Supplementary Material—Bitstream-Corrupted JPEG Images are Restorable: Two-stage Compensation and Alignment Framework for Image Restoration

Wenyang Liu[1], Yi Wang[1*], Kim-Hui Yap[1*] and Lap-Pui Chau[2]
[1]*School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore*
[2]*Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong*
{wenyang001, wang1241}@e.ntu.edu.sg, ekhyap@ntu.edu.sg, lap-pui.chau@polyu.edu.hk

This document supplies more detailed information and visual comparisons of our method. We first introduce the employed pix2pix network and loss function in Section 3.3 of the manuscript. Then, we show more visual results in different experimental conditions.

1. Pix2pix Network

Pix2pix [3–5] networks were widely used to achieve an image-to-image translation. In this paper, the guided compensation and alignment (GCA) stage casts this image restoration problem as a task of image-to-image translation under the guidance of the extracted thumbnail. The employed pix2pix network can be regarded as a coarse-guided restoration network in the GCA. The architecture of the pix2pix network is shown in Fig. 1 which consists of multi-resolution generators (i.e., $G1$ and $G2$) and multi-resolution discriminators (i.e., $D1$ and $D2$). The aim is to generate a coarsely color-compensated and aligned image by fusing the self-compensated image (from the self-compensation and alignment (SCA) stage) and the bicubic upsampled low-resolution thumbnail (from the JPEG file's header).

Multi-resolution generators. The multi-resolution generators consist of two generators: a global generator $G1$ and a local generator $G2$ as shown in Fig. 1. Both generators consist of a convolutional downsampling front-end, three residual blocks, and a transposed convolutional upsampling back-end, where the details of each component are shown at the bottom of Fig. 1. The input of $G2$ is the concatenated images of the self-compensated image and the upsampled thumbnail, and the input of $G1$ is 2x downsampled images from the input of $G2$. The corresponding output of $G1$ is element-wise added with the feature maps from the downsampling front-end of $G2$. Ref. [5] proved that the multi-resolution generator structure is able to effectively integrate the learned global and local information from the image inputs to generate high-resolution image synthesis. In our paper, the global information of the upsampled thumbnail, i.e., color and structure information, can coarsely and implicitly guide the compensation and alignment of the self-compensated image that suffers from color cast and block shifts. The final high-resolution image is restored in structure and color except for realistic details, which will be sent to a refine-guided Laplacian pyramid fusion network to refine details (see Figure 2 of the manuscript).

Multi-resolution discriminators. The multi-resolution discriminators contains two discriminators $D1$ and $D2$, which have an identical architecture. The real and synthesized high-resolution images are downsampled by a factor of 2, such that the two-scale real and synthesized images are employed to train $D1$ (with low scale) and $D2$ (with high scale), respectively. The multi-resolution discriminators encourage the generators to produce both globally and locally consistent images with different scales of the receptive field.

2. Loss Function

Here we introduce the adversarial loss L_A, the feature matching loss L_{FM}, and the perceptual loss L_{VGG} in Eq. (6) of the manuscript.

Adversarial loss. The adversarial loss is defined as a multi-task learning loss:

$$L_A = \min_{G} \max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{\mathrm{GAN}}\left(G, D_k\right) \tag{1}$$

where $\mathcal{L}_{\mathrm{GAN}}\left(G, D_k\right)$ is the adversarial loss of the k-the discriminator of D_k, expressed as:

$$\mathcal{L}_{\mathrm{GAN}}\left(G, D_k\right) = E_{(X)}\log D_k(X) + E_{(X)}\log D_k(G(X_s, T)) \tag{2}$$

where X, X_s, and T denote error-free real images, self-compensated images, and the extracted thumbnail. $G(X_s, T)$ represents the output by the pix2pix's generator.

