SCCA-Net: A Novel Network for Image Manipulation Localization Using Split-Channel Contextual Attention*

Yan Xiang[1][0009−0007−0607−0762], Kaiqi Zhao[2,*][0009−0009−3436−2249], and Haichang Yin[1][0009−0000−3512−7739]

[1] Guangdong Polytechnic of Science and Technology, Guangdong 519090, China;
2009853gia30006@student.must.edu.mo;
2009853gia30003@student.must.edu.mo
[2] School of Cyberspace Security, Shandong University of Political Science and Law, Shandong, 250014, China;
2009853dii30002@student.must.edu.mo

Abstract. This paper introduces SCCA-Net, an advanced end-to-end network designed specifically for Image Manipulation Localization (IML). SCCA-Net comprises four critical modules: Split-Channel Contextual Attention (SCCA), Extractor, Encoder, and Decoder. The SCCA module fuses dynamic frequency contextual features and similarity features extracted from RGB feature pyramids of images, overcoming the common shortcomings of existing attention-based IML technologies that typically overlook the frequency adaptation of contextual information. The SCCA module's pivotal component, the Parallel Dynamic Frequency Aggregator (PDFA), integrates Parallel Low-pass (PL) and Similarity Attention (SA) blocks to merge contextual and similarity vectors. The Extractor produces an RGB feature pyramid, channeling varied frequency features into the SCCA. The Encoder, utilizing Transformer, establishes a robust global feature representation. To reconstruct the predicted mask, the Decoder employs uniquely designed cascaded upsampling-convolution (Up-Conv) blocks. Rigorous testing demonstrates that SCCA-Net surpasses conventional models, achieving F1 score improvements of +14.3% on Coverage and +11.8% on CASIA, matching top performances on NIST2016. SCCA-Net pushes the field's boundaries and redefines the benchmarks for assessing IML.

Keywords: Split-Channel Contextual Attention (SCCA) · Parallel Dynamic Frequency Aggregator (PDFA) · Image Manipulation Localization (IML).

* This work was supported by the Key Scientific Research Project of Ordinary Universities in Guangdong Province (Grant No. 2022ZDZX4075), Guangdong Provincial Innovation Team Project for Ordinary Higher Education Institutions (Grant No. 2023KCXTD062), and Key task projects of rural science and technology commissioners in Guangdong Province(Grant No. KTP20240638 and KTP20240406).

1 Introduction

Due to the widespread availability of digital image manipulation software, photographs are increasingly susceptible to undetectable changes. Image enhancement operations, such as blurring, contrast enhancement, and scaling, are widely used to improve the visual appeal of images without altering their semantic representation. However, content editing manipulations pose significant social risks, particularly in finance, insurance, legal forensics, and journalism. It is crucial to develop robust models capable of detecting manipulated images, often invisible to human visual inspection.

Traditional techniques for IML rely on passive forensic methods, including overlapping patches [9], Scale-Invariant Feature Transforms (SIFT) [3], Speed-Up Robust Features (SURF) [16], Camera Filter Arrays (CFA) [17], pattern noise-based [15], and statistical feature analyses [19]. Although these methods are effective against certain image manipulations, the emergence of more complex and composite image tampering techniques often involves multiple operations. The applicability of traditional detection methods in real-world tampering scenarios has been reduced.

In recent years, deep neural networks (DNN) have excelled in dense prediction tasks in computer vision (CV), including object detection and semantic segmentation. These tasks are closely related to the goals of IML, as they all involve identifying and locating specific regions in images. Motivated by these similarities, numerous researchers have applied CV techniques to IML. CNN-based [27], object detection-based [13], and semantic segmentation-based methods [24] have been successfully applied to IML. Recently, with the success of Vision Transformers (ViT) [8] in the CV field, researchers have begun exploring their potential in IML, such as [14]. While DNN and ViT [8] excel in CV, it is important to note that their targets (such as object recognition, semantic segmentation, and pose recognition) may not fully align with the specific objectives of IML. In these tasks, targets typically have particular shapes or sizes along with semantics, whereas in IML tasks, the targets (manipulated regions) often have irregular sizes and shapes with semantic-agnostic.

In addressing the challenges of IML, the primary strategies include: (1) extracting superficial texture features using pre-processing blocks, as demonstrated in [4, 28]; (2) enhancing tampering traces with forensic attention modules, as cited in [13, 24, 28]; and (3) employing supervised contrastive learning to highlight feature discrepancies between tampered regions and backgrounds, as indicated in [26]. Specific methods include SATFL [28], which combines spatial and channel attention features to enhance forensic differentiation. TDA-Net [13] introduces tampering discriminative attention, focusing on the image's region of interest. SAPS-Net [24] employs a semantic-agnostic operational attention module, extracting multi-scale semantic features via parallel convolution, and then using residual structures and subtraction operations to reduce the impact of image semantics on detection. These methods focus on local feature similarity and ignore inter-pixel contextual features, leading to suboptimal effects.

