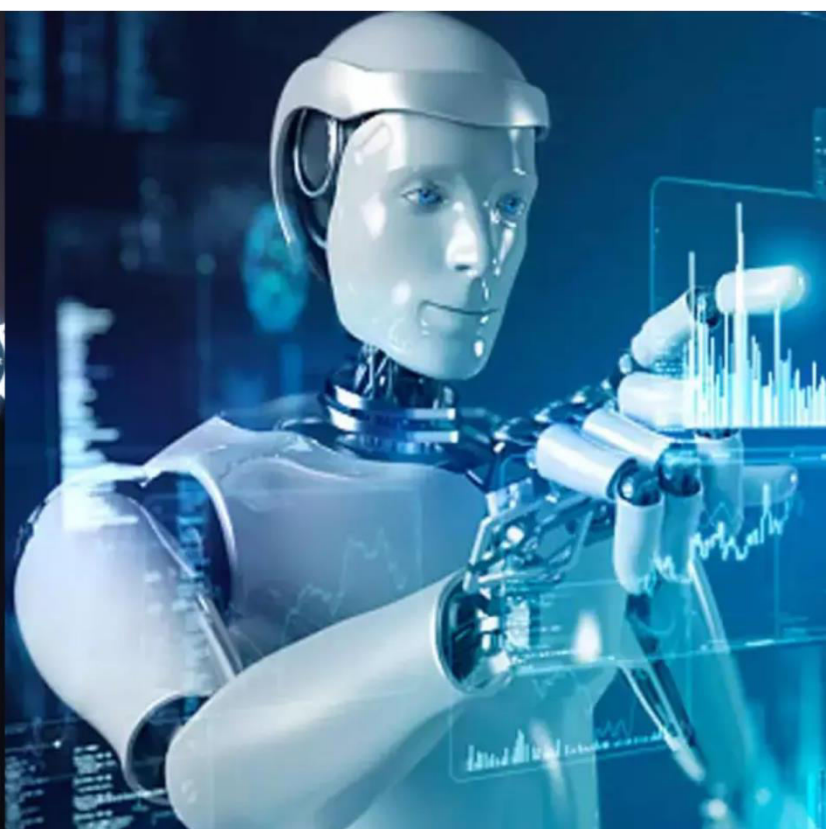




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 7, July 2026

Deep Learning-Based Image Forgery Detection and Localization: A Review of CNNs, Vision Transformers, Hybrid Models, and Weakly Supervised Methods

Er. Aarushi Thakur¹, Er. Neha Dhiman²

P.G. Student, Department of Computer Science and Engineering, Himachal Pradesh Technical University, Hamirpur,
Himachal Pradesh, India¹

Faculty, Department of Computer Science and Engineering, Himachal Pradesh Technical University, Hamirpur,
Himachal Pradesh, India²

ABSTRACT: The rapid advancement of image manipulation techniques, including copy-move, splicing, and AI-generated forgeries, has intensified the need for reliable image forgery detection and localization methods. Recent deep learning approaches have significantly improved performance through convolutional encoder-decoder architectures, multi-scale feature fusion models, transformer-based global attention mechanisms, contrastive learning frameworks, and edge-guided localization strategies. Although these methods achieve strong results on benchmark datasets such as CASIA and CoMoFoD, challenges remain in cross-dataset generalization, interpretability, and robustness to diverse manipulation types. Furthermore, heavy reliance on pixel-level supervision limits scalability in practical scenarios. This review systematically analyzes state-of-the-art techniques from 2023 to 2025, categorizing them based on architectural design, supervision strategy, and feature modeling mechanisms. A comparative evaluation of benchmark datasets and performance metrics is presented, highlighting emerging trends toward hybrid CNN-Transformer architectures, multi-scale contextual modeling, and weakly supervised frameworks. The paper concludes by identifying open research challenges and future directions for developing more robust and generalizable forgery localization systems.

KEYWORDS: Image forgery detection, Deep learning, Multi-scale modelling, CNN-Transformer architectures, Weakly supervised localization, Benchmark datasets.

I. INTRODUCTION

With the current rapid development of AI and the image editing technology, it is now extremely simple to play around with photos. The latest AI-generated models and generative art have the capability of producing images that appear close to the real one, making it increasingly difficult to tell the fake and the authentic photo. Due to it, the dissemination of modified images has gone viral in the recent few years.

