Deepfake

CS463/ECE424

University of Illinois

Outline

- Deepfake and Abusive Use
- Defense
  - Detection-based Method
  - Provenance-based Method

Deepfake in Practice

There's Something Fishy About Amazon's Anti-Union Twitter Army [Updated]

https://www.technologyreview.com/2021/03/31/1021487/deepfake-amazon-workers-are-sowing-confusion-on-twitter-thats-not-the-problem/

Deepfake in Practice

There's Something Fishy About Amazo
Twitter Army [Updated]

https://www.technologyreview.com/2021/03/31/1021487/deepfake-amazon-w

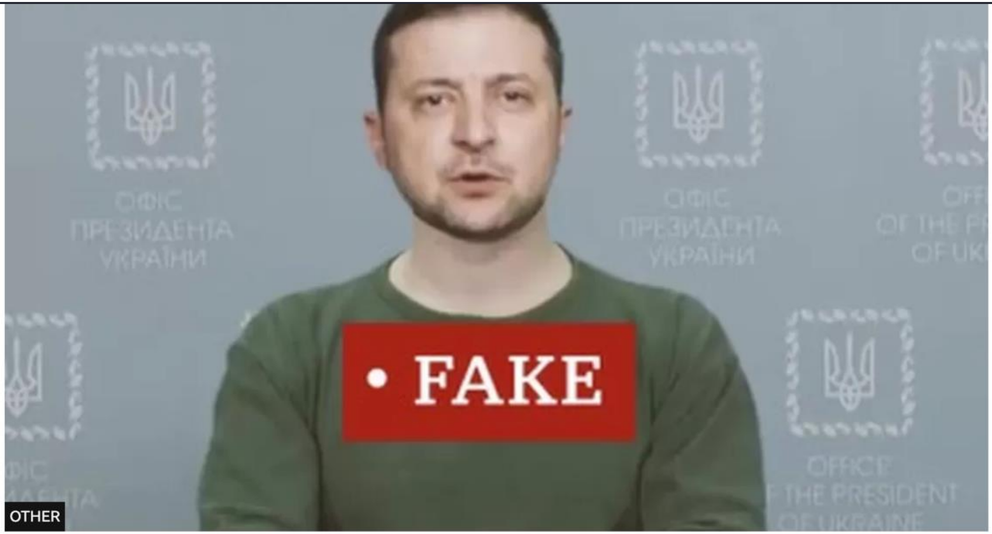
NEWS

Home | War in Ukraine | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech | Science

Tech

Deepfake presidents used in Russia-Ukraine war

Meanwhile, this week Meta and YouTube have taken down a deepfake video of Ukraine's president talking of surrendering to Russia.

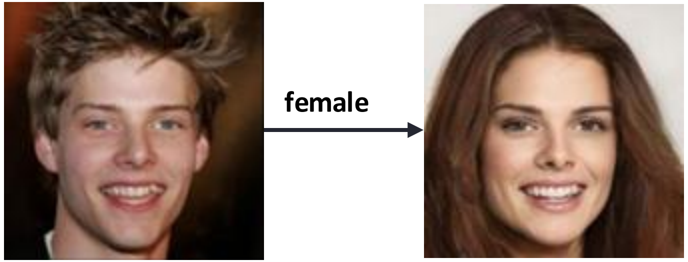


The deepfake appeared on the hacked website of Ukrainian TV network Ukrayina 24

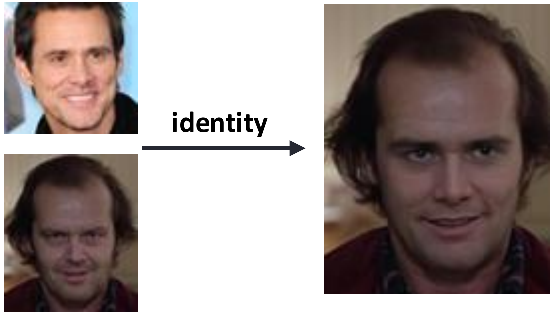
Background - Visual Deepfake Taxonomy

Expression Manipulation

Attribute Manipulation



Identity Swapping



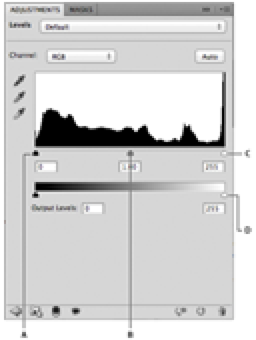
Full Face Synthesis

Face Swapping: What People Do in the Past

Input Image

Select Candidate Images

Similar in appearance and pose



Alignment, relight recolor

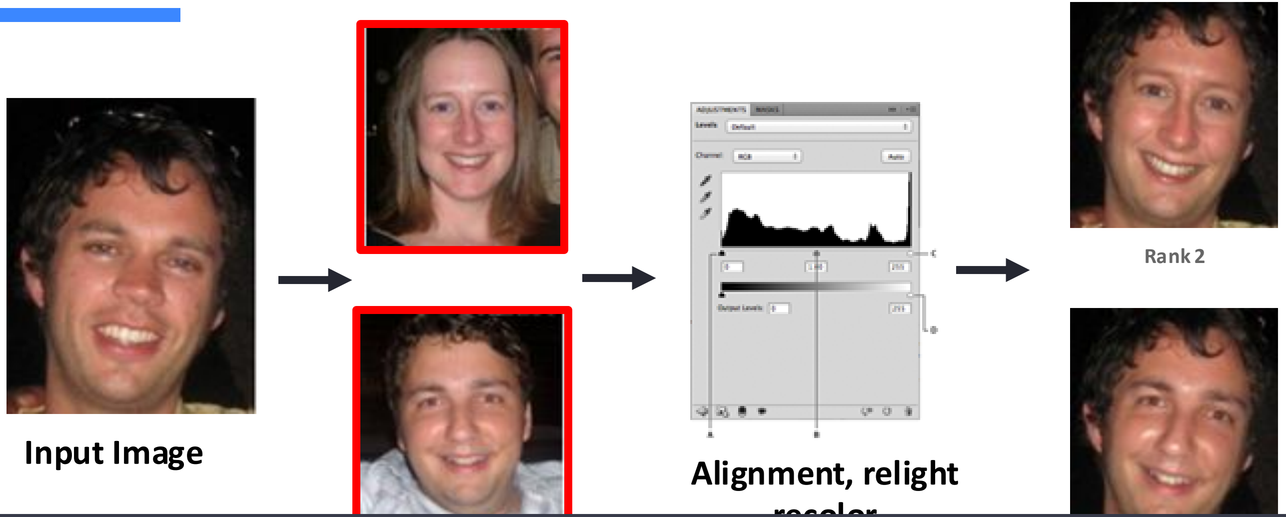
Rank 2

Rank 1

Ranked Images

Korshunova, Iryna, et al. "Fast face-swap using convolutional neural networks." ICCV, 2017

Face Swapping: What People Do in the Past



Limitations

- Candidate image have fixed facial expressions
- Unable to specify a certain target identity

Face Swapping via Style transfer

- Intuition: learn to transfer the style of an image based on another *reference image*

Face Swapping via Style transfer

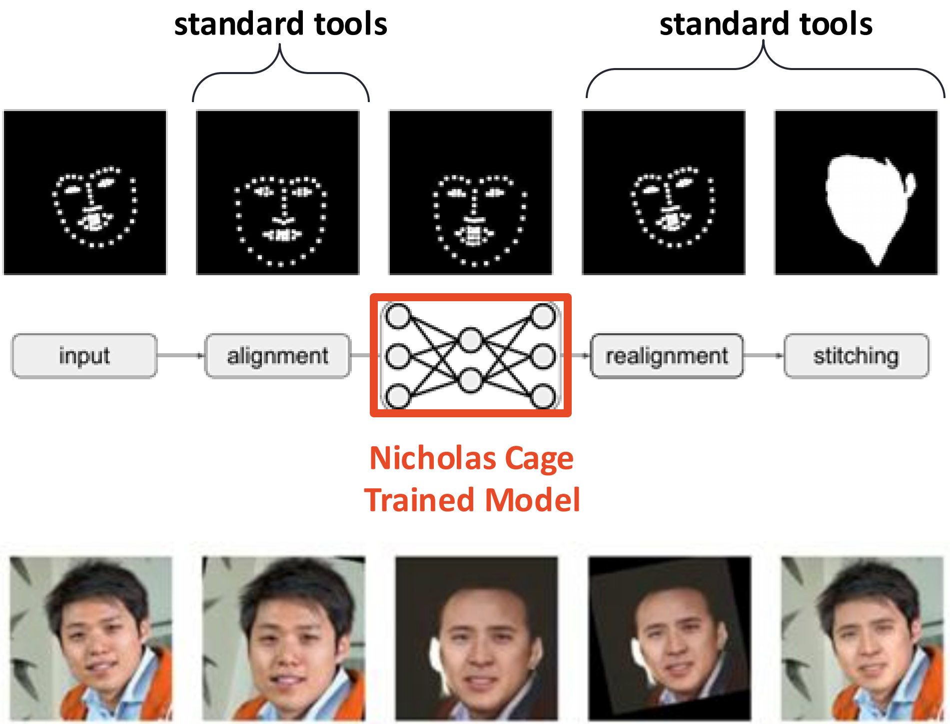
- Intuition: learn to transfer the style of an image based on another *reference image*

