Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.051705

ARTICLE

Research on Improved MobileViT Image Tamper Localization Model

Jingtao Sun[1,2], Fengling Zhang[1,2,*], Huanqi Liu[1,2] and Wenyan Hou[1,2]

[1]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China

[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China

*Corresponding Author: Fengling Zhang. Email: feng3838@stu.xupt.edu.cn

Received: 12 March 2024 Accepted: 18 July 2024 Published: 15 August 2024

ABSTRACT

As image manipulation technology advances rapidly, the malicious use of image tampering has alarmingly escalated, posing a significant threat to social stability. In the realm of image tampering localization, accurately localizing limited samples, multiple types, and various sizes of regions remains a multitude of challenges. These issues impede the model's universality and generalization capability and detrimentally affect its performance. To tackle these issues, we propose FL-MobileViT-an improved MobileViT model devised for image tampering localization. Our proposed model utilizes a dual-stream architecture that independently processes the RGB and noise domain, and captures richer traces of tampering through dual-stream integration. Meanwhile, the model incorporating the Focused Linear Attention mechanism within the lightweight network (MobileViT). This substitution significantly diminishes computational complexity and resolves homogeneity problems associated with traditional Transformer attention mechanisms, enhancing feature extraction diversity and improving the model's localization performance. To comprehensively fuse the generated results from both feature extractors, we introduce the ASPP architecture for multi-scale feature fusion. This facilitates a more precise localization of tampered regions of various sizes. Furthermore, to bolster the model's generalization ability, we adopt a contrastive learning method and devise a joint optimization training strategy that leverages fused features and captures the disparities in feature distribution in tampered images. This strategy enables the learning of contrastive loss at various stages of the feature extractor and employs it as an additional constraint condition in conjunction with cross-entropy loss. As a result, overfitting issues are effectively alleviated, and the differentiation between tampered and untampered regions is enhanced. Experimental evaluations on five benchmark datasets (IMD-20, CASIA, NIST-16, Columbia and Coverage) validate the effectiveness of our proposed model. The meticulously calibrated FL-MobileViT model consistently outperforms numerous existing general models regarding localization accuracy across diverse datasets, demonstrating superior adaptability.

KEYWORDS

Image tampering localization; focused linear attention mechanism; MobileViT; contrastive loss

1 Introduction

Images have become an omnipresent medium for information dissemination in today's society, largely due to their innate simplicity and ease of understanding. However, the recent advancements in artificial intelligence and deep learning technologies have given rise to a plethora of advanced techniques for image manipulation [1–4]. Unscrupulous individuals can misuse these techniques to fabricate tampered images, thereby posing a significant threat to personal, societal, and national security. Consequently, research dedicated on localizing image tampering carries profound practical implications.

Currently, deep learning methods have found extensive application in the field of image tampering localization. The current body of research primarily focuses on two main directions: localization methods tailored for specific types of tampering and those devised for multiple types of tampering. However, given that real-world scenarios frequently involve a blend of tampering techniques, methods that exclusively target a single type of tampering often find limited applicability. Research methodologies addressing multiple tampering types can be bifurcated into two categories: those based on convolutional neural networks and those rooted in visual transformers. Both methodologies are primarily applied in RGB and noise domains to extract subtle traces of tampering in RGB images. Methods grounded in convolutional neural networks exhibit superior generalization capabilities. This is attributable to their local correlation and transitional invariance. However, they are constrained by their ability to extract only limited information, which makes it challenging to obtain global contextual information. To tackle this issue, researchers have pivoted towards ViT-based methods, which are excellent at capturing global contextual information, but require a significant amount of data to achieve the desired localization results. Furthermore, the generalization performance of these two types of localization methods remains somewhat limited when it comes to localizing multiple types of tampering.

To improve the generalization performance of localization methodologies, researchers have begun to explore the underlying factors that contribute to performance limitations. It was discovered that the model's localization performance improved when specific tamper traces were prominently evident. Subsequent research revealed that this phenomenon stemmed from a tendency for the model training process to excessively focus on these specific tamper traces, resulting in overfitting and thereby limiting its generalization performance. Consequently, extensive research has been conducted to address the issue of overfitting in image tampering localization. Studies have indicated that the effectiveness of the cross-entropy loss model for image tampering localization is suboptimal. This can be attributed to different tampering techniques leaving distinct traces, making it prone to overfitting when extracting similar features from the tampered region. Thus, the model's generalization performance is limited. To address this issue, the researchers utilize contrastive learning techniques to introduce additional constraints during model training, thereby mitigating the risk of overfitting caused by focusing on specific tampering traces and ultimately improving the model's performance.

We propose FL-MobileViT, a novel model designed specifically for localizing image tampering. It utilizes a dual-stream architecture and constructs feature extractors from the RGB and noise domains. RGB feature extractor is mainly used to extract obvious tampering traces in RGB images. Within the noise domain, we employ the SRM filter to convert the image to high-frequency image to amplify the tampering traces that might be imperceptible in the RGB domain, and use SRM feature extractor to extract and capture these traces. The feature extractor, based on MobileViT, incorporates the Focused Linear Attention mechanism, thereby enhancing the Transformer attention mechanism inherent in MobileViT. Specifically, we substitute the original softmax function with

a new mapping function, aiming to reduce computational complexity. Furthermore, we introduce depthwise convolution (DWC) to tackle the row homogeneity issues that arise post the replacement of the mapping function, thereby enriching extracted features. The lightweight feature extractor effectively exploits image information to generate diverse and multi-scale feature outputs, thereby making it exceptionally suitable for localizing image tampering in scenarios with limited samples. To bolster the generalization performance, we introduce a contrastive learning module and devise a joint optimization training strategy. During the training phase, to better distinguish between tampered and untampered regions, the model calculates the contrastive loss at different stages of the feature extractor. This strategy avoids overfitting may arise from focusing on specific tampered traces, thereby bolstering generalization performance.

The remainder of this paper is structured as follows: Section 2 discusses the research methodology employed in this study. In Section 3, we provide a detailed exposition of the key methods incorporated in our model, including its overarching framework, feature extractor, and the contrastive learning approach. Section 4 presents a comprehensive overview of the experimental setup and results analysis. Finally, we conclude the paper in Section 5.

2 Related Work

With the advancement of deep learning technology, research on image tamper localization has achieved remarkable progress in recent years. The related research primarily concerns two apects. The first is the enhancement of the generality of the localization and detection methods. The second is the effective utilization of the tampering traces in the tampered images. Early studies mainly concentrated on devising methods for specific types of tampering [5–7]. However, such methods have limited applicability, as tampered images in real-world scenarios may incorporate various image processing techniques. Therefore, researchers have turned their attention to more general methods for image tamper localization or detection [8–10]. For example, literature [8] proposed a spatial pyramid attention network, which is based on the VGG network architecture. This network introduces a local self-attention mechanism and incorporates spatial position coding, which enables it to establish connections between image blocks at different scales and significantly improve the accuracy and generality of localization. In terms of the effective utilization of image tampering traces, researchers have conducted studies on both the image itself and feature extraction. The above methods have limitations in extracting unclear tamper traces in RGB images, leading to insufficient extraction of image tamper traces. Therefore, to extract and utilize subtle tamper traces in images more effectively, researchers concentrated on the image itself and proposed the RGB-N model [11]. The model is a dual-stream Faster R-CNN network that extracts features from the RGB and the noise domains. The high-frequency information in the noise domain helps to emphasize edge features. This dual-domain network design has inspired researchers to adopt dual-domain or multi-domain methods [12,13]. Furthermore, with regard to feature extraction, these convolutional neural network-based methods mainly focus on extracting local information, but they are insufficient in acquiring global context information, which leads to limitations in locating tampered regions of various sizes. To effectively locate image tampering regions, researchers have explored the global context modeling capability of Vision Transformer (ViT) [14,15]. For instance, The literature [15] proposed the TBFormer network, which consists of two feature extractors, each using differentiated superimposed Transformer layers to extract features from the RGB and the noise domains, respectively, aiming to mine more clues. However, despite its excellence in overcoming CNN limitations, ViT is inferior to CNN in terms of local information modeling and computational efficiency.