Photos are one of the most significant means of information exchange and recording of events, and it is of great importance that they remain authentic. False pictures have the potential to propagate falsehood and create major problems in the fields of journalism, social media, politics and even in forensic investigations. That is why the problem of detecting and locating forgeries has become a massive research agenda in digital image forensics.

A visual representation composed of pixels arranged into a grid that is two dimensional can be called an image. Image forgery can be defined as an act of deliberate alteration of an image to change or obscure intentional information to mislead the audience.

Common types of image forgery include image splicing, copy-move forgery, object removal (inpainting), image retouching, and AI-generated synthetic images. Among these, splicing and copy-move forgery remain widely studied in forensic research [1,14]. The major categories of image forgery are illustrated in Figure 1.

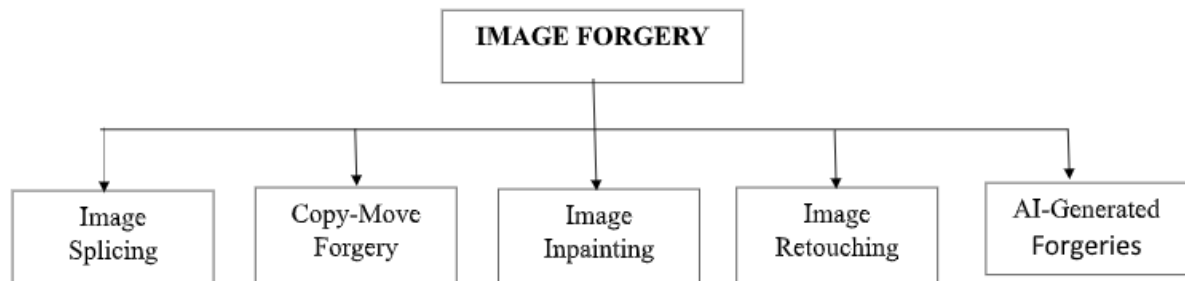


Figure.1. General Classification of Image Forgery Types

Traditional forgery detection techniques relied on manual creation of forensic features to identify forged documents. The methods showed good performance during controlled experiments but suffered from two major issues because they could not handle post-processing changes and their data-driven testing showed poor performance across various datasets. The past few years saw deep learning-based methods achieve significant progress in detecting and locating forged documents. Convolutional Neural Networks (CNNs) are widely used to learn hierarchical representations for manipulation detection [1,3,6], while multi-scale feature fusion mechanisms improve contextual understanding and pixel-level localization accuracy [1,4,13].

To improve feature discrimination, contrastive learning approaches have been employed to separate authentic and tampered feature representations in deep space [2]. Transformer-based architectures have further enhanced forgery localization by modeling long-range dependencies and global contextual relationships, particularly in hybrid CNN–Transformer frameworks [12-13]. Additionally, edge-guided and noise-aware multi-stream models have been proposed to strengthen boundary detection and manipulation trace extraction [10-11].

Most supervised deep learning models need pixel-level annotations as their primary requirement, which results in expensive annotation costs and creates architectural challenges that decrease model interpretability. The research community has started to focus more on weakly supervised methods and unsupervised methods. Weakly supervised approaches utilize image-level labels for localization [5,9], while unsupervised methods treat forgery detection as an anomaly or distribution inconsistency problem in feature space [8].

This review systematically analyses recent deep learning-based forgery detection and localization techniques, highlighting advances in multi-scale modelling, hybrid architectures, contrastive learning, and reduced-supervision frameworks, while identifying existing challenges and research directions.

II. REVIEW OF LITERATURE

The evolution of image forgery detection has transitioned from handcrafted forensic feature-based techniques to advanced deep learning architectures, including multi-scale CNN models, hybrid CNN–Transformer frameworks, and reduced-supervision approaches. To systematically analyze recent developments, the literature is categorized into (i) CNN-based and multi-scale architectures, (ii) contrastive and representation learning methods, (iii) hybrid CNN–Transformer and multi-stream models, and (iv) weakly supervised and unsupervised frameworks.

CNN-Based and Multi-Scale Architectures

Liang et al. [1] proposed MFAF-Net, a multi-scale feature attention fusion framework for image splicing detection. The model employs an encoder–decoder architecture in which hierarchical convolutional features are fused using attention mechanisms to enhance contextual representation. Evaluated on the CASIA dataset, the model achieved competitive pixel-level F1 and IoU scores. However, the method primarily relies on RGB-based feature aggregation and attention refinement without explicitly modeling semantic deviation between local and global representations.

