

Can We Get Rid of Handcrafted Feature Extractors?

SparseViT: Nonsemantics-Centered, Parameter-Efficient Image Manipulation Localization through Spare-Coding Transformer

Lei Su^{1,2}, Xiaochen Ma³, Xuekang Zhu^{1,2}, Chaoqun Niu^{1,2}, Zeyu Lei^{1,2,4}, Ji-Zhe Zhou^{1,2*}

¹College of Computer Science, Sichuan University, China

²Engineering Research Center of Machine Learning and industry intelligence, Ministry of Education of China

³Mohamed Bin Zayed University of Artificial Intelligence

⁴Department of Computer and Information Science, University of Macao, Macao SAR

Abstract

Non-semantic features or semantic-agnostic features, which are irrelevant to image context but sensitive to image manipulations, are recognized as evidential to Image Manipulation Localization (IML). Since manual labels are impossible, existing works rely on handcrafted methods to extract non-semantic features. Handcrafted non-semantic features jeopardize IML model’s generalization ability in unseen or complex scenarios. Therefore, for IML, the elephant in the room is: **How to adaptively extract non-semantic features?** Non-semantic features are context-irrelevant and manipulation-sensitive. That is, within an image, they are consistent across patches unless manipulation occurs. Then, spare and discrete interactions among image patches are sufficient for extracting non-semantic features. However, image semantics vary drastically on different patches, requiring dense and continuous interactions among image patches for learning semantic representations. Hence, in this paper, we propose a Sparse Vision Transformer (SparseViT), which reformulates the dense, global self-attention in ViT into a sparse, discrete manner. Such sparse self-attention breaks image semantics and forces SparseViT to adaptively extract non-semantic features for images. Besides, compared with existing IML models, the sparse self-attention mechanism largely reduced the model size (max 80% in FLOPs), achieving stunning parameter efficiency and computation reduction. Extensive experiments demonstrate that, without any handcrafted feature extractors, SparseViT is superior in both generalization and efficiency across benchmark datasets.

Code — <https://github.com/scu-zjz/SparseViT>

Introduction

With the rapid development of image editing tools and image generation technologies, image manipulation has become exceedingly convenient. To address this trend, researchers have developed Image Manipulation Localization (IML) techniques to identify specific manipulated regions within images. Due to the inevitable artifacts (manipulation traces) left on an image after manipulation, these artifacts can be divided into semantic and non-semantic (Semantic-Agnostic) features. Semantic-Agnostic Features refer to features that highlight low-level artifacts information, which

Model Name	Backbone	Extractor	Manual
ManTraNet <small>(CVPR19)</small>	VGG	Bayar+SRM	✓
SPAN <small>(ECCV20)</small>	VGG	Bayar+SRM	✓
MVSS <small>(ICCV21)</small>	ResNet-50	Bayar+Sobel	✓
CAT-Net <small>(IJCV22)</small>	HRNet	DCT	✓
ObjectFormer <small>(CVPR22)</small>	CNN+ViT	DCT	✓
NCL-IML <small>(ICCV23)</small>	ResNet-101	Sobel	✓
TruFor <small>(CVPR23)</small>	ViT	Noiseprint	✓

Table 1: Comparison and Summary of IML Models. We have indicated whether these methods rely on handcrafted feature extraction and specified the type of extractors used.

are independent of the image’s semantic content. These features show significant differences in distribution between manipulated and unmanipulated regions of an image. (Guillaro et al. 2023) Existing backbone networks (Simonyan and Zisserman 2014) (Wang et al. 2020) (Dosovitskiy et al. 2020), primarily designed for semantic-related tasks, are effective at extracting the semantic features of manipulated images. For extracting non-semantic features, most existing methods rely on handcrafted feature extractors (Zhou et al. 2018) (Bayar and Stamm 2018) (Cozzolino and Verdoliva 2019). As shown in Table 1, almost all existing IML models follow a design of ”semantic segmentation backbone network” combined with ”handcrafted non-semantic feature extraction.”

However, this approach requires custom extraction strategies for different non-semantic features, lacking adaptability in extracting these features. Consequently, this method is limited in improving the model’s ability to adapt to unknown scenarios. Unlike traditional methods that manually extract non-semantic features, we propose an adaptive mechanism to extract non-semantic features in manipulated images. We recognize that the semantic features of an image exhibit strong continuity and significant contextual correlation (Wang et al. 2018), meaning that local semantic features are often inadequate in representing the global semantics of the image. Thus, tight and continuous interactions between local regions are necessary to construct global semantic features. In contrast, the non-semantic features of an image,

*Corresponding Author: Ji-Zhe Zhou, jzzhou@scu.edu.cn

such as frequency and noise, are highly sensitive to manipulation and show greater independence across different regions of the image. This characteristic allows us to employ sparse coding to establish global interactions for non-semantic features, utilizing their sensitivity to detect manipulations.

Based on this concept, we introduce SparseViT, a novel Sparse Vision Transformer. SparseViT employs a sparse self-attention mechanism, redesigning the dense, global self-attention in ViT to better adapt to the statistical properties of non-semantic features. Through sparse processing, the self-attention mechanism selectively suppresses the expression of semantic information, focusing on capturing non-semantic features related to image manipulation. Using a hierarchical strategy, SparseViT applies varying degrees of sparsity at different levels to finely extract non-semantic features. We also designed a multi-scale fusion module (LFF) as the decoder, which integrates feature maps extracted at different sparsity levels, enriching the model’s understanding of non-semantic content across multiple scales and enhancing its robustness. This design enables SparseViT to focus on learning manipulation-sensitive non-semantic features while ignoring semantic features, allowing for adaptive extraction of non-semantic features from images.

To our knowledge, there are currently no models explicitly designed for adaptive extraction of non-semantic features. SparseViT can be considered a pioneering work in adaptive extraction of non-semantic features. All our experiments were conducted under the same evaluation protocol. All models were trained on the CAT-Net (Kwon et al. 2021) dataset and tested on multiple benchmark datasets. Our proposed method demonstrated outstanding image manipulation localization capabilities across several benchmark datasets, with our model achieving the best average performance compared to others. In summary, our contributions are as follows:

- We reveal that semantic features in an image require continuous local interactions to construct global semantics, while non-semantic features, due to their local independence, can achieve global interactions through sparse encoding.
- Based on the distinct behaviors of semantic and non-semantic features, we propose using a sparse self-attention mechanism to adaptively extract non-semantic features from images.
- To address the non-learnability of traditional multi-scale fusion methods, we introduce a learnable multi-scale supervision mechanism.
- Our proposed SparseViT maintains parameter efficiency without relying on feature extractors and achieves state-of-the-art (SoTA) performance and excellent model generalization capabilities across four public datasets.

