BlockShield: A TPM-Integrated Blockchain-based Framework for Shielding Against Deepfakes

VLSI-SoC 2024
Special Session: Security-by-Design (SbD)

Venkata K. Vishnu. V. Bathalapalli[1], A. Kumar[2] ,
S. Mohanty[3], E. Kougianos[4] , Venkata P. Yanambaka[5]

University of North Texas, Denton, TX, USA.[1,2,3,4] and
Texas Woman's University[5].

Email: **vb0194@unt.edu[1], aakarshankumar@my.unt.edu[2], saraju.mohanty@unt.edu[3], elias.kougianos@unt.edu[4], vyanambaka@twu.**

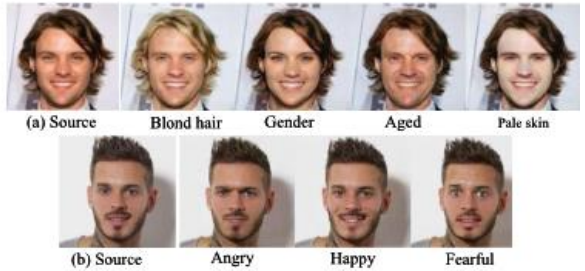
Outline

- Introduction to Deepfake Techniques
- Deepfake Mitigation
- Introduction to BlockShield
- TPM-Video Attestation
- Experimental Validation
- Conclusion & Future Research Directions