Feature matching loss. Feature matching loss is defined as the matching similarity of features in multiple layers of the discriminator between the error-free real images and the generated images, expressed as:

*Corresponding authors

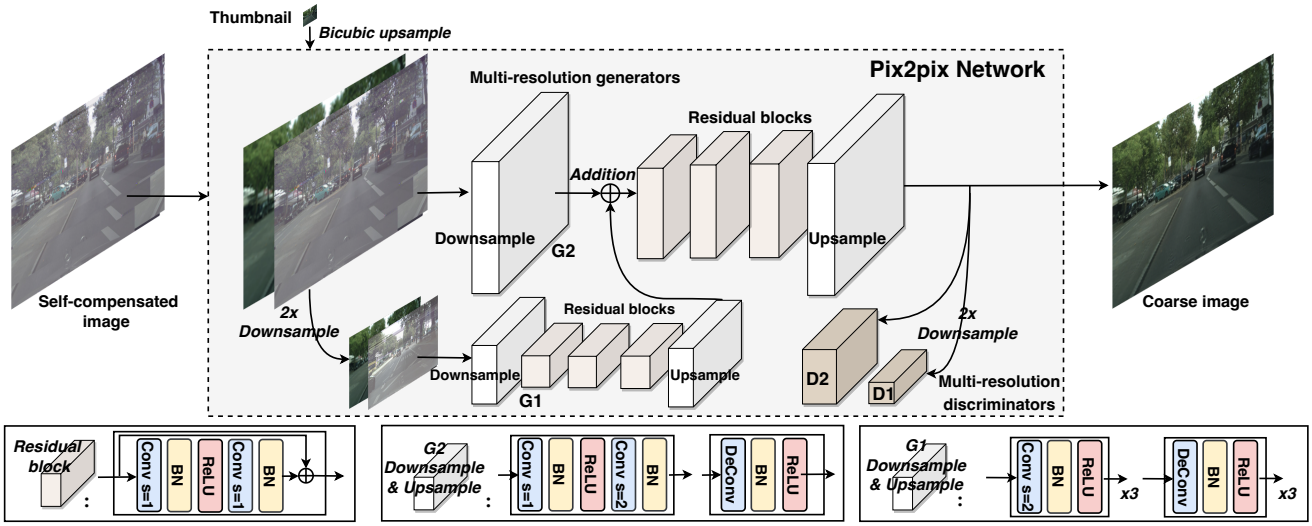

Figure 1. Architecture of the pix2pix network. The input consists of two images: the self-compensated image (from the self-compensation and alignment (SCA) stage) and the extracted thumbnail (from the JPEG file's header). The output is the coarse image, which is guided by the thumbnail. The coarse image is then sent to a refine-guided Laplacian pyramid fusion network to refine details (see Figure 2 of the manuscript). The details of each component are shown at the bottom of the figure. s means the stride of the convolution.

$$L_{FM} = \min_{G} \sum_{k=1,2} \mathcal{L}_{FM}\left(G, D_k\right) \tag{3}$$

where $\mathcal{L}_{FM}\left(G, D_k\right)$ is the feature matching loss with the k-th discriminator D_k, expressed as:

$$L_{FM} = E_{(X)} \sum_{i=1}^{L} \frac{1}{N_i} \left[\| D_k^{(i)}(X) - D_k^{(i)}(G(X_s, T)) \|_1 \right] \tag{4}$$

where L is the total number of layers used for feature extraction, N_i denotes the number of elements in the i-th layer, $D_k^{(i)}$ is the extracted feature maps of the i-th layer in D_k.

Perceptual loss. Perceptual loss is used to measure the high-level differences, e.g., content and style discrepancies, between images. It is defined by the differences between pre-trained VGGNet extracted feature maps, expressed as:

$$L_{VGG}^{\phi,i}(\hat{X}, X) = \frac{1}{C_i H_i W_i} \| \phi_i(\hat{X}) - \phi_i(X) \|_1 \tag{5}$$

where $\hat{X}$ is the final restored image of the network, H_i, W_i, and C_i are the height, width, and channel of the i-th layer in VGGNet. $\phi_i()$ denotes the output feature map of the i-th layer.

3. More Visual Results

Comparison of SCA with/without alignment. Fig. 2 shows a visual comparison of SCA with/ without block alignment processing. As we can see from the figure, although the proposed block alignment processing does not make the processed image fully aligned, this processing delivers better-aligned results than the SCA without the alignment, which proves its effectiveness.

Comparison of coarse and refined images. Fig. 3 shows a visual comparison of the coarse images by the pix2pix network and the refined images by the Laplacian fusion network. We can observe that the coarse images are not fully aligned and have some artifacts. After the refine-guided Laplacian fusion network processing, these artifacts are removed, and more texture details are generated, which proves the effectiveness of the proposed Laplacian fusion network.

