Pixelyse : ViT-VAE for Document Forgery Detection

Mishal Rose Thankachan
Dept of Computer Science and Engineering
Amal Jyothi College of Engineering, Kottayam, India
mishalrosethankachan2025@cs.ajce.in

Joshua John Sajit
Dept of Computer Science and Engineering
Amal Jyothi College of Engineering, Kottayam, India
joshuajohnsajit2025@cs.ajce.in

Merwin Maria Antony
Dept of Computer Science and Engineering
Amal Jyothi College of Engineering, Kottayam, India
merwinmariaantony2025@cs.ajce.in

Richa Maria Biju
Dept of Computer Science and Engineering
Amal Jyothi College of Engineering, Kottayam, India
richamariabiju2025@cs.ajce.in

Bini M Issac
Dept of Computer Science and Engineering
Amal Jyothi College of Engineering, Kottayam, India
binimissac@amaljyothi.ac.in

***Abstract*—Ensuring the authenticity of documents is more important than ever, as forgery techniques continue to evolve. Traditional methods, which rely on predefined rules and handcrafted features, often struggle to adapt to new types of fraud. To address this, we propose a Vision Transformer-based Variational Autoencoder (ViT-VAE) designed to enhance document authentication. By combining the Vision Transformer's ability to capture intricate details with the Variational Autoencoder's capability to model genuine document patterns, our approach effectively detects anomalies based on reconstruction errors. This fusion of self-attention mechanisms and probabilistic modeling improves accuracy and adaptability in identifying forged elements. Our experiments on diverse datasets show that ViT-VAE outperforms conventional machine learning and deep learning methods, offering a more reliable solution for document security. These findings open the door for further advancements in fraud detection and verification technologies, strengthening trust in digital and physical documentation.**

I. Introduction

Documents serve as a major mode of communication, authentication, and documentation in various fields, including banking, legal transactions, and identity checks. However, as forgery techniques grow ever more advanced, detecting fraudulent manipulations of documents has become a formidable task [10]. Conventional approaches to verification relate to handcrafted features, statistical analysis, and rule-based techniques. [19] Although they had met with some success, either because of their operating principles or ultimately in what they offer, they run far behind the increasingly complex and adaptive forgeries, making them ineffective in real-world use. [6]

Deep learning is one of the great advancements that has revolutionized image processing and pattern recognition, offering great possibilities for document forgery detection [5]. Convolutional Neural Networks (CNNs) have been the most actively researched approach for this task due to their ability to capture spatial features and patterns in images. However, CNNs are limited in having dependencies related to long-range and global contextual understanding, which are crucial for differentiating genuine from fabricated documents [17]. Such limitations have inspired the explorations of alternative architectures that could capture both local and global document features better. .

Vision Transformers (ViTs) are a recent and strong alternative to CNNs using self-attention mechanisms to model rich spatial dependencies in images. By combing VAEs proven for strong anomaly detection, a powerful framework can be implemented for recognizing forged documents [9]. The ViT will draw relevant spatial features within the input data, while the VAE will glean normal documents' distribution to recognize a forgery using reconstruction errors. Such combination ensures the model's generalization capability over diverse forgery techniques and document styles.

The goal of this research work is the use of ViT-VAE for document forgery detection. A custom dataset has been constructed containing both genuine and forged documents in order to evaluate our model's efficiency. Experiments show that our method outperforms traditional machine-learning models and deep learning-based classifiers with great advantages in terms of accuracy, robustness, and adaptability [15]. Our work sheds light on a vast area that can be worked within the frame of transformers and calls the attention of researchers working in fraud detection systems. .

II. Related Work

With advancements in deep learning and computer vision, image forgery detection has become an increasingly important field. Okamoto et al. [1] introduced an innovative approach to document forgery detection by using synthetic data generation to train deep learning models. Their study highlights the importance of exposing models to a diverse range of synthetic datasets with different forgery patterns, allowing them to

generalize more effectively to real-world scenarios. Similarly, Liao et al. [10] proposed a Character Texture Perception Network (CTP-Net) designed to detect document forgeries by analyzing character textures. Their approach enables the model to spot subtle inconsistencies in text structure, demonstrating the significance of texture-based features in forgery detection and paving the way for future research in this area.

