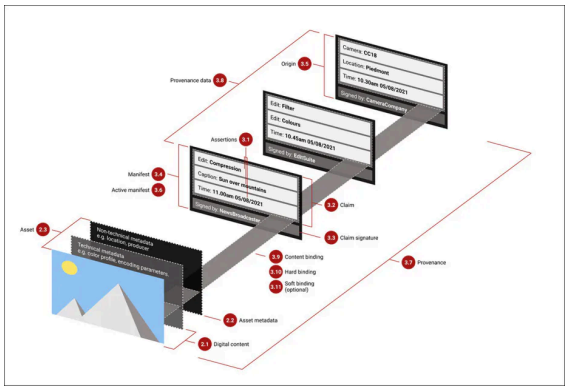
Restoring What Your Camera Captured Before AI Changed It

Originally published at https://www.unite.ai/restoring-what-your-camera-captured-before-ai-changed-it/

How can you protect the sanctity of a raw photograph from AI interference when it's already been automatically run through AI inside the camera? New research seeks to restore 'true' sensor data – also with AI!

The increase in the authenticity of AI images over the last year or so has caused many groups and individuals to rally against the ensuing erosion of trust in photography.

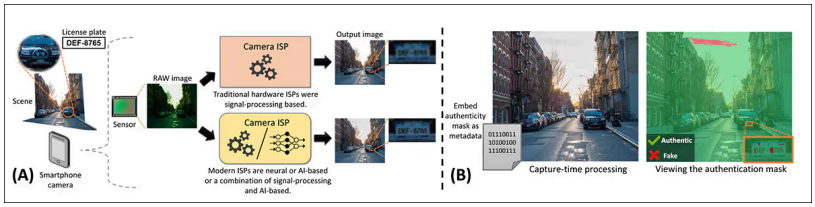
In the same period, the Coalition for Content Provenance and Authenticity (C2PA) has attempted to diffuse a semi-cryptographic standard that attaches metadata-based provenance information to an image, from the very instant that it is captured by a supported camera or device, hoping to unmask any subsequent use of generative AI on these 'original' pictures:



Schema of provenance in the C2PA system, where metadata written at the instance of capture can be added to like a diary, allowing for customary adjustments such as brightness and contrast, but recording major adjustments, so that a heavily AI-altered image will show up as such in media outlets that support this system. Source

Adoption of the standard has not been as widespread as the coalition had hoped, and currently only 14 cameras support in-camera imprinting of the authenticity information.

What's interesting about the C2PA's idea of giving a photo a 'passport' as soon as it comes into existence, is that by that time *it may already be too late* – because camera manufacturers now routinely bake AI-processing into the very creation of the image:



From the 2024 paper 'Advocating Pixel-Level Authentication of Camera-Captured Images': an illustration of how modern camera pipelines introduce hallucinated c pixel-level authentication metadata exposes it. In (A), a smartphone sensor image is processed by the ISP, where AI modules can invent details during digital zoor producing realistic images with errors such as misread license plate digits. In (B), an authentication mask is embedded as metadata and later overlaid to reveal no users to distinguish original data from AI-altered pixels. Source

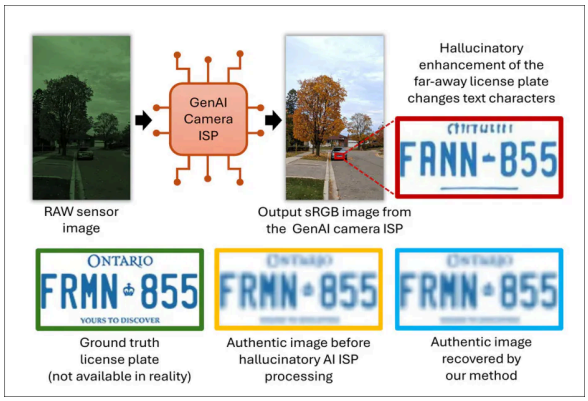
In fact, this AI 'interference' in the capture of raw data from the camera's sensor could eventually even become the governing process.