Deepfake

AI can be fooled by fake data

Attribute Manipulation



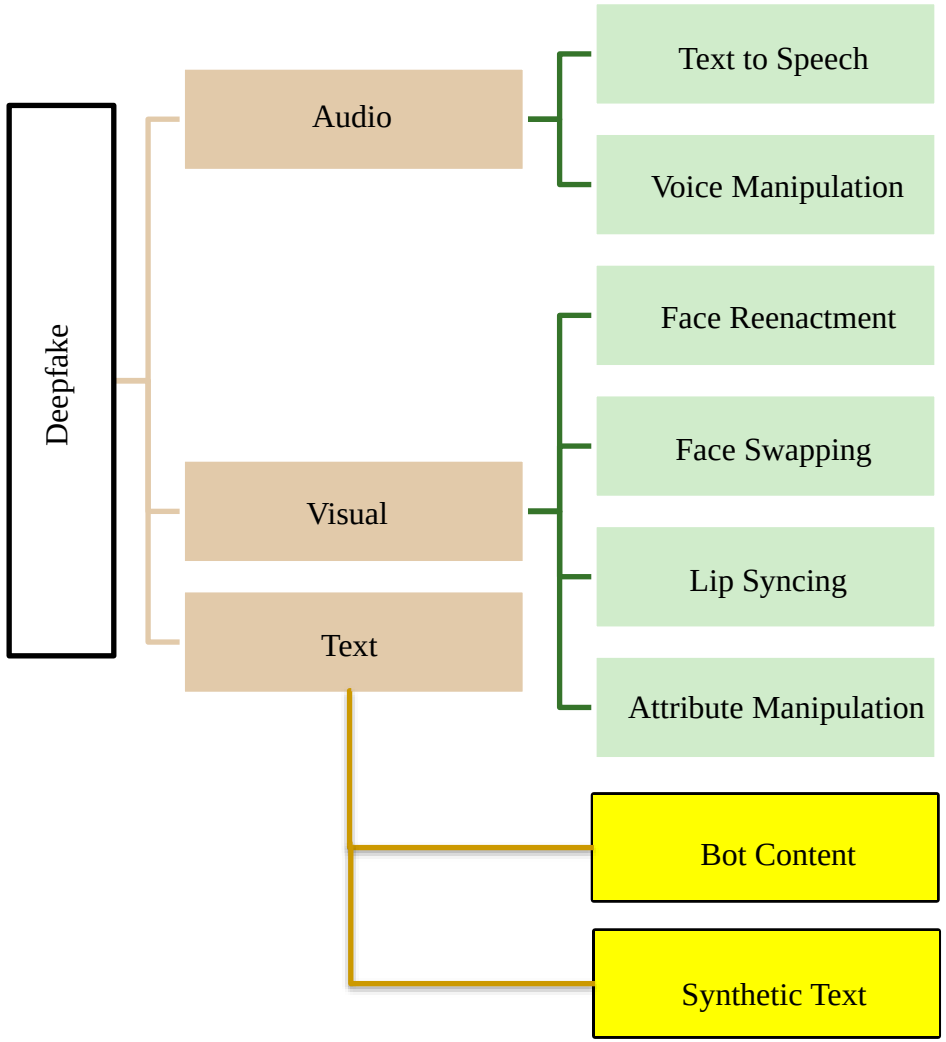
Identity Swapping

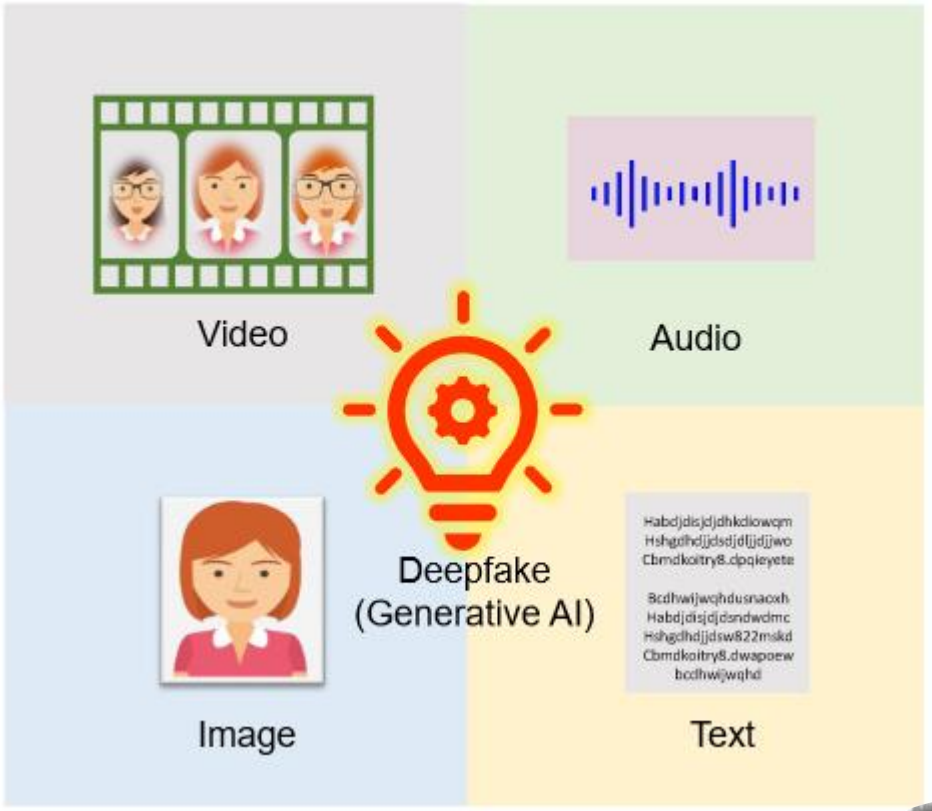
AI can create fake data (Deepfake)

1. Deepfake refers to super realistic, but fake images, sounds, and videos generated by machine learning methods.
2. Deepfake leverages a Generative adversarial network (GAN) which enables the modification of human faces in a video or image.
3. Deepfakes can be classified as Audio, Visual and Text

Source: A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in *IEEE Acc* 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

