Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053740

ARTICLE

IMTNet: Improved Multi-Task Copy-Move Forgery Detection Network with Feature Decoupling and Multi-Feature Pyramid

Huan Wang[1], Hong Wang[1], Zhongyuan Jiang[2,*], Qing Qian[1] and Yong Long[1]

[1]School of Information, Guizhou University of Finance and Economics, Guiyang, 550025, China

[2]College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang, 550025, China

*Corresponding Author: Zhongyuan Jiang. Email: jiangzydj@163.com

Received: 09 May 2024 Accepted: 13 August 2024 Published: 12 September 2024

ABSTRACT

Copy-Move Forgery Detection (CMFD) is a technique that is designed to identify image tampering and locate suspicious areas. However, the practicality of the CMFD is impeded by the scarcity of datasets, inadequate quality and quantity, and a narrow range of applicable tasks. These limitations significantly restrict the capacity and applicability of CMFD. To overcome the limitations of existing methods, a novel solution called IMTNet is proposed for CMFD by employing a feature decoupling approach. Firstly, this study formulates the objective task and network relationship as an optimization problem using transfer learning. Furthermore, it thoroughly discusses and analyzes the relationship between CMFD and deep network architecture by employing ResNet-50 during the optimization solving phase. Secondly, a quantitative comparison between fine-tuning and feature decoupling is conducted to evaluate the degree of similarity between the image classification and CMFD domains by the enhanced ResNet-50. Finally, suspicious regions are localized using a feature pyramid network with bottom-up path augmentation. Experimental results demonstrate that IMTNet achieves faster convergence, shorter training times, and favorable generalization performance compared to existing methods. Moreover, it is shown that IMTNet significantly outperforms fine-tuning based approaches in terms of accuracy and F_1.

KEYWORDS

Image copy-move detection; feature decoupling; multi-scale feature pyramids; passive forensics

1 Introduction

Malicious image forgeries, including techniques such as copy-move, splicing, and removal, can severely undermine the credibility and integrity of digital images. The copy-move tampering method, being one of the most prevalent image tampering techniques, is highly concealable and presents significant challenges for detection. Moreover, the characteristics (such as the saturation, light source, and noise) of the tampered areas can be adapted easily without affecting the original image properties. Therefore, CMFD methods have attracted the attention of forensic science scholars.

The traditional CMFD methods can be classified into two main categories: block-based methods and keypoint-based methods. In the block-based methods, discrete cosine transform (DCT) technologies are generally used in image processing because of their energy compaction properties. The work in

[1] analyzes the exhaustive search algorithms and proposes a block matching detection method based on DCT. This work is one of the landmark methods in CMFD methods. Each block contains 64 (8 × 8) features and any two feature vectors that are within a certain range should be matched to determine the duplicated regions. However, the proposed method cannot detect the small duplicated regions and the detection precision is dissatisfactory because of its lexicographically sorting algorithm. The study in [2] improves the method of [1] by reducing the number of features to a quarter that is located in the low frequency parts. However, the detection accuracy is unsatisfactory since some DCT coefficients that are located in the intermediate frequency parts are truncated. A Discrete Cosine Transformation (DCT) and Singular Value Decomposition (SVD) based technique is proposed to detect the copy-move image forgery in [3], the combination of DCT and SVD makes the proposed scheme robust against compression, geometric transformations, and noise. A hybrid method is reported to classify copy-move and splicing images based on the texture information of images in the spatial domain [4]. The proposed method divides the image into equal blocks to get scale-invariant features. The tampered image regions can be detected by matching the scale-invariant features. The proposed method is robust to most regular signal processing type attacks. However, it is less effective against some geometric transformation-type attacks.

In the keypoint-based methods, the underlying principle is that modifications made to the image, such as copy-move operations, will alter the local features and consequently impact the distribution and characteristics of the detected keypoints. By analyzing the changes in the keypoint-based representations, these techniques aim to identify the presence of tampering in the image. Amerini et al. in [5] propose a novel methodology based on a scale invariant features transform (SIFT) method. The proposed method can be used to determine whether a copy-move attack has occurred in an image. It also can be used to recover the geometric transformation that is used to perform cloning technologies. Furthermore, the proposed method can be used to individuate the altered areas and estimate the geometric transformation parameters with high reliability. The work in [6] reports an improved SIFT structure with inherent scaling invariance that is designed to enhance the capability of extracting effective keypoints in the homogeneous region. Zhong et al. in [7] analyze the structure and excavate the inherent characteristics of local descriptors (SURF) for feature extraction in the coarse and smooth regions. Subsequently, the proposed method utilizes kernel features for coarse feature matching to reduce matching costs. Following this, a smaller set of candidate keypoints is identified, which are then used in conjunction with complete features to conduct fine keypoint matching in order to identify suspicious candidate keypoint pairs. A method is proposed in [8] to find and locate the duplicated and pasted portions of a manipulated image by using the combination of Hessian and Raw patch features. In the proposed method, a parallelism condition is applied together with a random sample consensus method to eliminate mismatches. The proposed method was shown to be effective by obtaining high F_1 scores in images that are attacked with noise, JPEG compression, and scaling operations.

Traditional image tampering detection methods have faced growing challenges in keeping pace with the rapidly evolving landscape of image forgery techniques. In response, the increased pervasiveness of deep learning technologies has led to the development of contemporary copy-move forgery detection algorithms that predominantly leverage specialized neural network architectures, which are purposefully designed and trained for the task of image tampering identification. It is proposed in [9] that an end-to-end approach called BusterNet can identify the source and target regions by detecting image similarity through parallel branching. However, this method requires high accuracy on both branches. The work in [10] reports a serial branching network that is used to improve the drawbacks of BusterNet. The reported network consists of a copy-move similarity detection network and a source/target region distinguishment network. The branching network is simpler and more accurate compared with the BusterNet. However, generalization was powerless. The study in [11] proposes the Dense-InceptionNet that combines DenseNet and InceptionNet, by utilizing multiscale information and dense features. The study presented in [12] introduces the Spatial Pyramidal Attention Network which is designed to capture inter-block relationships across multiple scales through a pyramidal structure of locally self-attentive blocks. However, this approach demonstrates diminished effectiveness at lower image resolutions and overlooks the interplay between high-dimensional and low-dimensional features. The research in [13] reports a deep learning method for forgery detection at both image and pixel levels. In this method, authors used a pre-trained deep model with a global average pooling (GAP) layer instead of default fully connected layers to detect forgery. The GAP layer creates a good dependency between the feature maps and the classes. The study in [14] proposes the Laterally Linked Pixel (LLP) algorithm, which utilizes two-dimensional arrays and a single layer derived from a unit-linking pulsed neural network to detect copied regions. The method employs kernel tricks to identify multiple manipulations within a single forged image. The accuracy obtained through the LLP algorithm is about 90% and further forgery detection is improved based on optimized kernel selections in classification algorithm.

While extensive experimental evaluations have demonstrated the satisfactory performance of the aforementioned specialized neural network-based copy-move forgery detection methods, such approaches inherently compromise the broader generalizability of the underlying network architecture. This runs counter to the primary objective of neural networks, which is to learn robust, generalizable representations that can be effectively applied across a diverse range of related tasks and domains. Consequentially, the following problems arise: (1) Deep neural networks (DNNs) require a substantial quantity of high-quality labeled datasets [15]. (2) DNNs need hardware with higher computing capacity. (3) As a data-driven algorithm, DNNs produce individual results based on various types of data. However, collecting an enormous amount of data does not provide a comprehensive representation. (4) DNNs aim to create a general model that could cater to the needs of different users, environments, and devices. Consequently, there remains an urgent need to adapt and refine the general model to address personalized tasks effectively [16].

