Source and Target Region Discrimination on Copy-Move Image Forgery Using SegFormer Model

Imam Ekowicaksono*†
*Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
33220306@std.stei.itb.ac.id
†Teknik Informatika, FTI
Institut Teknologi Sumatera
South Lampung, Indonesia
imam.wicaksono@if.itera.ac.id

Rinaldi Munir*
*Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
rinaldi@staff.stei.itb.ac.id

Masayu Leylia Khodra*
*Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
masayu@staff.stei.itb.ac.id

***Abstract*—Copy-move image forgery is a type of image manipulation where a part of an image is copied and pasted onto another part of the same image. Detecting copy-move image forgery involves identifying duplicated region and discriminating between the source and target regions, which can provide investigators with insights into the purpose of the forgery. This research aims to discriminate between the source and target regions in copy-move image forgery using SegFormer. SegFormer is a transformer-based semantic segmentation model. This research train SegFormer architecture from scratch to differentiate source and target regions. Experiments on the CoMoFoD dataset show that the SegFormer model achieved a mIoU score over 60%, demonstrating its effectiveness in discriminating source and target regions on a publicly available dataset.**

***Index Terms*—copy-move image forgery, image forgery, deep learning, transformer, transfer learning.**

I. Introduction

In today's world, the rapid advancement of digital image processing technology has made it increasingly simple, quick, and affordable to manipulate digital images. The explosive growth of social media platforms has made it easier than ever to share these altered images, potentially blurring the line between fact and fiction. This situation creates both possibilities and obstacles in restoring the credibility of images as accurate, factual, and trustworthy sources of information.

A study by Lumoindong et. al. [1] in 2020 found that most people in Indonesia struggle to identify manipulated images from genuine ones. The survey showed that 85% of participants expressed a need for automated systems to detect falsified image content. In the academic realm, research [2] examining over 20,000 biomedical scientific publications from 1995 to 2014 discovered that roughly 3.8% contained problematic images. More than half of these papers with questionable images showed signs of image manipulation. This trend could potentially undermine academic integrity by presenting altered images as authentic data.

There are multiple prevalent techniques for image forgery, including image inpainting, image splicing, and copy-move. Image inpainting alters an image by reconstructing specific region through pixel interpolation based on surrounding pixels [3]. Image splicing modifies images by inserting patch from different sources [4]. Copy-move image forgery (CMIF) involves duplicating a section of an image and placing it elsewhere within the same image. The purpose of CMIF is to either remove an object from the image or create the illusion of multiple instances of an object within the same image. Figure 1 presents a visualization of image manipulation utilizing the copy-move technique.

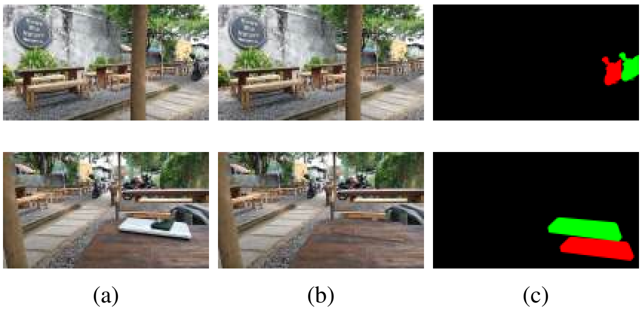


Fig. 1: Copy-move image forgery examples: (a) original image (b) forged image (c) ternary mask.

CMIF detection techniques can be categorized into three main approaches: block-based, keypoint-based, and deep learning-based methods. The block-based approach involves segmenting the image into smaller sections and analyzing the correlation between these segments to identify duplicated region [5]–[7]. In contrast, keypoint-based methods utilize specific feature extraction algorithms such as scale-invariant feature transform (SIFT) [8], speeded up robust features (SURF) [9], and oriented FAST and rotated BRIEF (ORB) [10] to extract image features.

Block-based CMIF detection generally employs a methodical approach. However, its process of patching and extracting features is inefficient and time-intensive. Furthermore, block-based techniques exhibit limitations in identifying CMIF in manipulated images with substantial rotational and scaling variations [11]. In contrast, keypoint-based detection methods demonstrate resilience to affine transformations. Nevertheless, these methods encounter challenges when analyzing images containing homogenous textures, as keypoints are difficult to extract in such instances. Consequently, keypoint-based approaches prove ineffective for CMIF detection in images characterized by homogeneous textures [12].