Comparison of other methods. Fig. 4 shows a 1k-resolution visual comparison of our robust decoder, SCA, and GCA methods with standard decoder and the EPDN [4] method. Figs. 5 and 6 show a 2k-resolution visual comparison. We can see that our method consistently has superior results over other methods in different-resolution image restoration.

Generalization of varying BERs of images. Fig. 7 shows more visual comparisons of the standard decoder, our SCA, and our SCA+GCA (two-stage model) on the AFHQ [1] dataset with different bit error rates (BERs). The training is under the BER=10^{-5}, and the testing is under various BERs. These results demonstrate the superior generalization ability of our two-stage method in handling varying degrees of BERs of the JPEG file without retraining.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2, 6

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for

semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4, 5, 6

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 1

[4] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 5, 6

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 1

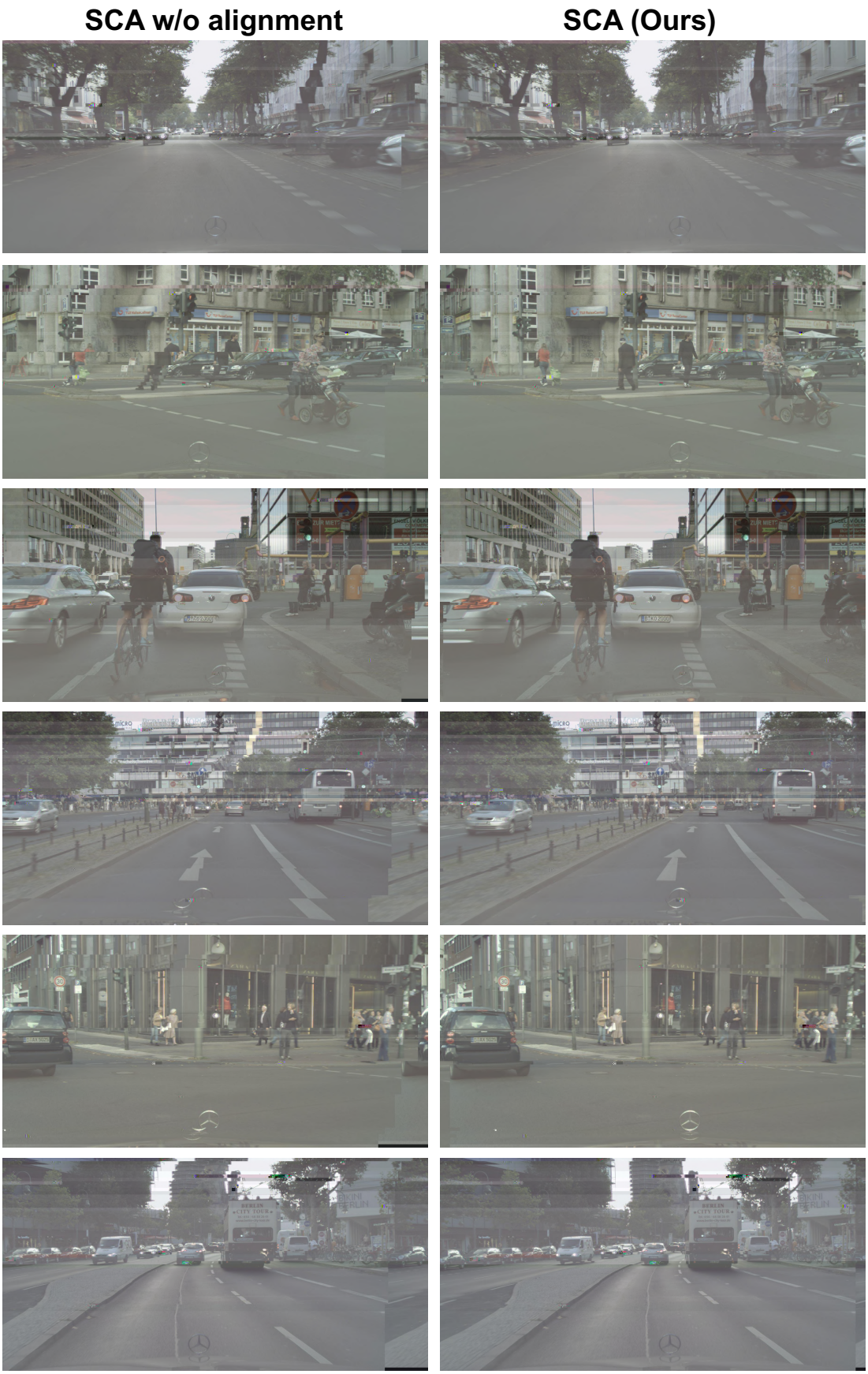

Figure 2. Visual comparison of the SCA with/without the block alignment processing on the 2k-resolution Cityscape [2] dataset.

