COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

INHA UNIVERSITY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Highlight Paper

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


Ju-Hyeon Nam


Dong-Hyun Moon


Sang-Chul Lee

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING OF INHA UNIVERSITY, REPUBLIC OF KOREA

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Image Manipulation

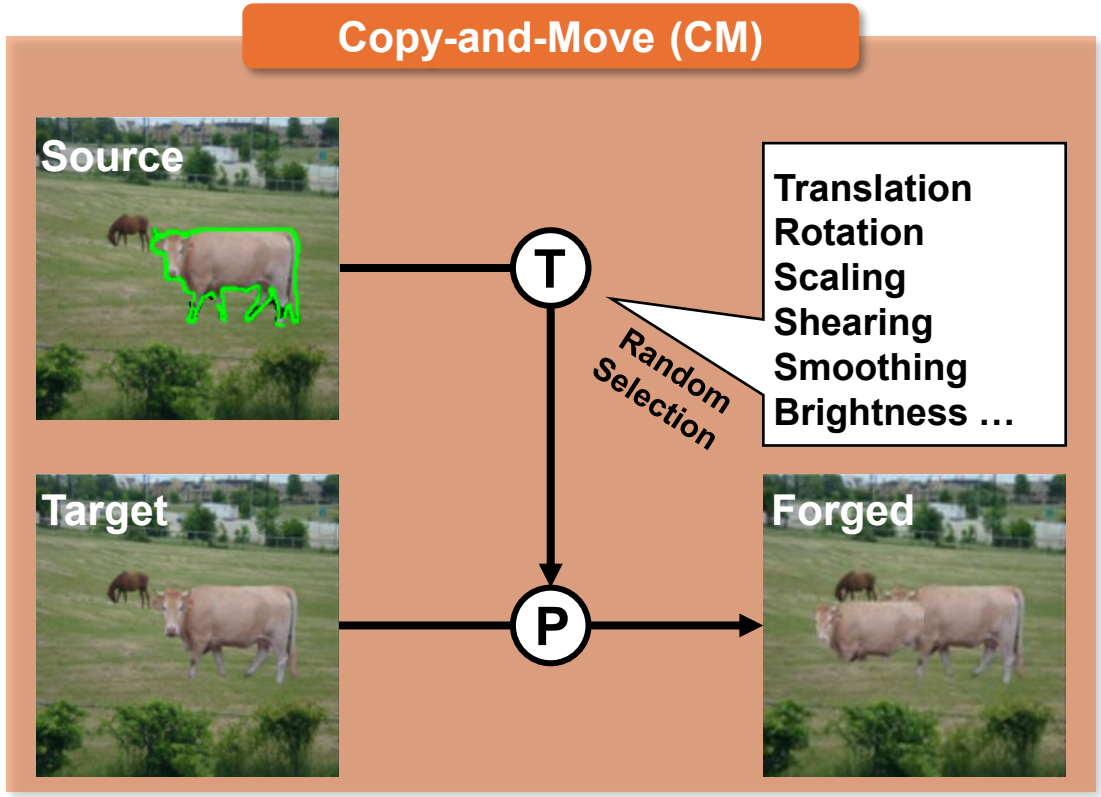
Forged Image Example

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

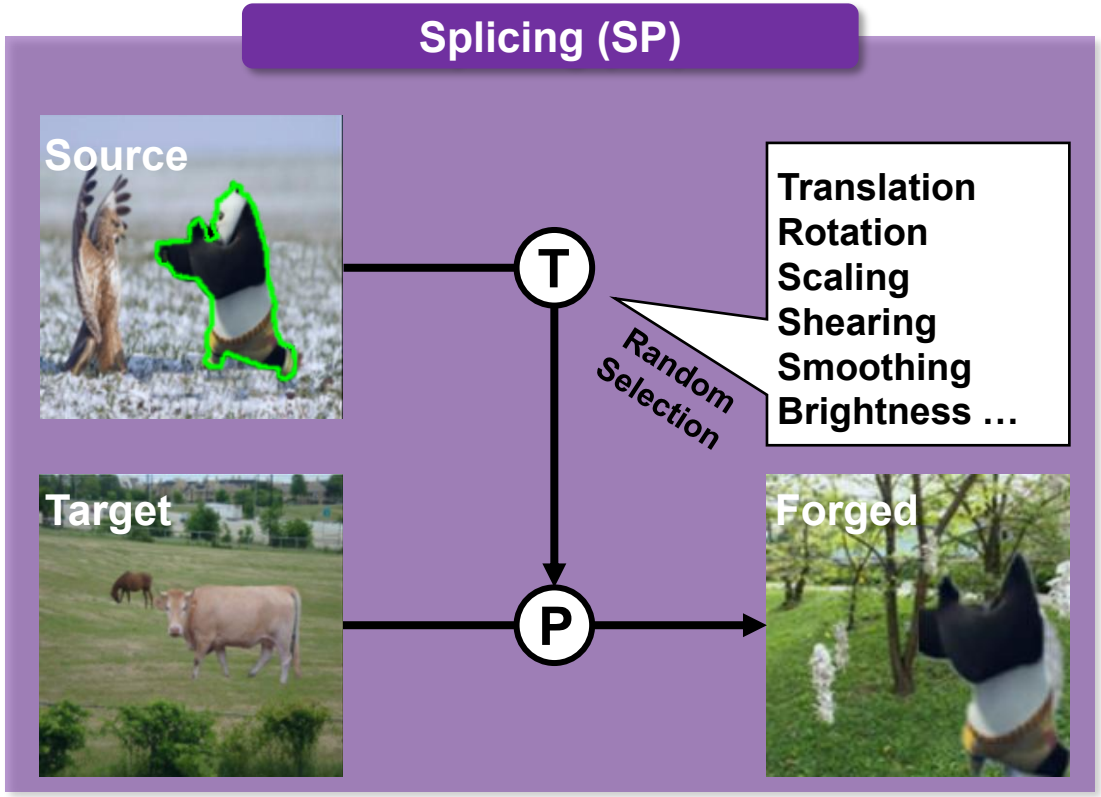
ICCV HONOLULU HAWAII
OCT 19-23, 2025

Image Manipulation

Forged Image Example



Source = Target



Source ≠ Target

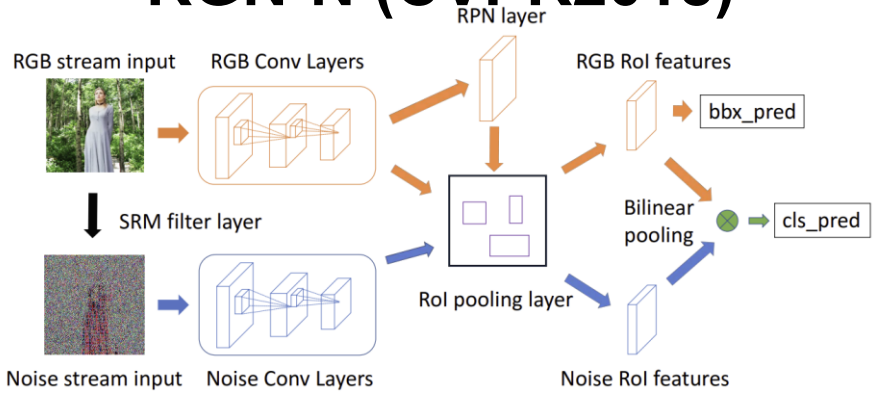
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Image Forgery Detection & Localization Method

CNN-based Approaches

RGN-N (CVPR2018)



RRUNet (CVPRW2019)

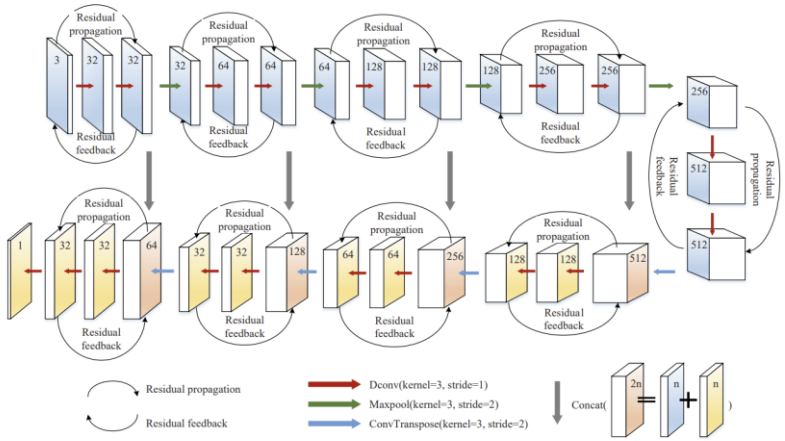
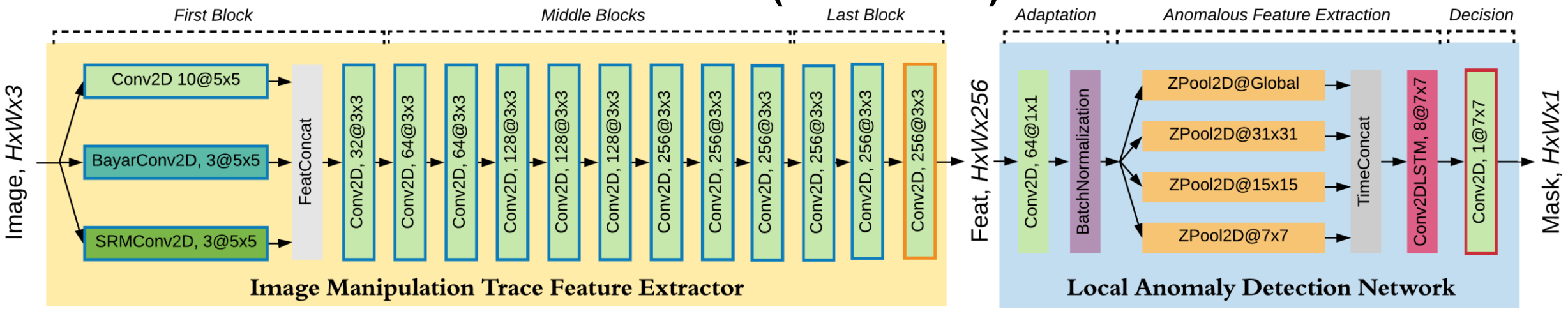


Figure 5. The network architecture of RRU-Net. The number on the box represents the number of features.