Related Work

Artifacts Extraction

Early image manipulation localization methods primarily relied on handcrafted convolutional kernels to extract non-semantic features from images. For example, BayarConv

(Bayar and Stamm 2018) designed a convolutional kernel with a high-pass filter structure to capture noise patterns in images. RGB-N (Zhou et al. 2018) introduced SRM filters to capture differences in noise distribution, thereby representing non-semantic features. With the success of deep learning in various computer vision and image processing tasks, many recent techniques have also adopted deep learning to address image manipulation localization (Zhou et al. 2018). However, due to the limitations of existing networks designed for semantic-related tasks in representing non-semantic features, nearly all manipulation localization methods currently rely on semantic segmentation backbone networks combined with handcrafted non-semantic feature extraction.

For instance, ManTra-Net (Wu, AbdAlmageed, and Natarajan 2019) and SPAN (Hu et al. 2020) both integrate BayarConv and SRM as the first layer of their models. ObjectFormer (Wang et al. 2022), based on the Transformer architecture, additionally employs a handcrafted DCT module to extract high-frequency features, enabling better capture of non-semantic characteristics in images. TruFor (Guillaro et al. 2023) uses the handcrafted Noiseprint (Cozzolino and Verdoliva 2019) feature extractor and, through contrastive learning, leverages these extracted features to enhance its manipulation detection and localization capabilities. NCL (Zhou et al. 2023) utilizes a Sobel-based (Dong et al. 2022a) non-semantic feature extractor to enhance its capability in identifying non-semantic features. The methods for extracting non-semantic features from manipulated images by each model are shown in Table 1.

Sparse Self-Attention in Vision Transformers

The Transformer was initially proposed to address natural language processing (NLP) tasks and was first applied to sequence data. The paper (Dosovitskiy et al. 2020) introduced a novel Vision Transformer (ViT) model, providing new insights for applying Transformers to the visual domain.

Since the introduction of Transformers in the visual domain, research on sparse attention has never ceased. The Swin Transformer (Liu et al. 2021b) aggregates attention using shifted windows within a hierarchical structure. The Sparse Transformer (Child et al. 2019) reduces computational complexity by limiting the number of non-zero elements in the attention weights. ResMLP (Touvron et al. 2022) incorporate local connections into the attention mechanism, while (Liu et al. 2021a) utilizes the non-linear properties of MLPs to replace traditional attention computation. ViViT (Arnab et al. 2021) and CSWin Transformer (Dong et al. 2022b) reduce computational cost and improve the model’s ability to handle long sequences by decomposing multi-head self-attention within the transformer. ViViT decomposes attention into temporal and spatial calculations, while CSWin Transformer splits multi-head self-attention into two parallel groups, one handling horizontal stripes and the other handling vertical stripes, forming a cross-shaped window. Focal Self-attention (Yang et al. 2021) sparsifies the attention pattern by combining fine-grained local and coarse-grained global interactions. In the field of IML, no current method has proposed using sparse attention to adapt

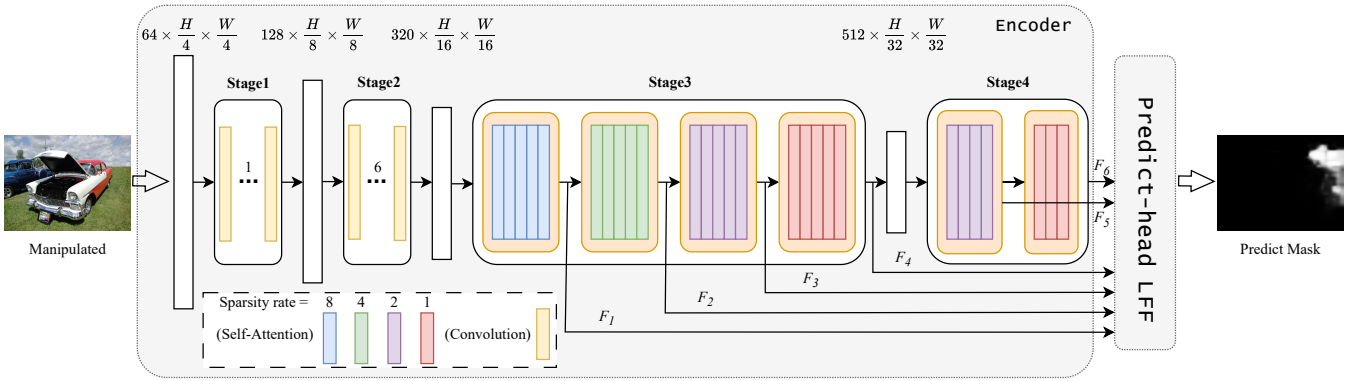


Figure 1: SparseViT. SparseViT consists of two key components: an encoder with a sparse self-attention mechanism and a prediction head (LFF) for multi-scale feature fusion. More detailed information about each module will be presented in Method.

tively extract semantic-agnostic information from manipulated images. Our work is pioneering in the IML field.

Method

Manipulated instances in current datasets often focus on operations such as moving, deleting, or copying entire objects. This allows existing models (Pun, Yuan, and Bi 2015) to identify manipulated regions relatively well by relying solely on semantic features. However, this over-reliance on semantic features neglects the importance of non-semantic features, limiting the model’s generalization ability in unfamiliar or complex manipulation scenarios. We observe that an image’s semantic information exhibits strong continuity and contextual dependency (Wang et al. 2018), necessitating global attention mechanisms to reinforce interactions between local and global regions (Vaswani 2017). In contrast, non-semantic information tends to remain consistent between local and global features and demonstrates greater independence across different regions of an image (Ulyanov, Vedaldi, and Lempitsky 2018). By leveraging this distinction, we can design a mechanism that reduces reliance on semantic information while enhancing the capture of non-semantic information.

To this end, we propose decomposing the global attention mechanism into a “sparse attention” form. Sparse attention, when representing an image’s semantic information, prevents the model from overfitting to it, allowing the model to focus more on non-semantic information in the image. As shown in Figure 1, we have improved the traditional attention calculation in Uniformer (Li et al. 2023) by replacing global self-attention with sparsely self-attention, featuring an exponential decay in sparsity.

