Multi-level Encoder-Decoder Architectures for Image Restoration (Supplementary Material)

Indra Deep Mastan and Shanmuganathan Raman
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
{indra.mastan, shanmuga}@iitgn.ac.in

In this supplementary material, we discuss more details of our multi-level encoder-decoder framework (med) and the network components. We also provide more image restoration results.

1. Image Restoration Procedure

In Algorithm 1, we have given the general process for image restoration. The function f prepares the network input z by adding noise or resizing the corrupted image $\hat{I}$. The network input z could also be generated randomly. Based on the network input preparation methods, we have different convergence times to the optimal solution [4].

In the multi-level encoder-decoder framework (Fig. 1), the generator is denoted by G and the enhancer is representation by E. The output of the image restoration is the generator output $G_{\theta^*}(z)$ instead of $\mathcal{F}_{\theta^*}(z) = E_{\theta^*} \circ G_{\theta^*}(z)$. The enhancer E_{θ^*} is used to improvise the output of the image restoration.

We have shown that the structure of network $\mathcal{F}_\theta$ depends upon the restoration task. For example, the network without skip connection is preferable for region inpainting (Fig. 10 of the manuscript) whereas for super-resolution the network with skip connections is desirable (Fig. 11 of the manuscript).

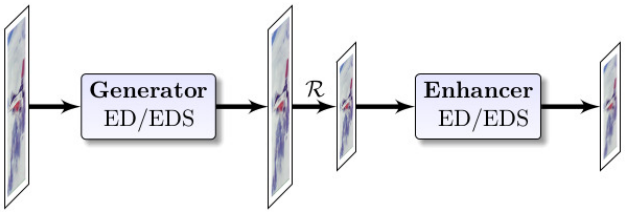


Figure 1: **Pictorial representation of multi-level encoder-decoder framework.** An abstraction of the multi-level network architectures where $\mathcal{D}$ is an operator to resize the tensor at the intermediate layer.

ImageRestoration **(corrupted image =** $\hat{I}$**)**

/* prepare the network input z from the corrupted image $\hat{I}$ using f */

$z = f(\hat{I})$

/* Create a image restoration network using framework $\mathcal{F}$ */

$\mathcal{F}_\theta = E_\theta \circ G_\theta$

/* Here, G and E are encoder-decoder network (ed) or a composition of ed */

/* Next *minimize* the loss function $\mathcal{L}$. */

$\theta^* = \arg\min_\theta \mathcal{L}(\mathcal{F}_\theta(z), \hat{I})$

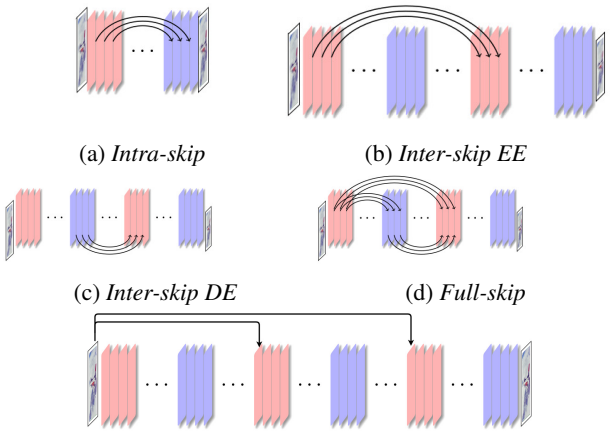
return $G_{\theta^*}(z)$

Algorithm 1: Image restoration algorithm. Given a corrupted image $\hat{I}$, *ImageRestoration* outputs $G_{\theta^*}(z)$ as the restored image. Pictorial representation of image restoration framework $\mathcal{F}$ is given in Fig. 1.

2. Classification of Skip Connections

The skip links are mainly classified into the following two types. (I) *Intra-skip*: skip links within an ed network and (II) *Inter-skip*: skip links between the layers of two different ed subnetworks. *Inter-skip* links can further be classified into the following types. (II.a) Skip links between the layers of the encoder/decoder of the first ed to the encoder layers of the second ed. We call them *Inter-skip EE*/*Inter-skip DE*. (II.b) Skip links between layers of the encoder/decoder of the first ed to the decoder layers of the second ed. We call them *Inter-skip ED*/*Inter-skip DD*. *Intra-skip* links, *Inter-skip EE*, and *Inter-skip DE* are pictorially shown in Fig. 2.

