Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Advocating Pixel-level Authentication of Camera-captured Images

Abhijith Punnappurath[1], Luxi Zhao[1], Abdelrahman Abdelhamed[2]†, and Michael S. Brown[1]
[1]Samsung AI Center Toronto, 101 College St, Suite 420, Toronto, Ontario M5G 1L7, Canada
[2]Google Research, 111 Richmond St West, Toronto, ON M5H 2G4, Canada

Corresponding author: Abhijith Punnappurath (e-mail: abhijith.p@samsung.com).

ABSTRACT The authenticity of digital images posted online and shared on social media is often questioned due to the ability of photo-editing software to alter image content and generative AI methods that can produce visually compelling *deepfakes*. Only images directly produced by cameras are deemed unaltered and beyond suspicion, as they have not undergone any modifications. However, there is a recent trend among camera manufacturers to integrate AI-based modules into the dedicated onboard hardware, specifically the image signal processor (ISP), responsible for processing the captured sensor image into the final saved image for users. Many of these AI modules utilize perceptual or generative losses during training, which can ‘‘hallucinate’’ image content. While this hallucinated content often manifests as small details and textures, there are instances where these regions unintentionally impact the interpretation of the entire image. This paper aims to bring attention to this issue and advocate for in-camera strategies to validate the authenticity of camera-captured images at a pixel level. We propose the creation of an "authenticity" mask that could be stored as additional metadata with each image. This information can be extracted and overlaid on the image to easily identify the hallucinated regions. Considering the widespread implications of image authenticity (e.g., in courtroom evidence, news broadcasts, and other media forms), we anticipate that authentication metadata will become a standard practice for any ISP utilizing AI.

INDEX TERMS Authenticity, AI camera, metadata, neural ISP, generative models, adversarial loss, perceptual loss, digital forensics.

I. INTRODUCTION AND MOTIVATION

DIGITAL image forensics is a branch of forensics that aims to validate whether digital images are authentic or fake. Traditionally, such ‘‘fakes’’ are created using professional image editing software, such as Adobe Photoshop. More recently, with the proliferation of generative AI methods, there are now tools available on the internet based on generative adversarial networks (GAN) [12] and diffusion models [13] that allow users to easily create *deepfake* images. *Deepfake* software is becoming increasingly user-friendly with reduced computational requirements, leading to widespread use.

Digital image forensics has an arsenal of tools to detect fakes, ranging from crosschecking conventional image metadata to analyzing image noise profiles, lens effects, compression patterns, scene lighting inconsistencies etc., to training deep-learning models for forgery detection, and more [1], [10], [11], [21], [24], [30]. These forensic analysis tools are applied to images suspected of alteration via external photo-manipulation software or images that are born digital using deepfake methods. In stark contrast, images saved directly from a camera are treated as authentic representations of the scene being imaged.

Unfortunately, the assumption that camera images are 100% authentic may no longer be valid. To understand why, it is necessary to examine the inner workings of a camera. A camera image starts with scene light passing through a lens, producing a photoelectric charge at each pixel site on the sensor. The sensor digitizes these electrical signals to produce what is referred to as a RAW image. This RAW image is unsuitable for viewing and must be processed before it can be displayed to the user. Cameras have dedicated hardware called an image signal processor (ISP) responsible for processing the RAW image to its final photo-finished image state. The ISP-rendered image is saved in a display-referred color space, such as standard RGB (sRGB) or Display-P3, suitable for viewing on monitors and mobile devices. Typical processing steps (or blocks of steps) on an ISP include demosaicing, noise removal, white balance, color

†Work done while with the Samsung AI Center Toronto.

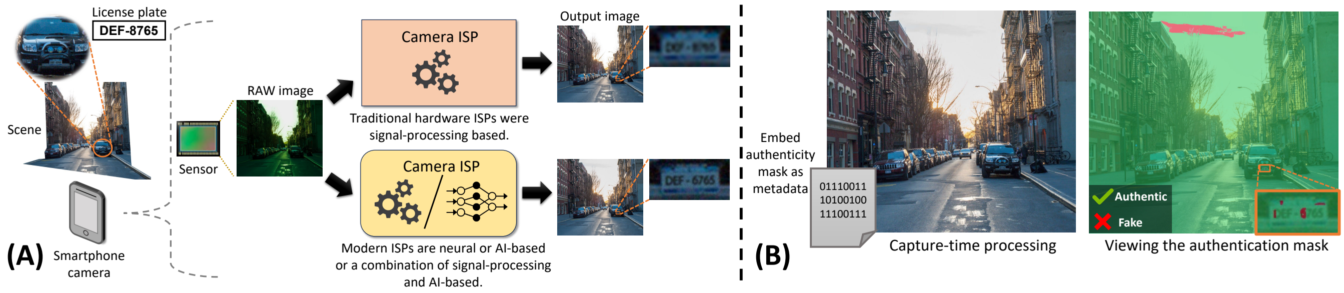

License plate
DEF-8765
Scene
Smartphone camera
(A)
Sensor
RAW image
Camera ISP
Traditional hardware ISPs were signal-processing based.
Output image
Camera ISP
Modern ISPs are neural or AI-based or a combination of signal-processing and AI-based.
Embed authenticity mask as metadata
01110011
10100100
11100111
(B)
Capture-time processing
Authentic
Fake
Viewing the authentication mask