To achieve a more balanced assessment of the respective advantages and disadvantages of ViT and CNN, as previously discussed, researchers have attempted to combine the two and propose a series of lightweight networks [16–18]. The literature [16] introduced the MobileViT network, which combines the advantages of both CNN and Transformer, overcoming their shortcomings, while maintaining lightweight and low-latency characteristics, and outperforming individual CNN or VIT networks. Subsequently, Mehta et al. optimized the MobileViT network by proposing MobileViTv2 [17] and MobileViTv3 [18]. Nevertheless, these hybrid networks still have high computational complexity. Researchers attempted to reduce the computational complexity of attention by replacing ViT's self-attention module with linear attention methods [19,20], but these methods led to a significant drop in model performance. Therefore, to address this challenge, the literature [21] proposed a Focusing Linear Attention module, which focuses on two aspects: improving the focusing ability and feature diversity, while reducing the computational complexity and maintaining the model performance. However, using traditional cross-entropy loss for training can easily cause overfitting on specific tamper types, thereby affecting the model's generalization ability.

To tackle this issue, researchers have proposed contrastive learning [22] to compute the contrastive loss and use it as an additional constraint with the cross-entropy loss to prevent overfitting. Inspired by contrastive learning, the literature [23] proposed a novel model that uses multi-scale and pixel-level supervised contrastive learning, which improves the model's ability for multi-scale perception and feature expression, thereby boosting localization accuracy and generalization. Moreover, the literature [24] introduced a new localization method named CFL-Net, which addresses the problem of lack of constraints on cross-entropy loss in tamper localization. This method effectively integrates supervised contrastive loss with cross-entropy loss to better distinguish tamper regions and improve the model's overall generalization performance.

We propose FL-MobileViT, an image tamper localization model that aims to address the issues of insufficient utilization of image information, overemphasis on local information, overfitting during model training, and failure to localize tampered areas of various sizes. Firstly, our model employs a lightweight feature extractor based on MobileViT that leverages the strengths of CNNs and Transformers. The dual-stream architecture separately extracts features in both RGB and noise domains to make full use of image information while adopting focused linear attention for enhanced feature richness and computational efficiency. This method improves tamper localization performance without relying on large amounts of data. Secondly, we incorporate the ASPP module [25] for multi-scale fusion of dual-stream output features. Finally, supervised contrastive learning is used during training to calculate contrastive loss at different stages of the feature extractor and design a joint optimization strategy with cross-entropy loss as the final objective function for boosting the generality, generalization ability, and model's localization accuracy in image tamper localization tasks under a limited sample condition.

3 Method

For the task of image tamper localization, we propose an improved model named FL-MobileViT based on the MobileViT. This chapter provides a comprehensive exposition of the model's three critical aspects: the overall architecture and devise concept, the construction of the feature extractor, and the utilization of contrastive learning with the joint optimization training strategy.

3.1 Methodology Philosophy and Overall Architecture

The field of image tamper localization presents a complex research challenge, encompassing several pivotal aspects: (1) how to extract subtle tamper traces from RGB images; (2) how to enable the model to locate tampered regions of various sizes precisely; (3) how to enhance the model's localization performance under limited data conditions; and (4) how to avoid the model is overfitting to specific tamper traces, thereby resulting in poor generalization. In response to these challenges, we devise a feature extractor based on MobileViT and utilize a dual-stream architecture to extract features from the RGB and noise domains. Moreover, we incorporate contrastive learning to boost the generalization capability of our model.

Fig. 1 illustrates the overall architecture of our proposed model. Firstly, we employ a dual-stream architecture that consists of an RGB feature extractor and an SRM feature extractor, which are responsible for extracting features from the RGB and the noise domain, respectively. These two extractors have the same architecture but do not share weights, which helps the model capture different kinds of tamper features. The noise stream takes the SRM-filtered [11] images as input, and the SRM-filtered images convert into high-frequency images, which enhance the edge information and suppress the semantic information, thus better revealing the inconspicuous tamper traces.

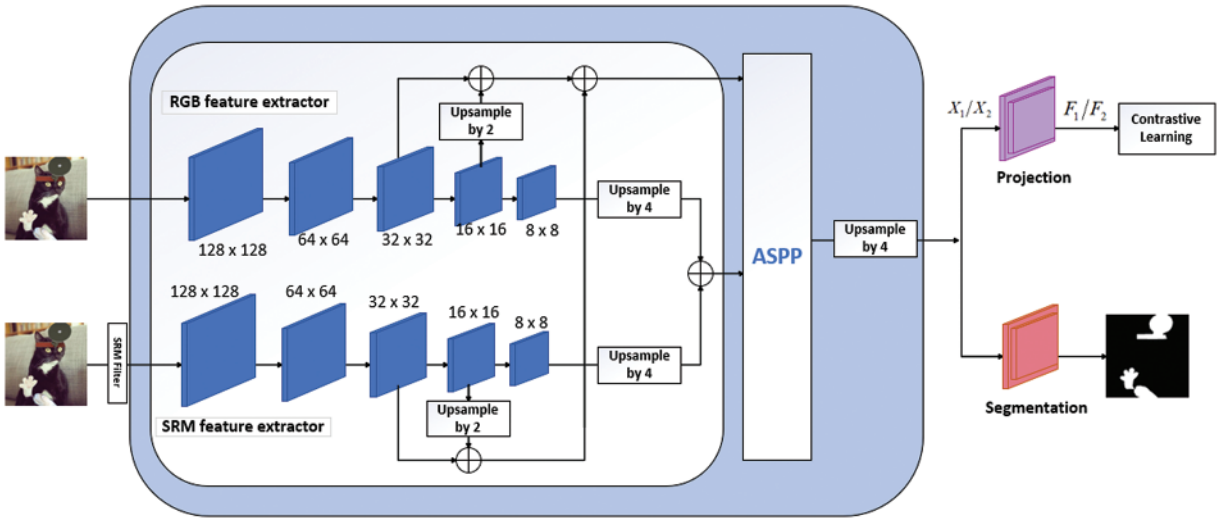


Figure 1: Overall model architecture

Subsequently, we integrate the dual-stream features distilled by our feature extractor, facilitating supervised contrastive learning and precise tamper region localization. To fuse the RGB and noise domain features effectively, we concatenate them along the channel dimension and feed them into the ASPP module for multi-scale fusion, which helps locate tampered regions of various sizes and provides more clues for image tamper localization. The fused features are used as inputs for both the projection head and the segmentation head. The projection head adopts a Conv-BatchNorm-Conv structure, and its output projection feature map is used for supervised contrastive learning. The segmentation head, designed following the DeepLab style, generates the final localization segmentation map.

Finally, we devise a joint optimization training strategy based on supervised contrastive learning. To improve the model's generalization ability, we devise a joint optimization training strategy that optimizes the contrastive loss for both low-level and high-level feature maps of the feature extractor. By

exploiting the feature distribution discrepancy between these levels, we effectively separate tampered and untampered regions. The details are given in Section 3.3.