MantraNet (CVPR2019)



COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

Image Forgery Detection & Localization Method

Transformer-based Approaches

TransForensic (ICCV2021)

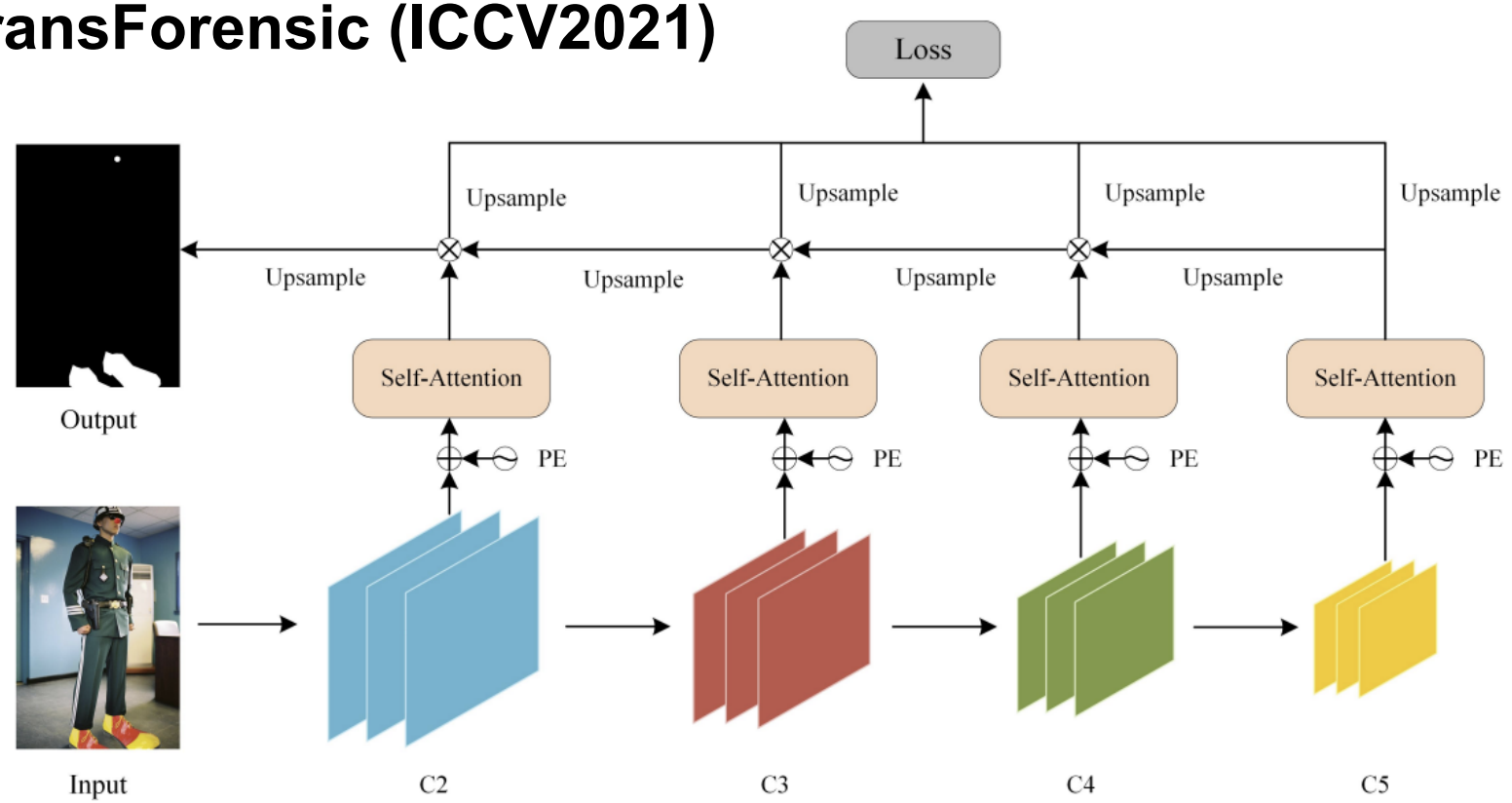


Image Forgery Detection & Localization Method

Frequency-based Approaches

FBINet (Access2022)

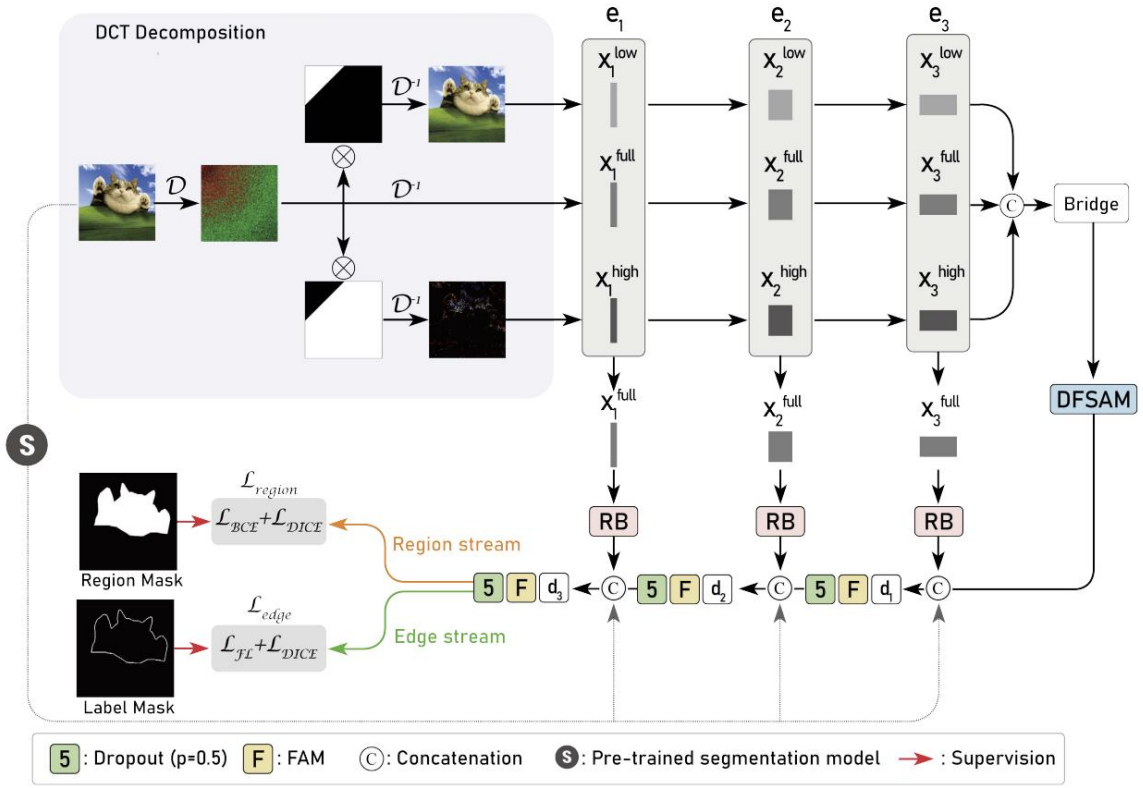
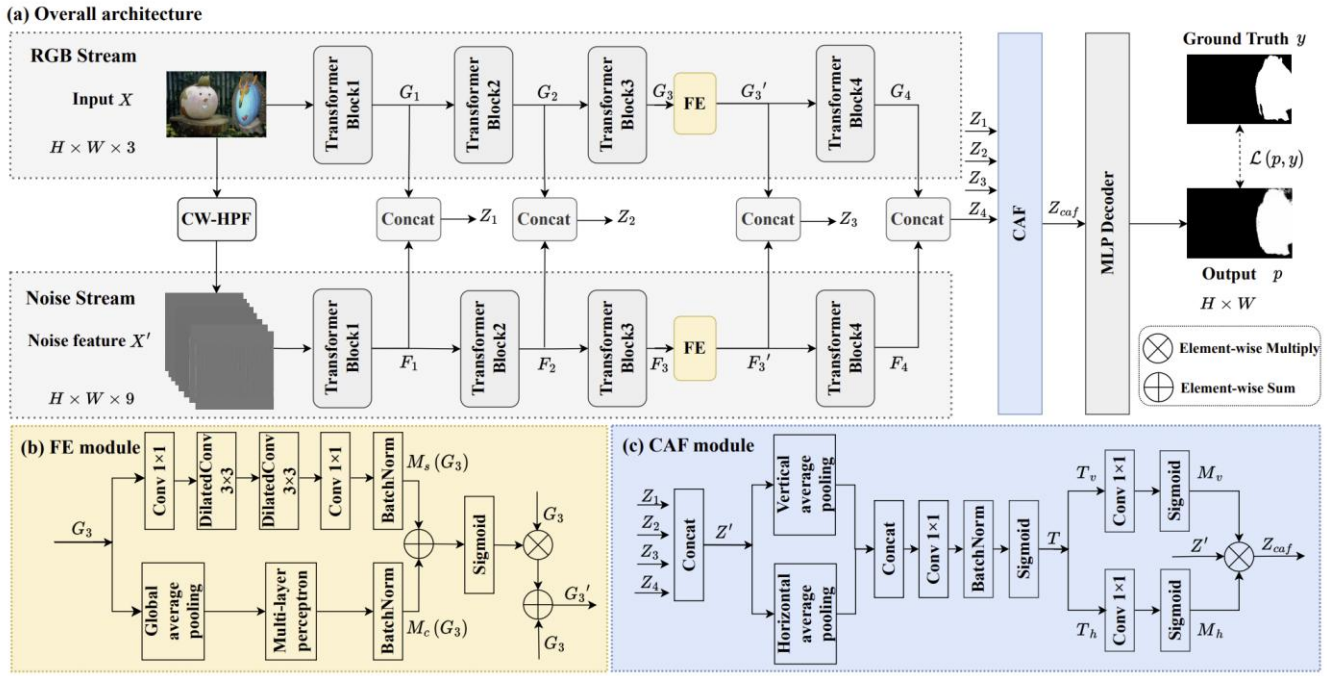