Deepfake Techniques

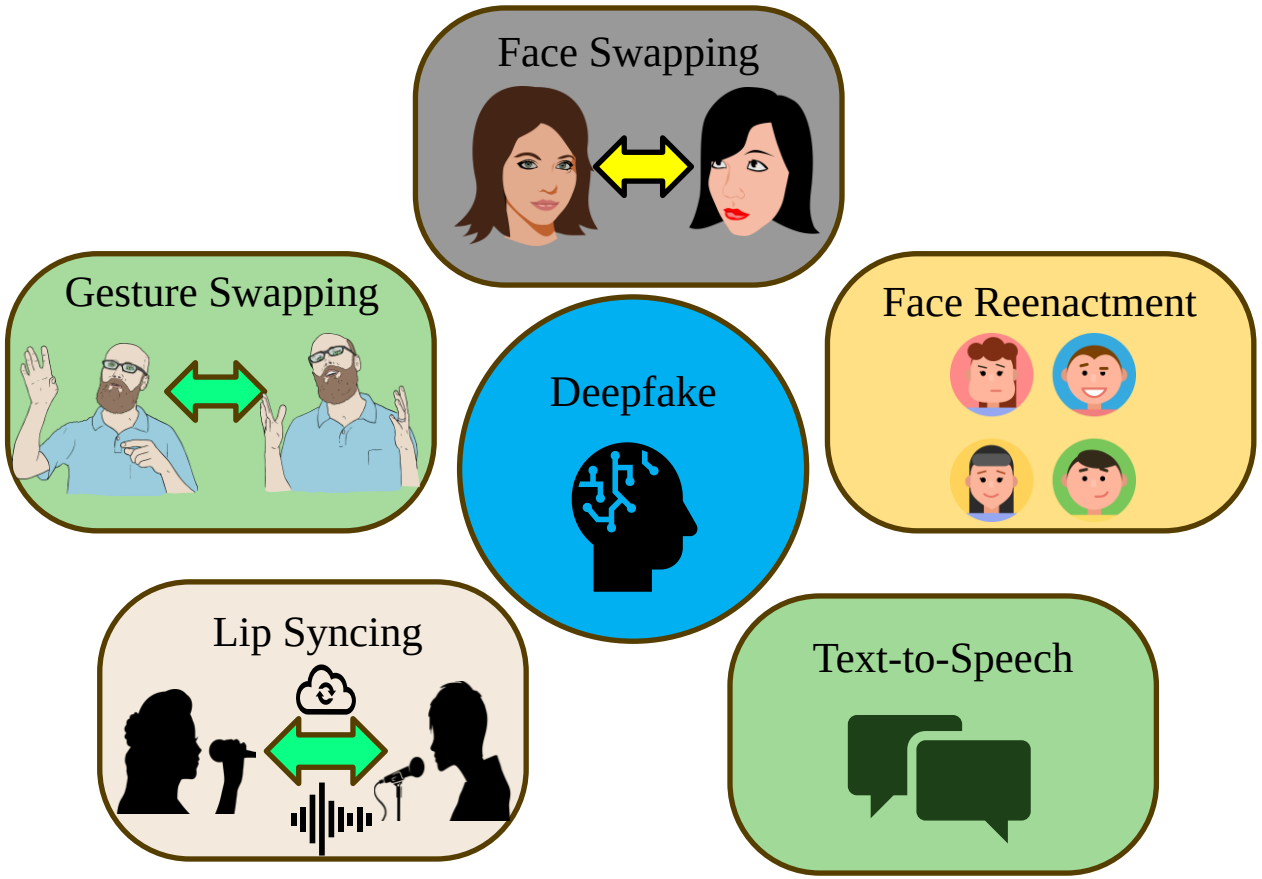




Source: A. Mitra, **S. P. Mohanty**, and E. Kougianos, "The World ' Generative AI: Deepfakes and Large Language Models", *arXiv Science*, arXiv:2402.04373, Feb 2024, 9-pages.

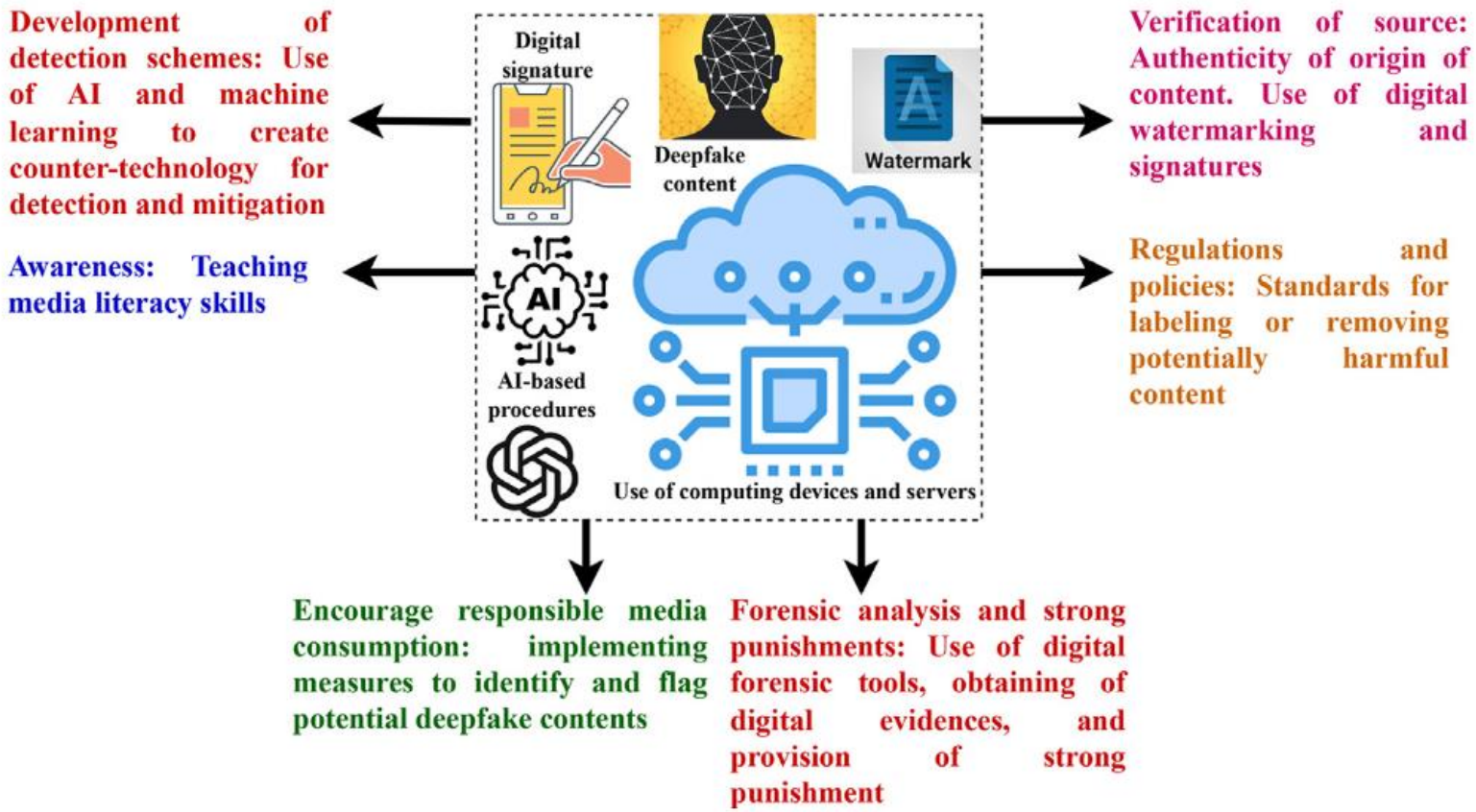
Source: V. K. V. V. Bathalapalli, V. P. Yanambaka, **S. P. Mohanty**, and E. Kougianos, "PUFshield: A Hardware-Assisted Approach for Deepfake Mitigation Thr Feature Attestation", in *Proceedings of the ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2024, pp. 676--681, DOI: https://doi.org/10.1145/3649476.36603