In response to prevailing challenges, we develop an efficient end-to-end network architecture for IML, denoted as SCCA-Net. Our key contributions are threefold: (1) We introduce an innovative end-to-end IML network depicted in Fig. 1, which utilizes a Split-Channel Contextual Attention (SCCA) module within an encoder-decoder framework to detect multi-scale forged regions accurately. (2) To our knowledge, SCCA is the pioneering attention mechanism that merges similarity features with adaptive frequency contextual features, overcoming the limitations of restricted contextual features in existing attention-based IML techniques. It efficiently fuses these features through a Parallel Dynamic Frequency Aggregator (PDFA) block. (3) Our approach is validated by comprehensive experiments across three benchmark datasets, demonstrating its superior efficacy. The paper is formed as follows: Sec. 2 presents the related works, Sec. 3 introduces our proposed method, and Sec. 4 describes the details of experimental settings, the performance evaluations, and comparisons with existing methods. Finally, we conclude the conclusion in Sec. 5.

2 Related Work

2.1 Image Manipulation Localization

Image manipulation techniques involving content changes are generally categorized into Copy-Move, Splicing, and Removal. Previous studies have focused on detecting and localizing these three types of manipulations. ManTraNet [22] combines multi-task learning to simultaneously perform manipulation detection and boundary localization by focusing on anomalous features, but it has poor performance. DFCN [27] is an encoding-decoding network utilizing dense connectivity in its convolutional network architecture, and it was the first to introduce a strategy for generating large-scale training data via Photoshop Scripting. MVSS-Net [6] proposes a multi-view supervised architecture for comprehensive feature analysis at various scales and perspectives, using the Sobel operator in IML for the first time. However, all three algorithms are not sensitive to Copy-Move manipulations. TDA-Net [13] operates as a tri-stream network with a tampering discriminative attention module focused on object detection, comprising RGB, local, and resampled feature streams merged through ROI pooling. SATFL [28] is a coarse-to-fine prediction network with a forensic attention module. Its experimental results are based on a two-stage training of specific datasets using a self-adversarial training strategy. SAPS-Net [24] designs semantic-agnostic manipulation attention that integrates pyramid features with parallel architecture, making it insensitive to Copy-Move manipulations. EMT-Net [14] focuses on residual approaches for highlighting edge characteristics. Recently, researchers have applied contrastive learning techniques to IML. NCL-IML [26] introduces a non-mutually exclusive contrastive learning method different from traditional techniques. It establishes a dual-branch pivot structure that continually toggles the role of contour patches in forged areas between positive and negative. Although these approaches are currently experimental and have not yielded optimal results, they offer insights into unsupervised learning practices in IML.

2.2 Vision Transformer

Transformer [20] is the state-of-the-art method for natural language processing (NLP) tasks. ViT [8] initially projects images onto sequences of flat patches, which are then fed into a Transformer encoder for classification purposes. SegFormer [23] introduces a novel hierarchical transformer encoder that generates multi-scale features without positional encoding, avoiding complex decoders. Transattunet [5] integrates Transformer self-attention with global spatial attention modules, boosting its capacity to manage and learn non-local feature interactions. We employ ViT [8] to encode RGB data into a robust, potent flattened representation.

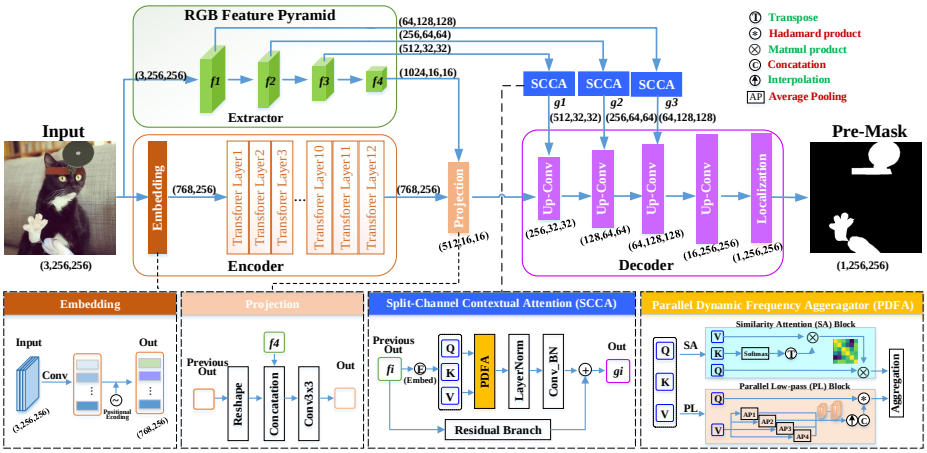

Fig. 1. Overview of the proposed SCCA-Net. SCCA-Net encompasses four core modules: SCCA, Extractor, Encoder, and Decoder. Initially, the input image is processed through the Embedding block and then routed to the Encoder. The Encoder's output is transferred to the Decoder via a Projection. The details of these operations are highlighted in the annotated dotted boxes. The SCCA module enhances tampering traces from the RGB feature pyramid, addressing the texture detail loss commonly observed during upsampling in the Decoder. The blue-head section provides a detailed explanation of the SCCA module's mechanism, while the yellow-head dashed box elaborates on the PDFA's details.

