Supplementary Material
Addressing Image Authenticity When Cameras Use Generative AI

Umar Masud[1†] um71000@gmail.com
Abhijith Punnappurath[2] abhijith.p@samsung.com
Luxi Zhao[2‡] lucyzhao.zlx@gmail.com
David B. Lindell[1] lindell@cs.toronto.edu
Michael S. Brown[2] michael.b1@samsung.com
[1]University of Toronto [2]AI Center–Toronto, Samsung Electronics.

This supplementary material contains additional results and experimental details that could not be included in the main paper due to space constraints.

S1. Implementation details of comparison methods

As mentioned in Section 4.2 of our main paper, we selected the hyper-parameters of the hashgrid [3] method such that their embedding and MLP are similar in size to our encoder and MLP, respectively. In particular, we chose the number of levels $L = 16$, the number of feature dimensions per entry $F = 4$, and the maximum entries per level (hash table size) $T = 2^9$. The remaining hyper-parameters are the same as recommended by the authors. Under these settings, the encoder of the hashgrid [3] method has roughly 32 K parameters (128 KB). The MLP model is chosen to be similar to our method with two hidden layers and 64 neurons per layer, making the overall hashgrid model size 184 KB.

For the blind image-to-image translation experiment, we use a large NAFNet [2] model with 36 blocks, same as the default configuration proposed by the authors. We change the default width from 32 to 24 to reduce overfitting. We use the Adam optimizer with β_1 = 0.9, β_2 = 0.999, weight decay 0, and train for 200 epochs with a learning rate of $1e^{-4}$. We train on patches of size 64×64 pixels using an l_1 loss and a batch size of 128. Because this is a no-metadata approach, we directly apply the trained models on the test set during inference and do not perform further finetuning on the test pairs.

S2. Additional qualitative results

Fig. S1 shows additional qualitative results of our method on the DIV2K [1] dataset described in Section 4.1 of our main paper. In the first example, the strap of the bag, that is missing in the hallucinated image, is correctly recovered by our method. In the second image of an aerial scene, the appearance of the crops is altered due to hallucinations. Our method is able to accurately recover the authentic image. Fig. S2 shows a text SR example of our method applied to a real image.

S3. Alternate metadata schemes

We propose to store the parameters of our encoder and our MLP decoder as metadata. Here, we compare this against directly saving the compressed *residual* between the hallucinated and authentic image pair as metadata. In particular, we applied lossy JPEG compression to the residual image with the quality factor (QF) chosen to yield a metadata overhead of around 180 KB, that is similar in size to our metadata, or higher. We perform this experiment on the DIV2K [1] dataset. In the main paper, we had down-

Method	Metadata size (KB)	PSNR (dB)
JPEG QF = 4	181	27.44
JPEG QF = 15	277	32.45
JPEG QF = 30	416	34.67
With binary mask	401	30.28
Error sampling	180	34.27
Ours	180	**35.12**

Table S1. Comparison of our method against residual JPEG compression at different quality factors, saving a binary mask of hallucinated pixels as additional metadata, and finetuning based on weighted error sampling.

[†]Work done during an internship at Samsung.
[‡]Work done while with Samsung.

Figure S1. Qualitative results for natural image super-resolution on the DIV2K [1] dataset. Inset shows zoomed-in region.

sampled the images by a factor of four and then upsampled them back up by a factor of four, adhering to the protocol set by [1]. For this comparison, we first downsample the images by a factor of two before super-resolving them by a factor of four such that the output images are approximately 12 MP resolution, which matches the resolution of most current smartphone cameras. We generate a paired dataset of hallucinated and authentic images using the RealESRGAN network [5], following the same procedure described in Section 4.1 of our main paper.

Table S1 shows the result of residual JPEG compression at different quality factors against our approach. Our results are obtained using the same pretrained DIV2K encoder and MLP decoder from our experiments in the main paper. It can be seen that our approach performs significantly better at comparable metadata size (QF = 4). Even at more than double the metadata size (QF = 30), our result is more accurate. Also note that our metadata overhead is independent of image resolution, while the memory footprint of the residual image increases with image size.

We also compare against a mask-based metadata approach suggested by [4]. Following [4], we threshold the residual between the hallucinated and authentic image pair to obtain a binary mask. Pixels that are flagged by the mask are considered hallucinated while the remaining pixels are authentic. This binary mask is saved as metadata, along with the encoder and MLP model weights. At inference, we finetune the MLP only on those pixels that are marked as hallucinated in the mask. The results are shown in Table S1. We see that this approach is not as accurate as our method. One drawback of this technique is that the mask adds considerable metadata overhead. As a proxy for the binary mask to flag fake pixels, we compare against a variant of our proposed method where we adapt the finetuning to focus on and correct those pixels that are more likely to be hallucinated. In particular, instead of randomly sampling pixels, we use a weighted sampling based on the per-pixel error map between the hallucinated and authentic image i.e., pixels with a higher error are sampled more frequently. However, we did not notice a significant improvement in the aggregate PSNR, as seen from the results in Table S1. Moreover, compared to random sampling, weighted error sampling is computationally more expensive and increases the time needed for finetuning.

