Graph-Augmented Hybrid Framework for Deep-Learning-Based and Feature-Based Detection of Image Tampering and Copy-Move Forgeries

Venkataraman
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
venkataraman.r2023@vitstudent.ac.in

Abraham Justin
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
abraham.justin2023@vitstudent.ac.in

Sudarshan Manikandan
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
sudarshan.manikandan2023@vitstudent.ac.in

Dr Geetha S
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
geetha.s@vit.ac.in

***Abstract*—Data manipulation and forgery are becoming a threat to honest data, blurring the line between fake and reality. Forgery is increasingly becoming common, from research papers to fake videos, increasingly spreading propaganda and fake information, causing confusion and mistrust in our society. Out of this, image forgery is becoming the most concerning. With text forgery already an age-old thing, images were considered to be the shield to truth. However, with an increasing number of images being forged, we soon realise that the truth is now becoming harder to identify, and untruth is being passed off as fact. There are multiple techniques to forge images, one such being copy-move forgery. What sets copy move forgery apart is the fact that truth is covered up by a part of itself, that is, a part of an image is covered up by another part of the very same image itself. This makes it harder to detect, as traditionally, to detect image forgery, foreign objects in an image were a tell-tale sign of forgery. However, with the elements of the image itself being a part of it, copy-move forgery becomes harder to detect. In this paper, we look at a hybrid framework that can reliably identify generic tampering using an EfficientNet-based binary classifier and a GRAD cam localisation system, and copy move forgery identification using the time-tested SIFT algorithm, combined with key point distribution analysis, feature matching and hierarchical clustering.**

Keywords—template, Scribbr, IEEE, format

I. INTRODUCTION

Image forgery detection is now increasingly becoming an important frontier in research, with use cases piling increasingly, with the introduction of image-generating AI systems, which, even though having ethical use cases, are now being used in hiding or manipulating facts and information, to spread hate, propaganda and mislead the masses. With tools to forge images becoming advanced, it has become impossible to identify manipulations and forgeries in images with the naked human eye. The requirement for frameworks to be usable and effective in identifying forgeries has increased, and their importance is paramount. The frameworks, models and algorithms used for this purpose must not only be effective, but also scalable, robust and lightweight. With the huge influx of forged images, the approach being lightweight attains the most importance, even trumping accuracy. Newer and modern approaches, diverging from generic algorithmic and machine learning approaches, to cutting-edge deep learning algorithms, requiring the most advanced GPU systems, have been proposed and are being implemented. However, the requirement for a model that is both generic enough to cater to a wider forgery umbrella and lightweight enough to be a widespread implementation remains elusive and is still a topic of research and discourse. Image forgery is typically considered to be of two types: image splicing and copy-move forgery [1]. Image splicing refers to images being tampered with using the help of other images. If the tampering is done by splicing parts of the same image, it is known as copy-move forgery. In this paper, we discuss a model that can identify both these forgeries effectively and in a lightweight fashion. The algorithms required to detect these must keep in mind the rotation, preprocessing and scaling of the spliced regions for detection. To detect these forgeries, image experts give us two approaches to deal with this matter. The active approach and the passive approach are two ways to detect images [2][8]. An active approach is a pre-tampering prevention or identification approach. It deals with methods such as steganography, watermarking, which, in the case of tampering, help us understand if tampering has occurred and, in some special cases, how and when it has occurred. On the other hand, a passive approach is used to identify if an already tampered image is indeed tampered, which is done by analysing for image inconsistencies, and any traces left behind on the image after the forgery has been made. This paper discusses a passive image forgery detection framework that identifies both splicing and copy-move forgery, which helps make it generic enough to handle all types of threats. Another motivation to go into this hybrid approach is the fact that CNN models do not pinpoint manipulation, which algorithms like SIFT fail to identify tampering like splicing, as they are only designed to identify a specific type of forgery, that being copy-move forgery. Current approaches rarely combine both, which raises a use case for this approach. To summarise, this approach comprises a binary classifier that uses a CNN model to identify or classify images as tampered or original, and also uses a GRAD-CAM for explainable localisation. This is combined with an enhanced SIFT algorithm that is combined with feature matching and hierarchical clustering to identify copy-move forgery, effectively covering all bases. It is to noting that this framework works on images regardless of the geometric orientation of the splicing that takes place, which makes this approach powerful [9].

