Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.055739

ARTICLE

# Image Copy-Move Forgery Detection and Localization Method Based on Sequence-to-Sequence Transformer Structure

**Gang Hao, Peng Liang*, Ziyuan Li, Huimin Zhao and Hong Zhang**

School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, 510630, China

*Corresponding Author: Peng Liang. Email: liangpeng@gpnu.edu.cn

Received: 29 April 2024; Accepted: 23 December 2024; Published: 06 March 2025

**ABSTRACT:** In recent years, the detection of image copy-move forgery (CMFD) has become a critical challenge in verifying the authenticity of digital images, particularly as image manipulation techniques evolve rapidly. While deep convolutional neural networks (DCNNs) have been widely employed for CMFD tasks, they are often hindered by a notable limitation: the progressive reduction in spatial resolution during the encoding process, which leads to the loss of critical image details. These details are essential for the accurate detection and localization of image copy-move forgery. To overcome the limitations of existing methods, this paper proposes a Transformer-based approach for CMFD and localization as an alternative to conventional DCNN-based techniques. The proposed method employs a Transformer structure as an encoder to process images in a sequence-to-sequence manner, substituting the feature correlation calculations of previous methods with self-attention computations. This allows the model to capture long-range dependencies and contextual nuances within the image, preserving finer details that are typically lost in DCNN-based approaches. Moreover, an appropriate decoder is utilized to ensure precise reconstruction of image features, thereby enhancing both the detection accuracy and localization precision. Experimental results demonstrate that the proposed model achieves superior performance on benchmark datasets, such as USCISI, for image copy-move forgery detection. These results show the potential of Transformer architectures in advancing the field of image forgery detection and offer promising directions for future research.

**KEYWORDS:** CMFD; self-attention; transformer; deep convolutional neural networks

## 1 Introduction

The pervasive presence of graphic editing software in current society has allowed for the effortless and inexpensive creation of many realistic counterfeit images. Maliciously manipulated images can have significant adverse consequences, including their use in fraudulent activities, the dissemination of misinformation, the fabrication of evidence, and the misguiding of public opinion. Consequently, it is imperative to develop effective image tampering detection methodologies to assist individuals in determining whether an image has been altered and identifying the specific areas of manipulation.

Among image tampering techniques, copy-move forgery [1] involves duplicating a region within an image and relocating it elsewhere within the same image. In recent years, deep learning methods have emerged as a prominent focus in copy-move forgery detection (CMFD) research due to their advantages of reduced hyper parameters and increased versatility. However, applying these models to CMFD tasks presents several challenges. These include significant loss of detailed information during the feature encoding process through convolution, particularly for small targets. Additionally, the limited size of the convolution

kernel restricts the receptive field of convolutional neural network (CNN) models, impeding their ability to effectively capture long-range dependencies. These challenges, among others, continue to pose obstacles in the field.

Recent research has demonstrated that the Transformer [2] model, widely utilized in natural language processing, can be effectively applied to various downstream tasks in computer vision. This development offers an alternative approach to the CMFD task beyond convolutional methods. Specifically, it maintains spatial resolution during image feature encoding, performs direct sequence-to-sequence self-attention calculations, and identifies task-relevant information from the global image context from the outset.

This paper introduces an innovative approach for detecting and localizing image copy-move tampering. The encoding process utilizes a conventional transformer encoder. By implementing a one-to-one feature matching module, it effectively distinguishes tampered features from similar background features, and employs a multi-scale contextual decoder to achieve more precise detection of image copy-move tampering. The primary contributions of this research are as follows:

1) A one-to-one feature matching module has been developed to mitigate the impact of similar background features on forged elements, while maintaining resilience to variations in image scale.
2) The multi-scale context decoder consolidates tampering feature information at various levels of granularity through straightforward element-wise addition. This approach incorporates a broader range of potential tampering details, thereby enhancing the precision of tampered region detection.

## 2 Related Work

Contemporary approaches for detecting image copy-move tampering predominantly fall into two categories: feature-based methods and deep learning-based methods [3]. Feature extraction-based methods involve extracting characteristics that represent an image's content or structure, and then utilizing the similarities or differences among these features to identify tampered areas. These methods can be further classified into block-based and key-point-based approaches [4]. The block-based method divides the image into overlapping or non-overlapping small blocks for feature extraction and comparison [5], employing techniques such as Zernike [6], DCT [7], and PCA [8]. The key-point-based method involves detecting salient or invariant key-points from an image, followed by feature extraction and matching for each key-point [9], using algorithms like SIFT [10], SURF [11], and FREAK [12]. While feature extraction-based methods generally offer faster detection speeds, they have limitations including sensitivity to feature selection and parameter settings, as well as limited adaptability to complex scenes and multiple tampering instances [13].

In the domain of image copy-paste tampering detection, contemporary research primarily employs deep learning techniques. Wu et al. [14] introduced BusterNet, an end-to-end deep neural network comprising two branches: Mani-Det and Simi-Det. The Mani-Det branch identifies tampered regions, while Simi-Det detects similarities between source and target areas, thus localizing them. However, Buster-Net's Simi-Det branch extracts only low-resolution feature information through convolutional networks, and both branches must accurately locate the target area for correct source and target classification. To mitigate these limitations, Chen et al. [15] advanced BusterNet by fusing CMSDNet with STRDNet. They utilized a single-branch dual-network architecture for detecting similarities in source/target regions and incorporated mechanisms such as spatial pyramid pooling, spatial attention and channel-wise attention to improve the model's similarity detection capabilities. Hu et al. [16] developed SPAN, which incorporates a spatial pyramid attention framework to analyze image regions across various resolutions using localized self-attention mechanisms. While SPAN leverages local correlations, it fails to comprehensively harness spatial correlations, thereby restricting the model's generalizability. DOA-GAN [17] employs a two-stage spatial attention mechanism to enhance the

capture of location information and discriminative feature information of copied and moved objects, refining localization results through a generative adversarial network. However, its detection effectiveness for small tampered regions remains suboptimal. Dong et al. [18] proposed MVSS-Net, comprising an edge supervision branch and a noise-sensitive branch. These branches aim to capture subtle differences at the boundaries between tampered and untampered regions, as well as noise inconsistencies. By extracting semantic-agnostic features through multi-view feature learning, MVSS-Net obtains more generalized features, facilitating tamper detection while reducing false positives for authentic images.

Presently, the majority of deep learning methods rely on Deep Convolutional Neural Network (DCNN) models. While convolutional down-sampling serves to significantly lower the computational demands of feature correlation processes, it also results in the loss of certain detailed information. This loss can lead to suboptimal detection performance, particularly for small target regions.

There remains substantial potential for enhancing robustness against diverse attacks and augmenting the proficiency in discriminating source regions from target regions within image forensics. Motivated by the advancements in Transformer architectures and self-attention techniques within Natural Language Processing (NLP), researchers have increasingly applied these concepts to Computer Vision (CV). This approach aims to address the limitations of traditional passive forensic techniques and DCNN models, potentially offering more effective solutions for detecting and localizing image manipulations.