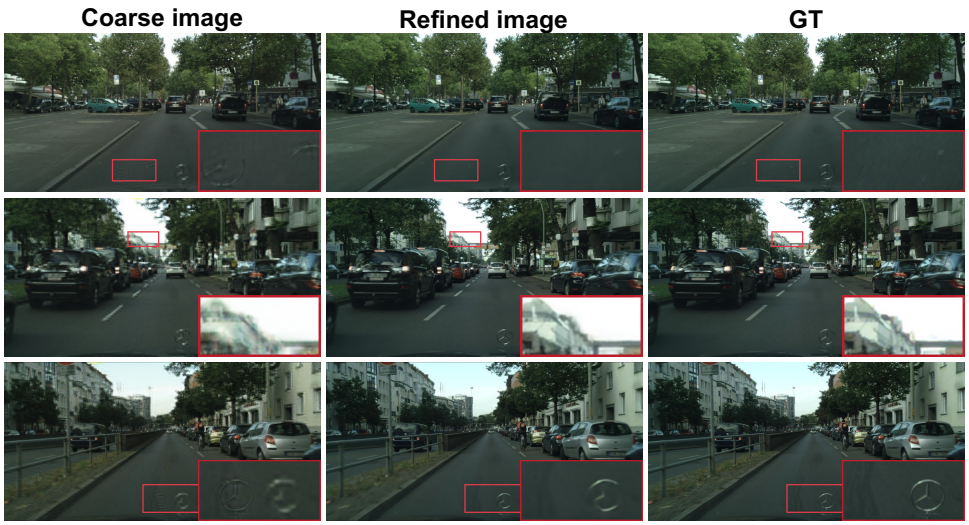

Figure 3. Visual comparison between coarse images obtained by the pix2pix network and refined images obtained by the Laplacian fusion network on 1k-resolution Cityscape [2] dataset.

