M2SFormer: Multi-Spectral and Multi-Scale Attention with Edge-Aware Difficulty Guidance for Image Forgery Localization

Supplementary Material

Dataset Name	Total Images	Copy-Move	Splicing
CASIAv2 [53]	5,123	3,274	1,849
CASIAv1 [14]	920	459	461
DIS25k [61]	24,964	0	24,964
Columbia [26]	180	0	180
IMD2020 [52]	2,010	-	-
CoMoFoD [62]	260	260	0
In the Wild [29]	201	0	201
MISD [30]	300	0	300

Table 5. Summary of the datasets used in this paper.

Method	Parameters (M)	FLOPs (G)
UNet [58]	29.9	104.3
SegNet [5]	15.7	44.3
MantraNet [67]	3.63	135.7
RRUNet [7]	5.1	45.1
MT-SENet [77]	10.1	34.2
TransForensic [23]	27.6	22.4
MVSSNet [13]	136.2	31.6
FBINet [18]	104.4	86.5
SegNeXt [20]	28.4	19.1
CFLNet [51]	49.2	52.2
EITLNet [19]	50.1	35.0
PIMNet [6]	73.3	17.9
M2SFormer (Ours)	27.4	14.2

Table 6. The number of parameters (M), and FLOPs (G) of different models.

6. Dataset Descriptions

- **CASIAv1** [14] and **CASIAv2** [53]: The CASIAv1 dataset consists of JPG images with a resolution of 384 × 256, including 459 copy move and 461 splicing images. CASIAv2 is more complex than CASIAv1, containing 5,123 tampered ones which consists of 3274 copy move images and 1849 splicing images. The image sizes range from 320 × 240 to 800 × 600 and are available in multiple formats, such as uncompressed BMP and TIFF.
- **DIS25k** [61]: The DIS25k dataset is a large-scale image splicing dataset designed to enhance the realism and complexity of manipulated images. It contains 24,964 spliced images, generated using image composition techniques, such as deep image matting and harmonization, to improve the seamlessness of manipulation. The dataset was created by leveraging the OPA dataset for rational object placement and refining the images with advanced blending techniques, making the forgeries harder to detect. The image sizes vary but generally range from 512 × 512 to 1920 × 1080, ensuring diversity in resolution.
- **Columbia** [26]: The Columbia dataset is predominantly altered using splicing and contains high-resolution images. The manipulated regions often span large portions of scenes. Image sizes range from 757 × 568 to 1152 × 768 and are provided in TIFF or BMP formats. In total, it includes 180 tampered images.
- **IMD2020** [52]: The IMD2020 dataset features a diverse set of tampered images collected from real-world sources on the internet, totaling approximately 2,010 manipulated samples. The images are provided in JPG and TIFF formats.
- **CoMoFoD** [62]: The CoMoFoD dataset is specifically designed for Copy-Move Forgery Detection (CMFD) and comprises 260 forged image sets, divided into two resolution categories: small (512 × 512) and large (3000 × 2000). The images are classified into five groups based on the type of manipulation applied: translation, rotation, scaling, combination, and distortion. Various post-processing techniques, including JPEG compression, blurring, noise addition, and color reduction, are applied to both the tampered and original images.
- **In the Wild** [29]: The In-the-Wild dataset is a collection of 201 manipulated images gathered from online sources such as THE ONION (a parody news website) and REDDIT PHOTOSHOP BATTLES (an online community focused on image manipulations). These images represent real-world, naturally occurring spliced forgeries. The images in this dataset come in various sizes, reflecting the diversity of online manipulations.
- **MISD** [30]: The Multiple Image Splicing Dataset (MSID) is the first publicly available dataset specifically designed for multiple image splicing detection. It contains 300 multiple spliced images, all in JPG format with a resolution of 384 × 256. The dataset was created by combining images from the CASIAv1 dataset and applying multiple splicing operations using Figma software. The spliced images feature various post-processing techniques such as rotation and scaling to enhance realism.

7. Technical Novelty of M2SFormer

M2SFormer introduces a unified framework that integrates multi-frequency and multi-scale attention in a single stream, addressing a longstanding challenge in forgery localization where frequency- and spatial-domain features were traditionally processed separately. By enriching the encoder–decoder pipeline with the *M2S Attention Block*, it effectively captures both local, fine-grained anomalies and global, context-aware cues within the same network pass. Furthermore, M2SFormer leverages a *Difficulty-guided Attention (DGA) mechanism* that quantifies each sample's complexity through a curvature-based global prior map, then automatically generates textual difficulty cues to guide the Transformer decoder's focus on challenging regions. This approach removes the need for extra metadata and provides adaptive attention control, ultimately boosting cross-domain generalization and boundary precision. The design also emphasizes computational efficiency, fusing multi-frequency representations within the feature space rather than the raw input space, thereby maintaining a balance between performance gains and inference speed. Consequently, M2SFormer stands out for its holistic combination of frequency- and spatial-domain attention, adaptive difficulty-based text guidance, and efficient, robust architecture—all contributing to significant improvements in forgery localization across diverse domains.

