

# IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer

Xiaochen Ma<sup>1</sup>, Bo Du<sup>1</sup>, Zhuohang Jiang<sup>1</sup>, Xia Du<sup>2</sup>, Ahmed Y. Al Hammadi<sup>3</sup>, and Jizhe Zhou<sup>1</sup>

<sup>1</sup> Sichuan University

<sup>2</sup> Xiamen University of Technology

<sup>3</sup> Mohamed Bin Zayed University for Humanities

## Abstract

Advanced image tampering techniques are increasingly challenging the trustworthiness of multimedia, leading to the development of Image Manipulation Localization (IML). But what makes a good IML model? The answer lies in the way to capture artifacts. Exploiting artifacts requires the model to extract non-semantic discrepancies between manipulated and authentic regions, necessitating explicit comparisons between the two areas. With the self-attention mechanism, naturally, the Transformer should be a better candidate to capture artifacts. However, due to limited datasets, there is currently no pure ViT-based approach for IML to serve as a benchmark, and CNNs dominate the entire task. Nevertheless, CNNs suffer from weak long-range and non-semantic modeling. To bridge this gap, based on the fact that artifacts are sensitive to image resolution, amplified under multi-scale features, and massive at the manipulation border, we formulate the answer to the former question as building a ViT with high-resolution capacity, multi-scale feature extraction capability, and manipulation edge supervision that could converge with a small amount of data. We term this simple but effective ViT paradigm IML-ViT, which has significant potential to become a new benchmark for IML. Extensive experiments on three different mainstream protocols verified our model outperforms the state-of-the-art manipulation localization methods. Code and models are available at <https://github.com/SunnyHaze/IML-ViT>.

## 1. Introduction

With the advances in image editing technology like Photoshop, Image Manipulation Localization (IML) methods have become urgent countermeasures to cope with existing tampered images and avoid security threats [33]. Effective IML methods play a crucial role in discerning misinformation and have the potential to contribute to the safety of multimedia world. As shown in Figure 1, the IML task aims to detect whether images have been modified and to localize the modified regions at the pixel level. Image manipulation can be

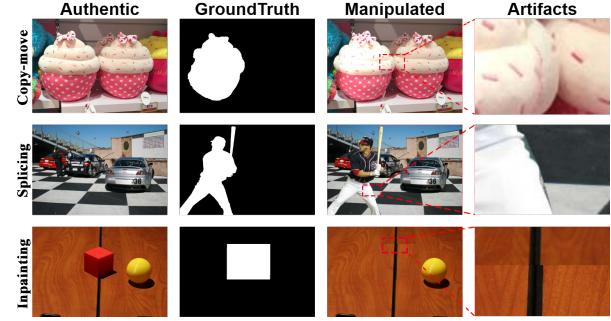


Figure 1. An example of three types of manipulations and their corresponding artifacts. Artifacts contain visible traces, including distortions, sudden changes, or anomalies caused by tampering operations. Artifacts are frequently found at the junction between two regions and appear in very detailed positions. For a better view, zooming in is recommended.

generally classified into three types [33, 36]: (1) *splicing*: copying a region from an image and pasting it to another image. (2) *copy-move*: cloning a region within an image. (3) *inpainting*: erasing regions from images and inpaint missing regions with visually plausible contents.

As shown in Table 1, most existing methods for IML tasks greatly benefit from tracing artifacts with various CNN-based feature extractors. “Artifacts” refer to unique visible traces (see Figure 1) and invisible low-level feature inconsistencies (e.g., noise or high-frequency) resulting from manipulation. As tampering aims to deceive the audience by creating semantically meaningful and perceptually convincing images, visual traces typically manifest at a non-semantic level, distributed in textures around the manipulated area. Additionally, low-level features, like noise inconsistencies introduced by different cameras, can also serve as crucial evidence to reveal manipulated regions within the authentic area. Thus, based on previous experiences, *the key to IML lies in capturing the artifacts by identifying non-semantic visible traces and low-level inconsistencies*.

However, convolution propagates information in a collective manner, making CNNs more suitable for semantic-

Method	Backbone		Resolution	Manipulation supervision	IML dataset thirsty	
	CNN	Tran.			Type	Amount
ManTra-Net [36]	✓	-	512x512	Noise (BayarConv+ SRM filter)	Private	102k
SPAN [15]	✓	-	224x224	Noise (BayarConv+ SRM filter)	Private	102k
CR-CNN [38]	✓	-	short side to 600	Noise (BayarConv+ SRM filter)	Public	5k
GSR-Net [44]	✓	-	300x300	Edge Prediction	Public	5k
MVSS-Net [2]	✓	-	512x512	Noise (BayarConv) Edge (sobel)	Public	5k
MM-Net [39]	✓	-	short side to 800	Noise (BayarConv)	Private	50k
TransForensic [13]	✓	✓	512x512	-	Public	10k
ObjectFormer [34]	✓	✓	256x256	High-frequency	Private	62K
Hifi-Net [12]	✓	-	256x256	Frequency	Public	1,710K
TruFor [11]	✓	✓	Crop 512x512	Noise (Contrastive learning)	Public	35K
IML-ViT (Ours)	-	✓	Zero-pad 1024x1024	Edge loss	Public	5k

**Table 1. Overview of State-of-the-Art End-to-End Models for Image Manipulation Localization.** *Tran.* stands for *Transformer*. *Manipulation supervision* serves as prior knowledge broadly acknowledged in the image manipulation detection field. Edge information effectively traces visible artifacts, while noise and high-frequency features primarily highlight low-level differences between tampered and authentic regions.

related tasks, such as object detection, rather than tracing non-semantic artifacts that often surround an object. Further, to identify low-level inconsistencies, we need to explicitly compare the relationships between different regions. But in deeper networks, CNNs may overlook global dependencies [26], rendering them less effective in capturing differences between regions. Given the weaknesses of CNN in non-semantic and long-distance modeling, we ask: *Is there any other optimal backbone for solving IML tasks?*

Considering the goal of capturing the feature discrepancies between the manipulated and authentic regions, we argue that self-attention should be a better solution regarding IML. As *self-attention can explicitly model relationships between any areas regardless of their visual semantic relevance, especially for non-adjacent regions*. The performance boost achieved by SPAN [15] highlights the effectiveness of integrating self-attention structures into convolutional layers. Furthermore, as artifacts are often distributed at the patch level rather than at the pixel or image level, Vision Transformer (ViT) [8] naturally becomes the ideal choice to trace artifacts and make comparisons.

While ViT may be suitable for IML tasks, directly applying the original ViT architecture is insufficient. We suggest that IML involves three key discrepancies from traditional segmentation tasks, which also have not yet received sufficient attention in previous IML methods, as supported by Table 1. These discrepancies are:

**High Resolution** While semantic segmentation and IML share similar inputs and outputs, IML tasks are more information-intensive, focusing on detailed artifacts rather than macro-semantics at the object level. Existing methods

use various extractors to trace artifacts, but their resizing methods already harm these first-hand artifacts. Therefore, preserving the *original resolution* of the images is crucial to retain essential artifacts for the model to learn.

**Edge Supervision** As mentioned earlier, IML’s primary focus lies in detecting the distinction between the tampered and authentic regions. This distinction is most pronounced at the boundary of the tampered region, whereas typical semantic segmentation tasks only require identifying information within the target region. From another perspective, it becomes evident that visible artifacts are more concentrated along the periphery of the tampered region rather than within it (as shown in Figure 1). Consequently, the IML task must guide the model to concentrate on the manipulated region’s edges and learn its distribution for better performance.

**Multi-scale Supervision** The percentage of tampered area to the total area varies significantly across different IML datasets. CASIAv2 [7] contains a considerable amount of sky replacement tampering, whereas Defacto [27] mostly consists of small object manipulations. On average, CASIAv2 has 7.6% of pixels as tampered areas, while Defacto has only 1.7%. Additionally, IML datasets are labor-intensive and often limited in size, which poses challenges in bridging the gap between datasets. Therefore, incorporating multi-scale supervision from the pre-processing and model design stages is essential to enhance generalization across different datasets.

In this paper, we present IML-ViT, an end-to-end ViT-based model that solves IML tasks. Regarding the proposed three key discrepancies, we devise IML-ViT with the following components: 1) a windowed ViT which accepts **high-resolution** input. Most of the global attention block is replaced with windowed attention as the trade-off for time complexity. We initialize it with Masked Autoencoder (MAE) [14] pre-trained parameters on ImageNet-1k [5]; 2) a simple feature pyramid network (SFPN) [20] to introduce **multi-scale supervision**; 3) a morphology-based edge loss strategy is proposed to ensure **edge supervision**. The overview of IML-ViT is shown in Figure 2.

In this manner, without any specialized modules, IML-ViT offers a general ViT structure for IML tasks. In other words, *IML-ViT proves that IML tasks can be solved without hand-crafted features or deliberate feature fusion process*, promoting future IML methods into a more generalizable design paradigm.

