EKILA: Synthetic Media Provenance and Attribution for Generative Art

Kar Balan[3] Shruti Agarwal[1] Simon Jenni[1] Andy Parsons[2] A. Gilbert[3] John Collomosse[1-3]
[1]Adobe Research [2]Content Authenticity Initiative (CAI)
[3]Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Recognizing and Rewarding Creatives

We present **EKILA** [1]; a decentralized framework that enables creatives to receive recognition and reward for their contributions to generative AI (GenAI).

EKILA combines **robust visual attribution** with a **content provenance standard (C2PA)** to address the problem of synthetic image provenance – determining the generative model and training data responsible for an AI-generated image.

EKILA extends **non-fungible tokens (NFTs)** – a decentralized way to track asset ownership via Blockchain (DLT) – to introduce tokenized rights, enabling **a triangular relationship** between the asset's **Ownership, Rights, and Attribution (ORA).**

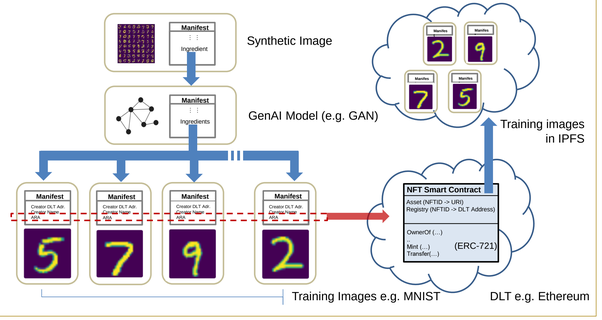
ORA enables creators to express **training consent** and, through our attribution model, to receive **apportioned credit, including royalties** payments for use of their assets in GenAI.

C2PA: Provenance for Synthetic Media

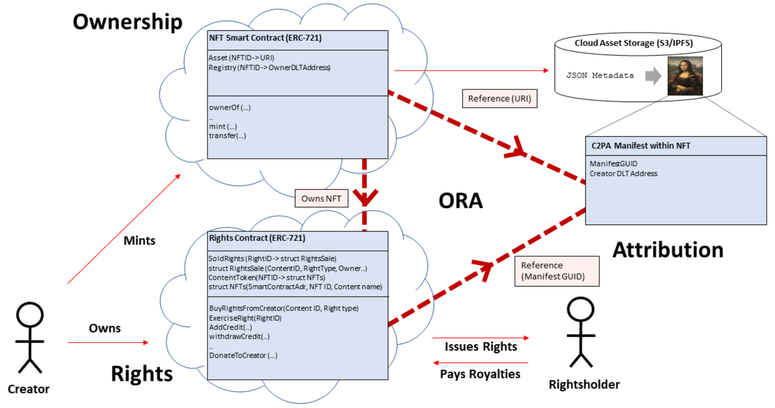
C2PA is an open emerging standard for media provenance [2]. C2PA describes facts about the **creation provenance** of an asset, e.g. who made it, how, and using **ingredient** assets. These facts are called **assertions** stored in a **manifest** in the **asset**. Manifests thus form a **provenance graph**.

We apply C2PA to describe **synthetic image provenance**: to describe within an image's manifest the GenAI model used to produce it and, within the GenAI model manifest, the ingredients used to train it.

Training assets are minted as NFTs. NFTs describe the ownership provenance of assets. EKILA **bridges creation provenance (C2PA) to ownership provenance (NFT)** using an assertion in each manifest.
Now we know **who owns** assets that contributed to the training of the model, and ultimately to the synthetic (GenAI) image.



ORA Triangle: Ownership, Rights, Attribution



EKILA introduces a way to assign rights to NFTs e.g. the **right to use an image to train GenAI**. Neither NFT nor C2PA contains mechanisms for specifying rights. We propose a smart contract --- the **Rights contract** -- to define rights and to distribute licenses to those rights, represented by rights tokens. Each creator manages a Rights contract, and issues tokenized rights to rightsholders via that contract.

The relationship between NFT, C2PA metadata in the asset, and Rights contract encodes the Ownership, Rights and Attribution (ORA) of the asset, **immutably bound in a triangular relationship**.

Visual Attribution

Our approach consists of 3 stages: 1) **partial matching based on image fingerprints** derived from dense multi-scale patches, 2) **pairwise verification** and scoring of attributed patches, and 3) apportionment based on a normalized **verification score** over verified matches.

Patchified Fingerprinting.
We adapt the fingerprinting approach in [3] to attribute patches using a contrastive learning objective.