FIGURE 1. (A) Traditional ISPs rely on signal processing-based routines to render the RAW sensor image to the final display-referred output of the camera. Modern ISPs, particularly on smartphone cameras, often employ neural or AI modules that may hallucinate scene content. In this illustrative example, the license plate of a car cannot be resolved by the camera's optics. The first digit '8' on the real license plate has been hallucinated as a '6' by the ISP's digital zoom AI module. Such a saved image would often be considered authentic. (B) In this paper, we advocate that capture-time metadata in the form of an authentication mask be saved with the image. The mask reveals pixels or regions in the image that may have been potentially hallucinated. When the user queries the image post-capture for authenticity information, the mask is recovered from the metadata and overlaid on the image to visualize non-authentic regions. In this example, the first digit '6' is flagged as fake, along with regions in the sky where clouds were hallucinated by the ISP's AI-based exposure correction module. (The original license plate number on the input sensor image was altered to maintain privacy.)

space transformation, global and local tone mapping, exposure adjustment, sharpening, rescaling (e.g., digital zoom), and final color space encoding [4], [7]. Traditionally, such ISP processing blocks were implemented using signal-processing algorithms. While signal-processing algorithms alter the colors and tonal qualities of the image, such methods are not prone to generating new image content. However, an increasingly common trend, particularly on smartphone cameras, is to replace conventional processing steps with AI-based (or neural) algorithms, owing to AI's improved performance. Such neural or AI modules are typically trained using perceptual and generative losses. The success of many generative methods lies in their ability to hallucinate or fake perceptually plausible details and content.

While "fake" image content produced by AI-based modules is usually in the form of texture and enhanced edge details, it is possible that the hallucinated content alters how we interpret the real scene. Fig. 1 (A) provides an illustrative example. In this scenario, the camera cannot optically resolve a car's license plate. Digital zoom applied by the ISP using a conventional signal-processing-based algorithm produces a less noisy but blurry license plate. However, an AI-based digital zoom module trained with generative losses, such as a GAN, can hallucinate sharp, visually plausible content. In this example, the AI-based zoom hallucinates the first digit as '6' unintentionally instead of the true value '8'. While seemingly innocuous, this image directly outputted by a camera could potentially be used as evidence against the car's owner of the hallucinated license plate.

On the surface, it would appear that the solution is to rely only on RAW images as authentic. Most modern cameras allow users to save RAW images encoded in formats such as Adobe's Digital Negative (DNG). RAW images arguably have much less processing than rendered display-referred images. However, many smartphones do not capture a single RAW image but instead capture a burst of RAW images, even in *single photo* mode. The RAW burst sequence is aligned and merged to produce a single composite RAW image. Composite RAW images are desirable because they tend to have better noise profiles and higher tonal range than possible with a single image. If the burst alignment and fusion module uses generative AI-based algorithms, there may be hallucinated content even in the composite RAW image.

As camera manufacturers replace conventional signal-processing components of the ISP with AI modules to improve image quality, there is a need to be mindful of the implications this has for image authenticity. This paper aims to shed light on this problem and suggest potential strategies to mitigate this issue via authentication metadata. Fig.1 (B) illustrates how the authentication mask will be saved as metadata and displayed to the user. We describe strategies for constructing pixel-level authentication masks in two common scenarios. In the first scenario, an AI module capable of image hallucination is used in a black-box manner with no ability to modify the underlying module. In this use case, a method to detect fake pixels is required. The second scenario is when a camera manufacturer is designing the AI module. For this situation, we suggest a training regime that separates the generative loss component of the AI module to readily identify hallucinated pixels. Finally, we outline how the authentication mask can be designed and incorporated as metadata. Given the proprietary nature of ISP design among different manufacturers and the rapid advancements in AI-based algorithms, the strategies described in this paper only serve as a guide. Our main contribution is to elucidate the problem of hallucinated image content within a camera ISP and encourage the development of pixel-level authentication masks.

II. RELATED WORK

Before outlining methods to produce an authentication mask, we briefly review various ISP architectures and modules, as well as common neural training regimes and loss functions.