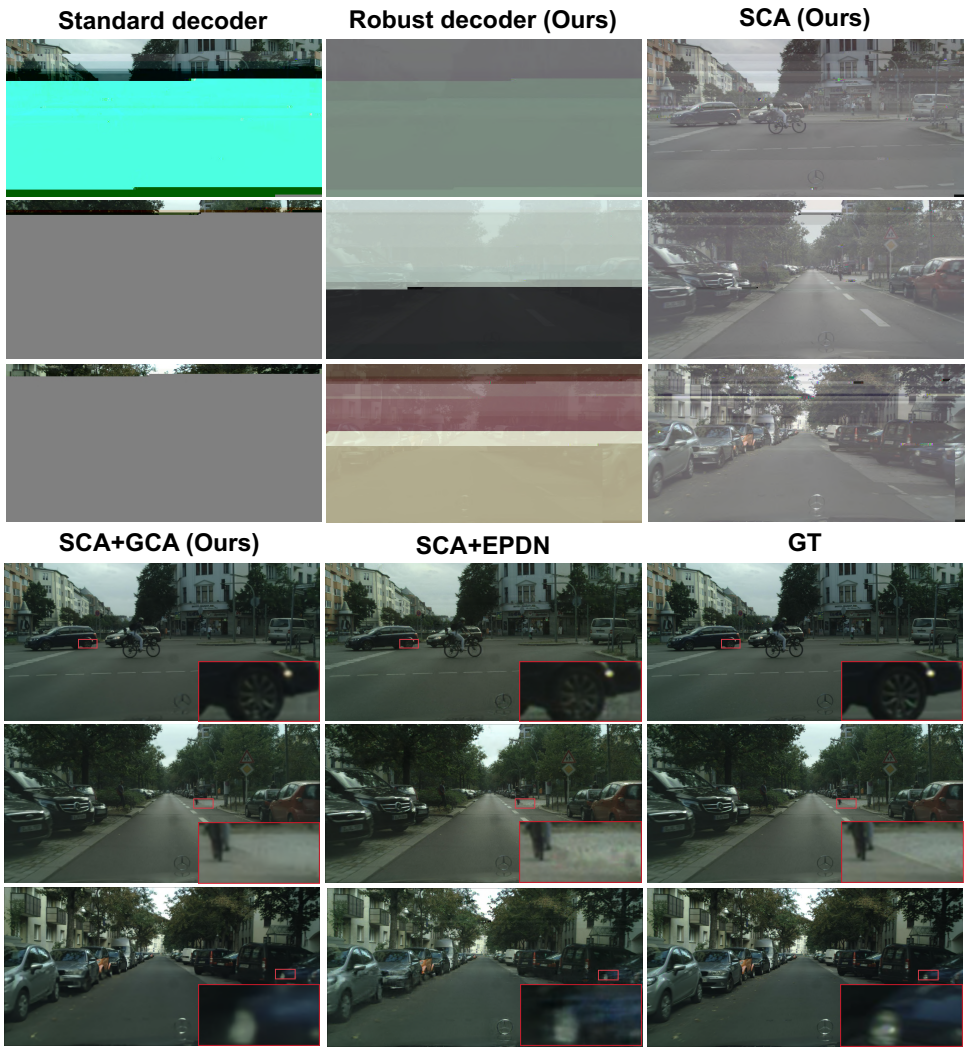

Figure 4. Visual comparison of the proposed robust decoder, SCA, and GCA methods with the standard decoder and the EPDN [4] method on 1k-resolution Cityscape [2] dataset.

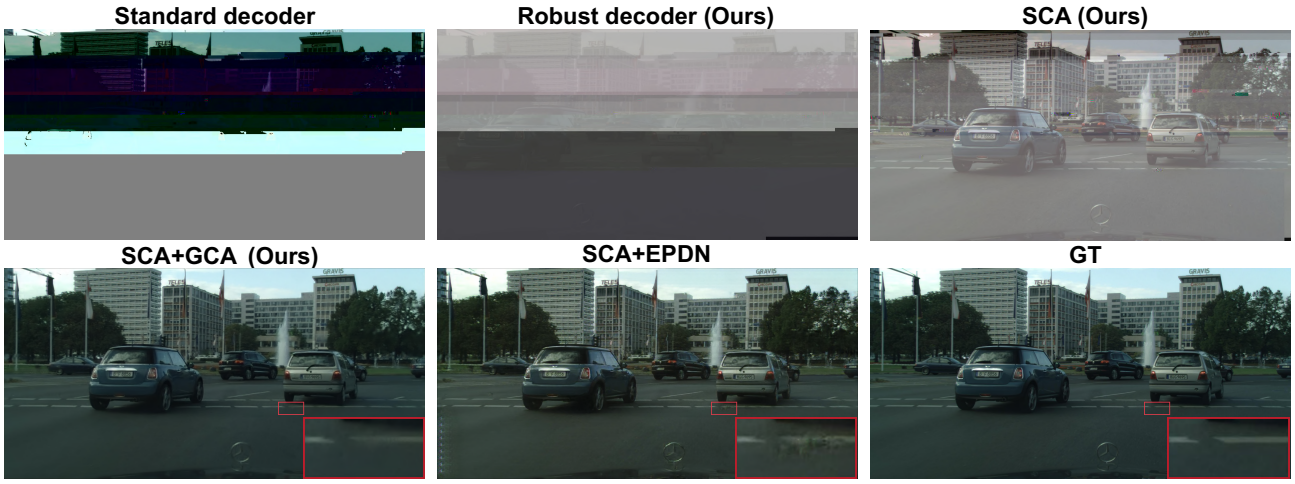

Figure 5. Visual comparison of the proposed robust decoder, SCA, and GCA methods with the standard decoder and the EPDN [4] method on 2k-resolution Cityscape [2] dataset.