This kind of post-processing is not the same as the current trend for altering photos in-camera, wherein a phone app or a camera app allows the user to re-think a photo at leisure before it is even downloaded from the device.

Rather, the processing happens in a 'black box' routine in the camera's Image Signal Processor (ISP), usually in a proprietary runtime that does not expose or make available the raw sensor data (and consider that the supposedly 'pure' camera RAW format is not all that 'raw').

Therefore, by the time you're able to see the photo at all, it may have been subject to AI-aided enhancements such as low-light enhancement, upscaling, or even moon replacement.

In many cases, this can lead to inaccurate reconstructions, for instance of text, that might invalidate the use of such an image for evidence, since a true 'raw' image would not be available:



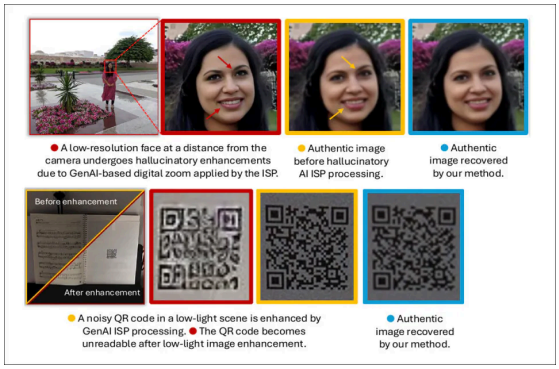
From the new paper – a RAW sensor image is processed by a GenAI-enabled ISP to produce a final sRGB output that appears clearer but may contain hallucinated details, as shown in the license plate example where characters are incorrectly inferred during digital zoom. The true scene, which is not accessible in practice, differs from both the AI-enhanced output and the intermediate authentic image prior to hallucination. The proposed approach enables recovery of this pre-hallucination image, restoring what the camera optics originally captured before AI-based enhancements modified the content. Source

The above examples come from a new research paper that offers a remedy to 'native AI photos', using alternate AI processes to reconstruct the estimated raw and unadulterated image from the processed image.

The authors state:

> *'When AI models trained with generative or perceptual losses are used in ISPs, they are prone to hallucinate content, potentially altering image [meaning]. The implication is that images directly output from the camera may now contain "fake" content, especially in smartphone cameras where AI-ISP modules are seeing increasing adoption.*
>
> *'The use of GenAI in camera hardware marks a paradigm shift in how we view camera images and challenges the traditional forensic view of camera-captured images as inherently trustworthy.'*

The new work uses a very light encoder and MLP decoder, which can be included inside the image at a weight penalty of only 180kb. The objective is the development of encoding systems fast enough to re-extract the original image in real time.



From the new paper: GenAI-based super-resolution within the camera ISP can subtly alter facial features, shifting appearance or perceived identity through changes in gaze and mouth shape. Low-light enhancement can similarly modify image content, affecting interpretation despite improving visual quality. In the QR code example, enhancement makes the image more appealing but renders the code unreadable. The method enables recovery of the authentic image prior to these hallucinations, restoring original facial details and a scannable QR code.

Alternatively, camera manufacturers could give users access to the truly unmolested sensor dumps; however, this seems likely to stay restricted to very high-end equipment. In the mobile and popular consumer space, sadly, access to non-processed photos may be considered a 'niche' or marginal pursuit.

While consumer cameras have always applied some level of post-processing, prior to the advent of edge AI, the algorithms used were minimally 'interpretive', and not likely to alter the content of a photo in the same meaningful way that current AI methods can.

Interestingly, considering the extent to which Samsung's 'moon-replacement policy' was subject to public criticism some years back, Samsung's AI Center at Toronto is one of the participants in the new work, which is titled *Addressing Image Authenticity When Cameras Use Generative AI*, and is led by contributions from five researchers at the University of Toronto.