Fig. 2 shows different possible ISP architectures and the common modules or blocks inside an ISP. A traditional signal-processing-based ISP has a number of processing blocks, as shown in (A), with the operations being applied in sequence. The ISP can broadly be divided into a Bayer

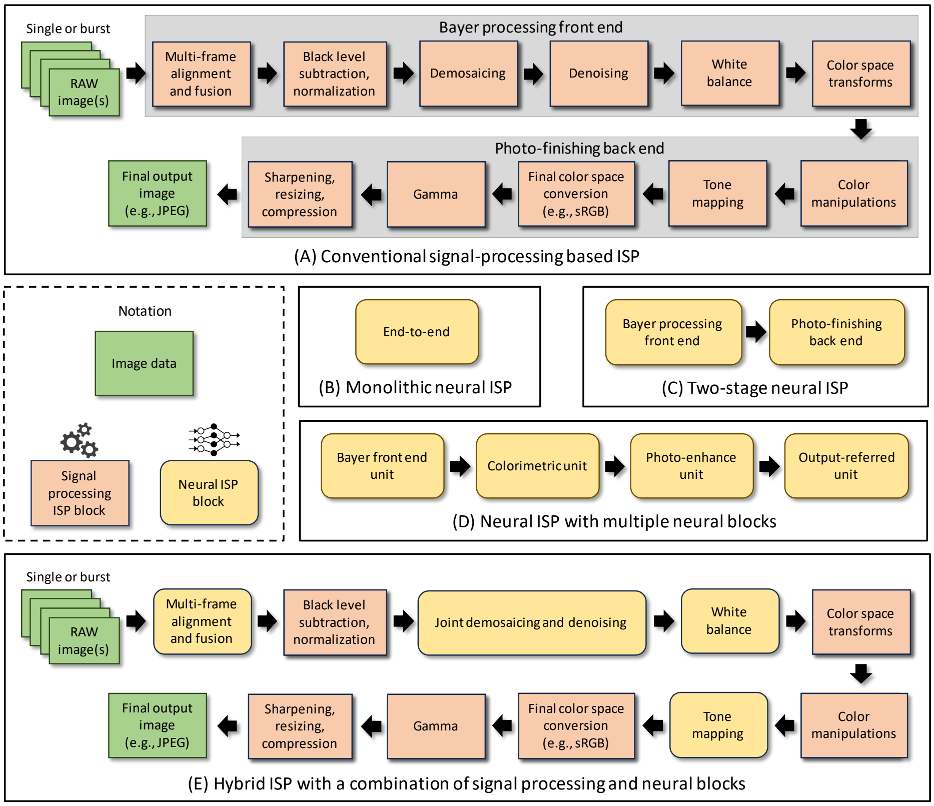

Single or burst
RAW image(s)
Bayer processing front end
Multi-frame alignment and fusion
Black level subtraction, normalization
Demosaicing
Denoising
White balance
Color space transforms
Photo-finishing back end
Final output image (e.g., JPEG)
Sharpening, resizing, compression
Gamma
Final color space conversion (e.g., sRGB)
Tone mapping
Color manipulations
(A) Conventional signal-processing based ISP
Notation
Image data
Signal processing ISP block
Neural ISP block
End-to-end
(B) Monolithic neural ISP
Bayer processing front end
Photo-finishing back end
(C) Two-stage neural ISP
Bayer front end unit
Colorimetric unit
Photo-enhance unit
Output-referred unit
(D) Neural ISP with multiple neural blocks
Single or burst
RAW image(s)
Multi-frame alignment and fusion
Black level subtraction, normalization
Joint demosaicing and denoising
White balance
Color space transforms
Final output image (e.g., JPEG)
Sharpening, resizing, compression
Gamma
Final color space conversion (e.g., sRGB)
Tone mapping
Color manipulations
(E) Hybrid ISP with a combination of signal processing and neural blocks

FIGURE 2. Examples of modern ISP architectures. (A) A traditional signal-processing-based ISP. (B) An AI ISP where a single monolithic neural block replaces the ISP in (A). (C) An AI ISP with two neural blocks for front- and back-end processing. (D) A modular AI ISP with multiple blocks. (E) A hybrid ISP with some signal-processing blocks replaced with neural units.

front end and a photofinishing back end. One possible AI ISP architecture is to replace the entire signal-processing-based ISP with a single neural unit trained end-to-end, as shown in (B). Often, a two-stage AI ISP, as in (C), with a front and back end, is preferred over the monolithic structure in (B). More modular architectures, such as in (D), are also possible since they offer more fine-grained control and better interpretability. Note that the number of blocks and their connections are a design choice—other configurations are also possible. The most common case is depicted in (E), where one or more blocks of a conventional signal-processing-based ISP may be replaced with neural units.

