DiffusionPrint: Learning Generative Fingerprints for Diffusion-Based Inpainting Localization

Supplementary Material

7. Augmentation Strategies

Contrastive learning relies heavily on data augmentation to generate diverse, positive views of the same underlying instance. However, in the context of image forensics, augmentations must be chosen carefully to avoid destroying the delicate traces left by the generative process. In this section, we evaluate the impact of different augmentation strategies during the pretraining of the DiffusionPrint backbone. Table 3 reports the linear probing accuracy across four configurations: no augmentation, geometric transformations, high-pass filtering, and JPEG compression.

Table 3. Ablation on pretraining augmentations. All metrics denote overall linear probing accuracy. While geometric augmentations improve representation quality, applying high-pass filters or JPEG compression severely degrades the learned forensic signal.

Augmentation Strategy	Overall Accuracy (%)
None	84.90
Geometric	**86.35**
Geometric + JPEG	83.40
Geometric + HP Filter	64.10

As a baseline, training the encoder with no augmentations yields an accuracy of 84.90%. Using purely geometric augmentations (random crops and horizontal/vertical flips) provides spatial diversity and improves the accuracy to 86.35%. These spatial transformations help the model learn a better representation without altering the local pixel statistics produced by the generative models.

In traditional image forensics, high-pass (HP) filters are often used to suppress image semantics and isolate high-frequency camera noise. However, applying an HP filter (blocking normalized frequencies $f_{\text{norm}} \leq 0.1$) during pretraining drops the accuracy to 64.10%. This suggests that although low-frequency content contains semantic data, it still holds useful forensic information about the generative process. Filtering it out simply removes these important traces. We also tested applying random JPEG compression (Quality Factor between 70 and 95, with a 50% probability) alongside geometric augmentations. While JPEG augmentation is commonly used in downstream tasks to improve robustness, applying it during pretraining reduces the accuracy to 83.40%. This performance drop likely occurs because the compression step removes important forensic information that the encoder needs to effectively learn the generative fingerprint.

8. Lite Baseline Architecture.

In addition to the state-of-the-art frameworks, we evaluate a custom lightweight two-stream baseline (Lite Baseline) to isolate the effectiveness of the forensic modalities with a simpler fusion mechanism. The RGB stream utilizes an ImageNet-pretrained Mix Transformer encoder (MiT-B2) from the SegFormer architecture [56]. For the forensic stream, the extracted feature map (either Noiseprint++ or our frozen DiffusionPrint) is passed through a lightweight convolutional secondary encoder utilizing residual blocks to produce feature maps at four corresponding scales (64, 128, 320, and 512 channels).

Unlike the complex cross-attention mechanisms from the cmx architecture [60] used in TruFor and MMFusion, our Lite Baseline employs a straightforward multi-scale concatenation fusion: at each of the four scales, the RGB and forensic feature maps are concatenated along the channel dimension and reduced via a 1×1 convolution, Batch Normalization, and a ReLU activation. The fused multi-scale features are then passed into a standard SegFormer All-MLP decoder to produce the final localization map. This streamlined architecture significantly reduces the computational overhead, containing approximately 36M parameters—nearly half the size of the TruFor network.

9. Training Details

For the integration of DiffusionPrint into the TruFor [23] and MMFusion [48] frameworks, we retain their original architectural designs and training protocols, referring readers to the respective papers for exhaustive network details. The lite baseline was adapted from the TruFor implementation. Across all frameworks, input images are randomly cropped to 512×512 and trained for 100 epochs using an SGD optimizer with an initial learning rate of 0.005 and a momentum of 0.9. To match TruFor's original effective batch size of 18, we utilize a physical batch size of 9 coupled with 2 gradient accumulation steps. Similarly, for MMFusion, we maintain the original effective batch size of 24 by employing a physical batch size of 8 with 3 gradient accumulation steps. All framework-specific hyperparameters remain strictly as originally proposed. Prior to extracting the forensic feature maps, we apply identical data augmentations to the RGB inputs across all models: random resizing in the range $[0.5, 1.5]$ and JPEG compression with a quality factor uniformly sampled between 30 and 100.