S4. Modality-specific encoder

As mentioned in Section 4.3 of our main paper, we pretrain a separate modality-specific encoder jointly with the MLP for each task, such as natural image SR, text SR, and low-light image enhancement. In Table S2, we compare

Encoder	PSNR (dB) on various datasets		
	DIV2K	**MARCONet**	**LOL**
Modality agnostic	31.54	29.91	36.32
Ours – Modality specific	**32.96**	**31.26**	**36.34**

Table S2. Comparison between a modality-agnostic encoder and our modality-specific encoder.

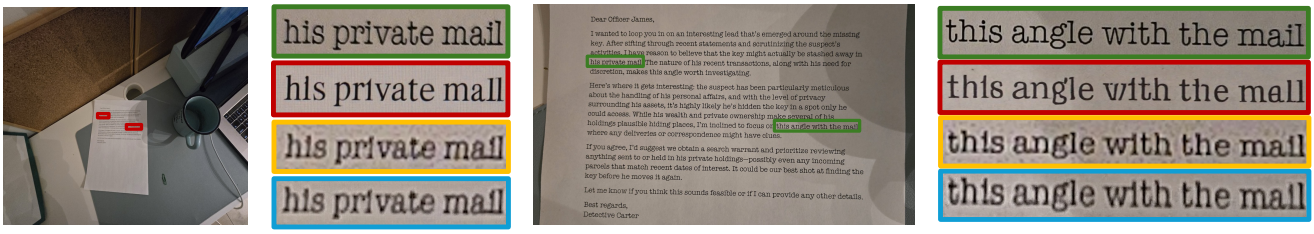

Figure S2. Example of text SR on real images. In the first image, far-away text is enhanced by the AI ISP causing characters to flip (mail to mall). The second image of the text region captured from a closer distance shows the actual text with no hallucinations (mail). Such a reference image would not be available in practice. The authentic image (yellow) before hallucinatory text enhancement shows that the third letter of the last word is an 'i', which the AI model mistakenly super-resolved to an 'l'. Our recovered authentic image (blue) can reverse this hallucination.

against a generic encoder that is pretrained with the MLP on a mix of the training splits from all three datasets described in Section 4.1. This modality-agnostic encoder is then evaluated on the test images from each dataset with the MLP decoder finetuning performed in an identical manner to our approach. Our results in the last row are reproduced from Table 1 of our main paper for ease of comparison. It can be observed that the modality-agnostic encoder performs worse than our modality-specific encoder with up to 1.5 dB drop in performance. This suggests that we need modality-specific pretraining to get the best performance.

We propose to save the encoder parameters as well as the MLP decoder parameters, with a combined size of 180 KB, as metadata along with each captured image. This makes the metadata self-contained and has the advantage that post-capture recovery requires just the image and its metadata. Alternately, since the encoder is frozen for a particular modality, only the MLP decoder weights, with a size of 53 KB, can be saved as metadata along with each image. However, this necessitates that the encoder's modality is known and its specific weights are available at recovery time through some other mechanism, such as retrieving it from a secure online repository. This may be less preferred in practice.

Description	Configuration	PSNR (dB)
Different hidden layers of MLP	MLP=64x1, 162 KB	32.73
	MLP=64x3, 196 KB	32.95
Only latent	K=64, no (x, y)	32.75
Random batch sampled every @iteration	Sampling @100	32.78
	Sampling @50	32.85
	Sampling @25	32.91
	Ours – Sampling @1	**32.96**

Table S3. Additional ablations of our model architecture and training settings.

S5. Additional ablations

Table S3 shows additional ablations of our model architecture and training settings. All experiments are performed on the DIV2K [1] dataset described in Section 4.1 of our main paper. Results reported are after finetuning. First, we vary the number of hidden layers of our MLP architecture. Increasing the number of layers to three from our choice of two increases the metadata size, but did not yield any notable difference in performance. We also tried not concatenating the (x, y) coordinates as input to the MLP and only decoding from the K-dimensional latent feature from the encoder. Further, we show some ablations on the frequency of drawing random samples during finetuning, where @p denotes that a random batch is drawn only every p iterations. Sampling less frequently could speed up the finetuning time, however, as shown, it comes with decreased performance. Our proposed approach of sampling at every iteration, shown in the last row, gives the best results.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 1, 2, 3

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 1

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4): 1–15, 2022. 1

[4] Abhijith Punnappurath, Luxi Zhao, Abdelrahman Abdelhamed, and Michael S. Brown. Advocating pixel-level authentication of camera-captured images. *IEEE Access*, 12: 45839–45846, 2024. 2

[5] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021. 2