The Vision Transformer (ViT) [19] pioneered the application of the standard Transformer model to image classification in computer vision, demonstrating that despite the self-attention mechanism's lack of inductive biases inherent to DCNN, it can match or surpass DCNN performance in image classification tasks with large-scale pre-training. Subsequent research has extended the application of self-attention and Transformer models to other tasks, including object detection (DETR [20] and Deformable DETR [21]) and image segmentation (SETR [22] and TransUNet [23]), while also enhancing image classification performance (DeiT [24] and Swin Transformer [25]). Wang et al. introduced ObjectFormer [26], successfully incorporating Transformer into image tampering detection. However, this approach merely concatenates CNN and Transformer sequentially without effectively integrating their strengths. Additionally, its use of frequency domain features provides minimal benefit for CMFD. Addressing these limitations, our model proposes that a standard Transformer module with a core self-attention mechanism can efficiently identify regions within an image that have identical forms but differ in edge artifacts. This module, when paired with an appropriate decoder, can be directly applied to feature encoding for CMFD tasks.

## 3 Methodology

To elucidate the model design in this paper, we first examine the CMFD methods within the DCNN framework. This framework bears resemblance to the encoder-decoder structure of the Fully Convolutional Network (FCN) [27] for semantic segmentation, comprising three primary modules: a feature extractor based on CNN, a module for feature matching, and a decoder. These components are employed to extract features, compute feature similarity, and generate tampering masks, respectively. To enhance detection efficacy, the model may incorporate additional post-processing modules or advanced designs, such as edge detection, feature pyramids, and multiple feature fusion techniques.

The model proposed in this paper adheres to the encoder-decoder framework, with a notable modification. To preserve the original image resolution during feature extraction, a Transformer encoder is employed. This design, centered on multi-head self-attention (MSA), effectively fulfills both feature extraction and feature matching requirements. Fig. 1 presents the complete architecture of the model.

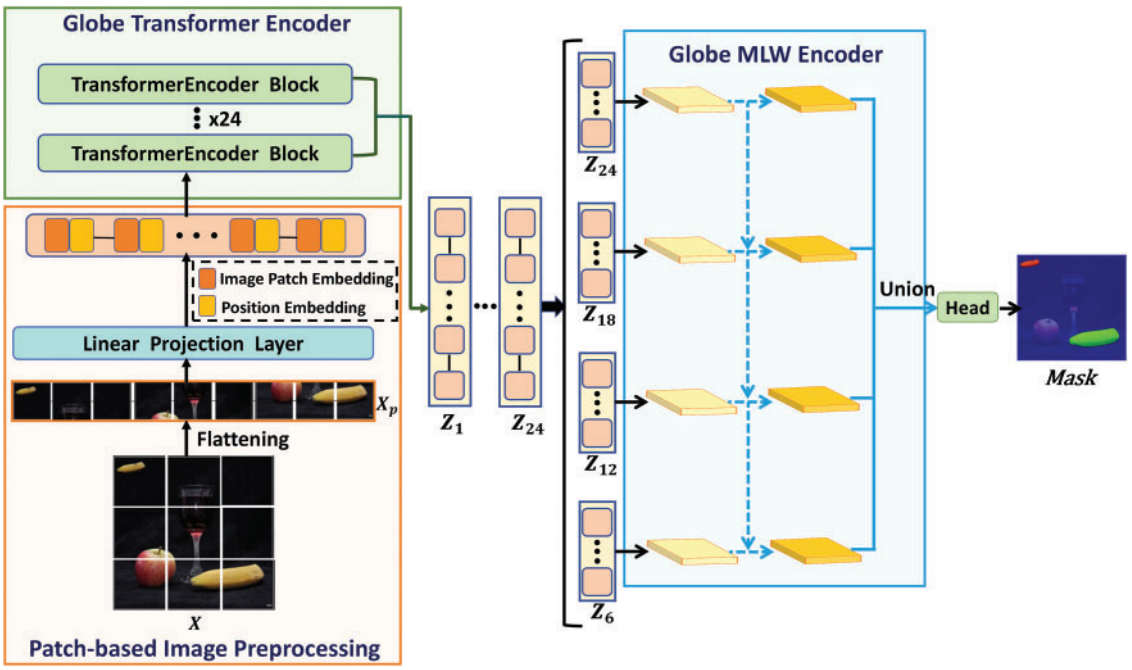


**Figure 1:** Illustration of proposed CMFD transformer (CMFDTR) model

### *3.1 Feature Encoder*

We represent the input image as $X \in \mathbb{R}^{H\times W\times 3}$, with $H$ and $W$ representing its height and width, respectively. The image $X$ undergoes an initial preprocessing step to transform it into a one-dimensional sequence suitable for input into a Transformer encoder. For feature extraction, a convolutional layer with a kernel size of $p \times p$ and a stride of $p$ is employed. This operation effectively partitions the image into $N$ blocks of size $p \times p$. The number of blocks $N$ is determined by the formula: $N = \frac{H}{p} \times \frac{W}{p}$. Following this, the image patches are flattened, and each vectorized patch is mapped to a latent d-dimensional embedding space (set to 1024 in this study) using a linear projection layer $E$, resulting in a one-dimensional sequence of image patch embeddings $S_p$.

To facilitate long-range modeling, encoding the spatial information of image patches is essential. Consequently, a learnable position embedding $Pos_i$ with identical dimensions is incorporated into the embedding sequence $S' \in \mathbb{R}^{(N+1)\times d}$. This process can be expressed by the following equation:

$$S_i^{'} = S_{p_i} + Pos_i \tag{1}$$

Subsequently, the embedded sequence $S'$ is inputted into the global Transformer encoder. The Transformer encoder is constructed with 24 stacked Transformer encoding blocks, each comprising multiple layers of MSAs and multiple layers of multi-layer perceptions (MLPs). The ultimate output produced by the global Transformer encoder is referred to as $Z_{24}$, while the features produced by each stacked layer of the Transformer during the encoding process are represented as $\{Z_1, Z_2, \cdots, Z_{24}\}$, as depicted in Fig. 1.

### *3.2 Feature Decoder*

The model presented in this paper adheres to the encoder-decoder framework, utilizing the Transformer architecture for feature encoding in the encoder component. However, the extracted features encompass

both similar backgrounds and forged regions. To mitigate interference from features of similar backgrounds, a common approach involves dividing the Transformer features into several sets of feature blocks, computing self-correlation scores for each set, and selecting the top k feature blocks with the highest scores. This method aims to reduce interference caused by similar backgrounds [9,15]. However, this empirical approach demonstrates sensitivity to variations in image size.

Considering the inherent nature of image copy-move tampering detection, which involves identifying nearly identical forged targets based on the original source, a one-to-one matching result is preferable to a one-to-many matching outcome. To address this, the proposed model integrates a one-to-one feature matching module and utilizes a multi-scale context decoder to validate the method's efficacy.

The feature matching module integrates the contextual information of each channel in the Transformer features through global average pooling, thereby reducing the feature map's dimensionality from $K \times w \times h$ to $K \times 1 \times 1$. Subsequently, the channel information weights are calculated using one-dimensional convolution, and the Sigmoid activation function is applied to constrain these weights within the range of (0–1). Lastly, the feature weight information is derived through vector multiplication with the original Transformer features, yielding the aggregated information of different channel weights.