Liu et al. [3] introduced a hierarchical progressive UNet-based method for forgery detection and localization. The architecture improves feature propagation through progressive refinement stages, enhancing mask prediction accuracy. Although the model demonstrates improved localization on CASIA, it remains heavily dependent on pixel-level supervision.

Niu et al. [4] proposed a guided-noise and multi-scale feature aggregation framework to enhance manipulation trace detection. By integrating noise-aware representations with multi-resolution fusion, the model improves boundary sensitivity. Nevertheless, the architecture increases computational complexity and lacks explicit semantic consistency modeling.

Qin et al. [13] developed a multi-scale deeply supervised network based on U²-Net. The method incorporates deep supervision at intermediate decoder stages and a transformer module to capture long-range dependencies. Experimental results on CASIA and COLUMB datasets demonstrate strong F1 performance. However, global contextual modeling is achieved through attention aggregation rather than statistical inconsistency modeling.

Contrastive and Representation Learning Methods

Niloy et al. [2] proposed CFL-Net, which applies contrastive learning to distinguish tampered and authentic regions in feature space. Patch-level embeddings are trained using contrastive objectives to enhance representation separability. While the approach improves feature discrimination and localization robustness, it requires careful pair construction and remains supervised. The method was evaluated on benchmark datasets such as CASIA, demonstrating improved localization consistency through contrastive region-level embedding separation.

Sheng et al. [9] introduced ViLaCo, a vision-language collaborative framework for weakly supervised forgery localization. By leveraging semantic alignment between visual and textual embeddings, the method achieves region-level prediction using image-level labels. Despite reducing annotation dependency, the integration of multimodal alignment increases model complexity. Experimental results show competitive performance under weak supervision, highlighting the effectiveness of cross-modal feature alignment in forgery localization tasks.

Hybrid CNN–Transformer and Multi-Stream Models

Sharma et al. [12] proposed FLTNet, a hybrid CNN–Transformer architecture combining convolutional feature extraction with window-based self-attention to capture both local and global dependencies. The model achieved strong IoU and F1 scores on CASIA and CoMoFoD datasets. However, transformer integration increases computational cost and model complexity. The integration of transformer blocks enables enhanced global dependency modeling, contributing to improved cross-dataset generalization performance.

Zhang et al. [10] presented a multi-scale edge-guided framework incorporating self-supervised forgery generation and adversarial training. The approach enhances boundary localization by focusing on transition regions between authentic and forged areas. While effective, the training pipeline is comparatively complex.

Yin et al. [11] proposed MMFD-Net, a multi-branch architecture integrating image-level classification, pixel-level localization, and edge-level supervision. Multi-stream feature fusion improves robustness across datasets, yet the architecture becomes heavy and less interpretable due to multiple auxiliary branches.

Weakly Supervised and Unsupervised Approaches

Tăntaru et al. [5] proposed a weakly supervised deepfake localization method relying solely on image-level labels. The framework reduces dependence on pixel-level annotations but may experience reduced localization precision compared to fully supervised methods. The method demonstrates the feasibility of region-level prediction without pixel-level masks, offering scalability advantages for large datasets.

Yin et al. [8] introduced an unsupervised ViT-VAE framework combined with Gaussian mixture modeling to detect forgery as an anomaly in latent feature space. The method eliminates the need for labeled masks; however, reconstruction-based detection may struggle with subtle manipulations.

Badar et al. [14] presented a multiscale feature fusion approach for copy–move forgery detection using CNN feature extraction and clustering mechanisms. While effective for duplication-based manipulations, the approach is limited to specific forgery types and does not provide end-to-end segmentation capability. The clustering-based detection mechanism effectively identifies duplicated regions but lacks generalized localization capability for complex manipulations.

Overall, recent literature demonstrates a shift toward integrating multi-scale contextual modeling with global dependency mechanisms. However, challenges remain in balancing localization precision, computational efficiency, and generalization across manipulation types.

Benchmark Datasets Used in Image Forgery Detection

There are several publicly available benchmark datasets that are commonly used to compare image forgery detection and localization algorithms. The datasets differ in terms of types of manipulations, annotation schemes, and protocols. Table 1 lists the most used datasets in recent literature.