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

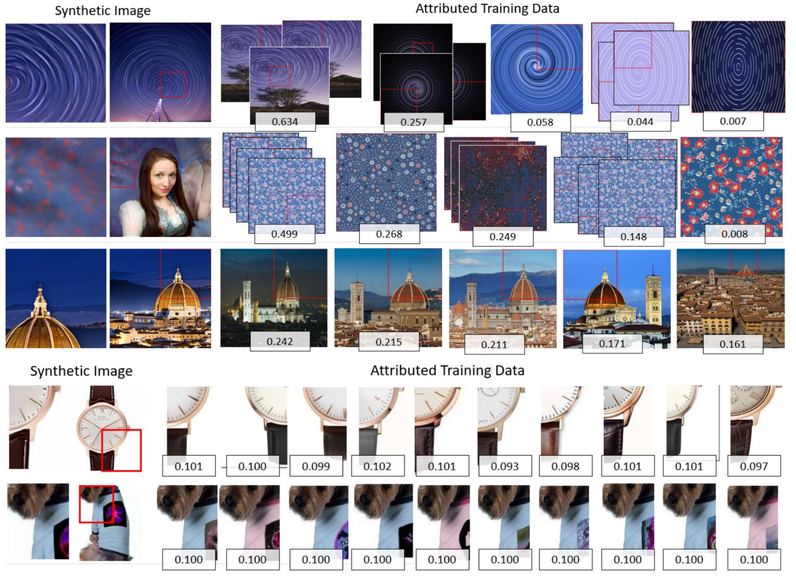
Where $\hat{\phi}_i$ is a set of augmentations applied to modelling GenAI appearance variation seen in near-memorization cases, $\mathcal{B}$ is a large random mini-batch and similarity is $d(a,b) := \exp\left(\frac{1}{\lambda}\frac{a^T b}{\|a\|_2\|b\|_2}\right)$

Pair-wise Verification.
We verify the top-k retrieved patches, comparing spatial feature maps $F_q \in \mathbb{R}^{H\times W\times D}$ from fingerprinting query patch x_q and $\{F_i\}_{i=1}^k$ be the k result maps. We process each feature map with a 1×1 convolution and extract various GeM-pooled [4] descriptors

$$\hat{F}_q = [f^q_{w_1},\ldots,f^q_{w_{|\mathcal{W}|}}] \in \mathbb{R}^{|\mathcal{W}|\times\frac{D}{4}}$$

A feature correlation matrix $C_{qi} = \hat{F}_q\hat{F}_i^T \in \mathbb{R}^{|\mathcal{W}|\times|\mathcal{W}|}$ passes to an MLP to predict verification score.

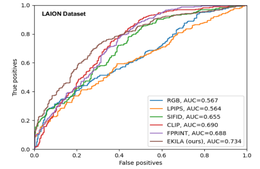
Visualization: Visual Attribution and Apportionment



Evaluation and Conclusion

We evaluate over two diffusion models trained on LAION-400M and on Adobe Stock. We show in both cases our attribution model to significantly outperform ViT-CLIP and recently proposed patch driven metrics for measuring attribution [5,6]. Visuals of each model are above, ROC curve below.

Dataset	Method	Image Attribution R@1↑	Image Attribution R@5↑	Image Attribution R@10↑	Image Attribution mAP↑	Patch Attribution R@1↑	Patch Attribution R@5↑	Patch Attribution R@10↑	Patch Attribution mAP↑
LAION	EKILA (ours)	**83.33**	**54.31**	**52.94**	**59.26**	**87.50**	**70.13**	**70.83**	**72.73**
	ViT-CLIP	56.25	47.92	43.75	47.61	75.00	57.64	56.02	59.91
	RGB-P	5.88	4.71	3.26	4.27	12.6	3.32	3.12	4.38
	ALADIN	37.39	33.74	18.29	28.87		—		
IPF-Stock	EKILA (ours)	**70.83**	**65.20**	**63.16**	**64.72**	**86.79**	**67.36**	**66.67**	**70.03**
	ViT-CLIP	60.22	58.60	51.61	58.38	67.74	62.9	60.93	63.10
	RGB-P	3.58	2.69	0.27	1.85	19.4	14.8	12.5	15.1
	ALADIN	29.03	23.23	21.86	23.50		—		



References

[1] Horniman Museum London. EKILA: The rules of sharing, 2014.
[2] Coalition for Content Provenance and Authenticity: c2pa.org
[3] Black et al. Deep image comparator… CVPR WMF 2021.
[4] Tolias et al. Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879.
[5] Carlini et al. Extracting training data from diffusion models. arXiv:2301.131188.
[6] Somepalli et al. Diffusion art or digital forgery? Investigating data replication in diffusion models. arXiv:2212.03860.

Funding Acknowledgement

Implementation of the EKILA prototype was supported in part by DECaDE: UKRI Centre for Decentralized Digital Economy under EPSRC Grant EP/T022485/1.