3.2 Construction of the Feature Extractor

In image tamper localization research, the CNN-based localization model faces the challenge of effectively extracting global context information, which hinders its ability to locate tampered regions with various sizes precisely. Conversely, the Transformer-based localization model has excellent global context modeling ability, it often neglects local information and has problems such as low computational efficiency and dependence on large amounts of data.

The proposed model aims to fuse the strengths of CNN and Transformer to improve the localization performance of tampered regions with various sizes under constrained sample conditions. Our proposed model aims to fuse the advantages of CNN and Transformer architectures, thereby improving the model's performance in accurately localizing tampered regions with various sizes under constrained sample conditions. To this end, we design the feature extractor based on MobileViT for both the RGB and noise domains, and integrate the Focusing Linear Attention mechanism. The architecture of our feature extractor is shown in Fig. 2. Specifically, the MV2 module denotes an Inverted Residuals block [26], which effectively alleviates gradient vanishing or exploding problems, thus enhancing the model's training efficiency and accuracy.

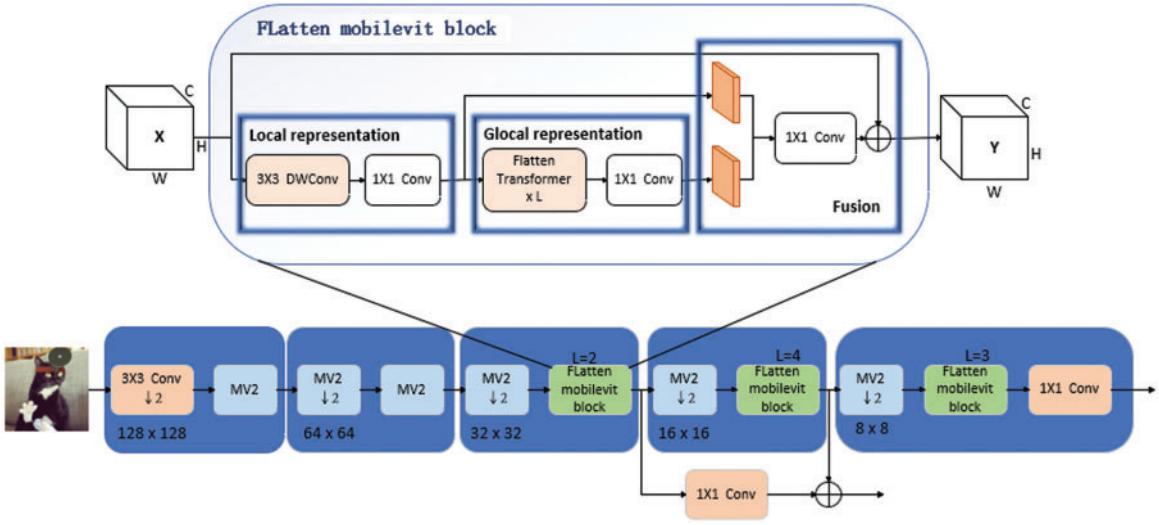


Figure 2: Feature extractor architecture

The FLatten MobileViT block is the core part of the feature extractor, inspired by the Focusing Linear Attention module proposed in [21]. We improve the global representation module in the MobileViT block by introducing a novel focusing function $f_p(x)$, as shown in Eq. (1), to replace the softmax attention function [27] used in Transformer. The softmax function computes the similarities between all query-key pairs to form an attention map, with a computational complexity of $O(N^2)$. In contrast, the focusing function $f_p(x)$ effectively adjusts the direction of each query-key pair (Q-K) by choosing a suitable parameter value p, bringing similar Q-K pairs closer and different Q-K pairs further apart, simulating the nonlinear weighting mechanism of softmax attention. This mode helps to concentrate attention on essential features and improves the localization performance of the linear attention model.

$$f_p(x) = \frac{\|x\|}{\|x^{**p}\|} x^{**p} \tag{1}$$

where x denotes the input N tokens, $x \in R^{N\times C}$; $\|\cdot\|$ denotes the norm of the feature; x^{**p} denotes the element-by-element power p of x, and p denotes the degree to which the focusing function pulls the vector (the value of p is set to 3 in this paper).

Furthermore, replacing the attention mechanism of the Transformer with linear attention leads to a rank reduction of the attention matrix, resulting in many rows in the attention map becoming homogeneous. Since the attention output is a weighted sum of the same set of V, this homogeneity may lead to a loss of diversity in the extracted features, affecting the localization of tampered regions. To solve this problem, we also choose to introduce a depthwise convolution (DWC) [16] into the attention matrix, which enriches the diversity of feature extraction while keeping low computational complexity. The specific implementation steps are as follows:

Step 1: We input the feature map matrix $X \in R^{C\times H\times W}$ into the Flatten MobileViT block, and after processing by the local representation block, we split it into N patches to obtain the feature matrix $x \in R^{N\times C}$. Then, we apply linear transformations to x to obtain the Q, K and V matrices, as shown in Eq. (2).

$$Q = xW_Q, K = xW_K, V = xW_V \tag{2}$$

where W_Q, W_K, $W_V \in R^{C\times C}$ denotes the learnable linear transformation matrix; $Q \in R^{N\times C}$, $K \in R^{N\times C}$, $V \in R^{N\times C}$ denotes the three matrices used to compute the attention weights in the self-attention mechanism.

Step 2: The function $ReLU$ is applied to guarantee the non-negativity and denominator validity of the input for $f_p(\cdot)$, as shown in Eq. (3).

$$\begin{aligned} F_p(Q) &= f_p(ReLU(Q)) \\ F_p(K) &= f_p(ReLU(K)) \end{aligned} \tag{3}$$

Step 3: The $F_p(\cdot)$ function is introduced to approximate the original similarity function, as shown in Eq. (4). The Transformer self-attention mechanism is reformulated by combining matrix multiplication with Eq. (4). A rearrangement of calculations is performed to reduce the computational complexity to $O(n)$, where $K^T V$ is computed before $Q\left(K^T V\right)$, as illustrated in Eq. (5).

$$Sim\left(Q_i, K_j\right) = F_p(Q_i) F_p\left(K_j\right)^T \tag{4}$$

$$\begin{aligned} O_i &= \sum_{j=1}^{N} \frac{Sim\left(Q_i, K_j\right)}{\sum_{j=1}^{N} Sim\left(Q_i, K_j\right)} V_j \\ &= \sum_{j=1}^{N} \frac{F_p(Q_i) F_p\left(K_j\right)^T}{\sum_{j=1}^{N} F_p(Q_i) F_p\left(K_j\right)^T} V_j \\ &= \frac{F_p(Q_i) \sum_{j=1}^{N} F_p\left(K_j\right)^T V_j}{F_p(Q_i) \sum_{j=1}^{N} F_p\left(K_j\right)^T} \end{aligned} \tag{5}$$

where $Sim(\cdot,\cdot)$ denotes the similarity function; $i, j \in \{1, 2, 3, \ldots, N\}$ denotes the indexes of the Q, K and V matrix elements; O_i denotes the attention weight corresponding to the *i-th* element.

Step 4: Add a depthwise convolution (DWC) module when computing the attention matrix using the focus function $f_p(x)$. The output of the depthwise convolution (DWC) module is shown in Eq. (6).

$$O = Sim(Q, K)\, V = F_p(Q)\, F_p(K)^T\, V + DWC(V) \tag{6}$$

where O denotes the attention matrix.