Visual Deepfake Techniques



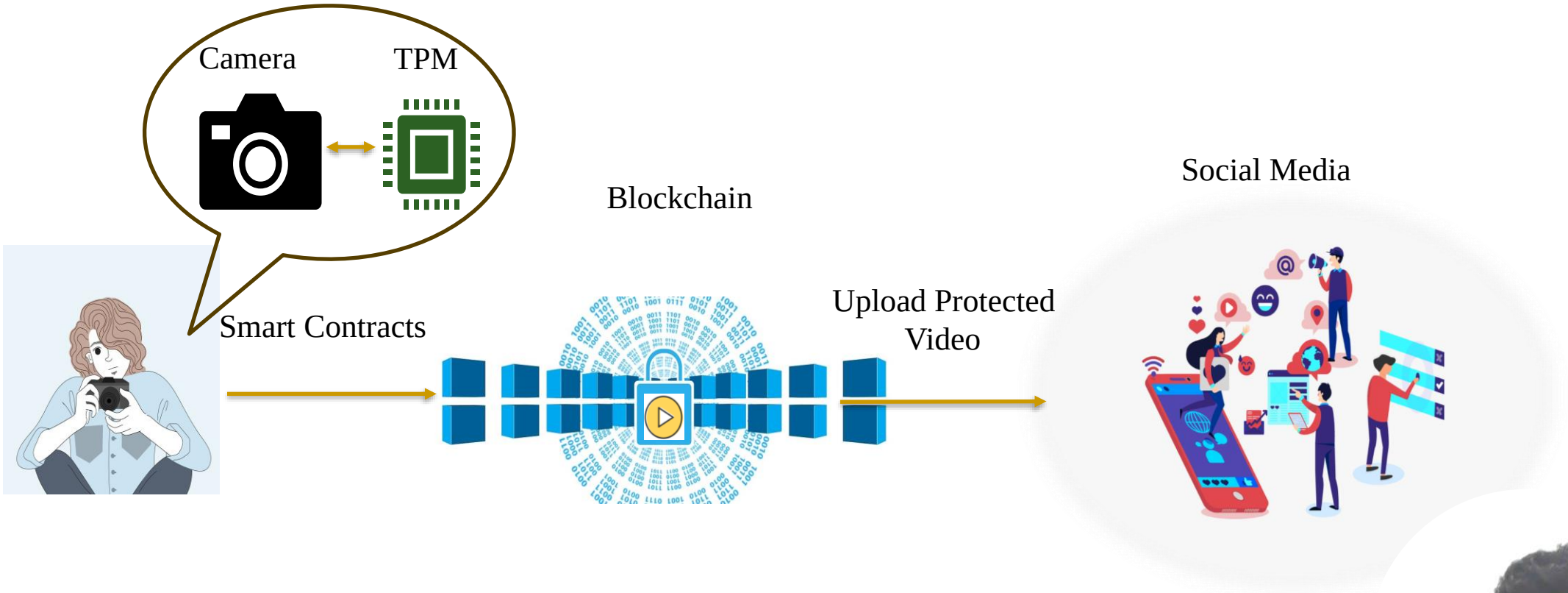
Source: V. K. V. V. Bathalapalli, V. P. Yanambaka, **S. P. Mohanty**, and E. Kougianos, "PUFshield: A Hardware-Assisted Approach for Deepfake Mitigation Through PUF Attestation", in *Proceedings of the ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2024, pp. 676--681, DOI: https://doi.org/10.1145/3649476.3660394.

Deepfake Mitigation



Source: Wazid, M., Mishra, A. K., Mohd, N., & Das, A. K. (2024). A Secure Deepfake Mitigation Framework: Architecture, Issues, Challen Impact. *Cyber Security and Applications*, 100040.