Sparse Self-Attention

Traditional deep models focus on detecting semantic objects, aiming to fit these semantic objects. Consequently, traditional self-attention employs a global interaction mode, where every patch in the image participates in token-to-token attention computation with all other patches (Liu et al. 2021b) (Yuan et al. 2021). However, in the domain of image manipulation localization, such global interactions in-

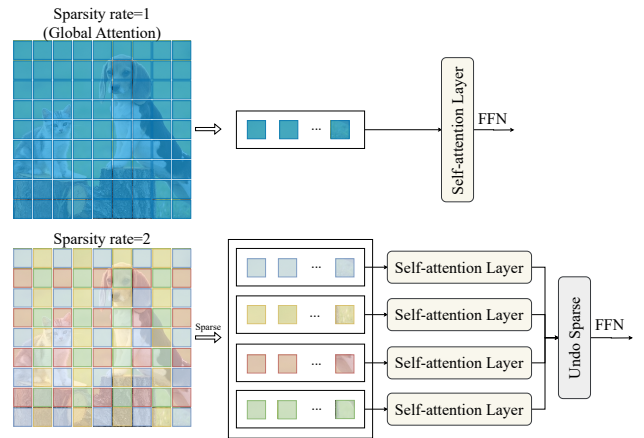


Figure 2: Sparse Self-Attention. A diagram illustrating the calculation of sparse attention. The self-attention computation occurs only between image patches of the same color.

roduce numerous irrelevant key-value pairs. Moreover, the model’s overemphasis on semantic information means that during global interaction, it takes into account features of all patches in the image, such as color and shape, leading to a comprehensive understanding of the image’s overall content. Since the model primarily focuses on the overall semantic structure of the image during global interaction, it tends to overlook the local inconsistencies in non-semantic information that arise after manipulation.

To address this issue, we propose using sparse attention to replace the original global attention. We introduce a new architectural hyperparameter called the “sparsity rate”, abbreviated as “ \mathcal{S} ”. Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, instead of applying attention to the entire $H \times W$ feature map, we divide the features into tensor blocks with a shape of $(\mathcal{S} \times \mathcal{S}, \frac{H}{\mathcal{S}} \times \frac{W}{\mathcal{S}}, C)$. This means that the feature map is decomposed into $\mathcal{S} \times \mathcal{S}$ non-overlapping tensor blocks of size $\frac{H}{\mathcal{S}} \times \frac{W}{\mathcal{S}}$, and self-attention computation is performed within these tensor blocks separately. As illustrated in the Figure 2, only tensor blocks marked with the same color will perform

self-attention computations. This design suppresses the expression of semantic information in sparse attention blocks, allowing the model to focus on extracting non-semantic features. Additionally, the sparsification of tensor blocks in the feature map eliminates the need for attention calculations involving a large number of irrelevant key-value pairs in manipulation localization, thereby reducing FLOPs.

Multi-scale Features

In the task of image manipulation localization, introducing multi-scale supervision with varying sparsity rates is crucial. Feature maps with smaller sparsity rates are rich in semantic information, which helps the model understand the global context and structure of the image. Conversely, feature maps with larger sparsity rates contain more non-semantic information, aiding the model in capturing image details and local features. This introduction of multi-scale supervision allows the model to adaptively extract various non-semantic features by suppressing semantic features to different extents, thereby enhancing its generalization ability across different visual scenes.

As shown in Figure 1, we introduce different sparsity rates in various blocks of Stage 3 and Stage 4. The calculation method for the sparsity rates of each block in Stage 3 and Stage 4 is as follows:

$$S3^{b_i} = 2^{(3-\frac{i}{5})}, \quad i = 0 \dots 19 \quad (1)$$

$$S4^{b_i} = 2^{(1-\frac{i}{4})}, \quad i = 0 \dots 6 \quad (2)$$

Here, the superscript b_i represents different layers within a Stage, where each layer is numbered starting from 0, and the subscript S indicates sparsity. We use the output of the last block in Stage 3 and Stage 4 at different sparsity rates as our multi-scale feature maps. Additionally, due to the sparsification of global attention, we can easily obtain multi-scale information. This approach not only significantly improves the model’s accuracy and performance without increasing computational burden but also makes the model more efficient and robust.

Lightweight and Effective Prediction Head LFF

Layer scale (Touvron et al. 2021) is a technique used in Transformers, where multiple layers of self-attention and feed-forward networks are typically stacked, with each layer introducing a learnable scaling parameter γ . This scaling parameter can learn different values, enabling more effective information transfer throughout the network. Currently, feature fusion methods are usually implemented through simple operations like addition or concatenation (Lin et al. 2017), which only provide fixed linear aggregation of feature maps without considering whether this combination is optimal for specific objects. For the model’s final prediction, our goal is to design a simple yet effective prediction head. Inspired by the Layer scale mechanism in Transformer architecture, we introduce a learnable parameter for each feature map to control the scaling ratio, allowing for more adaptive feature fusion.

The proposed LFF (Learnable Feature Fusion) prediction head is composed of five main parts, as shown in Figure 3.

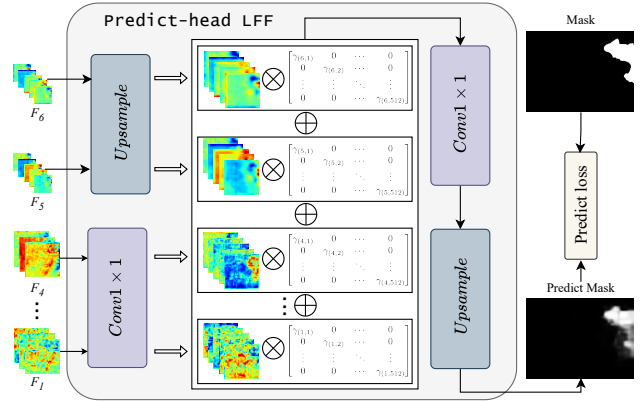


Figure 3: The Structure of LFF. By introducing learnable parameters γ , LFF dynamically adjusts the contribution of each feature map channel to the fusion result.