3.3 Supervised Contrastive Learning and Joint Optimization Training Strategy

In image tamper localization, models trained with cross-entropy loss are prone to extract similar features from tampered regions. However, different tampering techniques leave distinctive traces. Without additional training constraints, the model tends to overfit on specific tampering traces, limiting its generalization performance. To mitigate this issue, we introduce supervised contrastive learning in the training phase of the FL-MobileViT model. Based on the disparity in the feature distribution of the tampered images, we utilize the ground truth labels to compute the contrastive loss, which helps us separate between tampered and untampered regions. This method reduces the model's dependence on specific tampering traces and enhances its generalization capability. Furthermore, we design a novel joint optimization training strategy for our model, which optimizes the contrastive loss of both low-level and high-level feature maps simultaneously during the training phase. This strategy leverages feature maps at different levels to effectively capture tamper features and their distribution disparity, thereby further enhancing the model's generalization ability.

Assuming the given sample image $I \in R^{3\times H\times W}$, we input it into the SRM filter for processing to acquire the high-frequency image $I' \in R^{3\times H\times W}$. Subsequently, the high-frequency image I' is input into the SRM feature extractor, and the sample image I is input into RGB feature extraction, thereby extracting image features within a dual-stream architecture.

1) *The acquisition of low-level projected feature maps F_1*

We extract the feature maps from the third and fourth stages of the RGB feature extractor (as shown in Fig. 1, with sizes of 32 × 32 and 16 × 16, respectively). To retain more tampering traces, we use Padding technique to resize them to 32 × 32 and unify the channel number to 256 with a 1 × 1 convolution. Then, we concatenate these feature maps along the channel dimension and upsample them by a factor of four to obtain the feature map $F_r \in R^{512\times 128\times 128}$. We apply the same method to process the feature map from the SRM feature extractor, obtaining the feature map $F_s \in R^{512\times 128\times 128}$. We concatenate F_r and F_s along the channel dimension and feed them into an ASPP module for multi-scale fusion, obtaining a multi-scale context fusion feature $X_1 \in R^{512\times 128\times 128}$. Next, we input X_1 into the projection head of the Conv-BatchNorm-Conv architecture, as shown in Eq. (7), to obtain the projection map features $F_1 \in R^{256\times 128\times 128}$.

$$F_1 = Projection(X_1) \tag{7}$$

2) *The acquisition of high-level projected feature maps F_2*

We extract the feature maps of the fifth stage from both the RGB feature extractor and the SRM feature extractor (as shown in Fig. 1, with a size of 8 × 8). These feature maps are then upsampled by a factor of four to obtain a 32 × 32 RGB flow feature map F_R and an SRM noise flow feature map F_S, which are concatenated into the ASPP module along the channel dimension and further upsampled by a factor of four to acquire multi-scale context fusion feature $X_2 \in R^{512\times 128\times 128}$. Subsequently, we input this feature into the projection head to obtain

the projection feature map $F_2 \in R^{256\times128\times128}$, as shown in Eq. (8).

$$F_2 = Projection\,(X_2) \tag{8}$$

3) *Computation of the contrastive loss function*

Firstly, we select the low-level projection feature map F_1 and divide it spatially into $k \times k$ blocks of pixel embeddings. Then, we average all the pixel embeddings within each block to obtain $k \times k$ pixel embeddings $f_i^1 \in R^{256}$, $i \in \{1, 2, 3, \ldots, k^2\}$, forming the low-level feature pixel embedding feature map $f^1 \in R^{256\times k\times k}$. In the same way, we apply this method to the high-level projection feature map F_2 and obtain $k \times k$ pixel embeddings $f_i^2 \in R^{256}$, $i \in \{1, 2, 3, \ldots, k^2\}$, which constitute the high-level pixel embedding feature map $f^2 \in R^{256\times k\times k}$.
Secondly, we also partition the ground truth mask of the sample image $I \in R^{3\times H\times W}$ into $k \times k$ blocks, where a value of 1 represents the tampered region and a value of 0 represents the untampered region. Counting the number of 0s and 1s within the block, the highest number of labeled values are taken as the labels of the block to obtain the embedded block labels $m_i \in R$, $i \in \{1, 2, 3, \ldots, k^2\}$, which constitute the sample image truth label embedding $m \in R^{k\times k}$.

Finally, we adopt the contrastive loss function proposed in [24] (as shown in Eq. (9)) to separate the tampered and untampered regions. For the low-level feature pixel embedding feature map f^1, we split each pixel embedding ($k \times k$) in f^1 into positive pixel embedding z_1^+ and negative pixel embedding z_1^- according to the real label embedding m. By bringing f_i^1, z_1^+ and z_1^- into Eq. (9), we compute the contrastive loss L_i^1, $i \in \{1, 2, 3, \ldots, k^2\}$ of each pixel embedding in f^1. Likewise, for the high-level feature pixel embedding feature map f^2, we calculate the contrastive loss L_i^2 of each pixel embedding in f^2 using the same method.

$$L_i = \frac{1}{|A_i|}\sum_{z^+\in A_i} -log\frac{exp\,(f_i \cdot z^+/\tau)}{exp\,(f_i \cdot z^+/\tau) + \sum_{z^-} exp\,(f_i \cdot z^-/\tau)} \tag{9}$$

where z^+ denotes the positive pixel embedding of f_i and z^- denotes the negative pixel embedding of f_i; A_i denotes the set of all f_i corresponding positive pixel embeddings z^+; $i \in \{1, 2, 3, \ldots, k^2\}$ denotes the index of the corresponding pixel embedding in the sample image; and τ denotes the contrastive temperature.

4) *Jointly optimize the training strategy*

We employ the joint optimization training strategy, simultaneously optimizing the contrastive losses L_i^1 and L_i^2 of the two feature maps. Since the feature maps F_1 and F_2 for supervised contrastive have the same size and dimension, the loss $L_i^{'}$ can be computed for each embedding i of the two feature maps, as shown in Eq. (10).

$$L_i^{'} = \lambda L_i^1 + (1-\lambda)\,L_i^2 \tag{10}$$

where λ is a hyperparameter used to balance the effects of two contrastive losses in the jointly optimized training strategy.

We compute the average of all embeddings $L_i^{'}$, $i \in \{1, 2, 3, \ldots, k^2\}$ to obtain the contrastive loss L_{con} of a single sample image, which is used as an additional training constraint for the model. Combined with the cross-entropy loss, the model's final optimization objective L is formed, as shown in Eqs. (11) and (12).

$$L_{CON} = \frac{1}{k^2} \sum_{i \in k^2} L_i^{'} \tag{11}$$

$$L = L_{CE} + L_{CON} \tag{12}$$

where L_{CE} denotes cross-entropy loss.

In summary, the overall devise concept of supervised contrastive learning and joint optimization training strategy is shown in Algorithm 1.