Neural networks may be trained using standard reconstruction losses such as mean squared error (MSE) or mean absolute error (MAE). Perceptual losses such as VGG loss [20] or Learned Perceptual Image Patch Similarity (LPIPS) [31] are also common. Perceptual losses target image enhancement (rather than reconstruction) and are prone to hallucinate details or image content that is visually plausible and pleasing. Using generative loss functions can also lead to hallucinated content. GAN-based methods trained using an adversarial loss function are a common example of this category. It is important to distinguish that while the outputs of networks trained using only reconstruction losses such as MSE and MAE may contain visual artifacts, these artifacts differ from those created by perceptual losses. In practice, a reconstruction loss is used to recover an underlying signal that has become corrupted or distorted. Perceptual and adversarial losses aim to improve the perceptual quality of the result, regardless of the underlying signal. This decoupling from the underlying signal encourages the generation (i.e., hallucination) of detail and texture. As described in [3], these two losses are at odds. Typically, both losses are used when training a deep learning module, and balancing between the two criteria is at the discretion of the algorithm designers.

Replacing the conventional signal-processing ISP with a neural one has been a topic of significant interest in the past few years, with many ideas being explored in terms of architecture and loss functions and several competitions being organized yearly [8], [15]–[18], [26]. With the growing emphasis on improving smartphone image quality, the first RAW-to-sRGB challenges in [16], [18] aimed to process low-quality RAW images from a smartphone camera to look like high-quality rendered images from a DSLR camera. A large dataset of paired RAW-sRGB images [19] captured using a smartphone camera and a DSLR, respectively, was provided as part of the challenge. All methods used a monolithic architecture as in Fig. 2 (B). In addition to reconstruction losses, most works also used perceptual losses, while some also considered adversarial and style losses. Alongside image quality, runtime performance is also a crucial consideration on smartphone devices. Therefore, subsequent challenges [15], [17]

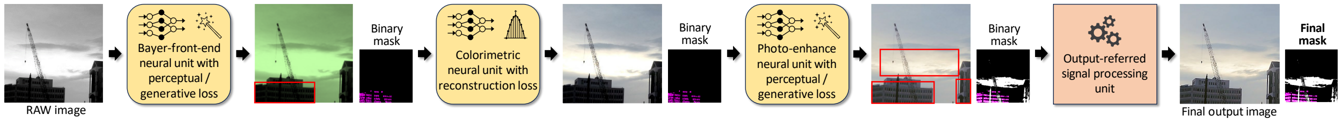

RAW image
Bayer-front-end neural unit with perceptual / generative loss
Binary mask
Colorimetric neural unit with reconstruction loss
Binary mask
Photo-enhance neural unit with perceptual / generative loss
Binary mask
Output-referred signal processing unit
Final mask
Final output image

FIGURE 3. An example showing how the binary authentication mask is computed and propagated at each stage of the ISP. The ISP is a hybrid ISP with a combination of signal processing-based and neural blocks. The neural blocks trained with perceptual and/or generative losses may introduce fake content, as indicated by the binary masks. Neither the signal processing-based block nor the neural block trained with pure reconstruction losses hallucinates details. The mask is carried forward at these stages. The final mask is saved as metadata.

placed more stringent constraints on on-device performance while providing a unified platform to evaluate proposed models on the target device. Similar trends were observed in these works regarding loss functions and a monolithic architecture, but notably, most methods could obtain close to real-time performance on the device. Works such as [6], [27], while still following a monolithic architecture, used additional metadata information such as ISO gain values and white-balance as input, similar to a conventional ISP.

A two-stage ISP with a Bayer processing front end and a photo-finishing back end similar to the structure in Fig. 2 (C) was proposed in [22]. The two networks, Restore-Net and Enhance-Net, were first trained separately using an MAE reconstruction loss and later jointly finetuned using a combination of MAE and perceptual losses. Recently, the more complex task of processing RAW images of nighttime scenes to sRGB was considered in the night photography rendering challenge [8], [26]. Due to the unique lighting environment in nighttime scenes, white balance, tone curves, and photo-finishing strategies vary significantly for nighttime rendering compared to daytime photography. Many methods adopted a more modular structure, similar to Fig. 2 (D). In particular, the winner of the challenge in [8], DeepFlexISP [23], used a three-stage network structure breaking down the full ISP into a denoising network, a white-balance network, and a Bayer-to-sRGB network. There were also user-controllable parameters for denoising strength, color cast, and overall brightness. Notably, many of the solutions proposed in [8], [26] used perceptual or adversarial loss functions for training. This widespread adoption of neural ISP modules raises the need for pixel-level authentication.

III. CREATING PIXEL-LEVEL AUTHENTICATION METADATA

In this section, we outline a framework for computing and propagating the authentication mask through the various stages of the ISP. As mentioned in Sec. I, there are two common scenarios faced by camera engineers. The first is when an AI-based ISP component is used in a black-box fashion. This arises when a third-party pre-trained solution is used or when an internally developed module cannot be altered (often due to lack of time). The second scenario is when camera engineers have control over an AI-module's design and training. For these two scenarios, we describe methods to detect hallucinated pixels to generate the authentication mask. Finally, we discuss how the metadata and the processed image can be efficiently and securely saved.