First, the channels of feature maps F_1 to F_4 are unified to 512 dimensions using an LFF layer. Feature maps F_5 and F_6 are upsampled to one sixteenth of the original size. Then, each feature map is multiplied by its corresponding γ scaling parameter, which is initialized to a small value like $1e-6$. After that, all scaled feature maps are summed using another LFF layer, and the channel dimension of the summed result is reduced to 1. Finally, the result is upsampled, and the upsampled $H \times W \times 1$ mask is used as the final prediction result. The LFF process can be formalized as follows:

$$F_i = \text{Linear}(C_i, C)(F_i), \quad i = 1 \dots 4 \quad (3)$$

$$F_i = \text{Upsample} \left(\frac{H}{16} \times \frac{W}{16} \right) (F_i), \quad i = 5, 6 \quad (4)$$

$$M_p = \text{Add}(F_i \times \gamma), \quad i = 1 \dots 6 \quad (5)$$

$$M_p = \text{Linear}(C, 1)(M_p) \quad (6)$$

$$M_p = \text{Upsample}(H \times W)(M_p) \quad (7)$$

By setting the feature map weight parameters, the model can dynamically adjust each feature map’s contribution to the fusion result, thereby enhancing the flexibility of feature fusion. Through this simple design, the model can better balance and integrate multi-scale features, highlighting important features while suppressing irrelevant or redundant ones.

Results

Experimental Setup

To ensure a fair comparison with existing state-of-the-art image manipulation localization methods, we trained our model on the dataset introduced by CAT-Net (Kwon et al. 2021) and then tested it on CASIAv1 (Dong, Wang, and Tan 2013), NIST16 (Guan et al. 2019), COVERAGE (Wen et al. 2016), Columbia (Hsu and Chang 2006), and DEF-12k (Mahfoudi et al. 2019) datasets. Similar to most previous works (Wei et al. 2023) (Ma et al. 2024), we used pixel-level F1 scores and AUC (Area Under the Curve) to measure the model’s performance. Unless otherwise specified, we reported results using a default threshold of 0.5. For detailed information on the experimental setup and the DEF-12k dataset, refer to Appendix A.

Version	Parameter	FLOPs	COVERAGE		Columbia		CASIAv1		NIST16		DEF-12k	
			F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Uniformer	1.00	1.00	0.378	0.911	0.873	0.938	0.789	0.971	0.326	0.848	0.182	0.810
Uniformer (Sparse)	1.00	0.83	<u>0.485</u>	<u>0.925</u>	0.936	0.976	<u>0.813</u>	<u>0.978</u>	<u>0.355</u>	<u>0.857</u>	0.187	0.809
Uniformer (LFF)	1.01	1.01	<u>0.473</u>	<u>0.923</u>	0.916	0.945	<u>0.808</u>	<u>0.972</u>	<u>0.338</u>	<u>0.850</u>	0.202	<u>0.811</u>
Uniformer (Sobel+LFF)	1.05	1.01	0.312	0.882	0.918	0.950	0.768	0.962	0.313	0.845	0.136	0.755
Uniformer (Bayar+LFF)	1.05	1.01	0.375	0.909	0.891	0.919	0.795	0.971	0.322	0.853	0.148	0.778
Uniformer (DCT+LFF)	1.05	1.03	0.472	0.922	<u>0.937</u>	<u>0.974</u>	0.805	0.974	0.336	0.851	0.173	0.800
Uniformer (SRM+LFF)	1.05	1.02	0.457	0.919	0.930	0.953	0.793	0.969	0.332	0.841	0.189	0.780
SparseViT (Sparse+LFF)	1.01	0.84	0.513	0.935	0.959	0.970	0.827	0.982	0.384	0.861	<u>0.197</u>	0.816

Table 2: Ablation Study of SparseViT. Trained on the CAT-Net joint dataset and validated on CASIAv1. The number of parameters and FLOPs are expressed as multiples of the backbone Uniformer. The best numbers in each column are highlighted in bold, and the second-ranked results are underlined. The consistently improving performance demonstrates the necessity of Sparse and LFF.

Version	Pixel-level F1			
	COVERAGE	Columbia	CASIAv1	NIST16
Single Scale	0.485	0.936	0.813	0.355
MLP	0.492	0.955	0.818	0.372
LFF	0.513	0.959	0.827	0.384

Table 3: Comparison of Multi-Scale Feature Fusion. The best numbers in each column are highlighted in bold.

Ablation Studies

To better assess the performance impact of each component, we adopt an incremental approach by gradually adding components and comparing them with the full model that includes all components. This method allows us to thoroughly measure and optimize the architecture of our proposed model. We examine the effects of using sparse attention versus global attention on model parameters and floating-point operations (FLOPs). Additionally, we compare the capability of manually designed feature extractors and sparse attention mechanisms in extracting non-semantic features. To explore the impact of the LFF prediction head, we compared its performance with the MLP prediction head from SegFormer (Xie et al. 2021) under the introduction of sparse attention. This comparison not only helped us assess the effectiveness of the prediction head design but also revealed the specific impact of different heads on overall model performance. Additionally, we compared the traditional single-scale supervision with our proposed multi-scale supervision method to investigate the advantages of multi-scale supervision and its contribution to model performance. The results for all these evaluations are reported based on training conducted on the dataset proposed by CAT-Net and tested on CASIAv1, NIST16, COVERAGE, Columbia, and DEF-12k. The experimental results are shown in Table 2 and Table 3.

Sparse attention is effective in capturing non-semantic information. In Table 2, we compared the performance of sparse attention and global attention across five datasets. Additionally, we reported the performance of manually ex-

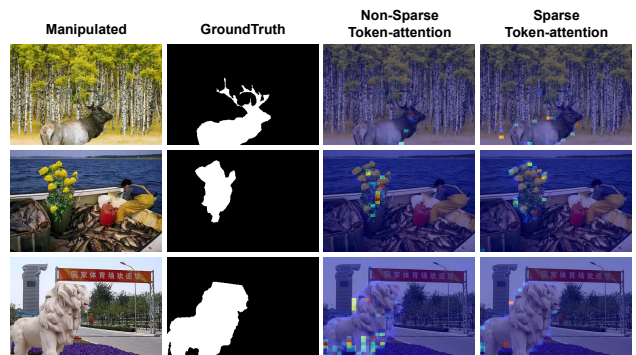


Figure 4: We select an anchor point in the manipulation region and observe how other labels contribute to its attention. After sparsification, the anchor point’s attention focuses more on the manipulation-related edge regions containing non-semantic information, rather than on the surrounding semantic regions.

tracted non-semantic features and sparse attention on these five datasets. The results consistently confirmed the significant advantage of the sparse attention mechanism in extracting non-semantic features from manipulated images. We observed that certain handcrafted feature extraction methods did not significantly enhance model performance on the datasets, and in some cases, even led to performance degradation. This raises questions about the effectiveness of manual non-semantic feature extraction, warranting further investigation. However, it is evident that the sparse attention mechanism significantly improves model performance across all datasets, achieving comprehensive enhancements on five different datasets.