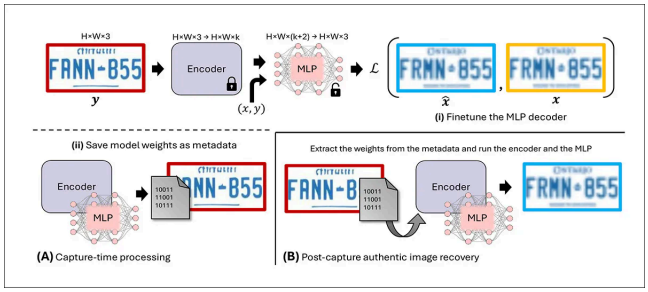
Method

The authors make use of the only other project to date that appears to have directly addressed the issue of perturbation-by-design: the 2024 paper *Advocating Pixel-Level Authentication of Camera-Captured Images*, which proposed a 'binary authentication mask' delineating areas changed by in-camera AI processes:

Rightmost, the 2024 paper's 'authenticity mask' reveals areas of the sky affected by AI 'smoothing' processes in a camera.

However, the system offered no method by which a 'true' image might be recovered, which the new work addresses, while acknowledging a debt to the earlier outing.

The objective of the new work is to enable users to recover an image as near as possible to what actually hit the sensor before the processing took place:



Overview of the proposed method. In (A), at capture time, the ISP output image containing hallucinations is passed through a frozen pretrained encoder, and its latent features are combined with spatial coordinates and fed into an MLP that operates per pixel to predict the non-hallucinated image, with training guided by a loss against the authentic image. The encoder and MLP weights are then saved as metadata alongside the image. In (B), at inference, these weights are retrieved from the metadata and used with the encoder and MLP to reconstruct the non-hallucinated image.

At capture time, in the new method, the processed image is passed through a frozen encoder that converts it into a compact latent representation. Subsequently, the relevant spatial coordinates are combined with these features and fed into a lightweight MLP that operates on a per-pixel basis, to predict the original image content – learning, effectively, to subtract the hallucinated elements through a reconstruction loss, against authentic targets.

The encoder and decoder are pretrained on paired authentic and hallucinated images, then quickly fine-tuned for each captured image, with their weights stored as metadata alongside the photo itself, adding only a small size overhead.

At viewing time, these stored weights are extracted and reused to run the same encoder and MLP, enabling recovery of an image that closely approximates what the camera sensor originally captured, without introducing new synthetic content.

Data and Tests

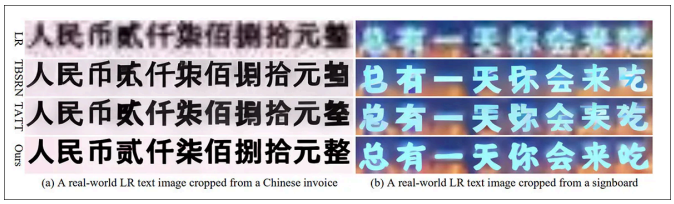
The authors tested the new method using two of the most commonly-implemented ISP post-processing tasks: super-resolution (SR, including for zoomed areas); and low-light photography.

For the general ('natural image') SR section of the tests, many examples of text were included in the data, since ISP SR routines are known to have altered text (for instance, of car license number plates, but see examples earlier in the article). Since text distortion is a discrete issue in its own right, this was treated as a sub-set of the SR tests, with dedicated data.

The aforementioned encoder was trained for each of the two modalities tested, and each was selected based on which AI ISP module was likely to be engaged during capture (i.e., a 'low-light' module, in dark conditions).

The authors used the DIV2K dataset for super-resolution training, powered by the popular RealESRGAN network. In line with the aforementioned 2024 work on ISP interference, the researchers generated paired data featuring unaffected and hallucination-affected content.