BlockShield: Conceptual Overview



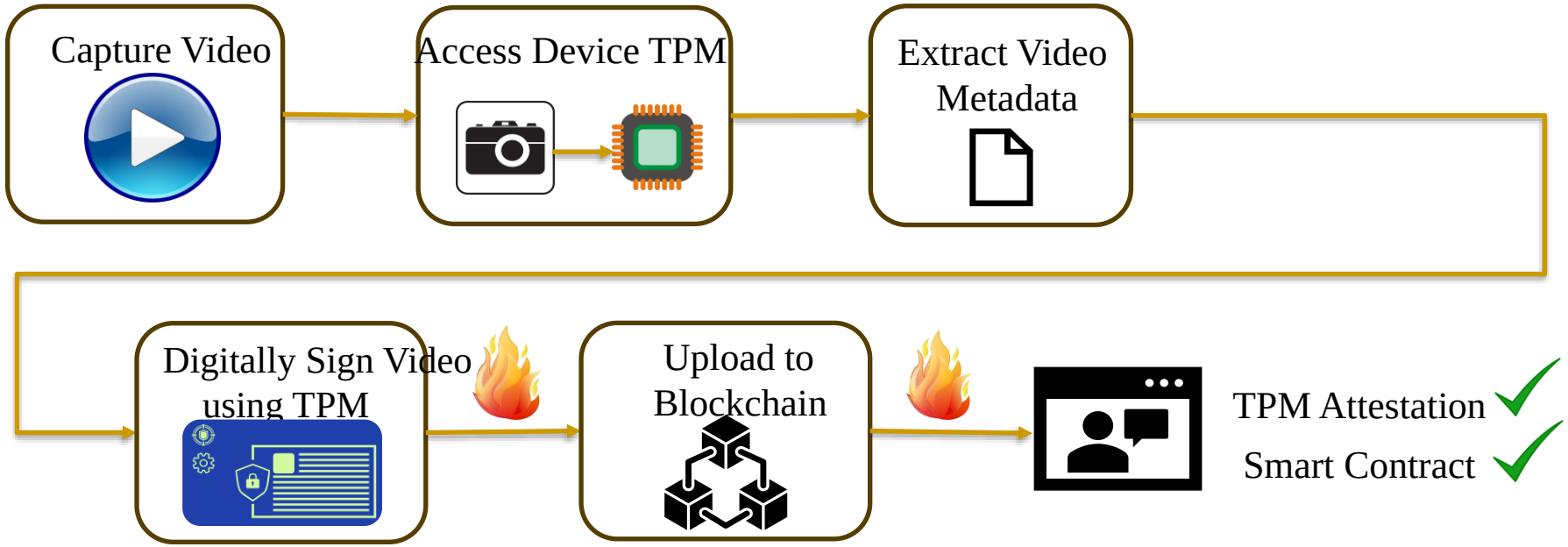
Related Research

Work	Technique	Methodology	Tools
Taeb et.al [8]	Detection	ML and Blockchain-Integrated Fake News Detection	Efficient Net, Smart Contracts
Bathalapalli et. al [14]	Mitigation(Image)	PUF and ML framework for facial feature attestation	Dlib 68 (Facial detection and keypoint prediction), PUF
Alattar et. al [10]	Fake news mitigation	Watermarking and Blockchain for Deepfake Video protection	IPFS, MTCNN algorithm, and Face Alignment Network (FAN) algorithm
Qureshi et. al [15]	Audio Deepfake Mitigation	Fragile speech watermarking with Blockchain	MTCNN, Wav2Lip
BlockShield	Visual Deepfake Mitigation	Blockchain and TPM-based video attestation	Hardware TPM, Sr Contracts