In Fig. 3, we show the effects of the skip connections on



(a) *Intra-skip* (b) *Inter-skip EE*

(c) *Inter-skip DE* (d) *Full-skip*

(e) *Cascading* of *network input*

Figure 2: **Network components.** Types of skip connections (from (a) to (d)) and cascading of the network input (e). Layers of the encoder are in *red* and layers of the decoder are in *blue*. (a) *Intra-skip*: the skip connections within EDS network. (b) *Inter-skip EE*: the connections from the first encoder to second encoder. (c) *Inter-skip DE*: the connections between the first decoder to second encoder. (d) *Full-skip*: both the *Intra-skip* connections and *Inter-skip* connections (*Inter-skip EE*, and *Inter-skip DE*) are present. (e) Cascading of the network input.

image inpainting when the corrupted image is prepared by removing 95% pixels uniformly at random. The MEDSF network has skip connections, and MED does not have skip connections. The effects from skip connections for single encoder-decoder (ed) network were shown in [4]; whereas we consider the case when the restoration is performed using a composition of the ed networks allowed by the med framework. In other words, the med framework provides the study of the image prior for various configurations of the skip connection (Fig. 2).

3. Perceptual Quality Comparison

As we see in Fig. 7 to Fig. 13, the perceptual quality of the generated images is comparable to the baseline methods *despite* we use very high capacity networks. This observation could also be validated using the SSIM index of our methods are close to the other methods given in Table 3 of the manuscript. We now provide further details of the image restoration tasks we have performed.

- **Inpainting.** We have given the perceptual qualitative comparison for inpainting 90% missing pixels in Fig. 7 and Fig. 8. The quantitative comparison is provided in the Table 1 and Table 2. After getting the restored image output, we perform the post-processing using the selective gaussian blur filter.

- **Super-resolution.** We have given a visual comparison of the generated images in Fig. 9, Fig. 10, and Fig. 11. The quantitative comparison between our method and baseline methods is given in Table 3 and Table 4. RGB images in Set14 dataset have three channels, and our MED and MEDSF networks also output RGB images having three channels. However, we downloaded the *super-resolution* output of DIP [4] from the project page had images with four channels (including the *alpha* channel). Therefore, to get a *fair comparison*, we have reconstructed DIP output before drawing the comparison. One could observe that the perceptual quality of the generated super-resolution output is comparable to baseline methods. We use the post-processing method proposed by Li et al. [1] to improvise the visual quality of the generated images.

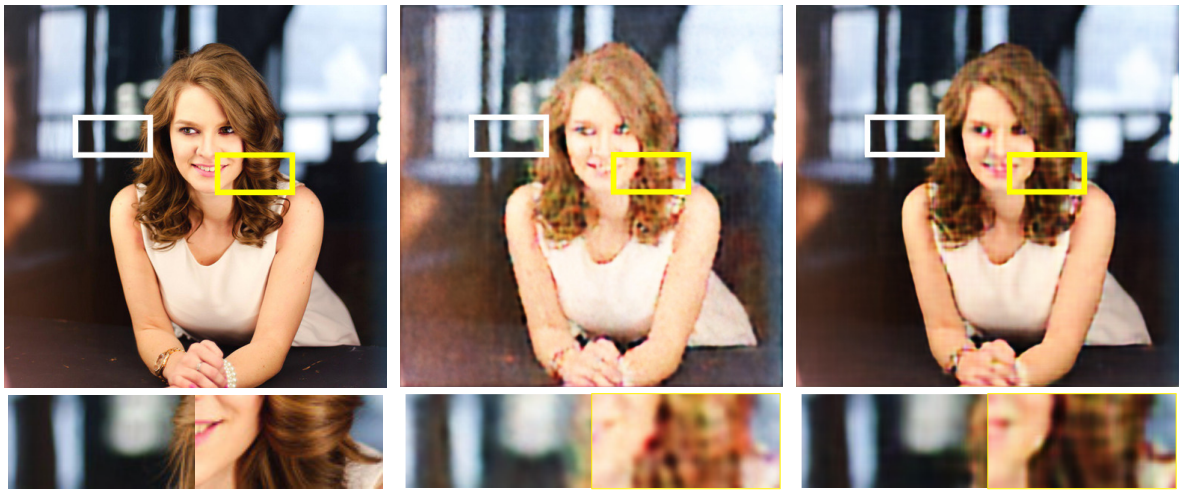
- **Denoising.** We give a visual comparison of the generated images in Fig. 12 and Fig. 13. The quantitative comparison is provided in the Table 5 and Table 6. Similar to super-resolution, we use the post-processing method proposed by Li et al. [1].

- **Flash-no flash.** In Fig. 4, we show how to control the image features provided by *flash* image and ambient illumination supplied by *no-flash* image for the *flash-no flash* based reconstruction. It is achieved using scaling factors λ_1 and λ_2. For example, a higher value to λ_1 will provide more features from *no-flash* image and a higher value to λ_2 will supply more features from *flash* image. The PSNR values are calculated using the reference image provided by Georg *et al.* [3][1].

4. Mean Square Error Loss

Here, we discuss a counter-intuitive result that the MSE loss performed better compared to the contextual loss [2] for super-resolution when working without the training data (Fig. 5). Our interpretation of this result is as follows. The contextual loss minimizes the difference between the context vector sampled from the feature space. Our restoration procedure iteratively learns the image prior by drawing a comparison between the network output and the corrupted image. Therefore, for initial iterations, the output of the network is not perceptually good because the network is yet to learn the prior of the target image. In such a scenario, the loss function which performs a direct comparison with the features of the target image (*i.e.,* MSE loss) is better than a loss function which performs comparison at the feature space (*i.e.,* contextual loss).