II. RELATED WORK

The approach discussed in the paper is inspired by the two approaches that are discussed in the subsequent sections. One uses a SIFT + entropy-based approach to identify copy-move forgeries, while the other discusses a CNN-based approach, which uses DeepPatchMatch and Pairwise ranking learning. These approaches led us to create a hybrid framework that utilises both the advantages of using a CNN-based model and an algorithmic model.

A. Entropy-based SIFT approach

In the approach discussed by Jiang and Lu [4], they try to offset the disadvantages of the classical SIFT algorithm by using entropy to increase the number of key points generated. Traditionally, SIFT comprises four steps: the determination of candidate points in the difference of Gaussian (DoG) space, selecting key points using contrast threshold, calculating the dominant orientation, then finally generating the feature descriptor. A key point is a distinctive and repeatable point in an image that is consistently recognisable regardless of any transformation, scaling or rotation being applied to the image. It is this property of a key point that helps SIFT approaches identify tampering. In traditional SIFT approaches, however, the identification of key points of a region in an image is hard or impossible if the said image is smooth or flat. This makes identifying forgeries hard, as without a good set of key points, it becomes hard to reliably identify if a region is tampered with or authentic. In the approach discussed by Jiang and Lu, they apply entropy to a grey-scaled version of the original image, in order to increase the number of key point generations in flat regions. Lu and Wang then apply hierarchical clustering to group key points and look for matches to identify a set of copy-move forged regions. They apply a binary mask to highlight the copy-move forged regions. This approach is purely algorithmic and doesn't require the overhead of training a model. However, the algorithm does need to be fine-tuned in order to make it effective enough to work in the real world. It is also unable to identify tampering due to splicing from different images, which makes this model unable to be scaled up for a real-world use case. Lu and Wang use the GRIP and CMH datasets for testing their approach.

B. CNN-based approach

Deep patch match and pairwise learning approaches were discussed by Li et al [3]. They designed this approach for the same reason as discussed in above, for copy-move forgery identification. Patch match is an efficient algorithm that is used to find the nearest correspondences across different images. Patch match is comprised of three steps. First, each pixel of the input image is assigned a random offset, which indicates the position of a corresponding pixel in the target image. The idea here is that a random large number of guesses will result in a certain number of optimal or sub-optimal offsets. In the second step, good offsets are mapped or propagated to their neighbours since they have good matches. This significantly prunes or reduces the search space. To reduce sub-optimality, the third step of patch match uses a random search. Steps two and three are often executed multiple times to increase optimality or till convergence is achieved. Multiple extensions of patch match have been proposed; however, the code idea is the above-discussed three steps. Li et al have tailored this patch match approach to fit their use case, which is copy-move forgery detection, devising a new module in patch match for this very purpose. They have two main branches: dense field matching via deep cross-scale patch match and source/target discrimination via pairwise ranking learning. These two branches make up the core of their model. The dense field mapping is used to match similar regions to identify copy-move forgeries. Pairwise ranking learning utilises prior knowledge to differentiate between the source or original to the target or forged regions. Hence, in this approach, Li et al cannot only identify copy move forgeries in their approach, but they are also able to reliably able to identify which part is the forged region and which is the tampered region, which is not done in the approach given by Jiang and Lu in their entropy-based approach. However, even here, as seen in the entropy approach, their model only works to identify a specific type of tampering, that being copy-move forgery. If a given image is tampered with via splicing from foreign images, this approach fails; hence, a hole is seen in their approach, and their model only remains as a solution to a niche region of image tampering [10].