The progress in deep learning has enabled automated CMIF detection. This approach employs a neural network structure consisting of an encoder for extracting features and a decoder for producing a mask that identifies copied region. Models based on convolutional neural networks [13]–[15] and transformers [18]–[20], which utilize deep learning techniques, have shown effectiveness in accurately identifying CMIF within images.

Distinguishing between source and target regions is a crucial aspect of CMIF detection. This process involves dividing the image into three distinct categories: source, target, and unaltered region. The task of differentiating source from target regions can be approached as an image segmentation problem. SegFormer [21], a highly effective transformer-based model for semantic segmentation, stands out in this field. This study applies the SegFormer model to the challenge of source and target region identification in CMIF.

This research has 2 main contributions:

- Employ the segformer model to differentiate source and target regions in CMIF.
- Experimental results indicate that SegFormer model demonstrates superior performance in terms of mean Intersection over Union (mIoU) compared to the state-of-the-art model for the source and target region discrimination task.

II. Literature Review

A. Copy-Move Image Forgery Detection

The primary objective in identifying CMIF is to localize the duplicated region within the manipulated image. This process involves categorizing the image into two distinct areas: the duplicated region and the unaltered region. The duplicated region refers to the section where content has been replicated and inserted, while the unaltered region represents the original, unchanged portion of the image. However, merely identifying the duplicated region does not provide sufficient information to comprehend the intent behind the image manipulation. Consequently, distinguishing between the source and target regions in CMIF plays a crucial role in effectively detecting this type of forgery.

The differentiation between source and target region on CMIF can be viewed as an image segmentation problem. This process involves separating the image into three distinct categories: the origin region, the destination region, and the unaltered/pristine region. The origin region represents the area from which content is copied, the destination region is where this copied content is inserted, and the pristine region encompasses the portions of the image that remain unchanged by these copying and pasting actions.

B. Deep Learning-Based CMIF Detection

CMIF detection techniques can be broadly classified into three categories: block-based, keypoint, and deep learning approaches. The block-based and keypoint methods utilize specific feature extraction techniques such as DCT, SIFT, and SURF to obtain relevant features. These extracted features are then compared to identify duplicated regions. In contrast, deep learning approaches employ a neural network architecture consisting of an encoder for feature extraction and a decoder for generating a mask of the duplicated area.

1) CNN-Based CMIF Detection: A CMIF detection approach using deep learning to distinguish between source and target region was introduced by [13] in 2018. This study extracts CMIF image features using the first four blocks of VGG16. It also incorporates and combines ManiDet and SimiDet branches to generate ternary masks for source, target, and unaltered regions. As auxiliary tasks, SimiDet and ManiDet produce duplication region and target regions. Nevertheless, this research employed several training strategies before end-to-end training to create a more effective model for differentiating between source and target region. The limitation of BusterNet is that it relies on SimiDet and ManiDet. If one of them fails, the model will not be able to detect CMIF.

In a separate investigation, researchers created a DOA-GAN [16] (dual-order attentive module combined with GAN) for CMIF detection. This model generates two types of attention maps: copy-move region and co-occurrence, leading to a more distinctive feature representation. The DOA-GAN also incorporates the ASPP module from DeepLabV3+ [17] to improve its CMIF detection capabilities. Nevertheless, the model shows limitations in identifying CMIF in very small images and in region with uniform background duplication.

2) Transformer-Based CMIF Detection: A CNN and Transformer-based discrimination model for CMIF detection was developed by Zhang [18]. Unlike previous studies that emphasized encoder design based on resulting feature representation, this research combines two existing encoders. The first is from the Vision Transformer (ViT) [22], which extracts global image features. The second is from the Feature Pyramid Network (FPN) [23], which uses a top-down architecture to extract multi-scale local image features. While this combination enhances model performance, it also leads to a substantial increase in parameters, requiring significant computational resources and time.