Transfer learning is concerned with the transfer of knowledge across domains, leveraging prior experience as a bridge to facilitate the adaptation from one scenario to another. Among its various subfields, feature decoupling stands out as a significant subclass, demonstrating a broad spectrum of applications. Feature decoupling is an approach to designing neural network architectures and training processes that aim to make the features learned from the network independent or uncorrelated from each other. The main purpose of feature decoupling is to improve the generalization ability and interpretability of the model. A feature decoupled training pipeline for describe-then-detect is designed for weakly supervised local feature learning [17]. Additionally, an introduced line-to-window search strategy enhances descriptor learning by explicitly utilizing camera pose information, and attained state-of-the-art performance across various tasks. The work in [18] reports a multi-scale single image deraining network, called the feature decoupling and reorganization network, which introduces a dilated pyramid split attention module to decouple input features and reorganize extracted features.

The aforementioned work exemplifies practical applications of feature decoupling, which often entails specific modifications to network architecture, such as the incorporation of regularization techniques or the addition of terms to the loss function to promote feature independence. By employing feature decoupling, the model is anticipated to acquire more robust and distinguishable feature representations, thereby enhancing its performance on novel and previously unseen data.

Image semantic information can be broadly categorized into three layers: the visual, object, and conceptual layers. The visual layer, often referred to as the shallow feature layer, encompasses detailed attributes such as color and shape. The object layer, or mid-level feature, usually contains object attributes. The conceptual layer, known as the high-level feature contains abundant semantic information. ResNet-50 [19] is specifically designed for image classification tasks. However, the tasks of CMFD and image classification have distinct focuses on the image, which necessitate the network to possess diverse feature extraction capabilities and representation capabilities. Meanwhile, quantitative analysis is performed experimentally on the pre-trained ResNet-50 to achieve structural risk minimization. It is worth mentioning that the 3461 model is a variant of the ResNet-50. It indicates that the ConvNet Layer of four modules in ResNet-50 is repeated 3, 4, 6, and 1 times, respectively.

Fig. 1 illustrates the comparison of the ResNet-50 deep module (fourth module) possessing varying numbers of convolutional layers, where the 3 4 6 1, 3 4 6 3 and 3 4 6 6 represent the numbers of stacks of residual blocks within the module, respectively. Multiple classes of models with slight variations were trained while keeping the first three modules frozen. The comparative analysis of visualization outcomes derived from the three distinct models suggests that increasing the depth of the network architecture enhances the ability to extract complex details. Nevertheless, based on the pertinent graphical representation, it can be inferred that integrating more complex structures into the model is inappropriate for tackling the CMFD task.

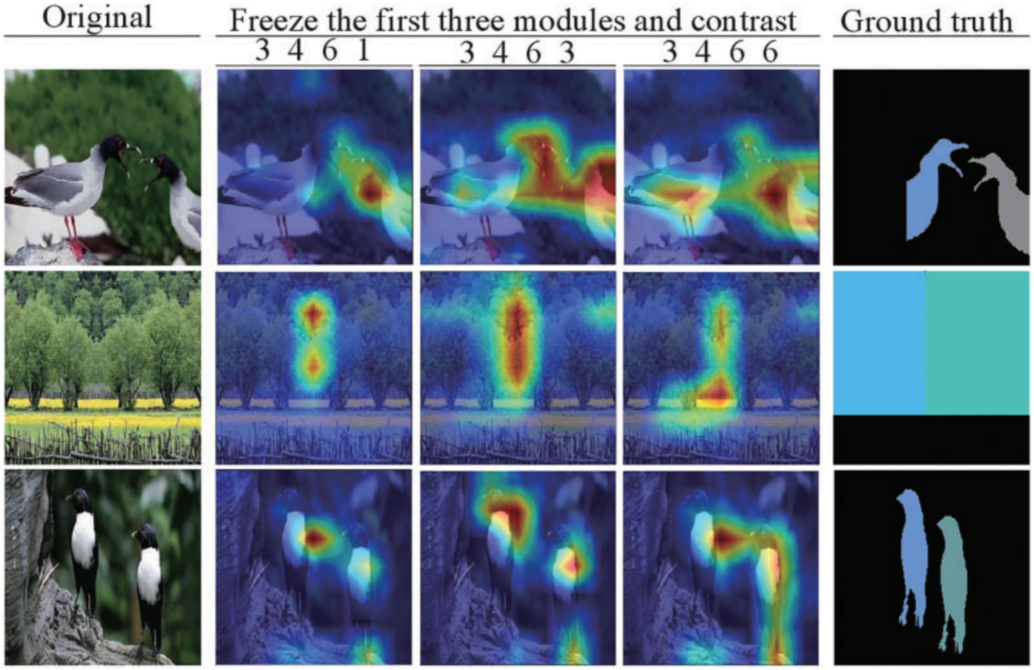

Figure 1: Contrast analysis. The first three module weights are frozen based on a pre-trained ResNet-50. The effect of the model is analyzed while the convolution layer of the fourth module is changed. Source and tampered areas are indicated in blue and gray

To address the aforementioned limitations, this paper introduces a feature decoupling approach within the context of transfer learning, leveraging knowledge migration to enhance model performance. When the source and target domains are similar, the main challenge in feature decoupling is determining which layers of knowledge in the source domain network should be fixed or fine-tuned [20].

In the proposed scheme, the relationship between the CMFD task and the number of deep feature repetition layers in the pre-trained ResNet-50 is explored. ResNet-50 is introduced for several compelling reasons: (1) the incorporation of identity shortcut connections within the residual branching structure (RBS) optimizes the process of backpropagation, thereby rendering ResNet-50, a simple and highly efficient method. (2) ResNet-50 employs a limited number of optimization techniques effectively, thereby minimizing potential interferences. Furthermore, the architecture incorporates two types of RBS. RBS-1 utilizes a step size of 2, which significantly reduces the output size and mitigates the risk of overfitting. On the other hand, RBS-2 serves as the major module in the ResNet network, primarily focused on enhancing the representational capabilities. (3) In addition, in the field of passive forensics, the scarcity of high-quality datasets leads to low model generalization ability and accuracy rates.

The objective of this paper is to streamline the multi-task framework of ResNet-50 to mitigate disaster forgetting and enhance task efficiency. To accomplish this, the multi-task objective is simplified into a single-task objective. Specifically, the focus is on maximizing the utilization of the solution space for the image classification task through pertinent experiments. Our goal is to enhance the model performance on the CMFD task when the image classification task has already reached its optimum. This paper proposes an improved multi-task copy-move forgery detection network by using a multi-feature pyramid module (MFPM) and features decoupling across tasks. The main contributions of this study are as follows:

1. An optimization problem is abstracted to establish a link between image classification and CMFD based on ResNet-50 using feature decoupling, wherein the relationship between the deep structure and task is illustrated during optimization.
2. This paper provides a quantitative demonstration of the similarities between image classification tasks and CMFD. Moreover, it addresses the challenge of limited high-quality data in CMFD through the transfer of frozen weights and retraining of the model.
3. The CMFD field saw the first introduction of the MFPM. It utilizes three matching maps to detect suspicious regions and enhances localization accuracy through the application of Feature Pyramid Networks and an optimized bottom-up pathway.

The rest of this paper is organized as follows. Section 2 describes the stages of the proposed method and briefly explains every step. Section 3 presents databases that are being used for experiments and an experimental setup and results in discussions for the proposed architecture. It offers tables and figures related to results calculated using the proposed architecture. At last, Section 4 provides conclusions for this study.

2 Proposed Method

To address the challenges in dataset building and training due to the difficulty of collecting high-quality annotated collections of tampered images in CMFD, IMTNet is proposed in this paper, the diagram is shown in Fig. 2. It comprises two components: copy-move forgery detection algorithm based on feature decoupling and tampered image localization algorithm named MFPN, which effectively leverages both high and low-dimensional image information.

