International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Multimodal Satellite Forensics Using Cross-Attention Transformer

Piyush Kumar Jha

IES
Ministry of Defence

Abstract

Satellite imagery is increasingly used in critical applications such as monitoring of environment, planning, disaster, and surveillance. However, rise of advanced generative models like Generative Adversarial Networks (GANs) and other image-editing tools, satellite images are exposed to manipulation. It is essential to ensure the authenticity and trustworthiness for decision-making. Multimodal Forensics (MMF) is a field dealing with the integrity of multimedia data. The paper presents a multimodal satellite imagery forensics that utilizes both Electro-Optical (EO) and Synthetic Aperture Radar (SAR) imagery to access the authenticity. The proposed model integrates dual Convolutional Neural Network (CNN) encoders with a Cross-Attention Transformer fusion module to detect tampering based on EO–SAR inconsistency. EO imagery are Panchromatic having high Spatial Resolution whereas SAR imagery has high Spectral Resolution. The model produces authenticity score and tamper heatmap.

Keywords: Generative adversarial networks (GANs); Multimodal Forensics (MMF); Electro-Optical (EO); Synthetic Aperture Radar (SAR); Convolutional Neural Network (CNN); Cross Attention Transformer; EO-SAR Consistency; Spatial Resolution; Spectral Resolution; Tamper Detection; Cross-Attention Transformer

1. Introduction

Satellite imagery is increasingly used in critical applications such as monitoring of environment, planning, disaster, and border surveillance for national security. However, rise of advanced generative models like Generative Adversarial Networks (GANs) and other image-editing tools, satellite images are exposed to manipulation and posing a significant risk to decision-making systems. Traditional forensic systems rely mainly on EO imagery, which is vulnerable to illumination, weather, and texture-related inconsistencies. SAR imagery, being independent of lighting and weather, provides an additional physical modality to enhance forensic analysis.

Despite their resemblance with digital pictures, satellite images represent signals with a very different lifecycle. They are inherently a multimodal data asset due to the presence of various families of sensors onboard satellites. Moreover, creating a satellite image involves a series of complex signal-processing operations. Data modalities describing different information than standard natural photographs and more

complicated processing pipelines imply that many forensic traces exploited in literature might be absent or not be performant enough to determine the authenticity of satellite images [1].

This paper introduces Cross Attention Transformer, which compares EO and SAR features extracted using CNN encoders to detect manipulation and produce authenticity score and tamper heatmap.

2. Literature Review

Remote sensing uses two common methods to measure electromagnetic radiation: (a) Active sensing, which emits their own energy to illuminate the object for sensing and (b) Passive sensing, which utlises sunlight energy for illumination of object. Electro-Optical (EO) imagery refers to images captured by passive sensors. Since, Passive sensors uses sunlight as source, it can capture wavelengths in the visible spectrum.

SAR imagery refers to images generated through active sensors. SAR system is a kind of radar that is installed on a moving platform, such as a satellite or aircraft. SAR system emits electromagnetic waves which is reflected by the object, which are then processed to generate a SAR image. EO imagery are Panchromatic having high Spatial Resolution whereas SAR imagery has high Spectral Resolution. Multimodal remote sensing has been widely studied for classification, segmentation, and change detection. However, multimodal forensics remains an emerging area.

The convolutional neural network functions much like a traditional feed-forward neural network, except that the operations in its layers are spatially organized with sparse (and carefully designed) connections between layers. The three types of layers that are commonly present in a convolutional neural network are convolution, pooling, and ReLU. The ReLU activation is no different from a traditional neural network. In addition, a final set of layers is often fully connected and maps in an application-specific way to a set of output nodes [7].

Transformer is a deep learning model which focus on the most relevant parts of the input. The Vision Transformer (ViT) is a modern deep learning model that applies the transformer framework—originally created for NLP—to visual tasks. Instead of using convolutions like traditional CNNs to learn local spatial patterns, ViT breaks an image into patches and uses self-attention to capture relationships across the entire image. By focusing on global interactions rather than only nearby features, Vision Transformers have achieved top-tier results in computer vision applications, including image recognition, object detection, and image segmentation [3].

