A Comprehensive Research of Image Tampering Detection Techniques Based on Deep Learning

Bosheng Yang

School of Communications and Information Engineering, Shanghai University
Shanghai, 200444, China
yb20ybs@shu.edu.cn

Abstract. This paper reviews the advancements in image tampering detection technologies driven by deep learning. Traditional methods, dependent on manually crafted features, often fall short when confronting sophisticated tampering techniques. In contrast, deep learning models such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Recurrent Neural Networks (RNNs) significantly enhance detection capabilities through robust feature extraction and pattern recognition. These models excel in various aspects of detection: CNNs in spatial analysis, GANs in improving robustness via adversarial training, and RNNs in capturing temporal sequences in data. Despite facing challenges like the necessity for extensive annotated datasets and issues with interpretability, ongoing research is dedicated to refining these models for better generalization and efficiency. Ongoing research is enhancing model efficiency and generalization, with future work focusing on integrating multimodal data and developing more interpretable deep-learning models to ensure the integrity of visual content. Future directions aim to expand upon current capabilities by integrating multimodal data and developing models that are easier to interpret, thus ensuring the integrity and authenticity of visual content across various digital platforms.

Keywords: Deep Learning, Image Tampering Detection, Convolutional Neural Networks (CNNs), Networks (GANs), Digital Forensics

1 Introduction

As social media and the internet have developed, image tampering has become increasingly prevalent. Image tampering detection technology aims to identify and locate tampered regions within images, ensuring the authenticity and integrity of visual content. Traditional image tampering detection methods mainly rely on manually designed features and rules. These methods are often limited in effectiveness when facing the diverse and complex tampering techniques seen today. The emergence of deep learning (DL) technology, particularly Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Recurrent Neural

Y. Yue (ed.), *Proceedings of the 2024 International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2024)*, Advances in Engineering Research 240,
https://doi.org/10.2991/978-94-6463-518-8_8

Networks (RNNs) has brought new solutions to the field of image tampering detection.

DL models, due to their powerful feature extraction capabilities, can automatically learn intricate patterns and anomalies in images that are indicative of tampering. This leads to significant improvements in the accuracy and robustness of tampering detection. Various deep learning architectures, such as CNNs, GANs, and RNNs, are applied to tackle different aspects of image tampering detection, from identifying specific types of tampering to precisely localizing tampered regions within images.

As the need for reliable image authentication continues to grow, the development and refinement of deep learning-based tampering detection methods are becoming increasingly critical. This review aims to provide a comprehensive overview of the current state of deep learning techniques in image tampering detection, highlighting key methodologies, datasets, evaluation metrics, and future research directions.

The various deep learning architectures have their different and similar advantages, which can be applied to different needs.

Despite these advancements, several challenges still remain. One significant issue is the need for large annotated datasets to train deep learning models effectively. The acquisition of such datasets can be resource-intensive and time-consuming. Additionally, while deep learning models can detect many types of tampering, they may still struggle with highly sophisticated or novel tampering techniques that were not present in the training data. Moreover, the interpretability of deep learning models is another concern, as it can be difficult to understand the reasoning behind their predictions.

To address these challenges, the research focuses on improving the generalization capabilities of deep learning models, developing more efficient training methods, and enhancing model interpretability. Integrating multimodal data, such as combining visual, textual, and audio information, is also being explored to provide a more comprehensive approach to tampering detection.

2 Basic Principles

The fundamental principle of image tampering detection is to identify anomalies and inconsistencies within digital images that suggest alteration. This process involves feature extraction, anomaly detection, and classification. Traditional methods rely heavily on handcrafted features, while modern approaches leverage deep learning models to automatically learn and extract features.

2.1 CNNs

Convolutional Neural Networks (CNNs) are highly effective for feature extraction and spatial data analysis. They excel in pattern recognition, making them ideal for detecting various tampering types such as splicing and copy-move forgeries. However, CNNs require large datasets for training and are computationally intensive, which can hinder their generalization across different tampering techniques. CNNs

make remarkable progress, particularly in image processing and video-related tasks, rekindling academic interest in deep learning. Numerous studies focus on enhancing CNN performance through improvements in activation functions, optimization techniques, regularization methods, and architectural innovations. Findings demonstrate that CNNs outperform traditional methods in classification, detection, and prediction tasks. The improvement in CNN performance primarily results from the transition from conventional layer structures to block-based architectures. These blocks serve as auxiliary learners within the network, utilizing spatial or feature-map information and enhancing input channels to boost performance. Furthermore, the block-based design of CNNs facilitates modular learning, simplifying the network structure and making it easier to understand [1].