FIGURE 1. The procedure of the proposed FBI-Net for forgery localization can be divided into three main parts. Firstly, we decompose an input image as a low-/high-frequency image with DCT. Then, each image is passed into the shared encoder and applied to DFSAM. Lastly, high dimensional feature maps are passed in the decoder, and we calculate the multi-task loss(region and edge loss).

EITLNet (ICASSP2024)



COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

Image Forgery Detection & Localization Method

Key Challenges in IFL Task

- Existing detectors overfit to datasets and ***fail on unseen forgeries and domains***.

  - ✓ **Different forensic fingerprints and feature targets**

    - CM duplicate content within the same image, so camera noise, CFA/demosaicing, and JPEG statistics remain consistent; detectors hunt for near-duplicate patches under small geometric changes.

    - SP merges content from different images, where cues are cross-image inconsistencies (noise/PRNU level, white balance, JPEG grids) and boundary resampling

  - ✓ **Editing pipelines and dataset bias → domain shift**

    - CM typically applies local affine transforms and mild blending while preserving the image's overall style

    - SP involves compositing, color matching/relighting, alpha matting, and often double compression

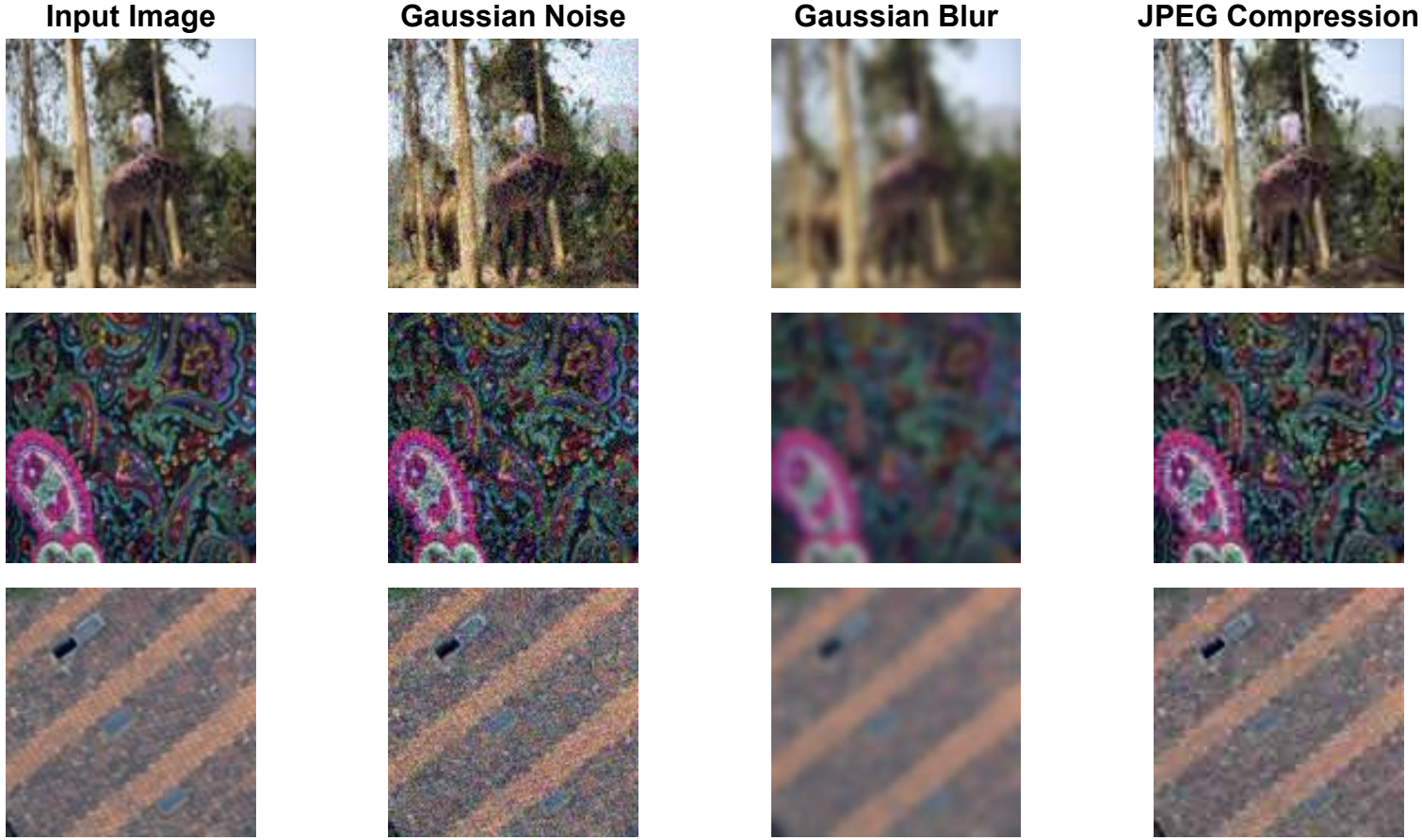
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Image Forgery Detection & Localization Method

Key Challenges in IFL Task

- Accuracy collapses under ***common corruptions*** leading to unstable localization.



COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

Contribution

- Propose a novel Transformer-based forgery localization framework, called *M2SFormer*, that efficiently integrates M2S attention block with Edge-Aware DGA-based Transformer decoder

  - M2S Attention Block → The integration of multi-spectral and multi-scale attention in the skip connection to better capture forgery artifacts.

  - Edge-Aware DGA-based Transformer Decoder → Difficulty-guided attention module after upsampling to preserve fine details in challenging regions.

- Extensive experiments on multiple benchmark datasets demonstrate that M2SFormer outperforms existing models, significantly improving generalization performance in forgery localization across unseen domains and common corruption.

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

Motivation: M2S Attention Block

- Problem: Subtle or complex manipulations are often missed when models rely only on spatial cues; frequency signals expose such artifacts more reliably.

  - **Biological Insight**: The HVS leverages *multiple frequency bands*; likewise, attending to selected DCT components can surface delicate tampering traces while keeping spatial context.

  - **Scale variation:** Forgery patterns appear at diverse sizes; scale-invariant cues (e.g., SIFT-style pyramids) help capture anomalous boundaries across resolutions

✓ **Fuse multi-spectral** (DCT-based channel attention) **and multi-scale** (pyramidal spatial attention) in the **skip connections** to inject richer, globally informed features into the decoder **without heavy compute**.

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

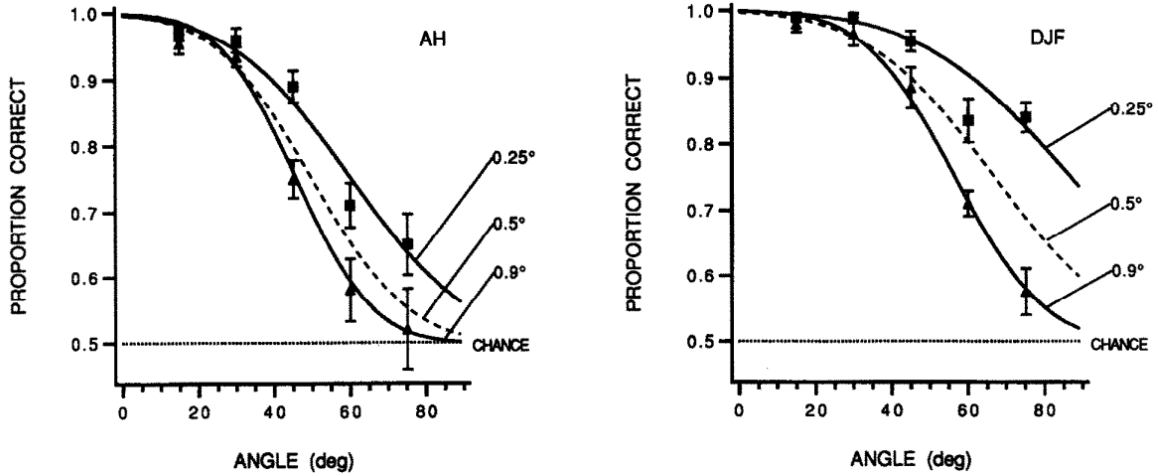
M2SFormer

Motivation: Edge-Aware DGA-based Transformer Decoder

- Problem: Upsampling in encoder–decoder pipelines tends to lose fine details, precisely where tampered edges/thin structures occur; the model needs guided emphasis during decoding.