Identity/style: Y

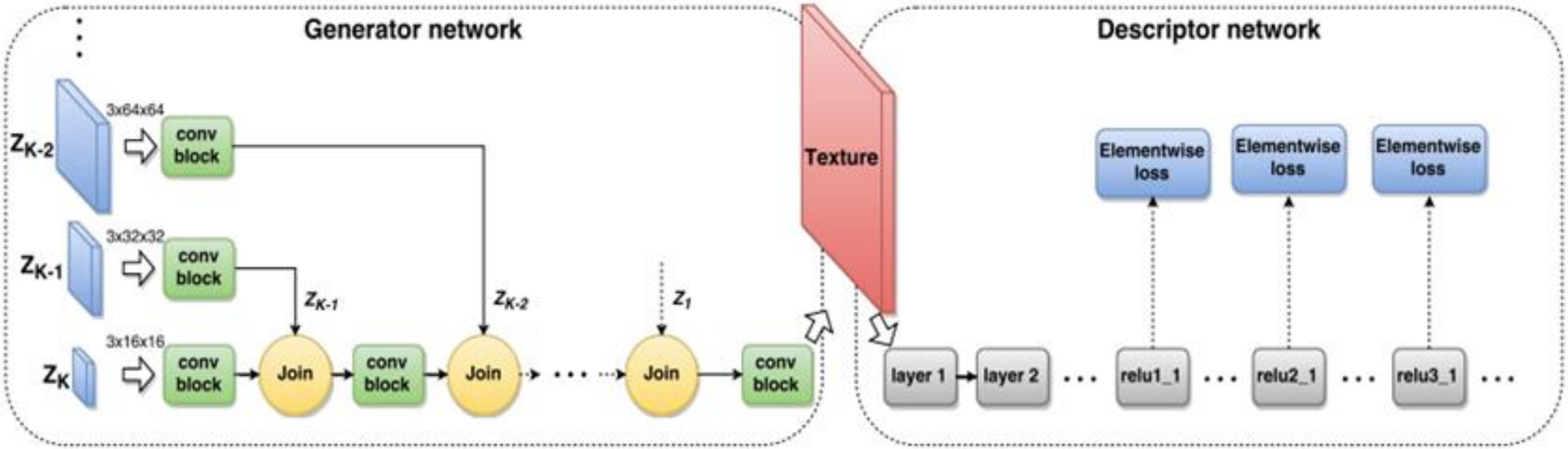
Content: x
(pose, impression)

Output: $\hat{x}$

Face Swapping Pipeline



Face Swapping Architecture



Multi-Scale Generative CNN

- **Goal:** Generate $\hat{x}$ for trained identity Y given content image x

Pre-Trained Discriminative CNN

- **Goal:** Provide loss needed to train generator
- **Latent Representation**:
  - **Lower Layers:** Textures, lines
  - **Upper Layers:** Objects, Structure

Optimize with Four Types of Loss

Y = Trained Identity

x = Content

$\hat{x}$ = Output

Content Loss: compare $\hat{x}$ and x

Style Loss: compare Y and $\hat{x}$

Light Loss: compare Siamese representation of x and Siamese representation of $\hat{x}$

Smooth Loss: penalize large color changes near each pixel of $\hat{x}$

$$\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{Y}) = \mathcal{L}_{content}(\hat{\mathbf{x}}, \mathbf{x}) + \alpha \mathcal{L}_{style}(\hat{\mathbf{x}}, \mathbf{Y}) + \beta \mathcal{L}_{light}(\hat{\mathbf{x}}, \mathbf{x}) + \gamma \mathcal{L}_{TV}(\hat{\mathbf{x}})$$

Face Swapping Results

With proper conditions, Decent!

- Often too smooth
- Skin tones often are off

Why Important?

- First automated method for targeted identity swapping
- Spawned a series of generative and defensive works

More Recent Results

FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping, CVPR 2020

Deepfake Detection based on Artifacts

Detection with Technique-induced Artifact

Heuristic-Based Features:
(manually engineered features)

Images:
- CNN using statistical properties
  - Rahmouni, 2017.
- Inconsistent eye color, missing reflections
  - Matern et al., 2019

Video:
- Lip and audio inconsistency
  - Korshunov and Marcel, 2018
- Head Movement
  - Yang et al., 2018

Deep Learned Features

(directly classify deepfake content from real content)

Images:
- Pure CNN
  - Bayar and Stamm., 2016
- Transfer learning via CNN XceptionNet
  - Rossler et al, 2019

Video:
- CNN + RNN
  - Guera Delp, 2018

Detection with Technique-induced Artifact

Heuristic-Based Features:
(manually engineered features)

Images:

- CNN using statistical properties
  – Rahmouni, 2017.
- Inconsistent eye color, missing

– Yang et al., 2018

Deep Learned Features
(directly classify deepfake content from real content)

Images:

- Pure CNN
  – Bayar and Stamm., 2016
- Transfer learning via CNN XceptionNet

Open Problems:

- Specific techniques may only work for specific types of deepfake
- Unsure of how methods generalize across different datasets
- Unsure of how methods compare against other methods
- Continuous cat-and-mouse game as deepfake generation improves over time

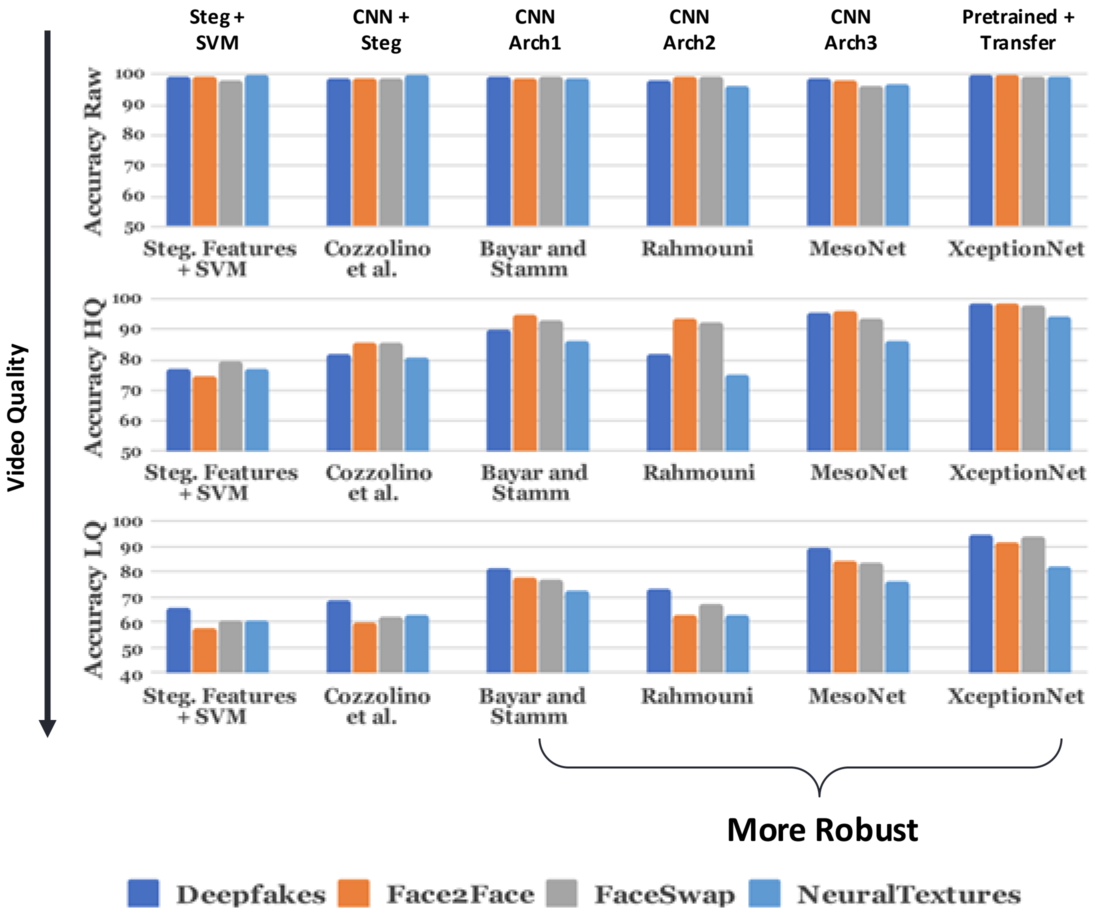
Video Facial Forgery Database

- Video Collected
  - 1,000 videos from Youtube
  - All front facing
- Insight: videos are often compressed
- 3 Video Quality Sets:
  - Raw: No Compression
  - HQ: Low Compression
  - LQ: High Compression

Low Quality Video Encoding

High Quality Video Encoding

Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." *ICCV* 2019.