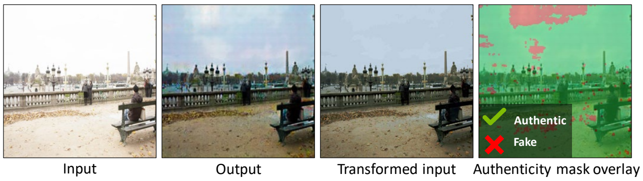


FIGURE 4. An example of a black-box neural ISP module used for exposure correction. The input and output of the module are shown in the first two columns. The third column shows the input transformed to the output using a global mapping function. Binarizing the difference between the transformed input and the output yields the authentication mask, shown overlaid on the output, in the fourth column. Cloud-like details hallucinated by the exposure adjustment module are flagged as fake in the authenticity mask.

We start by providing a high-level diagram of how an authentication mask is computed at different stages of the ISP in Fig. 3. In this example, the camera ISP is a hybrid ISP containing three neural blocks and one signal-processing-based block. A hybrid ISP mixing signal-processing and neural-processing components is currently the most common architecture used on devices. In this example, the first block is a Bayer front-end neural unit that inputs a single-channel Bayer-RAW image and outputs a denoised and demosaiced three-channel RAW-RGB image. The Bayer-front-end unit is a GAN-based network that hallucinates content around the edges of the building windows during demosaicing. These pixels are flagged as hallucinated (magenta color) in the binary mask. The second block is a colorimetric neural unit that takes the RAW-RGB image as input and applies white balance and color space transforms. This unit is trained using only reconstruction losses and does not introduce any fake content. As a result, the mask is carried forward from the previous stage. The third block is a photo-enhancement neural unit trained using a combination of perceptual and generative losses, and it hallucinates sky and building texture (white-colored regions in the binary mask) where there is under-/over-exposed content. The mask is now the union of the previous and the current stage (magenta + white). The final block is an output-referred unit, a conventional signal processing block that performs some sharpening and compression. The mask is carried forward to the final stage and saved as metadata.

For more granular annotation of which unit(s) of the ISP generated fake content, it is possible to store an n-bit mask, instead of a binary mask. In the example of Fig. 3, one could use a 4-bit mask since there are four ISP blocks. Each binary bitplane reveals the hallucinated pixels produced by

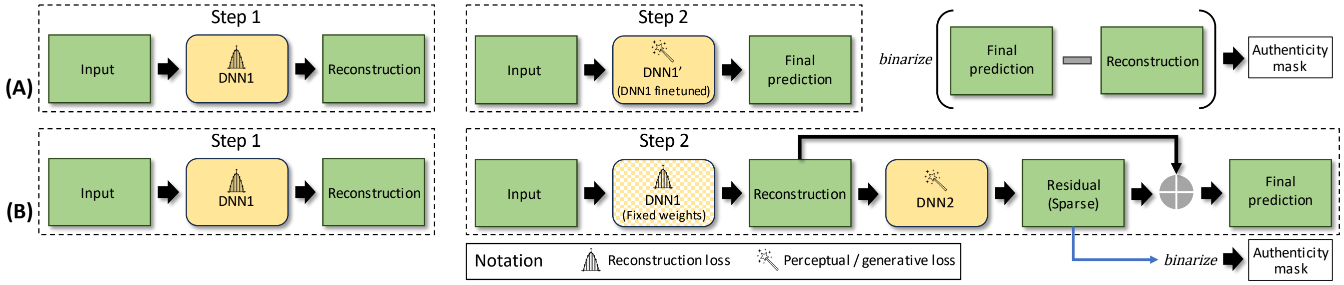


FIGURE 5. When the neural ISP modules are trainable, it is possible to adapt the training framework to automatically and accurately segment authentic and fake pixels. Two possible strategies are shown: (A) demonstrates an approach where we finetune the reconstruction model with perceptual/generative losses to obtain the final prediction. (B) demonstrates an approach where we cascade the reconstruction model and perceptual loss model to obtain the final prediction. In both cases, the authenticity mask is a natural outcome of the training process and does not need to be separately estimated.

the corresponding ISP module. However, note that this comes at the expense of larger metadata size.

A. BLACK-BOX NEURAL ISP MODULE

When working with a black-box neural module, we must determine which parts of the image have been hallucinated by examining only the input and output. One approach to this problem is to compute a global mapping between the input and output of the AI module. For example, we can fit a polynomial mapping (either globally or with smooth spatial interpolation) between the input and output RGB values. Such mappings have been successfully used to model color transforms between input/output pairs [14], [25]. This polynomial function is regressed using a reconstruction loss (e.g., MAE or MSE). The smooth nature of the polynomial mapping and the use of a reconstruction loss make it difficult for the estimated transform to model the hallucinated content. If we apply this smooth mapping to the input image and compare this with the output of the AI module, we can use simple thresholding operations to detect where hallucinated detail is in the AI module's output image. Different transformation functions can be used depending on the ISP stage and the complexity of the mapping required. This is a simple approach and has the benefit of minimal added overhead of mask generation.