  - **Perceptual cue: *Curvature*** reflects shape complexity and guides human visual attention; emphasizing high-curvature, edge-rich regions should improve localization of subtle manipulations.

✓ Compute a **global prior map**, derive **edge-aware curvature**, summarize difficulty and convert it into a **text embedding** that **gates channel attention** at each decoder stage (**DGA**).

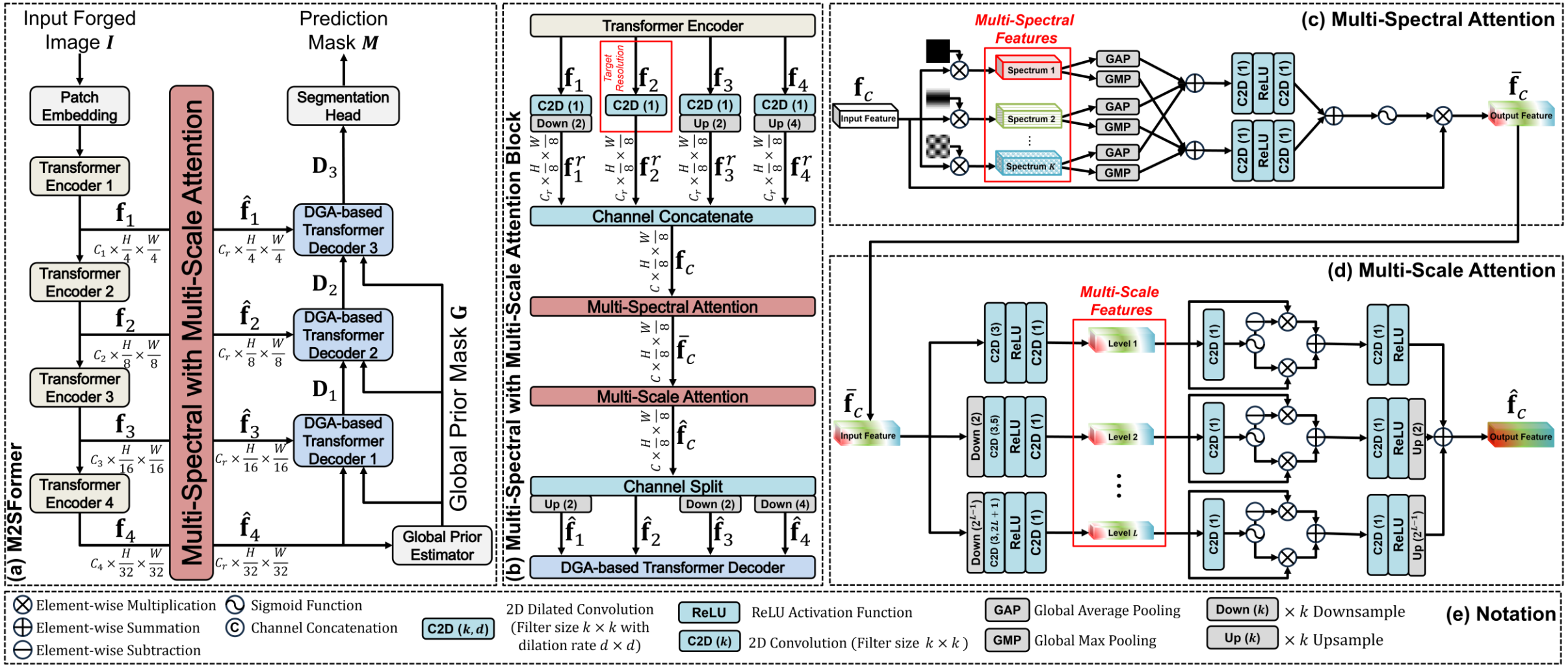


COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

Overall Architecture



COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

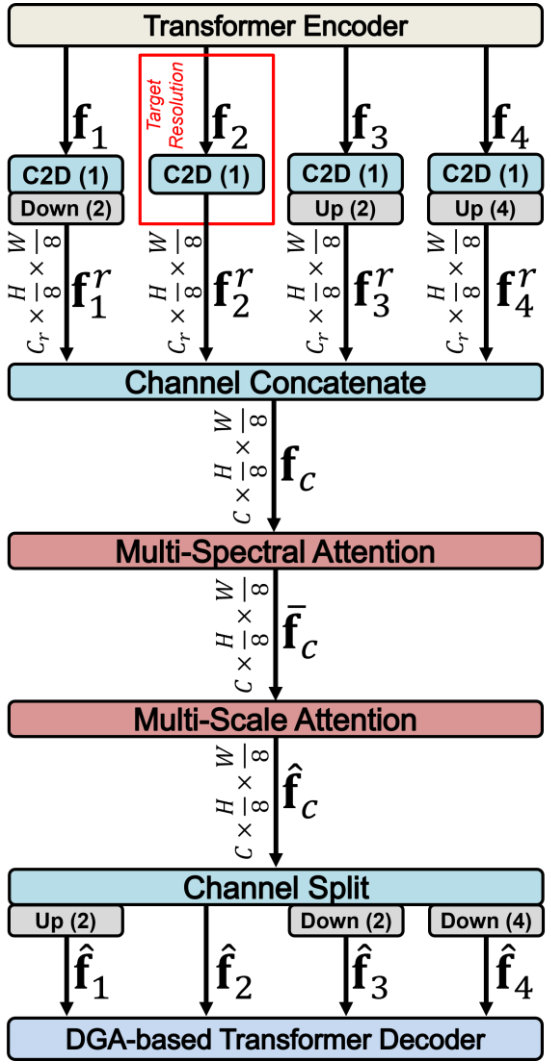
STEP1. [Feature Pre-processing] Align features & Control computational complexity

STEP2. [Multi-Spectral Attention] Highlight informative frequency bands

STEP3. [Multi-Scale Attention] Emphasize boundary cues across scales with low memory

STEP4. [Feature Post-processing] Return refined skip connection features to the decoder

✓ M2S unifies **multi-spectral** and **multi-scale** attention inside skip connections, delivering frequency-aware, edge-sharp features to the decoder for robust forgery localization.



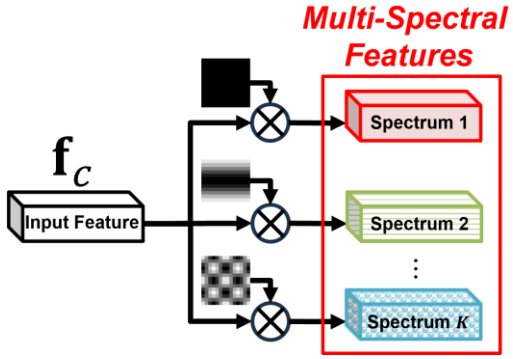
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

STEP2. [Multi-Spectral Attention] Highlight informative frequency bands



- Project f_c onto top-K 2D DCT basis images to obtain spectral components $\{f_c^k\}$

$$f_c^k = \sum_{h=0}^{H_t-1} \sum_{w=0}^{W_t-1} (f_c)_{:,h,w} D_{h,w}^{u_k,v_k}$$

  - $D_{h,w}^{u_k,v_k} = \cos\left(\frac{\pi h}{H_s}(u_k + \frac{1}{2})\right)\sin\left(\frac{\pi w}{W_s}(v_k + \frac{1}{2})\right)$ with top-K selection strategy for frequency indices (u_k, v_k)

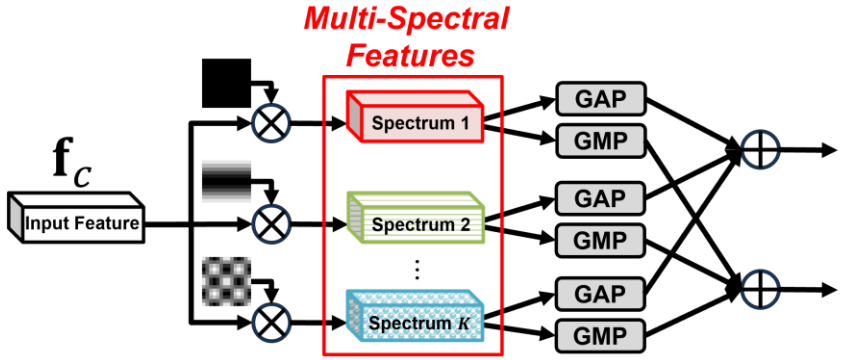
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

STEP2. [Multi-Spectral Attention] Highlight informative frequency bands



- For each $\mathbf{f}_c^k$, apply GAP/GMP to extract pooled statistics and aggregate them

