

DEPARTMENT: APPLICATIONS

To Authenticity, and Beyond! Building Safe and Fair Generative AI upon the Three Pillars of Provenance

John Collomosse , Adobe Research, San Jose, CA, 95110, USA. University of Surrey, Guildford, UK.

Andy Parsons, Content Authenticity Initiative, Adobe Inc., New York, NY, 10011, USA

Abstract—Provenance facts, such as who made an image and how, can provide valuable context for users to make trust decisions about visual content. Against a backdrop of inexorable progress in Generative AI for Computer Graphics, over two billion people will vote in public elections this year. Emerging standards and provenance enhancing tools promise to play an important role in fighting fake news and the spread of misinformation. In this paper we contrast three provenance enhancing technologies: metadata, fingerprinting and watermarking, and discuss how we can build upon the complementary strengths of these three pillars to provide robust trust signals to support stories told by real and generative images. Beyond authenticity, we describe how provenance can also underpin new models for value creation in the age of Generative AI. In doing so we address other risks arising with generative AI such as ensuring training consent, and the proper attribution of credit to creatives who contribute their work to train generative models. We show that provenance may be combined with distributed ledger technology (DLT) to develop novel solutions for recognizing and rewarding creative endeavour in the age of generative AI.

Fake news and misinformation are major societal challenges that impact everything from journalism, to healthcare, and our upcoming elections in which over two billion people will vote worldwide in 2024. Authentic storytelling frequently relies on graphics and visual content. Yet, whilst advances in Generative Artificial Intelligence (GenAI) have democratized access to graphics creation and manipulation [1], GenAI has also lowered the entry bar for the abuse of visual content to spread misinformation.

Digital forensics offers one solution to visual misinformation, through tools to detect synthetic or manipulated content. However, misinformation and manipulation only partially intersect. Visual content frequently undergoes some manipulation for editorial purposes. Moreover, some Human Rights Organizations report [2] that the majority of visual misinformation doesn't involve manipulation but rather misattribution – unaltered images, used out of context, to tell a different

story. Thus forensics may uncover only a subset of misinformation, and is engaged in a cat and mouse game with the ever quickening progress of GenAI.

Tools that explain the provenance of visual media, rather than decide its authenticity, help to put consumers in a more informed position to make trust decisions on visual content. A user encountering a photograph on social media, may be presented with information on who took it, when, and how it was modified. This additional context – likened to a 'digital nutrition label' – may provide a moment to reflect upon whether to accept that image as fact, or whether to reshare it. In the psychology of Nobel laureate Daniel Kahneman [3], this may enable a switch from a 'System 1' or reactionary response, to a rationalized or 'System 2' trust decision in light of the available facts.

Provenance also has a role to play beyond authenticity, to enhance creative attribution. The need for artists to be recognized and rewarded for the reuse of their intellectual property is fundamental to the creative economy. This issue is particularly acute with the emergence of GenAI. The provenance of synthetic media is the GenAI model that created it and, in turn,

APPLICATIONS




	Resilient to legacy platform stripping	Resilient to attacker stripping / spoofing	Deterministic lookup	Hard Binding	Detectable on client side	Standalone (no network)
Metadata 	✗	✗	✓	✓	✓	✓
Fingerprint 	✓	✓	✗	✗	✗	✗
Watermark 	✓	✗	✓	✗	✓	✗

FIGURE 1. Three complementary technologies for visual provenance. Metadata is easily stripped and replaced, but requires no network interaction to view, and can be cryptographically ('hard') bound to content. Visual fingerprinting (search by perceptual hash) is entirely complementary; being a passive signal, there is nothing to strip or replace. A robust fingerprint will survive rendition, but search using such 'soft' bindings is inexact. Watermarking shares the vulnerabilities of metadata to stripping and spoofing attacks. Yet a robust image watermark will survive renditions, and enables exact lookup using an embedded identifier.

the training data used to create that model. Legislation is now emerging in the US, EU and UK requiring clearer signaling of synthetic media provenance on the grounds of transparency. Going further, a solid basis for establishing provenance for generative content, may unlock new opportunities for value creation in the creative economy, such as auditing the re-use of images in GenAI training data to compensate contributing artists. We will discuss how such capabilities may be built as a technology stack, founded upon provenance.

Three Pillars of Provenance

Three core technologies for establishing media provenance are: metadata, fingerprinting and watermarking. Each of these technologies taken alone are insufficient, as each presents its own vulnerabilities. Yet when carefully combined, they may mutually reinforce one another's strengths to produce a solid foundation for media provenance (Figure 1).

Metadata provides a convenient channel to carry provenance information within a media asset, such as an image. Depending on the image format, the metadata may be carried inside the file or within a separate 'sidecar'. Public key infrastructure (PKI) may be used to sign the metadata for authenticity. Signed metadata can contain a cryptographic hash of the content to bind it to the asset. Nevertheless, the metadata may be stripped by legacy or non-compliant platforms, or substituted by an adversary to misattribute the image.