3 Proposed Method

3.1 Overview

To achieve precise pixel-level localization of tampered regions in images, we introduce an end-to-end encoder-decoder architecture comprising four main modules: SCCA, Extractor, Encoder, and Decoder. Fig. 1 delineates this framework, with

each module vividly color-coded for visual clarity. Initially, the input image I of dimensions $\mathbb{R}^{3\times W\times H}$ is resized to $\mathbb{R}^{3\times 256\times 256}$. The Extractor constructs a multi-scale RGB feature pyramid via ResNet50 [10]. Following this, the SCCA module refines these features through the Parallel Dynamic Frequency Aggregator (PDFA), incorporating a Parallel Low-pass (PL) block for capturing multi-scale contextual features and a Similarity Attention (SA) block for self-attention operations. It further enhances the Decoder's capability by transmitting crucial boundary and tampering details via lateral connections. The Encoder transforms RGB information into a flattened patch representation $\mathbb{R}^{256\times 768}$. Ultimately, the Decoder employs cascaded Upsampling-Convolution (Up-Conv) blocks to reconstruct the depth representation and predict the binary mask $\mathbb{R}^{1\times 256\times 256}$.

3.2 Split-Channel Contextual Attention

Motivation: IML tasks pose a sophisticated pixel-level classification challenge primarily due to irregular sizes and shapes of tampered areas, which often lack semantic content. Forgery traces consist of texture information from local features and inter-pixel contextual information from global features. Forensic Attention is often utilized to enhance these fragile tampered clues. Current attention-based IML approaches predominantly focus on local spatial or channel similarities, or both, neglecting the integration of contextual features, leading to focus inaccuracies. To address these limitations, we propose the SCCA module to enhance the forgery artifacts in the images.

Components: The core component of SCCA is the PDFA, which integrates dual-pathway PL and SA blocks to fuse contextual and similarity vectors. As shown in the blue-head dashed box in Fig. 1, the RGB feature pyramid $f_i \in \mathbb{R}^{c\times h\times w}$ extracted by the Extractor is initially linearly projected into vectors Q, V, and K ($Q, K, V \in \mathbb{R}^{n\times c}$, where $n = h \times w$). The internal mechanism of PDFA is further illustrated in the yellow-head dashed box in Fig. 1. The PL block extracts multi-scale low-frequency component features from V using a cluster of parallel average pooling, and the Hadamard product with Q generates the multi-scale contextual vector. The SA block employs a multi-head self-attention mechanism to extract a multi-frequency similarity vector, and the two vectors are finally aggregated by element-wise addition. The PDFA's formulation is given in Eq. (1).

$$\begin{cases} Q, K, V = \boldsymbol{E}(f_i) \\ \boldsymbol{D}_{PDFA}(Q, K, V) = \boldsymbol{D}_{PL}(Q, V) + \boldsymbol{D}_{SA}(Q, K, V) \end{cases} \tag{1}$$

Where $\boldsymbol{D}_{PDFA}(\cdot)$, $\boldsymbol{D}_{PL}(\cdot)$, $\boldsymbol{D}_{SA}(\cdot)$, and $\boldsymbol{E}(\cdot)$ respectively signify the PDFA, PL block, SA block, and linear embedding operations, as defined in Eq. (2), (3) and (4). The sign f_i represents the RGB feature pyramid produced by the Extractor. The symbol '+' indicates element aggregation. Q, K, and V symbolize the Query, Key, and Value vectors, aligned with Eq. (2) and (3). Notably, these

operators utilize a parallel structure, reducing computational costs by sharing weights.