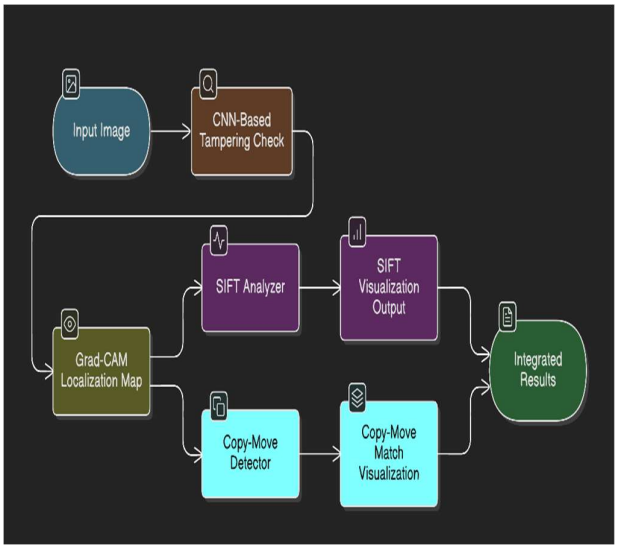
III. METHODOLOGY

This section describes the proposed hybrid framework that integrates deep-learning-based tampering classification, Grad-CAM–based localisation, SIFT-based statistical analysis, and classical copy-move forgery detection. We designed this system as a three-pronged approach to identifying tampering in images, one general purpose and two to identify copy-move forgeries, with one of them only being there as a visualisation tool to see the key points in the image.

A. System Overview

The proposed approach uses a multi-branch hybrid architecture to complement the disadvantages of classical algorithmic approaches and deep learning model-based approaches. For a given input image, the framework processes it via a three-pipeline parallel pipeline. The CNN-based binary classifier is used to determine if there is any tampering, particularly splicing from foreign images in the image, and thus the final classification by this model is authentic if no tampering is found, or manipulated if tampering is found. The second component is a visualisation tool, which is a SIFT-based forensic analyser that shows the key point distribution, anomalies in the image, and the texture. These are all required to find tampering, but here, they are used as a visualisation tool to verify if the key points are being generated uniformly throughout the image, helping us identify if the approach is working or not. The outputs from the branches or phases are combined into a singular report. This report consists of a Grad-CAM heatmap, which highlights the tampered region, and the copy move detection map, which maps the copy move forged regions, which helps us identify them. There is also a confidence rating given, which helps us understand how much the model believes a given image is forged or not. This hybridisation helps us to detect a wide range of tampering, from global tampering to localised self-copying. The reason for this hybridisation is because of the advantages of the individual components. CNNs are very helpful to identify tampering on a global scale. However, they require intense training and computational resources for identifying smaller edits or pinpointing tampering to a localised region. SIFT-based approaches are not powerful enough to identify generic anomalies; however, they are good enough to identify localised inconsistencies, which are required to identify copy-move forgeries. Combining the three approaches, we get a robust and effective model.

FIGURE 1. FLOW DIAGRAM OF THE FRAMEWORK



The above diagram shows the general flow of the process, from the user input to the integration of the results. This process flow is efficient and lightweight enough to be applied in a real-world scenario.

B. CNN-Based Tampering Classifier

1) EfficientNet-B0 model

The CNN-based tampering classifier uses the EfficientNet-B0 model, which is CNN-based. It is part of the zoo or group of models developed by Google for image analysis. It is lightweight enough to be used in our framework, and at the same time, being a convolutional neural network, it is accurate. It is fine-tuned to fit our use case, to classify tampered and authentic images, after being trained on a dataset. Another feature of this model is that it uses a lot fewer parameters compared to other traditional CNN architectures. Thus, the EfficientNet-B0 model is chosen as it is lightweight, accurate, and easily deployable. The input images to this layer are resized to 224X224 and are normalised. Then this model/network gives 2 class logits, which correspond to the authentic and tampered categories. These logits are then converted to a probability score (or confidence score) to evaluate how much the model believes the images have been tampered with. Since the dataset comprises mainly authentic images, a tampered image is paired with two authentic images while training to balance the data.

2) Grad-CAM-based localisation

To make the output interpretable or understandable after CNN is applied, the Gradient-weighted Class Activation Mapping (or Grad-CAM) is integrated into the CNN classifier. The Grad-CAM is used to visually identify the tampered regions, which are identified by the CNN model. Two activations take place here: Forward activation and Backward activation. These two activations are combined to generate channel-weighted localisation maps, which show which parts contribute to the classification of the input image to the tampered class, which is basically the tampered region. In our model, we refine the Grad-CAM output to use a shaded pink overlay circle to mark the high activation regions, which are the tampered regions in the image. This overlay helps analysts identify suspicious regions in the image quickly.