To the best of our knowledge, ObjectFormer [34], TransForensics [13], and TruFor [11] are the only Transformer-related models solving the IML tasks. However, their backbone distinguishes significantly from vanilla ViT, as will be explained in Section 2. Thus, IML-ViT can be regarded as the pioneering model utilizing a vanilla ViT as the backbone for IML tasks.

Currently, the evaluation protocol for IML tasks is rather

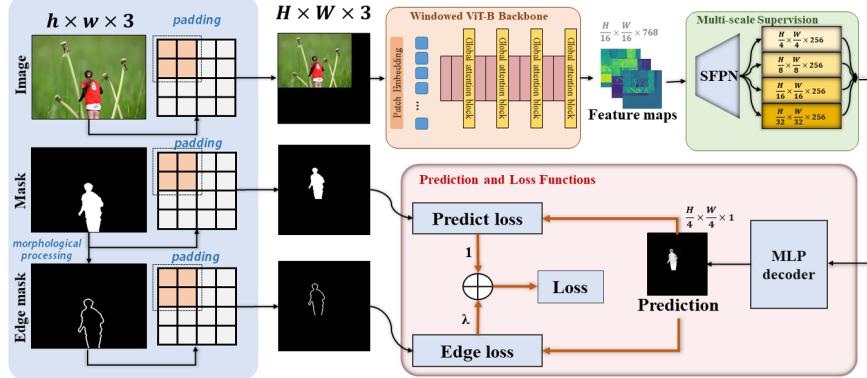


Figure 2. Overview of the general structure of IML-ViT.

chaotic. To bring faithful evaluations and establish IML-ViT as the benchmark model, we demarcate existing messy evaluation settings into three mainstream protocols and conduct comprehensive experiments across these protocols. The extensive experiment results demonstrate that IML-ViT has surpassed all SoTA (state-of-the-art) models, thereby validating the reliability of the three proposed key essences of IML. Thus, we believe that IML-ViT is a powerful candidate to become a new SoTA model for IML.

In summary, our contributions are as follows:

- We reveal the significant discrepancies between IML and traditional segmentation tasks by raising the three essences, which were overlooked by previous studies: high resolution, multi-scale, and edge supervision.
- Aiming at three essences, we modify the components of ViT and establish the IML-ViT, the first ViT-based model for image manipulation localization.
- Extensive experiments show that IML-ViT outperforms state-of-the-art models in both  $F_1$  and AUC scores on various protocols. This verifies the solidity of the three essences we proposed.
- We vanish the evaluation barrier for future studies by demarcating existing evaluation settings into three mainstream protocols and implementing cross-protocols comparisons.

## 2. Related Works

**Paradigm of IML** Research in the early years focused on a single kind of manipulation detection, with studies on copy-move [4, 30], splicing [3, 16, 17], and removal (In-painting) [45], respectively. However, since the specific type of tampering is unknown in practice, after 2018, general manipulation detection has become the focus. Many existing works follow the paradigm of “feature extraction + backbone inference”, especially extractors to exploit tamper-related information from artifacts. CR-CNN [38] has a noise-sensitive BayarConv [1] as the first convolution layer. RGB-N net-

works [43] develop an SRM filter to mine the difference in noise distribution to support decision-making. ManTraNet [36] and SPAN [15] combined SRM, BayarConv, and as the first layer of their model. Besides noise-related extractors, ObjectFormer [34] employs a DCT module to extract high-frequency features, which are then combined with RGB features and fed into a transformer decoder. And MVSS-Net [2] combines a Sobel-supervised edge branch and a BayarConv noise branch with dual attention to fuse them. Nevertheless, a feature may only be effective for a single type of tampering, e.g., noise is more sensitive to splicing from different images but less effective for copy-move from the same image. Recently, TruFor [11] and NCL [41] are the first to explore utilizing contrastive learning to extract features instead of manually designed filters. Proposed IML-ViT also aims to step out of the paradigm of “extraction + fusion” and let the model itself learn as much knowledge as possible from the datasets rather than rely on *priori knowledge*.

**Transformer-based IML method** At present, there are three Transformer-based models in the field of IML, namely ObjectFormer [34] TransForensics [13], and TruFor [11]. Though named “Trans” or “Former”, these models are hardly in line with vanilla ViT in overall structures and design philosophies. In particular, different from ViT directly embedding the patched images before encoding, the first two methods utilize several CNN layers to extract feature maps initially and subsequently employ Transformers for further encoding, leading to neglecting crucial first-hand low-level information. On the other hand, TruFor follows SegFormer [37]’s encoder, using convolution layers instead of position embedding to integrate the position information for Transformer blocks, which overlooked key global dependencies to capture differences between regions.

Moreover, in ObjectFormer’s encoder, the “query” inputs are learnable vectors representing object prototypes  $o_i$ , not image embeddings. As a result, it focuses on cap-

turing dependencies between object prototypes and image tokens, whereas a standard ViT encoder solely models the relationship between image embeddings. Besides, ObjectFormer is pre-trained with a large tampering-oriented synthesized private dataset, while IML-ViT achieves better performance with pre-training on the more accessible ImageNet-1k dataset.

Further, TransForensics has a different way to apply Transformer blocks. While ViT uses these blocks sequentially, TransForensics employs them in parallel, wherein each feature map of an FCN output is decoded with a Transformer block, and then fused for the final output.

In short, IML-ViT can be considered the first IML method with a vanilla ViT as its backbone and could easily benefit from recently advanced algorithms related to ViT, proving that IML tasks do not require complex designs.

### 3. Proposed Method

In this section, we introduce our powerful IML-ViT paradigm, as shown in Figure 2, it consists of three main components: (1) a windowed ViT to balance the high-resolution inputs and the space complexity; (2) a *simple feature pyramid network* (SFPN) to introduce multi-scale features; and (3) a lightweight MLP decoder head with additional edge supervision, which aids in focusing on artifact-related features and ensures stable convergence.

#### 3.1. ViT Backbone

**High Resolution** The ViT Encoder aims to mine the detailed artifacts and explore the differences between the suspicious areas. Thus, it is essential to preserve the **original** resolution of each image to avoid downsampling that could potentially distort the artifacts. However, when training in parallel, all images within a batch must have the same resolution. To reconcile these demands, we adopt a novel approach that has not been applied to any IML method before. Rather than simply rescaling images to the same size, we pad images and ground truth masks with zeros and place the image on the top-left side to match a larger constant resolution. This strategy maintains crucial low-level visual information of each image, allowing the model to explore better features instead of depending on handcrafted prior knowledge. To implement this approach, we first adjust the embedding dimensions of the ViT encoder to a larger scale.

**Windowed Attention** To balance the computation cost from high resolution, we adopt a technique from previous works [20, 21], which periodically replaces part of the global attention blocks in ViT with windowed attention blocks. This method ensures global information propagation while reducing complexity. Differing from Swin [23], this windowed attention strategy is non-overlapping.

**MAE Pre-train** We initialize the ViT with parameters pre-trained on ImageNet-1k [5] with Masked Auto Encoder

(MAE) [14]. This self-supervised method can alleviate the over-fitting problem and helps the model generalize, supported by Table 9.

More specifically, we represent input images as  $X \in \mathbb{R}^{3 \times h \times w}$ , and ground truth masks as  $M \in \mathbb{R}^{1 \times h \times w}$ , where  $h$  and  $w$  correspond to the height and width of the image, respectively. We then pad them to  $X_p \in \mathbb{R}^{3 \times H \times W}$  and  $M_p \in \mathbb{R}^{1 \times H \times W}$ . Balance with computational cost and the resolution of datasets we employ in Table 2, we take  $H = W = 1024$  as constants in our implementation. Then  $X_p$  is passed into the windowed ViT-Base encoder with 12 layers, with a complete global attention block retained every 3 layers. The above process can be formulated as follows:

$$G_e = \mathcal{V}(X_p) \in \mathbb{R}^{768 \times \frac{H}{16} \times \frac{W}{16}} \quad (1)$$

where  $\mathcal{V}$  denotes the ViT, and  $G_e$  stands for encoded feature map. The number of channels, 768, is to keep the information density the same as the RGB image at the input, as  $768 \times \frac{H}{16} \times \frac{W}{16} = 3 \times H \times W$ .

#### 3.2. Simple Feature Pyramid Network

To introduce multi-scale supervision, we adopt the *simple feature pyramid* network (SFPN) after the ViT encoder, which was suggested in ViTDet [40]. This method takes the single output feature map  $G_e$  from ViT, and then uses a series of convolutional and deconvolutional layers to perform up-sampling and down-sampling to obtain multi-scale feature maps:

$$F_i = \mathcal{C}_i(G_e) \in \mathbb{R}^{C_S \times \frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}}}, i \in \{1, 2, 3, 4\} \quad (2)$$

Where  $\mathcal{C}_i$  denotes the convolution series, and  $C_S$  is the output channel dimension for each layer in SFPN. This multi-scale method does not change the base structure of ViT, which allowed us to easily introduce recently advanced algorithms to the backbone.