Algorithm 1: Joint optimization training strategy

Input: the sample image I and the high frequency image I'

$F_r, F_R \leftarrow$ RGB feature extractor (I)
$F_s, F_S \leftarrow$ SRM feature extractor (I)
$X_1 \leftarrow$ AsppModel (concatenate (F_r, F_s))
$X_2 \leftarrow$ AsppModel (concatenate (F_R, F_S))
$F_1 \leftarrow$ projection (X_1)
$F_2 \leftarrow$ projection (X_2)
$f^1 \leftarrow$ AvgPool2d $(F_1, 2)$
$f^2 \leftarrow$ AvgPool2d $(F_2, 2)$
$m \leftarrow$ AvgPool2d $(M, 4)$
for each f_i **in** f^1, f^2 **do**
    $z_1^+, z_1^-, A_i^1 \leftarrow$ DividePositiveAndNegativeEmbeddings (f^1, m, m_i, f_i^1)
    $z_2^+, z_2^-, A_i^2 \leftarrow$ DividePositiveAndNegativeEmbeddings (f^2, m, m_i, f_i^2)
    $L_i^1 \leftarrow \frac{1}{\left|A_i^1\right|} \sum_{z_1^+ \in A_i^1} -log \frac{exp\left(f_i^1 \cdot z_1^+ / \tau\right)}{exp\left(f_i^1 \cdot z_1^+ / \tau\right) + \sum_{z_1^-} exp\left(f_i^1 \cdot z_1^- / \tau\right)}$
    $L_i^2 \leftarrow \frac{1}{\left|A_i^2\right|} \sum_{z_2^+ \in A_i^2} -log \frac{exp\left(f_i^2 \cdot z_2^+ / \tau\right)}{exp\left(f_i^2 \cdot z_2^+ / \tau\right) + \sum_{z_2^-} exp\left(f_i^2 \cdot z_2^- / \tau\right)}$
    $L_i^{'} \leftarrow \lambda L_i^1 + (1 - \lambda) L_i^2$
end for
for $i = 1$ **to** k^2 **do**
    $L_{CON} = \frac{1}{k^2} \sum_{i \in k^2} L_i^{'}$
end for
$L \leftarrow L_{CE} + L_{CON}$

4 Experiment

To evaluate the localization accuracy and generalization capability of the FL-MobileViT model, we conducted comprehensive experiments on three widely used image tampering datasets. We employe the Area Under Curve (AUC) scores at the pixel level as the evaluation metric, where higher scores indicate superior localization performance. These datasets encompass a variety of tamper types rather than being restricted to a singular one. Our experimental setup includes: (1) Comparative analysis of the FL-MobileViT model against other baseline models, along with visualization of its localization results; and (2) Ablation study of the FL-MobileViT model to validate the efficacy of the jointly optimized training strategy and feature extractor.

4.1 Implemntation Details

The PyTorch [28] framework is utilized to implement all codes involved in the experiment on the Pycharm software platform. The experimental PC was equipped with an Intel i7-11700K processor, a GeForce RTX 3060Ti graphics card, and operated on Windows10. The model parameters are configured as follows: the input image size is adjusted to 256×256 pixels; the projection feature maps are divided into 64×64 patches for computing contrast loss. To mitigate the impact of imbalanced sample labels on training, the tampered class was assigned a weight more than ten times when calculating cross-entropy loss. The Adam optimizer was selected with an initial learning rate set to 1e−4, and a decay strategy was implemented whereby the learning rate decreased by 20% every 20 training epochs. The training batch size is set to 8, and 150 training epochs are performed. The Linear Focused Attention parameter p is set to 3. The contrastive temperature τ value is set to 0.1, and the balance parameter λ range is defined as 1/2.

4.2 Datasets

In contrast to conventional methodologies, we did not use a large-scale synthetic tamper dataset for model pre-training, but rather trained and evaluated our model on a small-scale dataset. Specifically, we performed experimental analysis on five datasets: IMD-20 [29], CASIA [30], Coverage [31], Columbia [32] and NIST-16 [33]. Following the method in [14], we split each dataset into three subsets: training (train), validation (val), and testing (test). The IMD-20 dataset is a compilation of "real" image tampering datasets sourced from the Internet, encompassing various types of tampering. The CASIA dataset comprises two types of tampering: splicing and copy-move, with additional post-processing applied to the images, such as filtering and blurring. The NIST-16 dataset includes three types of tampering: splicing, copy-move, and removal, which is post-processed to conceal apparent traces of tampering. The Coverage dataset contains 100 images generated by copy-move techniques. The Columbia dataset focuses on splicing based on uncompressed images. Furthermore, these five datasets offer authentic ground truth (GT) masks for tampered regions, which we utilized for supervised contrastive learning to compute contrastive loss during training.

4.3 Comparison with Baseline Models

Under the same dataset, we conducted experiments to compare and analyze the performance between various baseline models and FL-MobileViT. The following is an overview of these baseline models:

RGB-N [11] employs a dual-stream parallel network architecture to mine tampering features from both the RGB and noise domains. SPAN [8] uses a pyramid structure and models the relationships between image blocks at different scales with self-attention mechanism. ManTraNet [34] utilizes feature extractors to detect tampering traces and localizes tampered regions with anomaly detection networks. TransForensics [14] combines visual transformers with dense self-attention encoders and dense correction modules, which model the interaction between the global context and the local blocks at different scales. PSCC-Net [35] adopts a lightweight backbone network and obtains both local and global information through a progressive mechanism. ObjectFormer [36] is based on the Transformer architecture and combines RGB features with high-frequency features to model the coherence of image blocks. TANet [37] introduces a stacked multi-scale Transformer (SMT) branches as a compensation for the feature representation of mainstream convolutional neural network branches. TBFormer [15] uses a dual-stream parallel network that extracts tampering features from both the RGB and noise

domains, by applying different stacked Transformer layers. CFL-Net [24] applies contrastive learning methods, combining contrastive loss and cross-entropy loss for model's training.

4.4 Comparative Study

To evaluate the efficacy of the FL-MobileViT model in image tamper localization, we performed a comparative analysis between FL-MobileViT and seven baseline models using the AUC score metric. It is worth mentioning that RGB-N and SPAN models were fine-tuned according to their respective papers. ObjectFormer, PSCC-Net, TANet, ManTraNet and TBFormer were all pre-trained on synthetic datasets, and the latter two did not receive further fine-tuning. Conversely, Transforensics and CFL-Net were not pre-trained on synthetic datasets. The detailed experimental results are shown in Table 1.

Table 1: AUC scores (in %) for FL-MobileViT model *vs*. 7 baseline models

Methods	IMD-20	CASIA	NIST	Colombia	Coverage
RGB-N [11]	–	79.5	93.7	85.8	81.7
SPAN [8]	75.0	83.8	96.1	93.6	92.2
MantraNet [34]	74.8	81.7	79.5	82.4	81.9
TransForensics [14]	84.8	83.7	–	–	88.4
PSCC-Net [35]	80.6	87.5	99.6	98.2	84.7
ObjectFormer [36]	82.1	88.2	99.6	95.5	92.8
TANet [37]	84.9	89.3	99.7	**98.7**	97.8
TBFormer [15]	86.3	95.5	99.7	–	–
CFL-Net [24]	89.9	86.3	99.7	–	–
Ours	**92.4**	**88.4**	**99.7**	95.3	**97.9**

As shown in Table 1, the FL-MobileViT model exhibits superior localization performance on the IMD-20 dataset, urpassing the performance of other baseline models. Specifically, the FL-MobileViT model achieves an AUC score of 92.40% on the IMD-20 dataset, which is 2.5% higher than the current state-of-the-art CFL-Net model. This result indicates that the FL-MobileViT model has significant advantages in locating tampered images in real-life scenarios. On the CASIA dataset, the FL-MobileViT model is slightly behind the TBFormer model, yet it still surpasses other baseline models. This can be attributed to the TBFormer model's utilization of a vast number of synthetic images generated from the CASIA dataset for pre-training, which are very similar to the distribution of the CASIA dataset. When applied to the NIST dataset, the FL-MobileViT model shows comparable localization performance with the TBFormer and ObjectForme models. On the Columbia and Coverage datasets, which contain only a single manipulation technique, the FL-MobileViT model is comparable to the best performing TANet. However, among these models, only our model and the CFL-Net model provide excellent localization performance without the need for pre-training with large-scale synthetic tampered datasets. These results suggests that the FL-MobileViT model is very suitable for situations with limited sample sizes.