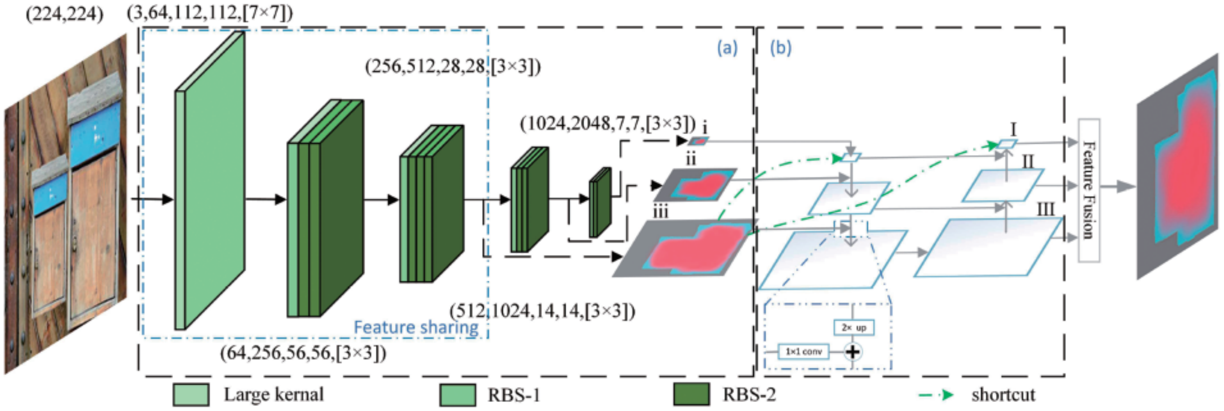

Figure 2: The diagram of IMTNet for copy-move forgery detection and localization. (a) Improved ResNet-50 backbone. (b) MFPN

2.1 Copy-Move Forgery Detection Algorithm Based on Feature Decoupling

ResNet-50 based on transfer learning. Due to the difficulty of collecting datasets in CMFD, the current number of datasets available is limited. Furthermore, the varying shallow weights obtained during different training epochs can potentially disrupt correlation ablation experiments. To tackle the aforementioned challenges, our proposal involves the adoption of a feature decoupling approach and enhancements to the network structure of the pre-trained ResNet-50. The primary objective is to maximize the leverage of the source domain as a feature extraction network for the target domain.

The pre-trained ResNet-50 combined with the feature decoupling approach empowers the IMTNet to acquire highly relevant implicit expression features. Furthermore, feature decoupling is employed as a regularization technique to mitigate dissimilarities between the source and target marginal distributions. This mathematical description can be formulated as

$$D_S = \left\{\left(x_i^S,\ y_i^S\right)\right\}_{i=1}^{n_s}, (x,\ y) \in (X,\ Y) \tag{1}$$

where S indicates the source domain and D_S denotes the sample space in the source domain, $X,\ \ Y$ denote the joint feature space and the corresponding label space.

$$D_T = \left\{\left(x_j^T\right)\right\}_{j=1}^{n_t} \tag{2}$$

in which T denotes target domain and D_T represents the sample space in the target domain. Besides, n_s is the number of samples in the source domain. There are also n_t samples in the target domain. S and T have different probability distributions, transfer learning transfers knowledge from S to T to execute specific tasks on T, the transfer process is shown as

$$\left\{\left(x_j^T\right)\right\}_{j=1}^{n_t} \underset{D_S}{\rightarrow} \left\{\left(x_j^T,\ \hat{y}_j^T\right)\right\}_{j=1}^{n_t} \tag{3}$$

in which $\hat{y}_j^T$ represents the predicted label after knowledge transfer. From this, feature decoupling is realized by transferring the features of classification to CMFD. Therefore, the features of CMFD can be extracted to verify the integrity of an image.

Optimization problem solution for multi-task based on ResNet-50. A ConvNet Layer i can be defined as a function: $Y_i = F_i\left(X_i\right)$, where F_i is the convolution operator, Y_i is the output tensor, X_i is the input tensor. A ConvNet N can be represented by a list of composed layers

$$N = F_k\Theta \ldots \Theta F_2\Theta F_1\left(X_1\right) = \underset{j=1\ldots k}{\Theta} F_j\left(X_1\right) \tag{4}$$

where Θ represents the multiplication operation. There are five stages in ResNet-50, and every layer in the last four stages has the same convolutional type, except for the first layer which performs downsampling. Therefore, ConvNet can be defined as

$$N = \underset{i=1\ldots s}{\Theta} F_i^{L_i}\left(X_{<H_i,W_i,C_i>}\right) \tag{5}$$

where N represents the abstracted whole network processing, i is the module serial number. F_i represents a convolution operator, $F_i^{L_i}$ means that the F_i operation is repeated L_i times in module i. X denotes the input tensor in stage i. $< H_i,\ W_i,\ C_i >$ indicates width, height and number of channels. Our purpose is to maximize the model accuracy for given resource constraints, which can be formulated as an optimization problem.

$$\max_d N = Accuracy\left(N_S\left(d\right)\right) + Accuracy\left(N_T\left(d\right)\right) \tag{6}$$

The multi-task optimization problem described above is transformed into a single optimization problem, where $\overline{N_S\left(d\right)}$ indicates that ResNet-50 has undergone optimization for image classification tasks, and $N_T\left(d\right)$ represents ResNet-50 employed for the CMFD task, d is a scaling factor, $\overline{L_i}$, $\overline{C_i}$are predefined parameters in ResNet-50 and $\left(\overline{H_1},\ \overline{W_1}\right) = (224,\ 224)$, denoted as

$$\begin{aligned}
&\underset{d}{max}\, Accuracy\left(N_T\left(d\right)|\overline{N_S}\left(d\right)\right)\\
&s.t.N\left(d\right) = \underset{i=1\ldots S}{\Theta} F_i^{d.\overline{L_i}}\left(X_{<H_i,W_i,C_i>}\right)\\
&Memory\left(N\right) \leq t\arg et\_memory\\
&FLOPs\left(N\right) \leq t\arg et\_flops
\end{aligned} \tag{7}$$

Initially, we assume that the model is the most optimal image classification solution. According to the steps above, this optimal solution can be adapted to the specifics of image copy-move tampering. In the deep model, the number of iterations between layers is scaled by d. Following feature decoupling and model optimization, the implementation of a multi-task ResNet-50 exhibits significant advantages. It effectively circumvents catastrophic forgetting while also improving the detection accuracy for the CMFD tasks.

2.2 Tampered Image Positioning Algorithm

The MFPN incorporates top-down and bottom-up bidirectional fusion branches, which combine low-level features information and high-level semantic information to improve the accuracy of semantic representation [21]. Therefore, IMTNet leverages the inherent multi-scale and hierarchical characteristics of the network to construct MFPN, which greatly enhances its representational capabilities and localization quality. Moreover, it reduces the cost of building MFPN. The output results of the three matching maps and their combination are shown in Fig. 3a,b show the forgery image and the ground truth mask, and Fig. 3c–f shows the results of the three up-sampled matching maps I, II and III and the final suspicious area location results.

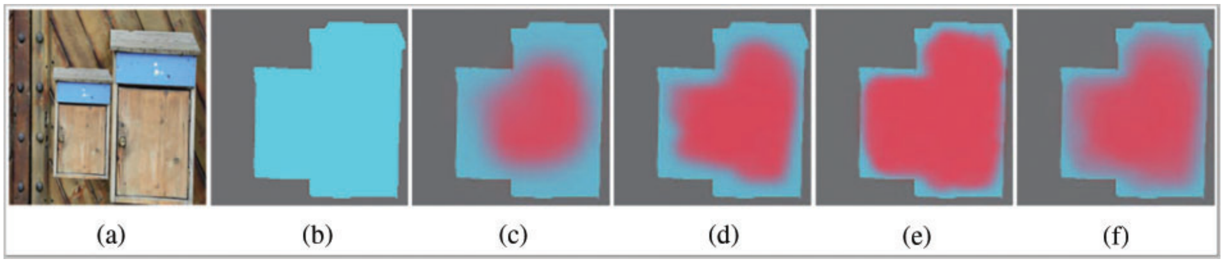

Figure 3: The output results of the three matching maps and their combination

The numerous untrained forgery classes or objects result in difficulty in applying a classic DNNs model to address those data. Therefore, an auxiliary image tampering localization model is proposed to learn the correlations between the rich hierarchical features. In the proposed scheme, P features are extracted for each image, which are the different features of the image after applying different convolution operators to the image. Assuming that sets of the feature point is $P = (P_1,\ P_2,\ \ldots, P_i,\ \ldots,\ P_{N\times N})$, the M-dimensional description operator of P can be expressed as