Video Quality
Steg + SVM
CNN + Steg
CNN Arch1
CNN Arch2
CNN Arch3
Pretrained + Transfer
Accuracy Raw
Accuracy HQ
Accuracy LQ
Steg. Features + SVM
Cozzolino et al.
Bayar and Stamm
Rahmouni
MesoNet
XceptionNet
More Robust
Deepfakes
Face2Face
FaceSwap
NeuralTextures

Summary

- Pros:
  - Views media in isolation
  - Cheap to implement and run

- Cons:
  - No limit to how close generated images can mirror real ones
  - Just as susceptible to adversarial ML
  - Can get provide short term benefit, but a quickly losing battle

Defense with **Provenance**

Provenance-Based Method

- Idea: cryptographically sign media
  - Private keys in camera hardware
  - Private keys of trusted entities/companies
  - Source is verifiable

- Where to store provenance?
  - Single trusted entity
  - Distributed trust

Hasan et al. - "Combating deepfake videos using blockchain and smart contracts.", IEEE Access 2019

P2P File System

Ethereum Chain

child of

parent of

Original
Artist

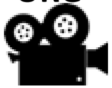
P2P File System

Ethereum Chain

child of
parent of

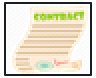
Original
Artist

ORG



- Video
- Device
- Original Addr
- Contract Addr

Original



- Contract Addr
- Original Addr
- File hash
- Functions
  - *requestPerm*
  - *grantPerm*
  - *Attestation*