#### 3.3. Light-weight Predict Head

For the final prediction, we aimed to design a model that is simple enough to reduce memory consumption while also demonstrating that the improvements come from the advanced design in the ViT Encoder and the multi-scale supervision. Based on these ideas, we adopted the decoder design from SegFormer [37], which outputs a smaller predicted mask  $M_e$  with a resolution of  $1 \times \frac{H}{4} \times \frac{W}{4}$ . The lightweight all-MLP decoder first applies a linear layer to unify the channel dimension. It then up-samples all the features to the same resolution of  $C_D \times \frac{H}{4} \times \frac{W}{4}$  with bilinear interpolation, and concatenates all the features together. Finally, a series of linear layers is applied to fuse all the layers and make the final prediction. We can formulate the prediction head as follows:

$$P = \text{MLP}\{\odot_i(W_i F_i + b_i)\} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 1} \quad (3)$$

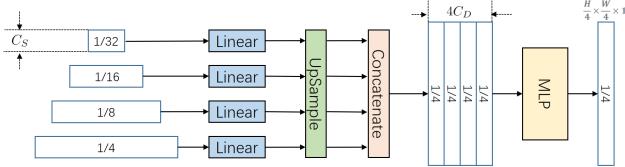


Figure 3. **Diagrams of the predict-head.** The rectangles on the left represent the output of SFPN. There is a normalization layer before entering the MLP block, which is fully discussed below.

Here,  $P$  represents the predicted probability map for the manipulated area;  $\odot$  denotes concatenation operation, and  $MLP$  refers to an MLP module. Detailed structure and analysis are illustrated in Figure 3.

### 3.4. Edge Supervision Loss

To account for the fact that artifacts are typically more prevalent at the edges of tampered regions, where the differences between manipulated and authentic areas are most noticeable, we developed a strategy that places greater emphasis on the boundary region of the manipulated area. Specifically, we generate a binary edge mask  $M^*$  from the original mask image  $M$  using mathematical morphology operations including dilation ( $\oplus$ ) and erosion ( $\ominus$ ) [32], followed by taking the absolute values of the result. The formula we use to generate the edge mask is:

$$M^* = |(M \ominus B(k)) - (M \oplus B(k))| \quad (4)$$

where,  $B(x)$  generates a  $(2x+1) \times (2x+1)$  cross matrix, where only the  $x^{th}$  column and  $x^{th}$  row have a value of 1, while the rest of the matrix contains 0s. The integer value  $x$  is selected to be approximately equal to the width of the white area in the boundary mask. Examples of the edge mask generated using this approach are shown in Figure 4.

**Combined Loss** To compute the loss function, we first pad the ground-truth mask  $M$  and the edge mask  $M^*$  to the size of  $H \times W$ , and refer to them as  $M_p$  and  $M_p^*$ , respectively. We then calculate the final loss using the following formula:

$$\mathcal{L} = \mathcal{L}_{seg}(P, M_p) + \lambda \cdot \mathcal{L}_{edge}(P * M_p^*, M_p * M_p^*) \quad (5)$$

where  $*$  denotes the point-wise product, which masks the original image. Both  $\mathcal{L}_{seg}$  and  $\mathcal{L}_{edge}$  are binary cross-entropy loss functions, and  $\lambda$  is a hyper-parameter that controls the balance between the segmentation and edge detection losses. By default, we searched the optimal  $\lambda = 20$  to guide the model to focus on the edge regions, which is supported by Figure 5. We choose a larger value for  $\lambda$  also for two reasons: (1) to emphasize the boundary region, and (2) to balance the significant number of zeros introduced by zero-padding.

While the proposed edge loss strategy is straightforward, as we will discuss in the Experiments section (Figure 8), it

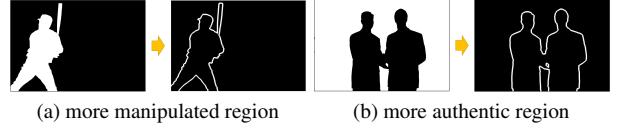


Figure 4. **Examples of generating the edge mask  $M^*$ .** White region represents for manipulated area,  $k$  is set to 7 while the image size is 1024×682. The absolute value operation ensures that whether the tampered region dominates or the non-tampered region dominates, the mask only emphasizes the junction of the two.

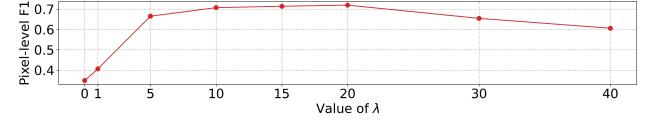


Figure 5. **Lambda selection, trained/test on CASIAv2/v1.**

Table 2. **All datasets in our experiments.**

Dataset	Type		Manipulation type			Resolution	
	Authentic	Manipulated	copymv	spli	inpa	min	max
CASIAv2 [7]	7491	5123	3274	1849	0	240	800
CASIAv1 [7]	800	920	459	461	0	256	384
NIST16 [10]	0	564	68	288	208	480	5616
COVERAGE [35]	100	100	100	0	0	158	572
Defacto-12k [27]	6000	6000	2000	2000	2000	120	640
Columbia [35]	183	180	0	180	0	568	1152
IMD-20 [29]	415	2010	-	-	-	176	4437
tampCOCO [18]	415	800000	600000	200000	0	51	640
JPEG RAISE [18]	24462	0	-	-	-	1515	6159

remarkably accelerates model convergence, stabilizes the training process, and mitigates potential NaN issues. Therefore, we consider this strategy a powerful prior knowledge for IML problems, deserving attention in future research.

## 4. Experiments

### 4.1. Experimental Setup

**Evaluation barrier for IML** While recent studies have introduced numerous SoTA models, comparing them on an equal footing remains challenging. This is due to the following reasons: 1) lack of publicly available code for the models and training processes [2, 15]; 2) utilization of massive synthesized datasets that are inaccessible to the wider research community [34, 36, 39]; 3) training and testing datasets often vary across different papers, also bringing difficulty for comparison.

**Datasets and Evaluation Protocol** To facilitate reproducibility and overcome the existing evaluation barrier, we demarcate existing mainstream IML methods into three distinct protocols based on different partitions of datasets. Subsequently, we compare IML-ViT against SoTA methods with these three protocols, as shown in Table 2 and Table 3. We followed MVSS-Net [2] to create Defacto-12k dataset. More details will be discussed in Section 4.2.

Table 3. The protocols we demarcate from the existing works.

Protocol	Details	Seminal paper
No.1	Train on CASIAv2. Test on other small datasets.	MVSS-Net [2]
No.2	Train on six large mixed datasets. Test on other small datasets.	CAT-Netv2 [18]
No.3	Random split train/test dataset on mixed small datasets.	ObjectFormer [34]

Table 4. Complexity of IML-ViT compared to open-sourced SoTA models. Inference time is measured on a per-image basis.

Method	Infer. Time(s)	Params.(M)	512x512 FLOPs(G)	1024x1024 FLOPs(G)
MVSS-Net[2, 6]	2.929	147	167	683
PSCC-Net [22]	0.072	3	120	416
HiFi-Net [12]	1.512	7	404	3470
TruFor [11]	1.231	68	231	1016
IML-ViT	0.094	91	136	445

**Evaluation Criteria** We evaluate our model using pixel-level  $F_1$  score with a fixed threshold 0.5 and Area Under the Curve (AUC), which are commonly used evaluation metrics in previous works. Both of them are metrics where higher values indicate better performance. However, it's worth noting that AUC can be influenced by excessive true-negative pixels in IML datasets, leading to an overestimation of model performance. Nevertheless, our model achieves SoTA performance in both  $F_1$  score and AUC.

**Implementation** We pad all images to a resolution of 1024x1024, except for those that exceed this limit. For the larger images, we resized them to the longer side to 1024 and maintained their aspect ratio. During training, following MVSS-Net [2], common data augmentation techniques were applied, including re-scaling, flipping, blurring, rotation, and various naive manipulations (e.g., randomly copy-moving or inpainting rectangular areas within a single image). We used the AdamW optimizer [25] with a base learning rate of 1e-4, scheduled with a cosine decay strategy [24]. The early stop technique was employed during training.

**Complexity** Training IML-ViT with a batch size of 2 per GPU consumed 22GB of GPU memory per card. Using four NVIDIA 3090 GPUs, the model was trained on a dataset of 12,000 images over 200 epochs, taking approximately 12 hours. For inference, a batch size of 6 per GPU required 20GB of GPU memory, with an average prediction time of 0.094 seconds per image. Reducing the batch size to 1 decreased the GPU memory requirement to 5.4GB. We also compare the number of parameters and FLOPs with SoTA models in Table 4 and achieve highly competitive results.