Global features elucidate inter-pixel contextual connections, predominantly present in low-frequency components. Through convolution, ResNet50 [10] extracts an RGB feature pyramid enriched with local textual details and various frequency components. Our goal is to select and amplify frequency components crucial for forgery analysis. To accomplish this, average pooling acts as a low-pass filter. However, the cutoff frequencies differ across images. To address this, we configure various kernel sizes and strides within a Parallel Average Pooling Cluster, named the Parallel Low-pass (PL) block. This block employs kernel sizes of $\{1 \times 1,\ 3 \times 3,\ 5 \times 5,\ 7 \times 7\}$, similar to multiple low-pass filters with different cutoff bandwidths, achieving frequency adaptation. Initially, V is projected to $\mathbb{R}^{c \times h \times w}$ and then split into four sub-features ($\mathbb{R}^{\frac{c}{4} \times h \times w}$) along the channels, which are used as inputs for four average pooling operations to extract inter-pixel contextual information. The outputs of the four average pooling operations vary in size and are interpolated to a uniform resolution of feature maps, then mapped to the dimension of $\mathbb{R}^{n \times c}$ ($n = h \times w$) after concatenation to produce the dynamic frequency contextual vector. Lastly, the inner product between Q and the dynamic frequency contextual vector strengthens the contextual information within images. To optimize the model size and computational complexity, the PL and SA blocks share inputs (Q, K, V) along with the parameters of their respective linear layers. The PL block's operational specifics are outlined in Eq. (2).

$$\boldsymbol{D}_{PL}(Q,V) = Q \cdot (\boldsymbol{T}^{'}(\boldsymbol{B}(\boldsymbol{P}_{s\times s}(\boldsymbol{T}(V))))) \tag{2}$$

where $\boldsymbol{B}(\cdot)$ refers to bi-linear interpolation, and $\boldsymbol{P}_{s\times s}(\cdot)$ denotes the Parallel Average Pooling Cluster with various kernel sizes, subscript $s \times s \in \{1 \times 1,\ 3 \times 3,\ 5 \times 5,\ 7 \times 7\}$. $\boldsymbol{T}(\cdot)$ is used for the translation operation to align dimensions, whereas $\boldsymbol{T}^{'}(\cdot)$ is its inverse.

Concurrently, we have developed the SA block focused on the similarity of the RGB feature pyramid, generally implemented via multi-head linear self-attention. The formulation of the SA block is defined as Eq. (3).

$$\boldsymbol{D}_{SA}(Q,K,V) = Q \times \frac{\boldsymbol{S}(K)^T \times V}{\sqrt{d}} \tag{3}$$

where $\boldsymbol{S}(\cdot)$, d, and $\times$ represents Softmax, dimensions of Q, K, and V vectors, and matmul product respectively.

After merging frequency similarity and the dynamic frequency contextual vector, they are normalized using LayerNorm and transformed by Conv-BN into feature maps that match the input size. Additionally, we retain RGB texture features, integrating them into a residual branch for the Decoder. The SCCA's formulation is detailed in Eq. (4).

$$\boldsymbol{g}_i = \boldsymbol{Conv}(\boldsymbol{LN}(\boldsymbol{D}_{PDFA}(\boldsymbol{E}(f_i))) + \boldsymbol{R}(f_i) \tag{4}$$

where $\boldsymbol{g}_i$, f_i, $\boldsymbol{Conv}(\cdot)$, $\boldsymbol{LN}(\cdot)$, and $\boldsymbol{R}(\cdot)$ represents the output of SCCA, RGB feature pyramid, Conv-BN, LayerNorm, and Residual Branch respectively.

3.3 Extractor

Our model incorporates hybrid ResNet50 [10] as its Extractor, retaining the original's 50 convolution layers while adapting each stage to suit our task-specific requirements better. This customization optimizes output resolution and channel counts. Specifically, the Extractor generates feature maps $f_1 \in \mathbb{R}^{64\times128\times128}$ from the Stem Layer before the max-pooling operation. The subsequent outputs f_i $(i = 2, 3, 4)$ emerge from Bottleneck layers, organized in sequences of $\{3, 4, 9\}$. Each feature map's resolution decreases by half sequentially, while the channel count progressively expands following the sequence $\{256, 512, 1024\}$. This results in the RGB Feature Pyramid that showcases higher resolution and shallower dimensionality.

3.4 Encoder and Decoder

Exploiting their inherent designs, CNNs are proficient in extracting local features, whereas Transformer [20] efficiently manages long-range dependencies. Following ViT [8], we project images into globally flattened patches, as illustrated in Fig. 1. Initially, the Embedding block converts RGB features $\boldsymbol{I} \in \mathbb{R}^{3\times256\times256}$ into vectorized patches $\boldsymbol{p}_0 \in \mathbb{R}^{256\times768}$, uniformly sized at 16×16. Deviating from standard ViT serialization, this method excludes the cls-token and incorporates cosine position codes $\boldsymbol{p}_{\cos} \in \mathbb{R}^{256\times768}$ to preserve spatial information in tampered areas.
In the decoder, encoded features $\boldsymbol{p}_{12} \in \mathbb{R}^{256\times768}$ are combined with the high-dimensional RGB feature pyramid $f_4 \in \mathbb{R}^{1024\times16\times16}$, and transformed into $\boldsymbol{x'} \in \mathbb{R}^{512\times16\times16}$ through Transpose, Reshape, and Conv-Bn operations. Up-Conv blocks feature Up-sampling-Concatenation and Conv-Bn. To mitigate texture and location detail loss during decoding, outputs from the SCCA module are integrated as lateral connections to enhance feature localization. These connections specifically highlight features sensitive to discontinuities at the boundaries in the final binary mask predictions.