Table 1: Benchmark Datasets Used in Image Forgery Detection

Dataset	Forgery Type	Annotation	Common Usage
CASIA v1.0 / v2.0	Splicing, Copy–Move	Pixel-level mask	Most widely used benchmark
MICC-F220 / F600	Copy–Move	Pixel-level mask	Cross-dataset evaluation
CoMoFoD	Copy–Move	Pixel-level mask	Robustness testing
Columbia	Splicing	Pixel-level mask	Splicing localization
NIST	Multiple manipulations	Pixel-level mask	Real-world evaluation

Comparative Analysis of Reviewed Methods

To provide a quantitative comparison of recent forgery detection and localization models, key performance metrics reported in the literature are summarized in Table 2, while pixel-level F1 scores on the CASIA dataset are illustrated in Figure 2. The comparison highlights performance variations among multi-scale CNN models, hybrid CNN–Transformer frameworks, and edge-guided architectures.

Table.2. Comparison of Reviewed Image Forgery Detection Methods

Ref	Year	Model Used	Category	Supervision	Dataset(s)	Performance
[1]	2024	MFAF-Net	Multi-scale CNN	Supervised	CASIA	AUC $\approx$ 0.91
[2]	2023	CFL-Net	Contrastive CNN	Supervised	CASIA	Improved AUC over baseline
[3]	2024	HPUNet	Hierarchical UNet	Supervised	CASIA	High localization F1
[4]	2024	Guided-Noise Network	Multi-scale + Noise	Supervised	CASIA, NIST	F1 $\approx$ 0.66
[5]	2024	Weakly-Supervised Deepfake	Weak Supervision	Weakly Supervised	Diffusion datasets	IoU $\approx$ 0.37 (weak supervision)
[6]	2024	VGG16-UNet	Transfer Learning CNN	Supervised	CASIA	IoU $\approx$ 0.80
[7]	2025	MF-Net	Multi-stream CNN	Supervised	FF++, Celeb-DF	AUC $\approx$ 0.99
[8]	2025	ViT-VAE-GMM	Unsupervised	Unsupervised	CASIA,	AUC $\approx$ 0.93

					DSO	
[9]	2025	ViLaCo	Weakly Supervised + V-L	Weakly Supervised	CASIA	F1 $\approx$ 0.59
[10]	2025	Edge-Guided MEAC	Edge-guided CNN	Supervised	CASIA, NIST	F1 $\approx$ 0.74
[11]	2025	MMFD-Net	Multi-branch CNN	Supervised	CASIA	AUC $\approx$ 0.87
[12]	2025	FLTNet	Hybrid CNN–Transformer	Supervised	CASIA, CoMoFoD	IoU $\approx$ 0.82
[13]	2025	MSDS-Net	Multi-scale + Transformer	Supervised	CASIA	F1 $\approx$ 0.89
[14]	2025	MSF-CMFD	Copy-Move Detection	Supervised	CASIA, Columbia	Accuracy $\approx$ 95%