8. Broader Impact in Artificial Intelligence

M2SFormer's unified approach to forgery localization—combining *multi-spectral and multi-scale attention* with the *text-guided difficulty attention*—not only enhances detection accuracy for unseen or subtle manipulations but also carries significant broader implications for AI. By capturing both spatial and frequency-domain cues, the model strengthens its cross-domain generalization, **assisting in the fight against misinformation across social media, journalism, and legal contexts**. Its Difficulty-guided Attention mechanism offers a scalable strategy for other AI tasks that require dynamic allocation of computational resources, potentially improving performance in areas like object detection, medical imaging, and fine-grained recognition. Furthermore, the integration of textual cues into a visual framework underscores the promise of cross-modal solutions, paving the way for more holistic approaches to content understanding. *While adversarial refinement of forgery techniques remains a concern, M2SFormer sets a higher bar for detecting tampered content, promoting authenticity, accountability, and ethical AI development*.

9. More Detailed Ablation Study on M2SFormer

In this section, we perform a more detailed ablation study on M2SFormer.

9.1. Real-world Scenario Evaluation

We conducted additional robustness tests on CASIAv2 by applying Gaussian Blur, JPEG Compression, and Gaussian Noise during the inference stage. As summarized in Tab. 7, M2SFormer still consistently outperformed recent models (EITLNet, PIMNet) across various digital distortions, demonstrating ***superior robustness in diverse real-world digital scenarios***.

Method	Clean	Gaussian Blur (k)			JPEG Compression (q)			Gaussian Noise (σ)		
		3	5	9	100	50	10	0.1	0.3	0.5
EITLNet	(54,76)	(***52***,75)	(***45***,***71***)	(***32***,***65***)	(15,56)	(8,53)	(4,51)	(***14***,55)	(***13***,***55***)	(***12***,***54***)
PIMNet	(***56***,***81***)	(***52***,***80***)	(30,67)	(19,60)	(***18***,***58***)	(***14***,***56***)	(***10***,***53***)	(***14***,56)	(12,54)	(10,52)
Ours	(**59**,**84**)	(**57**,**83**)	(**49**,**79**)	(**38**,**74**)	(**23**,**63**)	(**16**,**58**)	(**12**,**55**)	(**23**,**63**)	(**22**,**63**)	(**21**,**62**)

Table 7. Robustness performance using two metrics **(DSC, AUC)** on CASIAv2 under clean and three corruption types according to three corruption severity parameters.

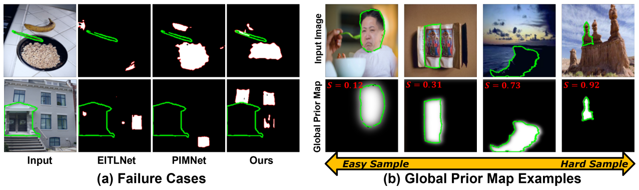

Figure 5. (a) Examples of failure cases indicating false positives in visually complex, non-manipulated regions. (b) Visualization of Global Prior Maps, clearly illustrating regions with varying detection difficulties (from easy to hard). S is difficulty level calculated by Algorithm 1.

9.2. Ablation Study on Backbone in M2SFormer

Network Type	Backbone	*Seen*		*Unseen*		Param (M)	FLOPs (G)
		DSC	mIoU	DSC	mIoU		
CNN	ResNet50	40.3	32.7	35.6	26.3	25.7MB	14.9GB
	Res2Net50	49.9	42.1	38.1	29.1	25.9MB	15.9GB
	ResNeSt50	51.1	43.2	38.5	29.5	27.6MB	18.0GB
Transformer	MiT-B2	***56.5***	***48.7***	***41.2***	***32.6***	26.2MB	12.3GB
	P2T-Small	55.8	48.1	40.6	32.1	25.6MB	13.5GB
	PVT-v2-B2 **(Ours)**	**57.8**	**50.8**	**43.0**	**34.3**	27.4MB	14.2GB

Table 8. Quantitative results on *Seen* (CASIAv2 [53]) and *Unseen* datasets (Other test datasets) according to backbone network. The efficiency metrics are measured at resolution 256 × 256.

In this section, we conduct an ablation study to evaluate the impact of different backbone models on the performance of M2SFormer. This experiment uses several popular CNN and Transformer architectures, including ResNet50 [24], Res2Net [17], ResNeSt50 [76], MiT-B2 [20], PVT-v2-b2 [65], and P2T-Small [68]. Notably, only the backbone network was changed, while all other architectural settings remained consistent with those in the main experiment. We reported the *mean* performance of five-fold cross-validation results for each *Seen* (CASIAv2 [53]) and *Unseen* datasets (Other test datasets) in Tab. 8. The datasets used in the *seen* and *unseen* datasets are the same as Tab. 1.

9.3. Examples of Failure Cases and Global Prior Maps

We analyzed failure cases (Fig. 5 (a)) and found false positives mainly in complex or high-frequency regions. Additionally, we also provide examples of inputs and corresponding global prior maps G (Fig. 5 (b)).

10. Metrics Descriptions

In this section, we describe the metrics used in this paper. For convenience, we denote TP, FP, and FN as the number of samples of true positive, false positive, and false negative between two binary masks A and B.