$$P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,i} & \cdots & p_{1,M} \\ p_{2,1} & \cdots & p_{2,i} & \cdots & p_{2,M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{i,1} & \cdots & p_{i,i} & \cdots & p_{i,M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{N\times N,1} & \cdots & p_{N\times N,i} & \cdots & p_{N\times N,M} \end{bmatrix} \tag{8}$$

where the parameter M represents the depth of the feature, which is the number of channels, $N \times N$ represents the number of candidate pixels or the size of the feature maps in the candidate matrix. In the matching localization algorithm, the feature correlation coefficient between the defined feature points is denoted as

$$\begin{aligned} P_{c_i} &= \left(P_{c_{i,1}}, P_{c_{i,2}}, \ldots, P_{c_{i,i}}, P_{c_{i,j}}, \ldots, P_{c_{i,N\times N}}\right) \\ &= \frac{1}{M}\begin{bmatrix} \left\|p_{i,1} - p_{1,1}\right\|_2 & \cdots & \left\|p_{i,M} - p_{1,M}\right\|_2 \\ \left\|p_{i,1} - p_{2,1}\right\|_2 & \cdots & \left\|p_{i,M} - p_{2,M}\right\|_2 \\ \vdots & \vdots & \vdots \\ \left\|p_{i,1} - p_{N\times N,1}\right\|_2 & \cdots & \left\|p_{i,M} - p_{N\times N,M}\right\|_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{aligned} \tag{9}$$

where the subscript i and j represent the localizations of feature point P_i and P_j in the corresponding matching map. $P_{c_{i,j}}$ is the matched measurement and represents the negative feature correlation coefficient between the feature point P_i and P_j. The definition of the correlation coefficient represents that the closer $P_{c_{i,j}}$ is to 0, the more similar P_i and P_j are.

In the IMTNet, the 2NN matching algorithm [22] is used to reduce the matching errors. Assume $P_{c_{i,j}}$ is the second minor characteristic correlation coefficient and $P_{c_{i,k}}$ is the third minor characteristic correlation coefficient, $P_{c_{i,j}}$ and $P_{c_{i,k}}$ satisfy the following condition:

$$\frac{P_{c_{i,j}}}{P_{c_{i,k}}} \le T_L \tag{10}$$

The relevant features can be filtered according to Eq. (10) while $T_L = 0.65$.

In Eq. (9), set $\alpha = P_{c_{i,j}}$, $P(X)$ is transformed into a binary classification problem by setting it to an activation function that approximates the sigmoid function as

$$P\left(X_{i,j}\right) = \begin{cases} \dfrac{2}{1+e^{\alpha}}, if \dfrac{P_{c_{i,j}}}{P_{c_{i,k}}} < T_L \\ \dfrac{2}{1+\beta \times e^{\alpha}}, others \end{cases} \tag{11}$$

where $P\left(X_{i,j}\right) \in [0, 1]$, $\beta = 2$. It is used to make the unmatched coefficient approach 0.

Finally, the feature matching coefficient $P\left(X_{i,j}\right)$ is filled in the localization of the matching map. In this way, other feature points search for the best matching coefficients and fill in the matching localizations matrix. The hyperparameter k, v, and l are based on the input feature depths of the MFPN blocks I, II and III. $L(X)$ stands for three matching map combinations. The data in blocks I, II and III of MFPM is obtained through processing by the second, third, and fourth modules of the ResNet-50, respectively, denoted as

$$L(X) = kL_{\mathrm{I}}(X) + vL_{\mathrm{II}}(X) + lL_{\mathrm{III}}(X) \tag{12}$$

where $L_{\mathrm{I}}(X), L_{\mathrm{II}}(X)$, and $L_{\mathrm{III}}(X)$ represent the numerical matrices of the first, second, and third modules, respectively, in the MFPM that is processed with the 2NN matching algorithm. The hyperparameter k, v, and l are obtained as

$$\begin{gathered} k = \frac{32}{(32\,(\mathrm{I}) + 48\,(\mathrm{II}) + 64\,(\mathrm{III}))'}, v = \frac{48}{(32+48+64)'}, l = \frac{64}{(32+48+64)'} \\ k + v + l = 1 \end{gathered} \tag{13}$$

where 32, 48, and 64 represent the number of channels in MFPM for locks I, II and III, respectively.

Task relevance can be trained based on the premise that similar tasks share the same model weights, and these tasks can be transformed or low-rank regularized to obtain richer representations. The feature decoupling approach used in this paper aims to capture multiple aspects of task relevance properties, such as sparsely and low-ranking of tasks. This is achieved by decomposing the model partial weights into the sum or product of different convolution operator components that capture information in addition to specific task information that is beneficial to each task. The flexibility of the feature decoupling technique provides a deeper understanding of the nature of multi-task, enabling feature sharing of model weights for both image classification and image tampering detection tasks [23].

3 Experiments

3.1 Experimental Sets

IMTNet is trained on two benchmark datasets: the CASIA2.0 [24] and CoMoFod_small [25] dataset. Meanwhile, the mentioned datasets are mixed as the target domain datasets. The blended dataset contains many manipulated images that have been attacked, which could enhance the “quality” of the dataset and improve its robustness. Moreover, MICC-F2000 [5], MICC-F600 [26], COVERAGE [27], and DEFACTO [28] are used to conduct generalization tests. Original and tampered images from the Ardizzone [29] dataset and the MICC-F2000 dataset are used. In addition, there are 140 images in the dataset for the attack resistance experiments, where 35 identical images come from the Ardizzone dataset that have undergone different tampering attacks, and in order to enlarge the size of the dataset, 35 images in the MICC-F2000 dataset are extracted that have also undergone different tampering attacks, totaling 70 tampered images. Thus, the interference due to different images is reduced and the model is realized for the anti-attack experiments.

Anchor points are used in the experiment to divide the semantic information of images, such as the visual layer, object layer, and concept layer. These points are chosen because each downsampling greatly enhances the representation of semantic information. Additionally, the shallow features of the images are generic, which is the reason that the shallow weights of the model are frozen.

3.2 Ablation Experiments

The ablation study is divided into two phases. The first phase assessed the similarity between the source and target domains to identify the appropriate number of layers to freeze. The second phase explored the correlation between the deep architecture of ResNet-50 and its performance on the CMFD task.

Similarity between the image classification domain and CMFD domain. The deep architecture of ResNet-50 is analyzed at various scales. An experimental comparison is conducted to evaluate the representational capabilities of ResNet-50 when trained directly or with specific modules frozen. Table 1 and Fig. 4 present the experimental results from alternative perspectives, respectively. The data in Table 1 are the value of F_1, accuracy (ACC), Precision (P) and Recall (R), which displays the average values that demonstrate the characterization ability of the models when different modules of ResNet-50 are frozen. The experiment images were obtained after data cleaning and mixing the MICC-2000 and MICC-600 datasets, the total number is 2600, and the valid data are 2560 after cleaning.

Table 1: Image level ablation experiments on ResNet-50 using MICC-F2000 and MICC-F600 datasets

Inter-layer numbers	Only training				Freeze all layers				Freeze 1 layer				Freeze 1, 2 layer				Freeze 1, 2, 3 layer			
	F_1 (%)	ACC (%)	P (%)	R (%)	F_1 (%)	ACC (%)	P (%)	R (%)	F_1 (%)	ACC (%)	P (%)	R (%)	F_1 (%)	ACC (%)	P (%)	R (%)	F_1 (%)	ACC (%)	P (%)	R (%)
3, 4, 2, 1	65.03	57.49	60.82	72.83	74.10	60.80	79.23	69.58	75.61	62.59	81.91	70.18	73.33	61.97	73.88	72.78	**76.35**	**62.59**	**85.34**	**71.08**
3, 4, 4, 2	62.69	53.79	56.82	71.18	74.29	61.52	78.58	70.43	75.60	62.55	81.99	70.13	71.67	60.20	71.18	72.18	**77.80**	**65.25**	**86.09**	**70.98**
3, 4, 6, 3	58.35	50.47	49.02	72.03	72.78	59.90	75.73	70.03	73.39	61.32	75.33	71.53	71.49	60.37	70.18	72.83	**75.60**	**63.20**	**80.54**	**71.23**

In Fig. 4, the inter-layer relationship refers to freezing the pre-trained parameters of different modules within the same model, obtained after training on the image classification task. And then the weight parameters of the unfrozen modules are initialized and trained in the image copy-move tampering detection task, while the generalization ability of the model in image copy-move tampering detection is later measured to achieve the comparison of results in Fig. 4.