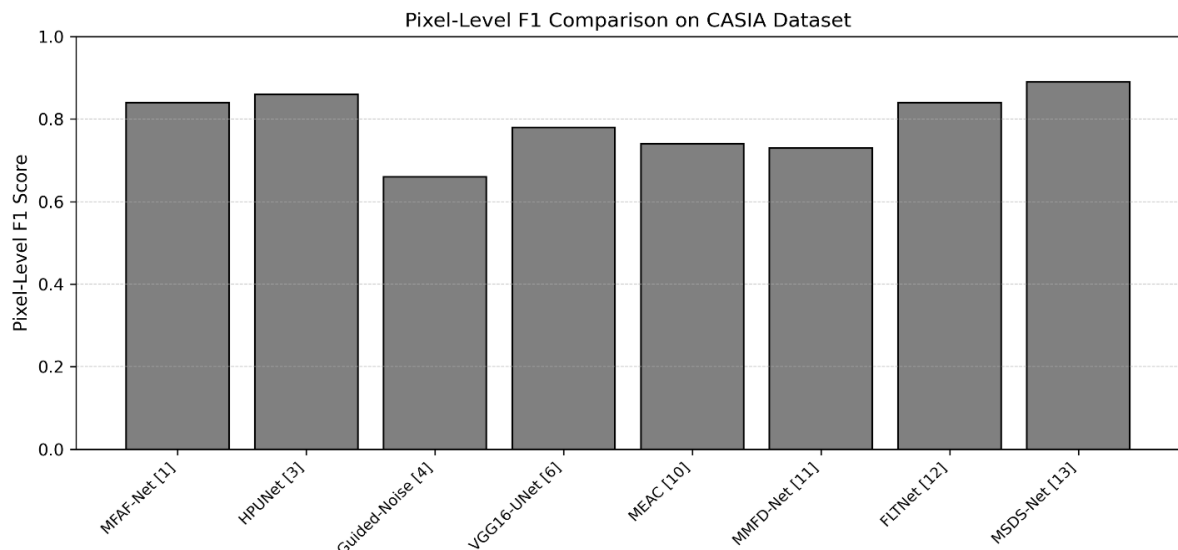


Figure.2. Pixel-Level F1 Score Comparison of Reviewed Methods on the CASIA Dataset

As shown in Figure 1, multi-scale transformer-enhanced architectures such as MSDS-Net [13] and hybrid models like FLTNet [12] demonstrate comparatively higher pixel-level F1 scores. Earlier multi-scale CNN models achieve competitive performance, whereas noise-guided and edge-focused architectures show moderate improvements. The comparison indicates that incorporating global contextual modeling contributes to enhanced localization accuracy.

III. ARCHITECTURE OF IMAGE FORGERY DETECTION SYSTEMS

Modern image forgery detection and localization systems typically follow a structured deep learning pipeline. Although architectural designs vary across different studies, most approaches share common stages including dataset preparation, feature extraction, classification or segmentation modules, and performance evaluation. This section presents a generalized overview of the architectural components commonly adopted in recent literature.

Data Preparation

Image forgery detection systems are typically trained and evaluated using benchmark datasets such as CASIA, MICC, CoMoFoD, Columbia, and NIST. These datasets contain authentic and manipulated images, often accompanied by pixel-level ground-truth masks for supervised localization tasks. Before training, images are resized to a fixed resolution and normalized. Data augmentation techniques, including rotation, flipping, scaling, and compression simulation, are commonly applied to improve generalization and robustness across varying manipulation conditions.

Feature Extraction Module

Feature extraction forms the core of modern forgery detection architectures. Most approaches employ Convolutional Neural Networks (CNNs) to learn hierarchical representations capable of capturing manipulation artifacts such as texture inconsistencies and boundary irregularities [1,3,6]. Multi-scale feature fusion strategies further enhance contextual modeling by aggregating features from different resolution levels [1,4,13]. Recent methods incorporate transformer-based modules to model long-range dependencies and global contextual relationships [12-13]. Additionally, edge-guided and multi-stream frameworks integrate auxiliary cues, such as noise or edge features, to improve sensitivity to subtle forgery traces [10-11].

Classification and Localization Module

Depending on the objective, image forgery detection systems perform:

1. **Image-level classification**, where the model predicts whether an image is authentic or forged.
2. **Pixel-level localization**, where a segmentation mask identifies manipulated regions.
3. **Multi-task learning**, where both classification and localization tasks are jointly optimized.

Encoder-decoder architectures such as UNet and its variants are widely used for pixel-level localization [3,6]. These architectures utilize skip connections to preserve spatial information during reconstruction. Loss functions commonly employed include Binary Cross-Entropy (BCE), Dice loss, Intersection-over-Union (IoU) loss, or hybrid combinations to improve segmentation accuracy.

Evaluation Metrics

The performance of forgery detection systems is evaluated using standard classification and segmentation metrics.

For **image-level detection**, commonly used metrics include Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC).

For **pixel-level localization**, evaluation typically relies on Pixel Precision, Pixel Recall, Pixel F1-score, and Intersection over Union (IoU). These metrics measure how accurately the predicted mask aligns with the ground-truth manipulated region.

Recent studies also emphasize cross-dataset evaluation to assess model generalization and robustness across different manipulation distributions.

IV. SUMMARY AND OPINION

Recent advances in image forgery detection and localization have shown a definite shift from traditional handcrafted forensic features to deep learning-based multi-scale and hybrid models. Encoder-decoder CNN architectures remain the backbone for pixel-level localization, with multi-scale feature aggregation improving global understanding. The addition of transformer layers further enhances global dependency modeling.