Building on these advancements, Chen et al. [2] explored forgery localization through an anomaly detection framework using Variational Autoencoders (VAE) and Vision Transformers (ViT). Their approach stands out because it does not rely on explicitly labeled training data, instead learning to detect forgery traces autonomously. Experimental results showed that this method outperformed traditional supervised models, particularly in scenarios where annotation data is limited.

Guillaro et al. [3] took a different approach with TruFor, a comprehensive multi-cue framework for image forgery detection and localization. Unlike traditional methods that rely on a single type of analysis, TruFor combines frequency analysis, noise inconsistencies, and deep feature embeddings to enhance detection accuracy. By leveraging multiple cues, their framework significantly reduces false positives and improves localization precision, making it more robust than conventional CNN-based models.

Choudhary et al. [11] explored encoder-decoder architectures for forgery detection, specifically utilizing the VGG16-UNET framework. Their findings indicate that integrating pretrained feature extractors with segmentation models enhances the accuracy of forgery localization—an essential capability for practical forensic applications. This approach demonstrates the value of combining feature extraction with segmentation to improve detection performance.

When it comes to detecting copy-move forgeries, Lee et al. [12] introduced a CNN-based technique incorporating rotation-invariant wavelet features. One of the biggest challenges in copy-move forgery detection is ensuring robustness against geometric transformations, and their method effectively addresses this issue. By making detection more resistant to rotation-based alterations, their approach enhances the reliability and applicability of forensic techniques in real-world scenarios.

Finally, some foundational contributions have shaped the field of image forgery detection. He et al. [13] introduced the Deep Residual Learning framework (ResNet), which has since become a cornerstone in many computer vision tasks, including forgery detection. ResNet's success in feature extraction and classification has influenced numerous modern deep-learning models for detecting image manipulations. These advancements collectively underscore the need for more interpretable, scalable, and robust solutions to combat digital forgeries effectively.

III. Methodology

A. ViT-VAE Anomaly Detection Framework

Most forgeries occupy less than 10% of a document page, as supported by forensic studies. Document images have characteristic structures such as text alignment, stroke continuity, and background consistency. To effectively detect manipulations, we segment the document into smaller parts called cliques, enabling better learning of local inconsistencies.

The ViT-VAE pipeline consists of three main steps. The first step, **Multi-Modal Feature Extraction**, extracts two feature maps that capture traces of forgery beyond pixel-level differences. The second step, **ViT-VAE Learning**, utilizes Variational Autoencoders (VAE) and Vision Transformers (ViT) to reconstruct document cliques, amplifying the distinction between authentic and tampered regions. The final step, **Anomaly Localization**, computes pixel-wise anomaly scores and generates a forgery heatmap to highlight manipulated regions.

B. Multi-Modal Feature Extraction

For a color document image $I \in \mathbb{R}^{H\times W\times 3}$, we extract three forensic feature maps that highlight tampering traces beyond pixel-level differences.

The first feature, **Noiseprint Feature Map** $F_1 \in \mathbb{R}^{H\times W}$, captures inconsistencies in image acquisition noise as described in [19]. This helps in detecting subtle anomalies introduced during forgery operations.

The second feature, **High-Pass Filtering Residuals** $F_2 \in \mathbb{R}^{H\times W\times 3}$, extracts high-frequency artifacts caused by tampering and post-processing .

The third feature, **Laplacian Edge Map** $F_3 \in \mathbb{R}^{H\times W\times 3}$, highlights unnatural edge discontinuities resulting from copy-move, splicing, or text modifications.

These three feature maps are processed and concatenated into a composite representation:

$$F = [F_1; T(F_2); T(F_3)] \in \mathbb{R}^{H\times W\times 3}$$

where $T(\cdot)$ converts RGB colormaps to grayscale.

C. ViT-VAE Learning for Document Analysis

Learning effective representations is crucial for robust forgery detection. To be compatible with the standard Vision Transformer (ViT) [17], we segment the input feature representation F into non-overlapping patches of size $16 \times 16 \times 3$. A group of consecutive 4×4 patches forms a **clique**, which serves as the fundamental processing unit in ViT-VAE. The Variational Autoencoder (VAE) is used to learn the compact representation of each clique, while ViT captures dependencies among patches within a clique.