4 Experiments and Analysis

To showcase our model's versatility, we evaluated its performance on CASIA [7], Coverage [21], and NIST2016 [2], comparing it against state-of-the-art methods using a defined experimental protocol.

4.1 Datasets

To ensure fair comparisons, our experiments use the same training and testing splits as the SOTA competitors [28, 13, 14, 24] for the three benchmark datasets.

CASIA [7] known for Copy-move and Splicing, comprises 920 tampered samples in V1.0 release and 5,123 in V2.0. Each image, manipulated using Adobe Photoshop, includes a corresponding binary ground truth. For our experiments, CASIA V1.0 is designated as the test set, while a subset of V2.0 serves as the training set.
Coverage [21] consists of 100 copy-move tampered images with TIFF format, which contains: migrate, scaling, rotate, distort, brighten, fusion, and other tampering operations. It is split into 75 training and 25 test samples.
NIST2016 [2] is a challenging dataset, that features 564 samples, with 68 copy-move, 288 splicing, and 208 removal images. A training-testing split of 404:160 is used.
PS-Synth Data DNN-based IML models are always limited by the availability of large forgery datasets for training. Inspired by SATFL [28] and DFCN [25], we use the Photoshop Scripting[3] tool to automatically generate 110,000 synthetic tampered samples for network training, including Copy-Move, Splicing, and Removal manipulations. For Splicing, two images are randomly selected, one as the background and the other as the donor. A source region is selected from the donor image and spliced into the background image. The Copy-Move and Removal operations use a single image. In Copy-Move, a source area is selected and copied to another area of the same image. The Removal operation uses Photoshop scripting's content-aware fill to remove and fill the source area, making the tampered area similar to the background. The tampering manipulations use Label-me[4] to obtain label information. Transformations such as distortion, rotation, and scaling are applied to the source area to simulate human-like forgery operations, along with post-processing techniques like blurring and lighting adjustments to disguise further tampering traces. The original images come from the VISION dataset [18], the KCMI dataset [1], and our photo collection.

4.2 Experimental Settings

Evaluation metrics: We follow the benchmark metrics of competitors [12, 15, 17, 28, 13, 14, 24]. Performance is evaluated using pixel-level F1 scores and Area Under the ROC Curve (AUC) for robust comparison against the ground truth, whose values closer to 1 indicate better. The F1 is computed by (5).

$$\boldsymbol{F1} = 2 \times \frac{Presion \times Recall}{Presion + Recall} \tag{5}$$

Implementation Details: Our model training utilizes PyTorch on three NVIDIA GeForce 3090 GPUs, employing the Adam [11] optimizer with a specific learning rate as detailed in Eq. (6). We apply Dice Loss and Binary Cross-Entropy Loss (BCE Loss) to address the size disparity between tampered regions and

[3] https://www.adobe.com/devnet/photoshop/scripting.html
[4] https://github.com/LabelMe/labelme.git

Table 1. The Details of experimental datasets. 'FT Set', 'CM', 'SP', and 'RE' represent Fine-tuned Set, Copy-Move, Splicing, and Removal respectively.

Datasets	Pre-training Set			FT Set			Test Set			Total
	CM	SP	RE	CM	SP	RE	CM	SP	RE	
PS-Synth Data	38,000	40,000	32,000	-	-	-	-	-	-	110,000
Coverage [21]	-	-	-	75	-	-	25	-	-	100
NIST2016 [2]	-	-	-	48	206	150	20	82	58	564
CASIAv1 [7]	-	-	-	-	-	-	459	461	-	920
CASIAv2 [7]	-	-	-	1,342	772	-	-	-	-	2,324

the background. ResNet50 [10] and Transformer [8] start with pre-trained ImageNet1K weights. The training consists of an initial pre-training on PS-Synth Data, followed by fine-tuning on a composite of benchmark datasets, with splits detailed in Tab. 1.

$$\boldsymbol{lr} = \boldsymbol{lr_0} \times (1.0 - \frac{iter}{Iter_{max}})^{0.9}, \quad iter \in [1,\ Iter_{max}] \tag{6}$$

Where $\boldsymbol{lr}$ stands for learning rate, $\boldsymbol{lr}_0$ denotes the initial learning rate, which is set to $1.2e$-4, $Iter_{max}$ is the maximum iteration during training, $iter$ is the number of iterations, ranging from 1 to $Iter_{max}$.

4.3 Baseline

We compared the performance of our method with various classic unsupervised [12, 15, 17] and DNN-based [13, 14, 24, 28] methods.

ELA [12] extracts and interprets image metadata to identify forgery areas by comparing compression levels.