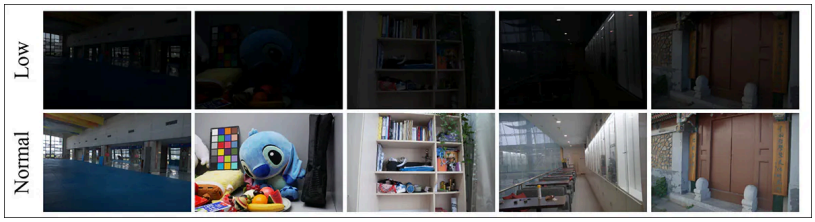
For the text SR section, the authors used the 2023 MARCONet text SR model:



From the 2023 MARCONet paper, examples of real-world low-resolution and equivalent upscaled texts. Source

To create paired data in this case, the researchers ran non-hallucinated images through MARCONet. Two thousand images were generated from the project's original code, with 200 set aside for validation, along with another 200 for testing.

For the low-light tests, the LOw-Light dataset (LOL) dataset from a 2018 Chinese paper was adopted:



From the 2018 Chinese LOL dataset, bracketed examples of the same pictures at different exposures and levels of darkness and degradation. Source

Rival Frameworks

To evaluate the method, comparisons were made against three specific baselines trained under matched conditions. First, SIREN and NeRF were pretrained on paired authentic and hallucinated images and then fine-tuned at capture time for the same duration as the proposed approach, offering a direct comparison to NeRF.

Second, an MLP with a learned encoding based on the *hashgrid* method from Instant-NGP was used, with the hash table entries and MLP jointly optimized.

The embedding size and network capacity were matched to the target encoder and MLP, with experiments covering both per-image optimization from scratch, and pretraining followed by finetuning.

Third, a blind image-to-image translation baseline was implemented using a 64MB NAFNet model, trained as a pixel-to-pixel regression system without access to metadata.

In training, the Adam optimizer was used over PyTorch, both for pretraining and fine-tuning. The encoder and MLP were trained for 50,000 epochs at a batch size of 32, with modality-specific encoders trained for each task (i.e., SR, text-SR, low-light).

Fine-tuning took place for around *three seconds* on a NVIDIA V100 GPU with 32GB of VRAM. The authors note that while on-device optimization is the target environment and scenario, it was not realistic to implement this for all the frameworks, and therefore all tests were conducted in a desktop environment:

Methods	Input type	PSNR (dB) on various datasets		
		DIV2K [10]	MARCO-Net [37]	LOL [56]
SIREN [49]	xyRGB	28.75	27.56	34.87
	xy	19.57	20.60	25.43
NeRF [40]	xyRGB	29.46	26.63	36.24
	xy	23.48	27.10	34.73
Hashgrid [41]	xyRGB	29.20	30.32	35.65
	xy	17.78	18.92	22.70
Blind NAFNet [12]	-	32.25	27.22	23.04
Ours	-	**32.96**	**31.26**	**36.34**

Performance comparison against metadata-assisted MLP-based baselines, including SIREN, NeRF, and the hash-grid method, alongside blind recovery using NAFNet. Results are reported as PSNR in decibels across three tasks: natural image super-resolution on DIV2K; text super-resolution on MARCONet; and low-light enhancement on LOL, with the authors' method achieving the highest scores in each case.

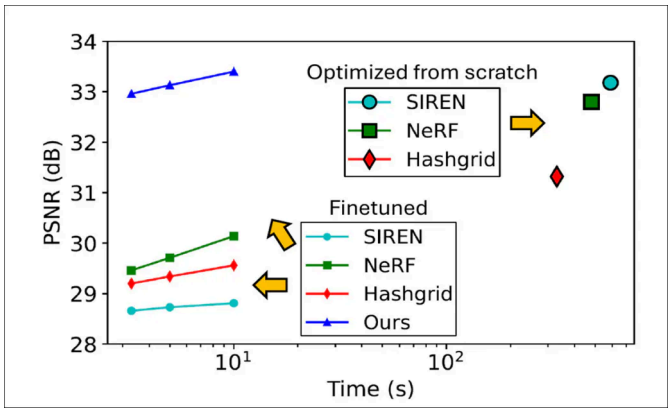
For MLP-based approaches, performance depended heavily on the input representation, where models trained using only spatial coordinates struggled during pretraining and failed to improve during the limited finetuning stage. Adding color information led to stronger results.