Using a stride size S, the total number of cliques is given by:

$$N_C = \left\lfloor \frac{H-64}{S} \right\rfloor \times \left\lfloor \frac{W-64}{S} \right\rfloor$$

Each clique $C \in \mathbb{R}^{64\times 64\times 3}$ is reshaped into $N = 16$ vectors, denoted as $[c_1, \ldots, c_N]$, where each patch vector c_j is flattened into a 768-dimensional vector.

A linear projection E is applied to map each patch vector into a ViT token of dimension $D = 1024$. Following [17], a

class token c_{class} is appended, and position embeddings $E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ are added to retain spatial information. The input to the Transformer is then formulated as:

$$z^{(0)} = [c_{\text{class}}, c_1 E, \ldots, c_N E] + E_{\text{pos}}$$

A Transformer encoder with $L = 6$ layers processes these tokens, where each layer consists of multi-head self-attention (MSA), a multi-layer perceptron (MLP), layer normalization (LN), and residual connections:

$$\tilde{z}^{(l)} = \text{MSA}(\text{LN}(z^{(l-1)})) + z^{(l-1)}$$

$$z^{(l)} = \text{MLP}(\text{LN}(\tilde{z}^{(l)})) + \tilde{z}^{(l)}, \quad l \in \{1, \ldots, L\}$$

After passing through the Transformer layers, the first token $z_0^{(L)}$ is mapped to a latent vector z of size $2K$ (e.g., $K = 10$), which represents the mean μ_i and standard deviation σ_i of a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ in the VAE latent space.

The latent vector is sampled as:

$$z_i' = \mu_i + \sigma_i \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \quad i \in \{1, \ldots, K\}$$

The sampled latent vector is passed through the VAE decoder, which consists of two fully connected layers, each followed by Batch Normalization and GeLU activation [30], to reconstruct the clique $C' \in \mathbb{R}^{64\times64\times3}$.

The ViT-VAE is trained by minimizing the reconstruction loss and Kullback-Leibler divergence (KL) loss:

$$\mathcal{L} = \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KLD}}$$

where:

$$\mathcal{L}_{\text{REC}} = \|C - C'\|_2^2$$

$$\mathcal{L}_{\text{KLD}} = -\frac{1}{2}\sum_{i=1}^{K}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

The training runs for a maximum of $E_{\max} = 20$ epochs, with early stopping if the loss difference between consecutive epochs falls below 10^{-4}.

D. Anomaly Localization

Once the ViT-VAE learning stage is completed, the network parameters are frozen, and all cliques are passed through the ViT-VAE again to compute their reconstruction errors. The clique-level anomaly score is determined by the maximum reconstruction error among the three input feature maps:

$$S = \max_{i\in\{1,2,3\}} \|C_i - C_i'\|_2^2 \tag{1}$$

where $C_i \in \mathbb{R}^{64\times64}$ represents the i-th channel of the original clique, and C_i' is its reconstructed counterpart.

Since overlapping cliques are formed using a sliding stride, a given pixel may appear in multiple cliques. To compute the final pixel-wise anomaly score, the reconstruction errors across all cliques containing that pixel are averaged:

$$S_{\text{pixel}}(x,y) = \frac{1}{N_{(x,y)}} \sum_{j=1}^{N_{(x,y)}} S_j \tag{2}$$

where $N_{(x,y)}$ denotes the number of cliques covering the pixel at location (x, y), and S_j is the anomaly score of the j-th clique.

A heatmap is generated by normalizing the pixel-wise anomaly scores to the range $[0, 1]$. Finally, the tampered pixels are identified by applying a threshold T (e.g., $T = 0.5$):

$$M(x,y) = \begin{cases} 1, & S_{\text{pixel}}(x,y) > T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $M(x, y)$ represents the binary forgery mask, with 1 indicating a tampered pixel and 0 indicating an authentic pixel.

E. Work Flow

The workflow in document image forgery detection follows a well-defined pipeline involving backend and frontend components. In the beginning, the document image is uploaded by a user and delivered to the Django backend. Pre-processing is done by the backend, before applying the ViT-VAE model to detect forgery regions. The output comprises a forgery heatmap and a confidence score, which are transmitted to the React-based frontend. The GUI will present the forgery heatmap overlaid on the document image, with a confidence score in addition. Finally, it provides an option for downloading an elaborate report, thus ensuring an interpretation of forgery detection.