## 4.2. Compare with SoTA (See Table 3 for protocols)

**Protocol No.1** Since MVSS-Net [2] has already conducted a detailed evaluation on a fair cross-dataset protocol and later works [41] followed their setting, we directly quote their results here and train our models with the same protocol. The results measured by  $F_1$  score are listed respectively in Table 5. We also compare this with some closed-source methods that only report their AUC tested on CASIAv1 in

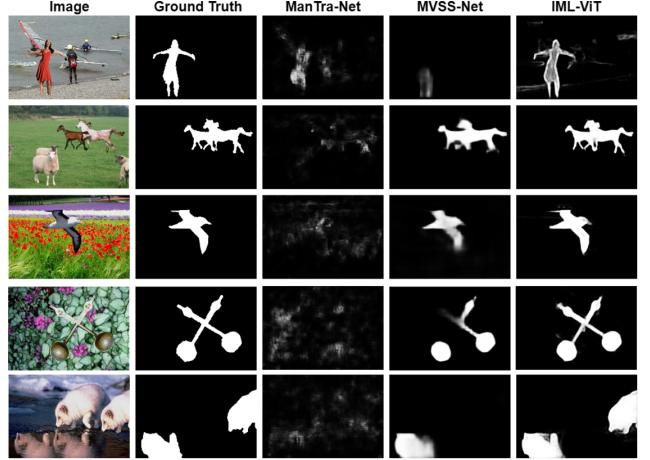


Figure 6. Qualitative results on Protocol No.1 of IML-ViT compared to ManTra-Net and MVSS-Net. For More results, see Appendix.



Figure 7. Qualitative results of IML-ViT for ablation Study. We remove each component to test their contribution.

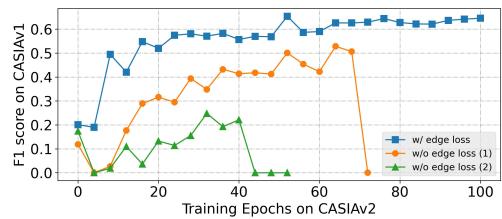


Figure 8. Training stability influenced by proposed edge loss.

Table 7.

Overall, our model achieves SoTA performance on this cross-dataset evaluation protocol. Figure 6 illustrates that our model portrays high-quality and clear edges under different preferences of manipulation types.

**Protocol No.2** TruFor [11] is a recent strong method with extensive experimental results, training on six relatively large IML datasets proposed by CAT-Netv2 [18]. In our aim to establish IML-ViT as the benchmark model, we adopt their protocol to compare our model. We outperform them on four benchmark datasets. Details are shown in Table 8.

**Protocol No.3** TransForensic [13], ObjectFormer [34], HiFi-Net [12] and CFL-Net [28] reported their performance based on mixed datasets. They randomly split these datasets

Method	Pixel-level $F_1$ score					
	CASIAv1	Columbia	NIST16	Coverage	Defacto-12k	MEAN
HP-FCN*, ICCV19 [19]	0.154	0.067	0.121	0.003	0.055	0.080
ManTra-Net*, CVPR19 [36]	0.155	0.364	0.000	0.286	0.155	0.192
CR-CNN*, ICME20 [38]	0.405	0.436	0.238	0.291	0.132	0.300
GSR-Net*, AAAI20 [44]	0.387	0.613	0.283	0.285	0.051	0.324
MVSS-Net*, ICCV21 [2]	0.452	0.638	0.292	0.453	0.137	0.394
MVSS-Net (re-trained)	0.435	0.303	0.203	0.329	0.097	0.270
MVSS-Net++, PAMI22 [6]	0.513	0.660	0.304	<b>0.482</b>	0.095	0.411
NCL-IML, ICCV23 [41]	0.598	0.704	0.231	0.383	0.066	0.396
<i>IML-ViT (ours)</i>	<b>0.721</b>	<b>0.780</b>	<b>0.331</b>	0.410	<b>0.156</b>	<b>0.480</b>

Table 5. **Evaluation results of Protocol No.1.** Except for ManTra-Net and HP-FCN, which were trained on a privately synthesized dataset, all the methods were trained on CASIAv2 datasets. The best scores are highlighted in bold. Symbol “\*” marks the results are quoted from MVSS-Net paper [2].

Table 6. **Evaluation results of Protocol No.3.** \* marks cross-dataset results. Metrics are quoted.

Method	Datasets (Train/validate/test split)	Pixel-level AUC				Pixel-level $F_1$	
		COVER	NIST16 <sup>1</sup>	CASIA	IMD-20	CASIA	COVER
TransForesinc, ICCV21 <i>IML-ViT(Ours)</i>	COVER + CASIA + IMD20 (8:1:1)	0.884	-	0.850	0.848	0.627	0.674
	COVER + CASIA + IMD20 (8:1:1)	0.912	0.821*	<b>0.961</b>	<b>0.943</b>	<b>0.825</b>	<b>0.815</b>
ObjectFormer, CVPR22 Hifi-Net, CVPR23	COVER(4:1); NIST(4:1); CASIA(v2:v1)	0.957	0.996	0.882	-	0.579	0.758
	COVER(4:1); NIST(4:1); CASIA(v2:v1)	<b>0.961</b>	0.996	0.885	-	0.616	0.801
CFL-Net, WACV23 <i>IML-ViT(Ours)</i>	NIST16 + CASIA + IMD20 (8:1:1)	-	<b>0.997</b>	0.863	0.899	-	-
	NIST16 + CASIA + IMD20 (8:1:1)	0.801*	<b>0.997</b>	<b>0.959</b>	<b>0.941</b>	<b>0.820</b>	0.505*

Method	Pre-train	$F_1$ (%)
RGB-N, CVPR18 [42]	ImageNet	40.8
SPAN, ECCV20 [15]	Private synthesized dataset	38.2
Objectformer, CVPR22 [34]	Private synthesized dataset	57.9
<i>IML-ViT(Ours)</i>	MAE on ImageNet-1k	<b>72.0</b>

Table 7. **Comparison with Closed-source methods with Protocol No.1.**

Table 8. **Pixel-level F1 on Protocol No.2.**

Method	CASIAv1	COVER	Columbia	NIST16
CAT-Netv2, IJCV22 [18]	0.752	0.381	0.859	0.308
TruFor, CVPR23 [11]	0.737	0.600	0.859	0.399
<i>IML-ViT(ours)</i>	<b>0.798</b>	<b>0.654</b>	<b>0.945</b>	<b>0.503</b>

into training/validation/testing splits, causing the random splits performed by others to potentially differ, leading to a certain degree of unfairness. Therefore, we do not recommend using this protocol in future work. However, for the sake of comparison with these state-of-the-art models, we also test IML-ViT following this protocol. Besides, note that HiFi-Net (1,710K images) and Objectformer (62k images) involve large IML datasets for pre-training, then tune on the specific small dataset, while we only pre-train with ImageNet-1k. Thus, we directly use results from mixed public IML datasets(14k images) to compare with them. Otherwise, it’s easy to overfit on small datasets. In summary, experiment results in Table 6 show that, under this

reasonable evaluation criteria, IML-ViT also outperforms these SoTA methods.

### 4.3. Ablation Studies

To evaluate the contributions of each component to the model performance, we conducted experiments with multiple settings and compared them with a *full setup* to test the four aspects we are most concerned about. For *initialization*, besides *full setup* with MAE pre-training on ImageNet-1k, we test Xavier [9] initialization and ordinary ViT pre-training on ImageNet-21k. To explore the impact of *high resolution*, we resized all images to 512×512 during training before applying our padding strategy. For *edge supervision*, we remove the edge loss for evaluation, while for *multi-scale supervision*, we replace the module with the same number of plain convolution layers.

To reduce expenses, we trained model *only* with **Protocol No.1** in the ablation study. Qualitative results are illustrated in Figure 7, which vividly demonstrates the efficacy of each component in our method. For quantitative results in Table 9, our findings are:

**MAE pretrain is mandatory.** Indeed, dataset insufficiency is a significant challenge in building ViT-based IML methods. As shown in Table 2, public datasets for IML are small in size, which cannot satisfy the appetite of vanilla ViT. As shown in *w/o MAE* aspects in Table 9, the use of Xavier initialization to train the model resulted in complete non-convergence. However, while regular ViT pre-training

Table 9. **Ablation study of IML-ViT.** Each model is trained for 200 epochs on the CASIAv2 dataset. Best scores are marked in bold. *H*-*Reso* refers to high resolution; *SFPN* refers to simple feature pyramid network; and *Edge* refers to proposed edge supervision.