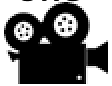
P2P File System

Ethereum Chain

child of
parent of

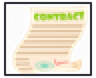
Original
Artist

ORG

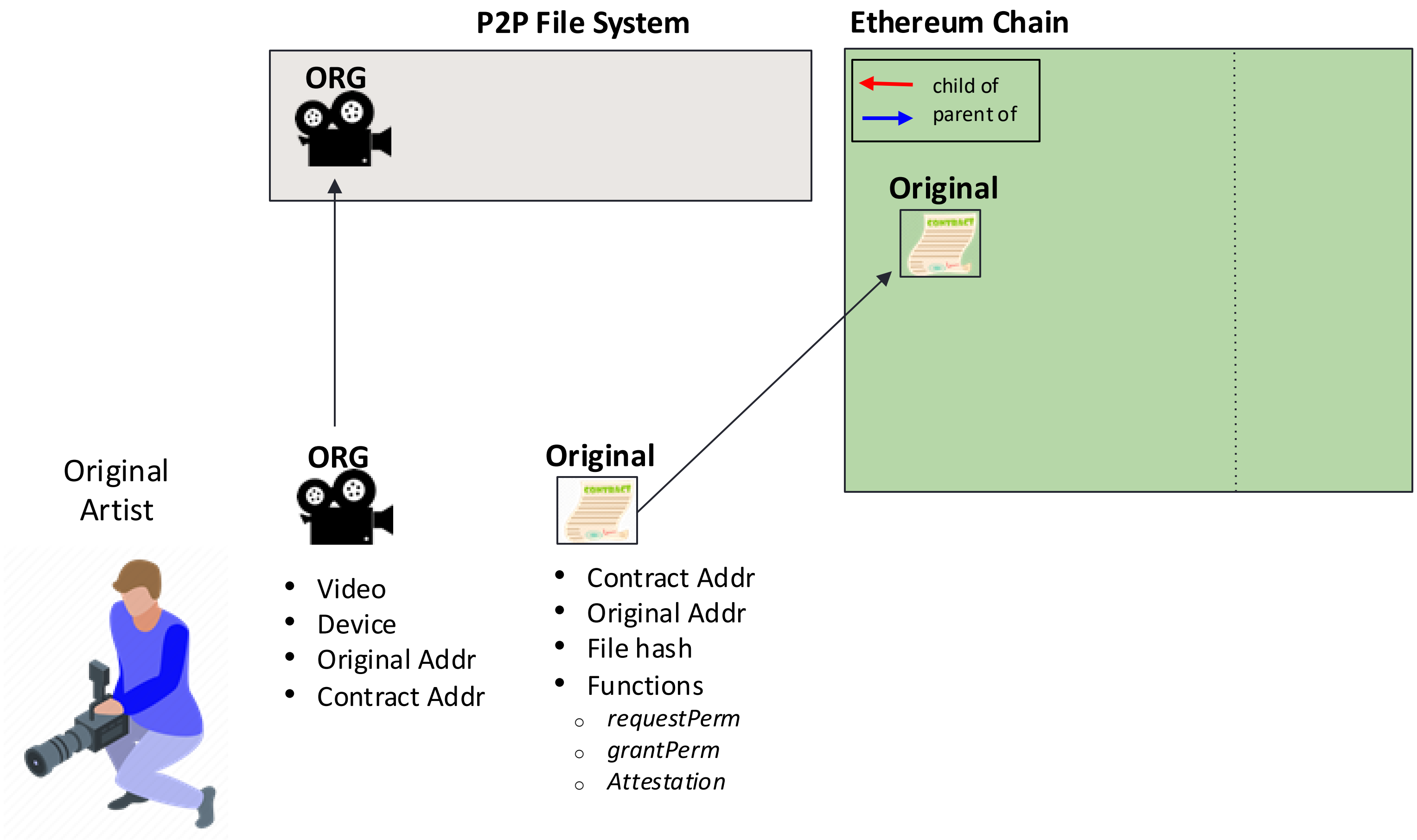


- Video
- Device
- Original Addr
- Contract Addr

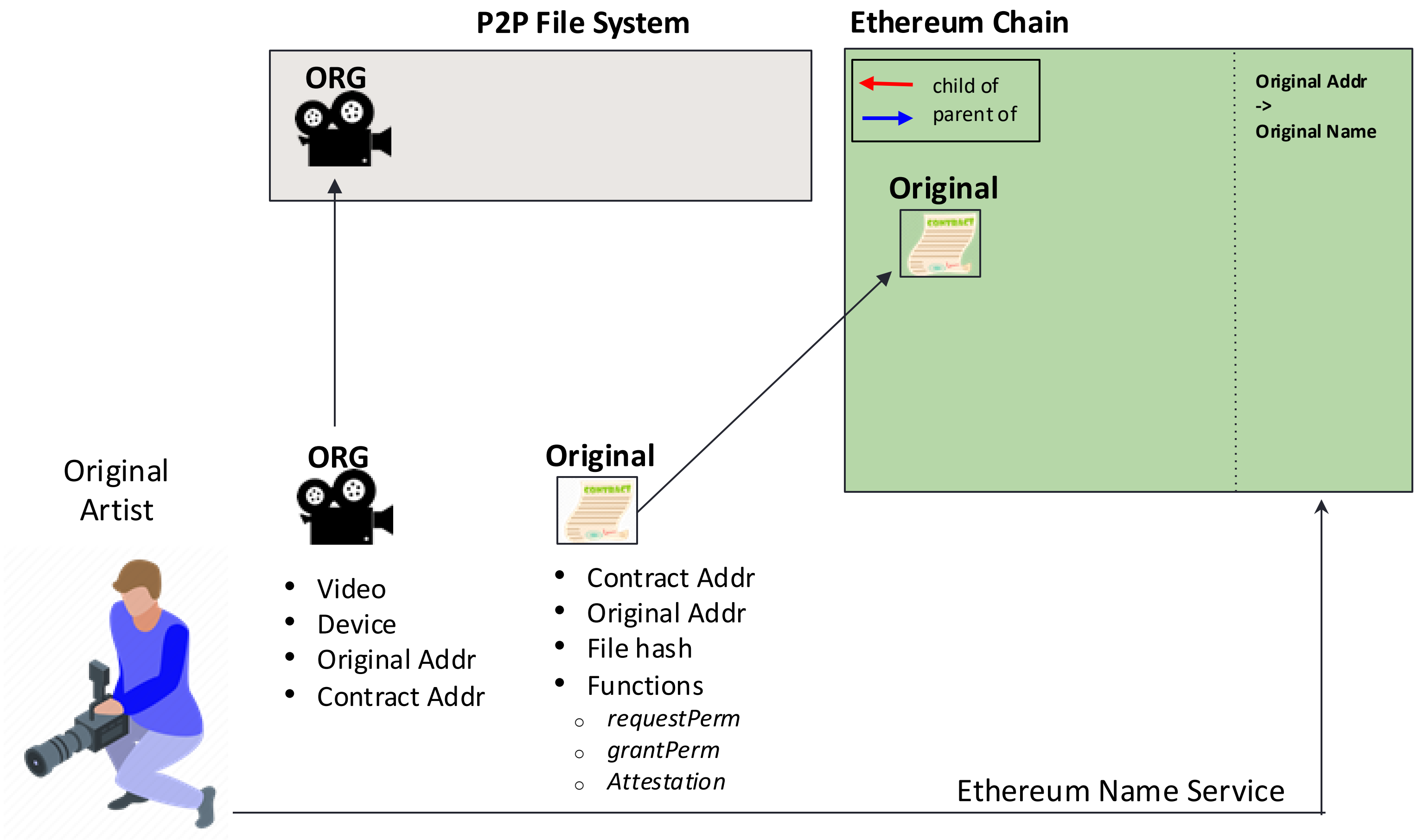
Original

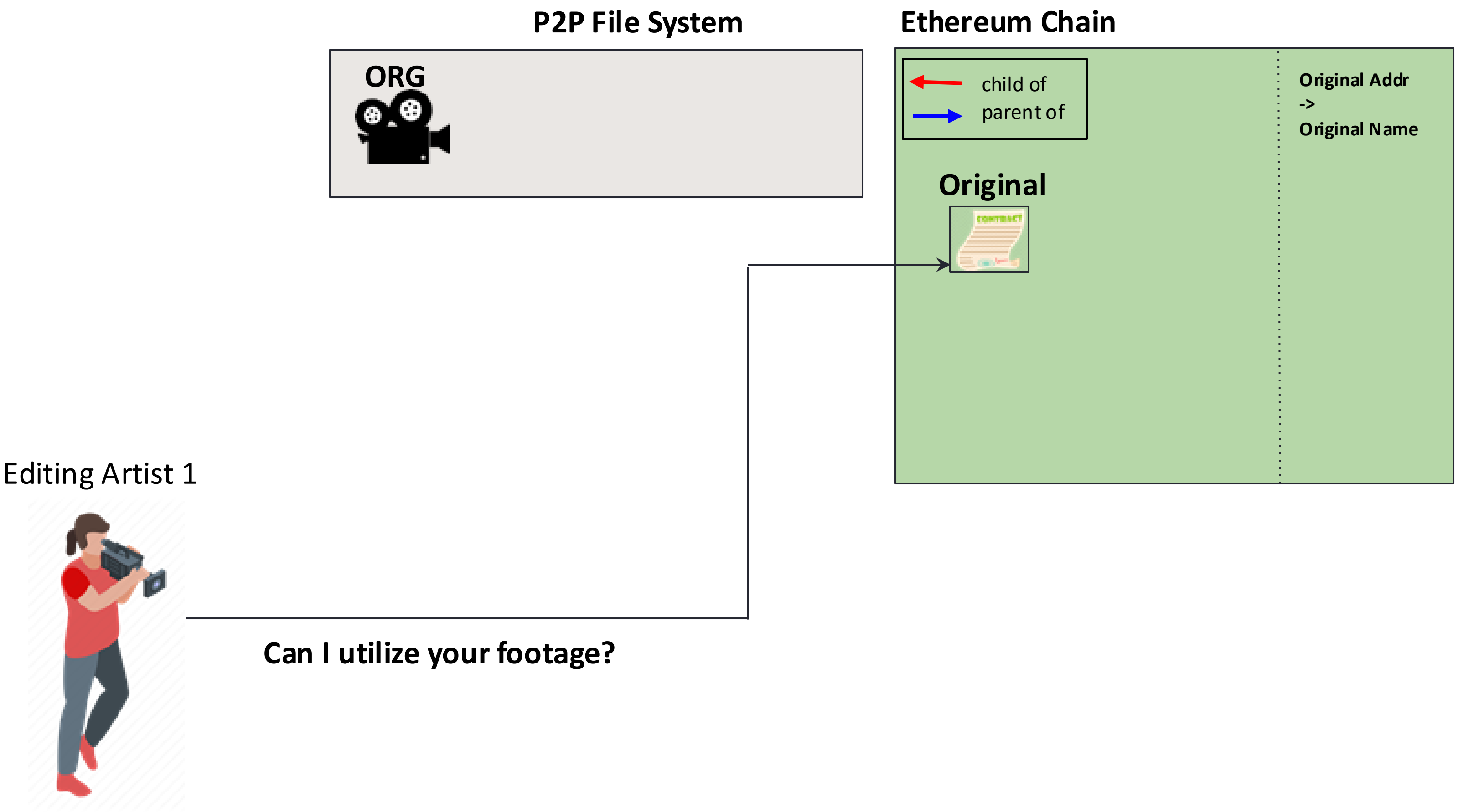


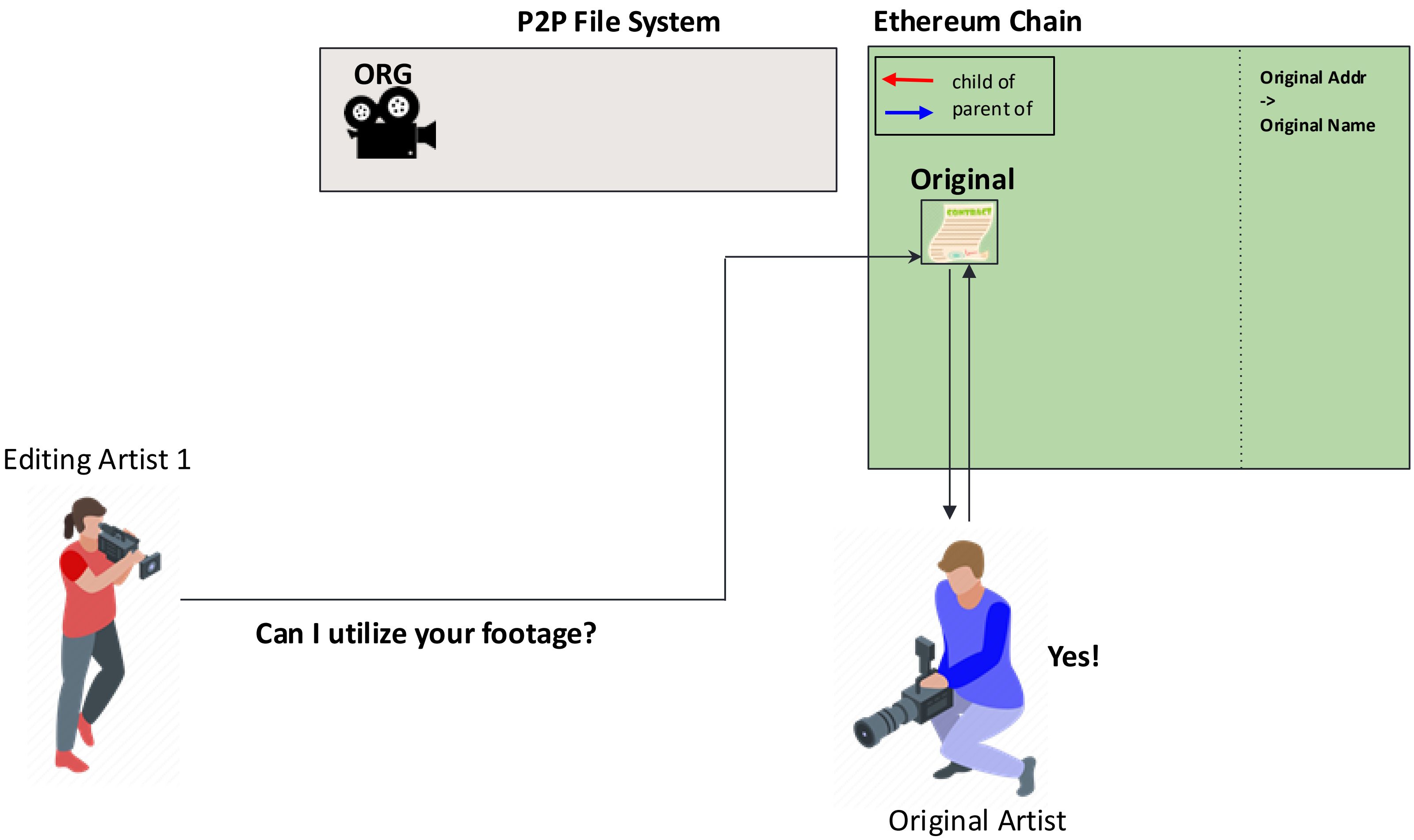
- Contract Addr
- Original Addr
- File hash
- Functions
    - *requestPerm*
    - *grantPerm*
    - *Attestation*