[1]URL: `http://hhoppe.com/proj/flash/`.



Original image MEDSF, 5%, 16.01 MED, 5%, 27.78

Figure 3: **Effects of skip connections.** The corrupted image is prepared by removing 95% pixels uniformly at random. The network with skip connection (MEDSF) achieves higher PSNR than the network without skip connections (MED).

(a) Flash image (b) No-flash image (c) (5,1), 18.5 (d) (2,1), 17.7 (e) (1,1), 16.6

Figure 4: **Flash-no flash.** This figure shows that using scaling factors (λ_1, λ_2) we can control features from no-flash and flash images in the reconstruction. (a) flash image I^F. (b) no-flash image I^{NF}. (c) $\lambda_1 = 5$ and $\lambda_2 = 1$. (d) $\lambda_1 = 2$ and $\lambda_2 = 1$. (e) $\lambda_1 = 1$ and $\lambda_2 = 1$. PSNR values for (c), (d), and (e) are also shown.

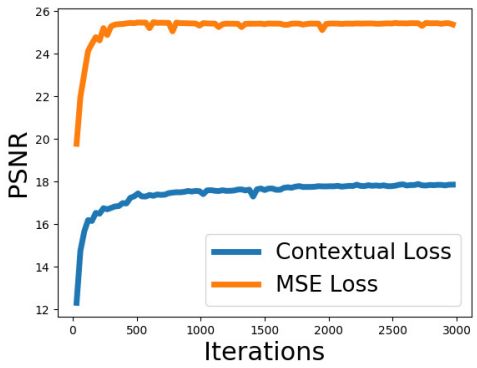


Figure 5: The comparision between contextual loss and MSE loss for $4\times$ super-resolution in a *learning-free* setting using a depth five encoder-decoder network.

5. Hyper-parameters (HP)

The learning-free image restoration using high capacity untrained networks is observed to be sensitive to the hyper-parameters (HP) (Fig. 6). This implies that a small modification of the hyper-parameters leads to a significant change in the restoration output. This could be because the network is not getting trained using a collection of images and it only depends upon the hand-crafted structure of the network and carefully chosen hyper-parameters, e.g., learning rate. In other words, the network design decisions and choice of the hyper-parameters are essential to learning the restoration task specific mapping between the network parameter space and the natural image space.

We have used the *adam* optimizer. The network parameters are randomly initialized and fitted to minimize the MSE loss between the network output and corrupted image. Our work builds upon the assumption that there exists a *learn-*

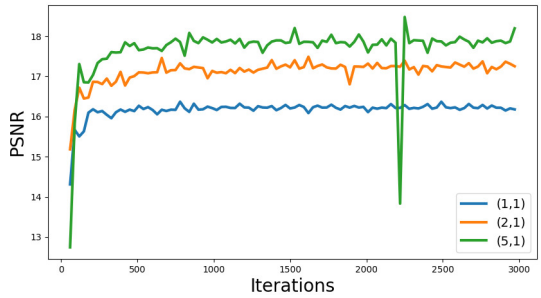


Figure 6: **Effects of Hyper-parameters (HP).** We have observed that there are sudden downfall of PSNR value during the execution. The experiments were performed for different values of (λ_1, λ_2) with the same set of HP (the experimental setup is the same). We have observed such events when learning rate or size convolutional kernel is high. However, a smaller value of learning rate would result in a slower convergence near optimum. The sensitivity of image restoration quality due to change in the learning-rate is studied in [4]). One way to control the sensitivity of the hyper-params is to minimize the total variation norm (TV norm) along with the MSE loss. TV norm is the sum of the absolute differences for neighboring pixel values of an image which measures how much noise is in the images. Minimizing the TV norm reduces the noise and provide more control over the image restoration procedure. The resulting image of the above experiment is shown in Fig. 4.

able mapping between the network parameter space and the natural image space [4].

The HP could be set for each image to get better results. However, we focused on determining HP which give satisfactory results for all the images in the dataset. The HP given in Listing 1 to Listing 6 were found using the *Tensorflow* implementation of our methods. One could also find different HP using a careful analysis.

References

[1] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

[2] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Learning to maintain natural image statistics. *arXiv preprint arXiv:1803.04626*, 2018.

[3] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. In *ACM transactions on graphics (TOG)*, volume 23, pages 664–672. ACM, 2004.

[4] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

	Barbara	Boat	House	Lena	Peppers	C.man	Couple	Finger	Hill	Man	Montage	**Avg**
DIP	22.30	25.66	21.97	25.48	23.57	26.78	29.77	29.42	26.17	21.48	22.96	**25.05**
Ours	22.30	25.85	23.14	26.18	22.42	25.78	26.78	25.54	24.80	22.83	25.21	**24.62**

Table 1: **Inpainting 90% of missing pixels (I).** The corrupted image is prepared by removing 90% pixels from the original image randomly. The performance comparison is done using PSNR values for DIP [4] and our MED. The perceptual quality comparison of the generated images for the above experiment is given in Fig. 7 and Fig. 8.