The research presented in [19] highlights the common challenges faced by existing techniques, particularly in dealing with significant dataset variations and accurately identifying forged region. To tackle these issues, the study introduces an innovative approach employing a convolutional neural network (CNN) that aims to achieve a balanced performance in both detecting forgeries and pinpointing source/target locations. This method focuses on extracting the inherent characteristics of manipulated regions rather than simply memorizing dataset patterns, thereby improving its efficacy. The researchers combine ResNet18 for feature extraction with Twins [24] as a decoder to emphasize altered region on the feature map.

They developed $Decoder_f$ to generate binary ground truth representing duplication region and $Decoder_d$ to produce ternary images showing source and target regions. While the study successfully identified duplication region, there remains room for developing a more robust and dependable model.

Another transformer-based model for detecting Copy-Move Image Forgery (CMIF) was introduced by Liu [20]. The research primarily focuses on identifying duplicated region in manipulated images, which is essential in digital forensic analysis. The proposed CMFDFormer employs a Transformer-based framework, aiming to improve both the precision and efficiency of copy-move forgery detection. To enhance its detection capabilities, the model's architecture integrates hierarchical feature integration and self-correlation computation. The implementation of continual learning allows the model to adjust to new information and challenges without losing previously learned knowledge, which is particularly beneficial given the evolving nature of forgery techniques. Experimental outcomes indicate that the CMFDFormer outperforms existing methodologies in terms of detection accuracy, especially when faced with various image alterations and attacks.

C. SegFormer Model

SegFormer, a semantic segmentation model based on transformer architecture, combines a transformer encoder with a simplified multi-layer perceptron decoder [21]. The model's encoder employs hierarchical feature representation to create multi-level feature maps. It processes input images with dimensions $H \times W \times C$, reshaping them into a sequence of 2D patches using overlap patch embedding. The Transformer block, consisting of efficient self-attention, Mix-FFN, and patch merging elements, generates multi-level features (F_i) with dimensions $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1} \times C_i}$. Each Transformer block output undergoes processing through an MLP layer to standardize the channel dimension. These unified features are then upsampled by a factor of 4 and merged. The combined features are further refined through an additional MLP layer. A final MLP layer is used to predict the resulting segmentation mask. The architecture of SegFormer is illustrated in Figure 2.

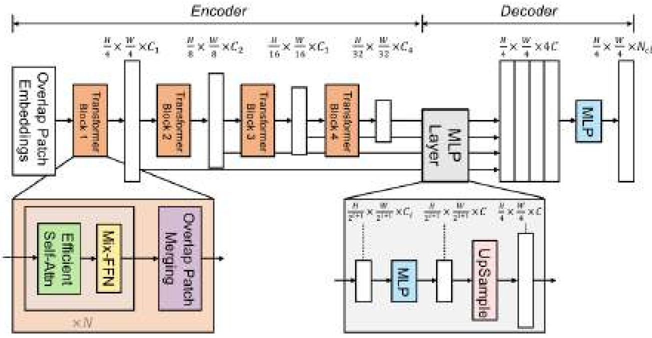

Encoder
Decoder
$\frac{H}{4} \times \frac{W}{4} \times C_1$
$\frac{H}{8} \times \frac{W}{8} \times C_2$
$\frac{H}{16} \times \frac{W}{16} \times C_3$
$\frac{H}{32} \times \frac{W}{32} \times C_4$
$\frac{H}{4} \times \frac{W}{4} \times 4C$
$\frac{H}{4} \times \frac{W}{4} \times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
×N
$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$
$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$
$\frac{H}{4} \times \frac{W}{4} \times C$
MLP
UpSample

Fig. 2: SegFormer Architecture

The efficient self attention method [25] reduces the spatial scale of $\mathbf{K}$ and $\mathbf{V}$, thereby decreasing computational complexity. In the $i-$ th stage, C_i represents the feature map channel, N_i is the number of attention layer heads, and n_i is the corresponding head index. The dimension of each head is given by $d_i = \frac{C_i}{N_i}$. The spatial scale reduction function $\Gamma(\cdot)$ for $\mathbf{K}$ and $\mathbf{V}$ is defined as:

$$\Gamma(\mathbf{x}) = \text{Norm}(\text{Reshape}(\mathbf{x}, R_i)\mathbf{W}_i^S). \tag{1}$$