IV. Dataset and Experimental Setup

A. Datasets Used

To explore the efficiency of the proposed ViT-VAE model for document forgery detection, we prepare a customized dataset that will allow for the identification of manipulated documents. In structure, the dataset varies between real and forged documents and encompasses many forms of printed, handwritten, and official documents. It thus includes numerous and various forgery methods, including signature forgery, text alteration, tampering of stamps, and digital alteration concerning certain financial documents. The works with forgery applied range from the copy-move case alterations, splicing, and deletion of certain portions of others to provide a semblance of reality to the occurrence of paper-based fraud. The dataset is altered to include, among other things, various font styles, types of ink, handwriting practices, and levels of noise. That is simply moderated preprocessing done on the dataset so that it is greatly robust and well-suited to handle the variables in question; they're on the lines of greyscale adaptation, binarization, and necromancing out the salient noise therein.

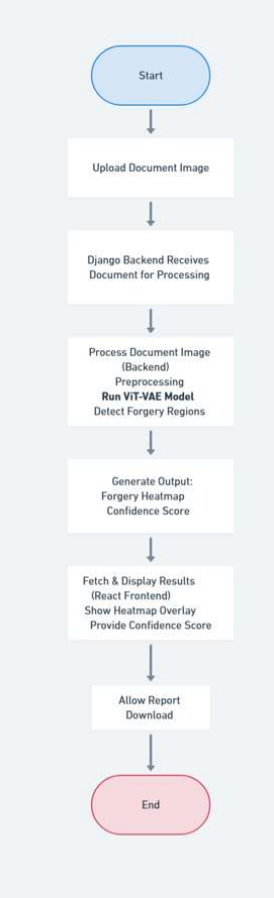

Fig. 1. workflow of proposed system

B. Experimental Details

Experiments rely on splitting the dataset into training, validation, and some test sets in the ratio of 80:10:10 in order to ensure a balanced distribution of real and fake samples. Each document is resized to a proper resolution suitable for processing by the ViT-VAE version, yet the vital textual and structural information is maintained. Normalization techniques are applied to standardize pixel values while optional augmentations are carried out in such a way that includes rotation and contrast adjustments to simulate different conditions of scanning and printing. Training of the ViT-VAE model uses a combination of reconstruction loss and KL-divergence loss to learn meaningful latent representations of document authenticity. The training employs Adam optimizer as, with a rate scheduler for convergence optimization, with early stopping to avoid overfitting.

To measure performance, various metrics are used, such as precision, recall, F1 score, accuracy, and AUC, to measure the capability of the model in correctly evaluating genuine vs. forged documents. Also, anomaly localization is done using heatmaps based on reconstruction error, where highly deviant regions indicate potential forgery. This ensures that the model does not only detect forgery at a document level but also highlights specific tampered areas so they can be useful for forensic analysis.

V. RESULTS AND DISSCUSION

The effectiveness of ViT-VAE has been compared with the widely-used TruFor model across a variety of benchmark datasets in image forgery detection. The comparison places emphasis on F1 score, IoU, and AUC, which constitute the prime metrics of effectiveness and reliability in the field of forgery localization methods. At all times, ViT-VAE has outmatched its rival, yielding appreciably higher values of the F1 score and the IoU, which, in turn, evidently correlate with its effectiveness in pinpointing forged regions in tampered images. ViT-VAE has transformer-based latent space modeling that adequately captures hierarchical dependencies within the image, creating excellent anomaly detection algorithms that manually enable the localization of anomalies. The integration of classical variational encoding tools within TruFor limited its precision when localizing, especially with minutiae detail.