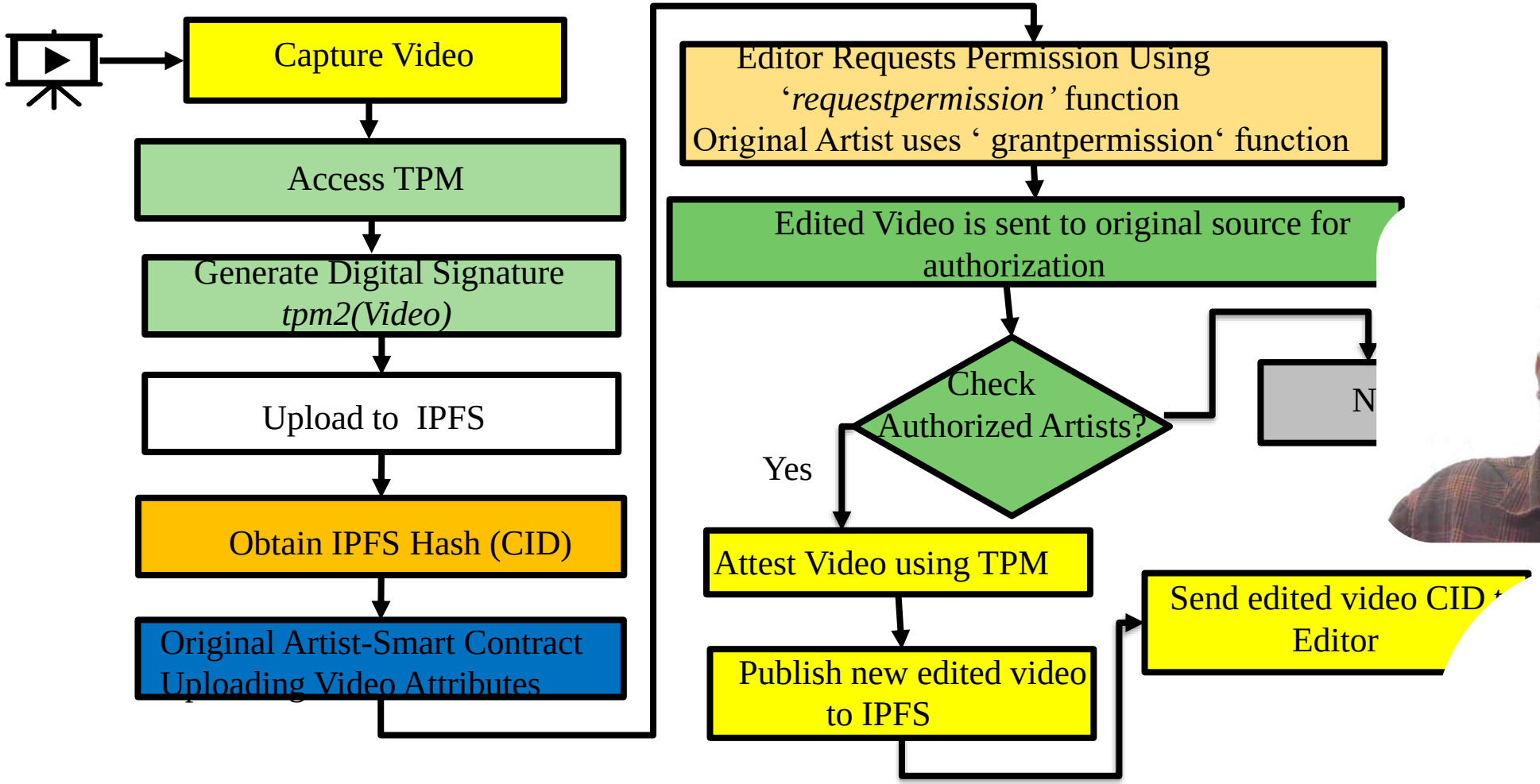
Novel contributions

- A sustainable Deepfake mitigation approach using state of-art TPM and Blockchain technologies.
- A secure visual Deepfake mitigation approach for individual content privacy and security on social media.
- An energy efficient solution that integrates TPM and Blockchain using smart contracts
- A secure digital content sharing framework using Blockchain to provide integrity and authenticity.
- An approach based on TPMs digital signature mechanism facilitating hardware root-of-trust for the video/image.

BlockShield: Proposed Deepfake Mitigation Technique



Working Flow



TPM Video Attestation Workflow

- 1: Access TPM hardware security module at the camera
- 2: tpm2 createprimary -C e -c primary.ctx

  \\ Create Primary Key
- 3: tpm2 evictcontrol -C o -c primary.ctx 0x81010001

  \\Assign a unique identifier in TPM NV RAM to make it persistent
- 4: tpm2 create -G rsa -u rsa.pub -r rsa.priv -C 0x81010001
- 5: tpm2 load -C 0x81010001-u rsa.pub -r rsa.priv -c rsa.ctx.
- 6: tpm2 evictcontrol -C o -c rsa.ctx 0x81010002

  \\Create RSA keys using primary key and make it persistent
- 7: Load the video file and Hash it Fi→ SHA256(Fi)→ Fi.hash

  \\Hash video File
- 8: tpm2 sign -c 0x81010002 -g sha256 -o sig.rssa Fi.hash

  \\Digitally sign the video hash file using TPM
- 9: tpm2 hash -C e -g sha 256 -o sig.rssa.hash -t ticket.sig.rssa sig.rssa

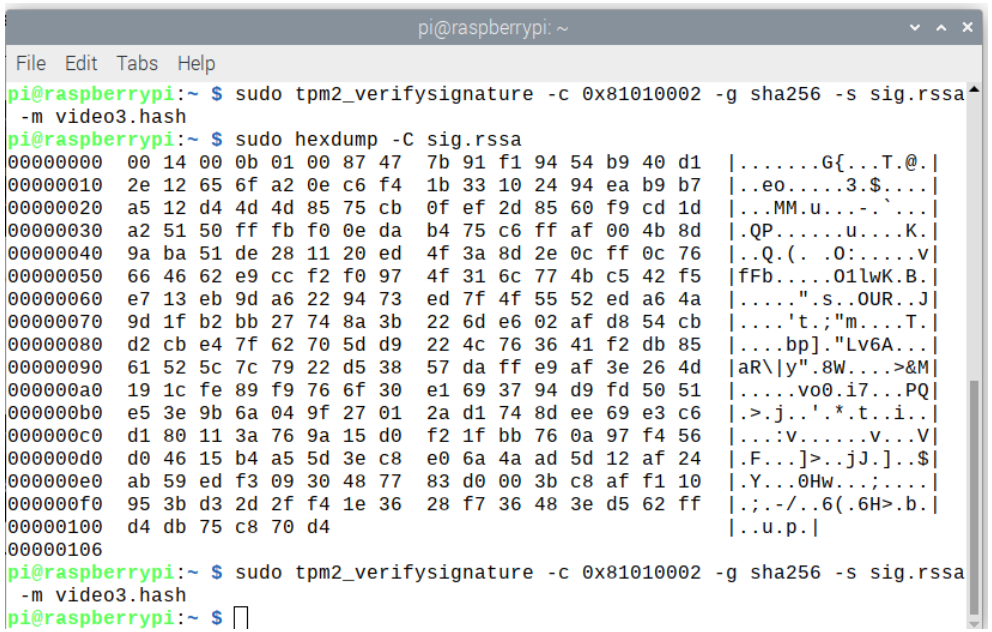
  \\Generate SHA 256 hash of Digital signature for video file