Fig. 4 shows an example of the task of exposure adjustment. We selected the GAN-based exposure correction approach of [9] for this experiment. We used the pre-trained model released by the authors [1]. The input image in the first column is taken from the MIT-Adobe FiveK dataset [5]. The image has a lot of saturated pixels, particularly in the sky region. The second column shows the output of [9]. Using a GAN loss for this module results in cloud-like details being hallucinated where there were originally only saturated pixels (i.e., no information). We fit a global polynomial mapping function [14] to transform the input into the output. The transformed input image, after the mapping function is applied, is shown in the third column. We compute the absolute difference between the transformed input and the output and threshold the result to obtain the binary authenticity mask. We show the mask overlaid on the output image in the last column. Image details in the sky region hallucinated by the GAN are flagged as non-authentic regions in the binary mask.

B. TRAINABLE NEURAL ISP MODULE

In scenarios where the camera manufacturer has the flexibility to modify the AI module's design and training method, we propose an approach that trains two networks, one focused on reconstruction only and one focused on perceptual and/or generative losses. Training requires a two-step procedure as shown in Fig. 5 (A). In the first step, we train a deep neural network, denoted as DNN1, using a standard reconstruction loss between the network's output and the ground truth. Once training has converged, we save the weights of this reconstruction-only model. As the second training step, we finetune the same network using perceptual and/or generative losses. In the case of adversarial training, an additional discriminator network can be added during this step. At test time, the output of the finetuned model DNN1$'$ is the desired output image, while the difference image between the output of the finetuned model and the output of the reconstruction-only model reveals the hallucinated content. Our authentication mask can be directly obtained by binarizing the difference image with an appropriate threshold. Optionally, simple morphological operations can also be applied to the binary mask to remove spurious noise.

Fig. 6 shows an example of this strategy for the super-resolution task. For this experiment, we chose the Real-ESRGAN method [28]. We used the official implementation from the authors [2]. The input image is from the DIV2K dataset [2]. We selected a super-resolution factor of ×4. The input and the ground truth are shown in the first row. We first train the model of [28] using a pure reconstruction loss. The result of this model is shown on the bottom left. Next, we finetune the model with an adversarial loss. This is the final output and is shown on the bottom right. The authenticity mask is obtained by binarizing the difference between this final output and the output of the reconstruction-only model. After the model is finetuned using a GAN loss, the

[1] https://github.com/yamand16/ExposureCorrection

[2] https://github.com/xinntao/Real-ESRGAN

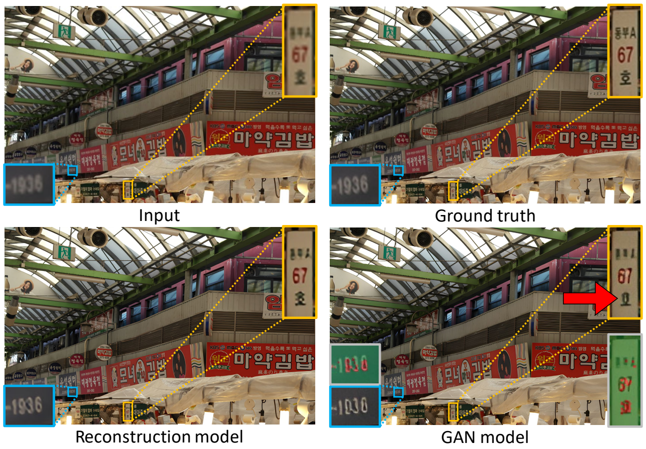


FIGURE 6. An example of a trainable neural ISP module for super-resolution implemented by finetuning the reconstruction model with perceptual/generative losses. The input and the ground truth are shown in the top row. The bottom left image shows the output of a model trained using a pure reconstruction loss. This model is then finetuned using an adversarial loss, and the final output of the GAN model is shown on the bottom right. Alongside the zoomed-in regions, our authenticity mask is also shown for the GAN model's output. Hallucinated pixels are accurately detected and flagged in the authentication mask.

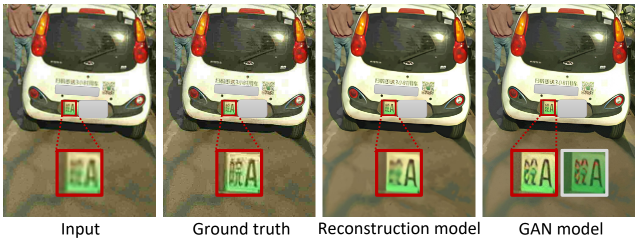


FIGURE 7. An example of a trainable neural ISP module for super-resolution implemented by cascading the reconstruction model and the perceptual/generative loss model. The input and the ground truth are shown in the first two columns. The third column shows the output of a first network trained using a pure reconstruction loss. A GAN-based second network predicts a residual image that is added to the reconstruction network's output, and this final result is shown in the fourth column. Alongside the zoomed-in region, our authenticity mask is also shown for the final output. Hallucinated pixels are accurately detected and flagged in the authentication mask. Note that the number plate has been grayed out to maintain privacy.

final output is sharper and has more detail than the output of the reconstruction model. However, from the zoomed-in regions, it can be observed that the GAN model is prone to hallucinations – the Korean Hangul character 호 is changed to what could be interpreted as either 오 or 모 (indicated by the red arrow), and the digits 9 and 6 appear as 0. Our authenticity mask correctly highlights these hallucinated regions.