Moreover, the design of sparse attention also shows its advantage in reducing computational burden. Compared to global attention, sparse attention reduces the model’s floating-point operations by approximately 15%, which is especially valuable in large-scale image processing tasks. In summary, sparse attention enhances the model’s sensitivity to subtle artifacts by precisely extracting non-semantic in-

Method	Pixel-level F1					Pixel-level AUC				
	COVERAGE	Columbia	CASIAv1	NIST16	AVG	COVERAGE	Columbia	CASIAv1	NIST16	AVG
ManTraNet	0.196	0.462	0.327	0.193	0.295	0.566	0.724	0.643	0.709	0.661
PSCC-Net	0.379	0.864	0.592	<u>0.369</u>	0.551	0.884	0.946	0.893	0.828	0.888
MVSS	0.439	0.740	0.583	<u>0.348</u>	0.528	0.845	0.934	0.915	0.792	0.872
CAT-Net	0.428	<u>0.915</u>	0.808	0.252	0.601	<u>0.921</u>	0.946	<u>0.978</u>	0.824	<u>0.917</u>
TruFor	<u>0.457</u>	0.885	<u>0.818</u>	0.348	<u>0.627</u>	0.846	0.992	0.897	<u>0.845</u>	0.895
Ours (SparseViT)	0.513	0.959	0.827	0.384	0.671	0.935	<u>0.970</u>	0.982	0.861	0.937

Table 4: Pixel-level Performance. The results show the F1 scores calculated under a fixed threshold of 0.5 and the AUC values obtained using the optimal F1-weighted configuration. The top-ranked results are highlighted in bold, and the second-ranked results are underlined.

formation in manipulated images, thereby significantly improving the model’s generalization ability.

As shown in Figure 4, we qualitatively demonstrate that after sparsification, the model successfully suppresses semantic features that require dense encoding and long-range context dependencies, while being able to extract non-semantic features that do not require dense encoding. In the Appendix C, we conduct a qualitative analysis of sparse attention and handcrafted feature extractors.

Influence of LFF. In Table 3, we report the performance of single-scale features, LFF, and MLP (Xie et al. 2021) prediction heads on the dataset. The experimental results show that regardless of using single-scale or multi-scale features, or adopting different feature fusion strategies, the F1 score on the CASIAv1 dataset exhibits high consistency. We attribute this phenomenon to the fact that CASIAv1 and CASIAv2 are sourced from the same dataset, thus the performance on the CASIAv1 dataset is not sufficient to reflect the model’s generalization ability (Ma et al. 2023). Further analysis reveals that both the LFF prediction head and the MLP prediction head achieve significant improvements in average F1 scores across the five datasets compared to using only single-scale features. This indicates that effective feature fusion strategies can significantly enhance the model’s performance in detecting image manipulation. Specifically, the LFF also achieves an improvement in mean F1 compared to the MLP prediction head, validating that learnable feature fusion outperforms simple feature addition in terms of performance.

The advantage of LFF lies in its ability to adaptively learn the optimal fusion weights between different feature maps, rather than just adding them. This learning mechanism allows LFF to more precisely handle multi-scale features, thereby better capturing manipulation traces in images. Additionally, the use of multi-scale features has proven beneficial, as it provides different levels of semantic and non-semantic information, aiding the model in making more accurate predictions under various operational conditions.

State-of-the-Art Comparison

To ensure fairness in the evaluation, we only considered models whose code is publicly available online. We followed the same protocol as CAT-Net, retrained these models, and tested them on public datasets. In this study, we con-

Method	Size	Parameter	FLOPs
ManTraNet	256×256	3.9M	274.0G
PSCC-Net	256×256	3.7M	45.7G
MVSS	512×512	147.0M	167.0G
CAT-Net	512×512	114.0M	134.0G
TruFor	512×512	68.7M	236.5G
Ours (SparseViT)	512×512	50.3M	46.2G

Table 5: Comparison with the State-of-the-Art on Parameter and FLOPs.

sidered a variety of methods and ultimately included four approaches that rely on handcrafted extraction of non-semantic features for manipulated images: ManTraNet, MVSS, CAT-Net v2, and TruFor. Additionally, we included one method that does not use handcrafted feature extraction: PSCC-Net (Liu et al. 2022). A brief summary of these methods is provided in Table 1 for reference. Our goal is to provide a comprehensive and fair comparison to gain deeper insights into the performance and potential of different approaches in image manipulation localization.

Localization results. In Table 4, we present the performance of various methods in pixel-level localization. Our method stands out with its superior average F1 scores, ranking best across all datasets. A detailed analysis of these results reveals that our model outperforms both traditional methods based on handcrafted non-semantic feature extraction and models that do not rely on handcrafted features. The reason our model excels among many others lies in its innovations in feature learning and representation. By deeply exploring the intrinsic structures of manipulated images, our model can accurately capture the subtle traces left by manipulation. Even when facing complex and varied manipulation techniques, it maintains high accuracy in detection.

Detection results. We selected the weight parameters that performed best in terms of the Pixel-F1 metric to evaluate the model’s AUC performance. By analyzing the data in Table 4, we observe that our SparseViT model achieved the best performance across nearly all tested datasets and exhibited the highest average AUC value. This result indicates that the SparseViT model outperforms existing baselines across a broad range of performance evaluation points.