The stacked architecture of the Transformer encoder enables each image patch's feature representation to incorporate information from other patches, allowing different encoding layers to capture forgery features at various levels of granularity. To integrate these multi-granularity forgery features, we have designed a CMFDTR-MLW decoder. As illustrated in Fig. 2, the CMFDTR-MLW decoder implements a unidirectional feature fusion strategy, facilitating information integration through a top-down pathway. This approach effectively enhances the flow of information within the Transformer encoder, thereby improving the model's ability to perceive and synthesize multi-level features. Specifically, we divide the 24 Transformer encoder blocks evenly into four groups and extract the embedded features ($Z_6, Z_{12}, Z_{18}, Z_{24}$) from the final block of each group as input. These features are then reshaped into 3D features ($Z_6^{'}, Z_{12}^{'}, Z_{18}^{'}, Z_{24}^{'}$) with dimensions $\frac{H}{p} \times \frac{W}{p} \times C$. Subsequently, these reshaped features are processed through a feature matching module before being input into the CMFDTR-MLW decoder.

Within the MLW decoder, the high-level feature maps are combined with the feature information of the sub-high-level feature maps through element-wise addition. The resulting aggregated feature map is then added element-wise to the next lower-level feature map. The fusion process maintains a copy of the highest layer feature map $Z_{24}^{'}$ (denoted as $P_{24}$), which is subsequently added to the next highest layer feature map. This result is then added to the next feature map. After each addition, the resulting feature map is preserved (i.e., $P_{18}, P_{12}, P_6$). Subsequently, the four fused feature maps undergo processing through two successive 3×3 convolutions to halve the number of channels and reduce it to 3, as well as two 4x up-sampling operations. This process yields four candidate three-class tampering masks ($R_6, R_{12}, R_{18}, R_{24}$) with different hierarchical information and dimensions matching the original image. By obtaining four feature maps with distinct mixed conditions through this approach, each of the four feature maps is decoded separately to produce four decoding results. Finally, a union of the four decoding results is taken to enhance the detection accuracy of tampered regions.

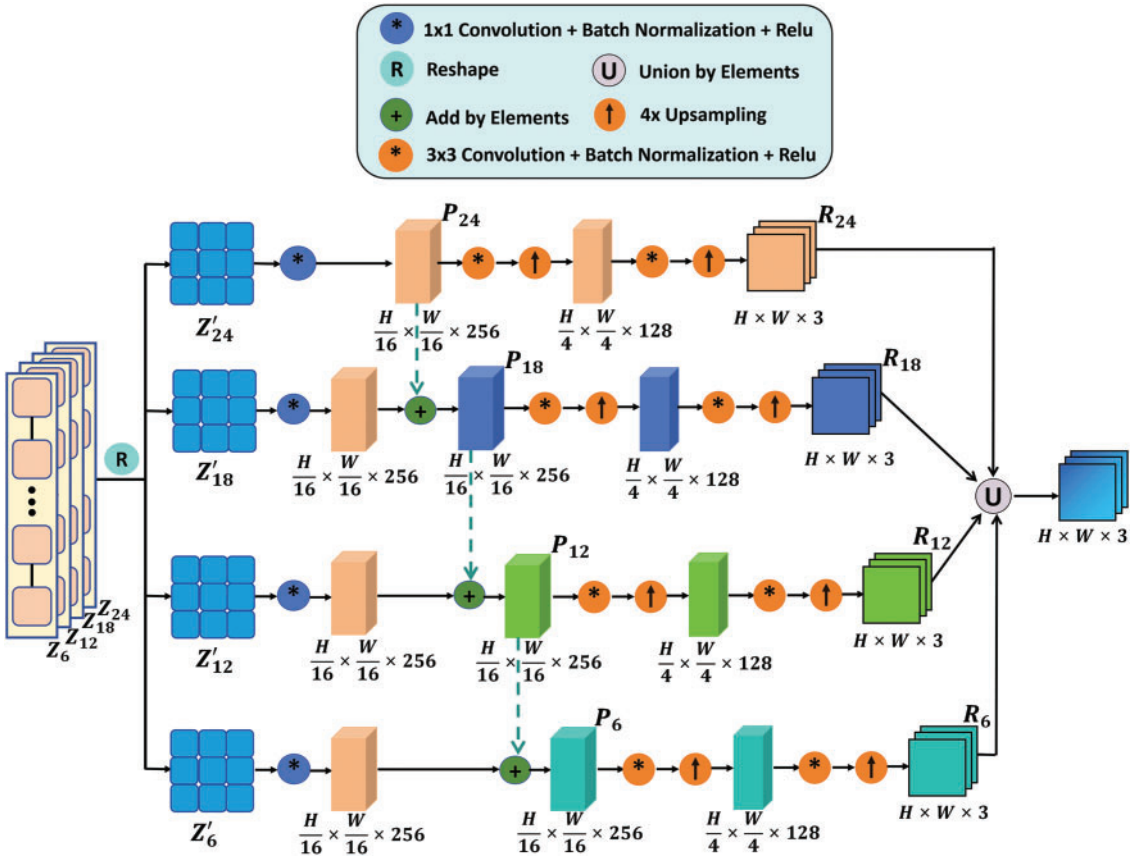


**Figure 2:** Illustration of CMFDTR-MLW decoder

### *3.3 Loss Function*

In the context of image copy-move tampering detection and localization, the model proposed in this paper conducts binary classification at the pixel level and utilizes the binary cross-entropy loss function (*BCELoss*) for network updates. The model's prediction mask is guided by pixel labels from a ground-truth mask, which corresponds to the original image's dimensions. In this mask, pixels labeled as 0 are classified as original, and those labeled as 1 are classified as tampered. The calculation formula for *BCELoss* is presented as follows:

$$BCELoss = -(ylog\,(y) + (1-y)\log\,(1-y)) \tag{2}$$

where $y$ denotes the label value of the ground-truth mask of the image.

The loss for the primary task of the proposed model is denoted as $decode\_loss$. Furthermore, the model includes an auxiliary decoding head designed to extract outputs from various layers of the Transformer encoder. These extracted results are processed through a fusion module and a decoder, and the auxiliary loss is computed by comparing the predicted mask with the ground-truth mask, represented as $au\_loss_i$, where $i$ denotes the $i$-th encoding block layer. Previous research has shown that incorporating auxiliary losses can enhance model training convergence [28]. The total loss $Loss$ of the model is calculated by combining the main task loss $decode\_loss$ with the auxiliary loss $au\_loss_i$. The formula for computing the total loss $Loss$

is presented as follows:

$$Loss = decode\_loss + \sum au\_loss_i \tag{3}$$

This study incorporated corresponding auxiliary decoding heads at the $Z_{10}, Z_{15}, Z_{20}$, and $Z_{24}$ layer of the Transformer encoder.

The MLW decoder ultimately generates four candidate three-category tampering masks $(Z_6, Z_{12}, Z_{18}, Z_{24})$, each containing hierarchical information and matching the original image in size, by integrating the $Z_6, Z_{12}, Z_{18}, and\ Z_{24}$ features from the Transformer encoder. These masks are collectively incorporated into the loss function during model training, where they undergo a weighted summation to obtain a composite cross-entropy. This approach enables more effective model training through back-propagation. The weighted summation process is expressed by the following formula:

$$decode_{loss} = \alpha loss_{R_6} + \beta loss_{R_{12}} + \gamma loss_{R_{18}} + \delta loss_{R_{24}} \tag{4}$$

where $\alpha$, $\beta$, $\gamma$, and $\delta$ are coefficients representing the proportional contribution of the four loss functions to the total decoder loss. In this study, these coefficients are assigned values of 0.1, 0.2, 0.3, and 0.4, respectively.