Based on the aforementioned analysis, the outstanding performance of the FL-MobileViT model could be largely attributed to the joint optimization training strategy we designed. This strategy combined contrastive losses for low-level and high-level feature maps as extra training constraints

besides cross-entropy loss, enabling the model to learn more rich and diverse features and thus enhancing its generalization performance. To verify this conclusion, we trained the model separately on IMD-20, CASIA, NIST-16 and Coverage datasets and evaluated its generalization performance on different test sets. Furthermore, we examine the generalization performance across datasets with the CFL-Net model, which employs a contrastive learning. The specific experimental results are shown in Table 2.

Table 2: FL-MobileViT model AUC scores across datasets (in %), 'w/o' indicates training without joint optimization strategy, 'w' indicates training with joint optimization strategy. '∗' indicates that the data is from the corresponding model

Datasets	Methods		$IMD20_{test}$	$CASIA_{test}$	$NIST_{test}$	$Colombia_{test}$	$Coverage_{test}$
$IMD20_{train}$	Our	w/o	87.6	75.2	78.67	89.2	76.6
		w	**92.4**	**77.3**	**93.0**	**95.8**	**80.3**
	CFL-Net		89.9*	75.6*	91.8*	92.6	78.5
$CASIA_{train}$	Our	w/o	77.6	86.4	80.6	81.4	80.8
		w	**80.5**	**88.4**	**82.1**	**85.2**	82.4
	CFL-Net		77.8*	86.3*	79.9*	84.1	81.2
$NIST_{train}$	Our	w/o	67.4	68.3	98.5	66.5	67.1
		w	**70.2**	**68.9**	**99.7**	**68.7**	**68.3**
	CFL-Net		69.8*	67.6*	99.7*	66.9	67.6
$Coverage_{train}$	Our	w/o	60.2	62.8	64.5	63.1	95.6
		w	**63.6**	**63.5**	**67.3**	**65.3**	**97.9**
	CFL-Net		68.3	66.7	69.7	69.4	96.2

As shown in Table 2, the FL-MobileViT model achieved significant improvement in generalization performance across datasets after applying the joint optimization strategy. With this strategy, the model bolstered the localization performance in all training and testing scenarios. These results confirmed the efficacy of the joint optimization strategy in enhancing the model's generalization ability. By analyzing the model's training on the IMD-20 dataset and its evaluation on five test sets, we found that the joint optimization strategy increases the AUC score by 4.8% for the IMD-20 test set, by 2.1% for the CASIA test set, by 4%∼6% on the Columbia and Coverage datasets, and by a remarkable 14.3% for the NIST test set. We compared the training and test evaluation results of the model on the CASIA and NIST datasets and found that the performance improvement was most significant when the model was trained on the IMD-20 dataset. This could be attributed to the IMD-20's ability to collect real-life image tampering cases, which enabled the FL-MobileViT model to learn more generalizable features. Furthermore, given the limited number of tampered images in the NIST and Coverage dataset and the Coverage dataset employs a single tampering technique, the evaluation results on other datasets were relatively lower when using this dataset for training. However, the model performance still improves after applying the joint optimization strategy. By comparing the performance of our model with the CFL-Net model across datasets, our model achieves better results

under all five test sets. This result further confirmed the effectiveness of this strategy in enhancing the model's generalization ability.

4.5 Ablation Study

We performed a series of ablation experiments on the IMD-20 dataset to evaluate the impact of each module on the FL-MobileViT for image tamper localization. The experimental settings are divided as follows: (A)∼(E) are FL-MobileViT models with MobileViT's Transformer self-attention mechanism; (F)∼(J) are based on (A)∼(E) and incorporate the Focusing Linear Attention mechanism; (A) and (F) use only cross-entropy loss function for training; (B) and (G), (C) and (H), and (D) and (H) are trained with different additional constraints. (E) and (J) are FL-MobileViT models under the joint optimization strategy without using the ASPP module. Table 3 shows the ablation results of the loss function in the joint optimization training strategy and the Focusing Linear Attention mechanism in the feature extractor of the FL-MobileViT model, as well as the ablation results of the ASPP module in the model.

Table 3: Results of ablation experiments on the IMD-20 dataset with different loss combinations in the joint optimization training strategy and the focused linear attention mechanism in the feature extractor. L_{con1} is the contrastive loss calculated using the low-level feature map. L_{con2} is the contrastive loss calculated using the high-level feature map. L_{con} is the contrastive loss calculated using the joint optimization strategy

ID	L_{CE}	L_{con1}	L_{con2}	L_{con}	ASPP	AUC (%)
(A)	√				√	**86.3**
(B)	√	√			√	88.2
(C)	√		√		√	88.6
(D)	√			√	√	**90.2**
(E)	√			√		86.6
(F)	√				√	**87.6**
(G)	√	√			√	90.5
(H)	√		√		√	90.9
(I)	√			√	√	**92.4**
(J)	√			√		88.3

We evaluated the localization performance of FL-MobileViT models, which incorporate the Focusing Linear Attention mechanism, under different combinations of loss functions. By comparing the results of (A) and (F), (B) and (G), (C) and (H), and (D) and (I), we found that the Focusing Linear Attention mechanism significantly improved the localization performance of the FL-MobileViT. As stated in [21], this mechanism retained the global context modeling capability of Transformer's self-attention and effectively increased the diversity of feature computation, thus improving the performance of downstream tasks. Further analysis revealed that, under the joint optimization strategy, (D) and (I) showed greater performance improvements than (A) and (F), which did not employ the joint optimization strategy. Also, (D) and (I) show better localization performance than (E) and (J) without ASPP. These findings indicated that the combination of the Focusing Linear Attention mechanism, the joint optimization strategy and the ASPP module not only improved the tamper localization accuracy,

but also achieved the optimal performance in the FL-MobileViT model composed of module (H), thereby confirming that each module made a significant contribution to the localization accuracy.

To evaluate the efficacy of the joint optimization strategy for different combinations of loss functions, we conducted comparisons between (A) and (B), (F) and (G). The results show that applying only the contrastive loss from the low-level feature map as an additional training constraint significantly improved the model's AUC score. Similarly, when comparing (A) with (C) and (F) with (H), we found that using the contrastive loss of the high-level feature map alone as an extra training constraint also substantially increased the AUC score. These findings indicate that the tampering information in both low-level and high-level feature maps positively affected the model's performance, which contributes to distinguishing tampered and untampered regions. Furthermore, through the comparison of (B) (C) with (D) and (G) (H) with (I), we observed that our joint optimization strategy, which integrated contrastive losses from low-level and high-level feature maps as extra training constraints, can further boost the AUC score. This confirms that our joint optimization strategy effectively integrated both contrastive losses and significantly improved the model's localization accuracy.

We followed the method of [16] and further evaluate the impact of the focused linear attention mechanism on the inference time and localization performance of the FL-MobileViT model, with a comparison made to the original MobileViT without the focused linear attention mechanism. The result is presented in Fig. 3. It can be observed that our model achieves a significantly higher AUC score and a shorter inference time. This result confirms the effectiveness of the Focused Linear Attention mechanism in reducing the computational complexity and improving the localization performance of the model.