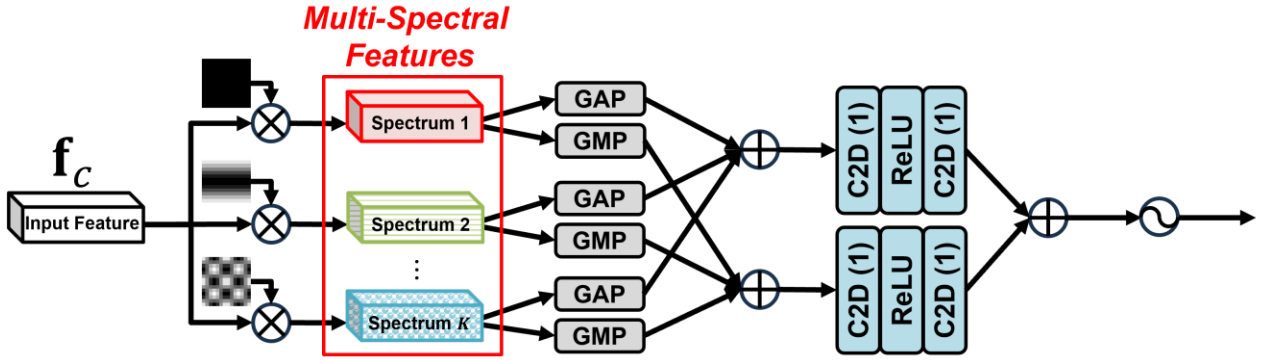
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

STEP2. [Multi-Spectral Attention] Highlight informative frequency bands



- For each aggregated statistic, sum over k to obtain a channel attention map $\mathrm{M}^{\mathrm{spectral}}$ by conv-ReLU-conv Block and Sigmoid operation

$$\mathrm{M}^{\mathrm{spectral}} = \sigma\left(\sum_{\mathrm{d}\in\{\mathrm{avg},\mathrm{max}\}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{C2D}_{1\times1}(\delta(\mathrm{C2D}_{1\times1}(\mathrm{f}_{\mathrm{c}}^{\mathrm{k}})))\right)$$

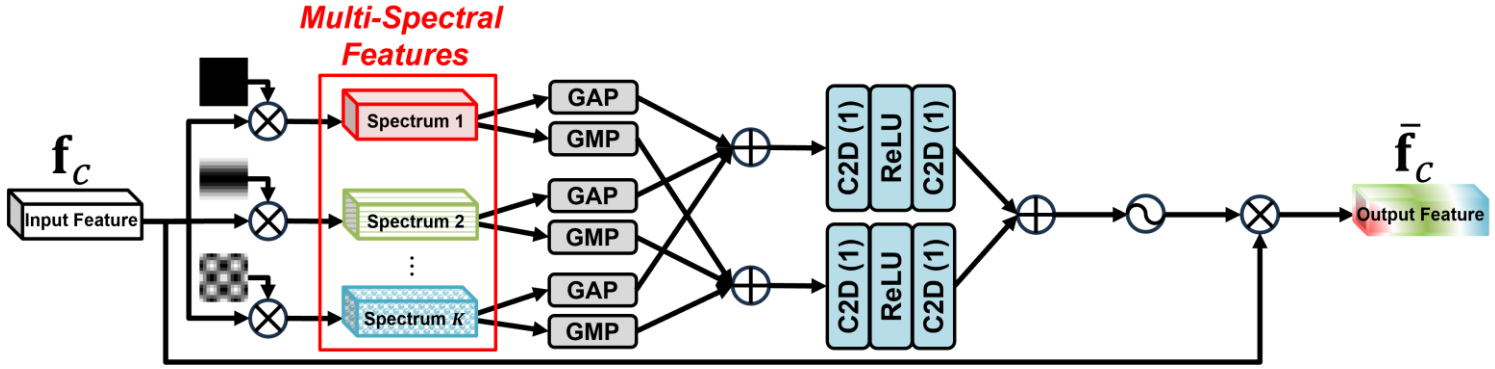
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

STEP2. [Multi-Spectral Attention] Highlight informative frequency bands



- Recalibrate channels of $\mathbf{f}_c$ using multi-spectral channel attention map $\mathbf{M}^{\text{spectral}}$

$$\bar{\mathbf{f}}_c = \mathbf{f}_c \times \mathbf{M}^{\text{spectral}}$$

✓ Subtle manipulations often manifest in specific frequencies; channel-wise weighting surfaces these cues while keeping spatial context.

➢ Channel recalibrated features are forwarded to Multi-Scale Attention [STEP3]

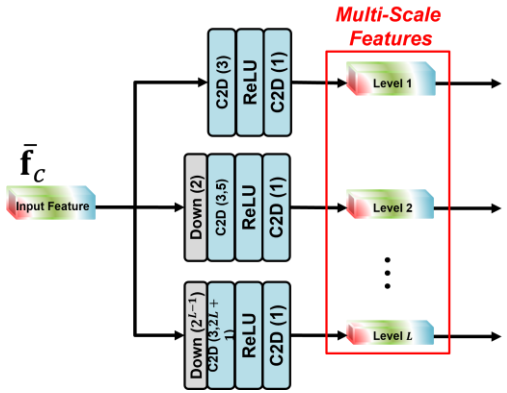
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

M2S Attention

STEP3. [Multi-Scale Attention] Emphasize boundary cues across scales with low memory



- Build a **feature pyramid** by downsampling $\bar{f}_c$ to L levels by inspired SIFT feature extractor
- At each level l, use 3×3 dilated convolution with dilation $2l + 1$ then 1×1 convolution to preserve spatial information

$$\bar{f}_c^l = \mathrm{C2D}_{1\times1}(\mathrm{DC2D}_{3\times3}^{2l+1}(\mathrm{Down}_l(\bar{f}_c)))$$

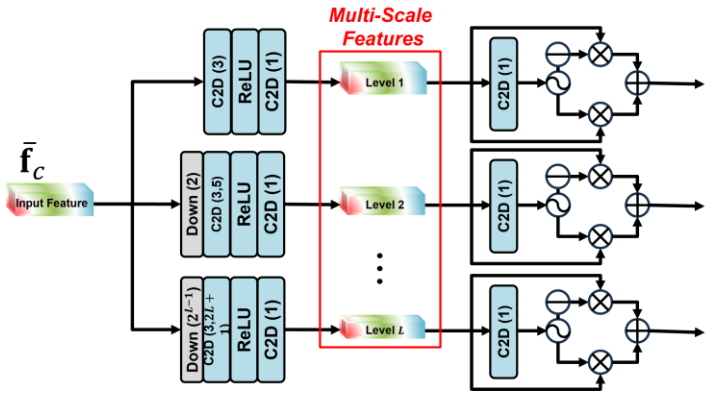
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

M2SFormer

M2S Attention

STEP3. [Multi-Scale Attention] Emphasize boundary cues across scales with low memory



- Compute a foreground map F_l and background map $B_l = 1 - F_l$ and blend them via learnable parameters α^l and β^l

$$\hat{f}^l = C2D_{3\times3}(\alpha^l(\bar{f}_c \times F_l) + \beta^l(\bar{f}_c \times B_l))$$

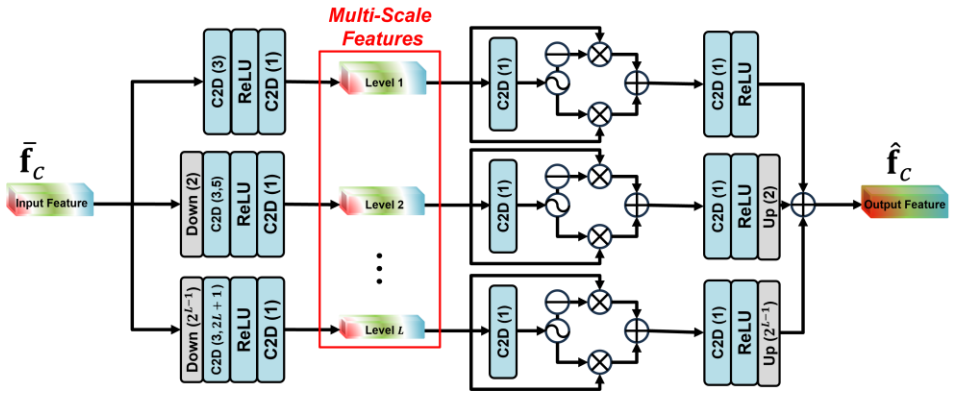
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

M2SFormer

M2S Attention

STEP3. [Multi-Scale Attention] Emphasize boundary cues across scales with low memory