We also envision a variation of our two-step training approach (shown in Fig. 5 (B)) where in the first step, we train a neural network DNN1 using a standard reconstruction loss, as performed first. Once this network is trained, we freeze the weights of DNN1, and proceed to the second step. Here, we introduce a second network, DNN2, which receives the output of DNN1 as input. DNN2 predicts a residual image that is added to its input to produce the final image. DNN2 is trained using perceptual and/or generative losses. The residual image output by DNN2 represents the hallucinated content. The advantage of training the two networks, DNN1 and DNN2, in this manner is that it once again directly decouples the fake pixels from the authentic pixels—at test time, pixel locations with a value of 1 in DNN2's residual output after binarization represent fake content. Optionally, a sparsity constraint can be imposed on DNN2's residual output during training since, in most cases, only a small percentage of pixels will be non-authentic. As before, for adversarial training, an additional discriminator network can be added to the second training step such that DNN2 is encouraged to produce a residual, which, when added to DNN1's output, is effective at fooling the discriminator.

Fig. 7 shows an example of this second approach once again for the task of super-resolution. We use the Real-ESRGAN method [28] as before. The input image is from the CCPD dataset [29]. The super-resolution factor is ×4. The input, the ground truth, and the result of the first network DNN1 trained using a pure reconstruction loss are shown in the first three columns, respectively. The second network DNN2, trained using an adversarial loss, predicts a residual image that is then added to the output of DNN1. This final output is shown in the last column. The authenticity mask is directly obtained by binarizing DNN2's residual output. As shown in the zoomed-in region, the Chinese character 皖 is changed to what could be interpreted as 铃 in the final result. Our authenticity mask correctly highlights this hallucinated character.

C. SAVING THE AUTHENTICATION METADATA

To reduce the metadata size, the binary mask can be downsampled and compressed. For example, a standard sized 12-mega-pixel image (i.e., 3000×4000) will require around 2 to 4 MB of storage after lossy compression using JPEG or HEIC. In comparison, a binary mask at full resolution compressed using a lossless binary image compression algorithm will be around 150 to 300 KB, depending on the sparsity of the mask. If the binary mask is downsampled to half the resolution 1500 × 2000 pixels, the file size further reduces to around 64 KB to 96 KB, representing a nominal overhead of 2 to 3%.

It is also important to prevent tampering with the mask itself. While securing information within a digital document is its own research topic, we envision methods for storing the authentication mask based on manufacturer encryption and steganography.

IV. CONCLUDING REMARK

This paper sheds light on a concern surrounding the authenticity of images directly outputted by cameras. Traditionally, camera images have been considered reliable, as they are assumed to be unaltered at the point of capture. However, a new challenge has emerged with the integration of AI-based algorithms into the ISPs of modern cameras. These AI modules, often trained using perceptual or generative losses, can potentially introduce unintended alterations in the form of "hallucinated" content within images. Consequently, the authenticity of images captured by cameras is not guaranteed, a critical aspect often overlooked in modern digital image forensics.

This paper has advocated for the implementation of strategies within the camera's ISP to enable the validation of image authenticity at a pixel level. Specifically, we have proposed the inclusion of capture-time metadata in all outputted images. This metadata serves as a spatial mask that can identify pixels potentially affected by AI-driven hallucination, allowing users to visualize and assess the authenticity of an image at a pixel level. Due to the wide-ranging implications of image authenticity, we envision the adoption of authentication metadata as a standard practice for any ISP incorporating AI.

REFERENCES

[1] Shruti Agarwal and Hany Farid. Photo forensics from rounding artifacts. In *The ACM Workshop on Information Hiding and Multimedia Security*, 2020.

[2] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Michael S. Brown. Color processing for digital cameras. In *Fundamentals and Applications of Colour Engineering*, chapter 4, pages 81–98. Wiley, 2023.

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[7] Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual Review of Vision Science*, 7(1):571–604, 2021.

[8] Egor Ershov, Alex Savchik, Denis Shepelev, Nikola Banić, Michael S. Brown, Radu Timofte, Karlo Koščević, Michael Freeman, Vasily Tesalin, Dmitry Bocharov, et al. NTIRE 2022 challenge on night photography rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[9] F. Irem Eyiokur, Dogucan Yaman, Hazım Kemal Ekenel, and Alexander Waibel. Exposure correction model to enhance image quality. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[10] Hany Farid. Image forgery detection. *IEEE Signal Processing Magazine*, 26(2):16–25, 2009.