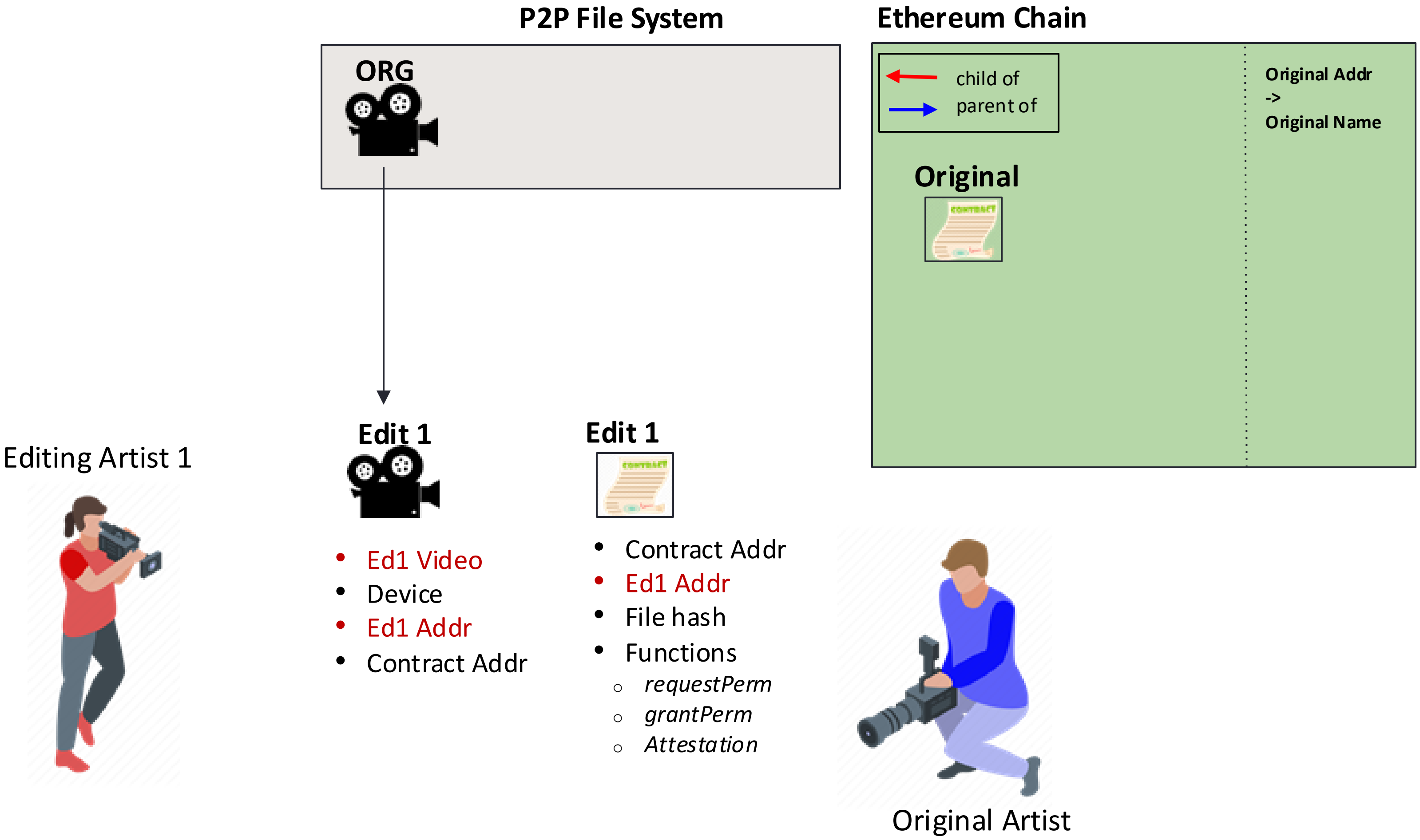


P2P File System
Ethereum Chain
ORG
child of
parent of
Original
Original
Artist
ORG
Video
Device
Original Addr
Contract Addr
Original
Contract Addr
Original Addr
File hash
Functions
requestPerm
grantPerm
Attestation

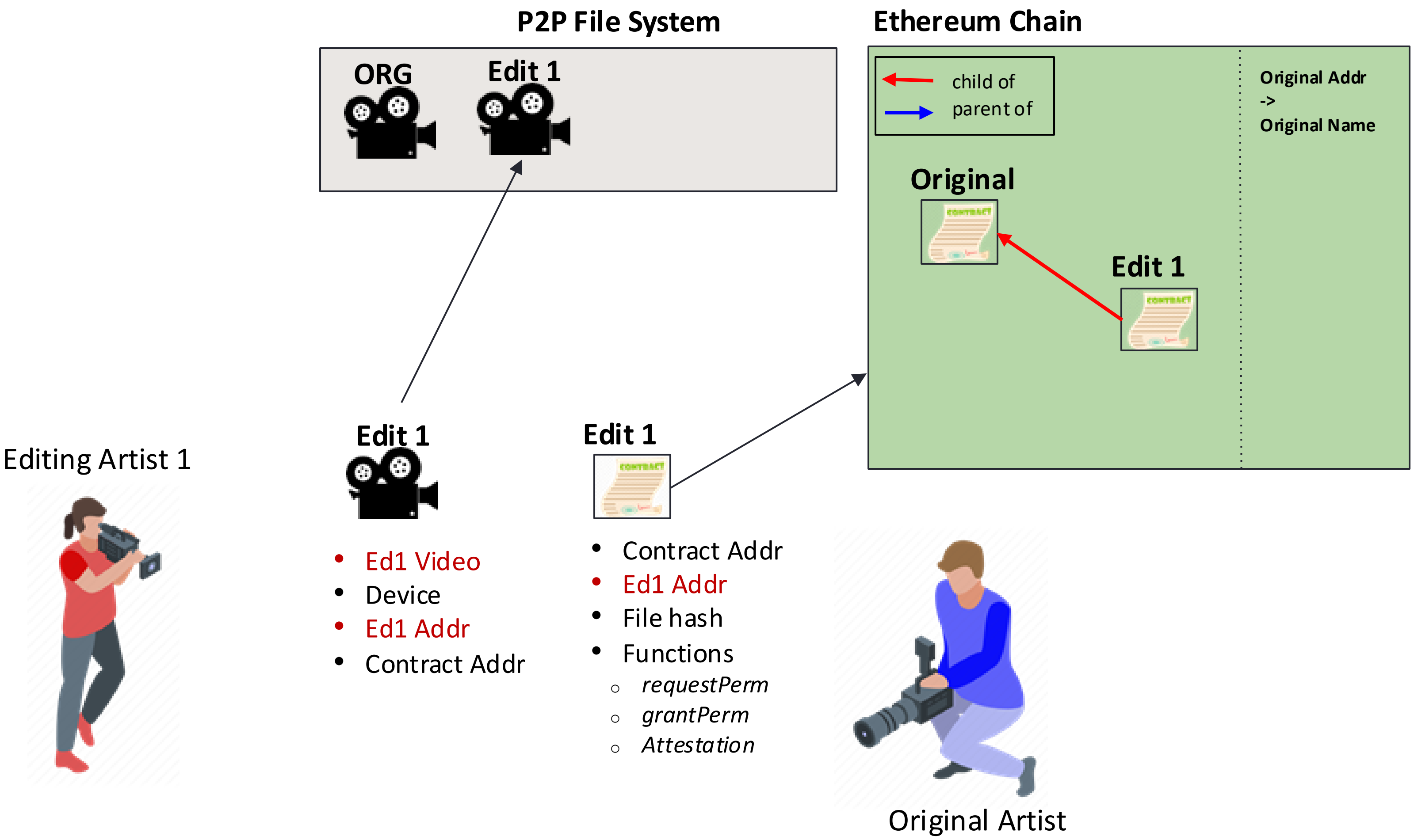

P2P File System
ORG
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Original
Original
Artist
ORG
• Video
• Device
• Original Addr
• Contract Addr
Original
• Contract Addr
• Original Addr
• File hash
• Functions
  ◦ *requestPerm*
  ◦ *grantPerm*
  ◦ *Attestation*
Ethereum Name Service