- Upsample and sum all refined scales

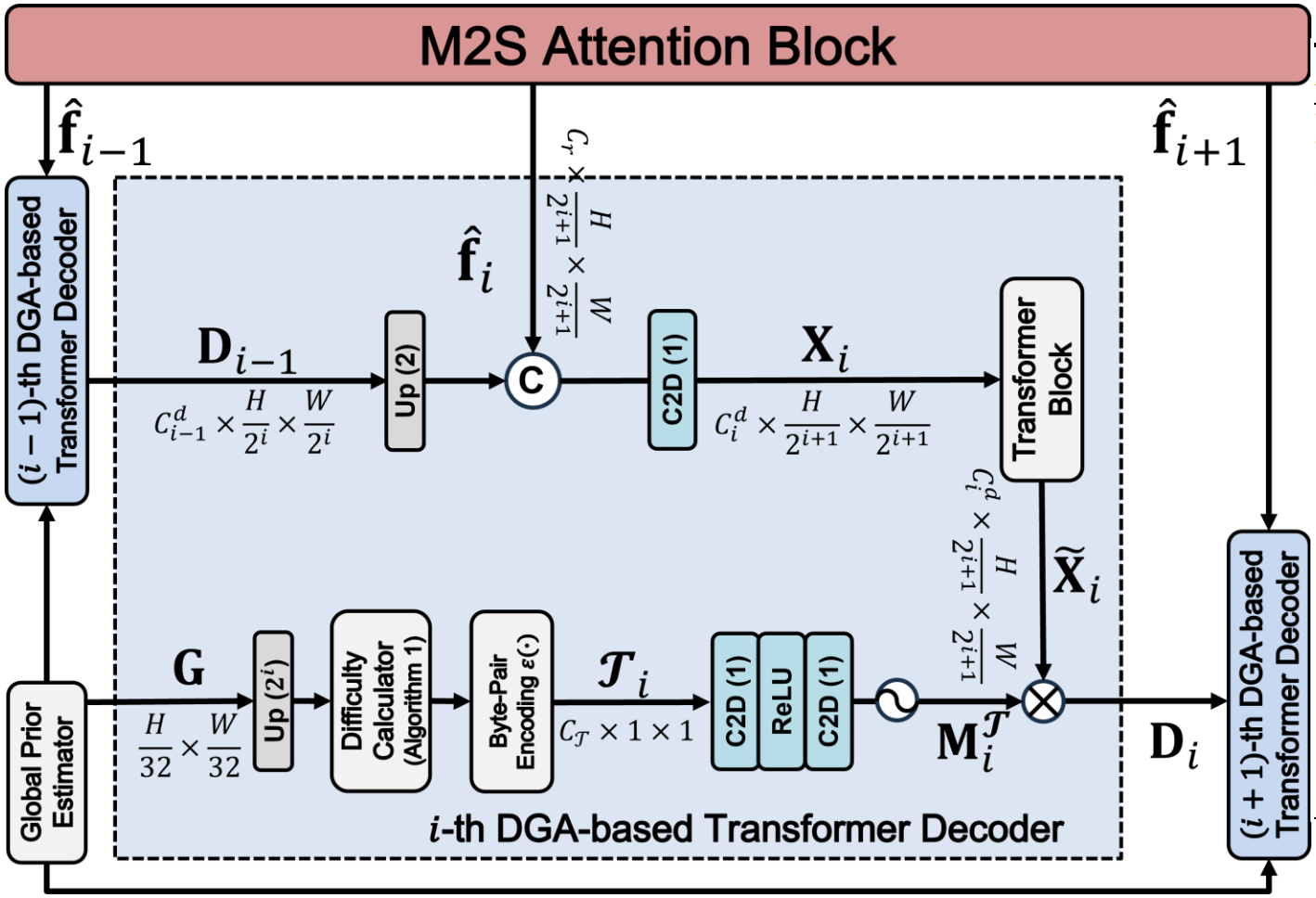
✓ Forged boundaries appear at diverse sizes; scale-aware spatial attention sharpens edges while keeping memory in check.

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

Edge-Aware DGA-based Transformer Decoder



Algorithm 1 Edge-Aware Difficulty Calculator

Input: Global prior map $\mathbf{G}$
Output: Difficulty level

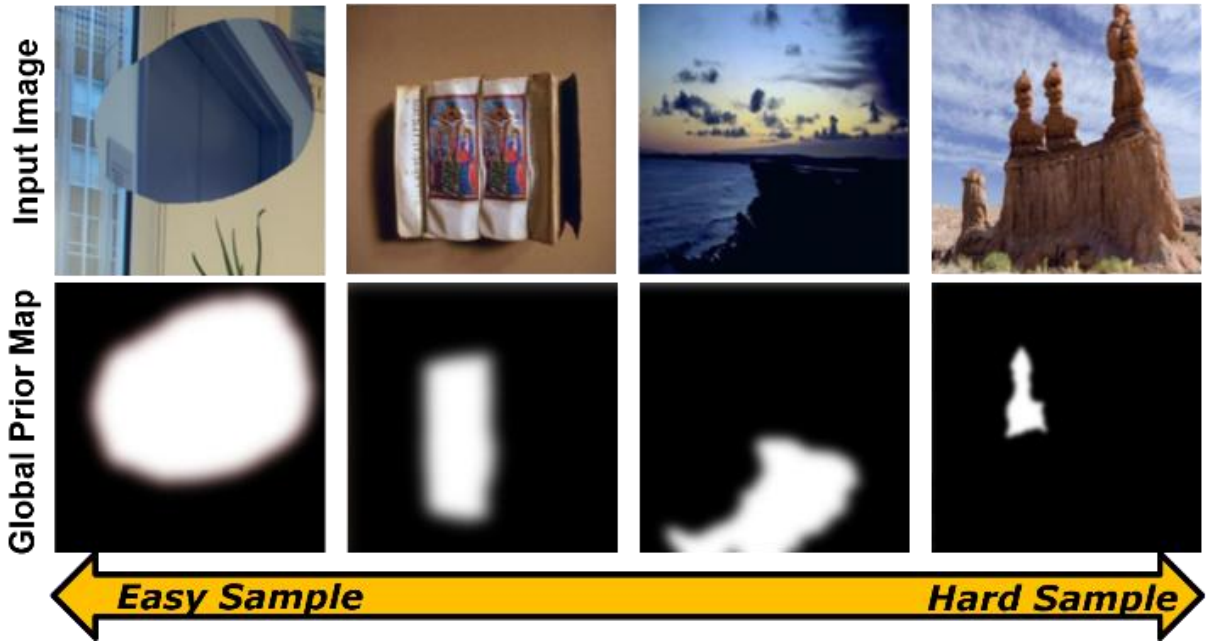
$\mathbf{G}_x, \mathbf{G}_y \leftarrow \mathbf{Sobel}(\mathbf{G})$
$\mathbf{G}_{xx}, \mathbf{G}_{xy} \leftarrow \mathbf{Sobel}(\mathbf{G}_x)$
$\mathbf{G}_{yx}, \mathbf{G}_{yy} \leftarrow \mathbf{Sobel}(\mathbf{G}_y)$
$\kappa \leftarrow (\mathbf{G}_x^2\mathbf{G}_{yy} - 2\mathbf{G}_x\mathbf{G}_y + \mathbf{G}_y^2\mathbf{G}_{xx})/(\mathbf{G}_x^2 + \mathbf{G}_y^2)^{1.5}$
$\mathbf{E} \leftarrow \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$
$\mathbf{S} \leftarrow \sigma(\sum(\kappa \otimes \mathbf{E})/\sum \mathbf{E})$
if $\mathbf{S} \geq 0.5$ **then**
    return "hard"
else
    return "easy"
end if

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

M2SFormer

Edge-Aware DGA-based Transformer Decoder



Why Difficulty Helpful?

Difficulty reveals where ***models are most likely to fail,*** ***turning uncertainty*** into guidance for better localization

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Quantitative Results

CASIAv2 Training Scheme Results

Method	Seen Domain		Unseen Domain													
	CASIAv2 [53]		DIS25k [61]		CASIAv1 [14]		Columbia [26]		IMD2020 [52]		CoMoFoD [62]		In the Wild [29]		MISD [30]	
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
UNet [58]	32.3 (10.9)	25.8 (9.2)	8.7 (1.2)	5.5 (0.8)	25.0 (1.9)	19.5 (1.8)	19.8 (2.5)	12.2 (1.7)	14.9 (1.0)	9.7 (0.6)	12.2 (1.0)	7.6 (0.7)	18.6 (1.4)	11.9 (1.0)	47.3 (2.5)	34.8 (2.2)
SegNet [5]	8.1 (6.6)	6.1 (5.0)	1.1 (0.7)	0.7 (0.4)	5.0 (1.4)	3.7 (1.0)	7.0 (3.7)	4.3 (2.5)	6.3 (2.4)	4.4 (1.6)	3.5 (2.0)	2.3 (1.5)	5.8 (2.8)	3.9 (1.7)	20.4 (5.7)	13.8 (4.2)
MantraNet [67]	18.8 (8.0)	11.9 (5.7)	12.7 (0.8)	7.4 (0.6)	19.8 (1.0)	12.1 (0.8)	25.0 (1.8)	14.9 (1.2)	14.1 (0.5)	8.2 (0.3)	11.2 (0.6)	6.5 (0.4)	18.2 (1.1)	10.6 (0.8)	30.7 (3.5)	19.2 (2.5)
RRUNet [7]	21.8 (10.4)	15.8 (8.4)	10.8 (2.2)	6.9 (1.5)	26.7 (2.4)	18.9 (1.7)	21.3 (7.2)	13.9 (5.3)	14.5 (1.9)	9.4 (3.3)	13.7 (2.2)	9.0 (1.7)	18.3 (4.2)	11.7 (3.0)	32.8 (3.4)	21.4 (2.5)
MT-SENet [77]	19.9 (9.9)	14.6 (7.6)	7.8 (1.0)	4.8 (0.7)	18.4 (1.6)	12.8 (1.3)	9.4 (1.1)	5.3 (0.6)	11.4 (1.3)	7.2 (1.0)	10.6 (2.1)	6.5 (1.5)	10.6 (2.1)	6.5 (1.5)	22.3 (2.5)	14.0 (1.8)
TransForensic [23]	40.1 (15.7)	32.0 (13.9)	32.3 (2.9)	24.1 (2.5)	44.2 (1.6)	35.0 (1.7)	*35.9* (5.7)	*25.0* (4.6)	27.2 (2.1)	19.1 (1.6)	21.7 (2.0)	14.3 (1.5)	*31.9* (3.8)	22.4 (3.0)	60.0 (1.8)	46.5 (2.1)
MVSSNet [13]	31.2 (13.4)	23.4 (11.1)	24.9 (3.6)	17.4 (2.9)	36.6 (1.6)	27.3 (1.5)	33.8 (3.7)	23.3 (3.0)	22.8 (2.4)	15.2 (1.8)	17.2 (1.2)	10.9 (0.9)	27.0 (3.1)	18.3 (2.2)	53.9 (3.6)	17.4 (2.9)
FBINet [18]	35.3 (14.3)	29.2 (12.8)	26.3 (2.3)	20.3 (1.9)	37.5 (2.1)	30.8 (1.9)	18.2 (2.7)	11.7 (2.1)	24.2 (1.6)	17.5 (1.3)	22.3 (0.8)	15.2 (0.5)	25.0 (2.7)	17.8 (2.1)	48.0 (2.2)	35.1 (2.0)
SegNeXt [20]	12.8 (4.2)	8.6 (3.0)	9.6 (1.1)	5.6 (0.7)	14.8 (2.0)	9.4 (1.5)	12.8 (1.4)	7.3 (0.9)	11.2 (1.3)	6.6 (0.8)	8.4 (1.1)	4.8 (0.6)	13.9 (2.3)	8.1 (1.5)	21.1 (3.0)	12.7 (2.0)
CFLNet [51]	40.4 (15.8)	33.4 (14.2)	25.5 (1.7)	18.9 (1.3)	38.0 (0.8)	31.1 (0.5)	16.7 (2.8)	10.6 (2.0)	22.7 (1.4)	15.8 (1.0)	20.5 (2.0)	14.0 (1.5)	25.0 (2.3)	17.3 (1.7)	*61.6* (1.5)	*48.3* (1.6)
EITLNet [19]	54.0 (14.7)	47.9 (14.7)	30.8 (2.8)	25.6 (2.6)	*52.9* (1.7)	*46.5* (1.5)	28.0 (4.6)	20.9 (4.0)	25.3 (2.6)	19.7 (2.2)	18.1 (1.8)	12.4 (1.4)	24.3 (3.6)	19.0 (3.1)	58.8 (1.8)	45.9 (1.8)
PIMNet [6]	*55.8* (15.1)	*48.5* (14.6)	*37.5* (2.4)	*30.1* (2.3)	49.7 (0.8)	42.2 (1.0)	32.5 (5.2)	23.1 (4.3)	*29.6* (2.7)	*22.2* (2.3)	*24.7* (1.6)	*16.8* (1.4)	31.2 (2.5)	*22.9* (2.2)	61.1 (0.9)	48.2 (0.9)
M2SFormer	58.8 (12.8)	50.8 (12.8)	38.5 (2.4)	31.3 (2.3)	58.4 (0.7)	50.1 (0.6)	42.4 (5.8)	32.4 (5.3)	32.6 (2.2)	24.9 (1.9)	24.9 (1.3)	16.8 (1.0)	35.0 (1.8)	27.4 (1.6)	69.1 (0.7)	56.9 (0.8)