Test Goal	Init Method	Components			CASIAv1		Coverage		Columbia		NIST16		MAEN	
		H-Reso	SFPN	Edge	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
w/o MAE	Xavier	+	+	+	0.1035	-	0.0439	-	0.0744	-	0.0632	-	0.0713	-
w/o high resolution	ViT-B ImNet-21k	+	+	+	0.5820	0.9037	0.2123	0.7898	0.5040	0.8335	0.2453	0.7939	0.3859	0.8302
w/o multi-scale	MAE ImNet-1k	-	+	+	0.5747	0.9121	0.2622	0.7889	0.5150	0.8028	0.3292	0.7950	0.4153	0.8247
w/o edge-supervision	MAE ImNet-1k	+	-	+	0.6504	0.9306	0.3877	0.8829	0.7096	0.8816	0.2847	0.7771	0.5081	0.8681
Full setup	MAE ImNet-1k	+	+	-	0.6177	0.9240	0.3176	0.8789	0.6843	0.9161	0.2648	0.8045	0.4711	0.8809

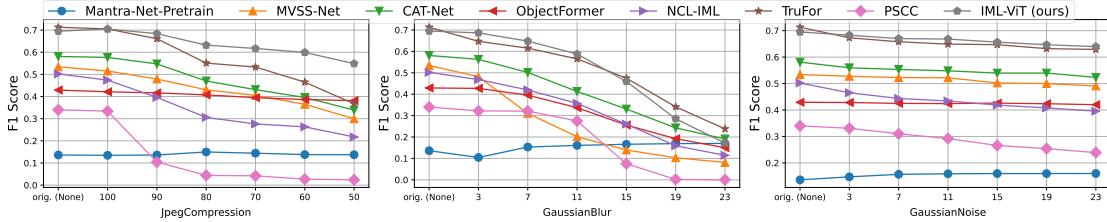


Figure 9. Robustness evaluation by three kinds of attacking across various SoTA methods under Protocol No.1.

initialization with Imagenet-21k achieves acceptable performance on CASIAv1, which is homologous to CASIAv2, it exhibits poor generalization ability on other non-homology datasets. This indicates that MAE greatly alleviates the problem of non-convergence and over-fitting of ViT on limited IML datasets.

**Edge supervision is crucial.** The performance of IML-ViT without edge loss shows significant variability with different random seeds, all leading to gradient collapse eventually, where the F1 score reaches 0, and the loss becomes *Nan*, as shown in Figure 8. In contrast, when employing edge loss, all performance plots exhibit consistent behavior similar to the blue line in Figure 8, enabling fast convergence and smooth training up to 200 epochs. Furthermore, Table 9 confirms the effectiveness of edge loss in contributing to the final performance. In summary, these results demonstrate that edge supervision effectively stabilizes IML-ViT convergence and can serve as highly efficient prior knowledge for IML problems.

**High resolution is effective for artifacts.** The improved performance shown in Table 9 for the *full setup* model across four datasets validates the effectiveness of the high-resolution strategy. However, it is essential to note that the NIST16 dataset shows limited improvement when using higher resolutions. This observation can be attributed to the fact that the NIST16 dataset contains numerous images with resolutions exceeding 2000, and down-sampling these images to 1024 for testing may lead to considerable distortion of the original artifacts, consequently reducing the effectiveness of learned features. Nevertheless, when considering the SoTA score achieved, it becomes evident that IML-ViT can flexibly infer the manipulated area based on the richness of

different information types.

**Multi-scale supervision helps generalize.** All these datasets exhibit significant variations in the proportion of manipulated area, particularly where CASIAv2 has 8.96% of the pixels manipulated, COVERAGE dataset has 11.26%, Columbia dataset has 26.32%, and NIST16 has 7.54%. Nevertheless, the comprehensive improvements in Table 9 with the aid of multi-scale supervision indicate that this technique can effectively bridge the gap in dataset distribution, enhancing generalization performance.

#### 4.4. Robustness Evaluation

We conducted a robustness evaluation on our IML-ViT model following MVSS-Net [2]. We utilized their protocol with three common types of attacks, including JPEG compression, Gaussian Noise, and Gaussian Blur. As shown in Figure 9, IML-ViT achieved very competitive results among SoTA models, which proved to possess excellent robustness.

### 5. Conclusions

This paper introduces IML-ViT, the first image manipulation localization model based on ViT. Extensive experiments on three mainstream protocols demonstrate that IML-ViT achieves SoTA performance and generalization ability, validating the reliability of the three core elements of the IML task proposed in this study: high resolution, multi-scale, and edge supervision. Further, IML-ViT proves the effectiveness of self-attention in capturing non-semantic artifacts. Its simple structure makes it a promising benchmark for IML.

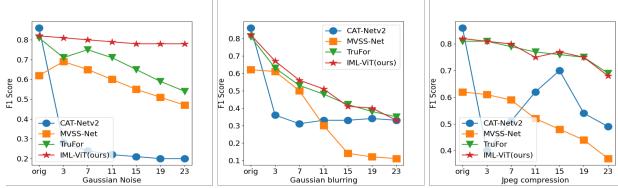
## References

- [1] Belhassen Bayar and Matthew C. Stamm. 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. 3
- [2] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 14165–14173, Montreal, QC, Canada, 2021. IEEE. 2, 3, 5, 6, 7, 8, 11, 12
- [3] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, page 1–6, Roma, Italy, 2015. IEEE. 3
- [4] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11):2284–2297, 2015. 3
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255, Miami, FL, 2009. IEEE. 2, 4
- [6] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–14, 2022. 6, 7
- [7] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, page 422–426, Beijing, China, 2013. IEEE. 2, 5
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. (arXiv:2010.11929), 2021. arXiv:2010.11929 [cs]. 2
- [9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 7
- [10] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhan, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, page 63–72, Waikoloa Village, HI, USA, 2019. IEEE. 5
- [11] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 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*, pages 20606–20615, 2023. 2, 3, 6, 7, 11
- [12] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2, 6
- [13] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: Image forgery localization with dense self-attention. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 15035–15044, Montreal, QC, Canada, 2021. IEEE. 2, 3, 6
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15979–15988, New Orleans, LA, USA, 2022. IEEE. 2, 4
- [15] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. 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*, pages 312–328. Springer, 2020. 2, 3, 5, 7
- [16] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. *Fighting Fake News: Image Splice Detection via Learned Self-Consistency*, page 106–124. Springer International Publishing, Cham, 2018. 3
- [17] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. 3
- [18] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 5, 6, 7
- [19] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 8300–8309, Seoul, Korea (South), 2019. IEEE. 7
- [20] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. (arXiv:2111.11429), 2021. arXiv:2111.11429 [cs]. 2, 4
- [21] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4794–4804, New Orleans, LA, USA, 2022. IEEE. 4
- [22] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Psccnet: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 6
- [23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9992–10002, Montreal, QC, Canada, 2021. IEEE. 4

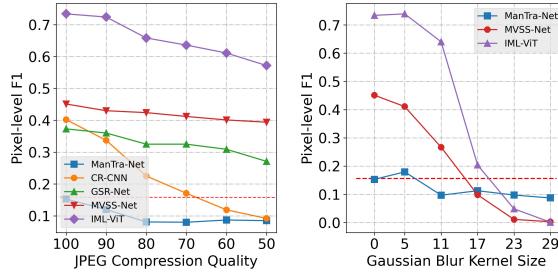
- [24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. (arXiv:1608.03983), 2017. arXiv:1608.03983 [cs, math]. 6
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. (arXiv:1711.05101), 2019. arXiv:1711.05101 [cs, math]. 6
- [26] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 2
- [27] Gael Mahfoudi, Badr Tajini, Florent Retraint, Frederic Morain-Nicolier, Jean Luc Dugelay, and Marc Pic. Defacto: Image and face manipulation dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)*, page 1–5, A Coruna, Spain, 2019. IEEE. 2, 5
- [28] Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S. Woo. Cfl-net: Image forgery localization using contrastive learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 4631–4640, Waikoloa, HI, USA, 2023. IEEE. 6, 12, 13
- [29] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, page 71–80, Snowmass Village, CO, USA, 2020. IEEE. 5
- [30] Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, page 1–6, Abu Dhabi, United Arab Emirates, 2016. IEEE. 3
- [31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 12
- [32] Jean Serra. Image analysis and mathematical morphology, 1983. 5
- [33] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5): 910–932, 2020. 1
- [34] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Object-former for image manipulation detection and localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2354–2363, New Orleans, LA, USA, 2022. IEEE. 2, 3, 5, 6, 7, 12, 13
- [35] Bihani Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage — a novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, page 161–165, Phoenix, AZ, USA, 2016. IEEE. 5
- [36] Yue Wu et al. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9535–9544, Long Beach, CA, USA, 2019. IEEE. 1, 2, 3, 5, 7
- [37] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 3, 4
- [38] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *2020 IEEE International conference on multimedia and expo (ICME)*, page 1–6. IEEE, 2020. 2, 3, 7
- [39] Chao Yang, Zhiyu Wang, Huawei Shen, Huizhou Li, and Bin Jiang. Multi-modality image manipulation detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Shenzhen, China, 2021. IEEE. 2, 5
- [40] Li Yanghao, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. (arXiv:2203.16527), 2022. arXiv:2203.16527 [cs]. 4
- [41] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22346–22356, 2023. 3, 6, 7
- [42] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 1053–1061, Salt Lake City, UT, USA, 2018. IEEE. 7
- [43] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 1053–1061, Salt Lake City, UT, USA, 2018. IEEE. 3
- [44] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13058–13065, 2020. 2, 7
- [45] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67:90–99, 2018. 3

## A. Futher Robustness Evaluation

JPEG compression, Gaussian Noise, and Gaussian Blur are the common attack methods for Image manipulation localization. Following the convention from TruFor [11] and MVSS-Net [2], we further carried out experiments on the resistance of these operations on Protocol No.1 and No.2. The evaluation results are shown in Figure 10 and Figure 11. The IML-ViT exhibited excellent resistance to these attack methods and consistently maintained the best performance of the models.



