Authentication of Art NFTs

Lior Hirschfeld (liorh@mit.edu)
Zhenbang Chen (zhenbang@mit.edu)
Zhenjia Chen (zhenjia@mit.edu)

Abstract

Non-fungible tokens (NFTs) have experienced a surge in popularity as a way to represent ownership of some idea or item. They have already been used to facilitate transactions involving music, event tickets, virtual video game assets, and artworks. As the price and value of NFTs skyrocket, it becomes increasingly important that there is a consensus on stable definition(s) of legitimacy for these digital assets. This paper proposes a decentralized trust network system for verifying NFTs using a chain of authentication methods ranging from automatic machine checking to manual expert curation.

1 Introduction

1.1 Background

In late 2017, a studio named AxiomZen launched CryptoKitties, a distributed application that allowed users to purchase digital cats. At the time, CryptoKitties received a lot of traffic for its use of NFTs stored in Ethereum's ledger to track the ownership of each cat. Each of these tokens was issued directly by AxiomZen and identified a crypto kitten's name and unique properties in associated metadata. Owners could then sell their tokens directly or use them to breed new crypto-kitties, all using Ethereum smart-contracts. Within just a few days of release, users spent over a million dollars, with some individual cats selling for over $100,000.[1]

Over the past few months, these non-fungible tokens have begun to garner mainstream attention. Unlike cryptocurrencies, like Bitcoin and Ether, which have ample liquidity, each NFT represents a unique concept and is very limited in supply. They have already been used to facilitate transactions involving ideas and items like music, event tickets, and virtual video game assets.[2]

One of the most common applications of NFTs is in the monetization of digital artwork. These images, whose full resolution is often available online, may be freely copied and distributed. However, NFTs, whose scarcity may be directly controlled by an artist, can be sold to designate

[1] Fitz Tepper, "People Have Spent over $1M Buying Virtual Cats on the Ethereum Blockchain," TechCrunch
[2] Kevin Roose, "Buy This Column on the Blockchain!," The New York Times

certain individuals as “owners.” Depending on the creator’s desires, an artwork’s NFT may convey certain redistribution rights, but it might also just serve as a badge of authenticity.

Recently NFTs for digital artwork have reached record prices. An NFT for “Everydays - The First 5000 Days,” a piece by Beeple, sold for over $69,000,000 at auction.[3] This is an impressive figure for any piece of art, physical or virtual, and reflects the high confidence of some consumers in a NFT’s proof of ownership.

1.2 Problem

NFTs provide a few guarantees which make them appealing for the use cases outlined above. Most notably:

- An NFT’s metadata, written on the blockchain, cannot be modified without compromising the entire ledger.
- Ownership of an NFT cannot change without access to the private key of it’s current holder’s wallet.

However, these promises obscure deeper problems in the schema which might result in disputed ownership claims. These arise from the fact that the legitimacy of a particular NFT’s claim is not well defined. “Legitimacy” is itself very difficult to define, but here we use it in its most abstract sense: Loosely, an NFT is legitimate if there is agreement between what the token’s owner claims it represents and the consensus opinion about what that token represents.

Disagreements are common and inevitable. NFTs may be issued by anyone, and there is no guarantee that an issuer has rights for the artwork in question. Even if they do, further questions remain: After selling an NFT for their artwork, an artist might easily produce a new one and discredit the token they’ve already sold. In that case, who is the proper “owner” of the artwork?

To answer these questions, most consumers currently look to centralized markets, which verify the authenticity of each token before issuance. For example, Foundation, an Ethereum-based NFT marketplace specializing in trading artworks, relies on a team of “curatorial staff” that is in direct contact with the artists that claim to have created the available art pieces. In this case, trusted and reputable curators attest to the authenticity of any NFTs that are permitted to be traded on their marketplace. However, this reliance on centralized authorities runs counter to the distributed nature of NFTs and introduces potential failure points: an authority might back a fraudulent ownership claim or suffer an attack which leaves their historical records compromised. Without access to this metadata, authentic and counterfeit tokens become

[3] Scott Reyburn, “JPG File Sells for $69 Million, as 'NFT Mania' Gathers Pace,” The New York Times

effectively indistinguishable. A better solution would introduce a distributed mechanism for this decision making process.

