Content Authentication in an Inauthentic World

Jacob Hobson, Lamyaa Aljuaid, James Rainey, Mohamed Elawady, Deepayan Bhowmik
deepayan.bhowmik@ncl.ac.uk

With the progression of advanced and easy-access toolsets and generative AI, media manipulation has become commonplace. The creation of manipulated media has been used in the entertainment industry as well as for criminal purposes. Such developments create issues around media privacy, where certain aspects of an image require anonymity (face) or access control (medical images), and security, which includes challenges around copyright/intellectual property rights. This work proposes the development of a framework that can be used for media privacy, security and provenance, leading to the establishment of trust in the media. The framework constitutes progress towards a larger framework/infrastructure that facilitates trusted media distribution. The proposed framework leverages a distributed media blockchain alongside modular components facilitating provisions for media integrity, authenticity and provenance. The framework also proposed a new self-embedding watermarking technique for manipulation detection that simultaneously provides a mechanism for content provenance.

Introduction

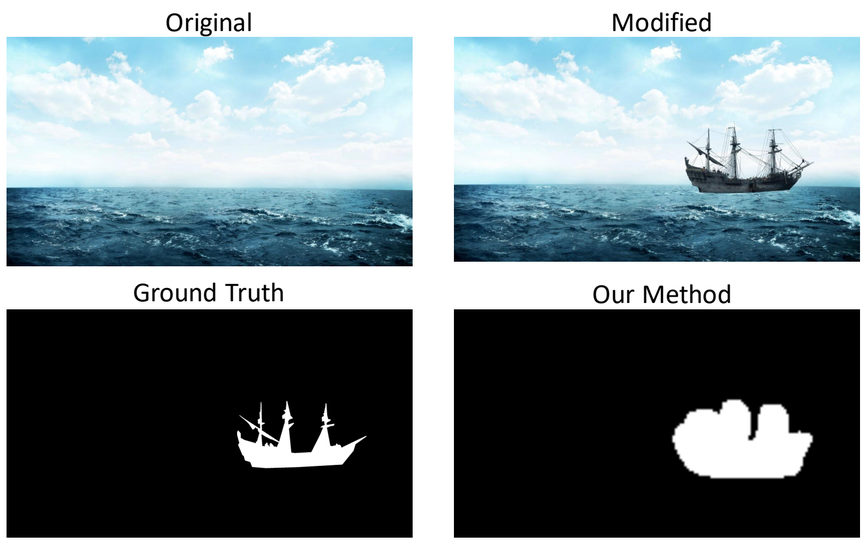

Figure: Example output from the proposed framework showing the original image, the modified version, the ground truth of the modification and the predicted mask.

- A multimedia blockchain-based content provenance framework incorporating a content integrity verification engine detecting manipulations.
- A self-embedding watermarking scheme, leveraging compressive sensing to detect and localise tampering/manipulation. The proposed algorithm enables the structural recovery of the original content.

TRAIT Framework

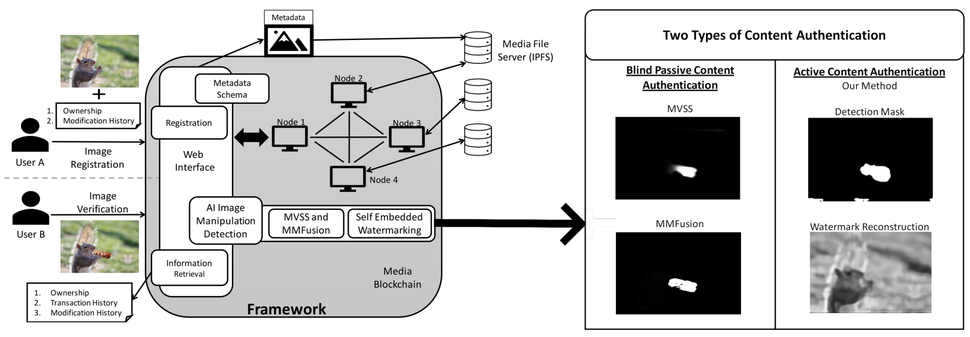

Figure: Overview of the proposed Framework with an expanded view of the methods used for Image manipulation detection.

The framework consists of four components:

- The **media blockchain** is responsible for recording the transactions performed on media assets as well as verifying the integrity of the assets.
- The **Content Authentication Engine** is responsible for the detection of modifications and the authentication of the image content. This framework will facilitate multiple manipulation detection algorithms, including a self-embedding watermarking method.
- A **distributed file server** manages the storage of registered images via the Inter Planetary File System (IPFS).
- A **graphical web interface** is provided to allow users to upload and register original or modified versions of media assets.