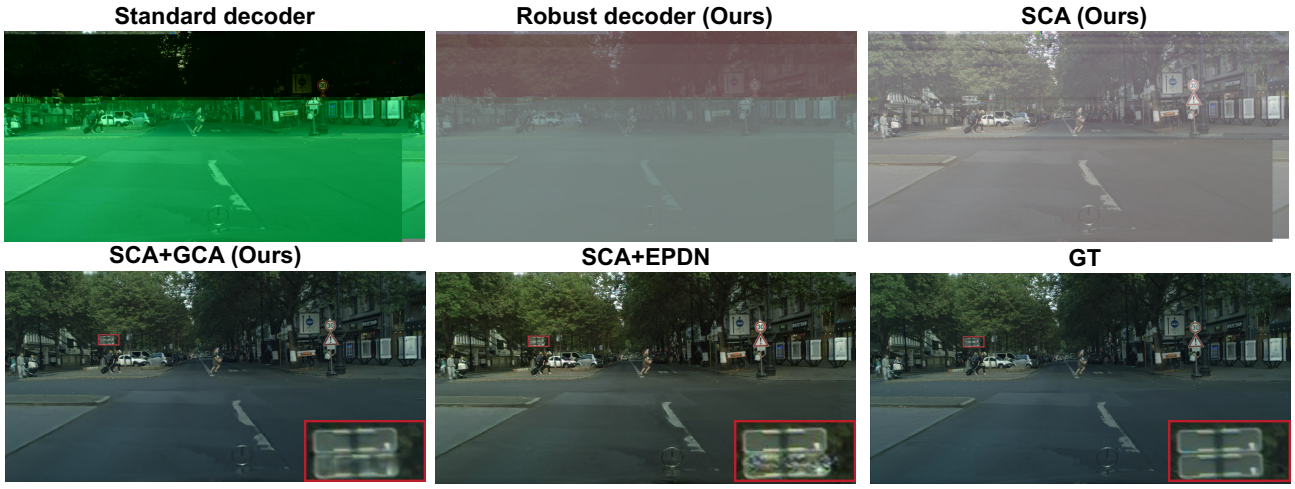

Figure 6. Visual comparison of the proposed robust decoder, SCA, and GCA methods with the standard decoder and the EPDN [4] method on 2k-resolution Cityscape [2] dataset.

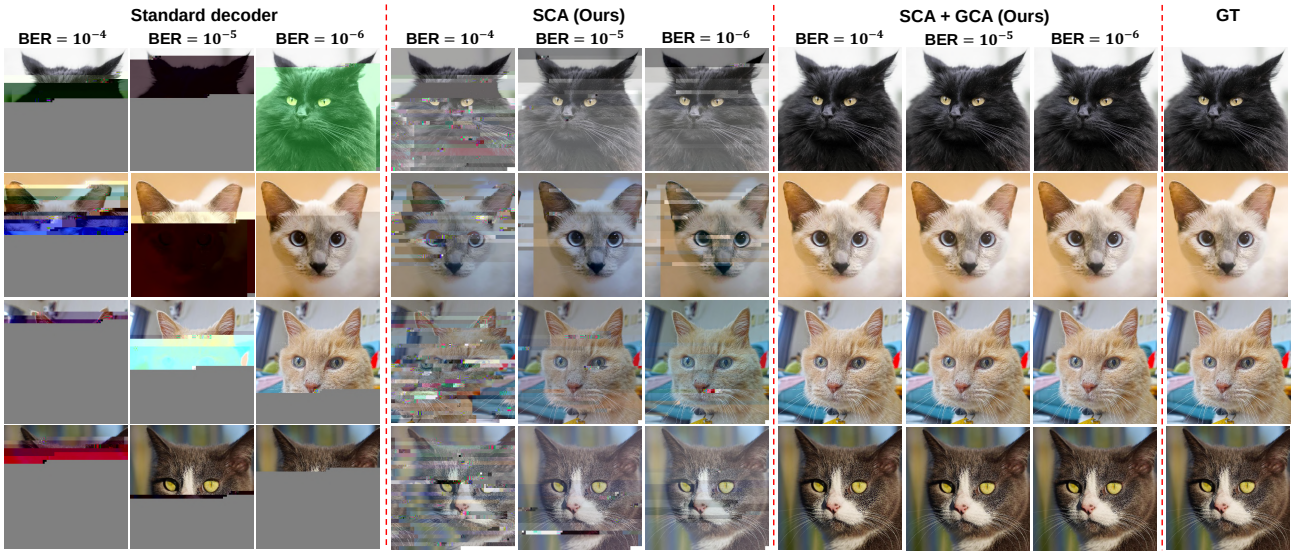

Figure 7. Visual comparison of the standard decoder's results (Left) with our SCA's (middle) and SCA+GCA's results (Right) on various BERs in the AFHQ [1] dataset.