**Figure 10. Robustness evaluation against common attack on Protocol No.2.** Results are quoted from TruFor paper. They searched for the optimal F1 score to report the results while we selected 0.5 as the threshold for the F1 score here, proving IML-ViT more suitable for real-world scenarios.



**Figure 11. Robustness Evaluation against JPEG compression and Gaussian blur on Protocol No.1.** The red dashed line represents the  $F_1$  score when all predictions are classified as positive. When the result is lower than this line, we consider the model to be less effective than random guessing and losing its localization ability. Performance against JPEG compression is quoted from MVSS-Net, while performance against Gaussian Blur is retested by us using three publicly available models. Our model has a later entry of the  $F_1$  score into the red-line value compared to other models, and it consistently maintains a relatively high position, proving its better resistance.

## B. More Implementation Details

### B.1. High-resolution ViT-Base

Mostly following the original Vision Transformer, we implemented our model with a stack of Transformer blocks stacking together. LN are employed in the self-attention head and MLP blocks. Every two windowed attention blocks

are followed by a global attention block. The windowed attention block only computes self-attention in a small, non-overlapped window, while the global attention block ensures global information propagation. Although we introduce the windowed attention, it only affects the self-attention manner but doesn't change the linear projection for Q, K, and V. Therefore, we can directly apply the MAE pre-trained parameters from a vanilla ViT-B with all global attention to this windowed ViT-B without any extra process. Detailed configuration are shown in Table 10.

Configs	Value
patch size	16
embedding dim	768
depth	12
number of heads	12
input size	$3 \times 1024 \times 1024$
window size	14
Norm layer	LN
Global block indexes	2,5,8,11
Output shape	$768 \times 64 \times 64$

Table 10. Detailed structure of windowed ViT-Base

### B.2. Simple Feature Pyramid

After obtaining the output from ViT-B, SFPN utilizes a sequence of convolutional, pooling, or deconvolutional (ConvTranspose2D) layers to downsample it into feature maps with 256 channels, scaling them to resolutions of  $\{4.0, 2.0, 1.0, 0.5, 0.25\}$  relative to the resolution of the input feature maps ( $768 \times 64 \times 64$ ). For example, the largest output feature map with a scale of 4.0 is shaped like  $256 \times 256 \times 256$ , while the smallest one with a scale of 0.25 is shaped like  $256 \times 8 \times 8$ . Each layer is followed by LayerNorm. Detailed structures of each scale can be seen in Table 11.

Scales	Layers & channels of feature maps
4.0	768 ConvT 384 ConvT 192 Conv(1,1) 256 Conv(3,3) 256
2.0	768 ConvT 384 Conv(1,1) 256 Conv(3,3) 256
1.0	768 Conv(1,1) 256 Conv(3,3) 256
0.5	768 maxpool2D 384 Conv(1,1) 256 Conv(3,3) 256
0.25	768 maxpool2D 384 Conv(1,1) 256 Conv(3,3) 256 maxpool2D 256

Table 11. Detailed structure of Simple feature pyramid  $ConvT$  denotes for ConvTranspose2D with kernel size of 2 and stride of 2;  $Conv(x,x)$  indicate that a Conv2D layer with kernel size of  $x$ ; and maxpool2D has also a kernel size of 2. The number shown between layers indicates the number of channels for its respective feature map between layers.

### B.3. Predict-head's norm & training including authentic images

The exact structure we applied in the predict-head is shown in Figure 3. There is a norm layer before the last  $1 \times 1$

convolution layer in the predict-head. We observed that when changing this layer may influence the following aspects: 1) convergence speed, 2) performance, and 3) generalizability.

In particular, Layer Norm can converge rapidly but is less efficient at generalization. Meanwhile, the Batch Norm can be generalized better on other datasets. However, when including authentic images during training, the Batch Norm may sometimes fail to converge. At present, a straightforward solution is to use Instance Normalization instead, which ensures certain convergence. Our experimental results are shown in Table 12.

Delving into the reasons, MVSS-Net [2] is the pioneering paper proposing the incorporation of authentic images with fully black masks during training to reduce *false positives*. We highly endorse this conclusion as it aligns more closely with the practical scenario of filtering manipulated images from a massive dataset of real-world images. However, in terms of metrics in convention, because the F1 score is meaningful only for manipulated images (as there are no positive pixels for fully black authentic images,  $F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$ , yielding  $F1 = 0$ ), we only computed data for manipulated images. This approach may result in an “unwarranted” metric decrease when real images are included.

Norm	Dataset	CASIAv1		Coverage		Columbia		NIST16		MEAN
		F1	Epoch	F1	Epoch	F1	Epoch	F1	Epoch	
Layer	CASIAv2-5k	0.686	168	0.347	168	0.760	128	0.231	104	0.506
Batch	CASIAv2-5k	0.702	184	0.421	184	0.730	184	0.317	184	0.543
Instance	CASIAv2-5k	0.719	184	0.419	176	0.792	140	0.263	150	0.547
Batch	CASIAv2-12k	0.715	176	0.352	128	0.767	150	0.263	124	0.524
Instance	CASIAv2-12k	0.721	140	0.362	100	0.784	136	0.258	68	0.531

Table 12. **Testing for norm layer in predict-head** Implementation is followed to ablation study in the main paper. CASIAv2-5k refers to manipulated images only, while 12k includes authentic images as well.

**Training settings** Since our model could only train with small batch size, we applied the *gradient accumulate* method during training, i.e. updating the parameters every 8 images during training. We select this parameter by experiments, details see Table 13.

Batchsize	GPUs	accum iter	CASIAv1		Coverage		Columbia		NIST16		MEAN
			F1	Epoch	F1	Epoch	F1	Epoch	F1	Epoch	
2	4	2	0.686	184	0.302	144	0.685	92	0.304	144	0.494
2	4	4	0.704	192	0.386	140	0.772	60	0.331	140	0.548
2	4	8	0.722	152	0.410	140	0.780	84	0.332	140	0.561
2	4	16	0.706	184	0.419	184	0.782	92	0.314	184	0.555
2	4	32	0.602	184	0.249	184	0.740	184	0.254	184	0.461

Table 13. **Test for best accumulate gradient parameter for IML-ViT**. Train/tested on CASIAv2/v1 with four NVIDIA 3090 GPUs.

Besides, we adopt the early stop method during training. Evaluate the performance on the F1 score for CASIAv1, and stop training when there is no improvement for 15 epochs. Other configs are described in Table 14.

Table 14. **Training settings for IML-ViT**

Configs	Value
batch size	2 (RTX 3090) or 4 (A40)
GPU numbers	4 (RTX 3090) or 2 (A40)
accumulate gradient batch size	8
epochs	200
warm up epochs	4
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
base learning rate	1e-4
minimum learning rate	5e-7
learning rate schedule	cosine decay
weight decay	0.05

## C. What artifacts does IML-ViT capture?

To investigate whether IML-ViT focuses on subtle artifacts as expected, we employ GradCAM [31] to visualize the regions and content the model focuses on, as shown in Figure 12. Additional results are in the Appendix E.4. We can observe that IML-ViT captures the traces around the manipulated region with the help of edge loss. Further, we can observe some extra subtle attention out of the manipulated region in the fourth image, proving the global dependent ability of ViT can help the model trace the tampered region.

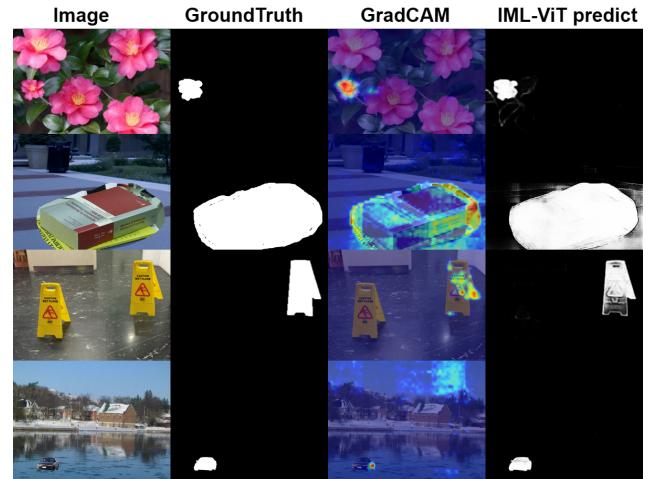


Figure 12. **GradCAM visualization of IML-ViT**.