Legitimacy is also non-binary and depends on context. Consider, for example, a famous photograph. There may simultaneously exist an original print, limited prints, and further reproductions. The last of these might be totally appropriate in a personal collection, but might not be appropriate for some museum exhibitions. An ideal solution for classifying and authenticating NFTs would recognize these nuances and allow different interested parties to function under different definitions of "legitimacy."

1.3 Design Goals

Introducing a rigid definition of legitimacy in this novel space is unlikely to succeed: The infrastructure which supports NFTs is changing rapidly, as is the perception of digital assets among consumers. Instead, our primary goal is to introduce a system which incentivizes both buyers and sellers, through popular consensus, to converge on a few, stable definitions. Consensus is important because it means the value of a token should be relatively consistent across communities, and stability is important because it means the value of the token should be relatively consistent over time. We hope that this will instill confidence in consumers that valuable tokens they purchase today will not be discarded as immaterial fakes at some point in the future.

Additionally, we propose a few methods of authentication, ranging from automated to manual, that might reasonably fit into a consensus definition of legitimacy.

2 Design

2.1 Existing Systems

Art authentication is a uniquely challenging task that revolves around providing a convincing legitimacy argument. For a new physical work of art, an artist can offer prospective buyers some form of evidence to establish that they are the author and rightful owner of the piece. This could include presenting concept/wireframe renderings, having recorded time lapse sessions, or simply claiming authorship and relying on the fact that accurately replicating physical artwork is impractical for many mediums. These methods rely on supplemental materials that only the true artist can readily provide to verify the initial ownership and authenticity of an artwork. While this authentication approach is not foolproof since evidence could be forged, it is typically sufficient to deter most false claimants.

In the case of well-known artworks by famous past artists, resellers may consult experts to stake their reputation on a particular piece's legitimacy. From a buyer's perspective, if the evidence is convincing, the relevant experts are trustworthy, and everyone else believes the piece is real, then the artwork is likely authentic. In this way, there is a general informal "consensus" regarding each artwork's legitimacy.

Within the context of digital artworks, there are additional considerations for authentication to work. While minting and trading a simple NFT of a digital artwork gives an unforgeable history tracing back to a creator address on the blockchain, it does not solve the issue of verifying that the token creator actually owns the rights to the piece. Any user can mint and sell illegitimate NFTs associated with existing, publicly accessible artworks. Currently, there are several large exchanges, such as Foundation[4] and SuperRare[5], that serve as authenticators for new NFTs. In this system, a private organization maintains a team of curators that work directly with artists interested in selling their artworks as NFTs. Once verified, the organization mints NFTs on the artist's behalf or provides their seal of approval for the artist to create their own verified NFTs. This centralized, middle-man approach to NFT authentication relies on user trust in the platform and its curators.

While centralized and curated marketplaces do allow NFT buyers to be reasonably confident that they are trading for legitimate artworks, there are still drawbacks to this approach. Marketplaces face a conflict of interest since they stand to profit from authenticating and selling illegitimate artworks. Another issue centers around the balance between openness and accurate verification. In order to achieve high verification accuracy and to remain profitable, marketplaces like SuperRare choose to only work with artists that already have large followings and lengthy sales histories[6]. This creates an exclusive club of established digital artists that are approved by private organizations to sell legitimate NFTs, which runs counter to the decentralized nature of blockchain technology. On the other end of the spectrum, some platforms, like OpenSea, forgo verification altogether in favor of an open marketplace where any user can mint and trade NFTs through their interface[7]. Here, the onus is on the buyer to make sure they are purchasing a legitimate artwork NFT.

[4] https://foundation.app/
[5] https://superrare.co/
[6] "Get On Our Radar," SuperRare Help Center
[7] https://opensea.io/

2.2 Trust Network

With these concerns in mind, we propose a trust network system that provides a distributed approach to NFT authentication. By using a distributed system, we hope to mitigate the downsides of the current authentication implementations employed by centralized exchanges. This distributed concept is similar to other cryptocurrencies such as Stellar in the form of trustlines and can be implemented as an application on top of the NFT blockchain. The main idea behind the trust network is to delegate the digital artwork authentication process to multiple agents.