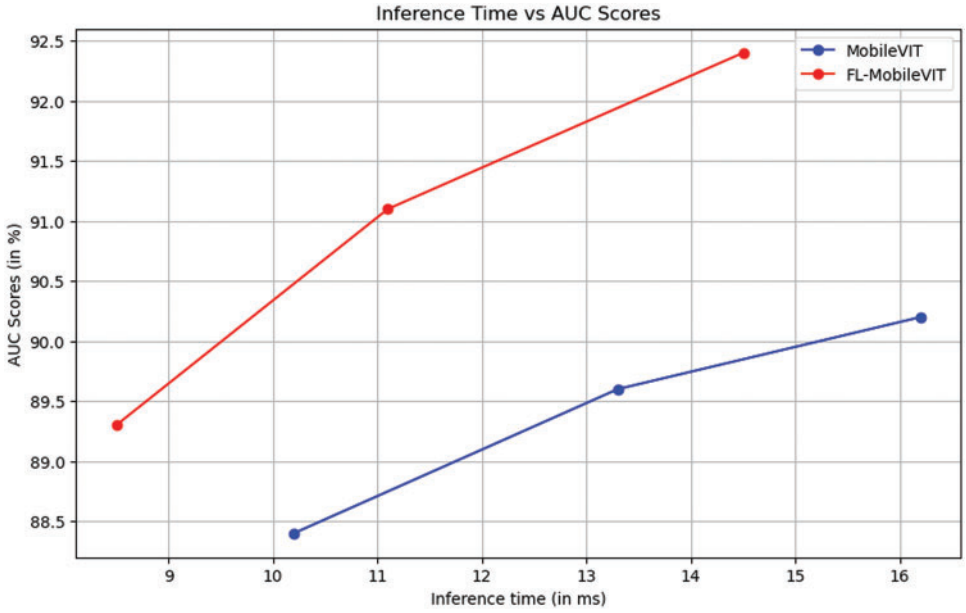
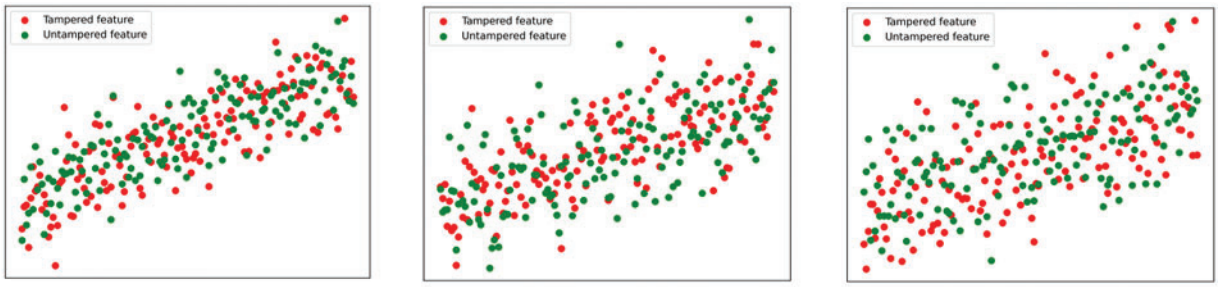


Figure 3: AUC-Inference time curveon IMD-20. Inference time is tested with image resolution 256×256

4.6 Qualitative Visualization Analysis

To provide a more intuitive demonstration of the efficacy of the joint optimization training strategy, we evaluated the performance of various combinations of loss functions in preventing the model from focusing on specific tampering types. Using the dimensionality reduction techniques, we conducted the experiment on the IMD-20 test set, which performed qualitative visual analysis by projecting the class features from the segmentation head output to a two-dimensional space (as shown in Fig. 4).



(a) L_{CE} (b) $L_{CE} + L_{con2}$ (c) $L_{CE} + L_{CON}$

Figure 4: Visualization of class feature distribution. Red = tampered features, green = untampered features

In Fig. 4a, the features corresponding to the tampered and untampered regions are densely clustered, indicating that the model trained only with cross-entropy loss tends to cluster similar class features. On the other hand, in Fig. 4b, the features of these two regions are more scattered, demonstrating that the addition of contrastive loss on the final feature map effectively avoids excessive feature clustering. Further observation of Fig. 4c reveals a higher dispersion of the features in both regions compared to Fig. 4b. This suggests that introducing contrastive loss on the low-level feature map further reduces the aggregation of class features, which makes different tampering traces more distinguishable, thereby improving the model's generalization performance. The aforementioned experimental results demonstrate that the joint optimization strategy can significantly enhance the model's generalization performance by dispersing the feature distribution, thereby mitigating the issue of cross-entropy loss driving the extraction of similar tampering features.

To visually demonstrate the efficacy of the FL-MobileViT model to locate tampered regions, we show the model's predicted masks on some tampered images from the IMD-20 dataset and compare them with the predicted masks of the CFL-Net, TBFormer and TANet (Fig. 5). The FL-MobileViT model exhibits its applicability in accurately localizing tampered regions of various sizes, demonstrating its practicality for real scenarios tampered image localization. It can be observed that our model exhibits superior localization accuracy in comparison to other models, with the exception of TANet, which outperforms it in the identification of region boundaries. For instance, in the first line of Fig. 5, the TBFormer incorrectly locates the kite line as a tamper region; In the third line, the CFL-Net model locates some areas inaccurately; The TANet model is more effective at identifying a clear boundary, whereas the FL-MobileViT model is more accurate in locating the tampered region, although the boundary is not as well defined.

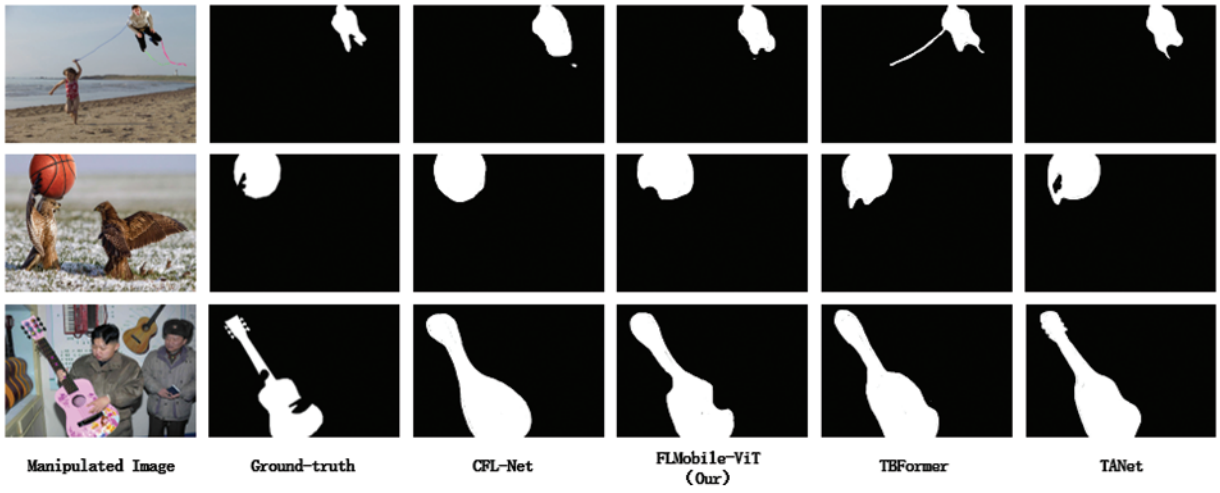


Figure 5: Example of image tampering localization on the IMD-20 dataset

5 Conclusion

We propose FL-MobileViT, an improved MobileViT model for image tamper localization, aiming to improve the localization capability of real-life tampered images. The architecture of FL-MobileViT is constructed by incorporating the Focusing Linear Attention mechanism into the MobileViT network. This novel design includes two feature extractors for extracting tamper features from different dimensions: one targets the RGB domain, while the other focuses on the noise domain. By leveraging these feature extractors, our approach significantly enhances the model's ability to localize tampered regions of various sizes, even with limited training samples. Moreover, our model adopts supervised contrastive learning and employs a joint optimization training strategy. By calculating contrastive loss on different layers' feature maps, it effectively discriminates between tampered and untampered regions, leveraging the disparities in image feature distributions. Consequently, it significantly bolsters the model's generalization performance. Experimental results on five commonly used tamper datasets demonstrate that our proposed model outperforms other high-level baseline models in terms of competitive advantages. Particularly noteworthy is its excellent applicability in locating tampered images in real-life scenarios, as demonstrated by experiments on the IMD-20 dataset.