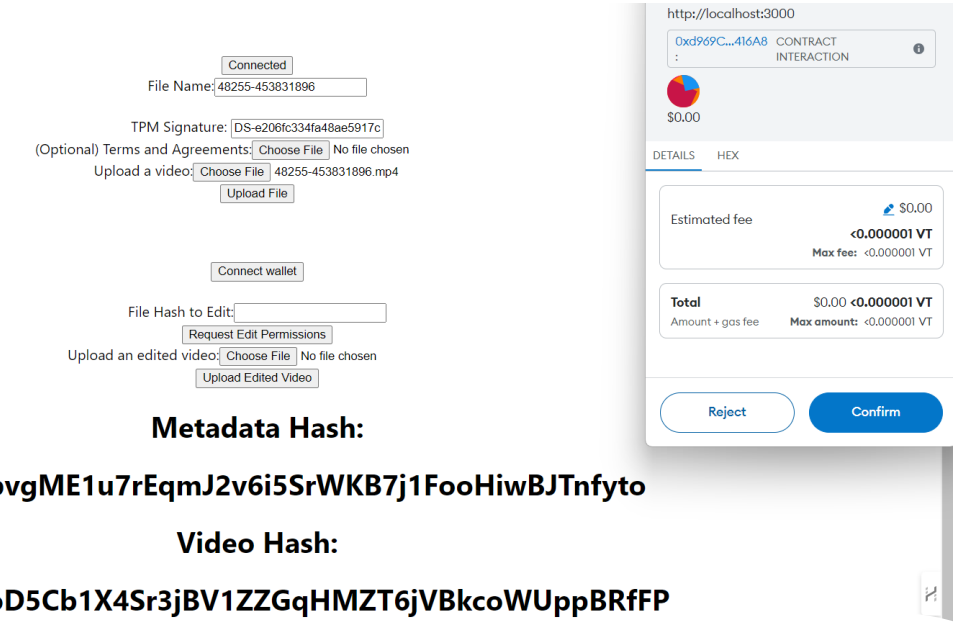
Smart Contract Validation Workflow

Input: Digital Signature of Video file D_{Fi} and Video File Fi
Output: Digital content is securely stored in Blockchain and secure accessed using smart contract

1: **for** Each Primary artist **do**
2: Individual contract is called by Primary artist to manage access
3: Primary Artist attested Video file information on to IPFS system
4: **for** Each Video File **do**
5: Upload video file Fi and Digital signature D_{Fi} on IPFS.
6: IPFSfile $Ii \leftarrow$ IPFS.upload(Fi, D_{Fi})
7: **end for**
8: ArtistContract.addIPFSHash(Ii)
Return hash is added as an attribute in the newly created Primary artist contract
9: **end for**
10: Share the contract address and IPFS hash of the video provide access to video file
11: Editor initiates a 'requestpermission' function to primary artist contract address which is accessible online.
12: Update list of artists at the primary artist side using 'grantpermission' function
13: Editor submits a new edited version of video file Fi to primary artist
14: Edited video file VFi is attested using TPM
$VFi \rightarrow$TPM$\rightarrow F = D_{FVi}$
15: Upload Edited and attested video file VFi and D_{FVi} to IPFS and share it with secondary artist or editor.