Cannas et al. (2025) highlight the importance of detecting tampered EO satellite images using deep learning [1]. Transformers, introduced by Vaswani et al. (2017), enable efficient global feature interaction and serve as a foundation for cross-attention fusion used in this work [4]. Datasets such as EuroSAT [5] and SEN12MS [6] have been extensively used for multimodal learning tasks. The proposed work extends these ideas into a forensic framework.

3. Dataset and Preprocessing

To validate our approach, experiments are conducted using imagery downloaded from the EuroSAT [5] and SEN12MS [6] datasets. A separate set of tampered samples is produced through region replacement, patch splicing, object removal, and inpainting. The results demonstrate that our model accurately identifies tampered regions and yields clearer and more reliable heatmaps than CNN-based multimodal fusion baselines. The combination of multimodal consistency learning and cross-attention fusion greatly enhances detection sensitivity and forensic interpretability. All the raw EO and SAR images are processed and resized to 256×256 and normalized. Finally, each pair of Clean and Tampered images are labeled as clean and tampered for training purpose.

4. Methodology

we introduce an architecture designed to detect tampering in EO–SAR satellite image pairs. The model consists of two parallel CNN encoders that extract modality-specific deep features. These feature maps are then tokenized into patch embeddings and enriched with positional encodings. A **Cross-Attention Transformer Fusion module** is employed to compare EO and SAR patches in a bidirectional manner—EO queries attend to SAR keys and values, and SAR queries attend to EO keys and values. This mechanism enables the model to learn fine-grained consistency patterns at the patch level. After fusion, feature aggregation and classification produce a global authenticity score through a Sigmoid activation. Additionally, a spatial decoder generates a pixel-wise tamper heatmap that highlights the manipulated regions.

The proposed framework addresses several limitations of earlier methods. First, it leverages multimodal physical consistency, making it robust to environmental variations and generative tampering. Second, the transformer architecture provides global context-aware fusion, outperforming CNN-only methods that rely on local receptive fields. Third, the cross-attention mechanism yields improved interpretability, as the learned attention maps reveal inter-modal dependency and regions of inconsistency. Fourth, the system integrates both detection (binary classification) and localization (heatmap generation), offering a complete forensic pipeline. Fig 1. presents a comprehensive framework for Multimodal Satellite Forensics Using Cross-Attention Transformer.

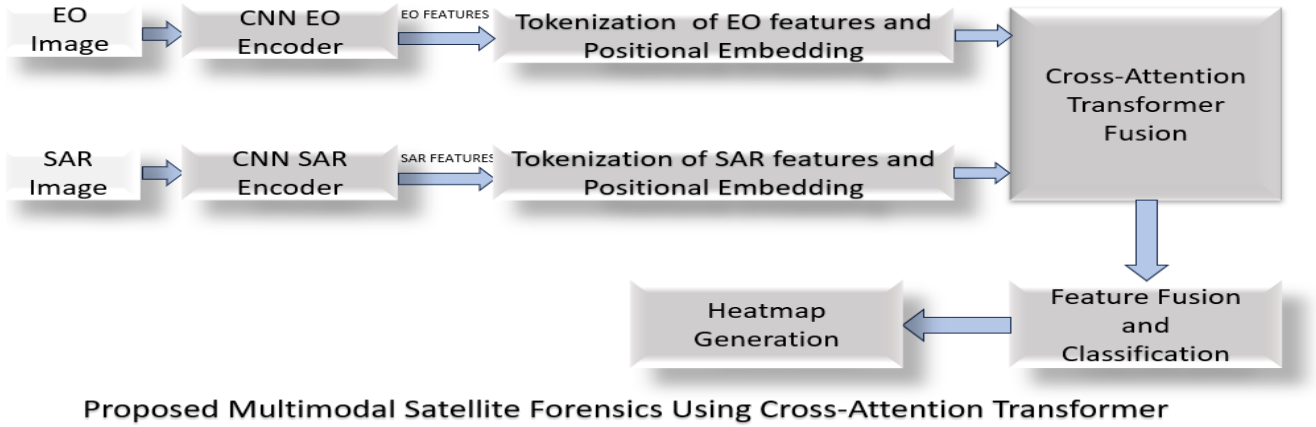

Fig 1. Block Diagram

4.1. CNN Encoders

Both EO and SAR images are passed through separate CNN encoders. The EO encoder extracts colour–texture features while the SAR encoder captures radar-intensity and structural patterns. Each encoder outputs a 512×8×8 feature map.