3) SIFT-Based Keypoint Distribution Analysis

Before the key points are extracted from the image (which is first converted to a grey-scale image), we do Contrast Limited Adaptive Histogram Equalisation (CLAHE) and multi-scale texture enhancement using variance filters [6]. These two steps are done to increase or amplify subtle texture irregularities and anomalies, which makes tampered regions more detectable, as they appear to have inconsistent illumination. Then the Scale Invariant Feature Transform is applied to the enhanced image to extract key points. We use the following parameters for this: nfreatures=5000, which is set to a high value to make it sensitive to small textures or cloned patches, which helps identify copy-move forgeries more easily. We also set the contrastThreshold to 0.02 and sigma to 1.2 for the same purpose. The key points are then analysed with different statistical measures, such as: Quadrant distribution analysis, which is the analysis of key points in the top-left, top-right, bottom-left, bottom-right and centre regions of the image region. We also find a uniformity metric for the region, where are lower value indicates the possibility of tampering or modifications. DBSCAN clustering [5] is applied to group key points, which is then used to identify unnatural patches or regions, irregular patterns and regions having similar anomalies. All the above are some symptoms of two or more regions being copies of each other, hence the possibility of copy-move forgery in the image.

4) Copy move forgery detection

The image is compared with itself to detect the forgeries. SIFT descriptors extracted in the previous steps are matched using a brute force KNN algorithm, applying a ratio test and a distance threshold to avoid trivial matches. The matched points are then passed to a hierarchical clustering module, which is then passed to an inconsistency threshold filter to remove associations that are unreliable. Outliers are also removed based on cluster frequency analysis. The final output is an image with lines connecting the paired matches. This complements the CNN model as it identifies the copy-move forgeries, which are mostly missed by the CNN model.

IV. EXPERIMENTAL SETUP

In this section, we discuss the dataset, the training protocol, implementation details, and the evaluation metrics that were used to validate the proposed hybrid forgery detection framework.

A. Dataset

The CASIA V2 dataset was used to train and test the model. This dataset is one of the most widely used for both training and evaluating tampered images for both splicing and copy-move forged tampering. The dataset contains both tampered and authentic images. There are approximately 7,491 authentic images and 5,123 tampered images. The tampered images are accompanied by a pixel-level ground truth mask, which is a binary image that indicates the manipulated regions. Since CASIA v2 uses an inconsistent filename convention, we had to implement an automated GT annotation matching strategy, which is given below:

1. Extract the base filename of each tampered image (e.g., TAMPERED_001).

2. Search for GT files with patterns matching:
    - {filename}_gt.png
    - {filename}_gt.jpg / .jpeg / .bmp / .tif
3. If no GT file is found, label the sample as "TAMPERED_NO_GT", allowing the training pipeline to proceed without interruption.
4. Cache all mappings using joblib to speed up repeated training sessions.

This matching strategy ensures that the pairing between the input images and their ground truth masks is robust for both training, evaluation and testing.

B. Train-Test split

The dataset was split into two, 70% training and 30% for testing, for both authentic and tampered images. Since the distribution of the data is skewed, we needed to adopt a hybrid sampling strategy. That is, for each tampered image included in the sample, two authentic images are sampled which resulting in a hybrid ratio of 2:1 (authentic: tampered). This stabilises training by reducing the class imbalance, improving tampered recall, and preventing the classifier from overfitting to authentic textures. Thus, the final dataset used in the training consisted of 70% of the hybrid dataset, with the remaining 30% used for testing.

C. Training Parameters

The CNN tampering classifier was trained using the following hyperparameters.

- Model Architecture: EfficientNet-B0
- Epochs: 12
- Batch Size: 16
- Optimiser: Adam
- Learning Rate (LR): 1×10^{-4}
- Weight Decay: 1×10^{-5}
- Loss Function: Cross-Entropy
- Input Resolution: 224×224

To prevent overfitting and reduce the training cost, all the layers in the EfficientNet model features, from 0 to -3, were frozen, and only the final convolutional blocks and the classification head were updated. This helps keep the model lightweight and suitable for deployment. The entire system was implemented in PyTorch.

D. Evaluation metrics

The performance of the model were accessed with the help of classical standard classification and forensic evaluation metrics such as: accuracy, precision, recall, and F1 score for each class (authentic and tampered), macro averaged metrics for dataset wide evaluation, tampering specific metrics, tampered recall (the true positive rate) which measures the ability to correctly identify manipulated images, tampered precision which measures the reliability of tampered predictions, false positive rate and false negative rate. A confusion matrix was also plotted for each epoch.