## 4 Experiment

The experiments described in this paper were conducted using the following experimental setup: Ubuntu operating system, Intel i7-11700K @3.60 GHz CPU, and an NVIDIA GeForce RTX 3090 GPU. The implementation was developed using the PyTorch framework and the MMCV library.

### *4.1 Experimental Setup*

**Datasets:** The performance of our model is evaluated on four widely-recognized datasets in the field of image tamper detection: USCISI [14], CASIAv2.0 [29], DEFACTO [30], and COVERAGE [31]. These datasets not only include images but also provide corresponding ground truth masks that distinguish source, target, and background areas. The source, target, and background areas are denoted in green, red, and blue, respectively. Table 1 presents the specific characteristics of these datasets.

**Table 1:** Condition of datasets

| Datasets | Total number of tampered pictures | For training/ Verification/ Testing | GT mask that distinguishes source/ Destination |
|---|---|---|---|
| USCISI [14] | 100,000 | 80,000/10,000/10,000 | Yes |
| CASIA CMFD [29] | 1311 | 0/0/1311 | Yes |
| DEFACTO CMFD [30] | 7057 | 0/0/7057 | Yes |
| COVERAGE [31] | 100 | 0/0/100 | Yes |

The USCISI dataset, introduced by Wu et al. [14], is a synthetic compilation of digital image tampering instances focusing on copy-move forgeries. It comprises 100,000 samples, each associated with a binary classification mask that distinguishes between untampered and tampered areas for CMFD. In this study's experimental phase, 80,000 samples were randomly extracted from the USCISI dataset for training purposes, while 10,000 samples were allocated for validation, and an additional 10,000 samples were reserved for testing, adhering to an 8:1:1 division ratio.

The CASIA v2.0 dataset [29] comprises 5123 digitally manipulated images, categorized into two types of forgery: splicing and copy-move tampering. This dataset serves as a valuable resource for research in digital image forgery detection. Wu et al. [2] conducted a manual verification of 1313 copy-move forged images within the CASIA V2.0 dataset and generated corresponding binary classification masks, establishing the CASIA-CMFD dataset. In the experimental phase of this study, all samples from the CASIA-CMFD dataset are utilized as the test set to evaluate the efficacy of the proposed model.

DEFACTO [30] is a comprehensive dataset that employs public objects from a contextual database to generate semantically meaningful counterfeit images automatically. This dataset encompasses three categories of forged images: spliced forgeries, copy-paste forgeries, and repair forgeries. From this dataset, we meticulously verified and selected 7057 images containing accurately labeled copy-move tampered images. To enhance the precision of the annotations, we reprocessed the corresponding binary masks for these images, thereby establishing the DEFACTO-CMFD dataset. In the experiments conducted for this paper, the entire DEFACTO-CMFD dataset serves as the test set to evaluate the model proposed herein.

COVERAGE [31] is a dataset specifically designed for CMFD in digital images, consisting of 100 images. The dataset employs a technique of superimposing similar objects onto original authentic images, presenting a significant challenge to human visual recognition. The alterations are subtle and difficult to detect without meticulous examination, as the forged objects are seamlessly integrated. The tampered elements in the dataset encompass a diverse range of items, including merchandise, fruits, furniture, and signage. The intricacy of the forgery details poses a substantial challenge to the generalization capabilities of various copy-move tampering detection models. In this study's experimental phase, all instances from the COVERAGE dataset serve as the test set to assess the performance of the proposed model.

**Model parameters:** The parameters of the Transformer encoder were configured to match those of ViT_Large, as detailed in Table 2. Additionally, batch normalization was implemented as the normalization method for each decoding head.

**Table 2:** Parameter settings of transformer encoder

| Transformer encoder | Number of superimposed layers | Number of embedded channels in image block | Number of self-attention heads | Image block size |
|---|---|---|---|---|
| ViT_Large | 24 | 1024 | 16 | 16 × 16 |

For pre-training, this study utilizes the weights of vit_large_patch16_384 pre-trained on ImageNet, provided by the MMSegmentation [32] project, to initialize the image preprocessing and Transformer encoder modules. The decoder module, in contrast, is randomly initialized.

**Training strategy:** For the training data, we implement the following standard data preprocessing techniques from MM-Segmentation: (1) Random scaling with a ratio range of 0.5 to 2.0. (2) Randomly cropped to achieve dimensions of 256 × 256 pixels. (3) Random horizontal flipping. (4) Photometric distortion. (5) Image normalization.

**Training parameters:** The training parameters are standardized across all models. The batch size is consistently set to 8. We utilize the SGD optimizer, setting the momentum and weight decay parameters to 0.9 and 0, respectively. The initial learning rate is established at 1e-3. For learning rate adjustment, we implement a polynomial decay strategy, setting the polynomial power to 0.9 and the minimum learning rate to 1e-4.

**Test index:** Choosing suitable evaluation metrics is essential for accurately gauging model performance in experimental studies. To quantify localization and other performance aspects, this study employs the most widely used evaluation metrics in the CMFD field, as established in previous literature [26]. Precision measures the ratio of correctly identified positive instances to the total instances predicted as positive. Recall represents the proportion of actual positive instances correctly identified by the model out of all actual positive instances. By merging precision and recall, the F1 score delivers an all-encompassing measure of model effectiveness.

### *4.2 Variant Performance Comparison*

To effectively assess the text model's capability in distinguishing and locating source/target regions, we evaluate the precision, recall, and F-score for each model across three distinct categories: background, source, and target.

Table 3 presents the experimental results of various decoder variants of the text model on the USCISI test set. The variants include CMFDTR-MLW (multi-layer weighting scheme), CMFDTR-Naïve (one-step up-sampling), and CMFDTR-PUP (progressive up-sampling scheme). In the table, data presented in bold text indicate the best performance of the corresponding experimental indicators in the comparative experiments.

**Table 3:** Test results of different variants of CMFDTR on the USCISI dataset

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| CMFDTR-MLW | Background | 97.47 | 96.45 | 98.5 |
| | Source | 74.39 | 84.17 | 66.64 |
| | Target | 80.26 | **84.72** | 76.25 |
| CMFDTR-Native | Background | 97.38 | 96.22 | 98.57 |
| | Source | 73.07 | 84.5 | 64.36 |
| | Target | 79.26 | 84.43 | 74.68 |
| CMFDTR-PUP | Background | **97.74** | **96.86** | **98.64** |
| | Source | **77.29** | **86.86** | **69.62** |
| | Target | **82.94** | 84.39 | **81.55** |

The CMFDTR-Naive decoder utilizes a single-step up-sampling process. This process involves applying a $3 \times 3$ convolution to the feature map $Z_{24}$, followed by a 16-fold up-sampling and batch normalization. In contrast, the CMFDTR-PUP decoder employs a step-by-step up-sampling strategy. This approach processes the feature map $Z_{24}$ through sequential $3 \times 3$ convolutions, gradual up-sampling, and batch normalization. Each up-sampling step doubles the size of the feature map from the previous step, requiring four operations to transform the feature map from size $H/256 \times W/256$ to full resolution.