Fingerprinting involves the use of a hash to look up a copy of the signed metadata, should it become 'decoupled' (stripped, or substituted) from the image as it circulates in the wild. Fingerprinting assumes access to a trusted repository of metadata, available over a network. It also assumes that content is identifiable under the kinds of non-editorial transfor-

mations (renditions such as resizing, format and quality change) often performed by content platforms. To meet the latter criterion, fingerprinting often relies upon perceptual hashing and AI techniques. These more-forgiving techniques will equate two images if they are closely similar, unlike cryptographic hashes which require every pixel value to match.

Watermarks are largely imperceptible signals actively embedded within image that may be used to identify it, and so recover provenance metadata from a repository (as with fingerprinting). Watermarks present a similar attack surface to metadata, in that they may be stripped from images or spoofed (copied or substituted) by an determined adversary. Nevertheless, robust watermarks persist through non-editorial renditions on platforms that strip metadata. The presence of a watermark offers both a convenient way to perform deterministic recovery of the metadata, and to tell if lookup should be performed when metadata is absent.

Content Credentials Metadata

The Coalition for Content Provenance and Authenticity is a cross-industry standards group that has developed an eponymous open specification (C2PA) for encoding provenance metadata within media assets [4].

C2PA describes provenance within a data structure called a **manifest** (Figure 2). The manifest contains facts, called **assertions**, about the provenance of the asset such as who made it, how, where and when, and using which source assets (**ingredients**). The ingredients may themselves be digital works bearing C2PA manifests so forming a graph. Assertions are collected within a **claim**, and signed cryptographically using PKI. A rich ontology of assertions are available to optionally include within a claim, and is extensible by the community. The only compulsory assertion is a

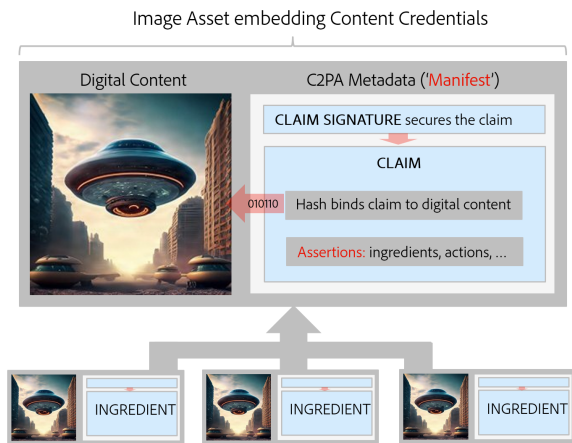


FIGURE 2. An image embedding Content Credentials (C2PA) metadata. The C2PA standard embeds metadata (a ‘manifest’) inside an image to describe its provenance. The manifest contains a signed ‘claim’, which in turn contains several ‘assertions’ (facts) about the provenance of the image, and ‘ingredients’ that reference the manifests of other images used to create it. The claim is bound to image pixels using a hash.

cryptographic hash of the image pixels that creates a **hard binding** between the manifest and that content.

Content Credentials (the name given to C2PA metadata) are now embedded by several camera hardware and creative tool vendors, and adoption is accelerating with some recent generative models (Adobe Firefly, OpenAI DALL-E) adding metadata to their synthetic images. Several apps (eyeWitness, Proof-Mode) also support embedding the metadata in photojournalism content. Yet as an emerging standard, many content platforms including – at the time of writing – all social media sites, routinely strip this metadata from images passing through them. C2PA provides open SDKs and a web app to view Content Credentials in images and video. Browser extensions are emerging to surface these in-line within images on web pages.

Fingerprint Recovery

A weakness of metadata is that it is easily stripped from assets. Recovering from stripped (or adversarially substituted) metadata requires reference to an external source of truth – a repository of manifests indexed by a key derived from the digital content. The hard binding (cryptographic hash) is unsuitable as such a key, because content is typically subjected to transformations as it is renditioned and reshared, changing the cryptographic hash. Instead, a **soft binding** may be derived from the digital content, either passively via a

perceptual hash (fingerprint) or actively via embedding (prior to distribution) a robust watermark capable of surviving such renditions.

We have implemented an image fingerprinting algorithm using a convolutional neural network (CNN) trained to produce a perceptual hash resilient to non-editorial change, but sensitive to editorial change. The latter prevents altered images being attributed to the provenance record of the original image. Our fingerprinting algorithm is applied to many millions of generated images in Adobe Firefly each day to enable recovery of stripped Content Credentials.