- The *Mean Dice Similarity Coefficient (DSC)* [48] measures the similarity between two samples and is widely used to assess the performance of segmentation tasks, such as image segmentation or object detection. **Higher is better**. For given two binary masks A and B, DSC is defined as follows:

$$\mathbf{DSC}(A, B) = \frac{2 \times |A \cap B|}{|A \cup B|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

- The *Mean Intersection over Union (mIoU)* measures the ratio of the intersection area to the union area between predicted and ground truth masks in segmentation tasks. **Higher is better**. For given two binary masks A and B, mIoU is defined as follows:

$$\mathbf{mIoU}(A, B) = \frac{A \cap B}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{9}$$

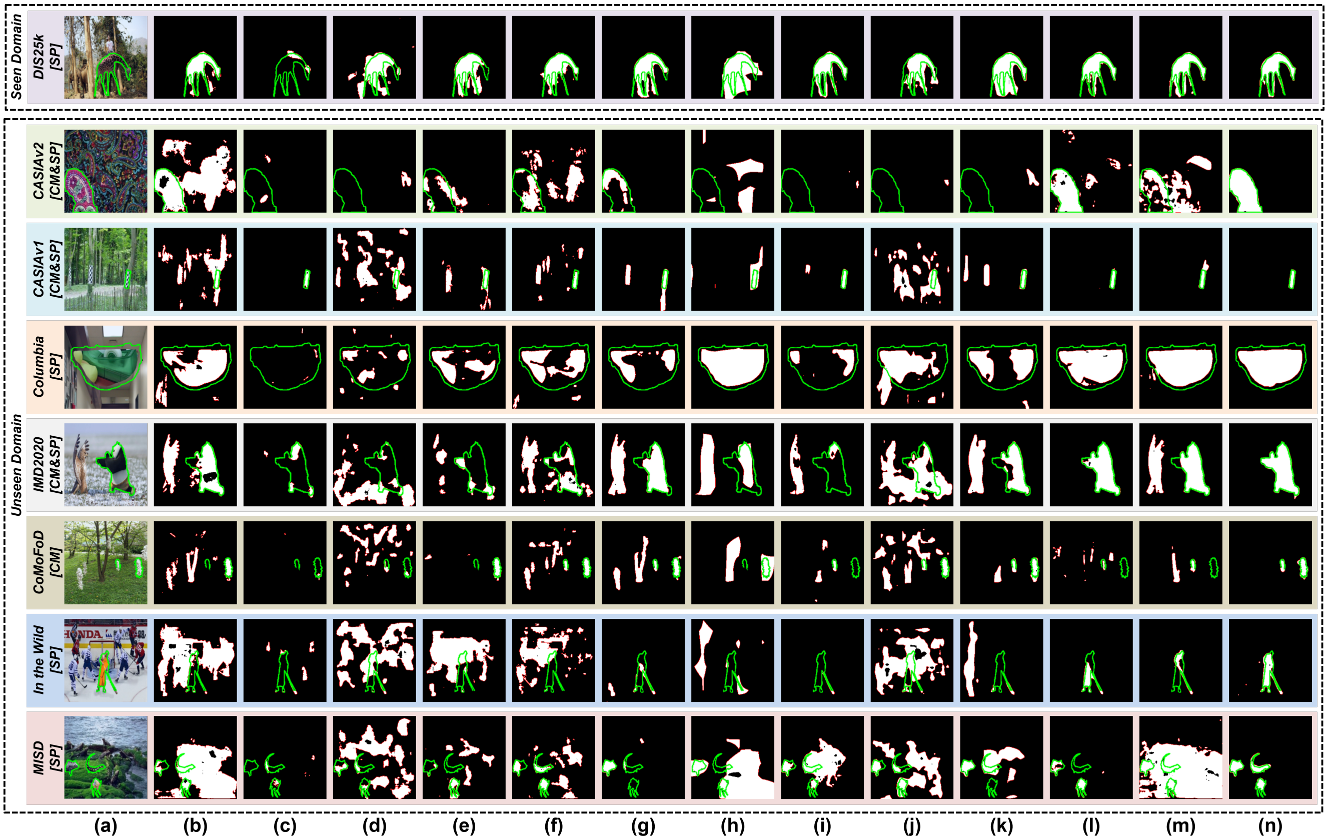

Figure 6. Qualitative comparison of other methods and M2SFormer with DIS25k training scheme. (a) Input images with ground truth. (b) UNet [58]. (c) SegNet [5]. (d) MantraNet [67]. (e) RRUNet [7]. (f) MT-SENet [77]. (g) TransForensic [23]. (h) MVSSNet [13]. (i) FBINet [18]. (j) SegNeXt [20]. (k) CFLNet [51]. (l) EITLNet [19]. (m) PIMNet [6]. (n) **M2SFormer (Ours)**. **Green** and **Red** lines denote the boundaries of the ground truth and prediction, respectively. And, the "SP" and "CM" in square brackets represent the types of forgery in the dataset: "SP" stands for Splicing, while "CM" denotes Copy-Move.