The experimental findings demonstrate that the Transformer encoder exhibits adaptability to various designed decoders, achieving comparable performance in source/target distinction tasks.

### *4.3 Comparative Analysis of Source and Target Differentiation in Advanced Technological Contexts*

Comparative experiments were performed to demonstrate that the proposed model outperforms current state-of-the-art methods in CMFD. The model was evaluated against four specialized CMFD models: BusterNet, DOA-GAN, CMSDNet, and MVSS-Net. These comparative experiments were performed using

four publicly available CMFD datasets: DEFACTO CMFD, USCISI, CASIA CMFD and COVERAGE. This comprehensive approach aimed to thoroughly test and compare the performance of the proposed model across diverse datasets in the Copy-Move forgery domain.

The USCISI dataset was employed by training the model on its training set and directly evaluating it on the test set. In contrast, for the DEFACTO CMFD, CASIA CMFD, and COVERAGE datasets, all samples were utilized as the test set for evaluation without any fine-tuning. This methodology enables a more comprehensive assessment of the model's generalization capabilities.

The pixel-wise localization performance of the compared models on the USCISI test set are shown in Table 4. In the table, data presented in bold text indicate the best performance of the corresponding experimental indicators in the comparative experiments. Owing to the substantial similarity in data distribution between the USCISI test set and the training data, each model has achieved satisfactory three-class localization performance. The results indicate that the CMFDTR proposed in this study outperforms BusterNet and CMSDNet on most metrics in the USCISI dataset, although it slightly underperforms compared to DOA-GAN and MVSS-Net. Notably, the USCISI dataset contains relatively few complex tampered samples. Most tampered samples in this dataset involve simple manipulations where source regions are scaled by a certain factor, subjected to minor random rotations, and then copied and moved to target areas. The DOA-GAN model, employing an adversarial training mechanism through the competitive process between the generator and the discriminator, captures the data's distributional characteristics more effectively. Consequently, DOA-GAN exhibits superior detection performance on the USC-ISI dataset, where the data distribution is largely consistent. However, its performance in detecting complex tampered regions may decline, indicating limited generalization capability. MVSS-Net jointly utilizes tampering boundary artifacts and noise views of the input image to extract semantic-agnostic features, better capturing lower-level features. It enhances detection specificity through multi-scale supervision, at the cost of reduced detection sensitivity, which is compensated for through multi-view feature learning. While MVSS-Net demonstrates excellent detection capabilities on datasets with highly consistent data distribution, its generalization performance may be relatively poor when faced with data that diverges from the training data distribution. By analyzing the detection results from the CASIA CMFD, DEFACTO CMFD, and COVERAGE datasets, it is evident that the CMFDTR's detection metrics generally surpass those of DOA-GAN and MVSS-Net, demonstrating that the proposed CMFDTR exhibits stronger generalization capabilities compared to DOA-GAN and MVSS-Net. As shown in Fig. 3, we compare the prediction results on the UISICI dataset using various methods, including BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and our proposed method.

**Table 4:** Source/target distinguishment test results of comparison model on the USCISI dataset

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| BusterNet [14] | Background | 96.03 | 94.35 | 97.77 |
| | Source | 60.33 | 65.86 | 55.66 |
| | Target | 77.76 | **84.72** | 71.87 |
| DOA-GAN [17] | Background | **97.94** | **96.89** | **99.0** |
| | Source | **81.83** | 84.18 | **79.6** |
| | Target | **86.27** | 84.08 | **88.57** |
| CMSDNet [15] | Background | 96.44 | 95.04 | 97.89 |
| | Source | 63.57 | 70.97 | 57.58 |
| | Target | 34.82 | 59.70 | 24.75 |
| | Background | 97.74 | 96.74 | 98.75 |

(Continued)

**Table 4 (continued)**

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| MVSS-Net [18] | Source | 80.42 | 81.6 | 79.27 |
| | Target | 85.32 | 84.42 | 86.24 |
| CMFDTR-MLW | Background | 97.47 | 96.45 | 98.5 |
| | Source | 74.39 | 84.17 | 66.64 |
| | Target | 80.26 | **84.72** | 76.25 |
| CMFDTR-Naive | Background | 97.38 | 96.22 | 98.57 |
| | Source | 73.07 | 84.5 | 64.36 |
| | Target | 79.26 | 84.43 | 74.68 |
| CMFDTR-PUP | Background | 97.74 | 96.86 | 98.64 |
| | Source | 77.29 | **86.86** | 69.62 |
| | Target | 82.94 | 84.39 | 81.55 |

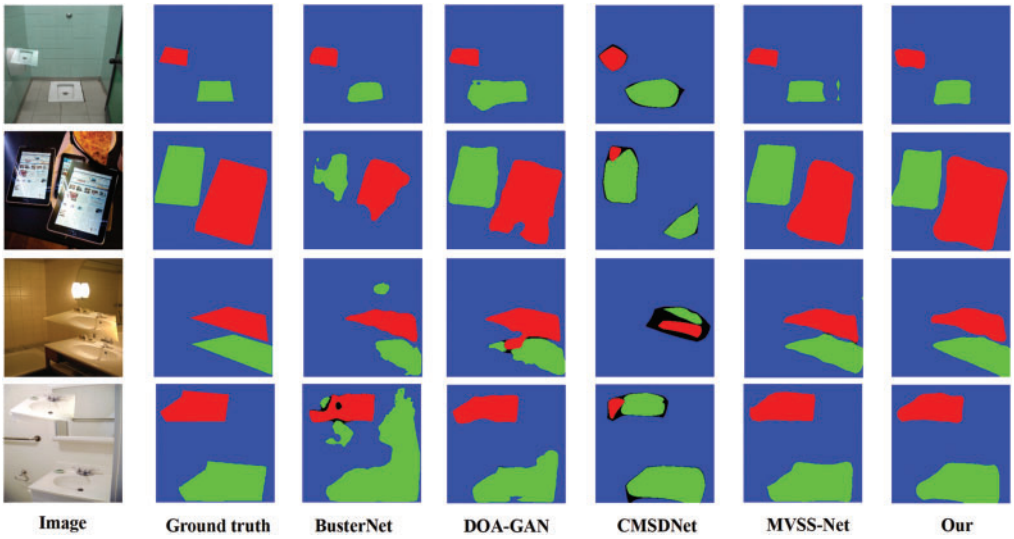


**Figure 3:** Comparison of prediction results on the UISICI dataset using BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and the proposed method (Blue represents the background, green represents the source, and red represents the target)

Table 5 displays the pixel-wise localization results for the models evaluated using the CASIA CMFD test set. In the table, data presented in bold text indicate the best performance of the corresponding experimental indicators in the comparative experiments. The dataset incorporates samples with Copy-Move forgeries in highly similar or semantically ambiguous background regions, as well as instances where copied and pasted regions overlap. In contrast, the USCISI dataset seldom contains such intricate tampered samples. Consequently, this disparity may result in suboptimal performance when a model trained exclusively on the USCISI training set is evaluated on the CASIA CMFD test set.

The results of the experiments confirm that the CMFDTR proposed herein surpasses other models in most performance metrics when evaluated on the CASIA CMFD dataset, surpassing the comparison models BusterNet, CMSDNet, DOA-GAN, and MVSS-Net. This indicates enhanced generalization capability and improved performance in detecting tampered regions. As shown in Fig. 4, we compare the prediction results

on the CASIA CMFD test set using various methods, including BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and our proposed method.