NOI [15] uses wavelet analysis and predicts that the noise in tampered areas will be inconsistent with the background noise.

CFA [17] estimates the statistical characteristics of the camera's internal CFA pattern for each patch in the image and identifies regions with abnormal regions.

SATFL [28] is a coarse-to-fine network with a forensic attention module, whose experimental results are based on two-stage training on specific datasets through the self-adversarial training strategy.

TDA-Net [13] proposes a triple-convolutional-stream deep architecture based on a tampering discriminative attention module. It extracts visual perception, resampling, and local inconsistency features from both spatial and frequency domains, combining them to form hybrid features to distinguish tampered regions from non-tampered parts.

EMT-Net [14] uses a Transformer branch to extract global and local noise features, employs a convolutional branch to capture local visual artifacts, and adopts an edge artifact enhancement module and edge supervision strategy to enhance fused tampering traces for detecting tampered regions in images.

SAPS-Net [24] designs semantic-agnostic progressive attention based on subtractive operations, which captures semantic feature associations through multi-scale convolution iterations and further uses residual structures and subtraction operations to mitigate the influence of rich image semantics on operation extraction.

Table 2. Ablation Study F1 Scores for SCCA Module with Varied Components: The Baseline configuration comprises solely the Encoder and Decoder. Results are presented with the highest scores in bold and the second-highest scores underlined for clarity.

Components	**Datasets**			**Average-F1**
	CASIA [7]	NIST2016 [2]	Coverage [21]	
Baseline	0.242	0.306	0.301	0.283
Baseline+Extractor	0.405	0.344	0.501	0.417
Baseline+Extractor+SA	<u>0.454</u>	0.353	0.505	0.437
Baseline+Extractor+PL	**0.458**	<u>0.370</u>	<u>0.513</u>	<u>0.447</u>
Baseline+Extractor+SCCA	0.449	**0.372**	**0.529**	**0.450**

4.4 Ablation Analysis

Our model leverages the SCCA module, validated through extensive testing. We conducted systematic ablation studies to assess the SCCA module. The initial baseline configuration consists only of the Encoder and Decoder, excluding the Extractor and SCCA modules. The baseline achieves F1 scores of 24.2% on CASIA [7], 30.6% on NIST2016 [2], and 30.1% on Coverage [21], with an average F1 score of 28.3%.

Progressive configurations incrementally integrate additional components: 'Baseline+Extractor' introduces the Extractor; 'Baseline+Extractor+SA' incorporates the Similarity Attention (SA) block; 'Baseline+Extractor+PL' adds the Parallel Low-pass (PL) block, crucial for extracting contextual features. The most comprehensive setup, 'Baseline+Extractor+SCCA', demonstrates marked improvements.

Uniform training across these configurations leads to significant performance gains. The SA block alone improves average F1 scores by 15.4% over the baseline, adding the 'PL' module further increases the average F1 score by 0.9%, and the fully optimized setup with the SCCA module enhances the score by an additional 0.3%, as documented in Tab. 2.

4.5 Visualization

In this section, we visually compare our proposed approach and SATFL[5] [28]. Their performance is tested on three benchmark datasets: Coverage [21], CASIA

[5] https://github.com/tansq/SATFL

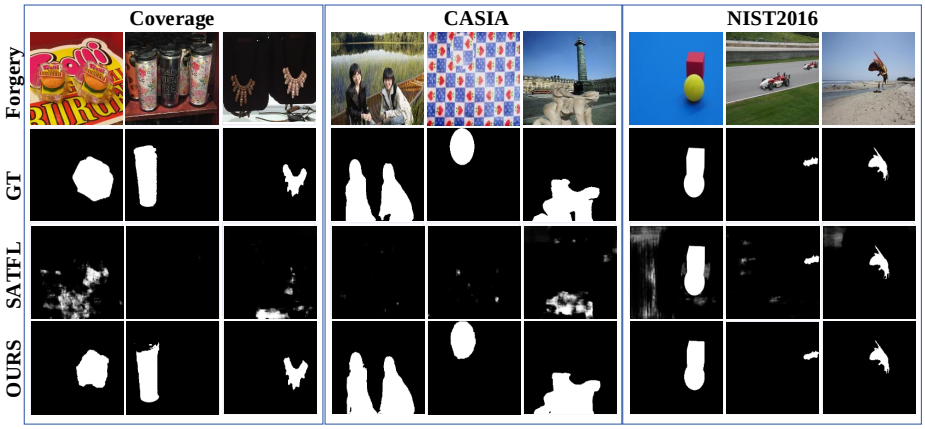

Fig. 2. Visualization of the detected results by different methods. From top to bottom, we show the forged images, GT labels, results of SATFL [28], and our SCCA-Net.

[7], and NIST2016 [2]. Fig. 2 presents the relevant results. It is worth mentioning that all binary masks are directly stored as masks of the model's predicted features using the Pillow API, without applying any post-processing techniques, such as morphological operations or edge enhancement.