	Barbara	Boat	House	Lena	Peppers	C.man	Couple	Finger	Hill	Man	Montage	**Avg**
DIP	0.79	0.85	0.82	0.84	0.84	0.83	0.92	0.91	0.85	0.91	0.87	**0.86**
Ours	0.77	0.86	0.84	0.85	0.81	0.85	0.91	0.90	0.86	0.91	0.88	**0.86**

Table 2: **Inpainting 90% of missing pixels (II).** The corrupted image is prepared by removing 90% pixels from the original image randomly. The performance comparison is done using SSIM values for DIP [4] and our MED. The perceptual quality comparisons of the generated images for the above experiment is given in Fig. 7 and Fig. 8.

	Baboon	Barbara	Bridge	Coastguard	Comic	Face	Flowers	Foreman	Lenna	Man	Monarch	Pepper	Ppt3	Zebra	**Avg**
DIP [4]	20.35	23.85	23.21	24.24	20.99	28.39	24.64	27.66	29.23	24.93	29.07	28.20	22.91	24.39	**25.14**
Bicubic	20.28	23.58	23.06	23.99	20.25	28.92	23.83	26.01	28.35	24.41	26.24	27.22	20.43	22.82	**24.24**
MEDSF	20.03	23.02	23.15	23.92	20.29	28.58	23.87	26.12	27.82	23.18	26.70	25.43	20.42	23.35	**23.99**

Table 3: **4× image super-resolution (I).** Performance comparison between DIP [4], Bicubic and ours MEDSF on Set14 dataset (PSNR values). The perceptual quality comparison of the generated images for the above experiment is given in Fig. 9, Fig. 10, and Fig. 11.

	Baboon	Barbara	Bridge	Coastguard	Comic	Face	Flowers	Foreman	Lenna	Man	Monarch	Pepper	Ppt3	Zebra	**Avg**
DIP [4]	0.59	0.78	0.72	0.70	0.74	0.83	0.83	0.92	0.88	0.80	0.94	0.88	0.90	0.82	**0.81**
Bicubic	0.55	0.76	0.69	0.68	0.68	0.83	0.81	0.90	0.87	0.78	0.91	0.88	0.85	0.78	**0.78**
MEDSF	0.57	0.77	0.71	0.69	0.72	0.82	0.82	0.91	0.87	0.79	0.92	0.88	0.88	0.81	**0.80**

Table 4: **4× image super-resolution (II).** Performance comparison between DIP [4], Bicubic and our MEDSF on Set14 dataset (SSIM values). The perceptual quality comparison of the generated images for the above experiment is given in Fig. 9, Fig. 10, and Fig. 11.

	House	Peppers	Lena	Baboon	F16	Kodak-1	Kodak-2	Kodak-3	Kodak-12	**Avg**
CBM3D	20.71	26.41	26.69	27.11	26.09	23.18	27.06	28.39	28.41	**26.00**
DIP	18.65	21.15	21.12	22.07	21.03	21.17	21.14	22.94	23.14	**21.37**
Ours	18.39	23.39	21.72	21.16	20.40	18.16	19.76	21.27	24.34	**20.95**

Table 5: **Denoising (I).** A detailed comparision for Denoising with strength $\sigma = 100$ using PSNR values. The perceptual quality comparison of the generated images for the above experiment is given in Fig. 12 and Fig. 13.

	House	Peppers	Lena	Baboon	F16	Kodak-1	Kodak-2	Kodak-3	Kodak-12	**Avg**
CBM3D	0.60	0.879	0.869	0.854	0.832	0.69	0.827	0.876	0.867	**0.810**
DIP	0.479	0.824	0.691	0.777	0.735	0.609	0.709	0.802	0.822	**0.716**
Ours	0.496	0.826	0.688	0.80	0.747	0.572	0.732	0.823	0.851	**0.725**

Table 6: **Denoising (II).** A detailed comparision for Denoising with strength $\sigma = 100$ using SSIM values. The perceptual quality comparison of the generated images for the above experiment is given in Fig. 12 and Fig. 13.

```
λ1, λ2, λ3=1,0,1
LR, η=0.0001,10
ch^e, ch^d=(64,128,128,128,128), (128,128,128,128,64)
f^ed, f^s=3,3
```

Listing 1: Hyper-parameters for super-resolution.

```
λ1, λ2, λ3=1,1,1
LR, η=0.0001,25
ch^e=ch^d=128,128,128,128,128
f^ed, f^s=3,1
```

Listing 2: Hyper-parameters for denoising.

```
λ1, λ2, λ3=1,1,1
LR, η=0.0005,10
ch^e, ch^d=(16,32,64,128,128), (128,128,64,32,16)
f^ed, f^s=3,3
```

