult-Adds.

We also visualize the results on NTIRE2020 Track2, which was captured with smartphones. The qualitative results are shown in Fig. 1. We can see that DiffIR S2 achieves the best performance.

2. Algorithm

The algorithm of DiffIR2 training is summarized in Alg. 1. The algorithm of DiffIR2 inference is summarized in Alg. 2.

3. More Training Details on Inpainting

We train our DiffIR for inpainting using the same loss functions of LaMa [12], which further introduce multiple perceptual losses and adversarial loss to the basic L1 loss.

For our experiments on image-inpainting in the paper Sec. 5.2, we used the code of LaMa [12] to generate synthetic masks. In training, we adopt the Adam optimizer with learning rates 0.0002 and 0.0001 for DiffIR and discriminator networks, respectively. All models are trained for 1M iterations with a batch size of 30. In addition, we use random crops of size 256 × 256 to train DiffIR on Places and CelebA-HQ. In testing, we use a fixed set of 2k validation and 3k testing samples from CelebA-HQ [5] and Places [23]. Moreover, we validate DiffIR S2 on crops of size 512 × 512 and 256 × 256 on Places and CelebA-HQ validation datasets, respectively.

4. More Training Details on SR

Compared with DIRformer for other IR tasks, we add a ×4 upsampling network [16] at the end of DIRformer for super-resolution (SR). We train our DiffIR for SR using the same loss functions of ESRGAN [16], which further introduce perceptual loss and adversarial loss to the basic L1 loss.

We train DiffIR for 1M iterations with a batch size of 64. In addition, we use random crops of size 256 × 256 to train DiffIR on DIV2K [1] (800 images) and Flickr2K [13] (2650 images) datasets for 4× super-resolution. We train our DiffIR using Adam optimizer with learning rates 0.0002 and 0.0001 for DiffIR and discriminator networks, respectively.

5. More Training Details on deblurring

Following previous works in single image motion deblurring [3, 19, 18], we train our DiffIR only using L1 loss for fair comparisons. We train DiffIR for 300K iterations with the initial learning rate 2^{-4} gradually reduced to 1^{-6} with the cosine annealing [7]. Following previous work [18], we progressively increase patch size and decrease batch size. Specifically, we start training with patch size 128 × 128 and batch size 64. The patch size and batch size pairs are updated to [(160 × 160, 40), (192 × 192, 32), (256 × 256, 16), (320 × 320, 8), (384 × 384, 8)] at iterations [92K, 156K, 204K, 240K, 276K].
Table 1. $4 \times 4$ SR quantitative comparison on real-world SR competition benchmarks. The Mult-Adds are computed based on an LR size of $256 \times 256$. Best and second best performance are marked in bold and underlined, respectively. The bottom two methods marked in gray adopt the diffusion model.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LPIPS↓</td>
<td>DISTS↓</td>
</tr>
<tr>
<td>BSRGAN [21]</td>
<td>1.18</td>
<td>0.3648</td>
<td>0.1676</td>
</tr>
<tr>
<td>Real-ESRGAN [15]</td>
<td>1.18</td>
<td>0.3629</td>
<td>0.1609</td>
</tr>
<tr>
<td>KDSRv2-GAN [17]</td>
<td>0.86</td>
<td>0.3610</td>
<td>0.1627</td>
</tr>
<tr>
<td>LDM [11]</td>
<td>37.25</td>
<td>0.4369</td>
<td>0.1982</td>
</tr>
<tr>
<td>DiffIR$_{S2}$ (Ours)</td>
<td>0.74</td>
<td><strong>0.3527</strong></td>
<td><strong>0.1588</strong></td>
</tr>
</tbody>
</table>

**Algorithm 1** DiffIR$_{S2}$ Training

**Input:** Trained DiffIR$_{S1}$ (including CPEN$_{S1}$ and DIRformer), $\beta_t (t \in [1, T])$.

**Output:** Trained DiffIR$_{S2}$.

1. Init: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$.
2. Init: The DIRformer of DiffIR$_{S2}$ copies the parameters of trained DiffIR$_{S1}$.
3. for $I_{LQ}, I_{GT}$ do
4. $Z = \text{CPEN}_{S1}(\text{PixelUnshuffle(Concat}(I_{GT}, I_{LQ})))$. (paper Eq. (5))
5. **Diffusion Process:**
6. We sample $Z_T$ by $q(Z_T | Z) = N((Z_T; \sqrt{\bar{\alpha}_T}Z, (1 - \bar{\alpha}_T)I)$ (i.e., diffusion process. paper Eq. (10))
7. **Reverse Process:**
8. $\hat{Z}_T = Z_T$
9. $D = \text{CPEN}_{S2}(\text{PixelUnshuffle}(I_{LQ}))$ (paper Eq. (12))
10. for $t = T$ to 1 do
11. $\hat{Z}_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( \hat{Z}_t - \epsilon_\theta(\text{Concat}(\hat{Z}_t, t, D)) \frac{1-\alpha_t}{\sqrt{\alpha_t}} \right)$ (paper Eq. (11))
12. end for
13. $\hat{Z} = \hat{Z}_0$
14. $\hat{I}_{HQ} = \text{DIRformer}(I_{LQ}, \hat{Z})$
15. Calculate $L_{diff}$ loss (paper Eq. (13)).
16. end for
17. Output the trained model DiffIR$_{S2}$.

6. More Visual Comparisons on Inpainting

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA inpainting methods (ICT [14], LaMa [12], and RePaint [8]). The results are shown in Fig 2. We can observe that our DiffIR$_{S2}$ can produce more realistic and reasonable structures and details than other competitive inpainting methods.

7. More Visual Comparisons on SR

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA GAN-based SR methods. The results are shown in Figs 3 and 4. Our DiffIR$_{S2}$ achieves the best visual quality containing more realistic details.

8. More Visual Comparisons on Deblurring

In this section, we provide more qualitative comparisons between our DiffIR$_{S2}$ and SOTA image motion deblurring methods. The results are shown in Fig 5. Our DiffIR$_{S2}$ has the best visual quality containing more realistic details close to corresponding HQ images.

References

Algorithm 2 DiffIR$_{S2}$ Inference

Input: Trained DiffIR$_{S2}$ (including CPEN$_{S2}$ and DIRformer), $\beta_t (t \in [1, T])$, LQ images $I_{LQ}$.

Output: Restored HQ images $\hat{I}_{HQ}$.

1: Init: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$.
2: Reverse Process:
3: Sample $\hat{Z}_T \sim \mathcal{N}(0, 1)$
4: $D = \text{CPEN}_{S2}(\text{PixelUnshuffle}(I_{LQ}))$ (paper Eq. (12))
5: for $t = T$ to 1 do
6: $\hat{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \hat{Z}_t - \epsilon_\theta(\text{Concat}(\hat{Z}_t, t, D)) \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \right)$ (paper Eq. (11))
7: end for
8: $\hat{Z} = \hat{Z}_0$
9: $\hat{I}_{HQ} = \text{DIRformer}(I_{LQ}, \hat{Z})$
10: Output restored HQ images $\hat{I}_{HQ}$.

Figure 1. Visual comparison of $4 \times$ real-world super-resolution methods. Zoom-in for better details.
Figure 2. More visual comparisons of inpainting methods. Zoom-in for better details.
Figure 3. Visual comparison of 4× image super-resolution methods. Zoom-in for better details.
Figure 4. Visual comparison of 4× image super-resolution methods. Zoom-in for better details.
Figure 5. Visual comparison of single image motion deblurring methods. Zoom-in for better details.