As mentioned, the NFT exchange, Foundation, performs artwork authentication by maintaining a team of professional curators who work directly with artists to ensure originality. Foundation operates on an invite-only basis in regards to which creators and artists they allow on to their platform. In this way, Foundation ensures that all artworks sold on its marketplace are created by individuals that meet its vetting standard and are unlikely to produce unauthentic works.

A trust network approach to security is not new to the cryptocurrency scene. The digital payment system Stellar employs a notion of trustlines and anchors to exchange real world assets and provide credibility to said assets. These anchors serve a similar function to traditional financial institutions such as PayPal or Venmo in that they accept deposits in return for Stellar tokens representing that deposit. The anchor also honors withdrawals, accepting Stellar tokens in exchange for the original asset. More pertinent to our NFT trust network is the notion of trustlines that emerge from these anchors. To begin exchanging assets on Steller, a user has to specify which assets they trust and from which anchors. This is done by adding trustlines, which correspond to different tokens, to their Stellar wallet. Once an user has specified their trustlines, they are able to freely trade those assets on Stellar.[8]

For our authentication system, we will take a similar approach to Stellar's trustlines. Within the trust network, users will declare sources of trust. These sources of trust can be individual NFTs that the user believes are authentic or other users that the user relies on to authenticate artwork. This will effectively divide the trust network into three roles: seller, authenticator, and buyer. The goal of a seller is to get their NFTs verified by authenticators. Once verified, other users will be able to check which authenticators carried out the verification process. A buyer's role will be to identify which authenticators they trust to authenticate NFTs. Now, a prospective buyer of a new NFT will be able to cross reference the NFT's authenticators with their own list of trusted authenticators to determine whether or not to believe the NFT's originality. With this network of trust, the authentication system no longer relies on centralized marketplaces to curate art NFTs.

[8] Kolten, "A Guide to Trustlines on Stellar," Medium

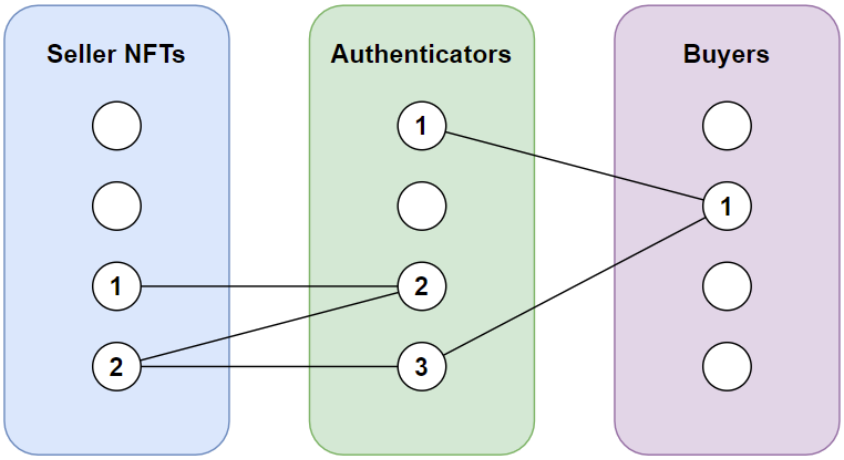

Figure 1. Trust network for authentication. Nodes on the network take on one of three roles. Seller NFTs are verified by Authenticators. Buyers trust certain Authenticators and the NFTs they have verified. In the figure, Buyer 1 trusts Authenticators 1 and 3 who have only verified NFT 2. Therefore, Buyer 1 believes NFT 2 is authentic.

In practice, each authenticator will be able to define their own metrics for what they consider to be authentic pieces and what they consider to be different or fall within the realm of parody. One authenticator may, for example, claim that recolorings of an artwork to be derivative works, while another may consider them to be new and unique expressions. Additionally, it is also possible for individual authenticators to gain reputation based on how many buyers trust them and how long they have been verifying NFTs. This means that more experienced authenticators with a larger buyer following will be ranked higher and be evaluated as more trustworthy when compared with newer ones.