Listing 3: Hyper-parameters for region inpainting, object removal, and text removal.

```
λ1, λ2, λ3=1,1,1
LR, η=0.0001,50
ch^e, ch^d=(32,32,32,32,32), (96,96,96,96,96)
f^ed, f^s=7,7
```

Listing 4: Hyper-parameters for restoration of BW image from 90% pixels.

```
λ1, λ2, λ3=1,1,1
LR, η=0.0001,5
ch^e, ch^d=(32,32,32,32,32), (64,80,96,112,128)
f^ed, f^s=5,7
```

Listing 5: Hyper-parameters for restoration of RGB image.

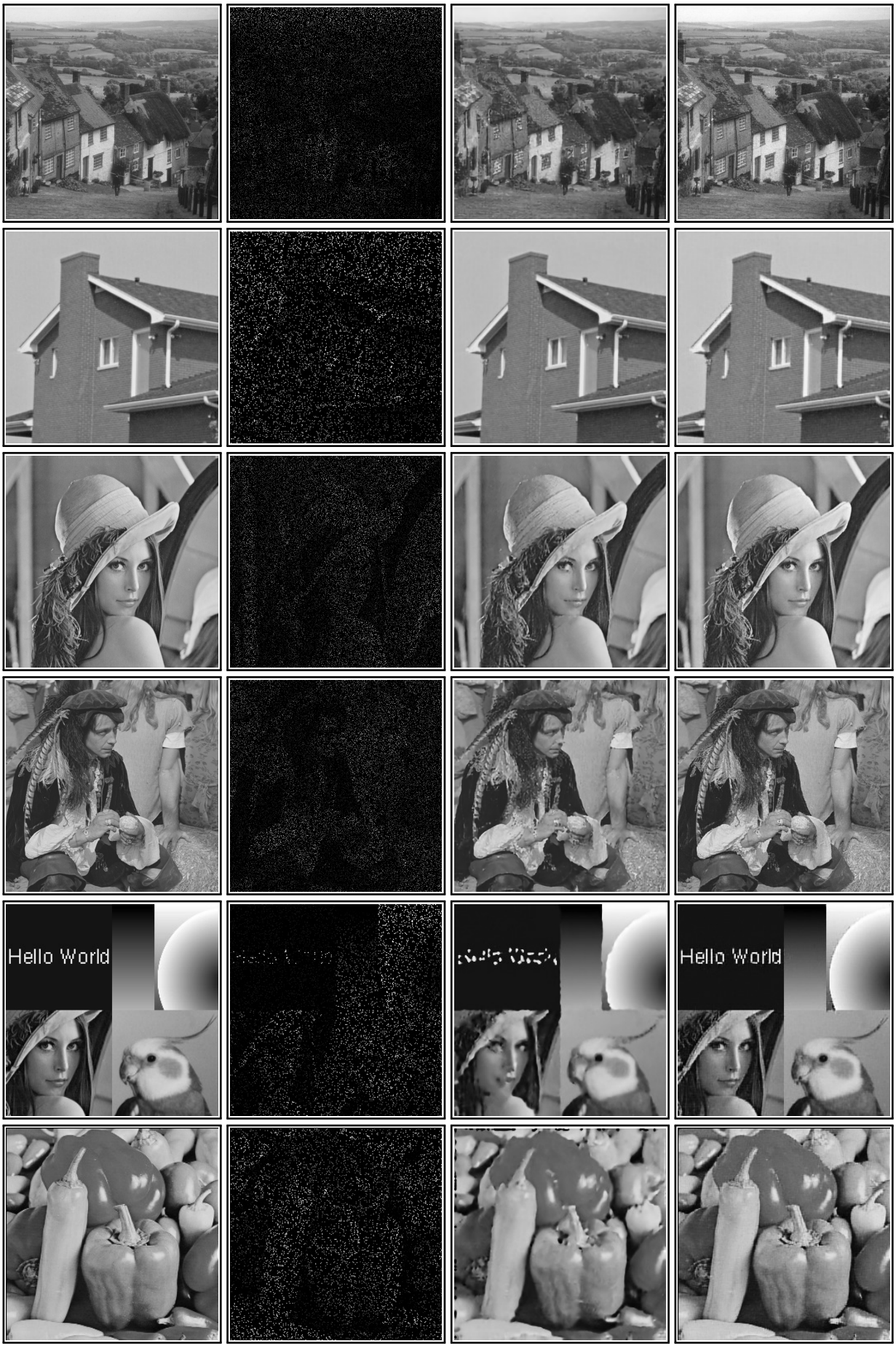
```
λ1, λ2=5,1
LR, η=0.0001,10
ch^e, ch^d=(32,64,128,128,128), (128,128,128,64,32)
f^ed, f^s=3,3
```

Listing 6: Hyper-parameters for flash-no flash.