Here, $\mathbf{x}$ is the input sequence, and R_i is the reduction ratio of the stage i attention layer. The Reshape$(\cdot)$ function transforms $\mathbf{x} \in \mathbb{R}^{(h_i, w_i, C_i)}$ into $\mathbf{x} \in \mathbb{R}^{(\frac{h_i \times w_i}{R_i}, R_i \times C_i)}$. $\mathbf{W}_i^S$ is a linear projection that reduces the input sequence x dimension to C_i. Layer normalization is represented by Norm$(\cdot)$. Figure 3 illustrates the efficient self attention process. The self attention function Att$(\cdot)$ can be expressed as:

$$Att(\mathbf{q_{n_i}}, \mathbf{k_{n_i}}, \mathbf{v_{n_i}}) = \text{Softmax}(\frac{\mathbf{q_{n_i}} \mathbf{k_{n_i}}^T}{\sqrt{d_k}})\mathbf{v_{n_i}}. \tag{2}$$

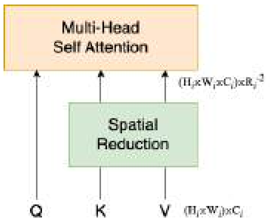

Multi-Head Self Attention
$(H_i \times W_i \times C_i) \times R_i^{-2}$
Spatial Reduction
Q
K
V
$(H_i \times W_i) \times C_i$

Fig. 3: Visualization of Efficient Self Attention [25].

SegFormer replaces the positional encoding (PE) used in ViT with Mix-FFN, as PE is not necessary for semantic segmentation tasks. Mix-FFN combines a 3×3 convolution and a multi-layer perceptron (MLP) with GELU activation function for each FFN. The Mix-FFN can be represented by the following equation:

$$\mathbf{x}_{\text{out}} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}\mathbf{x}_{\text{in}}))) + \mathbf{x}_{\text{in}}, \tag{3}$$

In this equation, x_{in} is the feature map from the efficient multi-head attention module, $\text{Conv}_{3\times3}(\cdot)$ represents a 3×3 convolution, and MLP$(\cdot)$ indicates a multi-layer perceptron with GELU activation function. SegFormer's all-MLP decoder, known for its lightweight design, operates through four distinct phases. To begin, an MLP layer standardizes the channel dimensions of multi-layer features (F_i) from each MiT decoder. Next, these standardized features undergo a 4x upsampling process before being merged. The combined features are then integrated using an additional MLP layer. In the final step, the ultimate segmentation mask is generated through the application of a concluding MLP layer.

III. Proposed Method

The objective of this study is to employ deep learning techniques to differentiate between source and target areas in CMIF. This task can be classified as an image segmentation problem. Among the various image segmentation models available, SegFormer emerges as a particularly efficacious transformer-based approach, demonstrating a superior IoU rate compared to its counterparts. In this investigation, the

SegFormer model is adapted to effectively distinguish source and target regions within CMIF.

Figure 4 delineates the research workflow employed in this study. The first stage of this research is preprocessing the input image. The input image is preprocessed by normalizing and resizing it to $512\times512\times3$. The next step is to divide the image dataset into training, validation and test images. The last stage is to train the SegFormer model using the training data that has been prepared beforehand so that the trained model can discriminate the source and target regions.

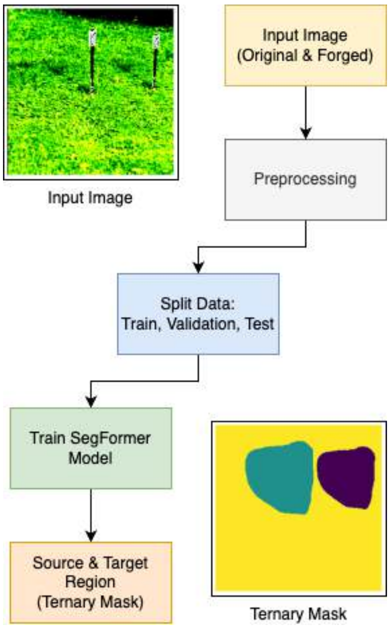


Fig. 4: The Reseach Workflow.