**Table 5:** Source/Target distinguishment test results of comparison model on the CASIA CMFD dataset

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| BusterNet [14] | Background | 94.58 | 90.64 | 98.89 |
| | Source | 13.73 | 25.96 | 9.33 |
| | Target | 1.32 | 20.76 | 0.68 |
| DOA-GAN [17] | Background | 95.17 | 91.29 | 99.05 |
| | Source | 20.17 | **45.02** | 13.0 |
| | Target | 13.23 | 37.45 | 8.03 |
| CMSDNet [15] | Background | 94.72 | 90.52 | **99.34** |
| | Source | 10.74 | 43.52 | 6.12 |
| | Target | 1.75 | 24.52 | 0.91 |
| MVSS-Net [18] | Background | 94.65 | 90.41 | 99.32 |
| | Source | 8.55 | 32.27 | 4.93 |
| | Target | 6.19 | 33.42 | 3.41 |
| CMFDTR-MLW | Background | **95.98** | **93.59** | 98.49 |
| | Source | **24.55** | 32.68 | **19.66** |
| | Target | 9.92 | 54.18 | 5.46 |
| CMFDTR-Naive | Background | 95.97 | 93.55 | 98.51 |
| | Source | 22.38 | 31.55 | 17.34 |
| | Target | 11.93 | 47.85 | 6.82 |
| CMFDTR-PUP | Background | 95.89 | 93.4 | 98.51 |
| | Source | 21.61 | 31.73 | 16.39 |
| | Target | **13.37** | **55.97** | **7.59** |

The pixel-wise localization performance for each benchmark model on the DEFACTO CMFD test set is detailed in Table 6. In the table, data presented in bold text indicate the best performance of the corresponding experimental indicators in the comparative experiments. Importantly, the DEFACTO CMFD dataset exhibits substantial differences in data distribution compared to the USCISI dataset, including more complex samples that are challenging for human visual perception, smaller target samples, and potentially misleading instances. Additionally, the DEFACTO CMFD dataset is more extensive than the CASIA CMFD dataset. The experimental results indicate that the CMFDTR proposed in this study outperforms the compared models (BusterNet, CMSDNet, DOA-GAN, and MVSS-Net) on the DEFACTO CMFD dataset for the majority of evaluation metrics. Notably, the recall is significantly higher than other comparison methods, suggesting that the proposed method, through a one-to-one matching strategy, effectively filters out incorrect matching features. This demonstrates that the CMFDTR possesses stronger generalization capabilities and superior performance in detecting tampered regions. As shown in Fig. 5, we compare the prediction results on the DEFACTO CMFD test set using various methods, including BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and our proposed method.

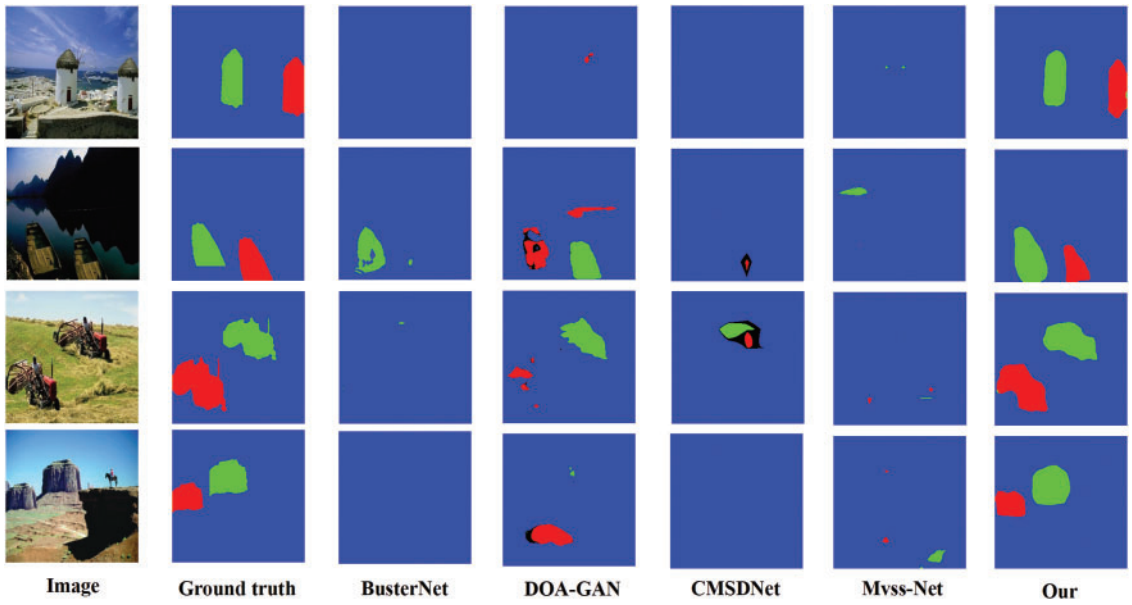


**Figure 4:** Comparison of prediction results on the CAISA CMFD dataset using BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and the proposed method (Blue represents the background, green represents the source, and red represents the target)

**Table 6:** Source/target distinguishment test results of comparison model on the DEFACTO CMFD dataset

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| BusterNet [14] | Background | 95.11 | 91.53 | 98.99 |
| | Source | 18.37 | 26.85 | 13.96 |
| | Target | 1.23 | 28.72 | 0.63 |
| DOA-GAN [17] | Background | 95.78 | 92.38 | **99.44** |
| | Source | 28.2 | **48.52** | 19.88 |
| | Target | 16.76 | 47.21 | 10.19 |
| CMSDNet [15] | Background | 95.22 | 91.51 | 99.25 |
| | Source | 11.29 | 32.7 | 6.82 |
| | Target | 3.47 | 29.24 | 1.85 |
| MVSS-Net [18] | Background | 95.12 | 91.27 | 99.32 |
| | Source | 14.46 | 37.09 | 8.98 |
| | Target | 11.19 | 49.39 | 6.31 |
| CMFDTR-MLW | Background | **96.14** | **93.75** | 98.67 |
| | Source | **31.07** | 36.67 | **26.96** |
| | Target | 14.39 | **61.97** | 8.15 |

(Continued)

**Table 6 (continued)**

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| CMFDTR-Naive | Background | 95.99 | 93.48 | 98.64 |
| | Source | 26.39 | 33.91 | 21.61 |
| | Target | 16.72 | 59.06 | 9.74 |
| CMFDTR-PUP | Background | 96.0 | 93.39 | 98.75 |
| | Source | 25.5 | 35.6 | 19.86 |
| | Target | **18.93** | 58.3 | **11.3** |