Hyper-parameters listing. The hyper-parameters in the Listing 1 to Listing 6 are as follows. λ_1, λ_2 and λ_3 are the coefficients of the loss function (Eq.4, Eq.5, Eq.6, and Eq.7 of the manuscript). LR is the learning rate. η is the noise strength to generate network input. ch^e and ch^d are the channels of encoder and decoder. f^{ed} is the kernel size of the encoder-decoder layers and f^s is the kernel size of the layer for skip connection.

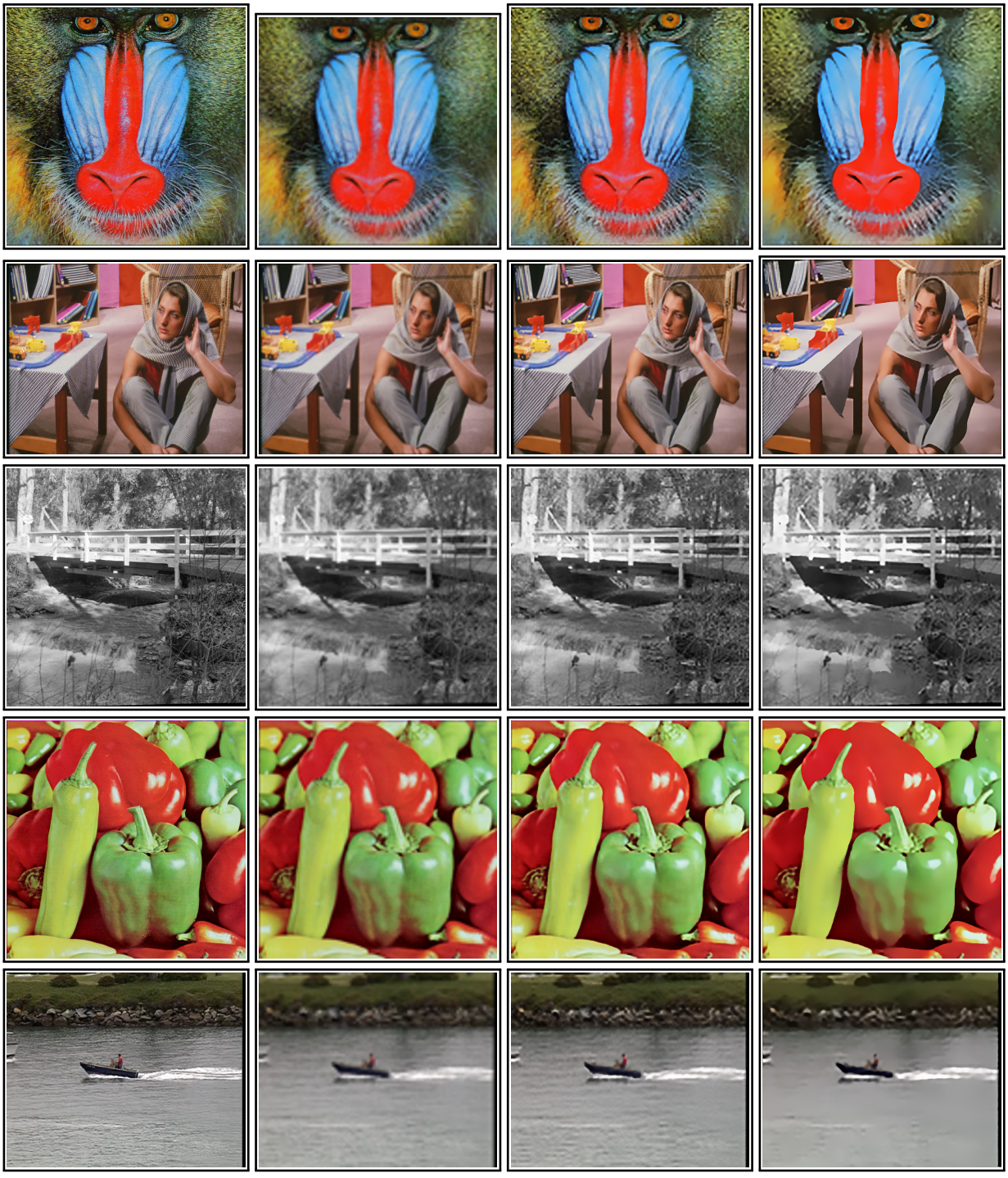
(a) Original image (b) Corrupted image (c) DIP [4] (d) MED (ours)

Figure 7: Restoration from 90% missing pixels *part-1*.