Results

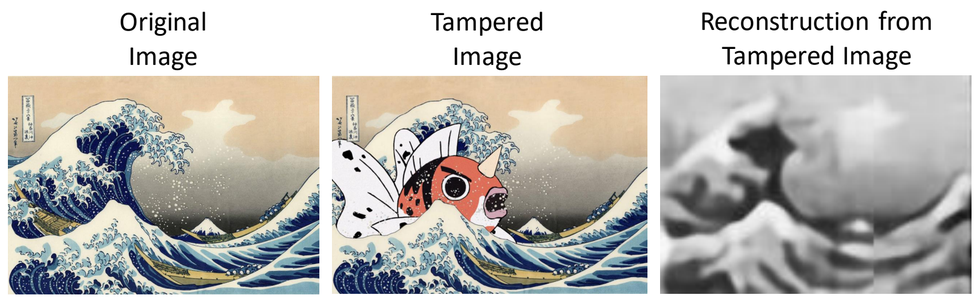

Figure: Compressed Sensing Reconstruction even after modification through our method. MVSS and MMFusion do not have this capability.

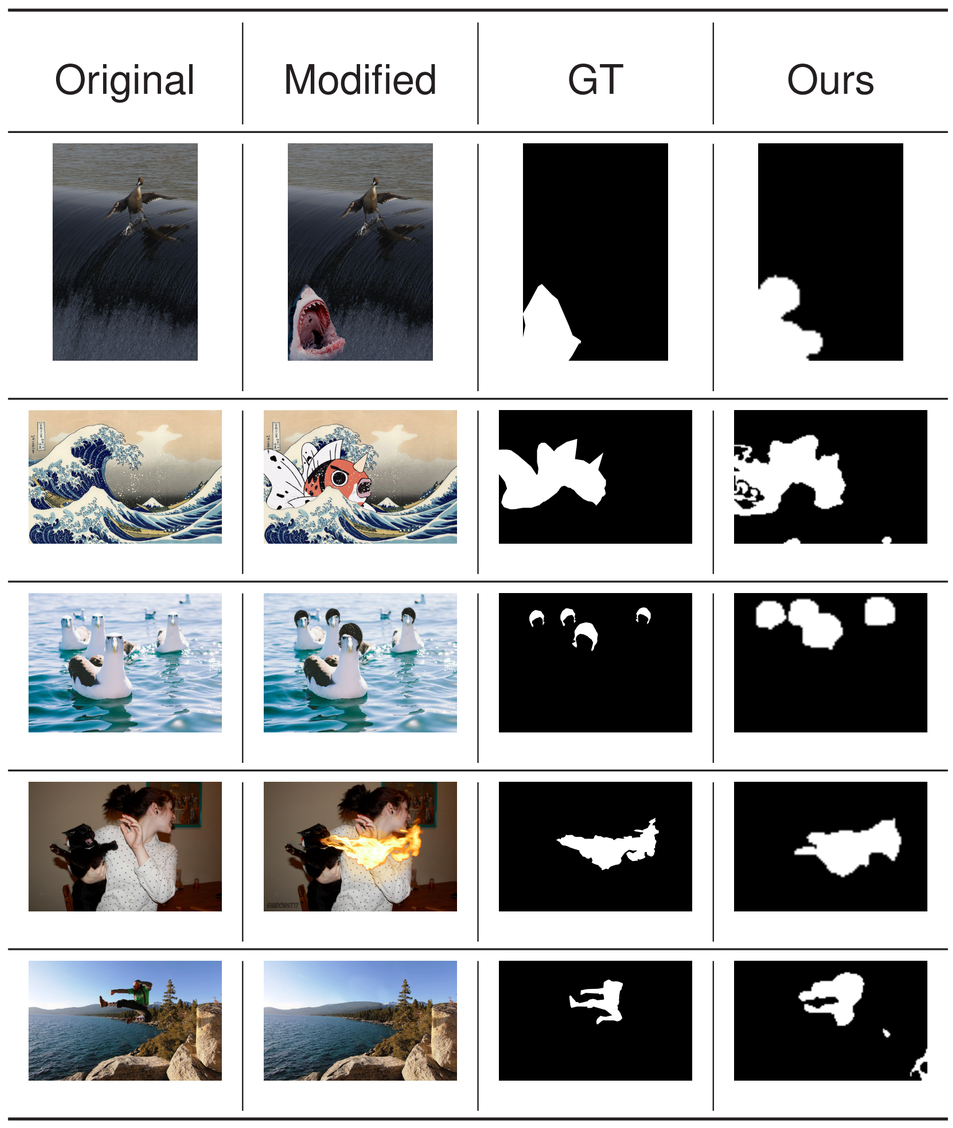

Figure: Comparison of the outputs on images 3, 11, 12, 7 and 18, from the IMD dataset with various modifications.

Conclusions

This work proposed a new framework facilitating content provenance for digital media assets. The framework consists of several parts, including a metadata schema, a media blockchain framework with an interplanetary distributed file system, and a content authentication engine. This framework provides capabilities to detect and localise manipulation and can structurally reconstruct the original media content even when it has been manipulated. The proposed framework aims to support the recent international standard JPEG Trust, and we envisage an impact on a large number of stakeholders, including end users across the world.

Acknowledgements
This work is jointly done with the JPEG international standardisation committee and supported by various funders, including EPSRC, Digital Catapult, Airbus Defence and Space, etc.