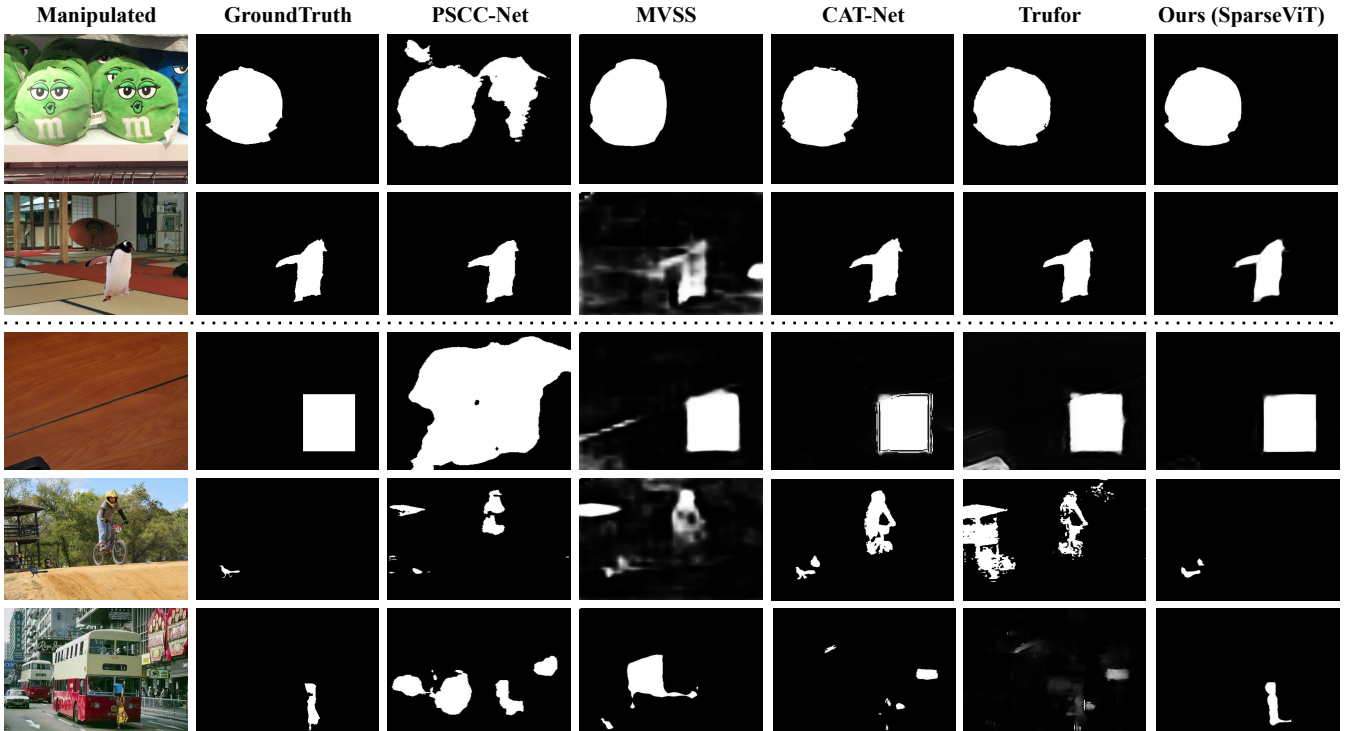


Figure 5: IML by the SoTA. Existing models exhibit noticeable semantic-related false positives in the last three rows. Our model, SparseViT, effectively ignores semantic-related distractions through its unique sparse self-attention mechanism, focusing on capturing features that are unrelated to the semantic content but crucial to the integrity of the image.

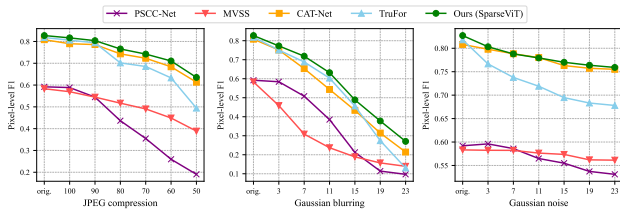


Figure 6: Robustness Analysis on CASIAv1. The results are presented in terms of pixel-level F1 scores.

Comparison of model size. Compared to the current top-performing TruFor, SparseViT not only achieves superior F1 and AUC performance with the same training data size (512×512 pixels) but also reduces the model size by over 80%. Additionally, even when compared to the ManTraNet, which uses smaller training data (256×256 pixels), SparseViT shows a significant advantage in reducing computational load. The specific data is shown in Table 5.

Robustness analysis. Following the guidelines of references (Wu, AbdAlmageed, and Natarajan 2019) and (Hu et al. 2020), we evaluated the robustness of the model against three common attack methods in image manipulation localization on the CASIAv1 dataset, namely JPEG compression, Gaussian blur, and Gaussian noise. The results are shown in Figure 6. Observations indicate that SparseViT outper-

forms existing state-of-the-art models in resisting these disturbances, demonstrating superior robustness.

Overall, compared to existing models tested under a fair cross-dataset evaluation protocol, our model achieves state-of-the-art performance. Figure 5 qualitatively illustrates a key advantage of our model: regardless of whether object-level manipulation is involved, our model effectively utilizes non-semantic features that are independent of the image’s semantic content to accurately identify manipulated regions, thereby avoiding semantic-related false positives.

Conclusions

Relying on handcrafted methods to enhance a model’s ability to extract non-semantic features often limits its generalization potential in unfamiliar scenarios. To move beyond manual approaches, we propose using a sparse self-attention mechanism to learn non-semantic features. Sparse self-attention directs the model to focus more on manipulation-sensitive non-semantic features while suppressing the expression of semantic information. Our adaptive method is not only parameter-efficient but also more effective than previous handcrafted approaches, with extensive experiments demonstrating that SparseViT achieves SoTA performance and generalization ability.

Acknowledgments

This work was jointly supported by the Sichuan Natural Science Foundation under grant 2024YFHZ0355, Sichuan Major Projects under grant 2024ZDZX0001, and the Science and Technology Development Fund, Macau SAR, under grants 0141/2023/RIA2 and 0193/2023/RIA3. Numerical computations were jointly supported by Hefei Advanced Computing Center and Chengdu Haiguang Integrated Circuit Design Co., Ltd. with HYGON K100AI DCU units.

References

- Arnab, A.; Deghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.
- Bayar, B.; and Stamm, M. C. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11): 2691–2706.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.
- Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Cozzolino, D.; and Verdoliva, L. 2019. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15: 144–159.
- Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; and Barnard, K. 2021. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3560–3569.
- Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022a. Mvssnet: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.
- Dong, J.; Wang, W.; and Tan, T. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 422–426. Beijing, China: IEEE. ISBN 978-1-4799-1043-4.
- Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022b. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12124–12134.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissborn, D.; Zhai, X.; Unterthiner, T.; Deghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 63–72. Waikoloa Village, HI, USA: IEEE. ISBN 978-1-72811-392-0.
- Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20606–20615.
- Hsu, Y.-f.; and Chang, S.-f. 2006. Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency. In *2006 IEEE International Conference on Multimedia and Expo*, 549–552. Toronto, ON, Canada: IEEE. ISBN 978-1-4244-0367-7.
- Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 312–328. Springer.
- Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 375–384.
- Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; and Qiao, Y. 2023. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12581–12600.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- Liu, H.; Dai, Z.; So, D.; and Le, Q. V. 2021a. Pay attention to mlps. *Advances in neural information processing systems*, 34: 9204–9215.
- Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Ma, X.; Du, B.; Liu, X.; Hammadi, A. Y. A.; and Zhou, J. 2023. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*.
- Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; et al. 2024. IMDL-BenCo: A Comprehensive Benchmark and Codebase for Image Manipulation Detection & Localization. *arXiv preprint arXiv:2406.10580*.
- Mahfoudi, G.; Tajini, B.; Reirant, F.; Morain-Nicolier, F.; Dugelay, J. L.; and Pic, M. 2019. DEFACTo: Image and