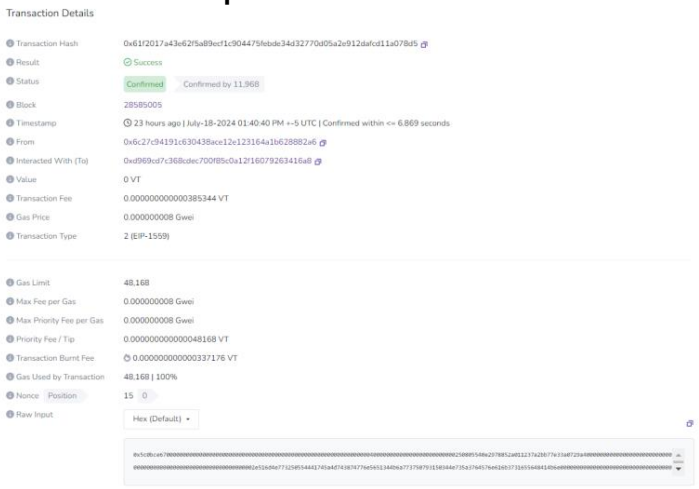
Experimental Validation of BlockShield





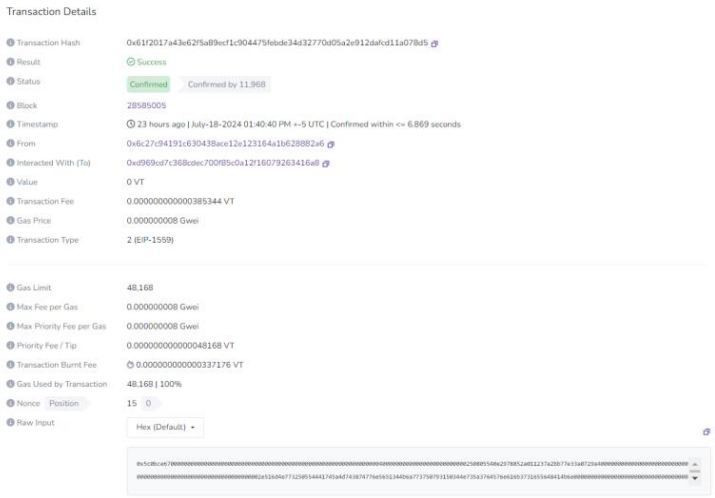
Smart Contract Validation

Upload Media

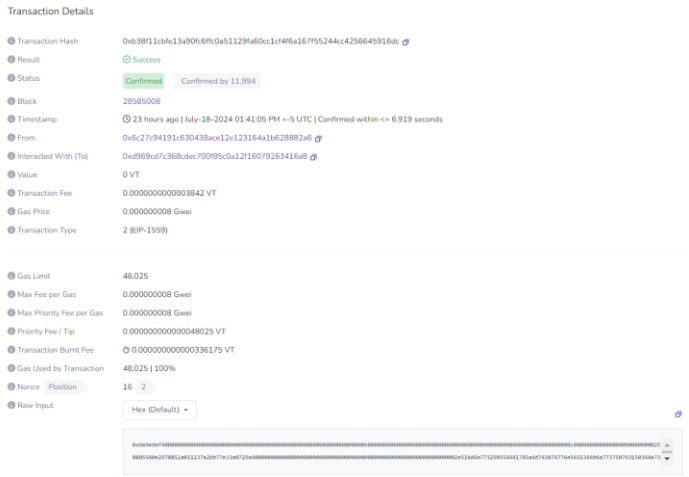


Request Permission

Grant Permission



Edit Media



Performance Analysis

Video	Duration (s)	Frame Rate	Bitrate	TPM-Signature	TX Hash	IPFS CID
48255-453831896.mp4	16	24	1633.922	e206fc334fa 48ae5917cac93dff 260d0fc0f0535f4e2 25c932466c 2291833df9	0x4c99d2c8f26 5498b09c53b372 e94f3cefc89d17be 12b401de25bf2b db892609b	QmQjxbmrjFbS7Xz4WXSigtkP msAKRi5FEXdDDpPK1Zn5g1
61299-498228517.mp4	26.56	29.97	3526.986	fa60e6faf0f64 c50846ac74ca185ffc d83d89fbd68fb 9d2985a6bb5a454eab1a	0x631719a0744dd 4d880924ac7ad 57b98d5d 385a73af7e8e5 4e55039d 4612e723b	QmS5PjDzYqGtfTCYYCm9QrW Fam6ZHZVKU2uw5vJgN6abqf
61706-500316063.mp4	15.65	29.97	937.89	e206fc334fa48a e5917cac93dff260d0fc0f0535f4e 225c93 2466c2291833df9	0x539dfe6fad 4015a6b4ed84 85717614f b9091f7ec0ff1 aa5fc7c93 3a76821b0fe	QmTtS4J4GHG4n2tMEPsfNbVT MuTEmw3uKni5x4djUPEqSN
73711-549547411.mp4	25.20	29.27	1098.274	9c1f7e38f1528cc18765b79e28fa 76f3fab662d0163cd309260939B 208d7dcee	0xbc61d12a 52c0815799da10 e0ea8806 28c287a538eb aae65fe857e8 aa0b1435cc	QmNw2PUDAtZ Mt8twnVQ4Kjw7Py 1P4NsZ7dWnak71eVHAK
44645-439940290.mp4	10.88	25	5596.436	9bee3bef81c8ac 3f596a 6f4c44b b218cb713d2 ba2541e89c487a 59362729f60f	0x067ee279182e24 372e49ecd7e5 22551169d23 1a30da152f7 651c1c4ac2 945a6d	QmWV/ 7NHtV RyqC

Conclusion and Future Research

- This research work presented and experimentally validated a Blockchain and TPM integrated approach for Deepfake mitigation through TPM-based hardware digital content attestation.
- The proposed work with state-of-art TPM-digital signature approach ensures hardware based digital content source attestation facilitated through Blockchain smart contract-based access control approach ensuring digital content authenticity.
- This is a novel work with TPM attestation and blockchain smart contract for access control and digital content sharing with the substantial performance indicators showcasing the robustness of the proposed Deepfake mitigation approach.
- Furthermore, proposed research work could be further applied to Deepfake mitigation for images with effective mechanism for facial feature and biometric-based user authentication.
- Additionally, this work could be extended to smart cities surveillance applications which work in untrusted environments to guarantee privacy, security and traceability to digital content.

Thank You!