2.3 Authenticating Artwork

Since authentication requirements vary between different contexts, no single artwork verification method will be appropriate for every use case. It is impractical to consult an curation expert to determine if every newly published art piece is a pixel perfect copy of a previously published and verified artwork already written to the blockchain. Likewise, automatic schemes cannot catch certain derivative artworks that are intentionally modified to elude machine detection. Instead, we suggest authenticator nodes use a chain of authentication methods that incorporate automatic duplicate artwork detection and manual review depending on the degree of verification required. We rely on the assumption that there is a correlation between the difficulty of distinguishing derivative artwork and the effort required to produce such art. In general, derivative artwork that require sophisticated/costly means to identify are typically harder to produce and will make up the minority of legitimacy disputes.

2.3.1 Perceptual Hashing

When deciding the legitimacy of a new artwork NFT, an authenticator node should first use an automatic detection scheme to filter out obvious duplicates and minimize unnecessary manual reviews. For image artworks, perceptual hashing can provide detection of artworks derived from previously published NFTs. In contrast to cryptographic hash functions, perceptual hash functions produce analogous outputs for images with similar features.

In the context of art NFT authentication, an authenticator will first generate the perceptual hash of the submitted artwork. Next, the authenticator compares this hash to the hash values of similar artworks that it has previously encountered and stored in an optimized data structure, such as vantage point tree. The specific comparison operation will vary depending on the perceptual hash algorithm used, but typically involves a simple computation such as the cross correlation of the two hashes. Finally, if the submitted artwork closely correlates with an existing piece above a predefined threshold, they decline to authenticate the submission. Additionally, the system could define a second comparison threshold for image pairs that are similar enough to warrant a manual review to determine legitimacy and not outright reject.

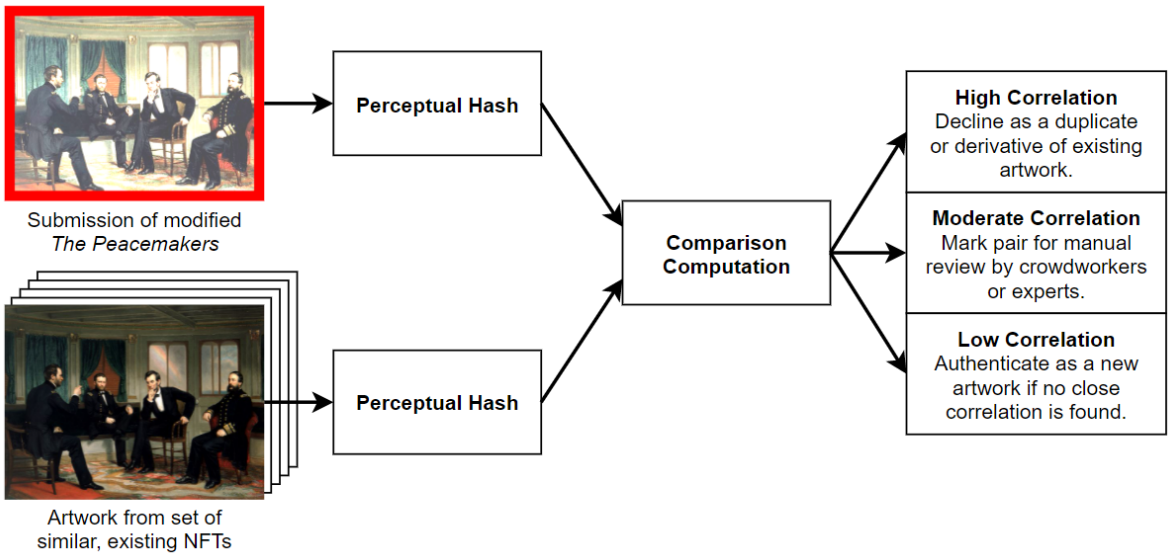

Figure 2. Procedure for automatic duplicate detection by comparing the perceptual hash of a modified submission against existing artworks. Depending on the authenticator's needs, any single thresholded outcome (i.e., high, moderate, or low correlation) can be omitted.