## D. Additional Experiments Results

Since the space limit, we place a part of our results in **Protocol No.1** here.

ObjectFormer [34] and CFL-Net [28] evaluate their models fine-tuning with CASIAv2 on AUC. Although this metric may overestimate the models, IML-ViT has still surpassed them, as shown in Table 15.

Method	CASIAv1	Coverage	Columbia	NIST16	MEAN
ObjectFormer [34]	0.882	-	-	-	-
CFL-Net [28]	0.863	-	-	0.799	-
<i>IML-ViT(Ours)</i>	<b>0.931</b>	<b>0.918</b>	<b>0.962</b>	<b>0.818</b>	<b>0.917</b>

Table 15. Comparison of AUC scores trained on CASIAv2.

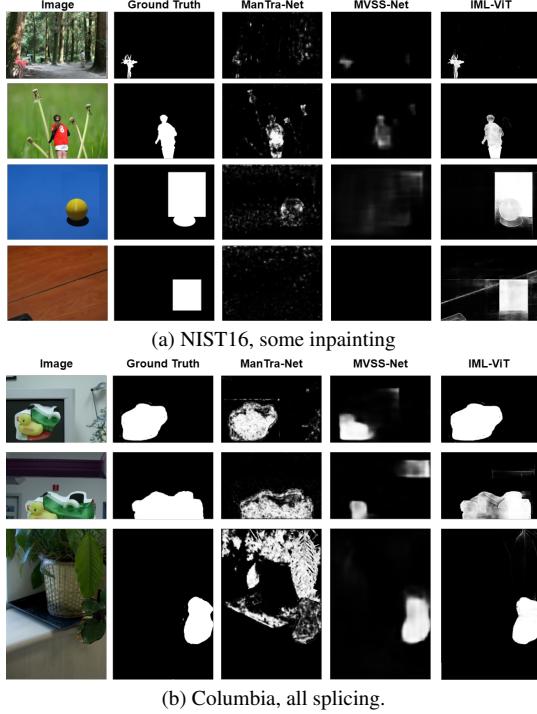


Figure 13. Testing results on Protocol No.1 of IML-ViT compare to ManTra-Net and MVSS-Net. Each dataset has its preference for manipulation types.

## E. Extra Visualization

### E.1. Visualization of Protocol No.1 on other datasets

Here we also present some of the predict masks under Protocol No.1, which was from dataset with other preference on manipulation types. Extended from CASIAv1 and COVERAGE datasets in the main paper, we present results in NIST16 and Columbia datasets here in Figure 13.

### E.2. Qualitative results for ablation study

The ablation study from Figure 7 evaluates the impact of various components on IML-ViT’s performance: 1) w/o multi-scale: Significant degradation with poor feature detection and blurred outputs. 2) w/o MAE: Improved over the absence of multi-scale, but still blurry with weak edge definition. 3) w/o high-resolution: Noticeable drop in detail and precision, with coarse boundaries. 4) w/o Edge Loss: Less defined edges, preserving overall shape but losing structural details. 5) Full Setup: Produces the most accurate and

detailed segmentation maps, capturing fine details and clear boundaries. In summary, the ablation study highlights the critical contributions of each component to the overall performance of IML-ViT. The multi-scale processing, MAE pre-training, high-resolution input, and edge loss each play a vital role in enhancing the model’s ability to produce a high-quality segmentation map.

### E.3. Extra results for CASIA datasets.

To provide a detailed showcase of IML-ViT’s performance on image manipulation localization tasks, we present additional image results on the CASIA dataset in Figure 14.

### E.4. Extra GradCAM results

Here we provide extra GradCAM results to verify if IML-ViT focuses on the artifacts we want it to trace. Artifacts are mainly distributed around the manipulated region with rapid changes. Figure 15 vividly shows that the IML-ViT can effectively discover the artifacts from the image and support its decision.

### E.5. Feature maps between each module

To gain a deeper understanding of IML-ViT, we present visualizations of feature maps between layers by calculating the average channel dimensions of the feature map. The outcomes are displayed in Figure 16. This visualization process allows us to shed light on the model’s functioning and provides valuable insights into its mechanisms.

## F. Limitation

We observe a rapid decline in IML-ViT’s performance on the Gaussian blur attack when the filter kernel size exceeded 11. We argue that this is mainly because our motivation is to make the model focus on detailed artifacts, but excessive Gaussian blurring can significantly remove these details, leading to a sudden drop in performance. However, from another perspective, this can actually prove that our model is able to effectively capture artifacts in tampering. Currently, the training does not specifically enhance blur, so we believe that adding enough blur data augmentation can compensate for this issue.

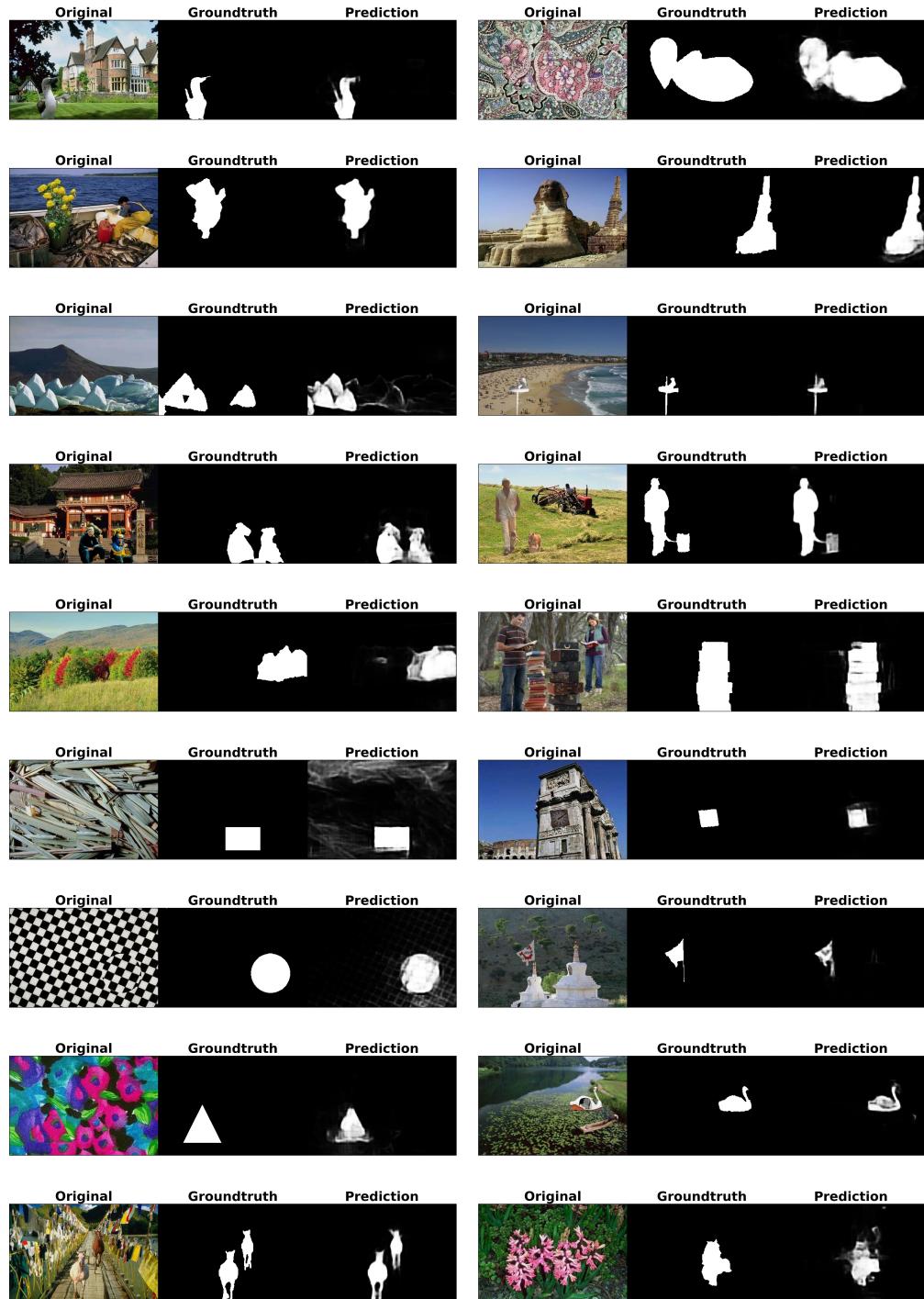


Figure 14. Additional CASIAv1 results of IML-ViT. Trained on CASIAv2.