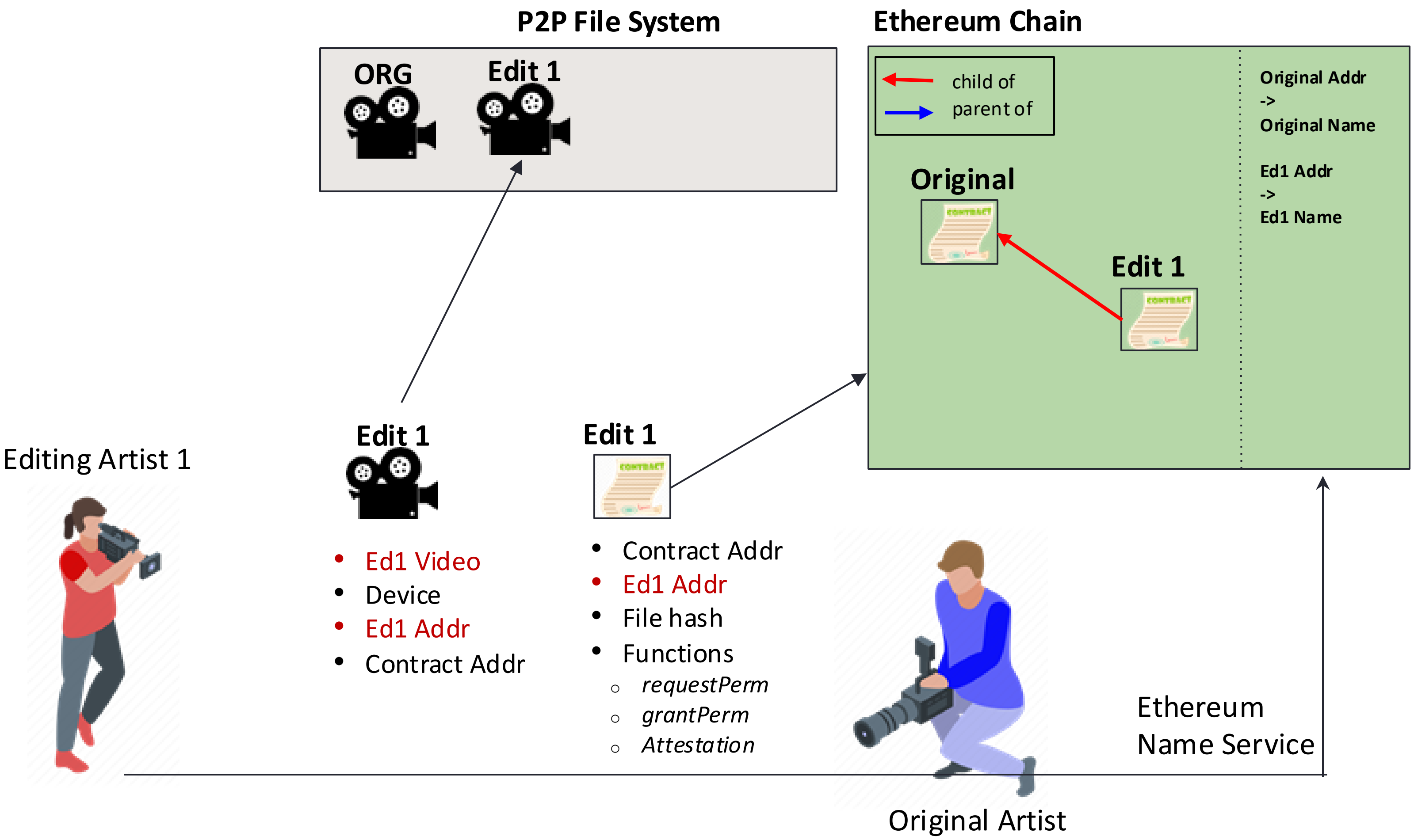

P2P File System
ORG
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Original
Editing Artist 1
Can I utilize your footage?

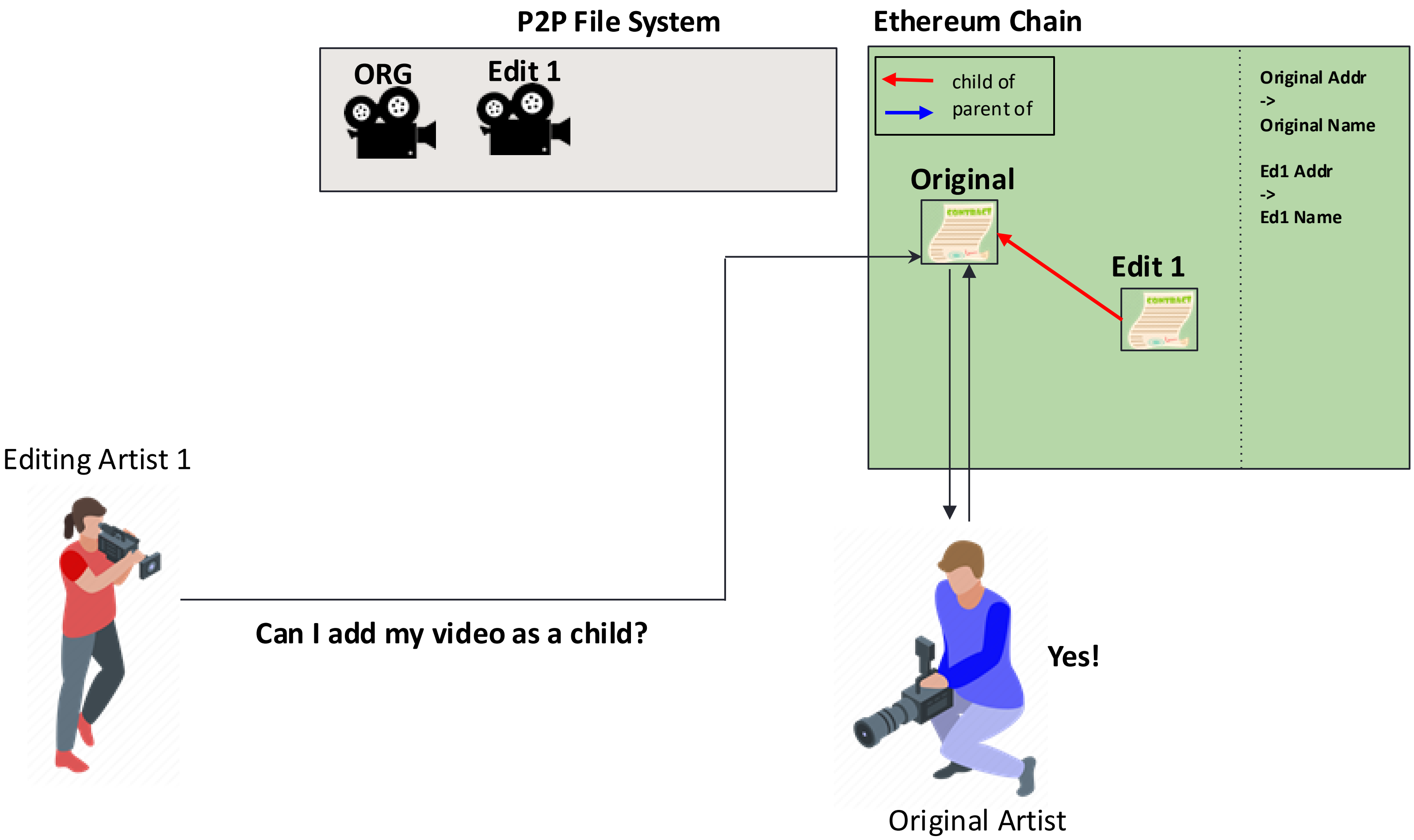

P2P File System
Ethereum Chain
ORG
child of
parent of
Original Addr
->
Original Name
Original
Editing Artist 1
Can I utilize your footage?
Yes!
Original Artist