2.2 GANs

Generative Adversarial Networks (GANs) utilize adversarial training between a generator and a discriminator to create realistic forged images, enhancing model robustness. GANs can generate diverse training data, improving detection models' performance against various tampering styles. They are advantageous in data-scarce scenarios but are complex to train and can generate false positives. GANs employ adversarial training between a generator and a discriminator, creating realistic forged images and enhancing model robustness. he rising popularity of GANs is attributed to their exceptional ability to learn highly nonlinear relationships between latent space and data space. Consequently, GANs can leverage vast amounts of unlabeled data, which are typically inaccessible to supervised learning methods. They generate diverse training data, improving detection models' performance against various tampering styles. GANs can also be trained with minimal labeled data, useful in data-scarce scenarios [2].

2.3 RNNs

Recurrent Neural Networks (RNNs), including Long Short-Term Memory Networks (LSTMs), are proficient in processing temporal data and capturing spatiotemporal relationships, aiding tampering detection across consecutive frames. They model long-term dependencies crucial for tasks requiring global image consistency, making them particularly useful in video tampering detection. Despite their strengths, RNNs are prone to vanishing gradient problems and require substantial computational resources. By using internal states to preserve a recollection of prior inputs, RNNs are built to handle sequential data. Sequence prediction, natural language processing, and time-series data challenges are among their most successful applications. Key features include their ability to capture temporal dependencies, process variable-length sequences, and model time-dependent phenomena. RNNs can suffer from issues like vanishing gradients, which LSTM and GRU architectures mitigate by incorporating gating mechanisms for better long-term memory retention [3].

3 Application Scenarios

Image tampering detection technologies are broadly applied across multiple fields. In digital forensics, these technologies help verify the authenticity of digital evidence, ensuring its reliability in legal proceedings. Media authentication, plays a crucial role in verifying the credibility of images and videos disseminated through news outlets and social media platforms, thus combating misinformation. In cybersecurity, tampering detection protects against unauthorized and malicious alterations of images within secure systems, preserving data integrity. In healthcare, validating medical images is essential to prevent misdiagnosis caused by tampered imagery, thereby safeguarding patient health.

4 Challenges and Future Directions

Despite these advancements, several challenges remain. One significant issue is the need for large annotated datasets to train deep learning models effectively. The acquisition of such datasets can be resource-intensive and time-consuming. Additionally, while deep learning models can detect many types of tampering, they may still struggle with highly sophisticated or novel tampering techniques that were not present in the training data. Moreover, the interpretability of deep learning models is another concern, as it can be difficult to understand the reasoning behind their predictions.

To address these challenges, ongoing research focuses on improving the generalization capabilities of deep learning models, developing more efficient training methods, and enhancing model interpretability. Integrating multimodal data, such as combining visual, textual, and audio information, is also being explored to provide a more comprehensive approach to tampering detection. Techniques such as transfer learning, few-shot learning, and the use of synthetic data are being investigated to reduce the dependence on large annotated datasets.

5 Evaluation Metrics and Datasets

The evaluation of deep learning-based tampering detection methods typically involves various metrics to ensure comprehensive assessment. Key metrics include:

Precision: Measures the proportion of true positive detections among all positive detections, indicating accuracy in identifying tampered regions without false alarms.

Recall: Evaluates the ability to detect all actual tampered regions, reflecting the model's sensitivity.

F1-Score: Combines precision and recall into a single metric to balance between false positives and false negatives.

Standardized datasets are crucial for training and benchmarking tampering detection models. Commonly used datasets include:

CASIA: Contains various manipulated images, widely used for benchmarking tampering detection algorithms.

Columbia: Focuses on splicing detection, providing a diverse set of tampered images.

COVERAGE: Includes images with multiple tampering types, used for comprehensive evaluation of detection techniques.

6 Key Methodologies

Several key methodologies are employed in deep learning-based image tampering detection, each leveraging different network architectures to optimize detection accuracy and robustness.