Each authenticator's unique definition of "legitimate artworks" will inform the parameters used to settle on the final legitimacy verdict. For example, an authenticator may advertise that they consider recolorings to be new distinct artworks. In this case, they may use a color-sensitive perceptual hash algorithm with a higher correlation threshold to reduce false positives. A robust authenticator system can make use of multiple specialized perceptual hash algorithms to handle

different types of image edits. These include Marr-Hildreth hash for edge features, color moment hash for color distributions[9], and radial hash and block mean hash for general image features[10][11].

2.3.2 Crowdworker Voting

Automatic methods like perceptual hash comparisons are incapable of handling every authentication case. In addition to machine verified checks, authenticators may employ human crowdworkers to settle questions of legitimacy with a system like Mechanical Turk[12]. This approach would be most effective in refining cases where the submitted artwork is a heavily modified derivative of an existing NFT and was flagged for review by an automatic duplicate detection process earlier.

Figure 3. Image pairs with moderate perceptual hash correlation flagged for manual review using block mean and color moment hash. The left pair is likely resolved as a valid parody work. The right pair is likely resolved as an illegitimate derivative.

The procedure will involve creating a task that asks human crowdworkers to identify if a newly submitted artwork can be considered an illegitimate derivative of another similar, published piece flagged by the automatic scheme. The task can also be augmented to provide results more in line with the authenticator's specific definition of legitimacy. For example, the authenticator may consider obvious parodies of existing art to be original work and instruct crowdworkers to ignore such examples. If applicable, crowdworkers can also be asked to consider specific licensing terms of the existing art that may allow certain derivative works.

---

[9] Tang, Zhenjun, Yumin Dai, and Xianquan Zhang. "Perceptual hashing for color images using invariant moments." Applied Mathematics & Information Sciences

[10] De Roover, Cédric., Christophe De Vleeschouwer, Frédéric Lefebvre, Benoît Macq. "Robust image hashing based on radial variance of pixels." IEEE International Conference on Image Processing

[11] Yang, Bian, Fan Gu, Xiamu Niu. "Block Mean Value Based Image Perceptual Hashing." 2006 International Conference on Intelligent Information Hiding and Multimedia

[12] https://www.mturk.com/

2.3.3 Expert Reviews

When even crowdworker voting fails to resolve a legitimacy dispute, the authenticator may turn to the advice of experts as the final verdict. Authenticators can defer a decision to an expert if the crowdworker voting does not reach a prespecified consensus threshold. This step is similar to the current method used by the centralized exchanges such as Foundation who maintain a team of curators to verify and authenticate new NFTs. Here, authenticators can follow conventional verification methods used for physical artworks like directly requesting authorship or ownership evidence from the NFT sellers.

2.4 NFT Metadata Standard

Currently, there is significant variance in the type of metadata associated with artwork NFTs depending on the artist and marketplace. To facilitate authentication, we propose a standardized format for useful information to include with NFT metadata:

- General artwork properties (e.g., author, title, description)
- Link to artwork on distributed storage (IPFS)
- Image cryptographic hash
- Image perceptual hashes (e.g., block mean hash, Marr-Hildreth hash)
- Licensing terms (e.g., CC-BY, CC-ND)

3 Evaluation and Discussion

3.1 Trust network

The most obvious disadvantage of the distributed trust network authentication system when compared with centralized marketplaces is the ease with which new authenticators can be created. As the NFT verifiers in the trust network are simply users who specialize in granting legitimacy to new NFTs, it is fairly easy to set up a new user as an authenticator and subsequently claim to trust unauthentic NFTs.

The untrustworthy authenticator problem is similar to how there can be multiple anchors for a single asset, such as USD, on Stellar. The majority of these USD asset issuers are unreliable with a low reputation, but this is not a significant problem for Stellar as the system still works even when there are only a handful of trustworthy anchors. This is also the case with our trust network implementation for authenticating NFTs. As long as a number of reliable verifiers exists, the distributed authentication will work with just those authenticators. Authenticators which make reasonable decisions are more likely to retain trust and can be distinguished by their popularity. A possible extension to this system would be to provide an economic incentive for

authenticators to act in the best interest of buyers. One option would be to introduce a small, regular payment made by any buyer to maintain a trustline with an authenticator.