Our fingerprints are compact embeddings of a CNN. Let $\phi_i = E(x_i) \in \mathbb{R}^{7 \times 7 \times 256}$ be the feature map obtained as the output of a ResNet encoder E for an image x_i . We train E via a contrastive objective function (a “loss function”):

$$\mathcal{L}_C = - \sum_{i \in \mathcal{B}} \log \left(\frac{d(\phi_i, \bar{\phi}_i)}{d(\phi_i, \bar{\phi}_i) + \sum_{j \neq i \in \mathcal{B}} d(\phi_i, \phi_j)} \right), \quad (1)$$

where $\bar{\phi}_i$ represents an embedding of a differently augmented version of x_i , $d(a, b) := \exp\left(-\frac{1}{\lambda} \frac{a^T b}{\|a\| \|b\|}\right)$ measures the cosine similarity between flattened feature vector inputs a and b , \mathcal{B} is a large randomly sampled training mini-batch, and λ is a scaling constant to tune sensitivity. The augmentations applied during training mimic the non-editorial renditions and image degradations that the fingerprint should be invariant to. An identical but negated term is added to \mathcal{L}_C , applying editorial augmentations $\bar{\phi}_i$ that the fingerprint should be sensitive to. Fingerprints are projected and quantized to create a compact soft binding key. To recover a manifest for a given query image, nearest-neighbor retrieval is performed using this key. A shortlist of matches is obtained using efficient vector search within a sharded OpenSearch index.

In order to reduce false positives, all retrieved matches are passed through the Image Comparator Network (ICN), a further CNN that performs pair-wise verification on each query-candidate match pair [5]. We have found this trainable verification module essential to produce a practical score that may be thresholded to exclude false positives at scale. In effect, the module learns a non-linear operator to compare fingerprints, over a shortlist first computed by a simpler embedding distance on $\phi(\cdot)$. Further training steps mitigate adversarial attacks which, combined with our two stage matching approach, defend against images crafted to trick the system into retrieving incorrect manifests.

APPLICATIONS

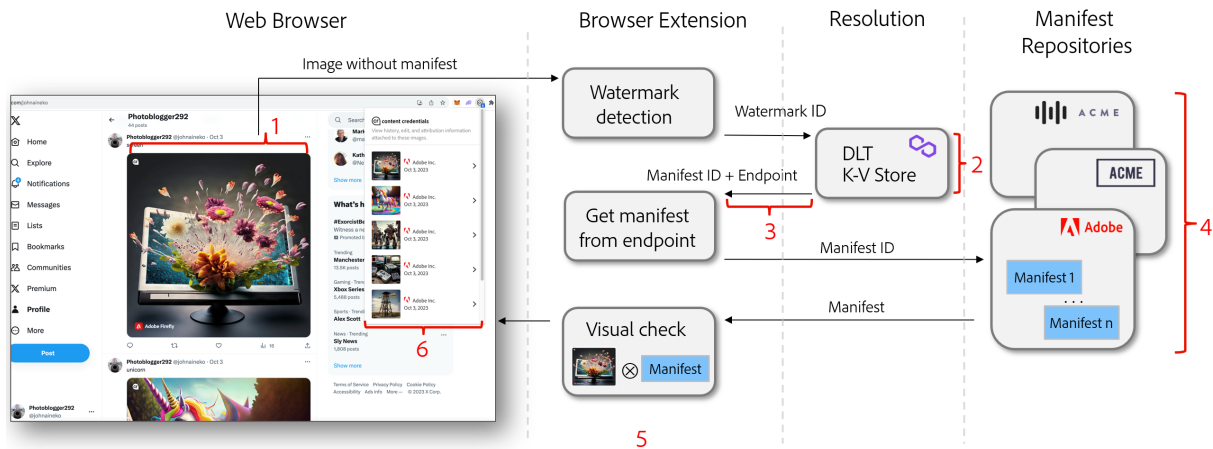


FIGURE 3. Tracing the provenance of a Generative AI image posted on social media, using a web extension that recovers Content Credentials from an image stripped of its metadata. Workflow: 1) Graphic encountered on social media feed without metadata; 2) Browser extension detects invisible watermark in graphic and extracts unique identifier (watermark ID); 3) A key-value (K-V) store on a Distributed Ledger (DLT) uses the watermark ID (key) to obtain a manifest ID and endpoint URI of the manifest repository (value); 4) Extension calls upon the manifest repository URI and retrieves the manifest; 5) Anti-spoofing check is performed using a visual fingerprint within the manifest; 6) The Content Credentials within the retrieved manifest are displayed next to the image within the page.

Watermarking to Complete the Triad

Fingerprinting alone is not practical for recovering metadata from many images. A browser displaying Content Credentials for each image on a web page would need to invoke a fingerprint lookup for every image found to lack metadata. This raises both scalability and privacy concerns. Using a passive method like fingerprinting, one cannot tell whether the metadata was stripped, or was simply never present.