Fig. 4 shows that the model trained directly is more stochastic and less stable compared to the pre-trained ResNet-50. It can be seen that the F_1, ACC and P values are higher while the 1, 2, 3 layers are frozen and the inter-layers are repeated 3, 4, 4, 2 times.

The experimental results in Fig. 4 indicate that the proposed IMTNet model achieves better performance in the CMFD task by applying the feature decoupling method, while freezing the first three module weights of ResNet-50.

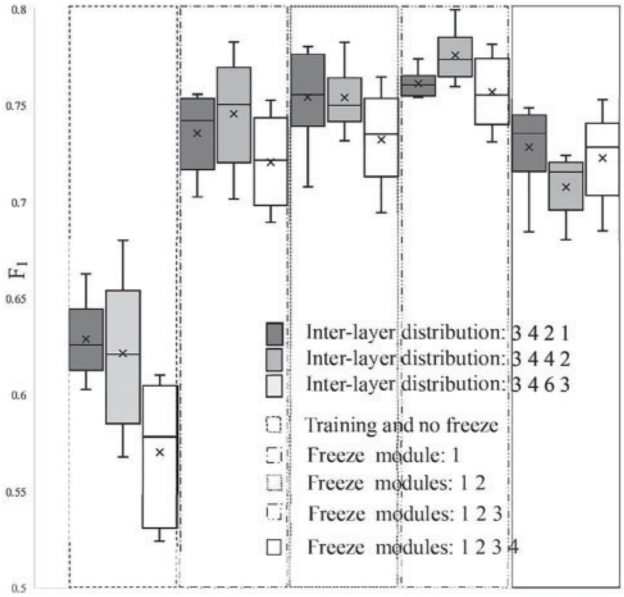

Figure 4: Interlayer relationships

The relationship between the deep structure of ResNet-50 and CMFD. In Fig. 5, the experimental result of accuracy is weighted and summed in proportions. The experimental datasets are the mixture of generalized experimental data based on MICC-F2000 and MICC-F600 datasets. The comparison experiment is repeated 35 times for each group that is a series of values derived from the same model after 35 experiments under the same conditions. Removing the two highest and the three worst results, the first three modules are frozen to avoid disruptions of the parameter change. The relationship between model depth and the image tampering detection task is revealed by the change in the probability distribution of *LOSS* and *ACC*, conditioned on the number of repetitions of the residual block in the model's last layer.

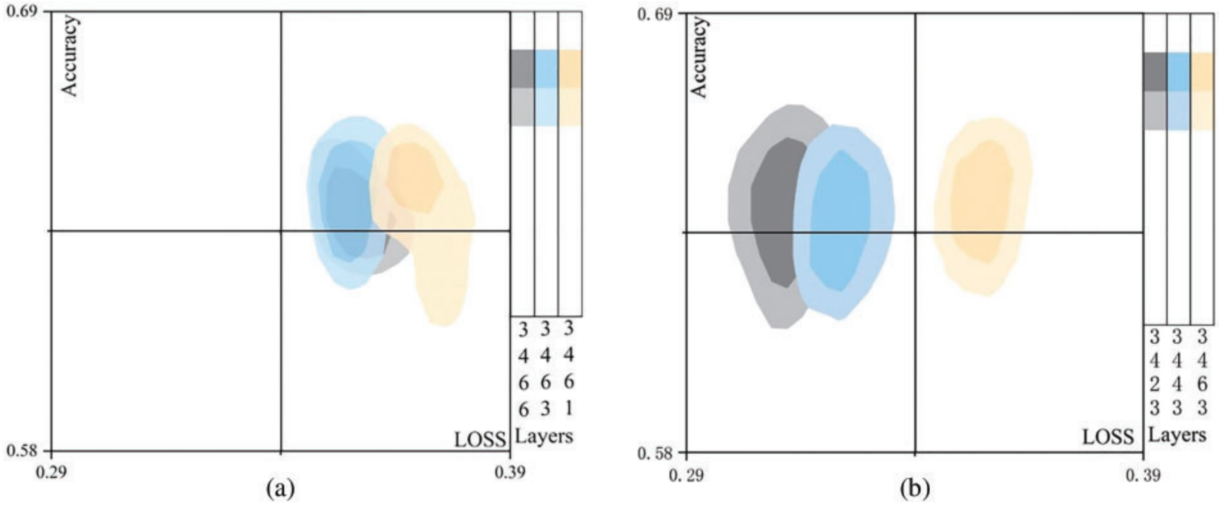

Figure 5: (Continued)

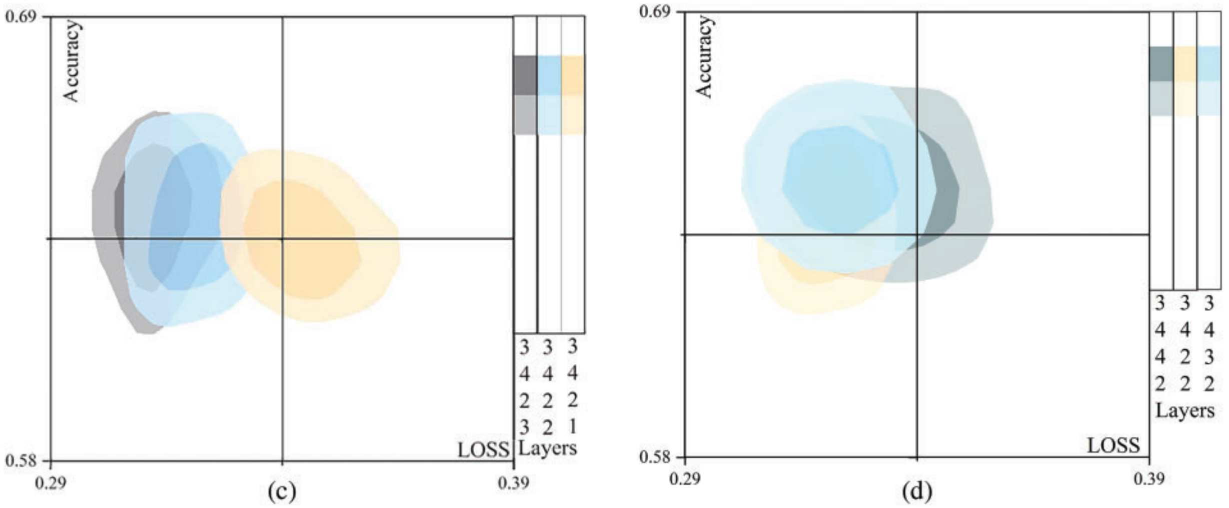

Figure 5: Ablation experiment of model deep layers. (a) Last module comparison. (b) Penultimate module comparison. (c) Refined comparison. (d) Final detail comparison

Fig. 5a characterizes the relationship between the model generalization ability and the model loss function. It can be seen from the distribution of model layers that model structure 3,4,6,6 does not perform as well as expected while consuming more hardware resources. In addition, the generalizability of the model is also reduced. The problem is mainly due to the model using concrete concepts to represent targets rather than abstract regions in tampered images. Furthermore, model structures 3,4,6,1 have insufficient characterization ability and the upper bound of the generalization error and the loss function error is large, which highlights the weakness of the model characterization ability and the instability of the characterization.