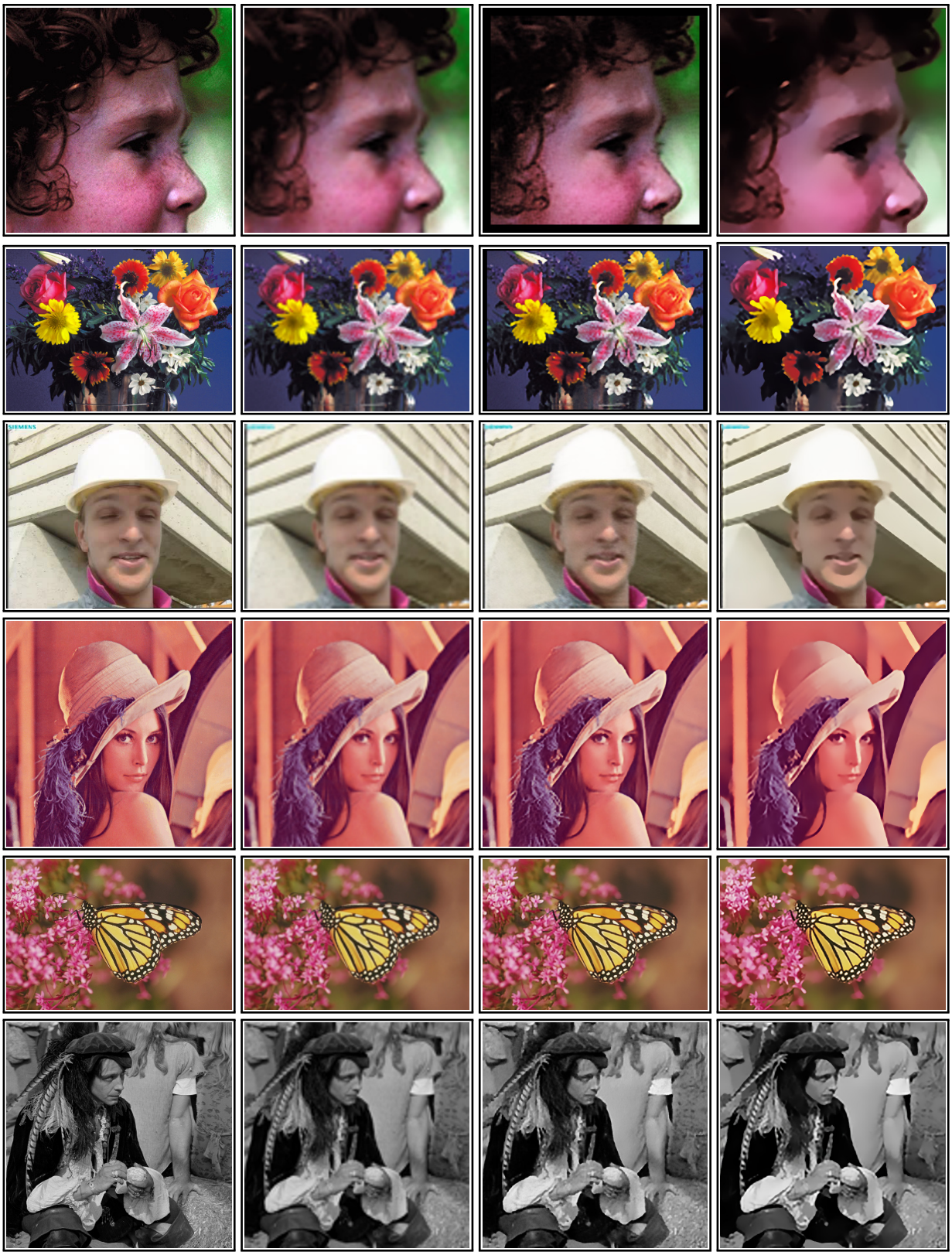
(a) Original image (b) Corrupted image (c) DIP [4] (d) MED (ours)

Figure 8: Restoration from 90% missing pixels *part-2*.



(a) HR image (b) Bicubic (c) DIP (d) MEDSF

Figure 9: $4\times$ Super-resolution on Set14 dataset *part-1*.

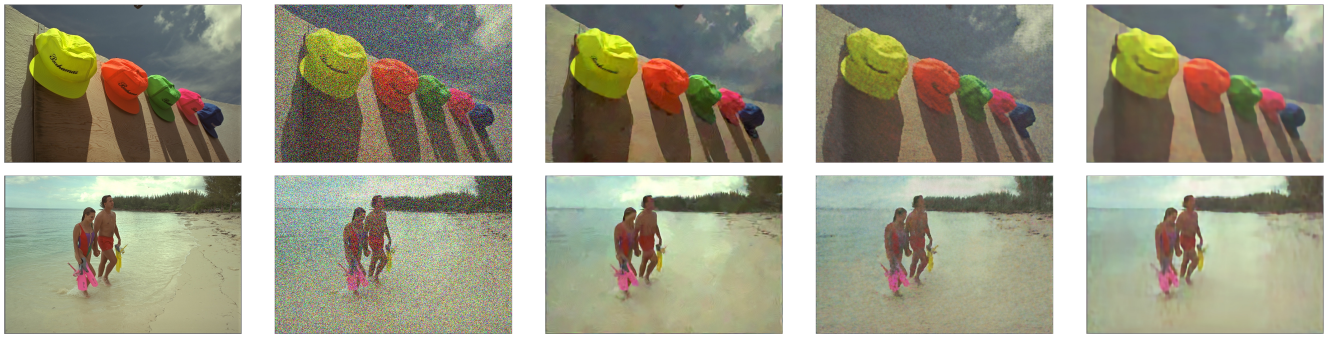


(a) HR image (b) Bicubic (c) DIP (d) MEDSF

Figure 10: $4\times$ Super-resolution on Set14 dataset *part-2*.