Nevertheless, the current methods still have certain trade-offs. The transformer attention and multi-stream fusion methods are also more localized but with more computational complexity and less interpretable. The dependence on annotations can be mitigated by using weakly supervised and unsupervised methods. Nevertheless, these methods never have as much superiority to the use of fully supervised methods in localizing boundaries. The fact remains that cross-dataset generalization is a problem, and demonstrates that it is sensitive to dataset-specific patterns of manipulation.

The latest trend in studies favours hybrid modeling which integrates local feature extraction alongside global contextual reasoning. The key to further advancement is in the areas of robustness, reduced supervision, and efficiency without compromising localization accuracy.

V. CONCLUSION

This review examines the latest advances in deep learning techniques for image forgery detection and localization. It focuses on the latest developments in multi-scale modeling, hybrid CNN-Transformer models, and weakly supervised learning. In fact, the detection accuracy has improved on the conventional benchmarks, but some of the issues such as generalization, annotation dependency and computational complexity still exist.

The rising areas of research suggest the need for more robust architectures that can effectively process various types of manipulation. Future directions, specifically improvements in cross-dataset evaluation, the encouragement of lightweight methods, and the design of interpretable models, will be crucial in the development of forgery localization systems.

REFERENCES

- [1] E. Liang, K. Zhang, Z. Hua and X. Jia, "Multi-Scale Feature Attention Fusion for Image Splicing Forgery Detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 1, Article 18, 2024.
- [2] F. F. Niloy, K. K. Bhaumik and S. S. Woo, "CFL-Net: Image Forgery Localization Using Contrastive Learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 3435–3444.
- [3] Y. Liu, X. Li and J. Zhang, "Hierarchical Progressive Image Forgery Detection and Localization Method Based on UNet," *Big Data and Cognitive Computing*, vol. 8, 2024.
- [4] Y. Niu, P. Chena, L. Zhanga, L. Tana and Y. Chena, "Image Forgery Localization via Guided Noise and Multi-Scale Feature Aggregation," *Journal of Information Security and Applications*, 2024.
- [5] D.-C. Țăntaru, E. Oneață and D. Oneață, "Weakly-Supervised Deepfake Localization in Diffusion-Generated Images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.
- [6] R. R. Choudharya, S. Paliwala and G. Meenaa, "Image Forgery Detection System Using VGG16 UNET Model," *Procedia Computer Science*, vol. 235, pp. 735–744, 2024.
- [7] H. Duan, Q. Jiang, X. Jin, M. Wozniak, Y. Zhao, L. Wu, S. Yao and W. Zhou, "MF-Net: Multi-feature fusion network based on two-stream extraction and multi-scale enhancement for face forgery detection," *Complex & Intelligent Systems*, 2024.
- [8] H. Yin, K. U, J. Wang and W. Ma, "Toward Adaptive Unsupervised and Blind Image Forgery Localization with ViT-VAE and a Gaussian Mixture Model," *Mathematics*, vol. 13, 2025.
- [9] Z. Sheng, J. Wu, W. Lu and J. Zhou, "ViLaCo: Weakly-Supervised Image Forgery Localization via Vision-Language Collaborative Reasoning," *arXiv preprint*, 2025.
- [10] H. Zhang, J. Zeng and J. Yang, "Multi-Scale Edge-Guided Image Forgery Detection via Improved Self-Supervision and Self-Adversarial Training," *Electronics*, vol. 14, 2025.
- [11] H. Yin, K. U, J. Wang and Z. Gan, "MMFD-Net: A Novel Network for Image Forgery Detection and Localization via Multi-Stream Edge Feature Learning and Multi-Dimensional Information Fusion," *Mathematics*, vol. 13, 2025.
- [12] S. Sharma, B. K. Singh, and H. Garg, "FLTNet: Robust Image Forgery Localization Using Hybrid CNN–Transformer Synergy," *Computers, Materials & Continua*, vol. 82, no. 3, pp. 3659–3676, 2025.
- [13] S. Qin, C. Liang, Y. Luo, J. liu, Q. Fu and X. Ouyang, "Multi-Scale and Deeply Supervised Network for Image Splicing Localization," *Frontiers in Artificial Intelligence*, vol. 8, 2025.
- [14] P. Badar, G. Geetha and T.R. Mahesh, "Multiscale Feature Fusion for Robust Copy-Move Forgery Detection in Digital Images," *Journal of Information Systems Engineering and Management*, vol. 10, no. 2, 2025.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details