An invisible watermark embedded within an image provides a convenient signal that a lookup is required. Moreover, watermarks can also carry a short payload in which an identifier may be embedded as a key to perform such a lookup. Watermark payloads are often on the order of a few hundred bits, due to the tradeoff between capacity and both imperceptibility and robustness. The former is critical to acceptance in creative use cases; the latter because the watermark must survive non-editorial transformations of the asset by content platforms that stripped the metadata. This also precludes directly encoding the Content Credentials.

Provenance is a non-steganographic use case for watermarking, where public detection and decoding is required on the client side. Recent studies have proven that invisible watermarks are strippable, using image reconstruction attacks [6]. Given access to a public client-side decoder, we may also expect that adversarial code will be run against the decoder to

spooft it by inserting fresh or copied watermarks into the image. Spoofing is often mitigated by encrypting watermarks, gating ability to encode or decode upon the knowledge of a shared secret. However in our use case, there can be no such secret as watermarks may be openly read or written by anyone. This places closed and open sourced watermarking technologies on a similar footing, since the oft-relied-upon defence of ‘security through obscurity’ is much reduced.

Spoofed watermark payloads may be countered by binding the payload to the content of the image using our fingerprint. One might ideally sign the fingerprint into the watermark payload using PKI, however robust watermarks lack the bit capacity. Fortunately, Content Credentials are signed via PKI, and so the fingerprint may equivalently be stored within the C2PA manifest referenced by the identifier within the payload.

Three core technologies for establishing media provenance are: metadata, fingerprinting and watermarking. Each taken alone presents its own vulnerabilities. Yet when carefully combined, they may mutually reinforce one another's strengths to produce a solid foundation for provenance.

The triad of watermarking, metadata and fingerprinting thus provides our solution (Figure 3). A watermark detector running on a client (such as a web browser extension) triggers a deterministic lookup over the network, using a watermarked identifier (ID). The watermark ID is used to query a manifest repository to retrieve a manifest. Upon receipt of the manifest, a visual check is performed by comparing the image fingerprint stored within the returned manifest to a fingerprint extracted from the query image. Should a watermark be spoofed to trigger the return of an unrelated manifest, the visual check will fail.

TrustMark

The expectation that watermarks may be openly stripped and spoofed implies only a limited risk, versus a major adoption upside, of an open-sourced watermarking algorithm that places the ability to encode and decode watermarks in the public domain. We have done this via TrustMark, a state of the art method for universal watermarking of arbitrary resolution images.

TrustMark [7] is a learned method that comprises an embedder, extractor and noise simulator module. The embedder is an encoder-decoder CNN, adapted from the content encoder of MUNIT [8] – a model more commonly used for the artistic stylization of images, where content preservation is also paramount. Given an image and bit string (payload), the embedder predicts a residual (watermark pattern) that is upsampled and linearly blended with the image at a user selectable strength. The embedder is trained adversarially with an extractor and detail preserving losses, that balance image quality via Peak Signal-to-Noise Ratio (PSNR) vs. original, with watermark recovery (bit accuracy) in the presence of complex noise simulation (18 sources, modelling typical image renditions); see Table 1.

$$\mathcal{L}_{total}(\mathbf{x}, \mathbf{y}) = \alpha \mathcal{L}_{quality}(\mathbf{x}, \mathbf{y}) + \mathcal{L}_{recovery}(\mathbf{w}, \bar{\mathbf{w}}) \quad (2)$$

where α controls the weighting between two loss functions, $\mathcal{L}_{recovery}$ and $\mathcal{L}_{quality}$. $\mathcal{L}_{recovery}(\mathbf{w}, \bar{\mathbf{w}})$ is a binary cross-entropy loss that measures the bit error between the embedded \mathbf{w} and extracted $\bar{\mathbf{w}}$ watermarks. $\mathcal{L}_{quality}$ measures the quality loss between the original (\mathbf{x}) and encoded (\mathbf{y}) image is defined as:

$$\begin{aligned} \mathcal{L}_{quality}(\mathbf{x}, \mathbf{y}) = & \beta_{YUV} \mathcal{L}_{YUV}(\mathbf{x}, \mathbf{y}) + \beta_{LPIPS} \mathcal{L}_{LPIPS}(\mathbf{x}, \mathbf{y}) \\ & + \beta_{FFL} \mathcal{L}_{FFL}(\mathbf{x}, \mathbf{y}) + \beta_{GAN} \mathcal{L}_{GAN+GP}(\mathbf{x}, \mathbf{y}) \end{aligned} \quad (3)$$

where β weights the four loss terms. $\mathcal{L}_{YUV}(\mathbf{x}, \mathbf{y})$ is the mean-squared error in YUV color coordinate

TABLE 1. TrustMark image watermarking [7] versus other watermark and steganographic baselines. Comparing imperceptibility (PSNR vs. original) and bit accuracy on the MetFace benchmark. 96% accuracy requires sacrificing 28% payload to BCH error correction; < 95% is impractical. A PSNR < 40 dB is typically unacceptable to creatives.