(a) HR image (b) Bicubic (c) DIP (d) MEDSF

Figure 11: $4\times$ Super-resolution the Set14 dataset *part-3*.



Original Image Noisy Image CBM3D DIP Ours

Figure 12: **Denoising (I).** A comparison between CBM3D, DIP [4], and our MEDSF for noise strength of $\sigma = 100$.

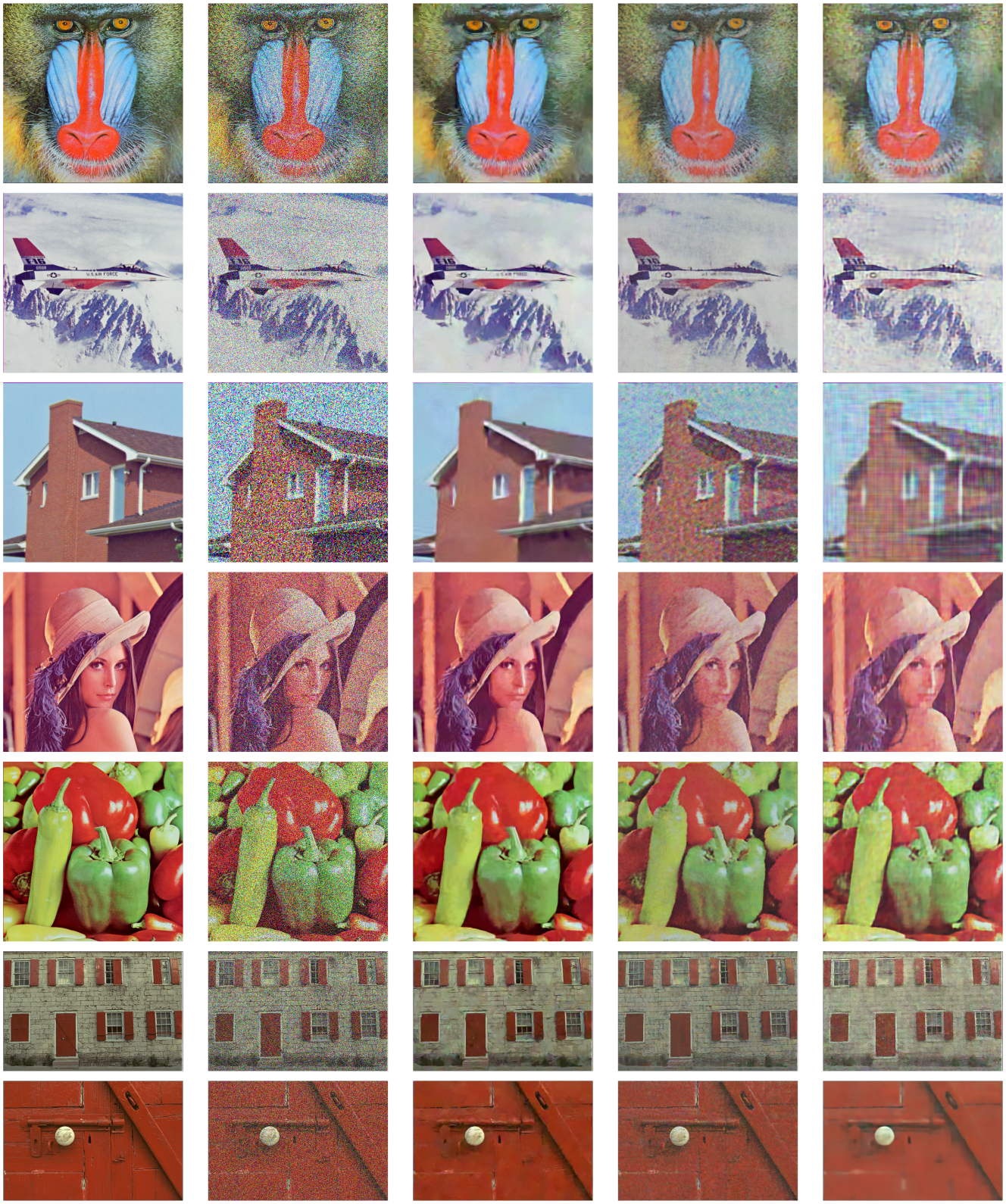


Figure 13: **Denoising (II).** A comparison between CBM3D, DIP [4], and our MEDSF with noise strength of $\sigma = 100$.