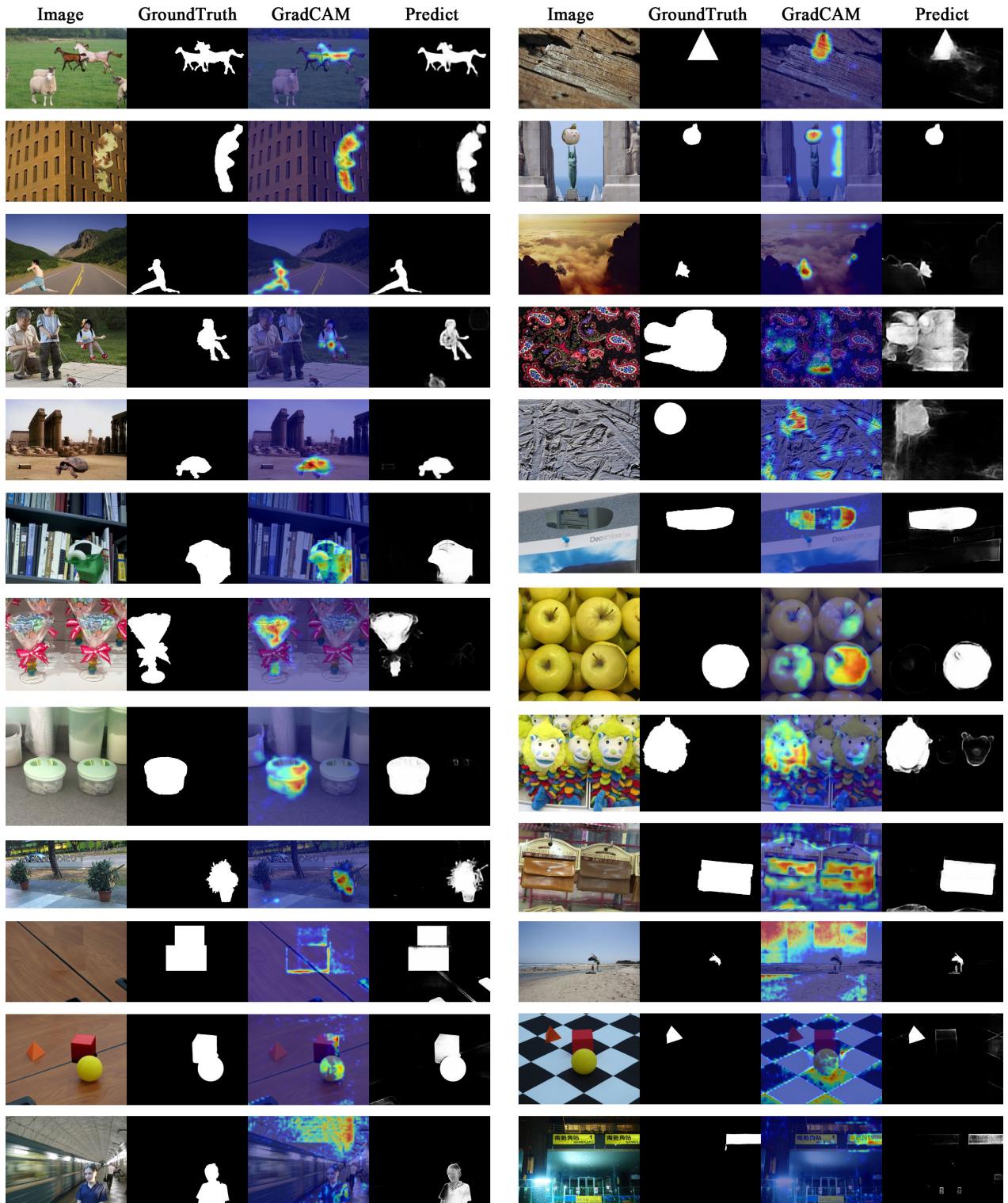
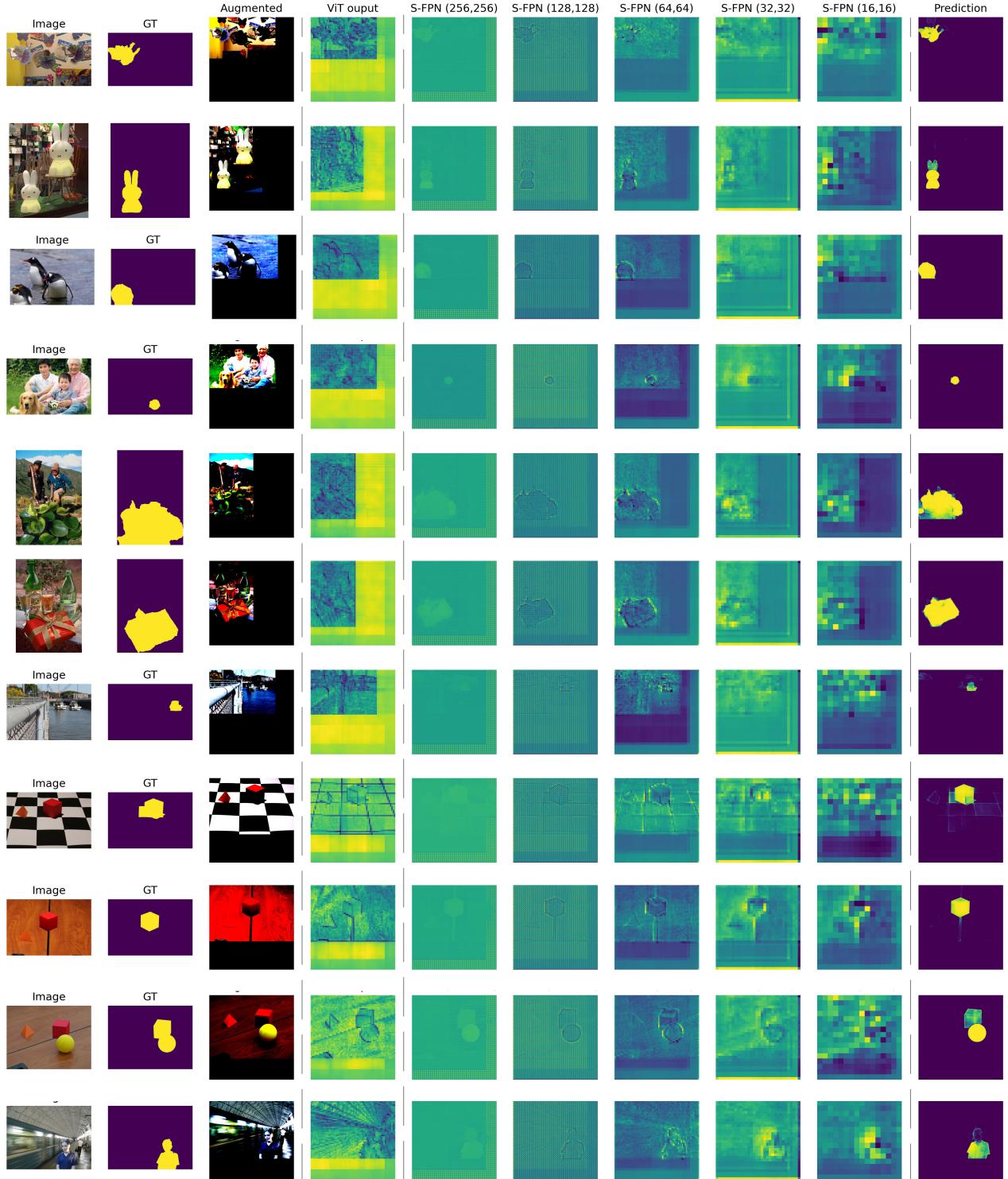


Figure 15. **Additional GradCAM results.** Datasets are collected from CASIAv1, NIST16, Coverage and Columbia. Attention around the manipulated region and long-range dependency could be observed, which is in line with our motivation to force the model to capture the artifacts and compare the relationships between regions explicitly.



**Figure 16. Visualization of the outputs from each component.** *GT* denotes ground truth; *Augmented* refers to the image after padding and normalize; *ViT output* is the feature map for the output of ViT backbone; *S-FPN* denotes the respective output of each resolution for different outputs in *simple feature pyramid*. For the output of ViT, since visualization takes the average of all channels, we cannot effectively observe the discrepancies between the manipulated region and the authentic region. However, we are pleased to see different types of feature expressions in the multi-level output of *S-FPN*. In the high-resolution output, more representation is given to larger, region-level “contrast differences”, while in the (64x64) feature map, we see the image focusing more on edge details and artifacts. This result is in line with the design logic of IML-ViT, which tracks tampering detection from both the perspective of comparing regional low-level differences and capturing detailed visible traces, proving the rationality and effectiveness of our IML-ViT.