- Face Manipulation Dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)*, 1–5. A Coruna, Spain: IEEE. ISBN 978-90-827970-3-9.
- Pun, C.-M.; Yuan, X.-C.; and Bi, X.-L. 2015. Image forgery detection using adaptive oversegmentation and feature point matching. *IEEE transactions on information forensics and security*, 10(8): 1705–1716.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 5314–5321.
- Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.
- Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.
- Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.
- Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.
- Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.
- Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.
- Wei, Y.; Xiao, B.; Bi, X.; Ma, Z.; Liu, Y.; and Ma, Z. 2023. Secondary Labeling: A Novel Labeling Strategy for Image Manipulation Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8225–8232.
- Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-T.; Shen, X.; and Winkler, S. 2016. COVERAGE — A novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, 161–165. Phoenix, AZ, USA: IEEE. ISBN 978-1-4673-9961-6.
- Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Mantranet: 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*, 9543–9552.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.
- Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; and Gao, J. 2021. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*.
- Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.
- Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; and Yan, S. 2022. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 6575–6586.
- Zhou, J.; Ma, X.; Du, X.; Alhammedi, A. Y.; and Feng, W. 2023. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22346–22356.
- Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1053–1061.

Appendix

Appendix A. Details of the Experimental Setup

Datasets. To ensure a fair comparison with current state-of-the-art Image Manipulation Localization (IML) methods, our model was trained on a dataset provided by CAT-Net (Kwon et al. 2021). Subsequently, we tested the trained model on widely recognized public datasets in the image manipulation localization field. These datasets include CASIAv1 (Dong, Wang, and Tan 2013), NIST16 (Guan et al. 2019), COVERAGE (Wen et al. 2016), Columbia (Hsu and Chang 2006), and DEFACTO (Mahfoudi et al. 2019). Specifically, given that the DEFACTO dataset lacks real images as negative samples, we employed the approach proposed by MVSS (Dong et al. 2022a) to address this issue. We randomly selected 6,000 images from the DEFACTO dataset as positive samples and similarly extracted 6,000 images from the MS-COCO dataset as negative samples. These 12,000 images collectively form our DEF-12k dataset for testing. This approach ensures that during evaluation, the model not only demonstrates its performance on diverse datasets but also undergoes effective testing even in the absence of standard negative samples.

Evaluation Criteria. In our evaluation process, as with most previous studies, we used pixel-level F1 score and AUC (Area Under the Curve) as key metrics to measure model performance. We acknowledge that using the optimal threshold for evaluation may lead to overly optimistic performance estimates, as the ideal threshold is often unknown in real-world applications and may vary across different scenarios. To avoid this and provide a more practical and comparable performance assessment, we employed a fixed threshold in the evaluation report unless otherwise specified. Specifically, we chose 0.5 as the default threshold for reporting the model’s performance metrics.

Implementation. Our SparseViT model was carefully implemented in the PyTorch framework and efficiently trained on an NVIDIA RTX 3090 GPU. During training, we selected a batch size of 16 and set 200 training epochs to ensure that the model could fully learn and converge. For optimization, we used the Adam optimizer with an initial learning rate of 1×10^{-4} , which was then periodically decayed to 1×10^{-7} using a cosine annealing strategy. This approach helps the model to finely approach the optimal solution during training. Similar to MVSS-Net, we performed data augmentation before training to enhance the model’s generalization capability. The data augmentation techniques used include image flipping, blurring, compression, and simple manipulation operations, which help to simulate various transformations and manipulations that images may undergo in the real world. Additionally, to further improve the model’s performance, we employed a pre-training strategy. Specifically, we initialized our SparseViT model using the UniFormer (Li et al. 2023) weights pre-trained on the ImageNet-1k dataset.

Appendix B. The Combination of Sparsity rates

Although we have introduced the hyperparameter “sparsity rate” to achieve sparsity in global self-attention for extract-

Sparsity rate	Pixel-level F1			
	COVERAGE	Columbia	CASIAv1	NIST16
2	0.461	0.967	<u>0.819</u>	0.348
4	0.479	0.945	0.813	0.357
8	<u>0.484</u>	0.949	0.810	<u>0.368</u>
SparseViT	0.513	<u>0.959</u>	0.827	0.384

Table 6: Pixel-level F1 Performance with Different Sparsity rates. All single sparsity rates utilize the LFF prediction head.

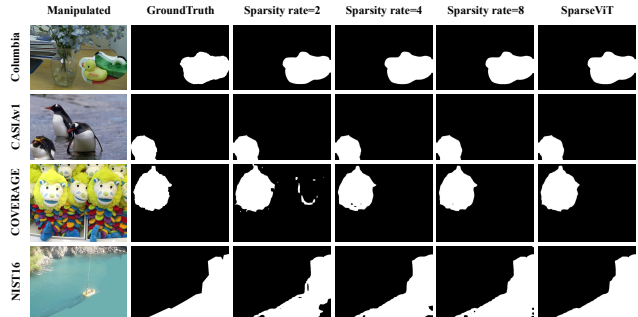


Figure 7: Qualitative Analysis Under Different Sparsity rates. We randomly selected one image from each of the four datasets to demonstrate the localization capability of different sparsity rates on these datasets.

ing non-semantic features, different levels of sparsity in the attention mechanism can identify non-semantic features at varying degrees. Therefore, selecting the “sparsity rate” for our model becomes crucial for extracting non-semantic features.

In our study, we conducted a series of experiments focusing on the combination of sparsity rates within the model. First, we explored the impact of a single sparsity rate on the extraction of non-semantic features. As shown in Table 6, we tested the model’s pixel-level F1 scores under different sparsity rates (2, 4, 8) across four different datasets. The experimental results indicate that on the CASIAv1 and Columbia datasets, models with lower sparsity rates achieved similar or even higher F1 scores compared to those with higher sparsity rates, whereas their performance on the NIST16 and COVERAGE datasets was inferior to that of the high sparsity rate models.