Fig. 5b illustrates the training of the last layer while keeping the first three modules frozen, and it also involves reducing the number of convolutional layers in the third module. Experimental evidence reveals that the abundance of highly specific semantic features extracted from the middle and deep layers hinders the detection of tampered images.

As demonstrated in Fig. 5c, the comparison of the generalization performance of the model structures 3,4,2,3 and 3,4,2,2 reveals that they are comparable. Fig. 5a,c shows that the deep parameter reduction of the model has little effect on the model's ability in image tampering detection. This leads to the conclusion that lightweighting the model without compromising its representational capabilities can significantly reduce operational resource consumption and enhance computing speed. At present, the preference is given to the model with larger loss as it is believed to possess better generalization ability. Additionally, this model is more compact and consumes fewer resources.

Based on the results depicted in Fig. 5d, the stable and well-generalized model structure 3432 has been selected as the optimal structure. By modifying the parameters of the last two layers of the model, the network can better characterize the problem of "whether the image has been tampered with" with an abstract entity description.

An exploration of the relevant properties of IMTNet. The mean values from multiple experiments by using MICC-F2000 and MICC-F600 datasets are presented in Table 2. The pre-training datasets we used are ImageNet-1K and ImageNet-21K. The ImageNet dataset is indeed a very common and important source of pre-trained models in the field of transfer learning. The ImageNet-1K is the most commonly used subset of the ImageNet dataset, containing 1000 classes and approximately 1.3 million images. The ImageNet-21K is the full ImageNet dataset, containing 21,841 classes and approximately 14 million images.

Table 2: Image level generation experiments on IMTNet (1), (2)

Image level generation experiments on IMTNet (1)												
Pretained dataset	Fine-tune											
	4 modules				3, 4 modules				2, 3, 4 modules			
	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)
ImageNet-1K	75.11	62.26	77.21	73.25	78.28	65.29	82.64	74.52	77.56	64.92	82.87	73.22
ImageNet-21K	76.56	64.13	77.53	74.76	78.60	66.01	82.91	74.83	77.16	63.97	81.94	73.89
Image level generation experiments on IMTNet (2)												
Pretained dataset	Fine-tune after feature decoupling											
	Only transfer				2, 3 modules				3 modules			
	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)	F_1 (%)	*ACC* (%)	*P* (%)	*R* (%)
ImageNet-1K	78.28	65.29	85.39	73.23	77.49	64.89	84.96	71.93	**78.39**	**66.93**	**84.79**	**72.88**
ImageNet-21K	78.60	66.01	85.33	74.96	78.52	65.74	85.44	73.58	**79.67**	**68.10**	**85.39**	**75.53**

From the perspective of the generalization dataset, feature decoupling and fine-tuning are compared in the CMFD task quantitatively. The experimental data is utilized to measure the similarity between the image classification domain and the CMFD domain. Then, the comparison of the number of frozen layers *vs.* the number of fine-tuned layers is made. The experimental results have led to the determination that the most optimal approach involves fine-tuning the third module subsequent to the feature decoupling process.

Fig. 6 depicts the correlation between the number of training parameters in a model and its representation capabilities. The trainable parameters are derived by subtracting the model's frozen parameters from the model's full parameters, while the model detection accuracy is derived from the model's performance in the generation experiments. This reveals that IMTNet exhibits outstanding generalization performance by employing a trace of parameters.

Additionally, Fig. 7 presents the comparison between feature decoupling and fine-tuning across five distinct dataset sizes: 166, 766, 2046, 4171, and 5337. These datasets are part of the training dataset. After subtracting 5337 images from the training dataset, the test dataset is uniformly the remaining 15,328 images. The highest value is taken in 25 runs. The results of the experiment indicate that, for datasets with less than 1000 samples, fine-tuning is generally a more effective approach compared to feature decoupling followed by fine-tuning. Fig. 7a,b portrays the relationship from different perspectives.

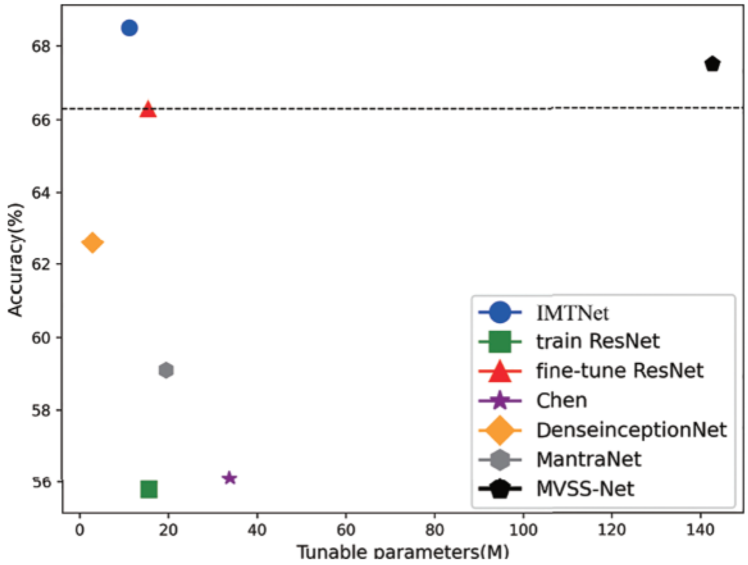

Figure 6: Generalization accuracy-training parameter

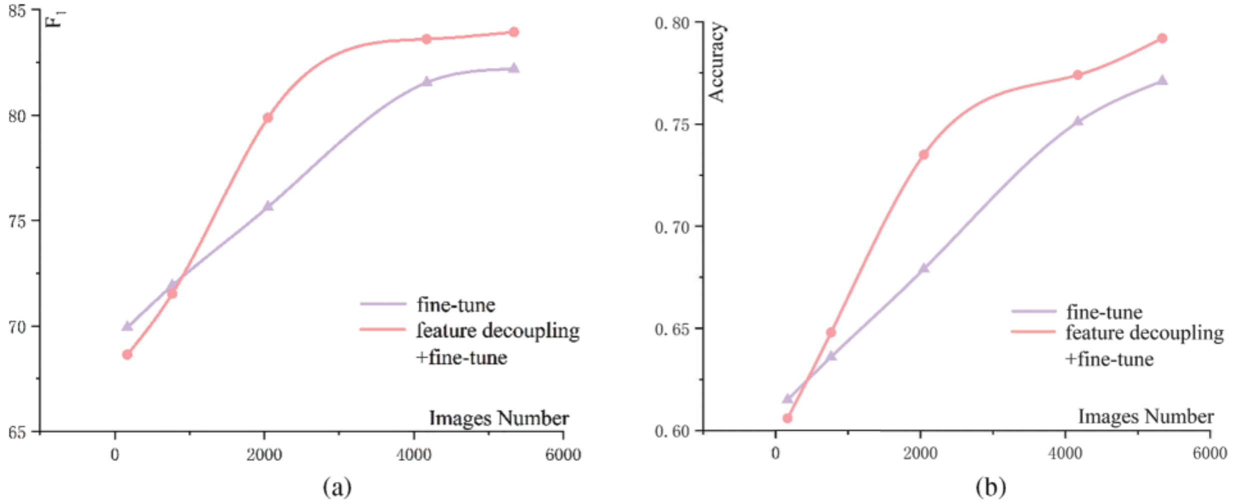

Figure 7: Comparative analysis of fine-tuning and feature decoupling. (a) F_1-images number. (b) Accuracy-images number

The following observations, derived from the comparison of the results obtained from these experiments, have been listed below: (1) The findings from the quantitative analysis experiments indicate that the CMFD task necessitates shallower layers for the ResNet-50 model as compared to the image classification task. (2) Fine-tuning achieves superior performance compared to feature decoupling when the dataset is limited in size. (3) For the ResNet-50 model, the CMFD task achieves better performance when reducing the number of convolutional layers in the third module, compared to reducing the number of convolutional layers in the last module.