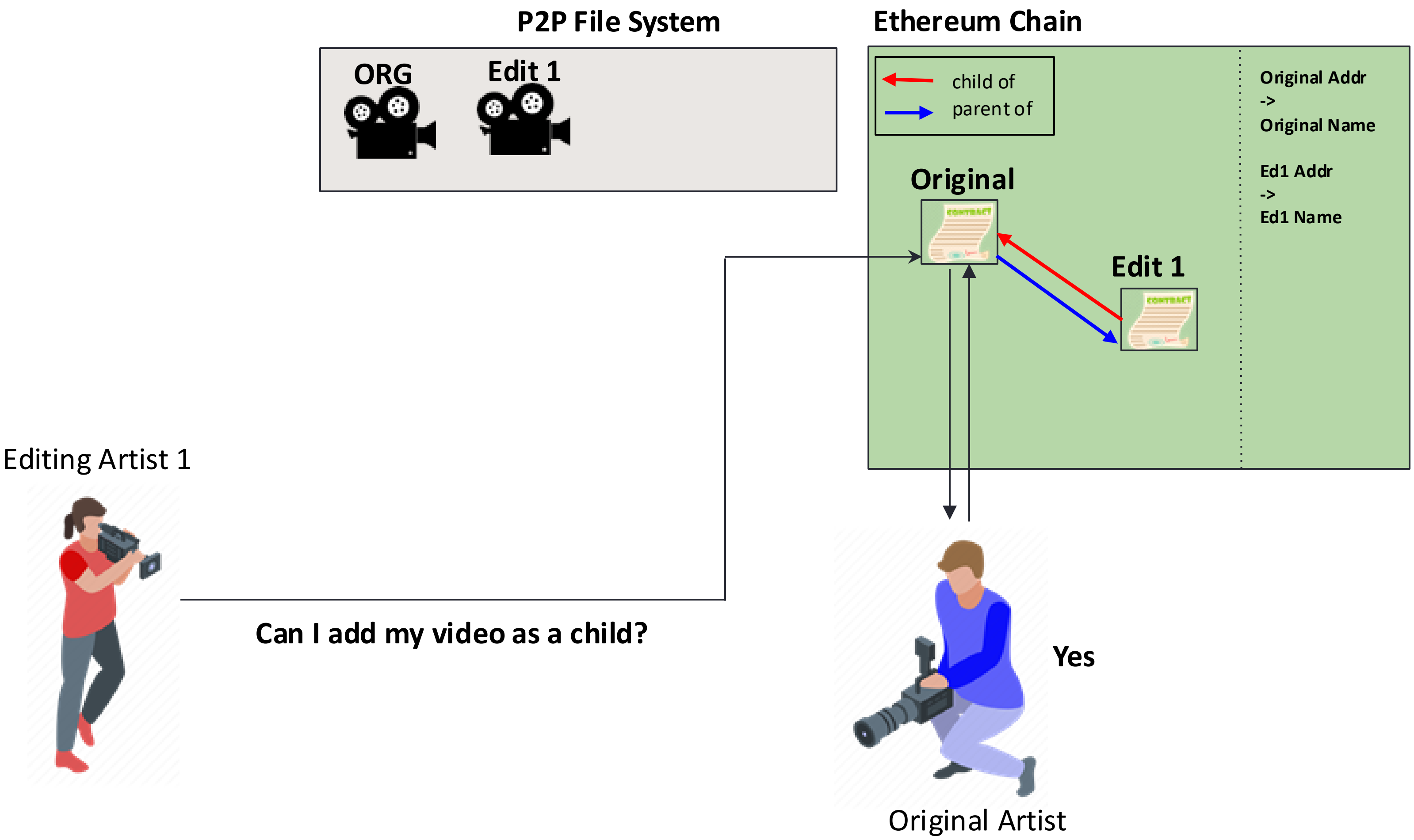
P2P File System
ORG
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Original
Edit 1
Edit 1
Editing Artist 1
Ed1 Video
Device
Ed1 Addr
Contract Addr
Contract Addr
Ed1 Addr
File hash
Functions
requestPerm
grantPerm
Attestation
Original Artist


P2P File System
ORG
Edit 1
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Original
Edit 1
Editing Artist 1
Edit 1
Ed1 Video
Device
Ed1 Addr
Contract Addr
Edit 1
Contract Addr
Ed1 Addr
File hash
Functions
requestPerm
grantPerm
Attestation
Original Artist

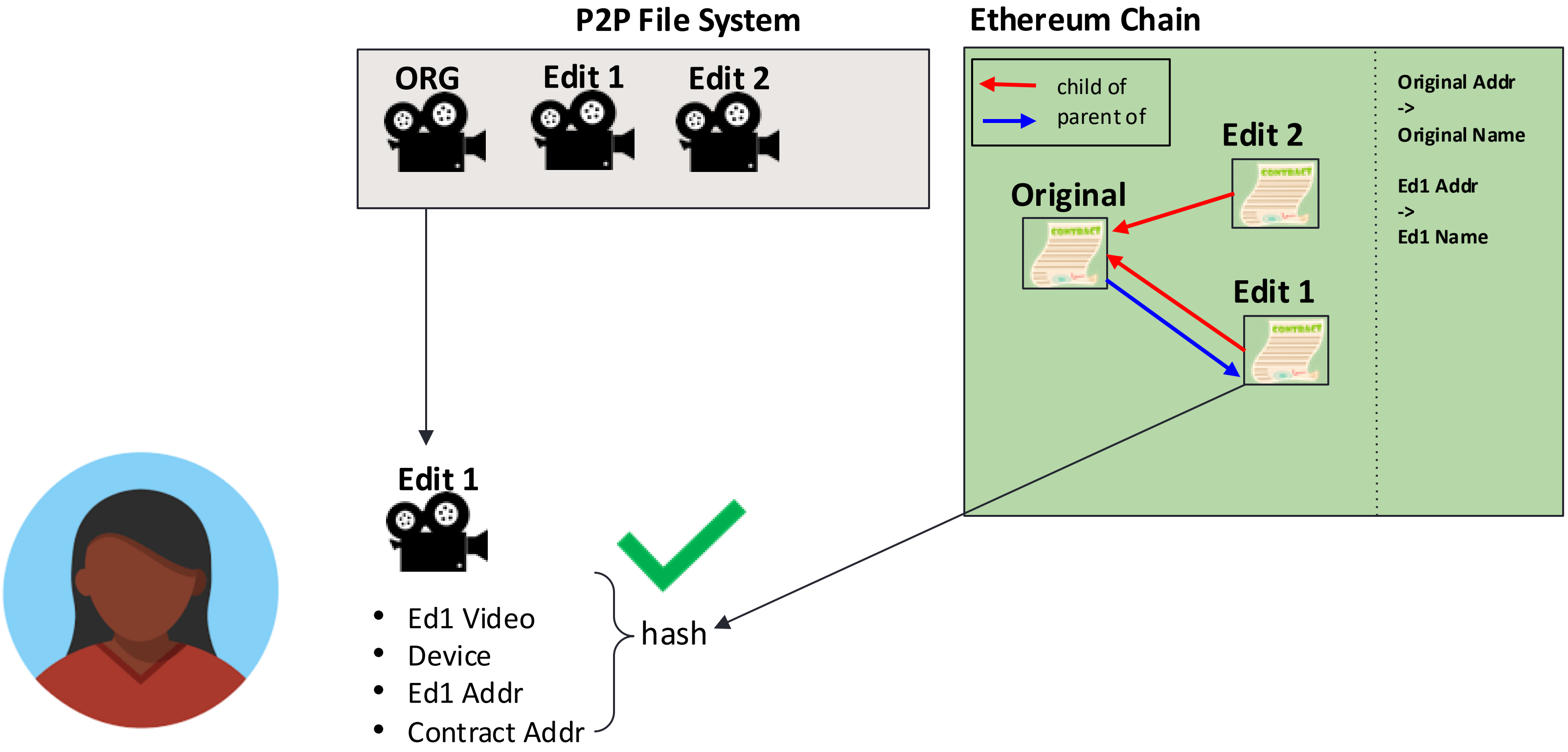

P2P File System
ORG
Edit 1
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
Original
Edit 1
Editing Artist 1
Edit 1
Edit 1
Ed1 Video
Device
Ed1 Addr
Contract Addr
Contract Addr
Ed1 Addr
File hash
Functions
requestPerm
grantPerm
Attestation
Ethereum
Name Service
Original Artist