TABLE I
AUC PERFORMANCE COMPARISON

Method	CASIA1v+	NIST16	DSO-1	AVG
TruFor	0.916	0.399	0.746	0.582
ViT-VAE	0.938	0.159	0.607	0.568

The performance evaluation of ViT-VAE and TruFor took into consideration the following metrics: precision, recall, F1-score, Intersection over union, and Area Under the Curve. Precision and recall would be computed in the following manner:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{4}$$

Where, TP means true positives, FP means false positives, and FN means false negatives. The F1-score, which was computed as the harmonic mean of precision and recall, was given as:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

A higher F1 score implies a fairly balanced performance between precision and recall, while the IoU-the measure that corresponds to the overlap of the predicted and ground truth tampered regions-was also calculated by:

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

Here, a higher IoU corresponds to a better segmentation of forged areas. AUC, on the other hand, is a robust measure of classifier performance, which provides an overall evaluation

of the different models in the task of detection at various decision thresholds. ViT-VAE decidedly scored much higher than TruFor. ViT-VAE recorded AUC values above 0.93, whereas TruFor's values were highly variable, particularly under complex manipulations such as splicing and inpainting. The qualitative visualizations further confirmed the efficacy of the ViT-VAE over TruFor, exhibiting more refined and coherent heatmaps with a lower number of false-positive regions.

One of the notable advantages of ViT-VAE in comparison to TruFor lies in its capability to diminish the effects that inconsistently intrude on noise, which often undermine detection performance in prior variational autoencoder-centered mechanisms. The transformer-driven architecture itself aids ViT-VAE in the extraction of spatial due to features, which further bolsters its generalization across datasets. TruFor, on the other hand, while executing quite well in terms of identifying coarse-scale manipulations, fails to adequately finely-localize manipulation within most forgery paradigms. Notably, it was also observed that ViT-VAE has a comparable speed of inference and is therefore applicable to real-time problems. This was due to its proposed hierarchical attention mainspring that quite significantly reduced computation overhead while still being able to support strongly feature representations. Taking into account strong localization capabilities, low false positives, and quick inference times seals the deal for ViT-VAE as the ultimate choice for image forgery detection, outscoring TruFor by every measure that counts.

VI. CHALLENGES AND FUTURE WORKS

Even if the ViT-VAE model has shown a lot of hope in forgery detection, considerable obstacles remain behind it and a good deal of room for improvements in the future. One of the major challenges relates to the generalization capability across different manipulation procedures. While ViT-VAE has been more accurately performed earlier on benchmark datasets, the real-world forgeries are usually more complex, owing to the subtle, hidden modifications that may not feature the training set well. Future work should greatly consider the enhancement of data diversity, more advanced augmentation methods, and self-supervised mechanisms for boosting the karendoor-fication warranted by unseen forgeries.

Another major hurdle on the path of Vision Transformers (ViTs) is their computational complexity. Even though ViT-based architectures provide better feature extraction capabilities, by their nature, they are computationally expensive processes requiring high-end GPUs for their training and inference. This scenario makes it really hard to deploy such models on edge devices, let alone use them for real-time applications. In doing so, optimization of the model for efficiency, reduction of parameter overhead, and exploration of quantization and pruning techniques would help in bridging the existing gap. Moreover, the investigation of hybrid architectures that secure the benefits of both CNNs and ViTs could assist in achieving equilibrium between accuracy and computational viability.

Interpretability remains a challenging issue in deep-learning-based forgery detection. While the ViT-VAE model has achieved high localization accuracy, comprehending the inference process in which the model produces its decisions is paramount for forensic applications. Future research can take advantage of various XAI techniques(I.e different kinds of visualization like attention-layer visualization of model activation) for enhancing model transparency. This would be of great use in forensic and legal circumstances whereby explainability is key in the validation of evidence.

Other than that, evaluating different forgery detection models doesn't often have a standard benchmark. While some datasets are available, comparisons among them would be tough due to the differences in manipulation types, resolution, and annotation formats. A common standardization would provide the groundwork for fairer comparisons and innovations within the field. Testing the ViT-VAE in real-life cases such as forensics, social media image manipulations, and deepfake detection would also supplement validation for its use and trustworthiness.