3.3 Anti-Attack Experiments

The anti-attack experiments of various algorithms are described below to evaluate the robustness of the pre-trained network. 1: Rotation attack. 2: Gaussian noise attack. 3: JPEG image compression attack. 4: Blurring attack. 5: Scaling attack.

As shown in Fig. 8, the value is used as an indicator to evaluate the detection capabilities of various methods in different tampering environments. The curve depicted is obtained through the quadratic interpolation of the relevant data points. In which, Fig. 8a–e show the result comparisons under various attacks. In anti-attack experiments, the following observations are presented: (1) under a variety of attacks, IMTNet exhibits excellent robustness compared to other methods. (2) IMTNet possesses the property conferred by feature decoupling as seen in the comparison of ResNet-50.

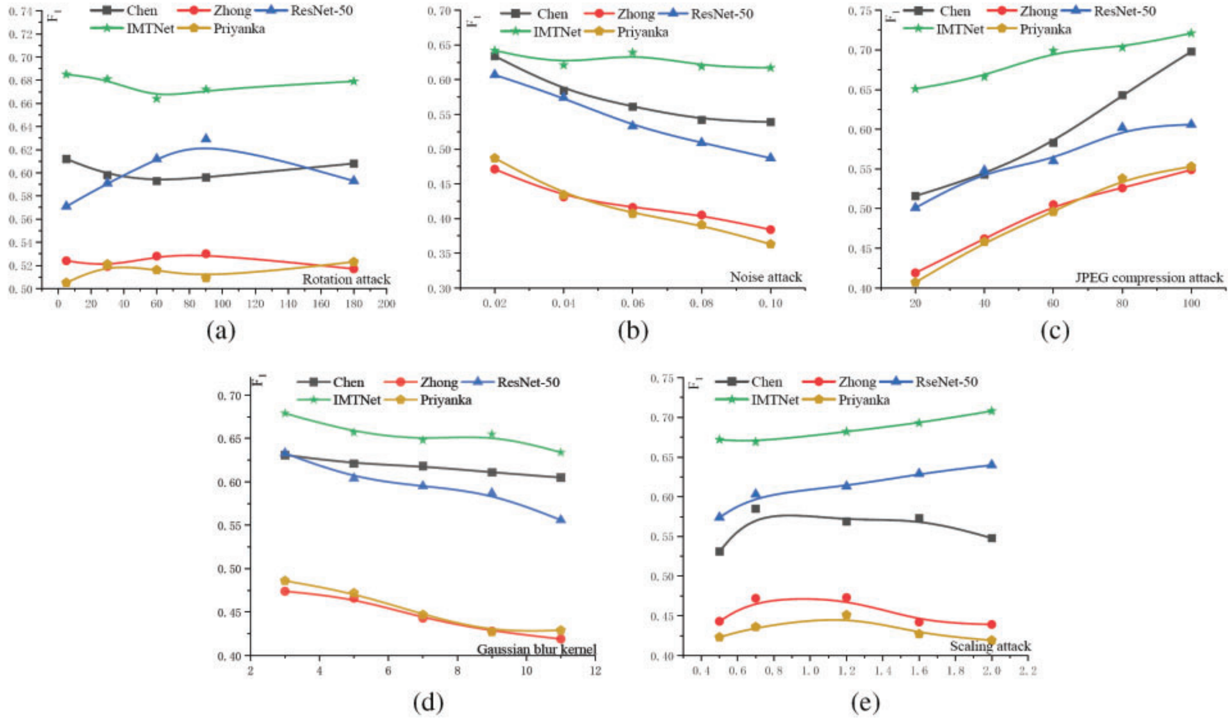

Figure 8: Resistance to attack performance. (a) Rotation attack. (b) Noise attack. (c) JPEG compression attack. (d) Blur attack. (e) Scaling attack

To test the performance of IMTNet, MICC-F2000, MICC-F600 and COVERAGE datasets are used to conduct generalization tests. Table 3 shows the value of F_1, accuracy (ACC), Precision (P) and Recall (R).

The algorithm in [19] introduces a classification task based on ResNet-50, the first row in Table 3 is the result of CMFD task by training directly on ResNet-50, it can be seen that applying the classification task model directly to the CMFD task does not work well. However, the accuracy and F_1 values of the proposed IMTNet combining feature decoupling and transfer learning are higher than other algorithms in different datasets. Therefore, it is believed that IMTNet exhibits better generalization than other proprietary methods. The outcomes presented in Fig. 9 showcase the robust localization proficiency of IMTNet in generalization experiments. The 1st and 2nd rows show the forgery images and corresponding detection results.

Table 3: Image generation experiments

Pretained dataset	MICC-F2000				MICC-F600				COVERAGE				DEFACTO			
	F_1(%)	ACC(%)	P(%)	R(%)	F_1(%)	ACC(%)	P(%)	R(%)	F_1(%)	ACC(%)	P(%)	R(%)	F_1(%)	ACC(%)	P(%)	R(%)
ResNet-50 [19]	54.29	46.48	45.42	67.48	59.63	51.76	48.57	77.18	51.91	46.28	57.97	46.97	52.11	46.53	48.32	56.57
Zhong [6]	48.47	46.33	50.52	48.53	46.67	47.55	47.68	41.83	41.93	46.36	49.78	55.05	47.98	50.44	52.57	52.03
Chen [10]	62.16	54.49	53.43	74.33	67.91	60.15	57.47	82.94	56.35	52.63	61.18	52.22	54.92	50.17	50.37	60.38
Priyanka [3]	47.34	46.67	48.59	46.78	50.09	49.56	47.46	48.29	49.01	49.92	50.03	51.49	49.98	41.29	45.69	46.46
IMTNet	**78.66**	**66.47**	**86.24**	**72.93**	**81.12**	**70.40**	**86.74**	**76.18**	**65.74**	**58.83**	**75.03**	**59.27**	**63.54**	**57.94**	**60.78**	**66.53**

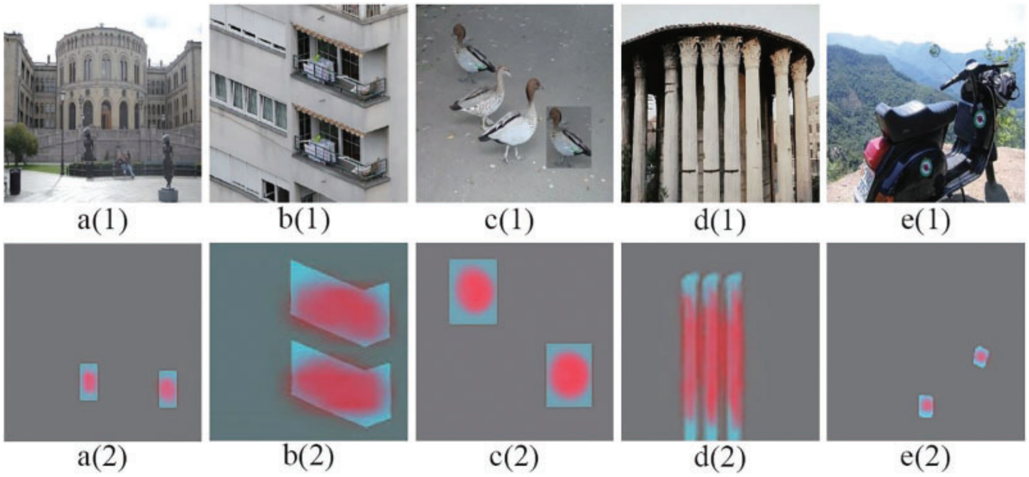

Figure 9: The CMFD results of IMTNet. a(1)~e(1) The forgery images. a(2)~e(2) Corresponding detection results

4 Conclusions

In the proposed scheme, the relationship between ResNet-50 and CMFD is thoroughly demonstrated through quantitative experiments. IMTNet is proposed by leveraging the image classification feature domains and reducing the deep architecture of ResNet-50. Firstly, the relationship between CMFD and deep network architecture is formulated as an optimization problem. In the CMFD task, IMTNet exhibits outstanding performance compared to ResNet-50 and other CMFD algorithms by reducing the deep structure of ResNet-50 and utilizing the feature decoupling method. Secondly, experiments demonstrate that the IMTNet reduced the number of ResNet-50 parameters while enhancing the generalization capability of the model. Furthermore, the integration of MFPN improved the capability of the proposed method in detecting suspicious areas in tampered images.

