Detecting AI-Generated Forgeries via Iterative Manifold Deviation Amplification

Supplementary Material

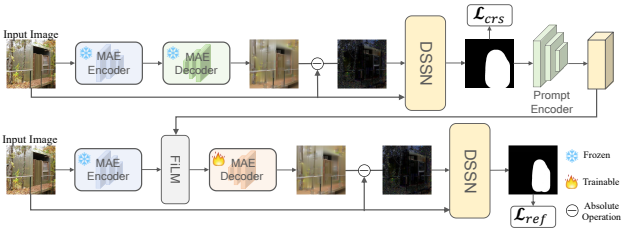

Figure 6. Detailed pipeline of IFA-Net.

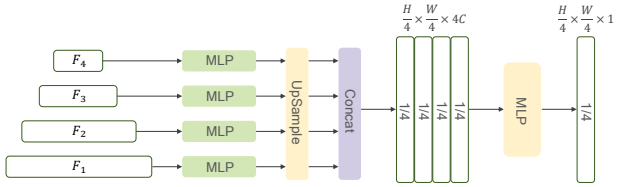

Figure 7. Architecture of the DSSN Decoder. The decoder aggregates multi-scale fused features F_1, F_2, F_3, F_4 to predict the final forgery mask.

7. Detailed Network Architecture

In this section, we provide a more detailed illustration of the proposed IFA-Net framework and the specific design of the Dual-Stream Segmentation Network (DSSN) decoder.

7.1. Overall Pipeline and Residual Computation

To supplement Figure 2 in the main paper, we present a detailed pipeline of IFA-Net in Figure 6.

As illustrated, the framework operates in two stages. A crucial design element is the explicit computation of the Residual Map (referred to as x_{rec} in the main paper). In both Stage 1 and Stage 2, the output of the MAE Decoder (the reconstructed image) is subtracted from the input image via an absolute difference operation to obtain the Residual Map (calculated as $|\mathbf{X} - \hat{\mathbf{X}}_{\text{MAE}}|$). This residual map serves as the artifact-rich input for the DSSN. In Stage 2, the Prompt Injection mechanism guides the trainable MAE Decoder to fail more significantly on forged regions, thereby amplifying the signals in the resulting Stage 2 Residual Map.

7.2. DSSN Decoder Architecture

The Dual-Stream Segmentation Network (DSSN) fuses semantic content and artifact features. Figure 7 details the decoder structure.

The inputs to the decoder, denoted as F_1, F_2, F_3, F_4, correspond to the fused features obtained from the Cross-Attention modules illustrated in Figure 2 of the main paper. These features represent the multi-scale integration of the semantic stream and the artifact stream at four hierarchical stages (scales 1/4, 1/8, 1/16, 1/32).

The decoder processes these features sequentially. First, each feature map F_i is passed through an MLP layer to unify the channel dimension. Next, the features are upsampled to a common resolution of $\frac{H}{4} \times \frac{W}{4}$. Following this, the upsampled features are concatenated along the channel dimension and fused via a linear projection. Finally, a prediction head maps the fused features to the binary mask.

8. Cross-Dataset Implementation Details

To robustly evaluate the generalization capability of IFA-Net, we adopt a rigorous cross-dataset evaluation protocol.

Pretraining. We first pretrain our model on the official training split of the OpenSDID dataset. This dataset contains images generated by Stable Diffusion 1.5. This stage allows the model to learn general realness priors and artifact patterns.

Evaluation on Unseen Generators. Without any further fine-tuning, the model pretrained on OpenSDID (SD1.5) is directly evaluated on the OpenSDID test sets generated by unseen models, including SD2.1, SD3, SDXL, and Flux.1. This setting tests the model's zero-shot transferability to advanced diffusion models.

Evaluation on Other Benchmarks. For the other six benchmarks (GIT10K, CocoGlide, Inpaint32K, IMD2020, NIST16, CASIA), we follow the official dataset splits if provided by the dataset authors. In cases where no official split is available, we randomly partition the dataset into training, validation, and testing sets with a ratio of 6:2:2.

Fine-tuning Strategy. When adapting the OpenSDID-pretrained model to these benchmarks, we employ a minimal fine-tuning strategy. Specifically, we fine-tune the model for only one epoch on the target training set. This constraint is imposed to prevent the model from overfitting to dataset-specific biases or memorizing specific forgery patterns. By limiting fine-tuning, the reported performance better reflects the model's generalized "realness" capability rather than its capacity to fit a specific distribution.

9. Additional Ablation Studies

In Equation (5) of the main paper, the total loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ref}} + \alpha\mathcal{L}_{\text{crs}}$, where α balances the supervision for the coarse mask relative to the refined mask. We conduct an ablation study to analyze the sensitivity of IFA-Net to α,

Table 3. Ablation on the balancing coefficient α. We evaluate the Average IoU and F1 scores on both GIT and TT benchmarks.

α	**GIT-AVG**		**TT-AVG**	
	IoU	F1	IoU	F1
0.0	0.615	0.702	0.410	0.525
0.1	0.684	0.765	0.495	0.612
0.2	0.742	0.821	0.551	0.672
0.3	0.771	0.849	0.578	0.701
0.4	0.775	0.852	0.583	0.705
0.5	**0.778**	**0.855**	**0.586**	**0.708**
0.6	0.774	0.851	0.581	0.703
0.7	0.769	0.847	0.575	0.698
0.8	0.762	0.840	0.568	0.692
0.9	0.755	0.831	0.560	0.685
1.0	0.748	0.825	0.554	0.678

varying it from 0.0 to 1.0 with a stride of 0.1. The results are reported in Table 3. As shown in the table, the model achieves optimal performance when $\alpha = 0.5$. Notably, when α is very small (e.g., $\alpha < 0.2$), the performance drops significantly. This validates that the coarse mask supervision is essential; without a reliable coarse mask to generate valid prompts, the subsequent amplification stage lacks direction. Conversely, when α is too large (e.g., $\alpha > 0.8$), the optimization over-prioritizes the coarse prediction, limiting the flexibility required for the Stage 2 amplification. The performance remains robust within the middle range $[0.4, 0.6]$, demonstrating that our method is relatively insensitive to small hyperparameter variations around the optimal point.