Method	Visual Quality (PSNR, dB)	Accuracy (bit-flip rate)
TrustMark	45.34 ± 1.33	0.96 ± 0.10
RoSteaLS	33.77 ± 2.37	0.93 ± 0.10
RivaGAN	40.98 ± 0.19	0.82 ± 0.14
SSL	42.84 ± 0.10	0.70 ± 0.13
StegaStamp	39.35 ± 1.57	0.70 ± 0.11
dwtDctSvd	41.14 ± 2.35	0.98 ± 0.08

space and $\mathcal{L}_{LPIPS}(\mathbf{x}, \mathbf{y})$ is a measure of the perceptual distance (using “Learned Perceptual Image Patch Similarity”) between \mathbf{x} and \mathbf{y} [7]. $\mathcal{L}_{FFL}(\mathbf{x}, \mathbf{y})$ is a focal frequency loss (FFL) to bridge the gap between \mathbf{x} and \mathbf{y} in the frequency domain, which in practice adds robustness by encouraging the watermark residual to fall within lower frequency bands, and $\mathcal{L}_{GAN+GP}(\mathbf{x}, \mathbf{y})$ is a discriminator loss that seeks to tell watermarked and unwatermarked images apart using the discriminator of a generative adversarial network (GAN) as well as introducing a gradient penalty (GP) to regularize training. TrustMark also co-trains a watermark remover module, with the embedder and extractor. Re-watermarking (erasing and encoding a fresh watermark) multiple times without loss is important to creative workflows where one may edit and export watermarked images iteratively, each time requiring a fresh watermark and C2PA manifest.

We apply Bose–Chaudhuri–Hocquenghem (BCH) error correction to the watermarked payload at a user-selectable level. Under BCH, 4 bit flips in 100 (96% accuracy) require 28 bits of parity to correct, leaving a space of 2^{72} unique identifiers. Trustmark enables this with high imperceptibility (45 dB PSNR); see Table 1.

Decentralized Watermark Lookup

Manifest recovery via the triad (Figure 3) relies upon a trusted manifest repository. It is impractical to assume a single repository for all the world’s images, and so federated lookup of multiple independent manifest repositories is desirable. Yet it is also desirable to have a single point of indirection from which to query.

Distributed ledger technology (DLT), colloquially ‘Blockchain’, offers one solution. DLT enables multiple

APPLICATIONS

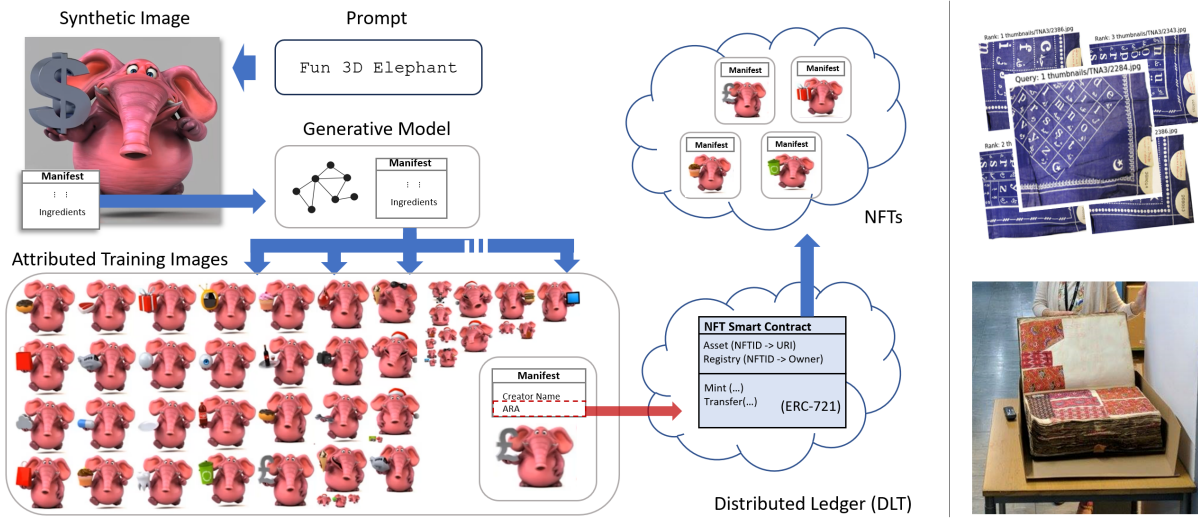


FIGURE 4. Fighting Image Misappropriation with Attribution. Left: Content Credentials (a C2PA manifest) within a synthetic image indicate the GenAI model as an ingredient. The model, in turn, indicates its training data as ingredients. Attribution is performed via fingerprint matching to find the training data responsible for the synthetic image. The Content Credentials within the attributed training images link (red arrow) to a Non-Fungible Token (NFT) collection (implemented by a NFT smart contract) enabling the owner of those responsible training images to be recognized and paid for the GenAI re-use of their image. Right: Fingerprinting was used in the ‘Deep Discoveries’ project for design provenance, to similarly to attribute design motifs in a textile collection to earlier uses of that design. Photographs taken by the author at The National Archives (UK). Used with permission.

independent parties that do not trust one another, to collaborate to produce a shared trusted datastore without recourse to any centralized point of trust. Originally developed as an append-only ledger to track who owns digital currency without any central bank, DLT may also be applied to non-cryptocurrency use cases.