False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate (FNR):

$$FNR = \frac{FN}{FN + TP}$$

These metrics are very important to identify the effectiveness and accuracy of our model, as in the real world, misclassifying a tampered image as authentic carries very severe punishments.

Epoch	Overall Accuracy (%)	Tampered Recall (%)	F1-Score
1	78.79%	65.67%	0.667
2	80.25%	67.83%	0.689
3	81.00%	73.83%	0.715
4	81.32%	72.17%	0.714
5	81.59%	73.50%	0.721
6	81.22%	75.00%	0.721
7	80.09%	77.17%	0.715
8	81.70%	78.00%	0.734
9	79.55%	80.00%	0.716
10	80.25%	80.67%	0.725
11	80.09%	72.50%	0.702
12	81.00%	75.50%	0.720

TABLE 1: MODEL METRICS PER EPOCH

V. RESULTS AND DISCUSSION

The model gives us the proper results, and the values attained as metrics are discussed in the previous table.

FIGURE 2. CNN CLASSIFIER

In the above image, we can see that the CNN classifier is able to identify the tampered region of the input image effectively. It utilises the Grad-CAM to create a pink circular overlay over the tampered region. There is also a confidence score given to the image. Here we see that the model is 68% sure that the image is tampered. This seems a little low, however, when we consider the fact that the image is tampered with using copy-move forgery, it becomes impressive as a deep learning model, such as this CNN classifier, almost always struggles to identify such tampering. Hence, even though this model was designed to identify only the splicing-based

tampering, this model was also able to identify the copy-move forged region. However, it still fails to detect the region from which the duplicate was copied. This is where our SIFT component comes into play.

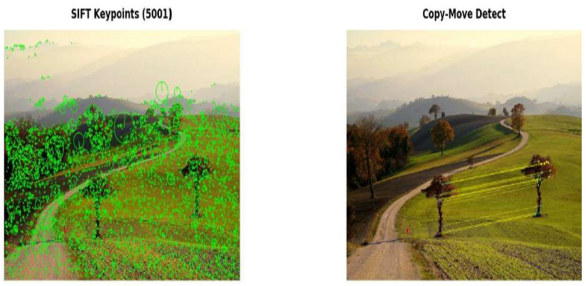


FIGURE 3. SIFT COMPONENT OUTPUT

The above image shows that the SIFT component is able to identify and map the copied region and the original region. The yellow lines indicate the match between the two regions. This shows that the model is robust and can reliably identify copy-move forgery, too. On the left, the image filled with green dots shows how the key points are distributed in the image. From that image, we can see that, even though there are some plain regions in the image, the modified SIFT can identify the key points reliably.

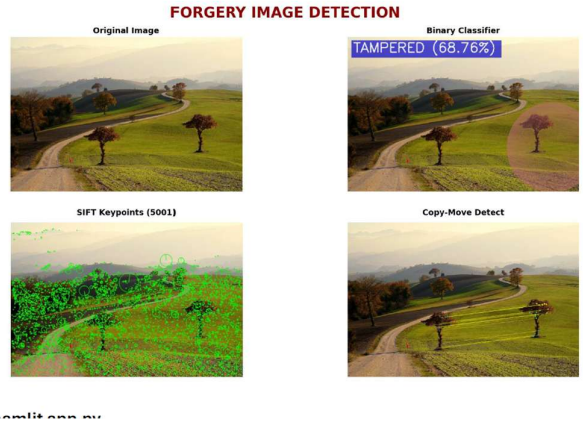


FIGURE 4. OVERALL OUTPUT

The above picture shows how the overall output is given by the model. The input image is in the top left, the CNN output on the top right, with the tampered region highlighted and the confidence metric shown. The key points generation and copy move forged detection are shown in the bottom left and right, respectively. Hence, the outputs are discussed.

VI. FUTURE WORK

Although the proposed method shows encouraging performance, several directions remain open for future exploration:

1. Improved Localisation using Transformer Models
Recent Vision Transformers (ViTs) and hybrid CNN–CNN-Transformer models can provide sharper manipulation boundaries. Integrating ViT-based attention maps may enhance fine-grained localisation accuracy.

2. Integration of Additional Forensic Cues
The current pipeline focuses on structural and duplication cues. Incorporating Noise residual analysis (PRNU), JPEG compression artefact inconsistencies, or illumination/lighting estimation could further strengthen detection against sophisticated manipulations.