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Quantitative Results

DIS25k Training Scheme Results

Method	Seen Domain		Unseen Domain													
	DIS25k [61]		CASIAv2 [53]		CASIAv1 [14]		Columbia [26]		IMD2020 [52]		CoMoFoD [62]		In the Wild [29]		MISD [30]	
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
UNet [58]	80.9 (13.2)	74.2 (14.1)	15.1 (6.4)	9.7 (4.6)	24.0 (0.4)	15.7 (0.3)	28.6 (1.1)	18.5 (0.8)	23.1 (0.1)	15.0 (0.1)	13.4 (0.7)	7.9 (0.5)	33.9 (0.8)	22.4 (0.5)	40.1 (1.3)	27.0 (1.1)
SegNet [5]	41.7 (8.6)	32.5 (7.4)	6.0 (3.1)	4.1 (2.3)	5.6 (0.8)	3.5 (0.5)	2.5 (0.5)	1.3 (0.3)	8.3 (0.6)	5.2 (0.4)	9.9 (1.2)	6.6 (0.8)	5.2 (0.4)	3.1 (0.3)	13.5 (1.0)	8.2 (0.6)
MantraNet [67]	69.6 (12.1)	59.1 (11.8)	12.7 (4.7)	7.6 (3.1)	18.1 (0.4)	11.0 (0.3)	26.5 (3.1)	16.3 (2.2)	18.6 (0.6)	11.3 (0.4)	11.9 (0.6)	6.8 (0.3)	26.2 (1.3)	16.0 (0.9)	31.9 (1.1)	20.1 (0.9)
RRUNet [7]	76.8 (11.3)	68.9 (11.5)	13.4 (6.6)	9.3 (5.1)	21.4 (0.9)	15.0 (0.7)	29.9 (2.0)	20.7 (1.6)	17.8 (0.7)	12.0 (0.5)	9.4 (2.1)	5.8 (1.5)	23.7 (1.2)	16.0 (0.9)	32.2 (1.6)	21.1 (1.2)
MT-SENet [77]	77.8 (12.7)	70.4 (13.4)	12.8 (5.7)	8.2 (4.1)	22.0 (0.7)	14.2 (0.5)	26.8 (1.6)	17.2 (1.2)	1.9 (0.2)	12.4 (0.2)	13.1 (1.0)	7.7 (0.6)	26.5 (0.9)	16.9 (0.6)	34.2 (1.2)	22.3 (0.5)
TransForensic [23]	83.4 (6.9)	76.2 (7.6)	21.2 (11.1)	16.4 (9.7)	30.4 (0.7)	23.9 (0.5)	33.0 (1.7)	23.4 (1.3)	27.0 (0.6)	19.8 (0.5)	18.1 (2.2)	12.1 (1.8)	31.0 (0.8)	22.3 (0.7)	45.7 (1.4)	33.2 (1.2)
MVSSNet [13]	70.0 (4.3)	60.1 (4.2)	19.8 (10.2)	14.4 (8.1)	30.0 (1.3)	22.6 (1.0)	35.7 (4.7)	26.4 (3.8)	24.7 (1.3)	17.4 (0.9)	17.5 (1.5)	11.3 (1.0)	30.7 (2.9)	22.0 (2.0)	47.7 (2.1)	34.8 (1.8)
FBINet [18]	81.6 (5.4)	74.7 (5.8)	17.9 (10.3)	13.8 (8.8)	21.0 (0.7)	15.9 (0.6)	13.9 (1.7)	8.7 (1.3)	21.6 (0.5)	15.4 (0.5)	15.0 (1.7)	9.9 (1.3)	17.9 (1.5)	12.3 (1.2)	32.4 (1.7)	21.4 (1.4)
SegNeXt [20]	72.9 (11.0)	63.7 (11.2)	13.4 (5.7)	8.4 (4.0)	23.7 (0.3)	15.3 (0.2)	35.5 (1.1)	24.1 (0.8)	21.5 (0.2)	13.8 (0.1)	13.6 (1.0)	8.1 (0.6)	*32.2* (0.5)	21.2 (0.4)	41.5 (0.9)	28.1 (0.8)
CFLNet [51]	79.8 (6.7)	71.5 (7.3)	20.2 (10.1)	15.0 (8.3)	29.0 (0.6)	22.2 (0.5)	30.0 (3.0)	20.9 (2.6)	26.1 (0.4)	18.5 (0.4)	17.9 (0.9)	11.6 (0.8)	29.8 (1.1)	20.9 (0.8)	*46.0* (1.5)	*33.4* (1.3)
EITLNet [19]	90.6 (3.9)	85.7 (4.1)	25.4 (13.7)	*21.3* (13.1)	*36.2* (0.3)	*31.1* (0.3)	31.6 (2.4)	24.0 (2.2)	*30.0* (0.4)	*23.8* (0.2)	19.9 (2.8)	14.2 (2.1)	29.2 (0.8)	23.1 (0.6)	39.0 (0.8)	27.2 (0.7)
PIMNet [6]	*88.8* (4.7)	*83.3* (5.0)	*25.4* (12.2)	20.6 (11.0)	34.5 (1.2)	28.1 (0.9)	38.6 (5.4)	29.6 (4.7)	29.7 (0.9)	22.7 (0.8)	*21.6* (0.2)	*14.9* (0.5)	33.4 (2.4)	25.0 (1.9)	42.5 (2.1)	30.2 (1.8)
M2SFormer	87.7 (3.2)	81.5 (3.2)	26.7 (14.3)	22.3 (13.6)	40.1 (1.3)	34.5 (1.1)	*37.3* (2.5)	*28.8* (2.0)	31.9 (0.8)	25.2 (0.6)	31.9 (0.6)	25.2 (0.4)	31.7 (1.1)	*24.4* (0.9)	43.2 (1.6)	30.7 (1.5)