The initial step entails resizing the input images to dimensions of $512 \times 512 \times 3$ and normalizing them. Both authentic and manipulated images that have undergone preprocessing serve as input for the SegFormer model. Subsequently, the image dataset is partitioned into separate sets for training, validation, and testing purposes.

The SegFormer model processes the input image, which is initially transmitted through the SegFormer encoder. This encoder comprises 4 hierarchically arranged transformer blocks, each generating a distinct feature representation. This architecture enables SegFormer to produce multi-level feature representations. In the context of CMIF, models capable of generating multi-level features exhibit superior performance in differentiating source and target regions across various scales compared to models utilizing deeper convolutional layers [15].

The SegFormer encoder generates multi-level features that are subsequently processed by Multi Layer Perceptron (MLP) blocks. These MLP blocks consolidate and integrate feature maps from each level, resulting in the generation of a predicted mask. The accuracy of this predicted mask is evaluated utilizing a cross entropy loss function, which quantifies the discrepancy between the predicted mask and the groundtruth mask.

In this investigation, the SegFormer decoder's classification head is modified to incorporate three categories: source, target, and pristine region. The revised classifier head integrates a softmax activation function and implements a 4x upsampling. These modifications are executed to generate a high-performance output mask that differentiates between source, target, and pristine regions across three classes. Visualization of the modified SegFormer classifier used in this research is depicted in Figure 5.

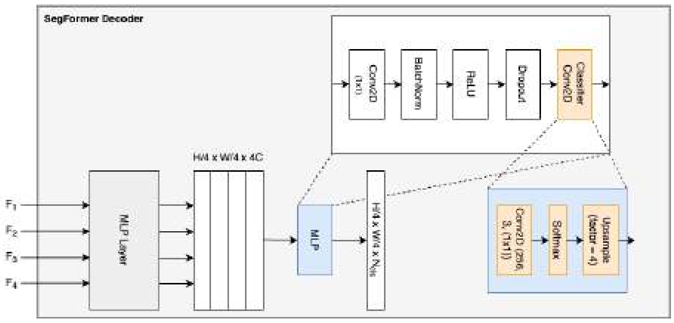


Fig. 5: Modified SegFormer's classifier.

A. Evaluation Metrics

This research utilizes various evaluation metrics, including precision, recall, F1-score, and Intersection over Union (IoU). The assessment is performed for each category of source, target, and pristine area. The calculations for precision, recall, and F1-score are based on the following equation:

$$\mathrm{p} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{4}$$

$$\mathrm{r} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{5}$$

$$\mathrm{F}_1 = 2 \times \frac{\mathrm{p}\times\mathrm{r}}{\mathrm{p}+\mathrm{r}}, \tag{6}$$

Intersection over Union (IoU) is an evaluation metric that quantifies the accuracy of image segmentation models. IoU is computed using the following formula:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{7}$$

B. Experimental Setup

This research employs the CoMoFoD dataset [26], which contains 200 groups of original images and 200 groups of images altered using copy-move forgery. The manipulations involve translation, rotation, scaling, distortion, and various combinations of these techniques. Both original and forged images undergo post-processing, resulting in a total of 10,000 images: 5,000 originals and 5,000 forgeries. For this study, the CoMoFoD dataset is divided into three subsets: 80% for training, 10% for validation, and 10% for testing.

This research employed various SegFormer models, including SegFormer-B0 through SegFormer-B5. The SegFormer model trained from scratch. We also finetune using pretrained models[1]. Adam optimization was used with a $6e-5$ learning rate. The training process lasted 100 epochs, utilizing a batch size of 64. To provide a basis for comparison, DOA-GAN[2] [16] were also trained and tested using the same dataset.

IV. Result and Discussion

Among the various SegFormer models, the pretrained B5 variant exhibits the best performance in terms of train and validation losses with mIoU score of 64.94% on average of 3 classes (source, target and pristine region). However, from the 5th to the 100th epoch, both loss values began to converge while remaining notably high. This trend indicates that the SegFormer model might not be learning effectively. Several factors could account for this suboptimal learning process, including a lack of diversity in the data, an insufficient dataset size, or an overly complex model architecture.