3. Enhanced Copy-Move Detection
The classical SIFT matching pipeline may fail under extreme geometric transformations. Future work can leverage: Deep PatchMatch networks, pairwise ranking learning, or Feature correlation layers, as demonstrated in recent research.

4. End-to-End Fusion Network
Currently, the three branches operate independently. A unified end-to-end fusion model could integrate all three forensic cues into a single learnable architecture, potentially improving robustness and inference speed.

5. Real-World Deployment and Adversarial Robustness
Testing the system on: social-media compressed images, collage-style edits, adversarial manipulated images, would help assess real-world resilience. Incorporating adversarial training or robust optimisation could be beneficial.

6. Larger and More Diverse Datasets
The system has been evaluated on CASIA 2.0; additional datasets such as CoMoFoD, Columbia, and PS-Battles would enable stronger cross-dataset generalisation and broader validation.

VII. CONCLUSION

In this work, we presented a hybrid forensic framework that integrates deep-learning-based tampering classification with classical copy-move detection and SIFT-based key point analysis. Unlike conventional single-stream approaches, the proposed system leverages the complementary strengths of three independent branches: EfficientNet-B0 with Grad-CAM for global tampering detection and localisation, an enhanced SIFT analyser for texture and structural irregularity identification, and a classical SIFT matching module for detecting copy-move forgeries. Extensive experiments on the CASIA 2.0 dataset demonstrate that the hybrid approach significantly improves tampered recall**,** localisation capability**,** and robustness across manipulation types**.** The Grad-CAM visualisation offers intuitive interpretability, while SIFT-based analysis provides additional forensic cues that are often missed by deep models alone. The copy-move detection branch effectively identifies duplicated regions, further enhancing the system's utility in forensic investigations. Overall, the results confirm that combining deep learning with feature-based forensics leads to a more reliable and interpretable tampering detection pipeline, suitable for use in digital journalism, legal evidence verification, and security-sensitive applications.

REFERENCES

[1] T. K. Huynh, K. V. Huynh, Thuong Le-Tien and S. C. Nguyen, "A survey on Image Forgery Detection techniques," The 2015 IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF), Can Tho, Vietnam, 2015, pp. 71-76, doi:10.1109/RIVF.2015.7049877.

[2] N. Kanagavalli and L. Latha, "A survey of copy-move image forgery detection techniques," *2017 International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, 2017, pp. 1-6, doi: 10.1109/ICISC.2017.8068703.

[3] Y. Li, Y. He, C. Chen, L. Dong, B. Li, J. Zhou, and X. Li, "Image Copy-Move Forgery Detection via Deep PatchMatch and Pairwise Ranking Learning," *IEEE Transactions on Image Processing*, vol. 34, 2025, pp. 424-440.

[4] L. Jiang and Z. Lu, "An Effective Image Copy-Move Forgery Detection Using Entropy Information, " IEEE Transactions on Image Processing, vol. 34, pp. 1-6, 2025

[5] Yuanman Li and Jiantao Zhou, "Fast and effective image copymove forgery detection via hierarchical feature point matching," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp.1307–1322, 2018.

[6] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra, "A sift-based forensic method for copy–move attack detection and transformation recovery," IEEE transactions on information forensics and security, vol. 6, no. 3, pp. 1099–1110, 2011.

[7] Y. Tan, Y. Li, L. Zeng, J. Ye, W. Wang, and X. Li, "Multi-scale targetaware framework for constrained splicing detection and localization in Proc. 31st ACM Int. Conf. Multimedia, vol. 4, Oct. 2023, pp. 8790–8798.

[8] Jessica Fridrich, David Soukal, and Jan Lukáš, "Detection of Copy-Move Forgery in Digital Images", Digital Forensic Research Workshop, Cleveland, Ohio, USA, 2003.

[9] Alin C Popescu and Hany Farid, "Exposing Digital Forgeries by Detecting Duplicated Image Regions", Dartmouth Computer Science Technical Report TR2004-515, USA, August 2004.

[10] Zhen Fang, Shuozhong Wang, Xinpeng Zhang, "Image SplicingDetection Using Camera Characteristic Inconsistency", MINES '09.International Conference on Multimedia Information Networking and Security, Hubei, 2009.