[11] Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[14] Guowei Hong, M. Ronnier Luo, and Peter A. Rhodes. A study of digital camera colorimetric characterization based on polynomial modeling. *Color Research & Application*, 26(1):76–84.

[15] Andrey Ignatov, Cheng-Ming Chiang, Hsien-Kai Kuo, Anastasia Sycheva, and Radu Timofte. Learned smartphone ISP on mobile NPUs with deep learning, mobile AI 2021 challenge: Report. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[16] Andrey Ignatov, Radu Timofte, Sung-Jea Ko, Seung-Wook Kim, Kwang-Hyun Uhm, Seo-Won Ji, Sung-Jin Cho, Jun-Pyo Hong, Kangfu Mei, Juncheng Li, et al. AIM 2019 challenge on RAW to RGB mapping: Methods and results. In *The IEEE International Conference on Computer Vision Workshops*, 2019.

[17] Andrey Ignatov, Radu Timofte, Shuai Liu, Chaoyu Feng, Furui Bai, Xiaotao Wang, Lei Lei, Ziyao Yi, Yan Xiang, Zibin Liu, et al. Learned smartphone ISP on mobile GPUs with deep learning, mobile AI & AIM 2022 challenge: Report. In *The European Conference on Computer Vision Workshops*, 2022.

[18] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. AIM 2020 challenge on learned image signal processing pipeline. In *The European Conference on Computer Vision Workshops*, 2020.

[19] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera ISP with a single deep learning model. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *The European Conference on Computer Vision*, 2016.

[21] Paweł Korus. Digital image integrity – a survey of protection and verification techniques. *Digital Signal Processing*, 71:1–26, 2017.

[22] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. CameraNet: A two-stage framework for effective camera ISP learning. *IEEE Transactions on Image Processing*, 30:2248–2262, 2021.

[23] Shuai Liu, Chaoyu Feng, Xiaotao Wang, Hao Wang, Ran Zhu, Yongqiang Li, and Lei Lei. Deep-FlexISP: A three-stage framework for night photography rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[24] Alin C Popescu and Hany Farid. Statistical tools for digital forensics. In *The International Workshop on Information Hiding*, 2004.

[25] Abhijith Punnappurath and Michael S. Brown. Spatially aware metadata for raw reconstruction. In *The IEEE Winter Conference on Applications of Computer Vision*, 2021.

[26] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan Ermakov, Nikola Banić, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, Simone Zini, et al. NTIRE 2023 challenge on night photography rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[27] Matheus Souza and Wolfgang Heidrich. CRISPnet: Color rendition ISP net. *arXiv preprint arXiv:2203.10562*, 2022.

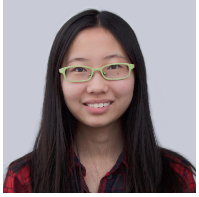
[28] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *The International Conference on Computer Vision Workshops*, 2021.

[29] Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, and Huan Huang. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *The European Conference on Computer Vision*, 2018.

[30] Marcello Zanardelli, Fabrizio Guerrini, Riccardo Leonardi, and Nicola Adami. Image forgery detection: A survey of recent deep-learning approaches. *Multimedia Tools and Applications*, 82(12):17521–17566, 2022.

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

ABHIJITH PUNNAPPURATH is a research scientist at Samsung AI Center, Toronto. He was a Postdoctoral Fellow at the Electrical Engineering and Computer Science department, York University, Toronto, Canada. He received his Ph.D. degree from the Electrical Engineering department, Indian Institute of Technology Madras, India, in 2017. His research interests lie in the areas of low-level computer vision and computational photography.



LUXI ZHAO is a machine learning engineer at Samsung AI Center, Toronto. She received her Bachelor of Applied Science in Computer Engineering from the University of British Columbia in 2020 and a Master of Science in Applied Computing from the University of Toronto in 2022. Her research interests include computer vision, computational photography, and machine learning.

ABDELRAHMAN ABDELHAMED is a research scientist at Google Research. Previously, he was a research scientist at Samsung AI Center, Toronto. He obtained his Ph.D. in computer science from York University, supervised by Prof. Michael S. Brown. He holds two MSc degrees from the National University of Singapore and Assiut University, Egypt. His research interests include computer vision, computational imaging, and machine learning.

MICHAEL S. BROWN is a professor and Canada Research Chair in Computer Vision at York University in Toronto. His research interests include computer vision, image processing, and computer graphics. He has served as program chair for WACV 2011/17/19 and 3DV 2015 and as general chair for ACCV 2014 and CVPR 2018/21/23. Dr. Brown holds a part-time position at the Samsung AI Center in Toronto as a senior research director.