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Quantitative Results

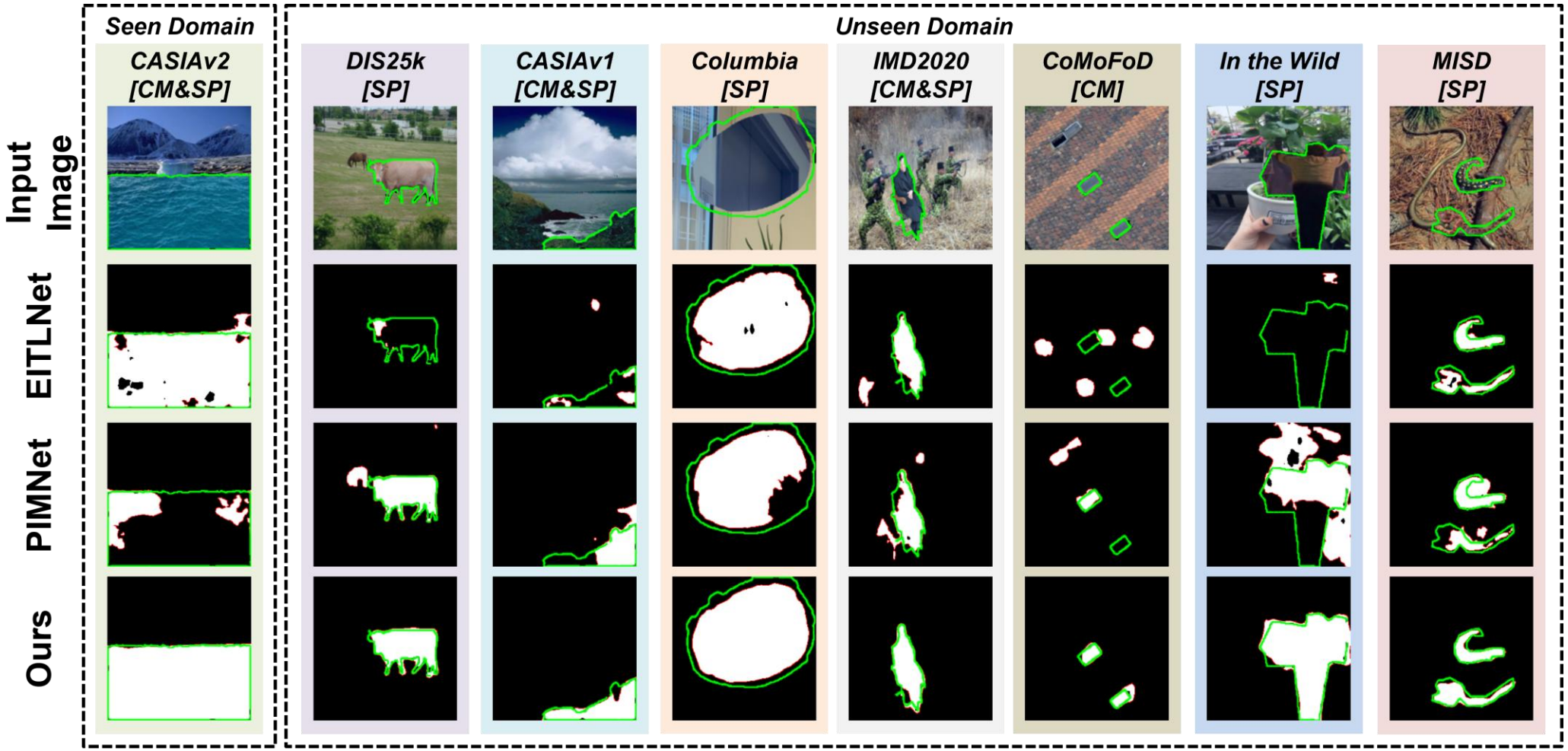
Corruption Robustness

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)

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

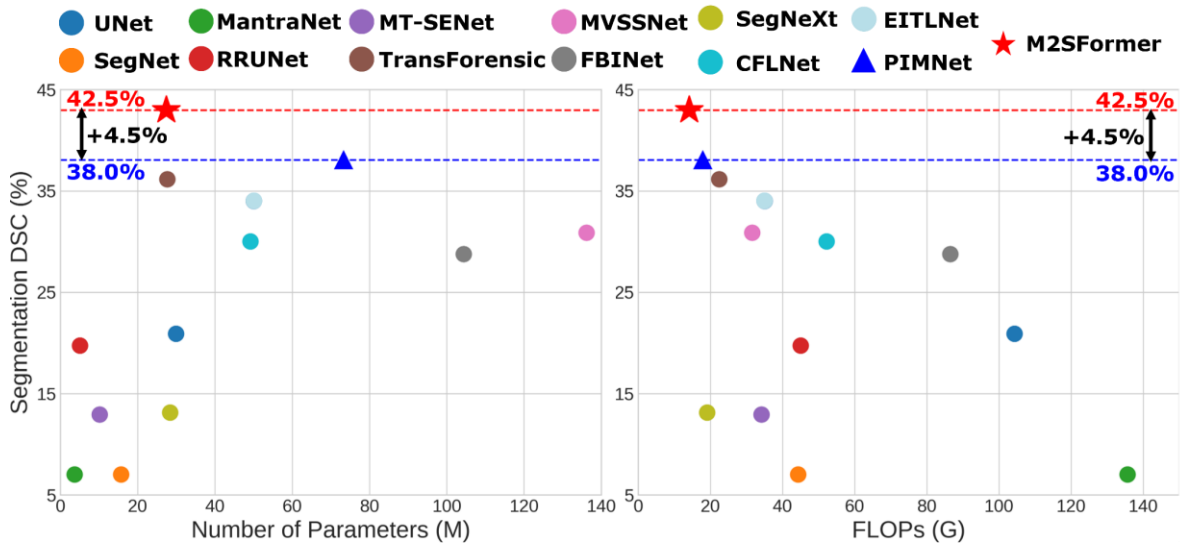
Qualitative Results



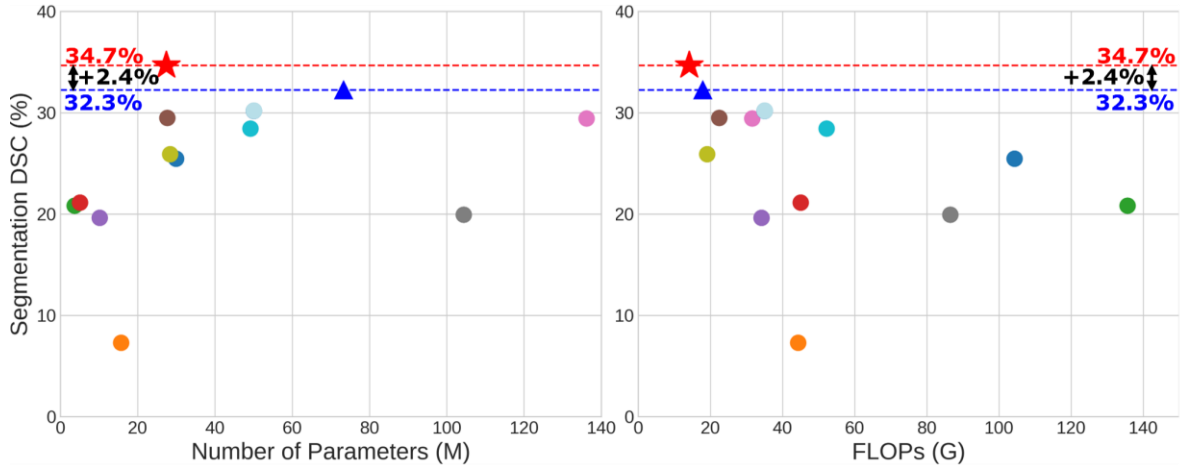
COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Efficiency Analysis



(a) CASIAv2 Training Scheme



(b) DIS25k Training Scheme

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Ablation Study

Setting Name	Spectrum S	Spectrum M	Scale S	Scale M	Seen DSC	Seen mIoU	Unseen DSC	Unseen mIoU	Param (M)	FLOPs (G)
S0	✓		✓		*56.3*	*49.1*	27.1	24.2	26.2MB	13.8GB
S1	✓			✓	55.5	48.6	33.6	28.1	27.4MB	14.2GB
S2		✓	✓		55.9	49.0	*36.1*	*31.3*	26.2MB	13.8GB
S3 (Ours)		✓		✓	**58.8**	**50.8**	**43.0**	**34.3**	27.4MB	14.2GB

Table 3. Ablation study of M2S attention block in skip connection on *Seen* (CASIAv2 [53]) and *Unseen* datasets (Other test datasets). "S" and "M" denote Single and Multi, respectively.

Setting Name	Seen DSC	Seen mIoU	Unseen DSC	Unseen mIoU	Param (M)	FLOPs (G)
No DGA	55.5	49.5	*32.3*	*26.1*	26.9MB	13.0GB
Simple DC + DGA	*56.1*	*50.2*	30.8	24.8	27.4MB	14.0GB
EADC + DGA	**58.8**	**50.8**	**43.0**	**34.3**	27.4MB	14.2GB

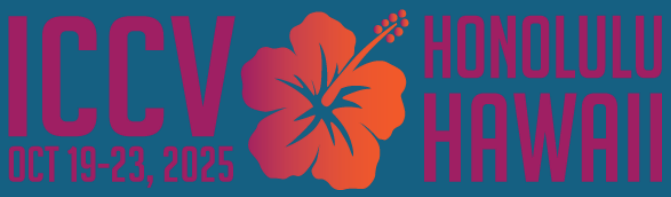
Table 4. Ablation study of DGA-based Transformer on *Seen* (CASIAv2 [53]) and *Unseen* datasets (Other test datasets). ECDC denotes Edge-Centric Difficulty Calculator.

Conclusion

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

ICCV HONOLULU HAWAII
OCT 19-23, 2025

- *Unified attention for robust localization*.
  - ✓ Fuse **multi-spectral channel attention** with **multi-scale spatial attention** in the **skip connections**.
  - ✓ Pair this with an **Edge-Aware DGA-based Transformer decoder**.
- *Difficulty-guided decoding.*
  - ✓ A curvature-based **global prior map** estimates sample difficulty (“easy”/“hard”) and converts it into a **text embedding** that gates channel attention during decoding
- *Cross-domain generalization & Efficiency.*
  - ✓ Across two training schemes (CASIAv2, DIS25k) and six external test sets, M2SFormer attains **state-of-the-art** pixel-level localization with stronger generalization to **unseen domains**.
  - ✓ The method balances accuracy and compute—**~27.4 M params** and **~14.2 GFLOPs**—while outperforming heavier baselines.



Thank you