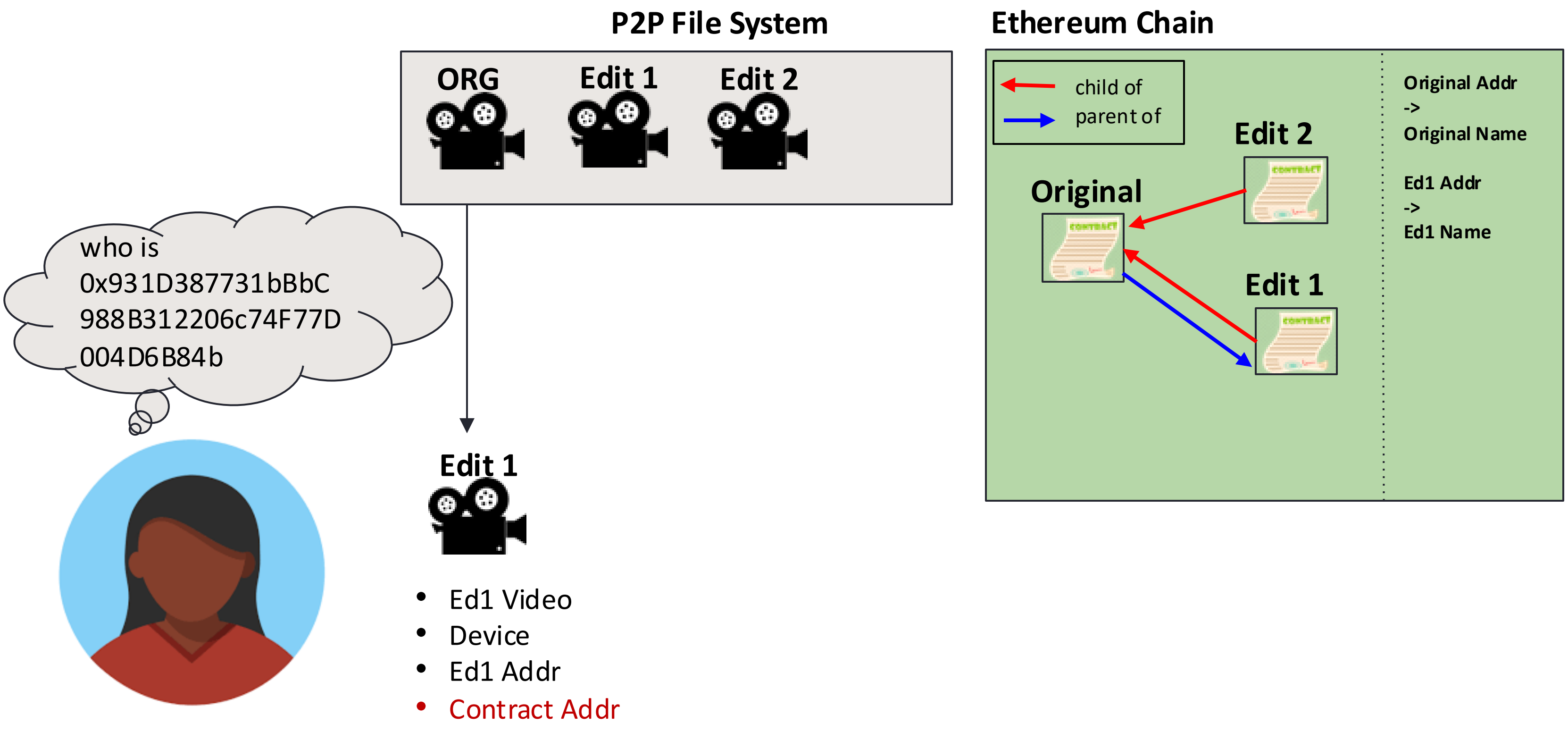


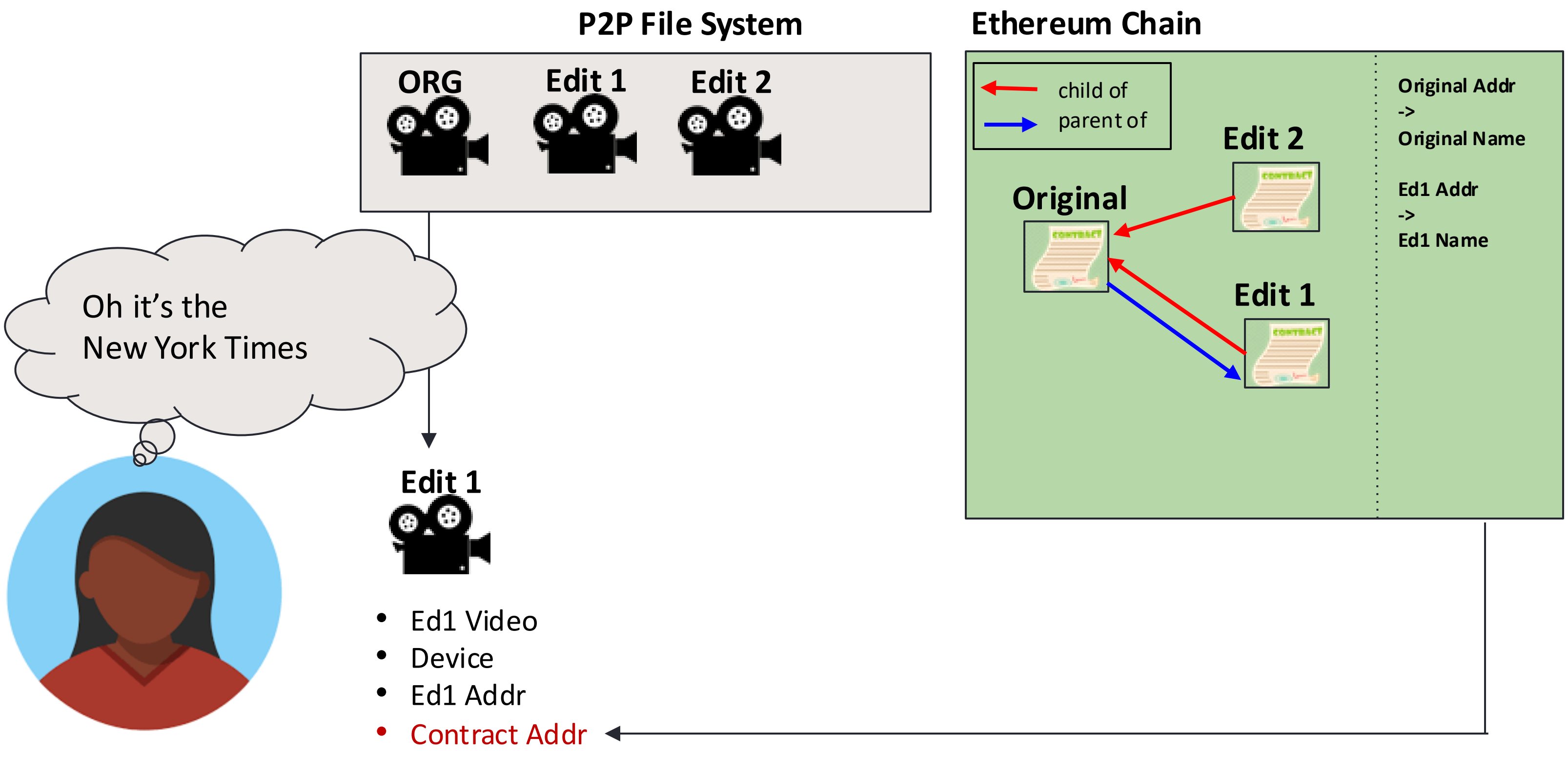
P2P File System
ORG
Edit 1
Ethereum Chain
child of
parent of
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
Original
Edit 1
Editing Artist 1
Can I add my video as a child?
Yes!
Original Artist



P2P File System
ORG
Edit 1
Ethereum Chain
child of
parent of
Original
Edit 1
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
Editing Artist 1
Can I add my video as a child?
Yes
Original Artist


P2P File System
ORG
Edit 1
Edit 2
Ethereum Chain
child of
parent of
Edit 2
Original
Edit 1
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
Edit 1
Ed1 Video
Device
Ed1 Addr
Contract Addr
hash


P2P File System
ORG
Edit 1
Edit 2
Ethereum Chain
child of
parent of
Edit 2
Original
Edit 1
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
who is
0x931D387731bBbC
988B312206c74F77D
004D6B84b
Edit 1
Ed1 Video
Device
Ed1 Addr
Contract Addr


P2P File System
ORG
Edit 1
Edit 2
Ethereum Chain
child of
parent of
Edit 2
Original
Edit 1
Original Addr
->
Original Name
Ed1 Addr
->
Ed1 Name
Oh it's the
New York Times
Edit 1
Ed1 Video
Device
Ed1 Addr
Contract Addr

Provenance-Based Defenses

Pros:

- Not dependent on media format
- Not dependent on forgery techniques
- Strong provenance guarantees

Cons:

- No guarantees on video authenticity
- Potentially impractical for generic media on social networks
- Organization-wide root of trust