Our method consistently excels in localization accuracy, outperforming the competitor. This superior performance is attributed to our integration of the SCCA module. When dealing with the Coverage dataset [21], our method accurately identified Copy-Move forgeries, a task where others faltered, often missing inter-pixel contextual features crucial for spotting similarities in manipulated regions. In contrast, SATFL [28] was less accurate in demarcating tampered boundaries and detecting edge inconsistency artifacts present in manipulated areas. The SCCA module further enhances contextual information processing for increased accuracy. We employ Grad-CAM for visual insights, effectively showcasing SCCA's ability to progressively pinpoint tampered regions (Fig. 3).

4.6 Comparison with the States-of-the-Art Methods

Our research leverages a series of classic unsupervised and influential deep learning algorithms for IML, including ELA [12], NOI [15], CFA [17], SATFL [28], TDA-Net [13], EMT-Net [14], and SAPS-Net [24]. Concise introductions of these comparators are provided in Sec. 4.3. We assess our method's efficacy against these state-of-the-art algorithms by conducting experiments on the combined 'Test Set' of [7, 21, 2], using fine-tuned weights on the 'FT set', detailed in Tab. 1. The results, measured by pixel-level F1 and AUC scores with the ground truth, are detailed in Tab. 3.

ELA [12], NOI [15], and CFA [17] are classic unsupervised methods based on hand-crafted features. The results show that our method significantly outperforms these classic unsupervised methods on CASIA [7], NIST2016 [2], and Cov-

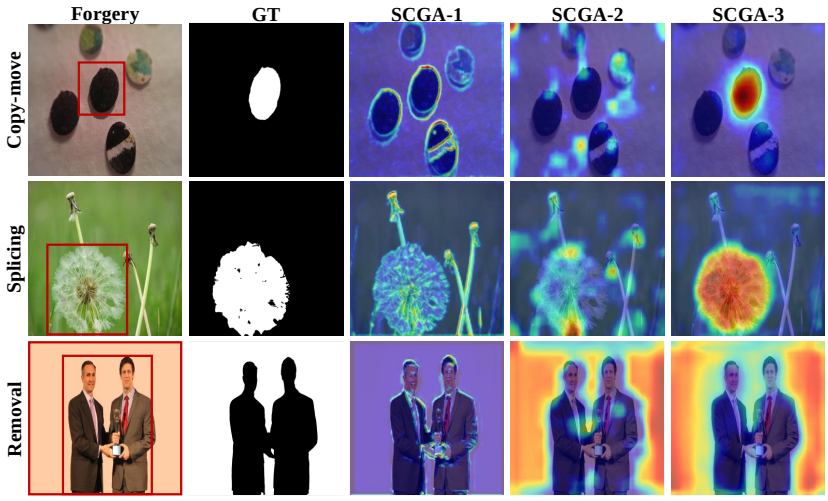

Fig. 3. Visualization of Grad-CAM features of three examples from SCCA (all from NIST2016 [2]).

erage [21]. This is because these classic methods use manually defined detection features that include only a few specific tampering artifacts, resulting in minimal detection performance. Our method also demonstrates substantial gains in F1 and AUC scores compared to DNN-based comparators on three standard datasets, with AUC improvements ranging from 2.9% to 9.7% and F1 scores increasing from 11.8% to 31.6%. Notably, our method surpasses EMT-Net [14] by 35.2% in F1 score on Coverage [21]. Although it ranks second on NIST2016 [2], it closely approaches the highest performance.

Table 3. Comparison with different models in terms of F1, and AUC scores on three standard datasets. The best outcomes are indicated in bold, while the second-best outcomes are underlined for emphasis.

Methods	CASIA [7]		NIST2016 [2]		Coverage [21]		Average	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
ELA [12]	0.214	0.613	0.236	0.429	0.222	0.583	0.224	0.542
NOI [15]	0.263	0.612	0.285	0.487	0.269	0.587	0.272	0.562
CFA [17]	0.207	0.522	0.174	0.501	0.190	0.485	0.190	0.503
SATFL [28]	0.384	0.788	0.622	0.943	<u>0.562</u>	0.856	0.523	0.862
TDA-Net [13]	<u>0.582</u>	0.831	0.756	0.948	0.474	<u>0.864</u>	<u>0.669</u>	0.881
EMT-Net [14]	0.459	<u>0.856</u>	0.825	<u>0.987</u>	0.353	0.812	0.546	0.885
SAPS-Net [24]	0.561	0.843	**0.878**	**0.996**	0.507	0.852	0.649	<u>0.897</u>
SCCA-Net	**0.700**	**0.885**	<u>0.852</u>	0.973	**0.705**	**0.938**	**0.752**	**0.932**