Blind recovery using NAFNet performed well on DIV2K, where the mapping from degraded to clean images was relatively stable, but broke down on MARCONet and LOL, where multiple plausible reconstructions existed and the model lacked the information needed to resolve this ambiguity.

This effect was most pronounced in low-light enhancement, where the original brightness of the scene could not be reliably inferred from the processed image alone.

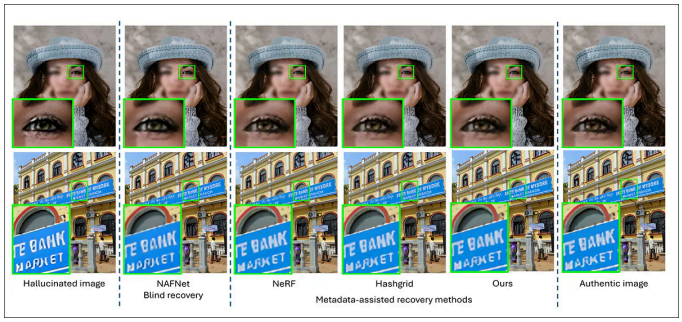
The authors state:

> *'[In] the synthetic MARCONet data, images with different blur strengths map to the same hallucinated image. It can be seen from the results that our proposed approach outperforms competitors across all datasets.'*



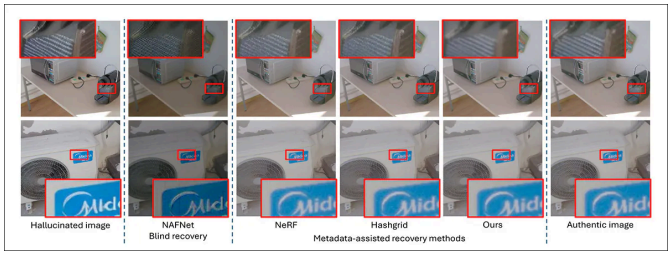
In the comparison above, we can see how well different methods perform depending on how much time they are given to run at the moment a photo is taken. Training a model from scratch for each image can produce strong results, as seen with SIREN, NeRF, and hash-grid – but this takes too long to be practical inside a camera.

Instead, the authors' method does most of the work in advance, with a quick adjustment at capture time, allowing it to deliver better results within tight time limits (3, 5, or ten seconds).



Performance comparison against metadata-assisted MLP-based baselines, including SIREN, NeRF, and the hash-grid method, alongside blind recovery using NAFNet. Results are reported as PSNR in decibels across three tasks: natural image super-resolution on DIV2K; text super-resolution on MARCONet; and low-light enhancement on LOL, with the proposed method achieving the highest scores in each case. Please refer to the source paper for (slightly) better resolution.

Above are shown qualitative results on DIV2K, where enhancement methods introduced visible hallucinations. A GAN-based super-resolution model altered eye color, and blind recovery struggled to reconstruct the original image. NeRF and hash-grid produced artifacts in structured regions such as windows and text, while the proposed method more closely matched the authentic image.



Finally, in the figure above, we see results on the LOL dataset, with brightness scaled for visualization.

Blind recovery could not resolve the unknown brightness scale, while the proposed method better reconstructed textures and restored altered characters, such as correcting a '1' back to 'i', without adding artifacts.

Conclusion

It is probably not arguable, nor ever was arguable, that 'the camera never lies'. Every decision about what to photograph and when to photograph it, along with how to present and contextualize it, is in effect a political or social decision.

Even the oldest methods of post-processing, such as dodging and burning (long-since transferred to Photoshop tools) are highly subjective acts of artistic decision and preference.

However, that's no reason to give up on at least the goal of 'objective' image captures; and it does seem reasonable that the average consumer should be allowed, even with some difficulty, to access the 'unmassaged' raw sensor dumps of the photos they take, if they want to; or at least, that they be permitted to restrict ISP post-processing to non-AI algorithms, as they might prefer.

First published Friday, April 24, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More