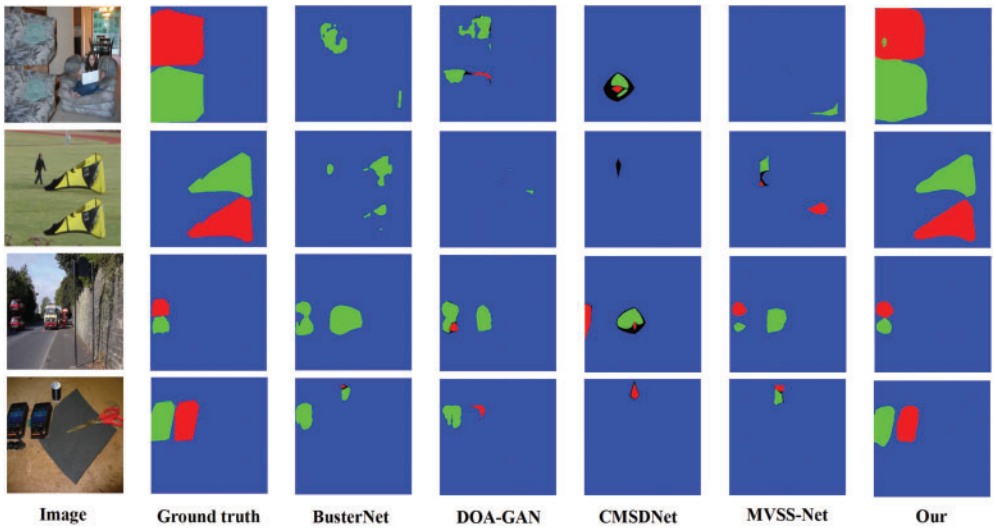


**Figure 5:** Comparison of prediction results on the DEFACTO CMFD dataset using BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and the proposed method (Blue represents the background, green represents the source, and red represents the target)

Table 7 summarizes the pixel-wise localization performance of each comparative model evaluated on the COVERAGE test set. In the table, data presented in bold text indicate the best performance of the corresponding experimental indicators in the comparative experiments. The forged images in the COVERAGE dataset are created by overlaying similar objects onto the original real images. This dataset poses a significant challenge to human visual recognition due to its finely detailed forgeries, and it is particularly demanding for image tamper detection and localization models to perform effectively. The experimental results demonstrate that all detection metrics of the proposed CMFDTR in this study surpass those of the compared models BusterNet, CMSDNet, and MVSS-Net, but are slightly lower than the DOA-GAN model. Analysis of the visualization results suggests that the marginally lower detection performance of the method proposed in this study compared to the DOA-GAN model is attributable to the misclassification of a portion of the forged source and target regions. However, considering the comprehensive detection results across the four test sets utilized in this study, it is evident that the proposed CMFDTR exhibits stronger generalization capabilities and superior performance in detecting tampered regions for unknown tampered images. As shown in Fig. 6, we compare the prediction results on the COVERAGE test set using various methods, including BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and our proposed method.

**Table 7:** Source/target distinguishment test results of comparison model on the CAVERAGE dataset

| Methods | Categories | F1-score | Precision | Recall |
|---|---|---|---|---|
| BusterNet [14] | Background | 87.29 | 78.33 | 98.57 |
| | Source | 21.06 | 38.12 | 14.55 |
| | Target | 1.44 | 34.37 | 0.73 |
| DOA-GAN [17] | Background | 88.43 | **81.26** | 96.99 |
| | Source | **33.84** | 49.08 | **25.83** |
| | Target | **22.24** | 49.89 | **14.31** |
| CMSDNet [15] | Background | 88.02 | 80.14 | 97.63 |
| | Source | 23.21 | 36.63 | 16.99 |
| | Target | 2.88 | 39.58 | 1.5 |
| MVSS-Net [18] | Background | 87.14 | 77.61 | 99.32 |
| | Source | 17.19 | 63.28 | 9.95 |
| | Target | 13.15 | 62.77 | 7.35 |
| CMFDTR-MLW | Background | **89.12** | 80.89 | 99.21 |
| | Source | 33.26 | 60.22 | 22.98 |
| | Target | 12.87 | **71.75** | 7.07 |
| CMFDTR-Naive | Background | 88.59 | 80.12 | 99.07 |
| | Source | 30.05 | 61.42 | 19.89 |
| | Target | 10.78 | 63.81 | 5.89 |
| CMFDTR-PUP | Background | 87.91 | 78.8 | **99.39** |
| | Source | 19.31 | **66.18** | 11.3 |
| | Target | 10.46 | 64.42 | 5.69 |

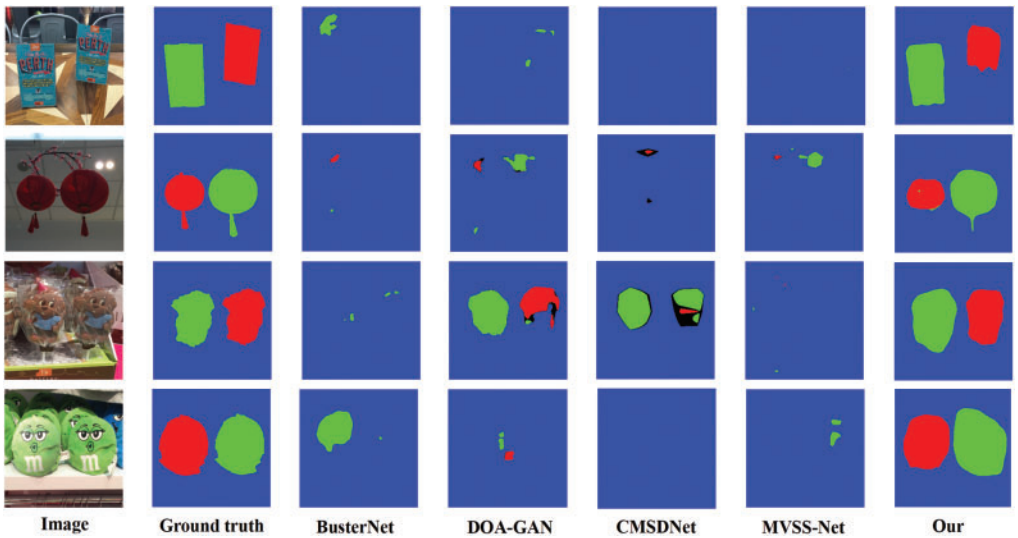


**Figure 6:** Comparison of prediction results on the COVERAGE dataset using BusterNet, DOA-GAN, CMSDNet, MVSS-Net, and the proposed method (Blue represents the background, green represents the source, and red represents the target)

## 5 Conclusion

This paper introduces a Transformer structure based on sequence-to-sequence modeling and proposes an end-to-end model called CMFDTR, specifically designed for the characteristics of the CMFD task. In comparison to existing DCNN methods, CMFDTR eliminates the reliance on image down-sampling throughout the feature encoding process, effectively mitigating the common issue of detail loss associated with DCNN approaches. Furthermore, the multi-head mechanism of MSA enhances the model's global context modeling capabilities. The self-attention mechanism demonstrates greater suitability for copy-move forgery characteristics than traditional feature correlation matching calculations. Experimental results demonstrate that our model outperforms other advanced techniques on USCISI, DEFACTO CMFD, CASIA CMFD, and COVERAGE datasets, indicating the high adaptability and promising potential of Transformer for the CMFD task. Future research endeavors will concentrate on expanding the current work, enhancing the Transformer encoder, and refining the model's capabilities for more precise detection and localization of copy-move forgery.

**Acknowledgement:** The authors would like to express appreciation to the National Natural Science Foundation of China, Department of Science and Technology of Guangdong Province, China, and Education Department of Guangdong Province, China for their financial support.