We use Polygon DLT¹ to create an append-only store of key-value (K-V) pairs that map watermark IDs (the key) to a tuple (the value) describing both a unique C2PA manifest ID and an endpoint at which that manifest is available [9]. Polygon is one of several modern proof of stake (PoS) networks that offer low cost high throughput and do not require energy intensive mining processes to operate. In our implementation, $\sim 10^5$ K-V pairs may be written by a DLT node each second, at less than US\$0.001 per pair. DLT service providers such as Alchemy² offer high read throughput solutions that may be used by the browser extension to resolve the watermark ID, prior to calling on the relevant endpoint to obtain the manifest, and finally perform the visual check. In order to deploy the necessary watermarking models on the browser client,

neural compression and quantization are performed to reduce their size via the Open Neural Network eXchange (ONNX) framework.

Synthetic media provenance

Generative images are a form of derivative work created by sampling and remixing training images. In a sense, today's use of GenAI echoes music sampling in the early 1980s, where ad-hoc re-use was at first widespread but then matured into a framework for rights clearance and residual payments to artists. Today GenAI has given rise to similar legal challenges around consent and fair use, as well as broader calls for the improved recognition, and reward, of artists whose work has been used to train models. However the much greater scale of image re-use raises practical challenges. GenAI models are typically trained on millions, if not billions, of images very few of which materially influence the synthesis any single image.

Explaining the generation of a given image by identifying this influential subset of training data (referred to as **attribution**) is therefore a necessary step in recognizing and rewarding that contribution. This is in addition to identifying the responsible model and its training data. Content Credentials can help with

¹www.polygon.technology

²alchemy.com

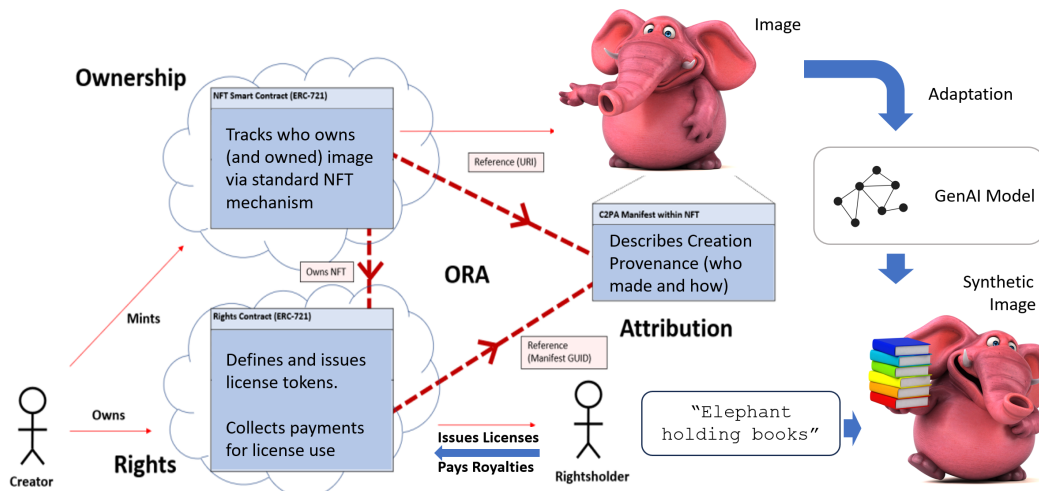


FIGURE 5. The Ownership-Rights-Attribution (ORA) triangle enables recognition and reward of creative contributions to GenAI [11]. ORA links creation provenance, via Content Credentials, with ownership provenance via Non-Fungible Tokens (NFTs). ORA extends NFTs by defining rights for image re-use, and issues of licences to exercise those rights via a smart contract. The tokens serve also as a means to collect royalties for the owner. Here an image is being used to adapt a GenAI model to create a custom generated image. ORA mediates a payment to the image owner for its use in GenAI.

both tasks. Models, like synthetic images, may bear Content Credentials, and so may be referenced as an ‘ingredient’, and in turn, reference their training data as ‘ingredients’ (Figure 4). Attribution algorithms then deduce which subset of those training images are responsible for the synthetic image. The Content Credentials of training images may then be used to obtain, if present, the identity or even financial details of their creators in order to pay a residual.