With this fee in place, authenticators would be encouraged to retain all established trust lines and convince new buyers that their method of authentication is sensible. Creators would be incentivized to construct NFTs that will be approved by popular authenticators, growing the desirability of their contract, but would be free to sell the NFT on any platform of their choice. Finally, buyers would be incentivized to purchase contracts that were approved by popular authenticators, in case they ever wish to resell their purchase.

3.2 Authentication methods

We proposed a chain of authentication methods that authenticator nodes could follow to determine the legitimacy of a new artwork NFT. This approach allows authenticators to balance accuracy and performance considerations. Automatic schemes will offer faster duplicate detection for simple cases while manual review will deliver accurate results for appropriate scenarios.

Deduplication with perceptual hashing has a significant performance advantage over the current manual verification used by exclusive marketplaces. Based on testing, we also found that even significant simple image modifications, such as cropping over half of the original image or inserting large random blocks of color, were detectable with perceptual hash comparisons. This gives us confidence that this automatic scheme can detect many common instances of derivative art. However, a system using perceptual hashes will still be vulnerable to gradient-based adversarial attacks that produce subtly modified images with unrelated hash outputs. While impossible to prevent entirely, these attacks can be mitigated with the use of multiple perceptual hash algorithms to match images[13].

Authentication methods involving manual review incur significantly higher costs, but offer better accuracy than automatic systems. Considering the difficulty of fooling human crowdworkers or art experts, we expect that manual reviews could offer certain NFT artworks a level of legitimacy assurance similar to what their physical counterparts have. Nevertheless, systems involving human input will likely be practical in scenarios where the artwork price justifies the additional verification cost, such as the sale of Beeple's "Everydays - The First 5000 Days."

[13] Dolhansky, Brian, Cristian C. Ferrer. "Adversarial collision attacks on image hashing functions." Cornell University

4 Conclusion

As NFTs become increasingly prevalent, the need to rethink its implementation and standards for longevity will become more important. By using a distributed verification system alongside a more rigorous authentication methodology, our proposed scheme aims to capture the decentralized spirit of cryptocurrencies while still maintaining a high degree of long-term legitimacy for NFTs.

5 References

De Roover, Cédric., Christophe De Vleeschouwer, Frédéric Lefebvre, Benoît Macq. "Robust image hashing based on radial variance of pixels." *IEEE International Conference on Image Processing,* Sept. 2005.

Dolhansky, Brian, Cristian C. Ferrer. "Adversarial collision attacks on image hashing functions." *Cornell University,* Nov. 2020.

Kolten. "A Guide to Trustlines on Stellar." Medium. A Medium Corporation, October 8, 2019. https://medium.com/stellar-community/a-guide-to-trustlines-on-stellar-8bc46091a86f.

Reyburn, Scott. "JPG File Sells for $69 Million, as 'NFT Mania' Gathers Pace." The New York Times. The New York Times, March 11, 2021. https://www.nytimes.com/2021/03/11/arts/design/nft-auction-christies-beeple.html.

Roose, Kevin. "Buy This Column on the Blockchain!" The New York Times. The New York Times, March 24, 2021. https://www.nytimes.com/2021/03/24/technology/nft-column-blockchain.html.

Tang, Zhenjun, Yumin Dai, and Xianquan Zhang. "Perceptual hashing for color images using invariant moments." *Applied Mathematics & Information Sciences,* April 2012.

Tepper, Fitz. "People Have Spent over $1M Buying Virtual Cats on the Ethereum Blockchain." TechCrunch. Verizon Media, December 3, 2017. https://techcrunch.com/2017/12/03/people-have-spent-over-1m-buying-virtual-cats-on-the-ethereum-blockchain/.

Yang, Bian, Fan Gu, Xiamu Niu. "Block Mean Value Based Image Perceptual Hashing." *2006 International Conference on Intelligent Information Hiding and Multimedia,* Dec. 2006.