4.2. Tokenization and Positional Encoding

The 512×8×8 encoder feature maps are reshaped into 64 tokens, each of size 512. So, each token is 512-D vector. Positional encodings are added to retain spatial information prior to feeding tokens into the transformer.

4.3. Cross-Attention Transformer Fusion

4.3.1. EO → SAR Cross Attention

Transformer learns which SAR regions correspond to which EO patch

EO Tokens = Query (Q)

Where, $Q = Wq_{EO} * EO\_Tokens$

SAR Tokens = Keys (K) and Values (V)

Where, $K = Wk_{SAR} * SAR\_Tokens$

$V = Wv_{SAR} * SAR\_Tokens$

4.3.2. SAR → EO Cross Attention

Transformer learns which EO regions correspond to which SAR patch

SAR Tokens = Query (Q)

Where, $Q = Wq_{SAR} * SAR\_Tokens$

EO Tokens = Keys (K) and Values (V)

Where, $K = Wk_{EO} * EO\_Tokens$

$V = Wv_{EO} * EO\_Tokens$

The cross-attention mechanism uses the standard Q – K – V formulation:

$$Q = X * W_q$$

$$K = X * W_k$$

$$V = X * W_v$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where, k is the dimension of Q/K/ V. Softmax normalises the score, so that all are positive and add up to 1.

EO queries attend to SAR keys and values, and vice versa. This allows comparison of multimodal relationships to detect inconsistencies.

$$F_{EO2SAR} = \text{Attention}(Q, K, V)$$
$$F_{SAR2EO} = \text{Attention}(Q, K, V)$$

4.4. Feature Fusion and Classification

$$F_{fused} = \text{concatenation}(F_{EO2SAR}, F_{SAR2EO})$$

Fused features are aggregated using mean pooling and passed through a fully connected classifier to produce a global authenticity score using ReLU (Sigmoid) activation.

Classification : Output Ɛ [0,1] using Sigmoid

4.5. Tamper Heatmap

The decoder upsamples fused token activations back to the original resolution and applies convolution operations followed by Sigmoid to produce pixel-wise tamper probability heatmaps.

128 Tokens → 128 Channels X 8 X 8

Convolution (Conv) → ReLU → Upsample x 2 → Conv → ReLU → Upsample x 4 → Conv → Sigmoid →Final Heatmap (256 X 256)

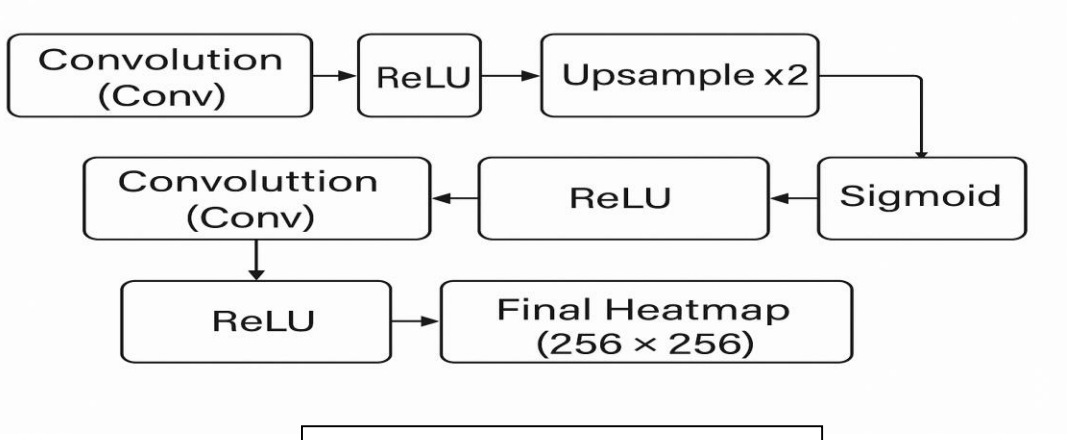

Fig 2. Heatmap Flow Diagram

5. Results

A Guided User Interface (GUI) interface built with Streamlit web application for real-time verification. Both EO and SAR image are uploaded on the web application to get the authenticity score and heatmap as output. The following figures show clean and tampered EO–SAR inputs along with prediction and heatmap:

Fig 3. EO_Clean_1

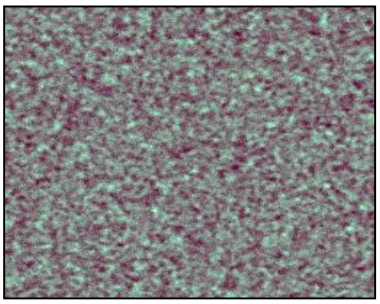

Fig 4. SAR_Clean_1

Fig5. Heatmap_Clean_1

Prediction: AUTHENTIC
Confidence Score: 0.0024

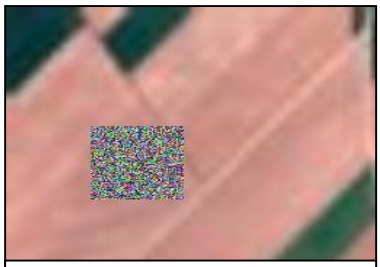

Fig 6. EO_Tampered_2

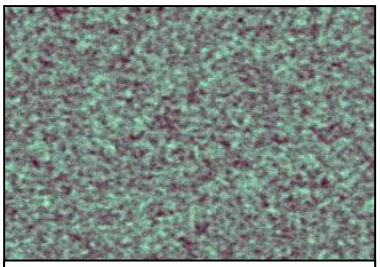

Fig 7. SAR_Tampered_2

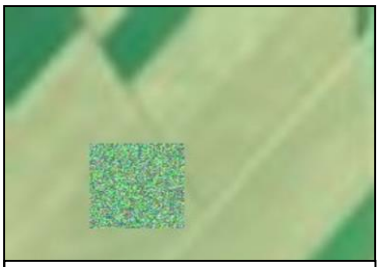

Fig 8. Heatmap_Tampered_2

Prediction: TAMPERED
Confidence Score: 0.9663

6. Conclusion

In summary, this paper presents a novel multimodal forensic framework leveraging cross-attention transformers to evaluate EO–SAR consistency for satellite image tamper detection. Through the integration of deep feature extraction, cross-modal alignment, and transformer-based fusion, the proposed model achieves improved robustness and accuracy compared to traditional approaches. The contributions of this work align closely with emerging research trends in multimodal learning and remote-sensing forensics, highlighting the importance of exploiting physical cross-modal relationships for image authenticity verification. Future work may include large-scale training, multimodal GAN detection, and temporal EO–SAR consistency analysis.

References

1. Cannas, E.D. (2025). Forensic Analysis of Satellite Imagery: Challenges and Solutions. In: Garatti, S. (eds) Special Topics in Information Technology. Springer Briefs in Applied Sciences and Technology. Springer, Cham (Published - 29 March 2025).
2. Cannas, E.D.: Multimedia forensics challenges in the multimodality data era. Ph.D. thesis, Politecnico di Milano (2024).
3. Vision Transformers (ViT) – Dosovitskiy et al., 2020.

4. Attention Is All You Need – Vaswani et al., 2017.
5. EuroSAT (EO dataset): https://zenodo.org/records/7711810
6. SEN12MS (SAR + EO pairs): https://mediatum.ub.tum.de/1474000
7. Aggarwal, Charu C. (2018). Neural networks and deep learning: A textbook ,Springer.
8. ESA: https://dataspace.copernicus.eu/data-collections/sentinel-data/sentinel-1
9. ESA: Sentinel-1 Mission User Guide. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar.
10. https://bhuvan-app3.nrsc.gov.in/data/download/index.php
11. Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In ACL, 2020.