We have explored the task of GenAI attribution through both correlation and causative methods. Our fingerprinting and watermarking technologies played a role in each, respectively.

GenAI often memorizes **concepts** in training data, such as styles or motifs. These are often visually evident in the synthetic image. Matching (correlating) a synthetic image to training images is therefore one approach to attribution. Figure 4 shows how such a correlation has been calculated using fingerprinting, computed over densely sampled image patches from GenAI training data. We showed a similar approach to be effective in our recent ‘Deep Discoveries’ project ³, which searched a digitized portion of the UK Registered Design catalogue in the National Archives. Design provenance is an active research area in the

arts, and mapping the origins of motifs and styles as they are remixed over time is analogous to remixing learned concepts in GenAI.

Whilst correlation can approximate causation, the two are not equal. Access to images prior to their use in model training offers an opportunity to leverage watermarking to achieve a *causal* attribution. We have explored this by clustering our training image using our fingerprint embedding $\phi(\cdot)$ and applying a unique watermark to each cluster. Duplication in training data often leads to memorization of duplicated concepts, and their regurgitation in synthesized images. We have found that applying watermarks to training data clusters can also induce memorization and regurgitation of that watermark to a useful end – to attribute synthesized images to those training clusters. Although the method is not as granular as correlation methods (clusters need to contain a hundred or so images to provoke watermark memorization), initial results show causative attribution of up to 2^{12} training concepts within a GenAI model trained on 1 million images [10].

Creative Agency and Generative AI

Content Credentials describe the **creation provenance** an an asset (how an image was created, by who, using what) but are silent, by design, on the matters of ownership and rights. Yet these issues are

³tanc-ahrc.github.io/DeepDiscoveries

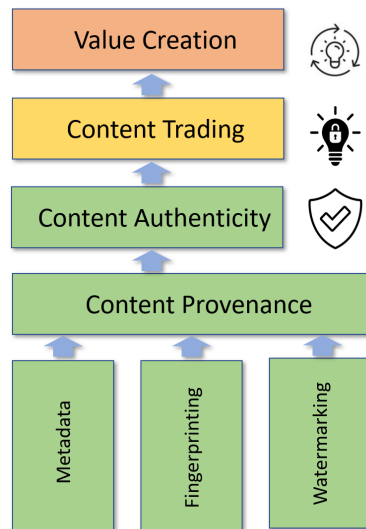


FIGURE 6. Provenance as a capability stack. A solid provenance layer underpinned by the three pillars (metadata, watermarking, fingerprinting) enables content authenticity. Authentic content enables value creation when coupled with ownership and rights, through decentralized creative markets. This may in turn unlock new value creation and ways to work.

particularly acute in the case of GenAI. Creatives often publish graphical work online (for example, to showcase portfolios and attract commissions) and may not consent to the use of that content for training models. One way to provide this agency is to combine C2PA with another provenance technology.

Non-fungible Tokens (NFTs) represent digital artwork as a token that can be bought and sold on a DLT. NFTs describe **ownership provenance** – who owns, and owned an artwork – and came to prominence during the global pandemic, which limited traditional monetization routes for digital artists. The value of a physical artwork is often defined, in part, by the history of asset ownership. It is one reason that the Mona Lisa hanging in the Louvre has value, and a perfect replica held elsewhere does not. *Catalogues Raisonné* are trusted physical ledgers recording the ownership provenance of artworks, maintained by authoritative sources such as the Wildenstein Institute. Digital ledgers (via DLT) do much the same for NFTs. However NFTs have serious limitations, not least the absence of any clear form of rights or licensing. For example, as a games publisher, if I purchase an animated character as an NFT, how can I be sure I have the right to use it in my game? This has led to NFTs being used mainly for digital speculation, rather than fulfilling their potential as a decentralized market for

re-usable creative assets.

We have developed an extension of NFT that unifies ownership, rights and attribution (ORA) to address this issue, and applied it to improve creative agency and compensation for GenAI [11].

ORA embeds Content Credentials within an asset – for example, a digital artwork – prior to its creation (‘minting’) as an NFT, so unifying creation and ownership provenance. ORA then creates digital tokens to represent licenses to use content, which may be issued or sold to rightsholders. Holders of these licence tokens can use them to issue a micropayment (residual) to the owner when that licence is exercised.

Figure 5 illustrates the end-to-end concept for a GenAI model that uses that digital artwork to adapt (customize) a GenAI model to create an image. This is common practice, using techniques such as Low Rank Adaptation (LoRA) [12] which enable zero-shot personalization of GenAI models to create images of specific people or products using an exemplar image. In this case the owner of the digital artwork has given consent for that image to be used as an exemplar, and receives a micropayment for its use at the time of generation. Coupled with our earlier described DLT framework to lookup (recover) the Content Credentials of assets via watermark or fingerprint lookup, ORA may be used to check for training consent and to pay creators when their images, even if circulating in the wild without metadata, are used for GenAI.