Convolutional Neural Networks are foundational tools in image tampering detection due to their powerful feature extraction capabilities. The fundamental principle behind CNNs is their use of convolutional and pooling layers to automatically learn and extract important features from images. Convolutional layers apply filters to the input image to create feature maps, highlighting key patterns such as edges or textures that indicate tampering. Pooling layers then down-sample these feature maps to reduce dimensionality and computational complexity, while preserving essential features. Through forward and backward propagation, CNNs iteratively adjust their parameters to enhance detection accuracy. This enables CNNs to excel in image classification, object detection, and the localization of tampered regions within images. For example, deep convolutional networks like ResNet utilize residual learning to effectively recognize and adapt to complex tampering patterns, thereby improving overall detection performance [4].

The foundation of Generative Adversarial Networks is the idea of adversarial training between a discriminator and a generator. While the discriminator tries to discern between actual and fraudulent images, the generator seeks to create plausible forged ones. Both networks continuously develop as a result of this adversarial process; the discriminator becomes better at spotting forgeries, and the generator produces increasingly convincing ones. The discriminator uses convolutional layers to examine these images for irregularities, while the generator up-samples latent vectors into high-resolution images using methods such as transposed convolutions.

For instance, CycleGAN can generate diverse training data by transforming images from one domain to another, thereby enhancing the robustness of detection models against various tampering styles. GANs are particularly advantageous in scenarios with limited labeled data, as they can generate synthetic data to augment the training set, thereby improving the robustness and accuracy of tampering detection models [5].

RNNs, and particularly Long Short-Term Memory Networks (LSTMs), are designed to handle temporal data and capture spatiotemporal relationships, making them ideal for video tampering detection. The fundamental idea underlying RNNs is their capacity to preserve a hidden state that stores details about prior inputs, enabling them to simulate long-term dependencies. LSTMs enhance this capability by incorporating memory cells and gating mechanisms to better manage long-term

information and mitigate issues like vanishing gradients. RNNs process sequences of image data, analyzing temporal patterns to detect inconsistencies across consecutive frames. This makes them particularly effective in identifying subtle tampering traces that static models might miss. For example, in video tampering detection, RNNs track changes in pixel intensities and spatial configurations across frames to identify discrepancies indicative of tampering. By continuously adjusting weights through backpropagation through time, RNNs refine their ability to capture temporal features and improve detection accuracy [6].

Hybrid Models combine the strengths of CNNs, GANs, and RNNs to address the multifaceted nature of tampering detection. For example, a hybrid model might use CNNs for spatial feature extraction, GANs for generating diverse training data, and RNNs for analyzing temporal sequences in video data. This multimodal approach leverages each architecture's advantages, resulting in a more comprehensive and robust detection system capable of handling a wide range of tampering techniques. In practical applications like DeepFake detection, a hybrid model could use CNNs to extract detailed facial features, GANs to generate varied fake facial data for training, and RNNs to analyze changes in facial expressions over time, leading to more accurate identification of forged videos [7].

These methodologies, grounded in their respective principles, significantly enhance the accuracy and robustness of image tampering detection systems. By leveraging the strengths of CNNs, GANs, and RNNs, these systems ensure the authenticity and integrity of visual content across various applications.

7 Case Study Analysis

Due to the success of deep learning in other fields, the domain of image tampering detection has also begun to focus on integrating tampering detection techniques with deep learning technologies. Deep learning models can automatically learn complex features during training, thereby avoiding the limitations of manually selected features, such as narrow applicability and poor robustness. Unlike traditional detection methods, deep learning-based detection methods are trained directly on datasets containing various tampering types and post-processing operations. In the development of deep learning-based tampering detection algorithms, Convolutional Neural Networks (CNNs) have received particular attention. By using CNNs to automatically extract features, a generalized tampering detection algorithm can be realized.

To address the robustness issues of traditional image tampering detection methods, an approach to two-stage hierarchical feature learning was presented by Zhang et al. [8]. This algorithm has some ability for tampering region localization and can detect copy-paste and image-splicing actions. To retrieve basic features as the original input, the image is first separated into non-overlapping 32x32 blocks. Each block is then subjected to wavelet treatment. A stacked autoencoder is utilized in the initial stage to extract intricate features from every initial input. The final detection result for each block is determined in the second stage by combining data from neighboring blocks.

On the CASIA dataset, the method's accuracy in classifying blocks was 91.09%. Nevertheless, the technique labels entire picture blocks, leading to a notable disparity between the sections that were tampered with and those that were not.