Unlike SegFormer, DOA-GAN show reduced train and validation loss values. However, the DOA-GAN model exhibits unstable train and validation loss patterns. This instability can be attributed to the separate objective functions of the generator and discriminator, which leads to a slower convergence rate in the GAN-based model compared to other deep learning approaches. Figure 6 illustrates the train and validation loss trends for the SegFormer and DOA-GAN models.

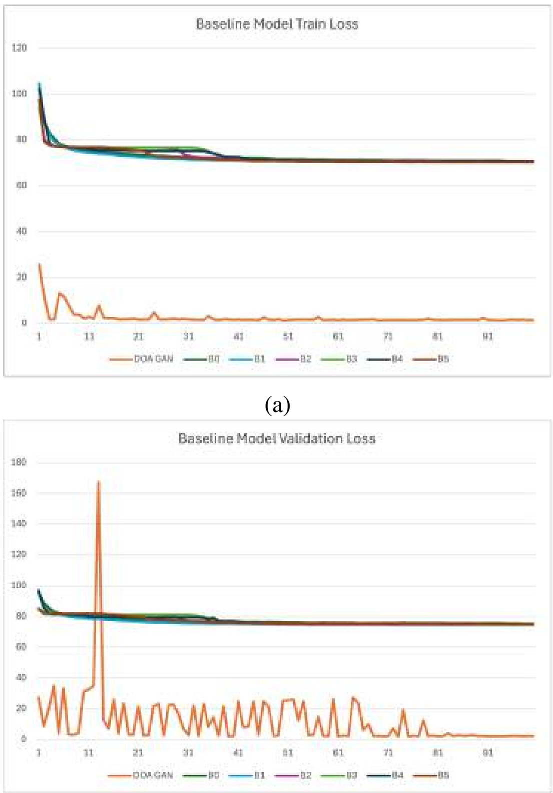
The performance of SegFormer and DOA-GAN models in distinguishing source, target, and pristine region at the pixel level is shown in Table I. DOA-GAN outperforms other models in identifying the target region with F1 score of 65.74%. The SegFormer model proves effective in identifying source and pristine regions with F1 score of 69.95% and 6.02% on CoMoFoD dataset.

The mean Intersection over Union (mIoU) scores for SegFormer trained from scratch and finetune using pretrained ADE20K and other state-of-the-art models on the CoMoFoD dataset are presented in Table II. Among trained from scratch model, SegFormer-B2 achieves mIoU of 35.03% in average of 3 classes (source, target and pristine region). SegFormer-B2 demonstrates superior performance in terms of mIoU for the source region of 53.24% while trained from scratch. SegFormer-B0 achieves 3.13% of mIoU score in pristine region and DOA-GAN acieves mIoU score of 48.96% in target region.

Among finetune with pretrained SegFormer Model using ADE20K dataset, SegFormer-B5 achieves best mIoU score of 64.94% of all classes. Almost all types of SegFormer model outperforms DOA-GAN in terms of mIoU using from scatch training (SegFormer B1-B5) and also finetune pretained model (SegFormer B0, B3-B5). Additionally, DOA-GAN exhibits the most favorable performance with the highest mIoU for the source region of 34.32%.

V. Conclusion

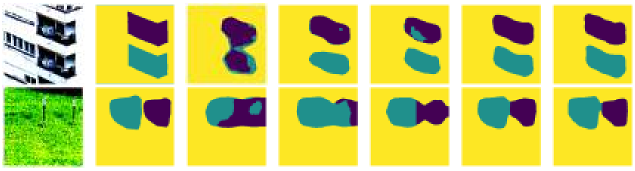
This research introduces the application of SegFormer for identifying source and target region in CMIF imagery. The SegFormer model on the CoMoFoD dataset, effectively distinguishes between source and target regions in CMIF. The experimental outcomes reveal that SegFormer outperforms current state-of-the-art models in terms of IoU when differentiating source and target region. The SegFormer model show promising results in identifying source and target region in CMIF. Future research aims to enhance SegFormer's capabilities to perform multiple task, including localize duplicated regions and discriminate source and target region.



(a)

(b)

Fig. 6: (a) Train Loss and (b) Validation Loss baseline model on CoMoFoD Dataset.