Conclusion

Generative AI is already transforming and democratizing computer graphics and visual storytelling. We must mitigate its associated risks by tackling misinformation, and ensuring that artists can express consent (and can be compensated) for use of their images in training. Provenance presents a foundational capability to address these risks, delivered through the three pillars of signed metadata, fingerprinting and watermarking. Provenance can enable users to contextualize trust decisions about content, helping to fight fake news. Building upon provenance and authenticity, provenance unlocks new opportunities to enable consent and monetization of GenAI through decentralized models for ownership, rights and attribution (Figure 6).

The creative economy is increasingly a decentralized, gig economy in which anyone can be a producer or consumer of content. These peer to peer engagements require strong provenance technologies to underpin the authenticity and value of images, as they are exchanged. As we enter an election year, it is encouraging to see growing adoption of provenance to

fight fake news and inform trust decisions on content. Yet authenticity is just the first step. To paraphrase Buzz Lightyear, provenance can take us "to Authenticity and Beyond!". It may ultimately be what underpins value in our future decentralized creative economy.

ACKNOWLEDGMENTS

We thank both the Content Authenticity Initiative (CAI) team at Adobe, and the DECaDE Centre at U. Surrey with whom this work was undertaken. We thank Simon Jenni, Shruti Agarwal and collaborators at Adobe Research, Tu Bui, and DECaDE PhD students Alex Black and Kar Balan. Deep Discoveries was funded by AHRC grant AH/T011114/1. This work was supported in part by DECaDE – the UKRI Centre for the Decentralized Creative Economy – under EPSRC grant EP/T022485/1.

REFERENCES

1. R. Po *et al.* "State of the Art on Diffusion Models for Visual Computing". *Comp. Graph. Forum* (Eurographics), September 2024.
2. S. Gregory *et al.*, "Ticks or it didn't happen," *Witness.org, Report*. December, 2019.
3. D. Kahneman, "Thinking fast and slow," Penguin. October, 2011. ISBN: 978-0374275631.
4. "Coalition for Content Provenance and Authenticity (C2PA) Technical Specification v1.4," Joint Development Foundation, Washington. <https://c2pa.org/>.
5. A. Black, *et al.* "Deep Image Comparator: Learning To Visualize Editorial Change," *Comp. Vision and Pat. Rec. (CVPR) W. Media Forensics*, IEEE/CVF. June 2021.
6. X. Zhao, *et al.* "Invisible Image Watermarks Are Provably Removable Using Generative AI," *arXiv preprint arXiv:2306.01953*, June 2023.
7. T. Bui, S. Agarwal and J. Collomosse. "TrustMark: Universal Watermarking for Arbitrary Resolution Images," *arXiv preprint arXiv:2311.18297*, November 2023.
8. X. Huang, *et al.* "Multimodal Unsupervised Image-to-Image Translation," *Euro. Conf. on Comp. Vision (ECCV)*, Springer. October 2018.
9. K. Balan, *et al.* "DECORAIT: DECentralized Opt-in/out Registry for AI Training," *Conf. on Visual Media Production (CVMP)*. ACM SIGGRAPH. December 2023.
10. V. Asnani, J. Collomosse, T. Bui, X. Liu and S. Agarwal. "ProMark: Proactive Diffusion Watermarking for Causal Attribution" *Comp. Vision and Pat. Rec. (CVPR)*, IEEE/CVF. June 2024.
11. K. Balan, *et al.* "EKILA: Synthetic Media Provenance and Attribution for Generative Art," *Comp. Vision and*

Pat. Rec. (CVPR) W. Media Forensics, IEEE/CVF. June 2023.

12. E. Hu *et al.* "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, June 2021.

John Collomosse is a principal scientist at Adobe Research, where he leads research for the Content Authenticity Initiative (CAI) and two cross-industry task forces within the C2PA open standards body for media authenticity. He is a professor at the University of Surrey, where he is the founder and director of DECaDE, the UKRI Research Centre for the Decentralized Creative Economy. He is a Chartered Engineer and holds a PhD in Computer Science from the University of Bath. Contact him at collomos@adobe.com

Andy Parsons is the Senior Director of the Content Authenticity Initiative (CAI), an alliance spearheaded by Adobe that brings together more than 2500 entities from industry, government, and academia. The CAI is dedicated to ensuring the provenance of content and is actively involved in developing open technologies to authenticate various types of media. Prior to his role at Adobe, Parsons held positions as a Founder, CTO, and VPE, achieving success in areas such as digital therapeutics, education technology, and hyperlocal news. He holds a BSEE in Electrical Engineering from Tufts University. Contact him at andyp@adobe.com.