Our analysis revealed that lower sparsity rates are less effective in suppressing semantic information compared to higher sparsity rates. This suggests that on datasets like CASIAv1 and Columbia, which contain more object-level manipulations, the model can still achieve good F1 scores even if it learns incorrect semantic associations. However, on meticulously designed datasets like NIST16 and COVERAGE, the model’s generalization ability is limited due to insufficient learning of non-semantic features. In Figure 7, we conducted a qualitative analysis of different sparsity levels. As revealed by the F1 scores, models with lower spar-

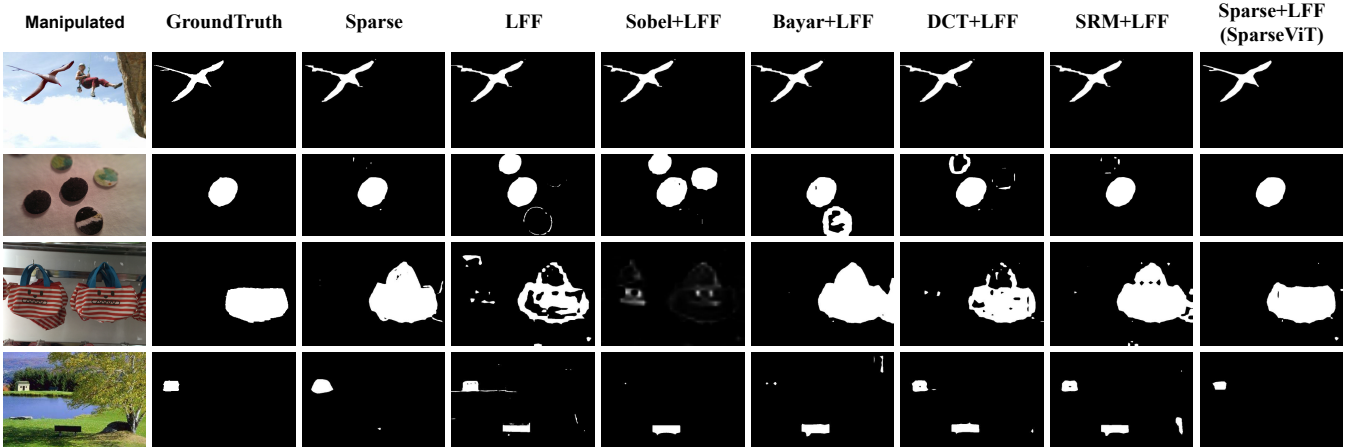


Figure 8: Examples of the ability of sparse self-attention and handcrafted feature extractors to localize manipulated regions.

sity rates underperformed in resisting semantic associations due to insufficient learning of non-semantic features. This resulted in poorer localization performance on high-quality datasets like NIST16 compared to models with higher sparsity rates.

To overcome this limitation and enhance the model’s learning of non-semantic features while improving its generalization ability, we propose a new strategy: applying sparsification to self-attention with exponentially decreasing sparsity rates across different layers of the model. This approach aims to balance the model’s learning of both non-semantic and semantic features, enabling the model to remain sensitive to non-semantic features while also capturing some semantic information, thereby achieving more balanced and robust performance across various datasets.

Appendix C. Qualitative Comparison Results

In Figure 8, we compare the ability of handcrafted feature extractors and the sparse self-attention method to localize manipulated regions in images. The results show that the DCT and SRM handcrafted feature extractors achieve some improvement in identifying manipulated areas. However, the Sobel and Bayar feature extractors, when combined with the LFF prediction head, do not surpass the localization performance of using the LFF prediction head alone. This raises the question of whether all handcrafted feature extractors can effectively extract non-semantic features from images. It is evident that the sparse self-attention mechanism, even without relying on the LFF prediction head, demonstrates superior localization ability compared to DCT and other handcrafted feature extractors. This finding confirms the capability of sparse self-attention in adaptively extracting non-semantic features from manipulated images, suggesting that, compared to traditional handcrafted methods, the sparse self-attention mechanism might be more effective in capturing non-semantic information within images.

Method	AVG F1	Parameter	FLOPs
ASPP	0.647	18.35M	5.46G
AFF	0.669	3.67M	2.06G
LFF	0.671	0.66M	0.68G

Table 7: The parameter efficiency and performance of different fusion techniques.

Appendix D. IoU Results Report

We report the Pixel-level IoU scores of the most advanced IML models, as shown in Table 8. SparseViT achieves the best results across all four datasets. Not only does SparseViT excel in pixel-level F1, but it also demonstrates high precision and robustness in overall image segmentation and recognition tasks. This is attributed to SparseViT’s unique sparse structure design, which significantly enhances the model’s ability to capture non-semantic features while maintaining parameter efficiency.

Appendix E. Implement sparse coding on other ViTs

We chose Uniformer because models like PVT (Wang et al. 2021) and Segformer (Xie et al. 2021) use overlapping patch partitioning, which may make sparse interactions between patches less controllable and lead to overfitting semantics. Additionally, Uniformer uses CNNs in the shallow layers to extract features, and we believe that CNN’s ability to capture basic features, such as edges, is beneficial for IML. Our approach is also compatible with vanilla ViT (Dosovitskiy et al. 2020), as shown in the Table 9. We implemented sparse attention on vanilla ViT and VOLO (Yuan et al. 2022) (without LFF), and the results demonstrate that our method is equally effective for vanilla ViT.

Method	Pixel-level IoU				
	COVERAGE	Columbia	CASIAv1	NIST16	AVG
PSCC-Net	0.301	0.814	0.459	0.294	0.467
MVSS	0.371	0.658	0.505	0.269	0.451
CAT-Net	0.388	0.895	0.754	0.213	0.563
TruFor	0.415	0.859	0.764	0.301	0.585
Ours (SparseViT)	0.472	0.938	0.775	0.331	0.629

Table 8: Pixel-level IoU. Consistent with the Pixel-level F1 results, SparseViT achieves the best performance in Pixel-level IoU compared to current state-of-the-art models.

Method		Pixel-level F1				
		COVERAGE	Columbia	CASIAv1	NIST16	AVG
ViT	non-sparse	0.417	0.858	0.693	0.345	0.578
	sparse	0.441	0.872	0.708	0.372	0.598
VOLO	non-sparse	0.382	0.776	0.646	0.235	0.510
	sparse	0.404	0.779	0.645	0.259	0.522

Table 9: Implement sparse coding on vanilla ViT and VOLO to validate the effectiveness of the proposed sparsification method.

Appendix F. The role of LFF in improving performance

One of the goals of designing LFF is to achieve both lightweight and efficient performance. Therefore, in the ‘‘Influence of LFF’’ section, we focus on comparing it with MLP, which is designed for lightweight purposes. To further highlight the advantages of LFF in terms of both lightweight design and efficiency, we provide additional comparisons with AFF (Dai et al. 2021) and ASPP (Chen et al. 2017) in Table 7. The results demonstrate that SparseViT outperforms these methods in terms of average F1 score and parameter efficiency, proving that LFF can significantly reduce model complexity and computational cost while maintaining performance.