**Funding Statement:** The research received financial support from the General Program of the National Natural Science Foundation of China (Grant No. 62072123), Key R&D Initiatives in Guangdong Province (Grant No. 2021B0101220006), the Guangdong Provincial Department of Education's Key Field Projects for Ordinary Colleges and Universities (Grant Nos. 2020ZDZX3059, 2022ZDZX1012, 2023ZDZX1008), Key R&D Projects in Jiangxi Province (Grant No. 20212BBE53002), and Key R&D Projects in Yichun City (Grant No. 20211YFG4270).

**Author Contributions:** The authors declare their contributions to the paper as encompassing the conception and design of the study: Gang Hao, Peng Liang; data collection: Gang Hao, Ziyuan Li and Hong Zhang; analysis and interpretation of results: Gang Hao, Peng Liang, Huimin Zhao, Ziyuan Li and Hong Zhang; draft manuscript preparation: Gang Hao, Ziyuan Li and Hong Zhang. 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 openly at: https://www.kaggle.com/datasets/defactodataset/defactocopymove, accessed on 10 December 2023; https://www.kaggle.com/datasets/divg07/casia-20-image-tampering-detection-dataset, accessed on 02 January 2024; https://drive.google.com/file/d/1gsx5c-oilsFEzX_j1zKTPP4yWEs6T385/view?usp= sharing, accessed on 15 January 2024.

**Ethics Approval:** No human or animal subjects were involved, and thus ethical approval was not required.

**Conflicts of Interest:** This research involves the detection and localization of copy-move forgery in digital images using publicly available datasets. All data utilized comply with relevant privacy and data protection regulations.

## References

1. Fridrich J, Soukal D, Lukas J. Detection of copy-move forgery in digital images. In: Proceedings of Digital Forensic Research Workshop (DFRWS). Cleveland: IEEE; 2003. p. 67–84.
2. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017:30.
3. Warif NBA, Idris MYI, Wahab AWA, Ismail N-SN, Salleh R. A comprehensive evaluation procedure for copy-move forgery detection methods: results from a systematic review. Multimed Tools Appl. 2022;81(11):15171. doi:10.1007/s11042-022-12010-2.
4. Tan W, Wu Y, Wu P, Chen B. A survey on digital image copy-move forgery localization using passive techniques. J New Media. 2019;1(1):11–25. doi:10.32604/jnm.2019.06219.

5. Gurunlu B, Ozturk S. Efficient approach for block-based copy-move forgery detection. In: Zhang YD, Senjyu T, So-In C, Joshi A, editors. Smart trends in computing and communications. Lecture notes in networks and systems. Vol. 286. Singapore: Springer; 2022. doi:10.1007/978-981-16-4016-2_16.
6. Ryu SJ, Lee MJ, Lee HK. Detection of copy-rotate-move forgery using Zernike moments. In: Information Hiding: 12th International Conference, IH 2010; 2010 Jun 28–30; Calgary, AB, Canada: Revised Selected Papers 12. Springer Berlin Heidelberg; 2010. p. 51–65.
7. Mahmood T, Nawaz T, Irtaza A. Copy-move forgery detection technique for forensic analysis in digital images. Math Probl Eng. 2016;2016(1):1–13. doi:10.1155/2016/8713202.
8. Deng-Yuan H, Ching-Ning H, Wu-chih H. Robustness of copy-move forgery detection under hign JPEG compression artifacts. Multimed Tools Appl. 2017;76(1):1509–30. doi:10.1007/s11042-015-3152-x.
9. Li Y, Zhou J. Fast and effective image copy-move forgery detection via hierarchical feature point matching. IEEE Trans Inf Forensics Secur. 2019 May;14(5):1307–22. doi:10.1109/TIFS.2018.2876837.
10. Amerini I, Ballan L, Caldelli R. A sift-based forensic method for copy-move attack detection and transformation recovery. IEEE Trans Inf Forensics Secur. 2011;6(3):1099–110. doi:10.1109/TIFS.2011.2129512.
11. Shivakumar BL, Baboo SS. Detection of region duplication forgery in digital images using SURF. Int J Comput Sci Issues (IJCSI). 2011;8(4):199–205.
12. Diwan A, Sharma R, Roy AK. Keypoint based comprehensive copy-move forgery detection. IET Image Process. 2021;15(6):1298–309. doi:10.1049/ipr2.12105.
13. Barad ZJ, Goswami MM. Image forgery detection using deep learning: a survey. In: Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, India: IEEE; 2020.
14. Wu Y, Abd-Alamgeed W, Natarajan P. BusterNet: detecting copy-move image forgery with source/target localization. In: Proceedings of the European Conference on Computer Vision (ECCV); Berlin, Germany: Springer; 2018. p. 168–84.
15. Chen B, Tan W, Coatrieux G. A serial image copy-move forgery localization scheme with source/target distinguishment. IEEE Trans Multimedia. 2020;23:3506–17. doi:10.1109/TMM.2020.3026868.
16. Hu X, Zhang Z, jiang Z. Span: spatial pyramid attention network for image manipulation localization. In: Proceedings of the European Conference on Computer Vision, ECCV 2020. Glasgow, UK; 2020 Aug 23–28.
17. Islam A, Long G, Basharat A. DOA-GAN: dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In: Proceedings of the IEEE/CVF Conference On Computer Vision And Pattern Recognition. Piscataway, NJ: IEEE; 2020. p. 4676–85.
18. Dong C, Chen X, Hu R, Cao J, Li X. MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans Pattern Anal Mach Intell. 2023 Mar 1;45(3):3539–53. doi:10.1109/TPAMI.2022.3180556.
19. Dosovitskiy A, Beyer L, Kolesnikov A. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations; 2021 May 4–8; New Orleans, LA, USA.
20. Carion N, Massa F, Synnaeve G. End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK: Springer International Publishing; 2020. p. 213–29.
21. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations; 2021 May 4–8; Vancouver, BC, Canada.
22. Zheng SX, Lu JC, Zhao HS. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2021. p. 6881–980.
23. Chen JN, Lu YY, Yu QH. TransUNet: transformers make strong encoders for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2021.

24. Touvron H, Cord M, Douze M. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. New York, NY: PMLR; 2021. p. 10347–57.
25. Liu Z, Lin YT, Cao Y. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE; 2021. p. 10012–22.
26. Wang J, Wu Z, Chen J. ObjectFormer for image manipulation detection and localization. In: Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA.
27. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2015. p. 3431–40.
28. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2017 Jul 21–26; Honolulu, HI, USA; p. 2881–90.
29. Dong J, Wang W, Tan T. CASIA image tampering detection evaluation database. In: Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing; 2017 Jul 6–10; Beijing, China. p. 422–26.
30. Mahfoudi G, Tajini B, Retraint F, Morain-Nicolier F, Dugelay JL, Pic M. DEFACTO: Image and face manipulation dataset. In: Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO); 2019 Sep 2–6; A Coruna, Spain. p. 1–5.
31. Wen BH, Zhu Y, Subramanian R, Ng TT, Shen XJ, Winkler S. COVERAGE—a novel database for copy-move forgery detection. In: 2016 IEEE International Conference on Image Processing (ICIP); 2016; Phoenix, AZ, USA. p. 161–5. doi:10.1109/ICIP.2016.7532339.
32. OpenMMLab. Welcome to mmsegmentation's documentation! 2022 Nov 04 [cited 2023 Jun 25]. Available from: https://mmsegmentation.readthedocs.io/en/latest/.