Acknowledgement: None.

Funding Statement: This work was supported and founded by the Guizhou Provincial Science and Technology Project under the Grant No. QKH-Basic-ZK[2021]YB311, the Youth Science and Technology Talent Growth Project of Guizhou Provincial Education Department under Grant No. QJH-KY-ZK[2021]132, the Guizhou Provincial Science and Technology Project under the Grant No. QKH-Basic-ZK[2021]YB319, the National Natural Science Foundation of China (NSFC) under

Grant 61902085, and the Key Laboratory Program of Blockchain and Fintech of Department of Education of Guizhou Province (2023-014).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Huan Wang, Zhongyuan Jiang; data collection: Huan Wang, Qing Qian and Yong Long; analysis and interpretation of results: Huan Wang, Zhongyuan Jiang and Qing Qian; draft manuscript preparation: Huan Wang, Zhongyuan Jiang and Yong Long. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, jiangzydj@163.com, upon reasonable request.

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] J. Fridrich, B. D. Soukal, and A. J. Lukas, "Detection of copy-move forgery in digital images," in *Proc. Digit. Forensic Res. Workshop*, Cleveland, OH, USA, 2003, pp. 19–23.

[2] Y. P. Huang, L. Wei, S. Wei, and D. Long, "Improved DCT-based detection of copy-move forgery in images," *Forensic Sci. Int.*, vol. 206, no. 3, pp. 178–184, 2011. doi: 10.1016/j.forsciint.2010.08.001.

[3] Priyanka, G. Singh, K. Singh, "An improved block based copy-move forgery detection technique," *Multimed. Tools Appl.*, vol. 79, no. 19, pp. 13011–13035, 2020. doi: 10.1007/s11042-019-08354-x.

[4] A. Akram, J. Rashid, A. Jaffar, F. Hajjej, W. Iqbal and N. Sarwar, "Weber law based approach for multi-class image forgery detection," *Comput. Mater. Contin.*, vol. 78, pp. 145–166, 2024. doi: 10.32604/cmc.2023.041074.

[5] I. Amerini, L. Ballan, R. Caldelli, A. BDel Bimbo, and G. Serra, "A SIFT-based forensic method for copy-move attack detection and transformation recovery," *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 1099–1110, 2011. doi: 10.1109/TIFS.2011.2129512.

[6] Y. Gan, J. Zhong, and C. Vong, "A novel copy-move forgery detection algorithm via feature label matching and hierarchical segmentation filtering," *Inf. Process. Manage.*, vol. 59, no. 1, pp. 167–178, 2022. doi: 10.1016/j.ipm.2021.102783.

[7] J. L. Zhong, J. X. Yang, H. Zeng, and Y. Q. Zhao, "A novel image copy-move forgery detection algorithm using the characteristics of local descriptors," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 36, no. 15, pp. 1–16, 2022. doi: 10.1142/S0218001422540192.

[8] Y. Aydın, "Automated identification of copy-move forgery using Hessian and patch feature extraction techniques," *J. Forensic Sci.*, vol. 69, no. 1, pp. 131–138, 2024. doi: 10.1111/1556-4029.15415.

[9] W. Yue, W. Abd-Almageed, and P. Natarajan, "Busternet: Detecting copy-move image forgery with source/target localization," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 170–186. doi: 10.1007/978-3-030-01231-1_11.

[10] B. Chen, W. Tan, G. Coatrieux, Y. Zheng, and A. Shafique, "A serial image copy-move forgery localization *scheme* with source/target distinguishment," *IEEE Trans. Multimedia*, vol. 23, pp. 3506–3517, 2020. doi: 10.1109/TMM.2020.3026868.

[11] 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.

[12] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," 2020, *arXiv:2009.00726*.

[13] F. Z. Mehrjardi, A. M. Latif, and M. S. Zarchi, “Copy-move forgery detection and localization using deep learning,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 37, no. 9, pp. 1–21, 2023. doi: 10.1142/S0218001423520122.

[14] K. K. Thyagharajan and G. Nirmala, “Image manipulation detection through laterally linked pixels and kernel algorithms,” *Comput. Syst. Sci. Eng.*, vol. 41, no. 1, pp. 357–371, 2022. doi: 10.32604/csse.2022.020258.

[15] F. Z. Zhuang *et al.*, “A comprehensive survey on transfer learning,” *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020. doi: 10.1109/JPROC.2020.3004555.

[16] S. Dong, P. Wang, and K. Abbas, “A survey on deep learning and its applications,” *Comput. Sci. Rev.*, vol. 40, no. 5, 2021, Art. no. 100379. doi: 10.1016/j.cosrev.2021.100379.

[17] K. Li, L. Wang, L. Liu, Q. Ran, K. Xu, and Y. Guo, “Decoupling makes weakly supervised local feature better,” 2022, *arXiv:2201.02861*.

[18] K. Li, J. Huang, H. Ren, W. Ran, and H. Lu, “Feature decoupling and reorganization network for single image deraining,” *Multimed. Syst.*, vol. 30, 2024, Art. no. 154. doi: 10.1007/s00530-024-01348-2.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conf. on Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[20] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “PANet: Few-shot image semantic segmentation with prototype alignment,” in *Proc. IEEE/CVF Conf. on Comput. Vis. Pattern Recognit.*, COEX Convention Center, Seoul, Republic of Korea, 2019, pp. 9197–9206. doi: 10.1109/ICCV.2019.00929.

[21] K. Zhou, Y. X. Yang, Y. Qiao, and T. Xiang, “Domain adaptive ensemble learning,” *IEEE Trans. Image Process.*, vol. 30, pp. 8008–8018, 2021. doi: 10.1109/TIP.2021.3112012.

[22] J. S. Beis and G. D. Lowe, “Shape indexing using approximate nearest-neighbor search in high-dimensional spaces,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Juan, PR, USA, 1997, pp. 1000–1006. doi: 10.1109/CVPR.1997.609451.

[23] J. Yu *et al.*, “Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model Eras,” 2024, *arXiv:2404.18961*.

[24] J. Dong, W. Wang, and T. N. Tan, “CASIA image tampering detection evaluation database,” in *Proc. IEEE China Summit & Int. Conf. on Signal and Inf. Process.*, Beijing, China, 2013. doi: 10.1109/ChinaSIP.2013.6625374.

[25] D. Tralic, I. Zupancic, S. Grgic, and M. Grgic, “CoMoFoD-new database for copy-move forgery detection,” in *Proc. Int. Symp. on Electron. in Mar. (ELMAR)*, Zadar, Croatia, 2013, pp. 49–54.

[26] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, L. Del Tongo and G. Serra, “Copy-move forgery detection and localization by means of robust clustering with J-Linkage,” *Signal Process.: Image Commun.*, vol. 28, no. 6, pp. 659–669, 2013. doi: 10.1016/j.image.2013.03.006.

[27] B. Wen, Z. Ye, R. Subramanian, T. T. Ng, X. Shen and S. Winkler, “COVERAGE—A novel database for copy-move forgery detection,” in *Proc. Int. Conf. on Inf. Photonics(ICIP)*, Phoenix, AZ, USA, 2016. doi: 10.1109/ICIP.2016.7532339.

[28] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. Dugelay and M. Pic, “DEFACTO: Image and face manipulation dataset,” in *Proc. 27Th Eur. Sig. Process. Conf. (EUSIPCO)*, Coruna, Spain, 2019, pp. 1–5. doi: 10.23919/EUSIPCO.2019.8903181.

[29] E. Ardizzone, A. Bruno, and G. Mazzola, “Copy-move forgery detection by matching triangles of keypoints,” *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 10, pp. 2084– 2094, 2021. doi: 10.1109/TIFS.2015.2445742.