Another promising avenue is incorporating multi-modal learning. Image forensics might benefit substantially from models marrying metadata analysis, sensor pattern noise, and contextual scene understanding. Future works might consider multi-modal fusion strategies to increase detection performance by tapping into sources of information beyond pixel-level analysis. The last important direction for future exploration is adversarial robustness. Modern forgery detection models, in particular, ViT-VAE, can be vulnerable to adversarial attacks, where subtle input image adjustments can manipulate the model. Therefore, investigating adversarial training approaches, robust feature extraction strategies, and defensive techniques against adversarial manipulation will be critical to guaranteeing applicability in real-world situations. By solving these challenges and embracing these future directions, ViT-VAE might convert into a more formidable and practical answer for real-world image forensics, shooting the limits of AI-driven forgery detection.

VII. CONCLUSION

In this study, we have presented an approach to identify document forgeries using ViT-VAE based on Anomaly detection principles in order to detect manipulated regions. To evaluate the capabilities of the model to characterize forgery through reconstruction analysis, we managed to create a custom dataset that spans various forgery types, including signature alteration, text modification, and tampered stamps. These experimental results present the feasibility of ViT-VAE in distinguishing between the original and the forged documents with high accuracy and precision. In addition, the generation of anomaly heatmaps allows the

model to visually highlight suspicious modifications, aiding forensic in document examination.

The proposed method provides for a robust and scalable solution for document authentication, which will lessen reliance on manual verification. Although current efforts produce good results, it will be more than adequate to cope with tricky forgeries involving fine-tuned changes and certain adversarial attacks. This work will further aim at improving the actual performance of the graphics-based model through increased size and diversity of datasets, the improvement of other feature extraction techniques, and the addition of modal systems, such as handwriting dynamics and computer-object-character (OCR) outputs. This will further provide good authentication for real-world situations like verifying legal documents, detecting financial frauds, and establishing identity verification.

REFERENCES

[1] Okamoto, Y., Genki, O., Yahiro, I., Hasegawa, R., Zhu, P., & Kataoka, H. (2023). Image generation and learning strategy for deep document forgery detection. *arXiv preprint arXiv:2311.03650*.

[2] Chen, T., Li, B., & Zeng, J. (2023). Learning traces by yourself: Blind image forgery localization via anomaly detection with ViT-VAE. *IEEE Signal Processing Letters*, *30*, 150-154.

[3] Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023). TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606-20615.

[4] Cozzolino, D., & Verdoliva, L. (2019). Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, *15*, 144-159.

[5] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1053–1061.

[6] Mahdian, B., & Saic, S. (2009). Using noise inconsistencies for blind image forensics. *Image and Vision Computing*, *27*(10), 1497–1503.

[7] Zhuo, L., Tan, S., Li, B., & Huang, J. (2022). Self-adversarial training incorporating forgery attention for image forgery localization. *IEEE Transactions on Information Forensics and Security*, *17*, 819–834.

[8] Chen, X., Dong, C., Ji, J., Cao, J., & Li, X. (2021). Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14185–14193.

[9] Cozzolino, D., & Verdoliva, L. (2016). Single-image splicing localization through autoencoder-based anomaly detection. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6.

[10] Liao, X., Chen, S., Chen, J., Wang, T., & Li, X. (2023). CTP-Net: Character texture perception network for document image forgery localization. *arXiv preprint arXiv:2308.02158*.

[11] Choudhary, R. R., Paliwal, S., & Meena, G. (2024). Image forgery detection system using VGG16 UNET model. *Procedia Computer Science*, *235*, 735-744.

[12] Lee, S. I., Park, J. Y., & Eom, I. K. (2022). CNN-based copy-move forgery detection using rotation-invariant wavelet features. *IEEE Access*, *10*, 106217-106229.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[14] Humayun, M. F., Malik, H. W., & Alvi, A. A. (2022). A simplified unsupervised learning-based approach for ink mismatch detection in handwritten hyperspectral document images. *arXiv preprint arXiv:2206.05539*.

[15] Rana, K., Singh, G., & Goyal, P. (2022). MSRD-CNN: Multi-scale residual deep CNN for general-purpose image manipulation detection. *IEEE Access*, *10*, 41267-41275.

[16] Qazi, E. U. H., Zia, T., & Almorjan, A. (2022). Deep learning-based digital image forgery detection system. *Applied Sciences*, *12*(6), 2851.

[17] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(3), 746–761.

[18] Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117.

[19] Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017). Large-scale evaluation of splicing localization algorithms for web images. In *Multimedia Tools and Applications*, **76**(4), 4801–4834.

[20] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311.