FL-MobileViT is used to locate tampered images, which is able to accurately localize the tampered region, but suffers from unclear boundaries of the localized region. Therefore, in the future, we can improve the extraction and utilization of boundary features of the model to further improve the boundary clarity of the tampered region. In addition, due to the rapid development of diffusion modeling, it is an interesting research direction to distinguish synthetic images from natural images. Therefore, we will try to apply the joint optimization strategy to synthetic image recognition.

Acknowledgement: This study was funded by the Science and Technology Project in Xi'an.

Funding Statement: This study was funded by the Science and Technology Project in Xi'an (No. 22GXFW0123), this work was supported by the Special Fund Construction Project of Key Disciplines in Ordinary Colleges and Universities in Shaanxi Province, the authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

Author Contributions: Conceptualization, Huanqi Liu; Data curation, Wenyan Hou; Formal analysis, Fengling Zhang and Jingtao Sun; Investigation, Fengling Zhang; Methodology, Fengling Zhang and Jingtao Sun; Software, Wenyan Hou; Supervision, Huanqi Liu; Writing—review & editing, Wenyan Hou. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: IMD-20 available at http://staff.utia.cas.cz/novozada/db, accessed on 20 June 2023. CASIA available at http://forensics.idealtest.org/, accessed on 20 June 2023. NIST available at https://www.nist.gov/itl/iad/mig, accessed on 20 June 2023. Columbia available at https://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm, accessed on 20 June 2023. Coverage is available at https://github.com/wenbihan/coverage, accessed on 20 June 2023.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Ding, J. Zhang, C. Wu, C. Cai, and G. Chen, "Real-time image inpainting using patchmatch based two-generator adversarial networks with optimized edge loss function," in *2022 IEEE Int. Symp. Circuits Syst. (ISCAS)*, Austin, TX, USA, 2022, pp. 3145–3149.

[2] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, NOLA, LA, USA, 2022, pp. 2426–2435.

[3] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for StyleGAN image manipulation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–14, 2021.

[4] A. Aminuddin and F. Ernawan, "AuSR3: A new block mapping technique for image authentication and self-recovery to avoid the tamper coincidence problem," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 9, pp. 101755, 2023. doi: 10.1016/j.jksuci.2023.101755.

[5] B. Liu and C. M. Pun, "Exposing splicing forgery in realistic scenes using deep fusion network," *Inf. Sci.*, vol. 526, no. 10, pp. 133–150, 2020. doi: 10.1016/j.ins.2020.03.099.

[6] J. L. Zhong and C. M. Pun, "An end-to-end dense-inceptionNet for image copy-move forgery detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 2134–2146, 2019. doi: 10.1109/TIFS.2019.2957693.

[7] H. Li and J. Huang, "Localization of deep inpainting using high-pass fully convolutional network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Republic of Korea, 2019, pp. 8301–8310.

[8] X. Hu *et al.*, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Comput. Vis.-ECCV 2020: 16th Eur. Conf.,* Glasgow, UK, Springer International Publishing, 2020.

[9] Y. Deng, "Utilizing sensitive features for image tampering detection," in *2022 IEEE 5th Int. Conf. Inf. Syst. Comput. Aided Educ. (ICISCAE)*, Chengdu, China, IEEE, 2022, pp. 109–112.

[10] A. A. Aminu, N. N. Agwu, and A. Steve, "Detection and localization of image tampering using deep residual UNET with stacked dilated convolution," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 9, pp. 203–211, 2021.

[11] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1053–1061.

[12] Y. Huang, S. Bian, H. Li, C. Wang, and K. Li, "DS-UNet: A dual streams UNet for refined image forgery localization," *Inf. Sci.*, vol. 610, no. 2, pp. 73–89, 2022. doi: 10.1016/j.ins.2022.08.005.

[13] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45.3, pp. 3539–3553, 2022.

[14] J. Hao, Z. Zhang, S. Yang, D. Xie, and S. Pu, "TransForensics: Image forgery localization with dense self-attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 15055–15064.

[15] Y. Liu, B. Lv, X. Jin, X. Chen, and X. Zhang, "TBFormer: Two-branch transformer for image forgery localization," *IEEE Signal Process. Lett.*, vol. 30, pp. 623–627, 2023. doi: 10.1109/LSP.2023.3279018.
[16] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," arXiv preprint arXiv:2110.02178, 2021.
[17] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," arXiv preprint arXiv:2206.02680, 2022.
[18] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," arXiv preprint arXiv:2209.15159, 2022.
[19] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proc. IEEE/CVF Winter Conf. App. Comput. Vis.*, 2021, pp. 3531–3539.
[20] D. Bolya, C. Y. Fu, X. Dai, P. Zhang, and J. Hoffman, "Hydra attention: Efficient attention with many heads," in *Eur. Conf. Comput. Vis.*, Cham, Springer Nature Switzerland, Tel-Aviv, 2022, pp. 35–49.
[21] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Paris, France, 2023, pp. 5961–5971.
[22] P. Khosla *et al.*, "Supervised contrastive learning," *Adv. Neur. Inf. Process. Syst.*, vol. 33, pp. 18661–18673, 2020.
[23] Y. Xu, J. Zheng, A. Fang, and M. Irfan, "Feature enhancement and supervised contrastive learning for image splicing forgery detection," *Digit. Signal Process.*, vol. 136, no. 5, pp. 104005, 2023. doi: 10.1016/j.dsp.2023.104005.
[24] F. F. Niloy, K. K. Bhaumik, and S. S. Woo, "CFL-Net: Image forgery localization using contrastive learning," in *Proc. IEEE/CVF Winter Conf. App. Comput. Vis.*, Waikoloa, HI, USA, 2023, pp. 4642–4651.
[25] L. C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 801–818.
[26] M. Sandler *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.
[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neur. Inf. Process. Syst. (NIPS'17)*, Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 6000–6010.
[28] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neur. Inf. Process. Syst.*, Red Hook, NY, USA, Curran Associates Inc., 2019, vol. 721, pp. 8026–8037.
[29] A. Novozamsky, M. Babak, and S. Stanislav, "IMD 2020: A large-scale annotated dataset tailored for detecting manipulated images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, Snowmass Village, CO, USA, 2020, pp. 71–80.
[30] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *2013 IEEE China Summit Int. Conf. Signal Inf. Process.*, Beijing, China, IEEE, 2013, pp. 422–426.
[31] B. Wen *et al.*, "COVERAGE—A novel database for copy-move forgery detection," in *2016 IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, IEEE, 2016, pp. 161–165.
[32] T. T. Ng, J. Hsu, and S. F. Chang, "Columbia image splicing detection evaluation dataset," 2009. Accessed: Jun. 20, 2023. [Online]. Available: https://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm
[33] NfFORMATION TECHNOLOGY LABORATORY of NIST, "NIST: Nimble 2016 datasets," Gaithersburg, MD, USA, 2016. Accessed: Jun. 20, 2023. [Online]. Available: https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation
[34] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 9543–9552.
[35] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32.11, pp. 7505–7517, 2022. doi: 10.1109/TCSVT.2022.3189545.

[36] J. Wang *et al.*, "ObjectFormer for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 2364–2373.

[37] Z. Shi, H. Chen, and D. Zhang, "Transformer-auxiliary neural networks for image manipulation localization by operator inductions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4907–4920, Sep. 2023. doi: 10.1109/TCSVT.2023.3251444.