Bayer et al. proposed a general image operation detection algorithm based on CNNs [9]. The method recognizes many image processing techniques, such as resampling, additive Gaussian white noise, median filtering, and Gaussian blur. The CNN model has been modified, which is innovative. Rather than learning features for detecting picture processes, CNNs typically learn features that represent image content. As a result, they applied more limitations to the filters in the first convoluted layer and cited previously conducted picture forensics research. The central value of -1 and the total of all surrounding values of 1 are the constraints for the 12 filters in the first layer. According to experimental data, this method has an average accuracy of 99.1% in detecting multiple picture manipulations.

Rao et al. proposed using CNNs to detect splicing and copy-paste tampering [10]. There are multiple steps in their algorithm. Using 128x128x3 image blocks, the CNN extracts hierarchical characteristics; the only distinction is that its initialization of parameters is not completely random. Rather, thirty high-pass filters are used to initialize the first layer to compute the residuals. Next, to extract features for the full image, a feature fusion stage and the sliding window method are used. To ascertain whether the image has been altered, the features are finally classified using an SVM classifier. On the CASIA v1.0 dataset, the algorithm yielded an accuracy of 98.04%, whereas, on the CASIA v2.0 dataset, it achieved 97.83%. It cannot locate tampered regions while having a high accuracy rate and the ability to identify both copy-paste and splicing tampering.

Bondi et al. proposed using CNNs to extract camera model-related features from image blocks to detect and localize splicing tampering [11]. An image has been stitched together if there are two distinct camera models visible in it. The image is divided into 64 x 64 non-overlapping blocks by the algorithm, and each block's camera model features are obtained using CNNs. Next, the full image is clustered using K-means to estimate the tampered region mask. To test the method's resilience to unfamiliar camera models, the test set comprises eight more camera models in addition to the ones used during CNN training. The technique showed that, even with unknown camera models, CNNs could identify splicing tampering and recover camera model attributes from image blocks.

Marra et al. proposed an end-to-end image tampering detection method [12]. Using this technique, a convolutional neural network receives the full image as input for end-to-end training. To enable CNN to learn from the full image, it is important to prevent high mistake rates at the image block level, which could result in high error rates for the entire image. Using the AUC measure, the approach was compared with existing methods on several datasets. The end-to-end strategy outperformed other methods on every dataset, according to experimental results.

8 Conclusions

As the need for reliable image authentication continues to grow, the development and refinement of deep learning-based tampering detection methods become increasingly critical. These methods offer significant improvements over traditional techniques, leveraging the power of deep learning to automatically detect and localize tampered regions in images and videos. However, challenges such as the need for large annotated datasets, the handling of novel tampering techniques, and model interpretability remain areas of active research. Future advancements will likely focus on enhancing the robustness, efficiency, and comprehensiveness of these detection systems, ensuring the integrity and authenticity of visual content in an increasingly digital world. The research will also explore the integration of multi-modal data and the development of more interpretable models, further advancing the field of image tampering detection.

References

1. Zhao, X., Wang, L., Zhang, Y., et al.: A review of convolutional neural networks in computer vision. Artif Intell Rev 57, 99–110 (2024).
2. Dash, A., Ye, J., Wang, G.: A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing. In: IEEE Access, vol. 12, pp. 18330–18357 (2024).
3. Sherstinsky, A.; Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306 (2020).
4. Al-Adwan, A., Alazzam, H., Al-Anbaki, N., et al.: Detection of deepfake media using a hybrid cnn–rnn model and particle swarm optimization (pso) algorithm. Computers 13, 99 (2024).
5. Warif, N. B. A., Idris, M., Idna, Y., et al.: A comprehensive evaluation procedure for copy-move forgery detection methods: results from a systematic review. Multimedia Tools and Applications 1-33 (2022).
6. Shi, C., Chen, L., Wang, C., et al.: Review of image forensic techniques based on deep learning. Mathematics (2023).
7. Patel, Y.: An improved dense CNN architecture for deepfake image detection. In: IEEE Access, vol. 11, pp. 22081–22095 (2023).
8. Shi, C., Chen, L., Wang, C., et al.: Review of image forensic techniques based on deep learning. Mathematics (2023).
9. Hall, D. A., Lynch, R. W., Hughes, M. T., et al.: Utilizing machine learning to predict limit cycle oscillation characteristics in aeroelastic wing. In: AIAA SCITECH 2024 Forum, pp. 2528 (2024).
10. Zhang, Y., Goh, J., Win, L. L., et al.: Image region forgery detection: a deep learning approach. SG-CRC 1-11 (2016).
11. Bayar, B., Stamm, M. C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security 5-10 (2016)
12. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE 1-6 (2016).