5 Conclusion

Our research proposes SCCA-Net, a cutting-edge end-to-end network featuring the SCCA module, a novel forensic attention mechanism designed for accurate pixel-level manipulation localization. Central to our approach, the SCCA module captures dynamic frequency contextual and similarity features from the RGB feature pyramid, significantly enhancing tampering clues and augmenting texture information critical for decoding. The network encodes RGB features via a Transformer, converting them into flattened patch sequences to ensure a comprehensive global representation. The decoder leverages cascaded Up-Conv blocks that adeptly utilize texture features provided by SCCA to reconstruct prediction masks accurately. Empirical tests demonstrate SCCA-Net's superior performance in pinpointing tampered regions across diverse datasets, establishing its innovative edge. However, despite outperforming other models, the model performs poorly on low-resolution, post-processed images, such as CASIA. The inter-pixel contextual information extracted from these tampered images is limited and insufficient to represent tampering artifacts. Tampering traces are often present in high-frequency components. In future work, we will consider combining high-frequency filtering operators with deep learning methods to adaptively filter and enhance high-frequency components that are beneficial for tampering detection and localization.

References

1. Camera model identification, https://www.kaggle.com/c/sp-society-camera-model-identification
2. Nist: Nist nimble 2016 datasets, https://www.nist.gov/itl/iad/mig/
3. Alberry, H.A., Hegazy, A.A., Salama, G.I.: A fast sift based method for copy move forgery detection. Future Computing and Informatics Journal **3**(2), 159–165 (2018)
4. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security **13**(11), 2691–2706 (2018)
5. Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K.: Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Transactions on Emerging Topics in Computational Intelligence pp. 1–14 (2023)
6. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
7. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China summit and international conference on signal and information processing. pp. 422–426. IEEE (2013)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Farid, H., Lyu, S.: Higher-order wavelet statistics and their application to digital forensics. In: 2003 Conference on computer vision and pattern recognition workshop. vol. 8, pp. 94–94. IEEE (2003)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), http://arxiv.org/abs/1412.6980
12. Krawetz, N., Solutions, H.F.: A picture's worth. Hacker Factor Solutions **6**(2), 2 (2007)
13. Li, S., Xu, S., Ma, W., Zong, Q.: Image manipulation localization using attentional cross-domain cnn features. IEEE Transactions on Neural Networks and Learning Systems **34**(9), 5614–5628 (2023)
14. Lin, X., Wang, S., Deng, J., Fu, Y., Bai, X., Chen, X., Qu, X., Tang, W.: Image manipulation detection by multiple tampering traces and edge artifact enhancement. Pattern Recognition **133**, 109026 (2023)
15. Mahdian, B., Saic, S.: Using noise inconsistencies for blind image forensics. Image and vision computing **27**(10), 1497–1503 (2009)
16. Pandey, R.C., Singh, S.K., Shukla, K.K., Agrawal, R.: Fast and robust passive copy-move forgery detection using surf and sift image features. In: 2014 9th International Conference on Industrial and Information Systems (ICIIS). pp. 1–6 (2014)
17. Popescu, A.C., Farid, H.: Exposing digital forgeries in color filter array interpolated images. IEEE Transactions on Signal Processing **53**(10), 3948–3959 (2005)
18. Shullani, D., Fontani, M., Iuliani, M., Shaya, O.A., Piva, A.: Vision: a video and image dataset for source identification. EURASIP Journal on Information Security **2017**(1), 1–16 (2017)
19. Stamm, M.C., Liu, K.R.: Forensic detection of image manipulation using statistical intrinsic fingerprints. IEEE Transactions on Information Forensics and Security **5**(3), 492–506 (2010)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Wen, B., Zhu, Y., Subramanian, R., Ng, T.T., Shen, X., Winkler, S.: Coverage—a novel database for copy-move forgery detection. In: 2016 IEEE international conference on image processing (ICIP). pp. 161–165. IEEE (2016)
22. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9543–9552 (2019)
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
24. Xu, D., Shen, X., Shi, Z., Ta, N.: Semantic-agnostic progressive subtractive network for image manipulation detection and localization. Neurocomputing **543**, 126263 (2023)
25. Zhang, Y., Goh, J., Win, L.L., Thing, V.L.: Image region forgery detection: A deep learning approach. SG-CRC **2016**, 1–11 (2016)
26. Zhou, J., Ma, X., Du, X., Alhammadi, A.Y., Feng, W.: Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22346–22356 (October 2023)
27. Zhuang, P., Li, H., Tan, S., Li, B., Huang, J.: Image tampering localization using a dense fully convolutional network. IEEE Transactions on Information Forensics and Security **16**, 2986–2999 (2021)

28. Zhuo, L., Tan, S., Li, B., Huang, J.: Self-adversarial training incorporating forgery attention for image forgery localization. IEEE Transactions on Information Forensics and Security **17**, 819–834 (2022)