(a) (b) (c) (d) (e) (f) (g)

Fig. 7: Visualization of SegFormer model discrimination results on selected sample CoMoFoD datasets. (a) Forged image, (b) Ternary mask, (c) SegFormer B0, (d) SegFormer B1, (e) SegFormer B2, (f) SegFormer B3, (g) SegFormer B4.

References

[1] C. W. D. Lumoindong, M. A. Aryadi, I. T. Wilyani, and A. Suhartomo, "Effectiveness of Probabilistic Image Sampling Techniques to Identify Hoax-related Images in Indonesia," Int. J. Innov. Technol.

[1] https://github.com/NVlabs/SegFormer

[2] https://github.com/asrafulashiq/doagan_clean

TABLE I: Baseline Model Performance on CoMoFoD dataset

Model	**Source**			**Target**			**Pristine**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
SegFormer B0	0.5233	0.8940	0.6602	0.4882	0.9982	0.6557	0.9734	0.0310	0.0602
SegFormer B1	0.5428	0.9382	0.6877	0.4885	0.9987	0.6561	0.9803	0.0308	0.0597
SegFormer B2	0.5653	0.9171	0.6995	0.4877	0.9990	0.6555	0.9819	0.0306	0.0593
SegFormer B3	0.5634	0.9039	0.6941	0.4876	0.9990	0.6554	0.9844	0.0306	0.0593
SegFormer B4	0.5755	0.8563	0.6884	0.4866	0.9991	0.6545	0.9856	0.0299	0.0581
SegFormer B5	0.5590	0.9296	0.6985	0.4880	0.9990	0.6557	0.9845	0.0304	0.0590
DOA-GAN	0.5377	0.9117	0.6765	0.4897	0.9997	0.6574	0.9960	0.0310	0.0601

TABLE II: Mean Intersection over Union (mIoU) on CoMoFoD Dataset

Model	**Source**	**Target**	**Pristine**	**mIoU** (Scratch)	**mIoU** (Finetune)
SegFormer B0	0.4991	0.4877	**0.0313**	0.3394	0.3744
SegFormer B1	0.5193	0.4881	0.0312	0.3462	0.2904
SegFormer B2	**0.5324**	0.4874	0.0310	**0.3503**	0.3367
SegFormer B3	0.5266	0.4873	0.0309	0.3483	0.5566
SegFormer B4	0.5173	0.4862	0.0302	0.3446	0.6346
SegFormer B5	0.5316	0.4877	0.0309	0.3500	**0.6494**
DOA-GAN	0.5087	**0.4896**	0.0312	0.3432	0.3432

Explor. Eng., vol. 9, no. 3S, pp. 125–131, Feb. 2020, doi: 10.35940/ijitee.C1029.0193S20.

[2] E. M. Bik, A. Casadevall, and F. C. Fang, “The prevalence of inappropriate image duplication in biomedical research publications,” MBio, vol. 7, no. 3, 2016, doi: 10.1128/mBio.00809-16.

[3] W. Li et al. ”Mat: Mask-aware transformer for large hole image inpainting.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[4] C. Yan, S. Li, and H. Li, ‘TransU 2 -Net: A Hybrid Transformer Architecture for Image Splicing Forgery Detection’, IEEE Access, vol. 11, pp. 33313–33323, 2023, doi: 10.1109/ACCESS.2023.3264014.

[5] S. Tinnathi and G. Sudhavani, ‘An efficient copy move forgery detection using adaptive watershed segmentation with AGSO and hybrid feature extraction’, Journal of Visual Communication and Image Representation, vol. 74, Jan. 2021, doi: 10.1016/j.jvcir.2020.102966.

[6] L. Darmet, K. Wang, and F. Cayre, ‘Disentangling copy-moved source and target areas’, Applied Soft Computing, vol. 109, Sep. 2021, doi: 10.1016/j.asoc.2021.107536.

[7] A. Kashyap, B. Suresh, and H. Gupta, ‘Robust Detection of Copy-Move Forgery Based on Wavelet Decomposition and Firefly Algorithm’, Computer Journal, vol. 65, no. 4, pp. 983–996, Apr. 2022, doi: 10.1093/comjnl/bxaa137.

[8] Y. Aydin, ‘Comparison of color features on copy-move forgery detection problem using HSV color space’, Australian Journal of Forensic Sciences, vol. 56, no. 3, pp. 294–310, 2024, doi: 10.1080/00450618.2022.2157046.

[9] B. Soni, P. K. Das, and D. M. Thounaojam, ‘Geometric transformation invariant block based copy-move forgery detection using fast and efficient hybrid local features’, Journal of Information Security and Applications, vol. 45, pp. 44–51, 2019, doi: 10.1016/j.jisa.2019.01.007.

[10] X. Tian, G. Zhou, and M. Xu, ‘Image copy-move forgery detection algorithm based on ORB and novel similarity metric’, IET Image Processing, vol. 14, no. 10, pp. 2092–2100, Aug. 2020, doi: 10.1049/iet-ipr.2019.1145.

[11] C. Wang, Z. Huang, S. Qi, Y. Yu, G. Shen, and Y. Zhang, ‘Shrinking the Semantic Gap: Spatial Pooling of Local Moment Invariants for Copy-Move Forgery Detection’, IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1064–1079, 2023, doi: 10.1109/TIFS.2023.3234861.

[12] X. Y. Wang, S. Li, Y. N. Liu, Y. Niu, H. Y. Yang, and Z. li Zhou, ‘A new keypoint-based copy-move forgery detection for small smooth regions’, Multimedia Tools and Applications, vol. 76, no. 22, pp. 23353–23382, Nov. 2017, doi: 10.1007/s11042-016-4140-5.

[13] Y. Wu, W. Abd-Almageed, and P. Natarajan, ’BusterNet: Detecting copy-move image forgery with source/target localization’, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 170-186.

[14] A. Diwan, S. Member, and A. K. Roy, ‘CNN-Keypoint Based Two-Stage Hybrid Approach for Copy-Move Forgery Detection’, IEEE Access, vol. 12, no. March, pp. 43809–43826, 2024, doi: 10.1109/ACCESS.2024.3380460.

[15] S. Weng, T. Zhu, T. Zhang, and C. Zhang, ‘UCM-Net: A U-Net-like tampered-region-related framework for copy-move forgery detection’, IEEE Transactions on Multimedia, vol. PP, pp. 1–14, 2023, doi: 10.1109/TMM.2023.3270629.

[16] A. Islam, C. Long, A. Basharat, and A. Hoogs, ‘DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image Copy-Move Forgery Detection and Localization’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2020, pp. 4675–4684. doi: 10.1109/CVPR42600.2020.00473.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ‘Rethinking Atrous Convolution for Semantic Image Segmentation’, Dec. 05, 2017, arXiv: arXiv:1706.05587. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/1706.05587

[18] Y. Zhang et al., ‘CNN-Transformer Based Generative Adversarial Network for Copy-Move Source/ Target Distinguishment’, IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 5, pp. 2019–2032, May 2023, doi: 10.1109/TCSVT.2022.3220630.

[19] S. Chang, ‘Can Deep Network Balance Copy-Move Forgery Detection and Distinguishment?’, May 2023, [Online]. Available: http://arxiv.org/abs/2305.10247

[20] Y. Liu et al., ‘CMFDFormer: Transformer-based Copy-Move Forgery Detection with Continual Learning’, Mar. 10, 2024, arXiv: arXiv:2311.13263. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/2311.13263

[21] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, ‘SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers’, Oct. 28, 2021, arXiv: arXiv:2105.15203. Accessed: Aug. 02, 2024. [Online]. Available: http://arxiv.org/abs/2105.15203

[22] A. Dosovitskiy et al., ‘An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale’, International Conference on Learning Representations, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[23] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, ‘Feature Pyramid Networks for Object Detection’, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.

[24] X. Chu et al., ‘Twins: Revisiting the Design of Spatial Attention in Vision Transformers’, Sep. 29, 2021, arXiv: arXiv:2104.13840. Accessed: Aug. 29, 2024. [Online]. Available: http://arxiv.org/abs/2104.13840

[25] W. Wang et al., ‘Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions’, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 548–558. doi: 10.1109/ICCV48922.2021.00061.

[26] D. Tralic, I. Zupancic, S. Grgic, and M. Grgic, ‘CoMoFoD - New database for copy-move forgery detection’, 2013. [Online]. Available: http://www.vcl.fer.hr/comofod.